Affordance-Based Object Recognition Using Interactions Obtained from a Utility Maximization Principle

Tobias Kluth(✉), David Nakath, Thomas Reineking, Christoph Zetzsche, and Kerstin Schill

Cognitive Neuroinformatics, University of Bremen, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany
tkluth@math.uni-bremen.de

Abstract. The interaction of biological agents within the real world is based on their abilities and the affordances of the environment. By contrast, the classical view of perception considers only sensory features, as do most object recognition models. Only a few models make use of the information provided by the integration of sensory information as well as possible or executed actions. Neither the relations shaping such an integration nor the methods for using this integrated information in appropriate representations are yet entirely clear. We propose a probabilistic model integrating the two information sources in one system. The recognition process is equipped with an utility maximization principle to obtain optimal interactions with the environment

Keywords: Affordance · Sensorimotor object recognition · Information gain

1 Introduction

The ability of humans to reliably recognize objects in the environment is still a largely unsolved problem for artificial systems. Usually, object recognition is understood as a classification problem where a fixed mapping from feature vectors to object classes is learned. This is in line with the classical view of perception as the input from world to mind and of action as the output from mind to world [6], which implies a dissociation between the capacities for perception and action. However, there is strong evidence that object recognition cannot be understood independently of the interaction of an agent with its environment [8]. "Active perception" approaches [1,2] take this partially into account, but actions are not merely means for acquiring new information, they also provide evidence themselves for the recognition [5]. What an agent perceives is thus also determined by what it does or what it is able to do [8].

Research in the direction of affordances by Gibson [3] also provides evidence that affordances are key ingredients of the perceptual process. A variety of studies regarding object recognition show that the visual information of a manipulable

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 406–412, 2015.
DOI: 10.1007/978-3-319-16181-5_29

object causes an activation of representations of actions which can typically be executed on the object [4]. The advantageous interplay between sensory and action information, which was also recognized by Neisser [7], should be considered in the recognition process.

The strong interrelation between motor actions and sensory perceptions is basis for the concept of a sensorimotor representation [8,10]. Similarly to the affordance point of view the processing stages for sensory and motor information are not separated. The approach including the actions in the representation gives the opportunity to choose the next action such that a specific objective is pursued. Generally, the problem of action selection can be solved in numerous ways, but as information gathering should be one major purpose of motor actions it is appropriate to consider an information-theoretic utility function. Prior research in this area often found that the principle of *information gain* is well suited to select an appropriate next action.

In this paper, we propose a system for object recognition which incorporates both the information gain principle from sensorimotor systems and the theoretical concept of affordances. Building upon the investigations in [11], we developed a sensomotoric probabilistic reasoning system for affordance-based object recognition. The design of our architecture is motivated by two main goals: i) to provide a clear relation to Bayesian inference approaches, and ii) to enable a comparison between the classic sensory approach and the sensorimotor, affordance-oriented approach within one common probabilistic framework.

2 Object Recognition System

The system described in the following is a generic architecture (see Fig. 1). The recognition loop starts out with a particular pose of an object which is perceived by a sensor. The sensor passes its raw data to the sensory processing module. After processing, the sensory data becomes part of a new sensorimotor feature, which is then fed into the probabilistic reasoning module. The processed sensory data are also used to obtain a set of possible interactions, i.e., the affordances offered by the sensory data related to the abilities of the agent. The Bayesian inference module calculates the new posterior distribution based on a previously-learned sensorimotor representation. This representation contains the learned perceptual consequences of an interaction in a given state for every object class. The posterior distribution constitutes the current belief of the system. This belief is used by the information gain strategy to choose an optimal next action from the set of possible interactions. The selected interaction then also becomes part of the sensorimotor feature and is subsequently executed by the agent. The whole process results in a new state, which in turn delivers new raw sensory data to enter the next cycle of the recognition loop.

More formally speaking, the system depends on an *agent*, which can be controlled such that it perceives information about a specific aspect of the world. In Fig. 1, the two arrows pointing from the states to the sensory processing module correspond to the mapping $A : U \times X \rightarrow R$, where U is the space of all

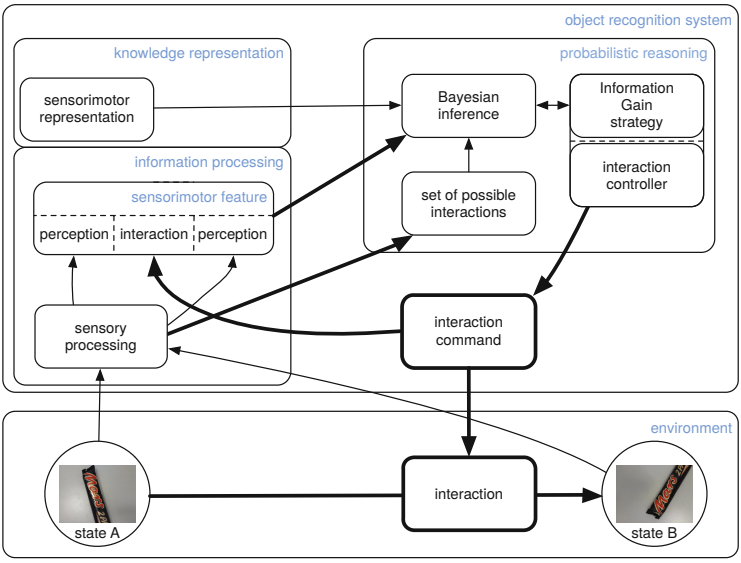


Fig. 1. Architecture of the object recognition system

interactions which are currently possible, X is the state space, and R is the raw sensor data space.

The system has no explicit knowledge about the actual state, and the currently possible interactions U. The possible interactions are of course dependent on the state but nevertheless both information must be obtained from the sensor data. The sensoric dependency on the state is formalized by the mapping $U : X \to \mathcal{P}(\Omega_U)$, where Ω_U is the set of all possible interactions and $\mathcal{P}$ denotes the power set. Note that U comprises the link from the state to the sensory processing module and the following link to the set of possible interactions in Fig. 1, i.e., the perceived affordances. Assuming that the output of the function U is given, we write U instead of $U(x),\ x \in X$, for convenience. Considering the state-agnostic behavior, the influence of the agent can be formally redefined to $A_x : U \to R$ with $A_x(u) := A(x, u) = r,\ x \in X,\ u \in U(x),\ r \in R$. The only time-dependent variables are the state x and the interaction u.

The raw sensor data $r \in R$ is fed into the *sensory processing* (SP) which mainly extracts the relevant features belonging to a feature space F, i.e., $SP : R \to F$. Subsequently, the quantization operation $Q_S : F \to S$ maps the features to a specific feature class in the finite space S. The possible interactions are mapped with $Q_M : \Omega_U \to M$ to the finite set of interactions M to yield a manageable product space of sensory information and actions. The results of these quantizations then become part of a sensorimotor feature (SMF). The single quantizations are represented in Fig. 1 by the arrows from the sensory processing module and the interaction command to the sensorimotor feature which is defined as the triple

$$SMF_i := (s_{i-1}, m_{i-1}, s_i), \tag{1}$$

where $m_{i-1} := Q_M(u_{i-1})$ is the interaction between the sensor information s_{i-1} and s_i at time step t_{i-1} and t_i. The whole chain of operations to obtain the sensor information at a time step t_i can be described by $s_i := (Q_S \circ SP \circ A_x)(u_{i-1})$.

The *knowledge representation* is comprised of the learned sensorimotor representation (SMR), which is a full joint probability distribution of SMFs and the classes represented by the discrete random variable Y. Every possible SMF is generated on a set of known objects in a training phase. This means that, from every possible state x, the sensory consequence of every possible action u is perceived, resulting in

$$SMR := P(SMF_i, Y) = P(S_{i-1}, M_{i-1}, S_i, Y). \tag{2}$$

The *probabilistic reasoning* module consists of a Bayesian inference approach accompanied by an information gain strategy. They rely on bottom-up sensory data and top-down information from the knowledge representation. The information gain strategy can choose an optimal next interaction for the agent, thus improving the input of the following Bayesian inference step.

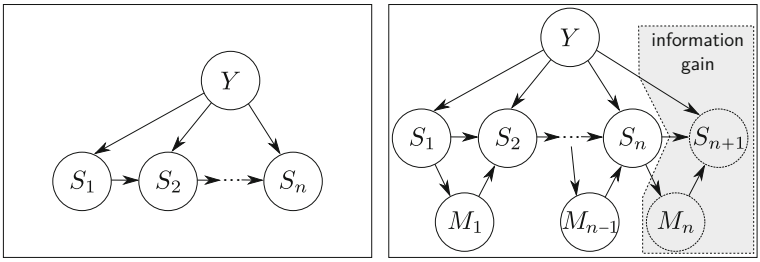
3 Model Implementation and Outlook

Based on the schematic outline presented above, we applied our system to object recognition using a robotic arm interacting with objects in 3D space. We used a discrete set of interactions M of a robotic arm with an object which comprise the relative position/pose of the visual sensor to the object ($\Omega_U = M$, $Q_M = Id$).

In the learning phase, features are extracted from the training data (images from every reachable state). GIST-features [9] are used to describe the sensory input, i.e., defining SP. The quantization Q_S is then learned by performing a k-means clustering on the extracted features. In order to build the individual SMFs, features are extracted and the results are assigned to clusters with the aid of the previously defined mapping Q_S. These labels are combined with the corresponding interactions in a set of SMFs. Finally, all generated SMFs are stored in a Laplace-smoothed SMR.

The probabilistic reasoning is comprised of a Bayesian inference module in the form of a dynamic Bayesian network (BN) and a corresponding information gain strategy. Two of these probabilistic reasoning modules were implemented to examine the difference between *sensor networks*, which only take into account sensory information (which also implies that no information gain strategy is used), and *affordance-based networks*, which integrate sensory perceptions and interactions. The object recognition in the sense of computer vision then takes place by classification which is performed by choosing the class with the maximum posterior probability.

The representative of the *sensor networks* is Bayesian network 1 (BN1) (see Fig. 2a), which resembles an extended naive Bayes model that additionally allows



(a) BN1 **(b)** BN2

Fig. 2. In Bayesian network BN1 (a) sensory information S_n is processed only to obtain the object class Y. Bayesian network BN2 (b) is equipped with the information gain startegy which takes also the action M_n into account.

for statistical dependencies between the preceding and the current sensor information, s_{i-1} and s_i, resulting in

$$P(y|s_{1:n}) \propto P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, y), \tag{3}$$

where $s_{1:n}$ is a short notation for the n-tuple $(s_1, \ldots, s_n)$.

Bayesian network 2 (BN2) (see Fig. 2b) uses the full information of the SMF and therefore belongs to the *affordance-based networks*. The assumption that the current sensory input s_i depends on the action m_{i-1} integrates sensory perceptions and actions in the recognition process and permits the application of the information gain strategy to choose the next optimal interaction. Additionally, it is assumed that the action m_{i-1} statistically depends on the preceding sensory input s_{i-1}. The inference can then be conducted by

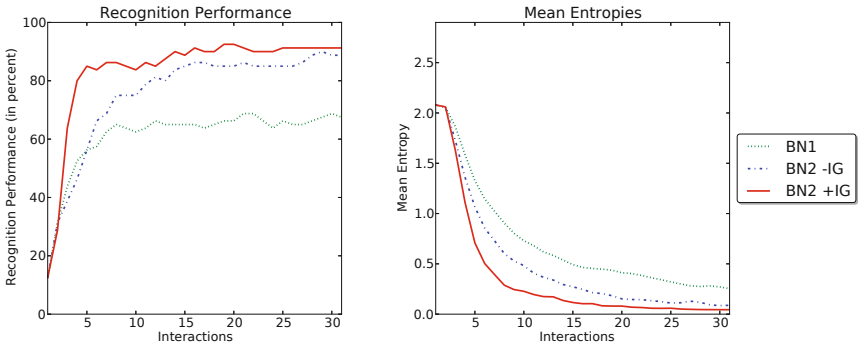
$$P(y|s_{1:n}, m_{1:n-1}) \propto P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, m_{i-1}, y)P(m_{i-1}|s_{i-1}). \tag{4}$$

The strategy for action selection should satisfy two main properties: (i) The strategy should adapt itself to the current belief state of the system and (ii) the strategy should not make decisions in an heuristic fashion but tightly integrated into the axiomatic framework used for reasoning. The information gain strategy presented in this paper complies with both of these properties.

The information gain IG of a possible next action m_n is defined as the difference between the current entropy and the conditional entropy,

$$IG(m_n) := H(Y|s_{1:n}, m_{1:n-1}) - H(Y|S_{n+1}, m_n, s_{1:n}, m_{1:n-1}). \tag{5}$$

This is equivalent to the mutual information of Y and (S_{n+1}, m_n) for an arbitrary m_n. As the current entropy $H(Y|s_{1:n}, m_{1:n-1})$ is independent of the next action



(a) Recognition performance **(b)** Mean entropies

Fig. 3. Results of the robotic arm evaluation (8 object classes, 10 objects per class, 30 discrete viewpoints). BN 1 and 2 -IG executed random movements while BN2 +IG executed information-gain-guided movements.

m_n the most promising action m^* can be calculated by minimizing the expected entropy with respect to S_{n+1},

$$m_n^* = \arg\min_{m_n}(\underset{S_{n+1}}{E}[H(Y|s_{1:n}, S_{n+1}, m_{1:n})]). \tag{6}$$

Because the sensory input s_{n+1} is not known prior to executing m_n, the expected value over all possible sensory inputs s_{n+1} is taken into account. The selected action $m^* \in M$ is integrated into the next sensorimotor feature. The inverse mapping of Q_M can then be used to obtain a top-down interaction command $u \in U$, which is then executed by the agent.

Preliminary results are shown in Fig. 3. In the future, we plan to conduct a more extensive evaluation of our approach (using different sensory features) by comparing it to established object recognition approaches on a larger data set. Furthermore we want to extend our approach by a saliency feature detector.

Acknowledgments. This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace], and DLR projects "EnEx" and "KaNaRiA".

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. International Journal of Computer Vision **1**(4), 333–356 (1988)
2. Bajcsy, R.: Active perception. Proceedings of the IEEE **76**(8), 966–1005 (1988)
3. Gibson, J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1992)
4. Grèzes, J., Decety, J.: Does visual perception of object afford action? Evidence from a Neuroimaging study. Neuropsychologia **40**(2), 212–222 (2002)

5. Helbig, H.B., Graf, M., Kiefer, M.: The role of action representations in visual object recognition. Experimental Brain Research **174**(2), 221–228 (2006)
6. Hurley, S.L.: Consciousness in action. Harvard University Press (2002)
7. Neisser, U.: Cognition and reality: Principles and implications of cognitive psychology. WH Freeman/Times Books/Henry Holt & Co. (1976)
8. Noë, A.: Action in Perception. MIT Press (2004)
9. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research **155**, 23–36 (2006)
10. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24**(5), 939–972 (2001)
11. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: top-down and bottom-up modeling. Journal of Electronic Imaging **10**(1), 152–160 (2001)