# Dynamisches Lernen für geometrische und graphische Objekterfassung – A Vision System for Interactive Object Learning

**Projektleitung**
Prof. Dr. Gabriele Peters

**Zeitraum**
2005 – 2007

**Kontakt**
Prof. Dr. Gabriele Peters
Fachbereich Informatik
Fachhochschule
Dortmund
Emil-Figge-Straße 42
44227 Dortmund
Tel.: (0231) 755-6796
E-Mail: gabriele.peters
@fh-dortmund.de

## Abstract

We propose an architectural model for a responsive vision system based on techniques of reinforcement learning. It is capable of acquiring object representations based on the intended application. The system can be interpreted as an intelligent scanner that interacts with its environment in a perception-action cycle, choosing the camera parameters for the next view of an object depending on the information it has perceived so far. The main contribution of this paper consists in the presentation of this general architecture which can be used for a variety of applications in computer vision and computer graphics. In addition, the funcionality of the system is demonstrated with the example of learning a sparse, view-based object representation that allows for the reconstruction of non-acquired views. First results suggest the usability of the proposed system.

## 1 Introduction

Both, computer vision as well as computer graphics are concerned with the visual appearance of objects of the real world. Major problems of computer vision are the recognition or classification of objects from images. An important challenge of computer graphics consists in the generation of internal models of objects from images, e.g., for the purpose of geometric modelling or graphic illustration. For both fields of research the acquisition of an object representation is necessary.The current situation is for the most part marked by a separation between object acquisition and further processing of the acquired information in a specific application, whether in the field of computer vision or in the field of computer graphics. This often leads to the fact that the recorded data do not meet the requirements of the application. A usual way to obtain reasonable results anyhow is the development of heuristics. Another approach, which has been used to different extents in the mentioned fields of application, is the active visual acquisition of objects. This means that the processing of the recorded data gives feedback to the acquistion part of the system. In analogy to human information processing the system should autonomously learn strategies of object acquisition on the basis of application-specific objectives only. We propose an architectural model for a responsive vision system based on techniques of reinforcement learning that takes these considerations into account. It is capable of acquiring object representations based on the intended application only and thus can be employed for a variety of tasks. The system can be interpreted as an intelligent scanner that interacts with its environment in a perception-action cycle, choosing the camera parameters for the next view depending on the information it has perceived so far. So, the subsequent input depends on the actions taken previously.

## 2 A responsive vision system

Figure 1 displays the general architecture of a vision system that learns object representations interactively depending on the intended application. The different components of the system are separated in three modules: Learning in the upper right part of the diagram, Acquisition in the lower left, and Application in the upper left. The resulting object representation is shown in the lower right part. In the diagram examples for concrete design decisions are printed in italics.

### 2.1 Module “Learning”

In traditional approaches of computer science a problem is solved by an algorithm that has been developed by the programmer who has reflected on the problem. In contrast to this, approaches exist which delegate also the discovery of solution procedures to the machine. Often principles of nature are a role model for such approaches. One example are behavior-based techniques such as Reinforcement Learning, which seem to be an appropriate approach to our problem of object learning. The principles of reinforcement learning are sketched in the following. An agent interacts with its environment} by perception and action. In an interaction step the agent receives information on the current state of the environment as input via perception. Then the agent chooses an action according to its policy function. The action is carried out and changes the state of the environment with the transition function. The agent is able to adapt its behavior to certain conditions. For this purpose the agent receives direct feedback for its last action by a scalar reward signal. In addition, a valuation of the state transition is conducted by a value function. The behavior of the agent should maximize the long term sum of the reward signals, i.e., the expected return. The value function is learned by systematic trial-and-error for which a bunch of techniques has been developed.

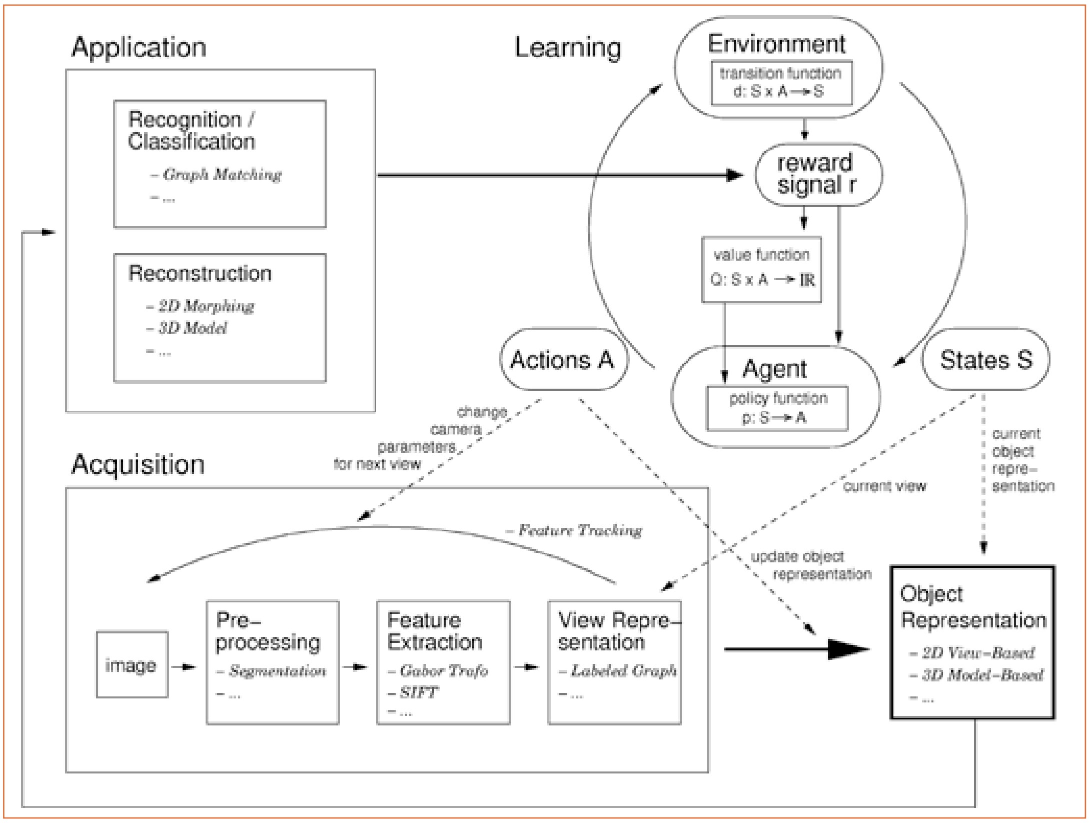

*Figure 1. Architecture of a vision system that learnes object representations.*

### 2.2 Module "Acquisition"

There are some steps in the acquisition process which are common to all kinds of applications and all kinds of data structures of the object representation. Whether a 3D model or a view-based representation should be learned, in any case there will be some steps of preprocessing of a perceived view of the object, such as image enhancement or segmentation techniques. As the perceived image is the only information about the environment available to the agent, there will also be some feature extraction technique. Examples for features are responses of a Gabor wavelet transform or the scale-invariant feature transform (SIFT). Depending on the specific design there may also be a calculation of a representation of a single view (e.g. in form of a graph labeled with local object information or in form of the 3d positions of object features) before the new information is incorporated into the object representation. (At last, to extract useful information from a sequence of views the common information between successive views has to be exploited. This can be realized for example by 2d feature tracking or by bundle adjustment techniques.)

### 2.3 Module "Application"

The two major fields of application for interactive object learning are computer vision with the goal of object recognition and classification and computer graphics with the purpose of object reconstruction. After the learning process and given a test view or a sequence of test views recognition and classification should be possible even if the agent has not experienced those test views. Analogously, the reconstruction of the complete object and the rendering from unfamiliar views are claimed. This module gives the crucial feedback to the Learning module utilizing the object representation learned so far. The dertails of the interaction between the modules are given in [ICVS 2006].

### 3 Application to a standard vision problem

We have applied the proposed system to a standard task in computer vision, namley the acquisition of a sparse, view-based object representation. To test whether the relevant information on the object has been captured by the learned scan path we reconstruct non-acquired views from per-

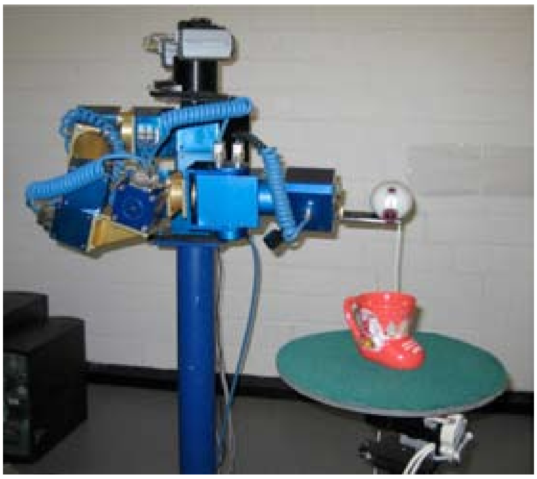

*Figure 2. Simulated setup with camera and object.*

ceived views by 2d view morphing. We simulate an eye-in-hand camera setup with the object on a table such as shown in figure 2.

The camera rotates around the object at a fixed distance and is oriented to the center of the object base. The observed object views are represented in a data base which contains views for 100 lines of longitude and 25 line of latitude on the upper view hemisphere resulting in 2500 views for one object. A sparse, view-based object representation consists of representations of key views of the scanned path (figure 3).

An unfamiliar view is morphed from those two consecutive key views which are closest to it. For view morphing we use a standard technique. A morphed view can then be compared to its original version by an error function. This yields an error for a reconstructed view. This technique is used for the calculation of the reward signal after each step of a scan episode as well as for the calculation of the total reconstruction error after a scan path has been learned. We carry out 32 steps per episode. Each episode starts at position (0, 0) on the view hemisphere. In each step the camera is moved one position on the quantized hemisphere. While tracking from step to step key views are determined. Each episode provides a scan path with associated key views. This learning process is stopped when the scan path has stabilized. To assess the quality of the learned path we calculate a total reconstruction error by choosing a set of 25 test views on the unquantized hemisphere. These views are reconstructed from the acquired key views of the learned path. Then the total reconstruction error for this path is the mean of the reconstruction errors of all test views.

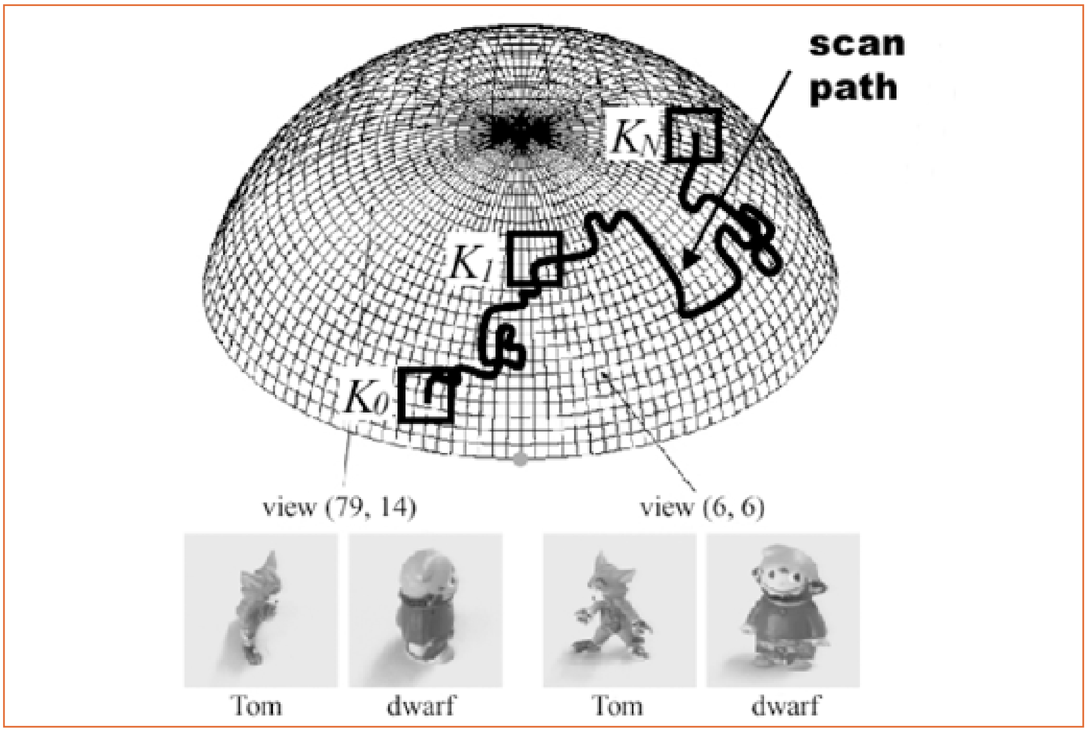

*Figure 3. Sample views of two objects and a possible scan path with three key views.*

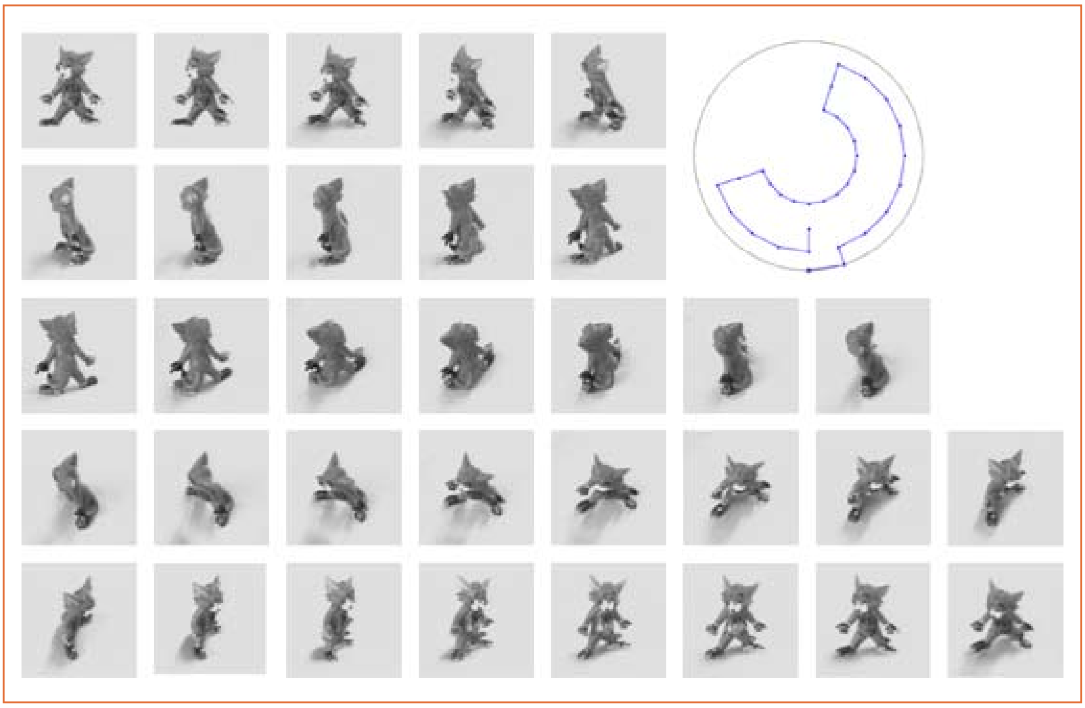

*Figure 4. Key views of the learned scan path.*

### 4 Results

The method described above has been carried out for the ``Tom" object (figure 3). The learned scan path stabilized after 2 million episodes which took a few minutes on a standard PC with pre-calculated values from the tracking and reconstruction modules. It yielded a significantly lower total reconstruction error than achieved with random scan paths of equal length. The mean reconstruction error for 100 random paths is 9.2, whereas the error for the learned path is 5.8. In figure 4 the key views of the stabilized, learned scan path are depicted. The inset shows the view hemisphere seen from above with view (0, 0) at the bottom. Only the key views of the path are displayed.

The resulting path has an even shape, going around the lower part of the view hemisphere from the front to the backside, turning up and moving back to the front in the upper part of the hemisphere. Those views of the backside of the object that haven't been covered are rather similar to the views where the agent turned up towards the top of the hemisphere. Thus it seems to make sense not to incorporate these redundant views into a sparse object representation.

### References

- Gabriele Peters, A Vision System for Interactive Object Learning, International Conference on Computer Vision Systems (ICVS 2006), 2006.

- Gabriele Peters and Thomas Leopold, Dynamic Learning of Action Patterns for Object Acquisition, International Workshop on Automatic Learning and Real-Time (ALaRT 2005), 2005.

- Gabriele Peters, Thomas Leopold, Claus-Peter Alberts, Markus Briese, Sebastian Entian, Christian Gabriel, Zhiqiang Gao, Alexander Klandt, Jan Schultze, Jeremias Spiegel, Jürgen Thyen, Martina Vaupel, Peter Voß, and Qing Zhu, Adaptive Object Acquisition, 18th International Conference on Architecture of Computing Systems (ARCS 2005), 2005.