

Interpreting Observed Action

Edited by

Susanne Biundo-Stephan¹, Hans Werner Guesgen²,
Joachim Hertzberg³, and Stephen R. Marsland⁴

1 Universität Ulm, DE

2 Massey University, NZ, h.w.guesgen@massey.ac.nz

3 Universität Osnabrück and DFKI RIC, Osnabrück, DE,
joachim.hertzberg@uos.de

4 Massey University, NZ, s.r.marsland@massey.ac.nz

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 12491 “Interpreting Observed Action”. The aim of the seminar was to get a coherent picture, which transcends the borders of applications and disciplines, of existing approaches and problems in interpreting observed action in semantic terms – primarily action by humans, but action by artificial agents may play some role, too. The seminar brought together, on the one hand, researchers from the different camps of AI, robotics, and knowledge-based systems who are working on the various aspects and purposes of interpreting observed action by humans, or occasionally, other agents; on the other hand, it added some researchers from cognitive science (psychology, neurosciences) working on human perception of behaviour and action. The main outcome of the seminar were a set of guidelines for setting up a workbench, which can be used to explore and test methods and techniques related to interpreting observed action.

Seminar 2.–6. December, 2012 – www.dagstuhl.de/12491

1998 ACM Subject Classification I.2 Artificial Intelligence

Keywords and phrases action, knowledge representation, plan recognition, symbol grounding, perception, behavior interpretation

Digital Object Identifier 10.4230/DagRep.2.12.1


1 Executive Summary

Susanne Biundo-Stephan

Hans Werner Guesgen

Joachim Hertzberg

Stephen R. Marsland

License  Creative Commons BY-NC-ND 3.0 Unported license
© Susanne Biundo-Stephan, Hans Werner Guesgen, Joachim Hertzberg, Stephen R. Marsland

For many applications of smart embedded software systems, the system should sense the footprint of a human or humans acting in the system’s environment, interpret the sensor data in terms of some semantic model about what the human is doing, and respond appropriately in real time. Examples of such applications include smart homes, human-machine or human-robot interaction, assistance, surveillance, and tutoring systems; given the current trend towards ambient intelligence, ubiquitous computing, and sensor networks, the number of systems in these categories can certainly be expected to rise in the next ten years or so.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Interpreting Observed Action, *Dagstuhl Reports*, Vol. 2, Issue 12, pp. 1–16

Editors: Susanne Biundo-Stephan, Hans Werner Guesgen, Joachim Hertzberg, and Stephen R. Marsland



DAGSTUHL
REPORTS

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The problem shares many features with classical object recognition and scene reconstruction from sensor data in terms of a static scene model. Interpreting in semantic terms sensor data from the environment has a long tradition in AI – arguably, it has been one of the original core problems put forth by AI’s founding fathers. However, the problem of interpreting observed action in the sense of this seminar differs in some aspects from what state-of-the-art AI or engineering approaches would allow to be tackled by routine:

- **Events in space-time** rather than static objects need to be characterized. This necessarily involves some representation and model of temporal and spatial data (e.g., the human put a saucepan *on* the cooker *and then* turned the cooker on).
- **Real-time processing** of the sensor data or percepts is required to keep track of what is happening. In fact, “real time” here is the pace of human action, i.e., relatively slow compared to CPU clock ticks. However, given a potentially rich stream of sensor data and a potentially large body of background knowledge, even this pace is demanding for the respective knowledge processing methods.
- **Willed human action**, be it planned, intended, or customary, is the domain of interpretation. In knowledge representation, this appears to be a relatively unexplored area, compared to, say, upper ontologies of household items, red wine, or pizza varieties.

Contemplating the three words that make up the title of this seminar (“interpret”, “observe”, and “act”), it becomes clear that there are a number of issues that need to be addressed in this context. Firstly, any interpretation is to some degree subjective and uses a particular repertoire of basic actions in its language. Secondly, an observation uses a particular type of sensor data and often is not possible without interpretation at the same time. Thirdly, there are issues around what actions are to be considered:

- Are only willed and physical action to be considered?
- Is avoidance an action?
- What constitutes an action in the first place?
- When does a particular action end?
- Is an unsuccessful action an action?

In summary, what precisely is observed action interpretation and what would be benchmark data for it?

To find an answer to this question, the participants of the seminar emerged themselves in a variety of activities: technical talks, working groups, plenary discussions, and a number of informal discussions. In the rest of this report, some of these activities and their results are discussed in more detail.

2 Table of Contents

Executive Summary

<i>Susanne Biundo-Stephan, Hans Werner Guesgen, Joachim Hertzberg, Stephen R. Marsland</i>	1
--	---

Overview of Talks

Recognizing Users' Intentions – A Key Competence of Companion-Systems <i>Susanne Biundo-Stephan</i>	5
How Do We Interact With the World? Objects, Spaces and Interactions <i>Martin V. Butz</i>	5
Learning Relational Event Models From Video <i>Krishna Sandeep Reddy Dubba</i>	6
STRANDS: Spatial-Temporal Representations and Activities for Cognitive Control in Long-Term Scenarios <i>Tom Duckett</i>	6
Visual Attention for Mobile Systems <i>Simone Frintrop</i>	7
Internal Simulations for Behaviour Recognition <i>Verena V. Hafner</i>	7
Mutual Understanding of Humans and Robots <i>Alexandra Kirsch</i>	8
Interpreting Observed Action in Dynamic Human-Robot Teams under Asymmetric Agency and Social Sentience <i>Geert-Jan M. Kruijff</i>	8
Model-free Behaviour Recognition <i>Stephen R. Marsland</i>	9
Activity Recognition with SCENIOR <i>Bernd Neumann</i>	9
Neural Mechanisms for the Analysis of Articulated Motion Sequences <i>Heiko Neumann</i>	10
Observing and Modeling the Embodiment of Attention <i>Lucas Paletta</i>	11
Hybrid planning and plan recognition <i>Bernd Schattenberg</i>	12
Can spatial partitioning help with interpreting observed action? <i>Sabine Timpf</i>	12
Towards Learning Activities From Kinect Data <i>Thomas Wiemann</i>	12
Qualitative Spatial Reasoning for Interpreting Action <i>Diedrich Wolter</i>	13
Working Groups	13

4 12491 – Interpreting Observed Action





Plenary Discussion 14

Participants 16

3 Overview of Talks

3.1 Recognizing Users' Intentions – A Key Competence of Companion-Systems

Susanne Biundo-Stephan (Universität Ulm, DE)





License     Creative Commons BY-NC-ND 3.0 Unported license
© Susanne Biundo-Stephan

Development of a Companion-Technology aims at enabling the realization of technical systems that provide their functionality in a completely individualized way. They adapt to a user's expertise, background, capabilities, and needs; furthermore, they take into account the current situation as well as the user's emotional state. The features that distinguish Companion-systems, namely individuality, adaptiveness, availability, cooperativeness, and trustworthiness, are realized by (the interplay of) cognitive processes. These include planning and decision making, interaction and dialog, and perception and recognition.

Intention recognition is crucial for Companion-systems. It serves three purposes: (1) monitor whether the user acts as expected; (2) detect and explain / interpret deviations from expected behavior; (3) initiate appropriate measures to avoid the break of communication. As Companion-systems are knowledge-based systems, intention recognition can rely upon various sources including background domain knowledge, interaction history, situation and action context. Intention recognition involves various cognitive levels. Elementary activities and emotional state of a user are recognized on the sensor data processing level. Having "perceived" basic actions of the planning level this way, higher-level plans of action can be identified to serve as hypotheses for the recognition of action strategies and goals that represent the user's intentions. Finally, predictions generated on the planning level are fed back into the sensor data processing to guide activity recognition.

3.2 How Do We Interact With the World? Objects, Spaces and Interactions

Martin V. Butz (Universität Tübingen, DE)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Martin V. Butz

Joint work of Butz, Martin V.; Ehrenfeld, S.; Herbort, O.

Main reference S. Ehrenfeld, M.V. Butz, "The modular modality frame model: Continuous body state estimation and plausibility-weighted information fusion," *Biological Cybernetics*, Vol. 107, Issue 1, pp. 61–82, Springer, 2012.

URL <http://dx.doi.org/10.1007/s00422-012-0526-2>

Various results are put forward that we do not really interact optimally with our environment – and particularly with objects. Other factors can rather easily influence our interaction kinematics and dynamics. Moreover, a model is presented in which such interactions can unfold and which integrates in a highly modularized manner body state estimations, sensory information, and sensorimotor predictions. Finally, by means of results from an eye-tracking experiment, it is shown how anticipations about object interactions guide our information search task-dependently. I conclude that we interact with the world in a highly anticipatory fashion, continuously integrating interaction knowledge and other biases as well as incoming sources of information in a weighted, integrative manner.

3.3 Learning Relational Event Models From Video

Krishna Sandeep Reddy Dubba (University of Leeds, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Krishna Sandeep Reddy Dubba

Main reference K.S.R. Dubba, A.G. Cohn, D.C. Hogg, “Learning Event Models From Complex Videos Using ILP,” in Proc. of the 19th European Conf. on Artificial Intelligence (ECAI’10), Frontiers in AI, Vol. 215, pp. 93–98, IOS Press, 2010.

Learning event models from videos has applications ranging from abnormal event detection to content based video retrieval. When multiple agents are involved in the events, characterizing events naturally suggests encoding interactions as relations. This can be realized by tracking the objects using computer vision algorithms and encoding the interactions using qualitative spatial and temporal relations. Learning event models from this kind of relational spatio-temporal data is particularly challenging because of the presence of multiple objects, uncertainty from the tracking and especially the time component as this increases the size of the relational data (the number of temporal relational facts is quadratically proportional to the number of intervals present).


Relational learning techniques such as Inductive Logic Programming (ILP) hold promise for building models from this kind of data, but have not been successfully applied to the very large datasets which result from video data. In this thesis, we present a novel supervised learning framework to learn relational event models from large video datasets (several million frames) using ILP. Efficiency is achieved via the learning from interpretations setting and using a typing system that exploits the type hierarchy of objects in a domain.

Positive and negative examples are extracted using domain experts’ minimal event annotations (termed deictic supervision) which are used for learning relational event models. These models can be used for recognizing events from unseen videos. If the input data is from sensors, it is prone to noise and to handle this, we present extensions to the original framework by integrating abduction as well as extending the framework based on Markov Logic Networks to obtain robust probabilistic models that improve the event recognition performance.

The experimental results on video data from two challenging real world domains (an airport domain which has events such as loading, unloading, passengerbridge parking etc. and a verbs domain which has verbs like exchange, pick-up etc.) suggest that the techniques are suitable to real world scenarios.

3.4 STRANDS: Spatial-Temporal Representations and Activities for Cognitive Control in Long-Term Scenarios

Tom Duckett (University of Lincoln, UA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Tom Duckett


Joint work of Hawes, Nick (STRANDS project Coordinator) and colleagues in the STRANDS consortium

“STRANDS” (Spatial-Temporal Representations and Activities for Cognitive Control in Long-Term Scenarios) is a new FP7 IP Project, which will run from April 2013 to March 2017, involving six academic institutes and two industrial partners across four European countries.

The project aims to enable mobile service robots to achieve robust and intelligent behaviour in human environments through adaptation to, and the exploitation of, long-term experience. Our approach is based on understanding 3D space and how it changes over time, from milliseconds to months. We will develop novel approaches to extract quantitative and qualitative spatio-temporal structure from sensor data gathered during months of autonomous operation. Extracted structure will include recurring geometric primitives, objects, people, and models of activity. We will also develop control mechanisms which exploit these structures to yield adaptive behaviour in highly demanding, real-world security and care scenarios.

3.5 Visual Attention for Mobile Systems


Simone Frintrop (Universität Bonn, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Simone Frintrop

Visual attention is one of the concepts of human perception that focuses the processing capabilities on the regions of a scene that are most promising. Such a mechanism is not only valuable for humans, but also for computer vision and robotic systems. Especially robots that act in an unknown, complex environment, have to prioritize which aspect of the sensory input to process first. Here in Dagstuhl, I presented an overview of our research on computationally modeling visual attention as well as some applications for intelligent vision system. For example, I introduced our work on saliency detection based on multivariate probability distributions and our current approach for detecting unknown objects in 3D scenes.

3.6 Internal Simulations for Behaviour Recognition

Verena V. Hafner (HU Berlin, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Verena V. Hafner

Joint work of Schillaci, G.; Lara, B.; Hafner, Verena V.


Main reference G. Schillaci, B. Lara, V.V. Hafner, "Internal Simulations for Behaviour Selection and Recognition," in Proc. of the 3rd Int'l Workshop on Human Behaviour Understanding (HBU'12), LNCs, Vol. 7559, pp. 148–160, Springer, 2012.

URL http://dx.doi.org/10.1007/978-3-642-34014-7_13

Humans are experts at recognising and identifying the behaviour of others. It is believed that they run internal simulations in order to simulate certain (sensorimotor) actions internally when the (visual) signal is noisy or delayed which is the case most of the times. In this study, we provide a computational model consisting of pairs of learned inverse and forward models on a humanoid robot. This allows the robot to choose its actions and recognise the reaching actions of a human.

3.7 Mutual Understanding of Humans and Robots

Alexandra Kirsch (Universität Tübingen, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Alexandra Kirsch

Joint work of Kirsch, Alexandra; Michael Karg; Christina Lichtenthäler; Thibault Kruse


When two partners interact, they have to understand the actions of the other. For a robot interacting with a human this means that it has to

- understand the situation and in particular the actions of the human partner,
- select appropriate actions and execute them in a legible way, and
- interpret the user’s reaction towards the robot actions as feedback.

I present examples from our work on all three aspects. In the area of plan recognition we try to identify everyday activities based on locations, durations of standing at the locations, and objects. These observations are relatively easy to obtain and can already lead to decent plan recognition and useful predictions, even without accurately recognizing single user actions. When choosing actions the robot must take into account the user’s needs and social conventions. In the area of robot navigation we have seen that typical robot navigation approaches can lead to illegible behavior. When explicitly taking into account the human approach direction and distance, a more natural robot behavior can be generated. There are currently no generally accepted metrics and test procedures for human-aware behavior. We attempt to develop objective measures that can be used to evaluate robot behavior, and have additionally the potential to be observed by the robot itself, so that the robot can improve its behavior based on implicit user feedback.

3.8 Interpreting Observed Action in Dynamic Human-Robot Teams under Asymmetric Agency and Social Sentience

Geert-Jan M. Kruijff (DFKI – Saarbrücken, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Geert-Jan M. Kruijff

Main reference H. Zender, M. Janicek, G.-J. Kruijff, “Situating Communication for Joint Activity in Human-Robot Teams,” in IEEE Intelligent Systems, Vol. 27, Issue 2, IEEE CS, 2012.

URL <http://dx.doi.org/10.1109/MIS.2012.8>

The talk considers human-robot teaming, particularly the aspect of how a robot can recognize, and decide how to act, when things go wrong. As they inevitably will. The talk introduces the notions of asymmetric agency and social sentience, to discuss how we could model robots as team members.

3.9 Model-free Behaviour Recognition

Stephen R. Marsland (Massey University, NZ)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Stephen R. Marsland

Joint work of Marsland, Stephen; Guesgen, Hans W.; the MUSE group

Main reference H.W. Guesgen, S. Marsland, (Eds.) “Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security,” ISBN 9781466636828, IGI Global, 2013.

URL <http://dx.doi.org/10.4018/978-1-4666-3682-8>

URL <http://www.igi-global.com/book/human-behavior-recognition-technologies/72160>

An overview of our approach to behaviour recognition in smart homes, including our unsupervised approach, which is based on the concept that behaviours are activities that are repeated, probably with minor variations.

3.10 Activity Recognition with SCENIOR

Bernd Neumann (Universität Hamburg, DE)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Bernd Neumann

Joint work of Bohlken, Wilfried; Hotz, Lothar; Koopmann, Patrick; Neumann, Bernd;

Main reference W. Bohlken, B. Neumann, L. Hotz, P. Koopmann, “Ontology-Based Realtime Activity Monitoring Using Beam Search,” in Proc. of the 8th Int’l Conf. on Computer Vision Systems, LNCS, Vol. 6962, pp. 112-121, Springer, 2011.

URL http://dx.doi.org/10.1007/978-3-642-23968-7_12

SCENIOR (SCENE Interpretation by Ontology-based Rules) is an implemented system for realtime recognition of multi-object activities in real-world scenarios. SCENIOR has been developed for monitoring aircraft turnarounds at Blagnac Airport in Toulouse, for example aircraft arrival preparation, unloading, loading, refuelling and other service operations. SCENIOR expects time-marked information about individual object locations as input and delivers activity descriptions as output.

Activity recognition is based on declarative models specified in the standardised ontology language OWL and the Semantic Web Rule Language SWRL. The recognition system is compiled automatically from this model base. Thus models can be modified, or new models can be added, without reprogramming the recognition process.

Input data can be obtained by several means, for example by object-centered GPS location transmission, or by visual tracking using cameras. In the prototypical implementation at Blagnac Airport, objects have been tracked by 6 cameras firmly installed at the border of the apron. The tracked object locations were used to generate elementary events relating the objects to fixed zones, such as “Loader-Positioned-In-Right-FWD-Loader-Zone” or “Tanker-Leaves-Left-Tanker-Zone”. These were transmitted as input to SCENIOR and interpreted in realtime as meaningful activities (or unrelated events).

Activity models are structured in a compositional hierarchy. This way, it is possible to recognise sub-activities contributing to a higher-level activity and generate predictions about the completion of a turnaround.

The system also comprises uncertainty management based on probability distributions for the temporal properties of all parts of a turnaround. This way, the system can cope with imperfect or partial data, and possible alternative interpretations can be given a preference rating.


Viewed as a support system for aircraft servicing, the following benefits can be expected:

1. The system gives a realtime account of completed service activities, thus providing progress control.
2. The system allows estimates about the completion of remaining activities, thus facilitating further scheduling.
3. The system also recognises unscheduled or unusual activities, and may thus provide security information.

Viewed as a technological framework, SCENIOR permits scene interpretation for diverse domains. When adapting SCENIOR to another domain, the main tasks are: (i) Developing sensor analysis up to primitive event recognition, and (ii) modelling higher-level concepts in an OWL ontology.

3.11 Neural Mechanisms for the Analysis of Articulated Motion Sequences

Heiko Neumann (Universität Ulm, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Heiko Neumann

Joint work of Neumann, Heiko; Layher, Georg; Giese, Martin A.

Main reference G. Layher, M.A. Giese, H. Neumann, “Learning representations for animated motion sequence and implied motion recognition,” in Proc. of the 22nd Int’l Conf. on Artificial Neural Networks – Part I (ICANN’12), LNCS, Vol. 7552, pp. 288–295, Springer, 2012.

The detection and categorization of articulated, or biological, motion is a crucial task underlying action recognition. Neural representations of perceived animate objects are built in STS (superior temporal sulcus) which is a region of convergent input from intermediate level form and motion representations in primate cortex. STS cell sub-populations are selectively responsive to specific action sequences. It is still largely unknown how and to which extent form and motion information contribute to the generation of representations specific to biological motion and what kind of mechanisms are involved in the learning processes.

A model architecture is proposed for the unsupervised learning of task specific articulated motion sequence representations. The processing builds upon two mainly separated pathways akin of the dorsal and the ventral streams in the visual cortex of primates. Along the model dorsal pathway image motion is processed and patterns of motion are represented [1, 4]. Respectively, contour and form information is processed and represented along the model ventral stream [2, 5]. Distinctive global level motion and form category representations are learned in independent pathways. Unsupervised Hebbian learning is employed to build such categorical representations which serve at input stage to the feedforward convergent processing at the level of model STS. How does the model automatically select significant motion patterns as well as meaningful static snapshot categories (keyposes) from video inputs? Such keyposes correspond to articulated postures which are particularly characteristic for a given motion sequence and thus decrease the ambiguity of a temporal prediction. It is shown how sequence selective representations are learned in STS by fusing form and motion input from the segregated bottom-up driving input streams [3]. We also emphasize the role of feedback signals propagated backwards along descending processing channels to make predictions about future input as anticipations generated by sequence-selective neurons. Network simulations demonstrate the computational capacity of the proposed model by reproducing several experimental findings from neurosciences and recent behavioral data.

References

- 1 C. Beck, H. Neumann. *Interactions of motion and form in visual cortex – A neural model*. Journal of Physiology Paris 104: 61–70, 2010
- 2 T. Hansen T, H. Neumann. *Neural mechanisms for the robust representation of junctions*. Neural Computation 16: 1013–1037, 2004.
- 3 G. Layher, M.A. Giese, H. Neumann. *Learning representations for animated motion sequence and implied motion recognition*. Proc. Int'l Conf. on Artificial Neural Networks, Part I, ICANN 2012, A.E.P. Villa et al. (Eds.), LNCS 7552, Springer, pp. 288–295, 2012
- 4 F. Raudies, E. Mingolla, H. Neumann. *A neural model for transparent motion processing*. Neural Computation 23: 2868–2914, 2011
- 5 U. Weidenbacher, H. Neumann. *Extraction of surface-related features in a recurrent model of V1-V2 interactions*. PLoS ONE 4(6): e5909, 2009

3.12 Observing and Modeling the Embodiment of Attention

Lucas Paletta (Joanneum Research – Graz, AT)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Lucas Paletta


Joint work of Paletta, Lucas; Santner, Katrin; Fritz, gerald; Mayer, Heinz

Main reference Visual recovery of saliency maps from human attention in 3D environments. Proc. IEEE International Conference on Robotics and Automation, ICRA 2013, Karlsruhe, Germany, May, 2013, in print.

Computational modeling of visual attention has recently been emerging as an important field of computer science and Artificial Intelligence. From human attention we know that many brain areas, including those processing motor signals, are involved in the computation of saliency as an indicator for which information a next action should consider. The concept of embodied attention understands attention processing as a meaningful system component within a perception-action cycle of autonomous systems where saliency computation should be operated according to the task at hand. Previous work (Paletta et al., 2005) developed a model for eye movements and belief aggregation for the task of object recognition where sequential attention strategies are adjusted in the frame of reinforcement learning. Furthermore, contextual rules (Perko et al., 2009) may prime the location of attention processing in the visual information. Current work targets at including physical actions such as body posture and position dynamics into the framework. In a first step, we extract ground truth data from human studies using eye tracking glasses and a tuned framework of SLAM (simultaneous localisation and mapping) that allows to map human gaze and integrated saliency measures directly onto the acquired three dimensional model of the environment, with high precision, with wearable interfaces that enable natural behaviours and without the use artificial markers (Paletta et al., 2013). Future work will use human ground truth to learn extended models of embodied attention from human behaviour.

3.13 Hybrid planning and plan recognition

Bernd Schattenberg (Universität Ulm, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Bernd Schattenberg

This presentation introduces plan recognition in the context of hybrid planning. The hybrid planning approach integrates the notion of action abstraction from hierarchical task network planning with the notion of means-end reasoning from partial-order causal-link planning into a common framework. We introduce the formal framework and show how search in the space of plan refinements generates solutions.

The talk also motivates central issues in recognizing plans from the point of view of hybrid planning for human users: plans provide context to the observed actions and hence any deviation of a user from a committed plan raises questions of whether the deviation compromises the overall causal structure wrt. the goals, whether the deviation can be interpreted as an ad-hoc (or: improvised) alternative task implementation, and the like.

3.14 Can spatial partitioning help with interpreting observed action?


Sabine Timpf (Universität Augsburg, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Sabine Timpf

In this talk I explore the notion of spatial partitioning for subdividing an activity such as navigating from a perron to another perron in a train station into several actions. Actions, essential actions and operations as subdivisions of an activity are introduced. The strategy for partitioning space is taken from earlier work on schematic geometry especially discussing the notion of a gateway as a place where one scene is left and another scene is entered. Wayfinding works by walking along a sequence of scenes through gateways. In order to implement this notion in an agent-based system, a change of perspective to an immersive, bottom-up one is necessary. The problem of implementing how a human agent would walk around an obstacle and the necessary parameters for this operation are discussed. The agent's movements can be interpreted as operations in the activity, subdividing the actions, but not exactly as a hierarchical subdivision.

3.15 Towards Learning Activities From Kinect Data


Thomas Wiemann (Universität Osnabrück, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Wiemann

In this talk we present an approach to recognize activities from Kinect data based on clustering in polygonal reconstructions of the sensed scene. The idea behind this research is to identify changes in a reference scene. If an activity is going on new objects will appear or disappear and a person's movement will cause shadows. The sequences of added clusters and detected shadows are logged over time. Our aim is to use machine learning techniques to identify activities in the logged data.

3.16 Qualitative Spatial Reasoning for Interpreting Action

Diedrich Wolter (Universität Bremen, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Diedrich Wolter

The claim of this work can be summarized as follows: qualitative representation and reasoning techniques provide good means to represent knowledge about actions and qualitative reasoning supports recognition and interpretation of observed actions.

Qualitative representations of space and time are acknowledged for their ability to capture cognitive concepts that underly spatial and temporal knowledge. With their according reasoning techniques, qualitative approaches provide a symbolic approach to representation and reasoning with spatial and temporal knowledge. Their primary feature is to bridge the gap between low-level data and conceptual knowledge. A drawback of qualitative approaches is that they are domain-dependent. Different applications require different representations and different reasoning algorithms. This can be a severe burden for application developers. To overcome this problem, we develop a versatile reasoning toolbox SparQ (<http://www.sfbtr8.uni-bremen.de/project/r3/sparq/>) that aims at making the various reasoning algorithms easily accessible to application developers.

More recently, qualitative approaches have been combined with expressive general logics. In case of combining a spatial representation with a logic, the term spatial logic has been coined. We use the term qualitative spatial logic to stress that a qualitative spatial representation is used. Our work revealed that the combination of a modal logics with a qualitative spatial representation is particularly attractive. Modal logics offer the expressivity to express temporal knowledge (linear temporal logic (LTL), for example) which allows purely spatial representation to be extended to represent spatio-temporal knowledge as required to represent actions. Good reasoning characteristics of the modal logic can be conveyed to the extended logic. Developing qualitative spatial logics has several interesting implications. Firstly, action representations based on qualitative spatial logics are comprehensible. Since the formulas are based on the qualitative concepts that relate to cognitive concepts, such formalism is well-suited for knowledge engineering. Secondly, suppose we are given a formula representing an action. Given observation data, one can approach action recognition as a model checking task. Sophisticated software tools are available that can easily cope with large data sets, making the overall approach efficient. Thirdly, this process can be generalized to reason about unknowns. For example, reasoning tools can be used to sensibly supplement missing information. Reasoning can be used to infer which pieces of information, if adjoined to the observations made, would allow an action to be identified. For example, by reasoning one can yield an interpretation stating that the observed actions suit the pattern of 'laying the dining table', given that an unidentified object X is plate. The ability to supplement missing pieces of information sensibly shows that qualitative spatial logics provide adequate means to tackle interpreting actions.

4 Working Groups

Most of Tuesday was dedicated to working groups, which were supposed to develop sets of criteria for benchmarks that can be used for research in formalisms, methods, techniques and systems related to interpreting observed action. The working groups took into consideration that there might be a need for different benchmarks for different sensor setups and that

the benchmarks should allow for different scenarios (or classes of scenarios), should include humans, robots, and animals, should be suitable for academic exercises as well as real-world applications, and should allow for comparing different approaches against each other.

The three working groups looked at three classes of benchmark scenarios: a plan recognition scenario (like helping users of technical systems or recognising that someone is having breakfast), a state monitoring scenario (like a robot waiter monitoring coarse-grained states of its customers or an electronic caretaker looking after an elderly person) and a scenario for activity recognition in dynamic environments (like road traffic or crowd dynamics as in airports, open-air concerts, and soccer stadiums).

The working group focussing on plan recognition first looked at what plan recognition means and came to the conclusion that this is an active rather than a passive process. The group then looked at characteristics of plans, which include causal structure, execution of plans, multiple plans, hierarchical structure, and the number of agents involved. Potential tasks in this context are identifying and anticipating activities, rating the normality of an activity, annotating video data streams, and generalisation and conceptualisation of activities.

The discussion in the working group on state monitoring focussed on getting a better understanding of what exactly a state is in the context of interpreting observed action. In this context, a state is predefined and specific, and often defined through a collection of cases. State monitoring might involve multiple indicators, and benchmarking might need to use artificial data.

The working group dealing with activity recognition in dynamic environments started with enumerating various relevant scenarios, which include road traffic and accident surveillance, autonomous vehicles and observation of the environment, soccer game observations, sea vessel observations, recognising abnormal behaviour, pedestrian/crowd behaviour recognition, collaborative activities (garbage collectors, rescue teams), and potentially gesture recognition. They emphasised that there is a difference between interpretation and recognition, where the latter (in contrast to the first) uses fixed sets of interpretations and learned concepts.

5 Plenary Discussion

The plenary discussion on the last day of the seminar focussed on two overarching questions that resulted from the working groups:

1. What are the structural dimensions of a benchmark scenario?
2. What should a benchmark website include?

One of the most important dimension of a benchmark scenario turned out to be the quality of the data. Are the data noisy and/or do they contain errors (e.g. produced by noisy channels)? Are data coming from different sources, and what are these sources (e.g. sensors, cameras, etc.)? Are the data complete or do they represent only partial information? Under which circumstances was the data obtained (e.g. night vs. day in the case of video data).

A second dimension revolves around issues related to the observed activities. Are the activities single activities or repetitive activities? Are we dealing with hierarchies of activities? Are multiple actors involved in the activity? Are actions performed in parallel? Is there diversity in the actions? Are there multiple scenes or just a single scene, and if the first is the case, is there diversity in the scenes?

Issues around time and space form another dimension of the benchmark scenarios. Are we dealing with a small or large scale space (in respect to the scene). Are there multiple

timescales, and what is the length of a typical episode? Are the data continuous or do they represent snapshots?

Last but not least we need to consider the complexity of the data as well as its availability. Does the benchmark provide various levels of complexity? Are the data real-world data (indoor or outdoor) or artificially generated data? Are data available for different domains?

As far as a potential benchmark website is concerned, there is a need for tools that enable users to upload new data, reports on results achieved with the data and experiences gained with them, software used to achieve the results and experiences, sets of use cases, and collections of success stories. It is also desirable to have a toolbox for manipulating data sets so that they can be inspected easily, transformed into different formats, etc.

In respect to the data sets, they should be available at different levels of granularity and abstraction, and potentially in different representations. The data sets should be provided with evaluation guidelines, annotations, and clear explanations of their aims. Metadata and ground truths should be available at least for some of the data sets.

Other aspects of the website include:

- Wiki and forum for discussing technical issues and results
- Tutorials for the material available
- Links to relevant conferences
- Repository with papers using the benchmarks

Overall, the value of a benchmark website for interpreting observed action seems to be without question. However, it is yet to be determined who is in charge of the website and what the timeframe is for setting it up.

Participants

- Sven Albrecht
Universität Osnabrück, DE
- Mehul Bhatt
Universität Bremen, DE
- Susanne Biundo-Stephan
Universität Ulm, DE
- Martin V. Butz
Universität Tübingen, DE
- Amedeo Cesta
ISTC-CNR – Rome, IT
- Krishna Sandeep Reddy
Dubba
University of Leeds, GB
- Tom Duckett
University of Lincoln, UA
- Frank Dylla
Universität Bremen, DE
- Simone Frintrop
Universität Bonn, DE
- Martin A. Giese
Universitätsklinikum
Tübingen, DE
- Hans Werner Guesgen
Massey University, NZ
- Verena V. Hafner
HU Berlin, DE
- Joachim Hertzberg
Universität Osnabrück, DE
- Alexandra Kirsch
Universität Tübingen, DE
- Geert-Jan M. Kruijff
DFKI – Saarbrücken, DE
- Stephen R. Marsland
Massey University, NZ
- Bernd Neumann
Universität Hamburg, DE
- Heiko Neumann
Universität Ulm, DE
- Lucas Paletta
Joanneum Research – Graz, AT
- Bernd Schattenberg
Universität Ulm, DE
- Aryana Tavanai
University of Leeds, GB
- Sabine Timpf
Universität Augsburg, DE
- Thomas Wiemann
Universität Osnabrück, DE
- Diedrich Wolter
Universität Bremen, DE

