FP6-004381-MACS

MACS

Multi-sensory Autonomous Cognitive Systems Interacting with Dynamic Environments for Perceiving and Using Affordances

Instrument: Specifically Targeted Research Project (STReP)

Thematic Priority: 2.3.2.4 Cognitive Systems

D5.4.2 Prototypical software for representing and learning visual affordance support

Due date of deliverable: July 31, 2006
Actual submission date: September 14, 2006

Start date of project: September 1, 2004 Duration: 36 months

Joanneum Research (JR_DIB)

Revision: Version 1

Project co-funded by the European Commission within the Sixth Framework Programme (2002–2006)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

EU Project MACS

Deliverable D5.4.2

Prototypical software for representing and learning visual affordance support

Lucas Paletta, Gerald Fritz, Ralph Breithaupt, Erich Rome, Georg Dorffner

Number: **MACS/5/4.2**
WP: *5.4*
Status: *version 1*
Created at: *July 13, 2006*
Revised at:
Internal rev: *v3 – September 13, 2006*

FhG/AIS	Fraunhofer Institut für Intelligente Analyse- und Informationssysteme, Sankt Augustin, D
JR_DIB	Joanneum Research Graz, A
LiU-IDA	Linköpings Universitet, Linköping, S
METU-KOVAN	Middle East Technical University, Ankara, T
OFAI	Österreichische Studiengesellschaft für Kybernetik, Vienna, A

This research was partly funded by the European Commission's 6th Framework Programme IST Project MACS under contract/grant number FP6-004381. The Commission's support is gratefully acknowledged.

Corresponding author's address:

Dr. Lucas Paletta
Joanneum Research
Institute of Digital Image Processing
Computational Perception (CAPE)
Steyrergasse 9
A-8010 Graz, Austria

Fraunhofer Institut für Intelligente Analyse- und Informationssysteme
Schloss Birlinghoven
D-53754 Sankt Augustin
Germany

Tel.: +49 (0) 2241 14-2683
(Co-ordinator)

Contact:
Dr.-Ing. Erich Rome

Joanneum Research
Institute of Digital Image Processing
Computational Perception (CAPE)
Steyrergasse 9
A-8010 Graz
Austria

Tel.: +43 (0) 316 876-1769

Contact:
Dr. Lucas Paletta

Linköpings Universitet
Dept. of Computer and Info. Science
Linköping 581 83
Sweden

Tel.: +46 13 24 26 28

Contact:
Prof. Dr. Patrick Doherty

Middle East Technical University
Dept. of Computer Engineering
Inonu Bulvari
TR-06531 Ankara
Turkey

Tel.: +90 312 210 5539

Contact:
Prof. Dr. Erol Şahin

Österreichische Studiengesellschaft
für Kybernetik (ÖSGK)
Freyung 6
A-1010 Vienna
Austria

Tel.: +43 1 5336112 0

Contact:
Prof. Dr. Georg Dorffner

Contents

1 Introduction

This deliverable provides the representation and the learning for visual affordance support. In extension to the development in WP3 on affordance cueing and recognition (verification), this deliverable describes the representation and the learning for visual affordance support with respect to *a concrete affordance of the MACS affordance scenario*, i.e., "liftability". All aspects of representations of affordance based cues and regarding the learning methodology have been described in detail and published in [1].

The outline of this report is as follows. Section 2 describes the intimate relationship between WP3 (Perception) and WP5 (Learning) in the context of representation and learning. Section 3 relates the definitions in the published paper (published at SAB 2006, see Appendix) with recently outlined updates in the definitions of the affordance relation. Section 4 lists the software components that were developed for this methodology and that are stored on the MACS server *gibson*.

2 Affordance Perception and Learning

So far function based representations (e.g., [2]) were basically defined by the engineer, while it is particularly important for affordance based representations to learn the structure and the features themselves from experience. There are affordances that are explicitly innate to the agent through evolutionary development and there are affordances that have to be learned [3]. Learning chains of affordance driven actions can lead to learning new, more complex affordances. This can be done, e.g., by imitation, whereby it is reasonable to imitate goals and sub goals instead of actions [4]. In the context of the proposed framework on affordance based perception, learning should play a crucial role in determining predictive features. In contrast to previous work on functional feature and object representations [5],[2], we stress the fact that functional representations must necessarily contain purposive features, i.e., represent perceptual entities that refer to interaction patterns and thus must be selected from an existing pool of generic feature representations.

This demonstrates that perception of affordance must necessarily be based on learning mechanisms as outlined in detail in the published paper in the Appendix. Feature representations, in particular, purposive ones - as required for affordance based representations - are first schematically developed in WP3, then fed into learning methods selected and applied in WP5. Vice versa, the results of the learning sessions from WP5 are used as representations in WP3. Hence, an intimate relationship exists between these two work packages which makes, e.g., deliverable D5.4.2 a result of efforts originating in resources from WP5 as well as WP3.

3 Definitions

Since the published paper (Appendix) consists of text that was submitted as camera ready version in spring 2006, it does not yet contain the update of definitions done in agreement by all MACS partners during summer 2006, as it is described in Deliverable D2.2.2 [6]. Therefore, we give here a table of definitions proposed in the paper and the corresponding notions described in D2.2.2 (see Table 1).

Table 1: Correspondence of notions between SAB 2006 text and Deliverable D2.2.2.

SAB 2006 (Appendix)		Deliverable D2.2.2	
Notion	page	Corresponding Notion	page
Affordance action	3	Affordance behavior	4
Affordance recognition	5	Affordance hypothesis verification	4
Affordance cue	6	Affordance cue event	19

4 Software Components

We describe in Appendix the concept, the theoretical framework and the experiments using the software for the representation and the learning for visual affordance support. Currently we have implemented the following software components namely:

Curiosity drive: a visual attention component that can focus of attention on regions of interest and center the image on the selected most salient region (implemented in C++).

Segmentation: basic segmentation functionality using a watershed method and post-processing (energy merge) developed in C++.

Feature extraction: several modules for the processing of color blob detection (C++), local SIFT detectors and descriptors (C++), SIFT histogram descriptors (C++), a SIFT histogram classifier (C++), and detectors of test object and gripper elevation type (Matlab).

Tracking: a tracking component was developed in order to calculate trajectories of local object description in time (C++), additional code for visualization of the tracking results was written in Matlab.

Simulator of robot experiment: in order to evaluate numerous activities of the root and its actuators (gripper, magnet), we implemented a software component that simulates all possible actions and outcomes, including perceptual state, reward and action generation (implemented in Matlab).

Based on the current implementations we have done experiments described in the Appendix to test the decision tree based decision making on the simulated imagery for the affordance “lift-ability”.

5 State of Work and Outlook

Ongoing as well as future work is dedicated to the following issues namely:

Reinforcement Learning of Visual Affordance Support: This task is in cooperation with Task 3.3 (see Deliverable D3.3.1 [7]) which already described aspects of MDP (Markov Decision Process) based affordance representations which underlie reinforcement learning methodologies. Furthermore, perception itself would seldom

provide results and therefore state definitions under Markovian constraints.Therefore it is crucial to extend the theoretical framework on reinforcement learning for Partial Observable Markov Decision Processes (POMDP). Investigations on which framework to use out of several existing ones are still ongoing.

Real World Experiments: Currently we are building up a playground in a real world scenario and plan to perform extensive experiments with various test objects there. In addition, the scenario for learning the affordance "stack-ability" is planned and will be implemented in the near future. Additional computer vision functionalities will be included in order to enable robustness of affordance recognition under real world conditions (image noise, illumination conditions, etc.).

References

[1] Gerald Fritz, Lucas Paletta, Manish Kumar, Georg Dorffner, Ralph Breithaupt, and Erich Rome. Visual learning of affordance based cues. In S. Nolfi, G. Baldassarre, R. Calabretta, J. Hallam, D. Marocco, J-A. Meyer, and D. Parisi, editors, *From animals to animats 9: Proceedings of the Ninth International Conference on Simulation of Adaptive Behaviour (SAB)*, LNAI. Volume 4095., pages 52–64, Roma, Italy, 25–29 September 2006. Springer-Verlag, Berlin. in press.

[2] L. Stark and K. W. Bowyer. Function-based recognition for multiple object categories. *Image Understanding*, 59(10):1–21, 1994.

[3] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA, 1979.

[4] Martin G. Edwards, Glyn W. Humphreys, and Umberto Castiello. Motor facilitation following action observation: a behavioural study in prehensile action. In *Brain Cognition*, volume 53, pages 495–502, 2003.

[5] A. Stoytchev. Behavior-grounded representation of tool affordances. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 3071–3076, Barcelona, Spain, 2005. April.

[6] Erich Rome, Erol Şahin, Ralph Breithaupt, Jörg Irran, Florian Kintzler, Lucas Paletta, Maya Çakmak, Emre Uğur, Göktürk Üçoluk, Mehmet R. Doğar, Gerald Fritz, Georg Dorffner, Patrick Doherty, Mariusz Wzorek, Hartmut Surmann, and Christopher Lörken. Delevopment of an affordance-based control architecture. Deliverable MACS/2/2.2 v1, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Sankt Augustin, Germany, 2006.

[7] Lucas Paletta, Gerald Fritz, Manish Kumar, Joachim Hertzberg, and Frank Schönherr. Top-down and bottom-up symbol grounding. Deliverable MACS/3/1.1 v3, Joanneum Research Institute of Digital Image Processing Computational Perception (CAPE), Graz, Austria, 2004.

A Appendix

(see next pages)

Visual Learning of Affordance based Cues

Gerald Fritz[1], Lucas Paletta[1], Manish Kumar[1], Georg Dorffner[2],
[3]Ralph Breithaupt, and [3]Erich Rome

[1]JOANNEUM RESEARCH Forschungsgesellschaft mbH,
Institute of Digital Image Processing, Computational Perception Group,
Wastiangasse 6, Graz, Austria

[2]Österreichische Studiengesellschaft für Kybernetik,
Neural Computation and Robotics, Freyung 6, Vienna, Austria

[3]Fraunhofer Institute for Autonomous Intelligent Systems,
Robot Control Architectures, Schloss Birlinghoven, Sankt Augustin, Germany

Abstract. This work is about the relevance of Gibson's concept of affordances [1] for visual perception in interactive and autonomous robotic systems. In extension to existing functional views on visual feature representations, we identify the importance of *learning* in perceptual cueing for the anticipation of opportunities for interaction of robotic agents. We investigate how the originally defined representational concept for the perception of affordances - in terms of using either optical flow or heuristically determined 3D features of perceptual entities - should be generalized to using *arbitrary* visual feature representations. In this context we demonstrate the learning of causal relationships between visual cues and predictable interactions, using both 3D and 2D information. In addition, we emphasize a new framework for cueing *and* recognition of affordance-like visual entities that could play an important role in future robot control architectures. We argue that affordance-like perception should enable systems to react to environment stimuli both more efficient and autonomous, and provide a potential to plan on the basis of responses to more complex perceptual configurations. We verify the concept with a concrete implementation applying state-of-the-art visual descriptors and regions of interest that were extracted from a simulated robot scenario and prove that these features were successfully selected for their relevance in predicting opportunities of robot interaction.

1 Introduction

The concept of affordances has been coined by J.J. Gibson [1] in his seminal work on the ecological approach to visual perception: "*The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill … something that refers both to the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment.*" In the context of ecological perception, visual perception would enable agents to experience in a direct way the opportunities for action. However, Gibson

remained unclear about how this concept could be used in a technical system. Neisser [2] replied to Gibson's concept of direct perception with the notion of a perception-action cycle that shows the reciprocal relationship of the knowledge (i.e., a schema) about the environment directing exploration of the environment (i.e., action), which samples the information available for pick up in the environment, which then modifies the knowledge, and so on. This cycle describes how knowledge, perception, action, and the environment all interact in order to achieve goals.

Our work on affordance-like perception is in the context of technical, i.e., robotic systems, based on a notion of affordances that '*fulfill the purpose of efficient prediction of interaction opportunities*'. We extend Gibson's ecological approach under acknowledgment of Neisser's understanding that visual feature representation on various hierarchies of abstraction are mandatory to appropriately respond to environmental stimuli. We provide a refined concept of affordance perception by proposing (i) an interaction component (*affordance recognition*: recognizing relevant events in interaction via perceptual entities) and (ii) a predictive aspect (*affordance cueing*: predicting interaction via perceptual entities). This innovative conceptual step enables firstly to investigate the functional components of perception that make up affordance-based prediction, and secondly to lay a basis to identify the interrelation between predictive features and predicted event via machine learning technology.

The outline of this paper is as follows. Section 2 describes the relevance of affordance-like representations in robot perception and argues for the importance to learn the features of perceptual entities. Section 3 focuses on the issues of affordance recognition, in contrast to the predictive aspect of affordance-like representations in affordance cueing presented in Section 4. Section 5 illustrates the experimental results that strongly support the proposed hypothesis on the relevance of generalized features that must be learned for successful affordance-like perception in robot control systems. Section 6 concludes with an outlook on future work.

2 Affordance Perception and Learning

Affordance-like perception aims at supporting control schemata for perception-action processing in the context of rapid and simplified access to agent-environment interactions. In this Section we argue that previous research has not yet tackled the relevance of learning in cue selection, and present a framework on functional components that enables to identify relevant visual features.

2.1 Related Work

Previous research on affordance-like perception focused on heuristic definition of simple feature-function relations to facilitate sensor-motor associations in robotic agents. Human cognition embodies visual stimuli and motor interactions in common neural circuitry (Faillenot et al.[3]). Accordingly, the affordance-based context in spatio-temporal observations and sensor-motor behaviours has been outlined in a model of cortical involvement in grasping by Fagg and Arbib [4], highlighting the relevance of vision for motor interaction. Reaching and grasping involves visuomotor

coordination that benefits from an affordance-like mapping from visual to haptic perceptual categories (Wheeler et al.[5]). Within this context, the MIT humanoid robot Cog was involved in object poking and proding experiments that investigate the emergence of affordance categories to choose actions with the aim to make objects roll in a specific way (Fitzpatrick et al.[6]). The research of Stoytchev [7] analysed affordances on an object level, investigating new concepts of object-hood in a sense of how perceptions of objects are connected with visual events that arise from action consequences related to the object itself. Although this work innovatively demonstrated the relation between affordance triggers and meaningful robot behaviours, these experiments involve computer vision still on a low level, and do not consider complex sensor-motor representation of an agent interaction in less constrained, even natural environments. In addition, they are restricted to using vision rather than exploiting the multi-modal sensing that robots may perform. In the biologically motivated cognitive framework of Cos-Aguilera et al. [15], object based affordances are set in the context of motivation driven behaviour selection. In contrast to our work, they do not learn visual feature extraction in a purposive manner (Section 2.2) but rather match sensory input with stored object features in a classical sense [16] and then associate object identities with appropriate interaction patterns.

Affordance based visual object representations are per se function based representations. In contrast to classical object representations, functional object representations (Stark and Bowyer [8], Rivlin et al. [9]) use a set of primitives (relative orientation, stability, proximity, etc.) that define specific functional properties, essentially containing face and vertex information. These primitives are subsumed to define surfaces and from the functional properties, such as '*is sit-able*' or '*provides stable support*'. Bogoni and Bajcsy [10] have extended this representation from an active perception perspective, relating observability to interaction with the object, understanding functionality as the applicability of an object for the fulfillment of some purpose. However, so far function based representations were basically defined by the engineer, while it is particularly important for affordance based representations to *learn* the structure and the features themselves *from experience* (Section 4).

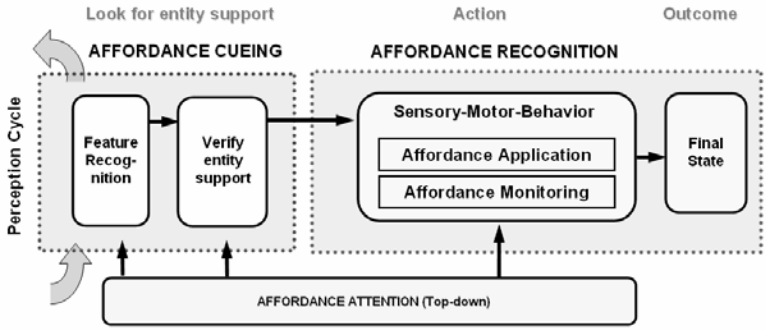

Fig. 1. Concept of affordance perception, depicting the key components of affordance cueing and recognition embedded within an agent's perception-action cycle (most left). While affordance cueing (left) provides a prediction on future opportunities of interaction on the basis of feature interpretation, affordance recognition (right) identifies the convergence of a perceptual patterns in a sensory-motor behavior towards the outcome of the overall process.

2.2 Predictive Features in Affordance based Perception

Fig. 1 depicts the innovative concept of feature based affordance perception presented in this paper. We identify first the functional component of affordance recognition, i.e., the recognition of the affordance related visual event that characterizes a relevant interaction, e.g., the capability of lifting (*lift-ability*) an object using an appropriate robotic actuator. The recognition of this event should be performed in identifying a process of evaluating spatio-temporal information that leads to a final state. This final state should be unique in perceptual feature/state space, i.e., it should be characterized by the observation of specific feature attributes that are abstracted from the stream of sensory-motor information.

The second functional component of affordance cueing encompasses the key idea on affordance based perception, i.e., the prediction aspect on estimating the opportunity for interaction from the incoming sensory processing stream. In particular, this component is embedded in the perception-action cycle of the robotic agent. The agent is receiving sensory information in order to build upon arbitrary levels of feature abstractions, for the purpose of recognition of perceptual entities. In contrast to classical feature and object recognition, this kind of recognition is *purposive* in the sense of selecting exactly those features that efficiently support the evaluation of identifying an affordance, i.e., the perceptual entities that possess the capability to predict an event of affordance recognition in the feature time series that is immediately following the cueing stage of affordance based perception. The outcome of affordance cueing is in general a probability distribution P_A on all possible affordances (Section 4.1), providing evidence for a most confident affordance cue by delivering a hypothesis that favors the future occurrence of a particular affordance recognition event. This cue is functional in the sense of *associating* to the related feature representation a specific *utility* with respect to the capabilities of the agent and the opportunities provided by the environment, thus representing *predictive features* in the affordance based perception system.

The relevance of attention in affordance based perception has first been mentioned by developmental psychologist E.J. Gibson [11] who recognized that attention strategies are learned by the early infant to purposively select relevant stimuli and processes in interaction with the environment. In this context we propose to understand affordance cues and affordance hypotheses as fundamental part in human attentive perception, claiming that – in analogy – purposive, affordance based attention could play a similar role in machine perception as well.

There are affordances that are explicitly innate to the agent through evolutionary development and there are affordances that have to be learned [1]. Learning chains of affordance driven actions can lead to learning new, more complex affordances. This can be done, e.g., by imitation, whereby it is reasonable to imitate goals and sub goals instead of actions [12]. In the context of the proposed framework on affordance based perception (Fig. 1), learning should play a crucial role in determining predictive features. In contrast to previous work on functional feature and object representations [8, 9], we stress the fact that functional representations must necessarily contain *purposive features*, i.e., represent perceptual entities that refer to interaction patterns and thus must be selected from an existing pool of generic feature representations.

Feature selection (and, in a more general sense, feature extraction) must be performed in a machine learning process and therefore avoid heuristic engineering which is always rooted in a human kind understanding of the underlying process, a

methodology which is necessarily both, firstly, error prone due to failing insight into statistical dependencies and, secondly, highly impractical for autonomous mobile systems. Our work highlights the process of learning visual predictive cues which to our understanding represents one of the key innovative issues in autonomous learning for affordance based perception.

3 Affordance Recognition

By affordance recognition we particularly refer to the process of identifying the relevant interaction events from perception that actually 'motivate' an agent to develop/learn perceptual cues for early prediction. In early infant development, the monitoring of affordances such as '*grasp-ability*' of objects or '*pass-ability*' of terrain [1] must be crucial to obtain an early as possible classification of the environment so that interaction behaviors can be initiated as fast and as robust as possible. In analogy, autonomous robotic systems should possess a high degree of flexibility and therefore be capable of perceiving affordances and therefore select appropriate functional modules as early as possible with respect to the goals of the robotic system. In this sense, goals and affordances are intimately related and make up a fully purposive perception system.

Fig. 2 illustrates the various stages within the affordance based perception process, in particular affordance recognition, for the example of the affordance '*fill-able*' in the context of the opportunities for interaction with a coffee cup.

(a) (b) (c)

Fig. 2. Affordance recognition (Section 3) in affordance based perception for the example of the affordance *fill-able* with respect to the impact of selecting appropriate features. The seemingly simple interaction of *filling up a coffee cup* can be partitioned into various stages in affordance based perception, such as, (a) affordance cueing by predictive features that refer to a *fill-able* object, (b) identifying perceptual entities that represent the process of the affordance related interaction (e.g., flow of coffee), and (c) recognizing the final state by detecting perceptual entities that represent the outcome of interaction (e.g., level of coffee in cup).

Fig. 2(a) schematically illustrates the detection of perceptual entities that would provide affordance cues in terms of verifying the occurrence of a cup that is related to the prediction of being *fill-able* in general. Fig. 2(b) shows in analogy entities that would underlie the process of interaction of an agent with the cup by actually filling it up. Finally, Fig. 2(c) represents the entities corresponding to the final state of the interaction with the outcome of a successfully filled coffee cup. These figures

illustrate that affordance cueing and affordance recognition must be conceptually separated and would involve different perceptual entities in general. While affordance *recognition* actually involves the recognition of the interaction process and its associated final state, affordance *cueing* will be solely determined by the capability to reliably predict this future event in a statistical sense.

4 Visual Cueing of Affordances

4.1 Feature based Cueing for the Prediction of Affordances

Early awareness of opportunities for interaction is highly relevant for autonomous robotic systems. Visual features are among the ones among multiple modalities from sensory processing that operate perception via optical rays and therefore support early awareness from rather remote locations. Although the necessity of affordance perception from 3D information recovery, such as optical flow, has been stressed in previous work [1], we do not restrict ourselves to any specific cue modality and intend to generalize towards the use of arbitrary features that can be derived from visual information, restricting only on the constraint that they enable reliable prediction of the opportunity for interaction processes from an early point in time.

The outcome of the affordance cueing system is in general expected to be – given a perceptual entity in the form of a multimodal feature vector - a probability distribution over affordance hypotheses,

$$P_A = P(A \mid F_t), \tag{Eq. 1}$$

with affordance hypothesis set A, and feature vector F_t at time t. It is then appropriate to select an affordance hypothesis $A_{max}(P_A(\cdot)=P_{Amax}(\cdot))$, with Maximum A Posteriori (MAP) confidence support for further processing.

From the viewpoint of a technical system using computer vision for digital image interpretation, we particularly think that complex features, e.g., local descriptors, such as the Scale Invariant Feature Transform (SIFT [13]), could support well the construction of higher levels of abstraction in visual feature representations. SIFT features are derived from local gradient patterns, and provide rotation, translation and – to some degree – viewpoint and illumination tolerant recognition of local visual information, and are therefore well suited for application in real world scenarios for autonomous robotic systems. Among other cues, such as color, shape, and 3D information, we are therefore interested to investigate the *benefit of using visual 2D* patterns for their use in affordance cueing.

Fig. 3 shows the application of local (SIFT) descriptors for the characterization of regions of interest in the field of view. For this purpose, we first segment the color based visual information within the image, and then associate integrated descriptor responses sampled within the regions to the region feature vector. The integration is performed via a histogram on SIFT descriptors that are labeled with ‘rectangular’ (a) and ‘circular’ (b) attributes, respectively. The labeling is derived from a k-means based unsupervised clustering over all descriptors sampled in the experiments, then by selecting cluster prototypes (centers) that are relevant for the characterization of

corresponding rectangular/circular shaped regions, and finally by determining histograms of relevant cluster prototypes that are typical in a supervised learning step (using a C4.5 decision tree [14]).

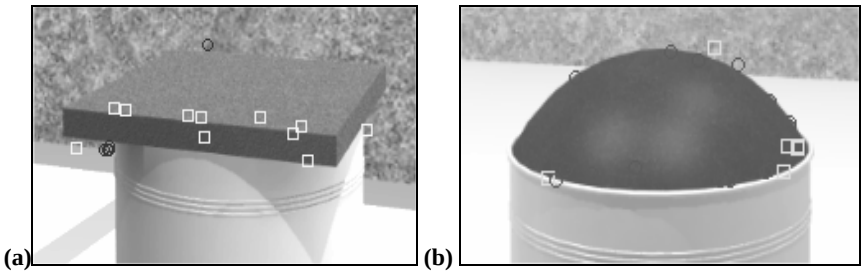

Fig. 3. Categories of local descriptor classes supporting affordance cueing. Classes of SIFT descriptors [13] occurring on (a) *rectangular* (favored by descriptors represented by squares) and (b) *circular* (favored by descriptors represented by circles) region boundaries, respectively. It should be noted that the descriptor classes support the classification of segmented regions. These classes are mandatory to discern affordance cues from 2D features.

Fig. 4 shows a sample cue-feature value matrix (in the context of the experiments, see Section V) that visualizes dependencies between feature attributes of the region information and a potential association to results of the affordance recognition process. We can easily see that the SIFT category information (*rectangular=R* and *circular=C* region characterization) together with a geometric feature (*top=T* region, i.e., representing a region that is located on top of another region) provides the discriminative feature that would allow to predict the future outcome (e.g., *lift-able*/*non lift-able*) of the affordance recognizer. The latter therefore represents the identification of the affordance and thereby the nature of the interaction process (and its final state) itself.

colour	G	R	M	R	Y	B	Bl	Gr
SIFT category	R	R	C	C	R	R	R	N
shape L/W	L	L	L	L	P	P	P	L
T/B	T	T	T	T	B	B	B	N
LIFTABLE	**Y**	**Y**	**N**	**N**	**Y**	**Y**	**N**	**N**
NOT LIFTABLE	**N**	**N**	**Y**	**Y**	**Y**	**Y**	**Y**	**N**

Fig. 4. Cue-feature value matrix depicting attribute values of 2D features (color G/green, R/red, M/magenta, etc., or SIFT category R/rectangular, C/circular, etc.)) and interaction results (left column) in dependence on various types of visual regions (top row). From this we conclude a suitable feature value configuration (i.e., SIFT categories to discriminate *lift-able*/*non lift-able* predictions) to support the hypothesis on *lift-able* object information.

4.2 Learning of Relevant Feature Cues from Decision Trees

The importance of machine learning methodologies for the selection of affordance relevant features has already been argued in Section 2.2. The key idea about our idea of applying learning for feature selection is based on the characterization of extracted perceptual entities, i.e., *segmented regions* in the image, via a feature vector representation. Each region that would be part of the final state within the affordance recognition process can be labeled with the corresponding affordance classifications. The regions can be back-tracked using standard visual tracking functionality to earlier stages in the affordance perception process. The classification label together with the feature attributed vectors of the region characterization build up a training set that can be input to a supervised machine learning methodology (using a C4.5 DT [14]).

5 Experimental Results

The experiments were performed in a simulator environment with the purpose of providing a proof of concept of successful learning of predictive 2D affordance cues, and characterizing affordance recognition processes.

The scenario for the experiments (Fig. 5) encompassed a mobile robotic system (Kurt2, Fraunhofer AIS, Germany), equipped with a camera stereo pair and a magnetizing effector, and some can-like objects with various top surfaces, colors and shapes. The purpose of the magnetizing effector was to prove the nature of the individual objects by lowering its rope-end effector down to the top surface of the object, trying to magnetize the object (only the body, *not* the top surface of the can are magnetizable) and then to lift the object. Test objects with well magnetizable geometry (with slab like top surfaces, in contrast to those with spherical top surface) are subject to a lifting interaction, while the others were not able to be lifted from the ground. This interaction process was visualized for several test objects and sampled in a sequence of 250 image frames. These image frames were referenced with multimodal sensor information (e.g., size of magnetizing and motor current of the robot, respectively).

5.1 Simulation

The scenario is split up into two phases (a) a *cueing phase*, i.e., the robot is moving to the object, and (b) a *recognition phase*, i.e., the robot tries to lift an object like shown in Fig. 5. In both phases parts of the objects are described by their regions and any region has different features like color, center of mass, top/bottom location and the shape description (rectangular, circular) already described in Section IV. Those features are extracted from the robot camera imagery. Additional information, such as, effector position are provided by the robot. Regions are the entities used in the experiment, no explicit object model is generated for the can-like objects.

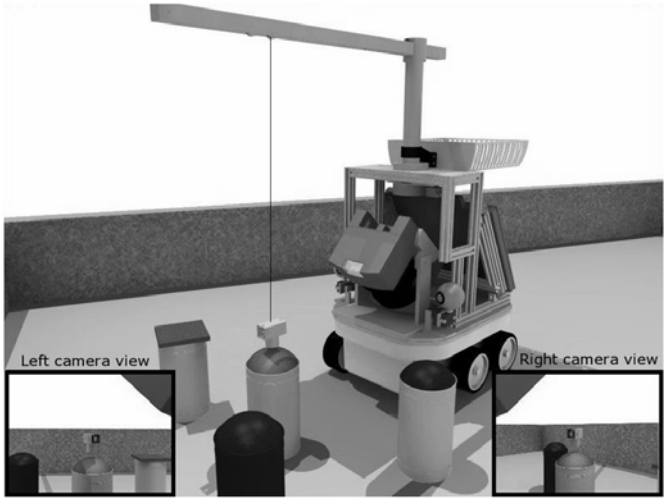

Fig. 5. Scenario of affordance based robot simulation experiments (Section 3). Birds view illustrating robot Kurt2 within a scene of objects of colored cans, using a magnetic effector at the end of a rope for interaction with the scene, described in more detail in Section V. The lower left/right corner shows the field of view of the left and right camera, respectively.

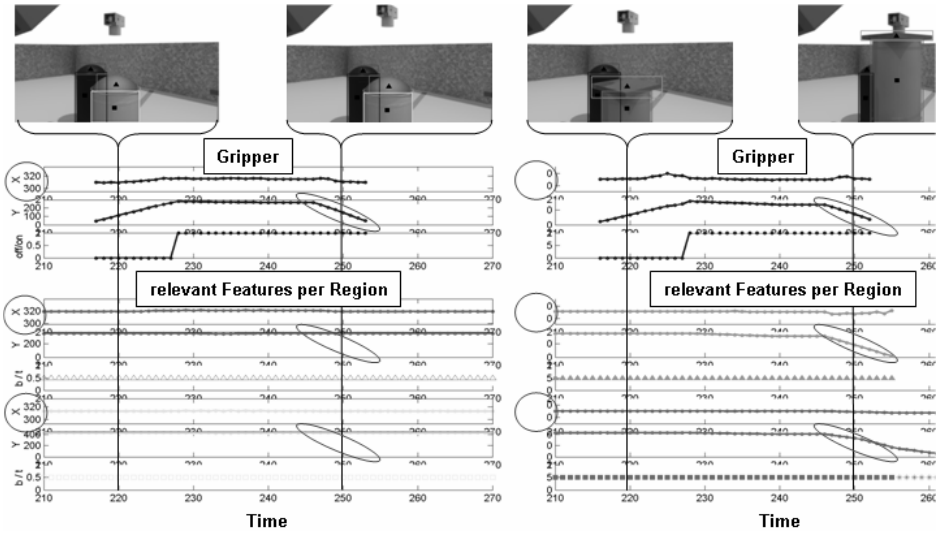

Fig. 6. Example of an affordance recognition process (here referring to '*lift-able*'): The upper image shows the right camera views of the robot while trying to lift a test object by means of a magnetizing effector at the lower end of a rope. The diagrams visualize the observation of robot relevant sensor information (e.g., status of gripper, magnet [on/off] and various features of test objects within the focus of attention. Using this sensor/feature information, the relevant channels to discriminate regions of interest that are associated to *lift-able* and *non-lift-able* objects are identified (highlighted by ellipsoids).

5.2 Affordance Recognition

The *recognition* of an affordance is crucial for verifying a hypothesis about an affordance *A* associated with a entity *E*. These entities are extracted out of the images as follows. Firstly, a watershed algorithm is used to segment regions of similar color together. After merging of smaller parts, every entity is represented by the average color value, the position in the image and the relation to adjacent regions (top/bottom). This information is also used for tracking entities over time. To verify whether or not an entity becomes '*lift-able*', the magnetizable effector of the robot is lowered until the top region of the object under investigation is reached, the magnet is switched on and the effector is lifted up. Fig. 6 shows the features of the effector (position and magnet status) over time (diagram of gripper features). If the entity is *lift-able* (Fig. 6, right column), a common motion between effector and region can be recognized. Additionally the magnet has to be switched on and the effector has to be placed in the center of the top region. These rules build up the affordance recognizer looking for *lift-able* entities in the recognition phase of the experiment.

5.3 Affordance Cueing

Cueing and recognition can require extraction of different kinds of features. Section IV already emphasized the need for some structural description of the top region, to separate the unequal shape of the top regions. In order to get structural information about an entity a histogram over prototypical SIFT descrptors is used to discriminate between circular and rectangular regions.

Classification of Relevant Descriptors. All local SIFT descriptors extracted in the region of the entities are clustered using the k-means (k = 100) method. For each specific entity, we generate a histogram over cluster prototypes, using a NN-approach to get the cluster label for each SIFT descriptor in that region. In a supervised learning step, every histogram is labeled whether it is or isn't associated with a rectangular or circular entity. A C4.5 decision tree of size 27 is then able to distinguish between these two classes. The error rate on a test set with 353 samples is ~ 1.4%. Table I shows the resulting confusion matrix for the test set.

TABLE I
CONFUSION MATRIX FOR C4.5 BASED STRUCTURE CLASSIFICATION

Classified as			
Rect.	*Circ.*		
256	1	*Rect.*	Class
4	92	*Circ.*	

Decision tree used for Affordance Cueing. The objects tested for the affordance '*lift-able*' in the recognition phase are members of the training set. The outcome of the recognition provides the class label ('*lift-able*' or '*non lift-able*'). The bottom region of the object is marked 'unknown' because this entity is not tested directly. As mentioned earlier, there exists no object model yet, therefore only *entities* exist in the system. Backtracking the object's entities over time allows additional training samples to be used with little more memory effort to remember the data. In our

experiment 30 frames are used from the beginning of the affordance recognition back, that means a recall of ~2.5 seconds from the past (12 fps are captured by the robot during simulation). The entity representation for the cueing phase contains the following features: (a) average color value of the region in the image, (b) top/bottom information, (c) the result of the structure classification, (d) the size of the segmented region. Fig. 7 depicts the structure of the decision tree. It is important to note that as a result from learning, the *relevant attributes* in the cueing process are *on top of the tree*, these are '*top*'/'*bottom*' and '*circ*'/'*rect*' here. The size attribute is located on the lowest level and only useful to separate 6 non *lift-able* samples from 474 *lift-able* ones. The error rate on the test set, containing the remaining entities which where not used for training, is 1.6%. Table II shows the confusion matrix for these data.

```
tb = bottom: unknown (1086.0)
tb = top:
|   structure = circ: non lift-able (552.0)
|   structure = rect:
|   |   size > 1426 : lift-able (402.0)
|   |   size <= 1426 :
|   |   |   size <= 1410 : lift-able (72.0)
|   |   |   size > 1410 : non lift-able (6.0)
```

Fig. 7. Structure of the C4.5 decision tree that maps attributes of the affordance feature vector f(A,t) to affordance capabilities (*lift-able*, *non lift-able* unknown). The number of samples that support the corresponding hypothesis are denoted in brackets.

TABLE II
Confusion Matrix for C4.5 based Descriptor Classification

Classified as				
lift-able	*non lift-able*	*unknown*		
95	11	0	*lift-able*	class
3	319	0	*non lift-able*	
0	0	471	*unknown*	

5 Conclusions

This work presented the perceptual cueing to opportunities for interaction of robotic agents in a general sense, in extension to the classical functional view on feature representations. The new framework for cueing and recognition of affordance-like visual entities is verified with a concrete implementation using state-of-the-art visual descriptors on a simulated robot scenario and proved that features are successfully selected that are relevant for prediction towards affordance-like control in interaction. The simulation was chosen in a realistic way so that major elements of a real world scenario, such as shadow events, noise in the segmentation, etc., characterized the results and thus enable a fundamental verification of the theoretical assumptions.

Future work will focus on extending the feature based representations towards object based prediction of affordance-based interaction, routing in the work on the

visual descriptor information presented here, and demonstrating the generality of the concept. Furthermore, we think that the presented machine learning component implemented by a decision tree can be enhanced by using reinforcement learning methodology to learn relevant events in state space for cueing to the opportunities for interaction.

Acknowledgments

This work is funded by the European Commission's projects MACS (FP6-004381) and MOBVIS (FP6-511051) and by the FWF Austrian joint research project Cognitive Vision under sub-projects S9103-N04 and S9104-N04.

References

[1] J.J. Gibson, *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979.

[2] U. Neisser, *Cognition and Reality. Principles and Implications of Cognitive Psychology*, San Francisco, Freeman & Co., 1976.

[3] E.J. Gibson, Exploratory behavior in the development of perceiving, acting and the acquiring of knowledge. *Annual Review of Psychology*. 39, 1-41. 1988.

[4] Faillenot, I., Toni, I., Decety, J., Grégoire, M.-C., & Jeannerod, M., Visual pathways for object-oriented action and object recognition: functional anatomy with PET. *Cerebral Cortex*, 7, 77-85. 1997.

[5] Fagg, A. H. and Arbib, M. A., Modeling parietal-premotor interaction inprimate control of grasping. *Neural Networks*, 11(7-8):1277-1303. 1998

[6] Wheeler S.D. and Fagg H.A. and Grupen R.A., Learning Prospective Pick and Place Behavior, *Proc. 2nd International Conference on Development and Learning*, Pages 197-202, IEEE Computer Society, Cambridge, MA, June, 2002.

[7] Fitzpatrick, Paul, Giorgio Metta, Lorenzo Natale, Sajit Rao and Giulio Sandini. "Learning About Objects Through Action - Initial Steps Towards Artificial Cognition*", In Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Taipei, Taiwan, May 12 - 17, 2003

[8] Stoytchev, A., "Behavior-Grounded Representation of Tool Affordances", *In Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain, April 18-22, 2005

[9] Stark L. and Bowyer, K. W., ``Function-based recognition for multiple object categories", *Image Understanding*, 59(10), 1--21.

[10] Rivlin, E., Dickinson, S.J., and Rosenfeld, A., "Recognition by functional parts," *Computer Vision and Image Understanding*, 62, pp. 64–176, 1995.

[11] Bogoni L. and Bajcsy R., "Interactive Recognition and Representation of Functionality", Computer Vision and Image Understanding: CVIU, 62(2), 194-214, 1995.

[12] M.G. Edwards, G.W., Humphreys, and U. Castiello, Motor facilitation following action observation: a behavioural study in prehensile action In *Brain Cognition*, volume 53, pp. 495-502, 2003.

[13] D. Lowe, Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60(2), pp. 91-110, 2004.

[14] J.R. Quinlan, *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[15] I. Cos-Aguilera, L. Cañamero, G. M. Hayes, and A. Gillies, Ecological integration of affordances and drives for behaviour selection. In Bryson J. et al. (eds.), *Proc. Workshop on Modeling Natural Action Selection*, pp. 225-228, AISB Press, 2005.

[16] I. Cos-Aguilera, L. Cañamero and G. M. Hayes, Using a SOFM to learn Object Affordances", *Proc. Workshop of Physical Agents*, WAF'04, Girona, Catalonia, Spain. March, 2004.