09431 Abstracts Collection
From Form to Function
— Dagstuhl Seminar —

Darius Burschka[1], Heiner Deubel[2], Danica Kragic[3] and Markus Vincze[4]

[1] TU München, D
burschka@cs.tum.edu
[2] LMU München, D
deubel@psy.uni-muenchen.de
[3] KTH - Stockholm, S
danik@nada.kth.se
[4] TU Wien, A
vm@acin.tuwien.ac.at

Abstract. From October 18 to October 23, 2009 the Dagstuhl Seminar 09431 "From Form to Function" was held in Schloss Dagstuhl - Leibniz Center for Informatics. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Recognition of structure, form and shape, affordances, perception action loop, Grasping

09431 Summary – From Form to Function

At present we are on the verge of a new era when technical systems expand from typical industrial applications with pre-programmed, hard-wired behaviors into everyday life situations where they have to deal with complex and unpredictable events. The increasing demand for robotic applications in dynamic and unstructured environments is motivating the need for novel robot sensing and adaptable robot grasping abilities. The robot needs to cope with a wide variety of tasks and objects encountered in open environments. Since humans seem to have no difficulty to estimate a rough function of an object and to plan its grasping solely from the visual input, robot vision plays a key function in the perception of a manipulation system.

Our hypothesis is that the form and shape of objects is a key factor deciding upon actions that can be performed with an object. Psychophysical studies with humans confirm that affordance of grasping includes information about object orientation, size, shape/form, and specific grasping points. Affordances

Dagstuhl Seminar Proceedings 09431
From Form to Function
http://drops.dagstuhl.de/opus/volltexte/2010/2419

are discussed as one ingredient to close the loop from perception to potential actions.

The aim of this seminar is to bring together researchers from different fields related to the goal of advancing our understanding of human and machine perception of form and function. We set out to explore findings from different disciplines to build more comprehensive and complete models and methods. Neuroscientists and experimental psychologists will provide initial conceptual findings on the selective nature of sensor processing and on how action-relevant information is extracted. Cognitive scientists will tackle the modeling of knowledge of object function and task relations. Computer vision scientists are challenged to develop procedures to achieve context-driven attention and a targeted detection of relevant form features. All participants will profit from the ideas and findings in the related disciplines and contribute towards establishing a comprehensive understanding of brain and computing processes to extract object function from form features.

- Computer vision and perception needs to detect relevant features and structures to build up the shape/form of objects, to determine their orientation and size, and to define good grasping points. Currently, appearance has been successfully used for recognizing objects and codebooks of features assist in object categorization. Our goal is to move the data abstraction higher to define object function from the perception of edges, contours, surface properties, and other structural features, which still remains less explored. A main task of the workshop is to bring the key experts together to discuss how to advance the state of the art.
- Attention is the mechanism to enable fast and real-time responses in humans. Studies with humans show that grasping can be performed independent of object recognition. Hence it is timely to investigate how this direct link or affordances can be modeled and replicated for exploitation in robotic and cognitive systems.
- Prediction and the integration of bottom-up and top-down data flow is often discussed. Primate vision has been largely studied based on passively recording neuron functions when observing patterns (bottom-up stream). Only recently the importance of top-down triggers has been more closely shown. For example, 85from the retina but other brain areas including what is thought to be higher brain regions. Recent neuro-scientific findings state that predictions are a primary function of these connections. This indicates that the human brain uses predictions to focus attention, to exploit task and context knowledge, and hence scale an otherwise too wide space of inputs. For example, prediction indicates how a shape will be perceived when a certain action is executed on the target object. The task will be to identify what is the relevant information and how can it be computed in a machine vision system.
- Finally, humans seem to build up extensive knowledge about typical shapes and forms of whatever is seen in daily life. Seeing a partly occluded object often immediately triggers the respective model to complete the shape. Also

in grasping it has been found that the grasping point on the backside of an object is typically invisible but it is inferred from a symmetry assumption. The search for objects (say cups) is focused on horizontal surfaces and exploits knowledge about object category to look in a kitchen rather than in the garage. Recent work created first databases and ontologies to describe such knowledge, yet it remains open to fuse these developments with the results listed above.

- In summary, the seminar brought together scientists from disciplines such as computer science, neuroscience, robotics, developmental psychology, and cognitive science to further the knowledge how the perception of form relates to object function and how intention and task knowledge (and hence function) aids in the recognition of relevant objects.

Keywords: Recognition of structure, form and shape, affordances, perception action loop, Grasping

Joint work of: Burschka, Darius; Deubel, Heiner; Kragic, Danica; Vincze, Markus

Modeling Cognitive Shape Processing: Issues and Goals

Thomas Barkowsky (Universität Bremen, DE)

In mental spatial information processing there are numerous cognitive functions that are directly or indirectly related to shape aspects. Besides shape being a distinctive type of spatial knowledge that informs about specific perceptual characteristics of objects, shape, for instance, influences the control of spatial and non-spatial mental processes; it enables the efficient perception of features, objects, and entire scenes; it works as a classification instrument for entities and structures; it allows for the transformation of information between sensual modalities, etc.

In my presentation, I will - from cognitive modeling point of view - argue for shape being a universal mental category that goes far beyond the visual appearance of objects and that calls for a specific form of representation that enables the functional spectrum sketched out above:

How can shape information (in the general sense) be represented to serve the indicated purposes? What are the processes that operate on this representation?

Accuracy and flexibility in human action

Eli Brenner (VU University - Amsterdam, NL)

Our proficiency in manipulating objects with our hands illustrates that human movements are extremely accurate and flexible. I will argue that the accuracy is achieved by a combination of two factors.

The first is to choose a strategy that minimizes the impact of any lack of precision. This is where function, shape, orientation, surface texture, motion and so on, all have to be considered together. The second is to continuously adjust the ongoing movement on the basis of any new information that could be useful. Whether information is useful doesn't only depend on how reliable it is, but also on how quickly it can be used. The need to respond quickly makes it advantageous to use diverse sources of information separately and independently. I will illustrate these issues with examples from our research on grasping, placing and intercepting objects.

Keywords: Human, grasping

Object Properties as "Action Filter" for Manipulation

Darius Burschka (TU München, DE)

He geometric shape of an object specifies two functionalities of the interaction with it: the way it is grasped as a direct relation of its own and the gripper's geometry, and the way it is handled as a result of usage of its geometric properties (containment, supporting surfaces) in the context of the current scene. A geometric shape property defines a functionality, like for example support on a horizontal planar surface can define a chair or containment of liquids can define a "cup functionality". The function defines restriction on motion parameters of the handling task. In our approach, the object does not define any actions since it is not in the "knowledge scope" of an object to decide about it, but it defines constraints on the motion from the derived functionality of its geometry. The action is defined in a planning task based on a current mission and the object merely modifies/restricts the motion parameters from arbitrary motion in space to possible orientational and acceleration constraints.

We present a framework, that allows to segment objects in typical manipulation environments and that assigns a-priori knowledge about possible grasping strategies and motion restrictions through match of the object with an Atlas information based on object's geometry. An instantiated object indexes with its incomplete *geomoetric 2.5D/3D* information from the sensor into the Atlas. The stored information completes the geometry into a physical entity with complete shape, appearance, and handling information. The initial indexing for the unknown object is based on geometric 3D information while later a simplification though direct usage of the estimated appearance information in image space can be used.

Keywords: Knowledge representation, appearance- and geometry-based indexing

Joint Visual and Motor Learning for Object Modeling and Recognition

Barbara Caputo (IDIAP Research Institute - Martigny, CH)

Object recognition is a key problem of artificial vision; in robotics, it is strongly connected to that of grasping. In fact, there is so far no general solution to either problem. Traditionally, visual features are evaluated from camera images and statistical methods are then trained on huge visual datasets in order to obtain a robust object classifier. The knowledge so obtained is then used to choose a model to perform a grasping action.

Inspired, among others, by the neuroscientific framework of mirror neurons, we hereby propose to enhance the model of an object by adding to its visual features a probabilistic description of the grasps chosen by human subjects to grasp it.

Since in a standard setting the grasps are not directly available to the system, they must be reconstructed from the visual features, and then used to augment the recognition system's input space. We achieve this by building a map from visual to motor features, which we call a Visuo-Motor Map (VMM), practically enforced via regression on a human grasping database.

We experimentally show that such a technique improves the recognition rate of a standard object classifier: in case the original motor features are used, the improvement is dramatic, whereas when we reconstruct them via the VMM we still obtain a statistically significant improvement.

The proposed system can be seen as an instance of a general framework for multi-model learning, in which an artificial system learns to reconstruct active sensory patterns ("how do I grasp a mug") from passive ones (the visual appearance of a mug), and then to use the former together with the latter to improve its understanding of the environment.

Function from spatio-temporal interaction

Anthony G. Cohn (University of Leeds, GB)

In this talk I will present ongoing work at Leeds, in collaboration with Krishna Sridhar and David Hogg on inducing functional object categories from (qualitative) spatio-temporal descriptions of the participating objects. First we mine event classes by building a complete activity graph of the video, represented as a layered graph of qualitative spatial and temporal relations, and then induce event classes from this, in an unsupervised manner. Having formed these event classes, object categories can be formed by clustering those objects which take the same roles in a particular event. We have experimented with these techniques in two domains: a kitchen scenario, and aircraft turnovers.

Keywords: Functional object categories, unsupervised event learning, qualitative spatio-temporal relations

Full Paper:
http://www.comp.leeds.ac.uk/qsr/pub/ecai08.pdf

See also: Learning Functional Object-Categories from a Relational Spatio-Temporal Representation, M. Sridhar, A. G. Cohn and D. C. Hogg ECAI, edited by M. Ghallab, C. D. Spyropoulos, N. Fakotakis and N. M. Avouris, Frontiers in Artificial Intelligence and Applications, 178 , pp 606-610, IOS Press, (2008).

Attentional landscapes in the preparation of reaching and grasping movements

Heiner Deubel (LMU München, DE)

It is now well established that during the preparation and execution of goal-directed movements, attentional resources are biased towards the movement goal. Most of the previous work in primates has focused on rather simple movements, such as single saccades or manual reaches towards a single target. Here I review recent behavioural studies, mostly from our lab, on manual actions that require to consider more than a single spatial location in the planning of the response, such as movement sequences, grasping, and movements around obstacles.

The experimental results provide compelling evidence that the planning of a complex movement enacts the formation of an "attentional landscape" which tags all those locations in the visual lay-out that are relevant for the impending action. Despite the basically serial nature of movement generation, the findings imply a concurrent deployment of attentional resources on multiple locations, rather than the sequential processing of the action-relevant locations and features by a serial mechanism. Additionally, the data show that more attentional resources are dedicated to the location of the immediately following movement goal, and to those parts that require more precise motor control. Thus it seems that more than just selecting the action-relevant locations, the distribution of attentional weights also mirrors further, motor-related aspects such as temporal instancy, required accuracy, and the difficulty of the future action. The findings help to clarify how perceptual processing is bound by action preparation and also offer new directions for understanding human motor control. Some consequences of these attentional processes for grasping and manipulating objects will be discussed.

Keywords: Attention, reaching, grasping, humans

Joint work of: Deubel, Heiner; Baldauf, Daniel

Towards Human reach-to-grasp generalization strategies: studies based a probabilistic approach for grasp exploration & sensor data combination & shape representation

Jorge Manuel Miranda Dias (University of Coimbra, PT)

In this work we present a Bayesian framework to describe the mechanisms underlying the human strategies that define the appropriate characteristics of the reach-to-grasp movements to specific contexts, objects and how these strategies can be extended and replicated to other contexts and objects. The Bayesian framework is suitable for use information extracted from different sensors about the pose of the hand, fingers and head acquired by a magnetic tracker device, finger flexure data acquired by a data glove, as well as, visual data about eye gaze and saccade movements of the subject.

In this work we present a probabilistic framework that allows to combine multisensory information by using a probabilistic volumetric map where is possible have a probabilistic representation of object model acquired from grasp exploration. Using sensors of electromagnetic tracking motion system on the fingertips (thumb, index and middle) is possible to perform the object contour following acquiring the 3D data from the fingers movement around the object. The information from stereovision contributes for shape and three-dimensional representation of the object. The occupancy of each individual voxel in the map is assumed to be independent from the other voxels occupancy. The posteriori achieved from Bayes' rule is the probability distribution on the occupations percentage for each voxel.

In this work the sensor data registration and geometric calibration is addressed. From images frames is possible to compute the 3D points of the object and then after a calibration between the sensors we can work in the same reference frame to combine the information from these different sensors. It is used the object referential for its representation.

Keywords: Robot handling, Probabilistic Representation, Multisensor Fusion, Artificial Grasping

Joint work of: Dias, Jorge; Martins, Ricardo; Faria, Diego

The Evolution of Object Categorization and the Challenge of Shape Abstraction

Sven Dickinson (University of Toronto, CA)

Object recognition systems have their roots in the AI community, and originally addressed the problem of object categorization. These early systems, however, were limited by their inability to bridge the wide representational gap between low-level image features and high-level object models, hindered by the assumption of one-to-one correspondence between image and model features. Over the

next thirty years, the mainstream recognition community moved steadily in the direction of exemplar recognition, effectively narrowing this representational gap. The community is now returning to the categorization problem, and faces the same representational gap as its predecessors did. We review the evolution of object recognition systems and argue that bridging this representational gap requires the development of abstract representations for both shape and structure, along with algorithms for their recovery from an image. I will argue that recovering the abstract shape of real object in a cluttered scene remains the greatest challenge to object categorization, and is a prerequisite for function-based object recognition. I will review some of my group's efforts along these lines, including learning an abstract shape mdoel from examples, characterizing the abstract "shape" of a configuration, and matching features many-to-many.

Keywords: Image and shape abstraction, object categorization, many-to-many matching

Full Paper:
http://www.cs.toronto.edu/∼sven/Papers/cat2009.pdf

Infant Development of Form and Function Knowledge

Frank Guerin (University of Aberdeen, GB)

This talk looks at the development of form and function knowledge in infants, in order to give some ideas for a developmental approach which may benefit artificial intelligences. Infants seem to make use of a two-way interaction between form and function. The investigation of affordances is a strategy which infants use to explore the world, and to find visual features which determine particular affordances. This then gives knowledge of shape fragments which are salient to the infant, and in turn influences how the infant perceives the world. In terms of learning about affordances, there is a progression: The infant initially learns subjective knowledge about the relationship between his actions and the objects in the world, and subsequently progresses to more objective knowledge about the objects and their possible relations with others. An important function of an object is to use it as a tool to act on another object, hence it is important to be able to look at an object's form and understand the possibilities for action on another object. Through the exploration of these more complex affordances infants begin to acquire more general knowledge about the underlying physical processes. Understanding processes on this more abstract level is essential to allow the recognition of complex affordances, because they are composed of relationships among possible physical processes which could occur on parts of an object. Understanding the process by which this abstract knowledge develops is a major challenge; it is probably similar to the process of scientific discovery.

Keywords: Developmental Psychology, tool use, affordance, Piaget

Active Vision for Detecting, Fixating, Manipulating Objects and Learning of Human Actions

Danica Kragic (KTH - Stockholm, SE)

The ability to autonomously acquire new knowledge through interaction with the environment is one of the major research goals in the field of robotics. The knowledge can be acquired only if suitable perception-action capabilities are present. In other words, a robotic system has to be able to detect, attend to and manipulate objects in the environment.

In the first part of the talk, we present the results of our longterm work in the area of vision based sensing and control. The work on finding, attending, recognizing and manipulating objects in domestic environments is discussed. More precisely, we present a stereo based vision system framework where aspects of Top-down and Bottom-up attention and foveated attention are put into focus and demonstrate how the system can be utilized for object grasping.

The second part of the talk presents our work on the visual analysis of human manipulation actions which are of interest for e.g. human-robot interaction applications where a robot learns how to perform a task by watching a human. A method for classifying manipulation actions in the context of the objects manipulated, and classifying objects in the context of the actions used to manipulate them is presented. The action-object correlation over time is then modeled using conditional random fields. Experimental comparison shows improvement in classification rate when the action-object correlation is taken into account, compared to separate classification of manipulation actions and manipulated objects.

Adaptive Grasping based on 3D edge information

Norbert Kruger (University of Southern Denmark - Odense, DK)

In this work, we design two grasping strategies based on 3D edge information, one for grasping known and the other for grasping unknown objects. Both strategies involve learning based on past experience and can actually be combined. Experiments are made in a physical set-up allowing for exploration with a high degree of autonomy.

The first grasping strategy is based on a simple association between co-planar pairs of 3D contours and four different grasping actions. We can show that by using such a simple mechanism already a surprisingly good performance can be achieved. Moreover, the system is able to evaluate the success of the grasping attempts by haptic information and by that is able to build up an episodic memory of triplets containing (1) the visual features that have triggered the grasp, (2) the pose corresponding to the grasp as well as its haptic evaluation. Based on this experience, a neural network learns to predict the success of a potential grasping action which is used to select the executable action with the highest likelihood of success.

In the second grasping strategy, object specific grasps become coded. Starting with a simple mechanism similar to the first strategy (however applied to an abstracted learned object representation), a number of potential grasps are represented in a grasp density. For this, each potential grasp is transformed into a 6D kernel and the grasp density is coded as a weighted sum of these kernels. Based on this hypothesis density, the object specific grasps are tested while the robot is 'playing' with the object and the successful grasps become coded in an empirical grasp density. We can show that multiple cycles of this grasp learning leads to grasp densities that allow for grasps of significantly higher success likelihood than the original hypothesis density.

Keywords: Grasp learning

Joint work of: Kruger, Norbert; Piater, Justus ; Detry, Renaud; Kraft, Dirk

Perceiving Scenes

Justus Piater (University of Liège, BE)

Scene reconstruction from a single image is an underconstrained problem. How is it that humans appear to be able to do this effortlessly? To address this question from a robotics perspective, I advance the following conjectures:

- Mental scene reconstruction is an illusion. We do not actually construct a detailed internal model of a scene. Instead, we parse a scene into constituents and relations that are meaningful to us.
- Constituents and relations are learned together with their meanings or utilities. We add them to our repertoire to the extent that they help us explain / interact with the world. For example, having grasped different objects by their handle, the concept of a handle emerges.
- Such constituents and relations are a vehicle for generalization.
  For example, handles of unseen objects are perceived as potential grasp locations.
- The link between 2D perceptual arrays and 3D interactive behavior is a regression problem on interactively gathered data.

Thus, interaction is essential for forming representations that allow us to make sense of images. As a result, image analysis in terms of objects and shapes is inseparably linked to scene interpretation in terms of semantics and interaction.

Motivated by such reasoning, we are developing representations that integrate action-relevant parameters with visual features. These representations are hierarchically-structured Markov networks that can be learned from interactively-acquired data and that draw on established inference algorithms for robust performance.

I will outline how this research might lead to vision systems that can "understand" images in a deeper sense than labeling and segmentation, and talk about some of the major open challenges.

Possibilities (proto-affordances) Between Form and Function

Aaron Sloman (University of Birmingham, GB)

I shall discuss the need for an intelligent system, whether it is a robot, or some sort of digital companion equipped with a vision system, to include in its ontology a range of concepts that appear not to have been noticed by most researchers in robotics, vision, and human psychology.

These are concepts that lie between (a) concepts of "form", concerned with spatially located objects, object parts, features, and relationships and (b) concepts of affordances and functions, concerned with how things in the environment make possible or constrain actions that are possible for a perceiver and which can support or hinder the goals of the perceiver.

Those intermediate concepts are concerned with processes that *are* occurring and processes that *can* occur, and the causal relationships between physical structures/forms/configurations and the possibilities for and constraints on such processes, independently of whether they are processes involving anyone's actions or goals.

These intermediate concepts relate motions and constraints on motion to both geometric and topological structures in the environment and the kinds of 'stuff' of which things are composed, since, for example, rigid, flexible, and fluid stuffs support and constrain different sorts of motions.

They underlie affordance concepts. Attempts to study affordances without taking account of the intermediate concepts are bound to prove shallow and inadequate.

A longer abstract is here
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/between-form-and-function.html

A closely related paper from a previous Dagstuhl seminar on vision (Feb 2008) is available here: http://drops.dagstuhl.de/opus/volltexte/2008/1656

Keywords: Objects, surfaces, structure, matter, stuff, process, possible process, constraint, proto-affordance

Full Paper:
http://drops.dagstuhl.de/opus/volltexte/2008/1656

See also: Sloman, Aaron, Architectural and Representational Requirements for Seeing Processes, Proto-affordances and Affordances, Dagstuhl Seminar 'Logic and Probability for Scene Interpretation' Feb 2008

Object Recognition Through Reasoning About Functionality: A Survey of Related Work and Open Problems

Melanie Sutton (University of West Florida - Pensacola, US)

This survey summarizes the past two decades of research on object recognition through reasoning about functionality. Across the subfields of AI, computer vision and robotics, it is clear that as systems scale up to more complex scenes, reasoning that is functionality-based (as opposed to category-based) holds great promise. The most successful systems developed share commonalities for architectures and representations supporting multiple sensors and object domains. However, across these domains, new approaches are being used to ensure the systems continue to evolve in ways that demonstrate continued scalability, efficiency, accuracy, and ability to learn. As we look to the future of these fields, we examine the strengths and open research areas of a subset of representative systems addressing each of these characteristics and potential impacts on related disciplines.

Keywords: Function-based object recognition, computer vision, artificial intelligence, robotics

Joint work of: Sutton, Melanie; Stark, Louise

See also: This is an extension of this work: Bowyer, K.W., Sutton, M.A., & Stark, L. (2009, in press). Object recognition through reasoning about functionality: A survey of related work. Accepted to appear in Dickinson, S. (Editor), Object Categorization: Computer and Human Vision Perspectives, Cambridge University Press.

From Form to Function: Related Problems, Related Solutions?

Marko Tscherepanow (Universität Bielefeld, DE)

Technical systems such as robots are expected to work in everyday environments which are usually open-ended and non-stationary. Therefore, they have to deal with a great variety of tasks and objects that are not known in advance. The challenges of open environments, incremental learning, deriving knowledge of functions from shapes, and deciding on appropriate actions based on visual information are a topic of ongoing research in several contexts. I want to give two examples for related vision-based problems and how they can be tackled. The first one originates from bioinformatics or rather proteomics and the second one from social robotics.

Example 1: Protein Function Determination

A protein's three-dimensional structure or rather its shape directly determines its function, as it specifies the possible interactions of a protein with other molecules. Similar to objects observed by a human, the exact three-dimensional structures of the overwhelming majority of proteins are not known. But tagged proteins can be captured by a camera as well (when they are observed through a microscope). Then the function of proteins is derivable from corresponding images showing their distribution within a cell and their position with respect to possible interaction partners. Due to the extremely high number of proteins (more than 1 million in humans) and cell dynamics, the number of protein locations that can be observed by a microscope is not fixed. Consequently, incremental learning methods are required. I will present a method that enables the recognition of known locations as well as the detection and incorporation of distribution patterns showing unknown locations. In order to circumvent the stability-placticity dilemma occuring in on-line learning, it utilises neural networks based on the Adaptive Resonance Theory (ART) which was introduced as a model for information processing in the human brain.

Example 2: Facial Expression Imitation

Imitating the facial expressions of another person is a meaningful signal within interpersonal communication: Specific formations of facial components transmit specific information. In the context of social robotics, it has been shown that an anthropomorphic robot with the capability of imitating an interactant was perceived as responding more adequately to a specific social interaction. In contrast to well-known techniques, I will present a novel approach to facial expression imitation which does not require observed expressions to be assigned to a set of basic emotional expressions or preselected action units. Rather, arbitrary expressions are directly imitated solely based on camera images of the interactant's face. The mapping from images to motor commands is performed by regression models. This increases the number of displayable expressions and renders artificial agents more appropriate for interactions with humans. New information on the social meaning of observed facial expressions (new functions) could be easily integrated into a robotic system by recording facial images showing relevant emotional expressions and imitating them in an appropriate social context.

What does Attention have to do with it?

John Tsotsos (York University - Toronto, CA)

I will overview several projects from my lab that illustrate the value of attentive processes. In stark contrast to the way attention is viewed - almost universally in computer vision - finding a region of interest is only a minor component. The full breadth of visual attention will be overviewed and I will speculate on its role in determining function from form.

Function and Form

Markus Vincze (TU Wien, AT)

Children learn what to do with objects and to link objects to certain tasks. When grown up to adults, we have no difficulty to determine the use of an object and to plan an action with the object solely from the visual input. Even with objects formed by designers or artists, we rather quickly discover its purpose and function. So why does it turn out to be so difficult for a robot or cognitive system to bring me a cup or discover for a not seen object what it could be? The hypothesis is that the form and shape of objects is a key factor deciding upon actions that can be performed with the object. Psychophysical studies with humans confirm that the affordance of grasping includes information about object orientation, size, shape/form, and specific grasping points. Affordances are discussed as one ingredient to close the loop from intended action over perception back to potential actions. The intended task directs attention towards likely percepts, where scrutinising then decides upon continuation of search or task execution. Hence, the target form or shape is not only needed when scrutinising the attended region, it is also relevant to the attention process. Recent studies on human vision indicate that 3D cues are used for attention and seem to be always present in orientation.

Building Hierarchical Scene Representations from and for Human-Robot Interaction

Sven Wachsmuth (Universität Bielefeld, DE)

The ultimate goal of human-robot interaction is to enable the robot to seamlessly communicate with a human about natural everyday environments. While most research in this area is concentrating on the communicative cues itself, it is frequently underestimated that the success of communication heavily relies on the compatibility of the representations behind it. If a speaker refers to an object or scene structure that the robot does not perceive or perceives differently, the robot cannot react, appropriately. In my talk, I will discuss different approaches how relevant scene structure (like functional room areas, tables, shelfs, doors, etc.) can be learned from coarse shape representations and human-robot interaction. The techniques are based on the processing of depth data and include holistic representations, the analysis of scene changes over time, verbal descriptions, and mixed-initiative dialog.

Keywords: Scene analysis, integrating visual and verbal information, human-robot interaction

Model-free learning of object-action relations in manipulation

Florentin Woergoetter (Universität Göttingen, DE)

How do infants learn that certain actions and objects belong together (affordance learning)? How can they recognize, copy, and imitate so efficiently? These questions are intriguing because initially children have very little comprehension about the cause effect relations in their world. Thus, their learning must be model-free. This, however, carries the danger of trying to match each and everything (combinatorial explosion). Due to this problem, conventional approaches in computer vision and robotics so far heavily rely on prior (object and action) knowledge, e.g. using features (SIFT) to describe objects and action-primitives to group actions. In contrast to this, I would like to present a novel, efficient, and model-free approach for detecting spatiotemporal object-action relations, leading to both, action recognition and object categorization. The method is based on a real-time computer vision front-end for stereoscopic scene segmentation and the model-free tracking of the segments. Using the tracked segments, semantic scene graphs are extracted and used to find the characteristic main graphs of the action sequence via an exact graph-matching technique, thus providing an event table of the action scene. These event tables are representative for the different object action relations that occur during a given manipulation and they are invariant to irrelevant context. As a consequence, actions are recognized without requiring prior object knowledge and objects are categorized solely based on their exhibited role within an action sequence. Thus, this approach is grounded in the affordance principle and provides a way forward for trial and error learning of object-action relations through repeated experimentation. It may therefore be useful for recognition and categorization tasks for example in imitation learning in developmental and cognitive robotics.

Keywords: Affordance learning

Learning to predict object behaviour

Jeremy L. Wyatt (University of Birmingham, GB)

Robots might benefit from the ability to predict how objects interact with one another when they are manipulated. In this talk I will describe a framework for learning to predict object behaviour. I outline the desiderata for such a learner. Manipulation will be very simple, involving pushes of objects. I will show how we can learn a forward model from data, and speculate as to how a particular representation of shape will enable generalisation to novel objects.

Keywords: Robot learning, density estimation, prediction learning, forward models, internal models

Joint work of: Wyatt, Jeremy L.; Kopicki, Marek; Stolkin, Rustam; Zurek, Sebastian

Vision as Prediction

Michael Zillich (TU Wien, AT)

Vision as a process of reconstructing the 3D scene is an ill-posed problem, yet humans seem to do it effortlessly. Within an instant we can recognise a scene and take in its essential 3D structure. Still there are enough every-day cases where even for humans scene reconstruction becomes quite impossible (e.g. very low light situations). Humans can however still employ vision successfully in such cases - how do we do that?

We will argue that prediction plays a key role, not only in the form of attention but in a broader sense. While reconstruction generally is a difficult problem the inverse problem - prediction - is often very simple. Concretely, reconstructing the 3D scene from a single image is close to impossible whereas predicting the visual appearance of a 3D scene (i.e. creating an image) is a rather trivial problem solved long ago in computer graphics.

The basic idea of this talk then is to view the vision problem as a prediction problem. We formulate a general framework for probabilistically fusing multiple cues (visual and possibly non-visual), where each cue predicts some aspect of object shape. Predictions and actual observations are then used in a recursive filter to update an estimation of object shape.

Keywords: Vision, 3D shape estimation, cue integration