From Images via Symbols to Contexts: Using Augmented Reality for Interactive Model Acquisition

Sven Wachsmuth[+], Marc Hanheide[+], Sebastian Wrede[+] and Christian Bauckhage[*]

[+]Bielefeld University, Faculty of Technology, D-33594 Bielefeld, Germany
{swachsmu,mhanheid,swrede}@techfak.uni-bielefeld.de

[*]York University, Centre for Vision Research, Toronto ON, M3J 1P3, Canada
bauckhag@cs.yorku.ca

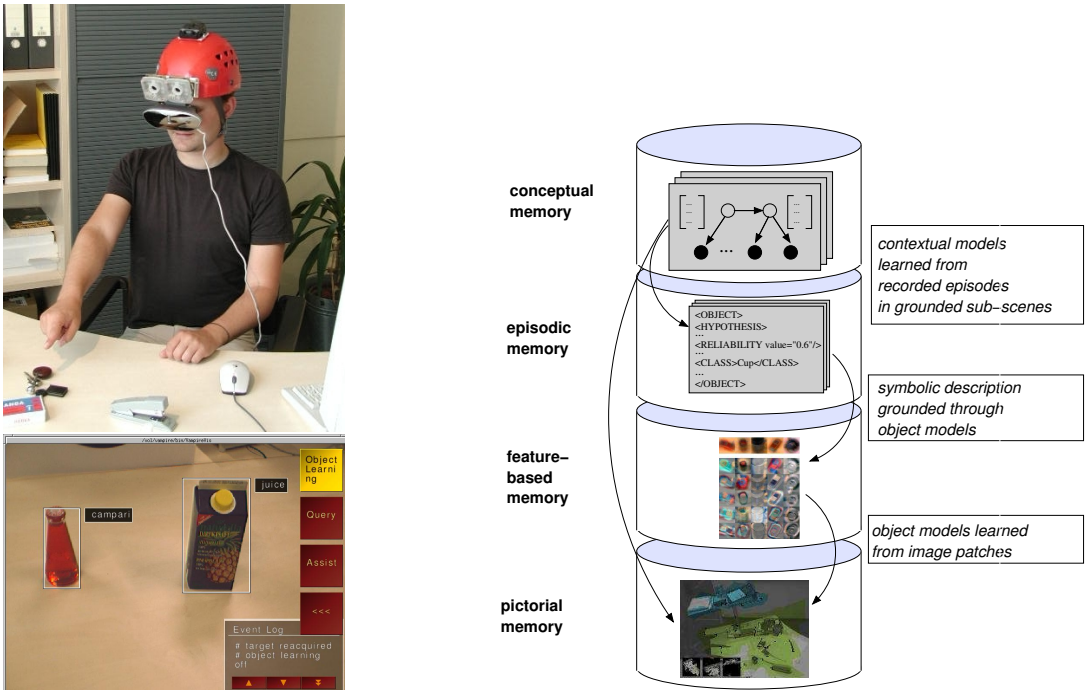
Abstract. Systems that perform in real environments need to bind the internal state to externally perceived objects, events, or complete scenes. How to learn this correspondence has been a long standing problem in computer vision as well as artificial intelligence. Augmented Reality provides an interesting perspective on this problem because a human user can directly relate displayed system results to real environments. In the following we present a system that is able to bootstrap internal models from user-system interactions. Starting from pictorial representations it learns symbolic object labels that provide the basis for storing observed episodes. In a second step, more complex relational information is extracted from stored episodes that enables the system to react on specific scene contexts.

1 Introduction

Mixed reality systems combine real world views with views of a virtual environment [4]. In the sub-field of augmented reality virtual augmentations are added to the real world view of the user. This is typically realized by using a setup with a head-mounted device which is equipped with cameras and a display. Most of the research on computer vision in this field is dedicated to the problem of aligning real and virtual objects (cf. e.g. [4, 8]). This is mostly based on pre-defined 3-d CAD models. The AR system is either used to present a virtually changed environment to the user or to support the user in a pre-defined task, e.g. [8]. In the VAMPIRE[1] project we take a different approach in that we focus on the problem of how a system can bootstrap its knowledge about an unknown real environment. By using Augmented Reality techniques, the computer vision system is embodied through the tight interaction with the user. In this kind of scenario, augmentations, like bounding boxes, text labels, or arrows, are used in order to close the feedback cycle to the user. In turn, the user is able to react based on the augmentations by changing the view or acting in the scene. Thus, the coupling between the user and the vision system is highly dynamic and depends on the interaction history.

The learning of visual models based on human feedback has been explored in several different scenarios. Roy uses video data from mother child interactions in order to learn the association between acoustic and visual pattern [9]. Steels introduces

[1] Visual Active Memory Processes and Interactive REtrieval – IST-2001-34401



(a) AR gear

(b) Memory organization

Fig. 1. AR system and memory organization

the term of social learning in a scenario where a human teaches different kinds of objects to an Aibo robot [10]. In [2], imitation learning is explored as social learning and teaching process with aims at socially intelligent robots. Finally, Heidemann et al. [7] present an augmented reality system for interactive object learning which was developed within the VAMPIRE project. However, most systems limit the learning capability on a single aspect, like learning a classifier for an individual object. A more general approach needs to deal with various kinds of data structures and needs to integrate different learning processes in a single framework.

2 AR interaction and formation of memory content

In Fig. 1(a) the scenario of the system is shown. The user is sitting at a regular office table and wears a head-mounted device which is equipped with cameras and a display. Information about recognized objects and results of user queries are visualized using augmented reality (AR). The head of the user is tracked using a CMOS camera and an inertial sensor that are mounted on the top of the helmet. The head pose is computed from an artificial landmark that is placed in the scene and defines a global coordinate system. The system is able to detect objects, and user activities, like moving an ob-

ject. It copes with varying lighting conditions as well as cluttered video signals. By selecting from a menu displayed on the right of the field of view by speech or a mouse wheel the user can trigger learning sessions or retrieve information.

In order to realize a bootstrapping behavior of the system starting from image-based representations to symbol-based representations, the organization of memory content plays a key role. The technical basis for storing and retrieving various kinds of information as well as the coordination of different visual behaviors is provided by the *Active Memory Infrastructure* which is also described in [11]. The persistence back-end is the native Berkeley XML DB. Binary data is stored directly in the underlying relational database and is referenced from stored XML documents. Thus, XML provides a unified data model for structured information that is exchanged between system components and stored in the memory.

On the conceptual level, we distinguish four different kinds of abstraction layers in the memory representation (see Fig. 1(b)). On the pictorial layer images and image patches are temporarily stored. The feature-based layer includes learned object models and configuration data of the object recognition components. In the episodic memory layer recognition results are stored that have been reliably detected during an interactive session with a user. Finally, the categorical layer consists of a couple of contextual models that e.g. describe typical configurations of objects. Each higher layer is grounded in a layer that is nearer to the signal. Object models in the feature-based memory are learned from image patches that are stored during system usage; detected objects and events are related to learned prototypes in the feature space; finally, contextual models are learned from episodic sequences that capture a spatial context, e.g. the user was looking around on the writing area of his or her desk.

Interpretation as well as learning processes are working asynchronously on memory representation. They are coordinated through memory event notification [11], e.g. the object anchoring component is triggered if a new object hypothesis is stored in the memory.

3 Image-based scene decomposition and acquisition of object views

In the Augmented Reality scenario, the user and the system share a common view. The images of the head-mounted cameras are directly shown on the head mounted stereo display, so that the user sees what the camera records and the system knows which part of the scene is focused by the user. Two different visual behaviors are used on this pictorial representation level.

Mosaicing: In indoor environments meaningful sub-scenes are typically defined by planes, e.g. table top, front side of a shelf, walls. However, if we are keeping a sufficient level of image detail these kind of contextual areas cannot completely be seen through a single view. In [5] we present an unique approach to create mosaics for

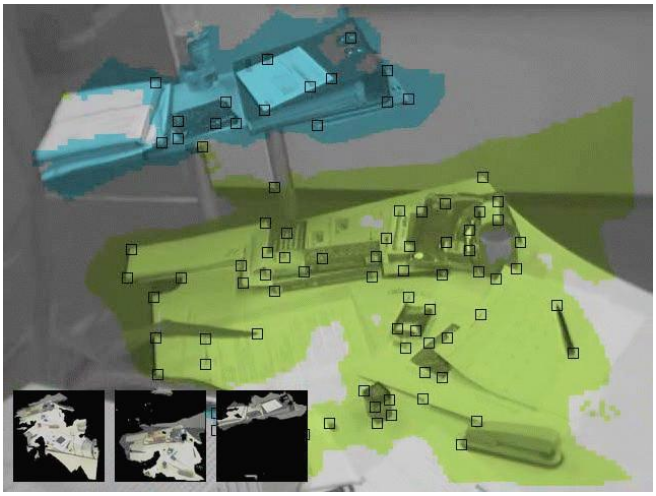
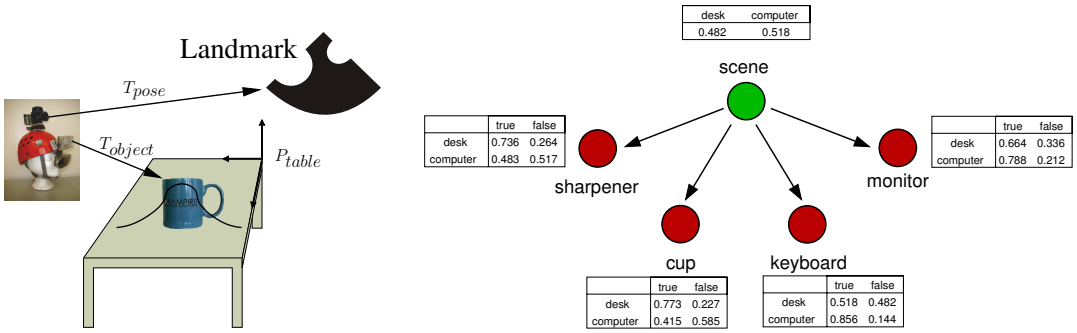


Fig. 2. Constructing and tracking of planar sub-scenes. The mosaicing approach has constructed three different planar sub-scenes that are stored in the pictorial memory. They were constructed from an image sequence of the head mounted cameras which is incrementally processed in soft real-time. The user turned his or her head from the right side of the table to the left side. The system has correctly identified the two different desk levels.

arbitrarily moving head-mounted cameras. It uses a three stage architecture. First, we decompose the scene into approximated planes using stereo information, which afterwards can be tracked and integrated to mosaics individually (see Fig. 2). This avoids the problem of parallax errors usually occurring from arbitrary motion and provides a compact and non-redundant representation of the scene. Each plane defines a coarse spatial context from which contextual models can be learned that interrelate objects that frequently co-occur in such a sub-scene.

Object tracking: The acquisition of object models is a key to higher-level descriptions of a scene. For object recognition an appearance-based VPL-classifier [1] is used that can directly be trained from image patches. These are automatically extracted while a user is focusing the target object. An entropy measure is used in order to segment unknown objects from a more or less homogeneous table plane. In the learning mode of the system the detected area is augmented to the view of the user. Once the first view is registered by the system a data-driven tracking technique [6] is started that provides additional views of the object. Each view that the system collects for learning is checked with the user so that he or she can control the learning process. The patches can be stored in the pictorial memory of the system for a fast online learning of objects as well as a more accurate object learning on a longer time scale [1]. A label is currently given by speech input based on a pre-defined lexicon.



(a) Object localization and anchoring based on 3-d pose.

(b) Bayesian network for scenery classification can be learned from anchored object hypotheses.

Fig. 3. Contextual models are learned from episodic memory content.

4 Object anchoring and the role of context

Object anchoring links corresponding object hypotheses that are detected at different points in times to the same symbol. This is essential for representing episodes over an extended period of time. In addition to the trajectory information from object tracking, a second strategy is applied for linking that takes the 3-d position of the object hypotheses into account. This can be estimated based on a self-localization of the cameras [3]. Currently, we assume that each object is lying on a table plane. Object hypotheses are fused over time if the 3-d positions are close enough to each other. A Gaussian curve models the probability that two hypotheses refer to the same object. (see Fig. 3(a)). For the final classification result the labels provided by the object recognition component is integrated over a short period of time. Thereby, the reliability value of a specific hypothesis is adapted. Only those hypotheses that have a highly rated reliability value are used for contextual model learning.

Based on such kind of episodic data, contextual models can be estimated that represent typical configurations of objects in a sub-scene. For that, we use simple Bayesian networks with discrete conditional probability tables. In Fig. 3(b) a learned parameterization of a Bayesian network is shown.

The contextual models in turn can be used to judge certain object hypotheses given their context as well as can be used to classify more general scene contexts, like 'office table' if a keyboard and computer mouse has been found. Thus, higher-level categories can be detected that are defined through relations between objects.

5 Conclusion and Outlook

In this paper we presented a bootstrapping approach for the acquisition of knowledge in unknown environments. Augmented Reality techniques are used in order to close

the interaction loop with the user. This acquisition process combines several visual behaviors that are integrated using the active memory infrastructure. It is shown how the tight coupling with the user can be used in order acquire grounded higher-level representations. The demonstration system is running on 5 different laptops allowing a soft real-time behavior. New objects can be learned in about 2-3 minutes acquiring between 4 to 6 object views. Contextual models are learned on a longer time scale. Parameters of Bayesian networks are estimated from about 5 minutes of regular system usage where the corresponding scenery label is given by the user. Further system development will focus on a further integration of the mosaiced sub-scenes and the structural learning of contextual models. We think that the triadic interaction between the system, the human, and the environment provides an ideal basis for pushing the cognitive development of artificial systems to a further level. Augmented reality offers strong interaction patterns for this purpose. On the other side, cognitive system capabilities will lead to a next generation of assistance technology offering a variety of applications.

References

1. H. Bekel, I. Bax, G. Heidemann, and H. Ritter. Adaptive Computer Vision: Online Learning for Object Recognition. In *Proc. Pattern Recognition Symposium (DAGM)*, 2004.
2. C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning from and about Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots,. *Artificial Life*, 2004. (Forthcoming 2004).
3. M.K. Chandraker, C. Stock, and A. Pinz. Real Time Camera Pose in a Room. In *Int. Conf. on Computer Vision Systems*, volume 2626 of *LNCS*, pages 98–110, 2003.
4. D. Drascic and P. Milgram. Perceptual Issues in Augmented Reality. In Mark T. Bolas, Scott S. Fisher, and John O. Merritt, editors, *Stereoscopic Displays and Virtual Reality Systems III*, volume 2653 of *SPIE*, pages 123–134, San Jose, California, USA, January - February 1996.
5. N. Gorges, M. Hanheide, W. Christmas, C. Bauckhage, G. Sagerer, and J. Kittler. Mosaics from Arbitrary Stereo Video Sequences. In *Proc. Pattern Recognition Symposium (DAGM)*, 2004.
6. Ch. Gräßl, T. Zinßer, and H. Niemann. Efficient Hyperplane Tracking by Intelligent Region Selection. In *Proc. IEEE Southwest Symposium on Image Analysis and Interpreta tion*, pages 51–55, 2004.
7. G. Heidemann, H. Bekel, I. Bax, and H. Ritter. Interactive Online Learning. *Pattern Recognition and Image Analysis*, 15(1):55–58, 2005.
8. G. Klinker, K. Ahlers, D. Breen, P.-Y. Chevalier, Ch. Crampton, D. Greer, D. Koller, A. Kramer, E. Rose, M. Tuceryan, and R. Whitaker. Confluence of Computer Vision and Interactive Graphics for Augmented Reality. *Presence: Teleoperations and Virtual Environments*, 6(4):433–451, August 1997.
9. D. Roy. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4(1), 2002.
10. L. Steels and F. Kaplan. AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2001.
11. S. Wachsmuth, S. Wrede, M. Hanheide, and C. Bauckhage. An Active Memory Model for Cognitive Computer Vision Systems. *Künstliche Intelligenz*, 19(2):25–31, 2005.