Project no.:	**027657**
Project full title:	**Perception, Action & Cognition through Learning of Object-Action Complexes**
Project Acronym:	**PACO-PLUS**

Deliverable no.: D2.1.3

Title of the deliverable: Demonstration on the robotic platform

Contractual Date of Delivery to the CEC:	January 31st, 2008
Actual Date of Delivery to the CEC:	March 25th, 2008
Organisation name of lead contractor for this deliverable:	JSI
Author(s):	Aleš Ude, Damir Omrčen, Tamim Asfour, Pedram Azad, Kai Welke, Rüdiger Dillmann
Participants(s):	JSI, UniKarl
Work package contributing to the deliverable:	WP1, WP2, WP8
Nature:	R/D
Version:	1.0
Total number of pages:	6
Start date of project:	1st Feb. 2006 **Duration:** 48 months

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)	

Abstract:

This deliverable is the continuation of the last year deliverables D2.1 and D2.2 in which we described algorithms and software that implements basic sensorimotor processes for learning of object-action complexes. In this deliverable we show the videos of the current state of implementation of these processes. More specifically, we show the following videos:

- oculomotor control implementing foveation,
- visual attention on a humanoid robot,
- learning appearance-based object representations for recognition, and
- human motion capture with the active head.

In the report we briefly describe the videos and the underlying algorithms. While some of the algorithms have been developed already in the first year of the project, we have now implemented all of them on a humanoid robot ARMAR and on the copies of the ARMAR's humanoid head, which was made by UniKarl for KTH and JSI. The videos are included on the supplied CD.

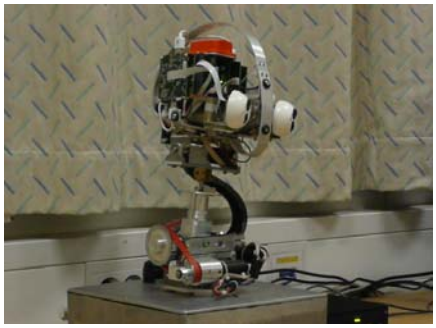
Keyword list: Sensorimotor primitives, oculomotor control, visual attention, object learning, human motion capture

Table of Contents

Introduction

In the second year of the project we worked on the further development of sensorimotor processes described in D2.1.1 and D2.1.2. The focus was on a reliable implementation of the proposed algorithms on a humanoid robot ARMAR.

Foveated Vision



The first video (**foveation.mpg**) shows the movement of the humanoid head during foveation and the simultaneous views from the four cameras. The results of the tracker enable the robot to direct its eyes towards potential objects of interest. The main task of the control system is to place a salient region over the field of view of both foveal cameras in order to enable image processing at greater detail (see image left above). Although the focus of the task is to bring an object into the center of the fovea, the control system uses the views from peripheral cameras as the basis for control. Data from peripheral images is more reliable for tracking and pursuit because objects can easily be lost from the foveal views. It is clear from the supplied video that the tracked object often disappears from the foveal views when the movement is fast, but the smooth pursuit can continue based on information from peripheral views.

We showed both theoretically and empirically what kind of accuracy we can expect from the foveated setup using two cameras per eye. A closed-loop control method that can be used to quickly direct the robot's eyes towards the object of interest was also presented. The method is described in D2.1 and the references therein. It does not need accurate knowledge of eye and body kinematics and can exploit the redundancies of the humanoid robot head.

Visual Attention

Attention and the pre-attentive processes involved in guiding the focus of selective attention are key processes in visual perception, and are important mechanisms for guiding and constraining man-machine interaction. In any interactive situation, attention is utilized as an initial mechanism to capture focus.

We implemented preattentive, bottom-up visual processing based on the model proposed by Itti, Koch, and Niebur and feature integration theory, which postulates that bottom-up preattentive processing explores the images using a number of feature processors, whose output is integrated into a global saliency map. The focus of attention emerges through competition across features in the integrated saliency map. Since this a computationally intensive process, we utilized distributed processing to implement it.

Although the system we presented last year was purely bottom-up, we note that the selectivity of

early vision processes is an important foundation for visual perception. It enables the system to filter out low-level unimportant information while attending to features indicated as important by higher-level processes by way of top-down modulation. We developed a novel way to integrate top-down and bottom-up processing for achieving such attention-based filtering [Moren et al. 2008].

The attached video **attention1.mpg** and **attention2.avi** respectively show the JSI and UniKarl version of the computer cluster and the simultaneous processing of visual information. **attention2.avi** also shows humanoid robot ARMAR at UniKarl reacting to the stimuli from the attention system.

Learning object representations by manipulation

Learning about new objects without any prior information about them is a difficult problem, which is not easy to solve by passive observers. It has been suggested that an active vision paradigm can resolve many of the ill-posed problems arising in static vision. Earlier research demonstrated that by actively exploring the environment, the robot can gain some knowledge about the objects in its world.

When learning about the new objects, the robot needs to first find interesting areas in the scene and generate the initial grasp hypotheses, which is followed by attempts to grasp the object. The problem of generating the initial grasp hypotheses and grasping itself is considered in WP4.1, whereas in WP2.1 we focused on the also difficult problem of acquiring snapshots of objects across a continuous portion of the view sphere without having prior information about the object appearance. The main idea is that by having control of the object, the robot can bring enough knowledge into the system to ensure that it can segment the object from the background, thus solving the figure-ground discrimination problem, and capture snapshots suitable for learning (see figure above).

The video **ObjectLearning.mpg** (or the identical **ObjectLearning.mp4** for QuickTime playback) shows sensorimotor processes necessary to realize the observation of objects from all relevant viewpoints of the view sphere on a humanoid robot. These include

- explorative movement primitive that can be used to determine an optimal placement of the object with respect to the robot's eyes so that the object will be in the image center and have appropriate size for learning, i. e. it will cover significant portion of the image while being away from the image boundary.
- A primitive motion that can be used to observe the grasped object from various viewpoints while keeping it centered in the image. Due to the limited manipulation capabilities of humanoid robots and arms, it is unavoidable to regrasp the observed object to ensure that the robot looks at it from all relevant viewpoints. However, the number of necessary regrasps can be reduced by performing the exploratory movements in an optimal way so that the redundancy of the humanoid is exploited and the joint limits are avoided.
- A Bayesian visual process that enables the robot to segment the object from its surroundings and acquire snapshots of the object without having prior information about its appearance.

The video shows the movement and the image area within the ellipse that the system detects as object area and that can be used for the acquisition of object representations. The algorithms are described in more detail in [Omrčen et al., 2007]. This paper was among the finalists (three papers) for the best paper award at Humanoids 2007.

Human Motion Capture with an Active Head

The perception of human motion is a capability that is crucial for human-robot interaction and human motion understanding. In particular, perceiving human motion in real-time on the humanoid's active head allows the robot to learn from humans *online*. In this context, the observation of arm movements during goal-directed actions operating on objects is of special interest. Additionally, information about the objects present in the scene, including their pose over time, is a valuable building block for perceiving OACs.

For this purpose, we have developed a stereo-based real-time human motion capture system tailored to the use with humanoid active heads, which has been presented in [Azad et al., 2007]. The input to the system is a stereo image sequence at a resolution of 640×480, which is captured by the wide-angle camera pair (4 mm focal length) of the active head of ARMAR III.

The scenario used to demonstrate the performance of our human motion capture system is as follows. A person is sitting at a table, on which a number of objects are located. In the presented video, these are a plate and two cups. The person grasps and moves the objects in order to achieve a new setup.

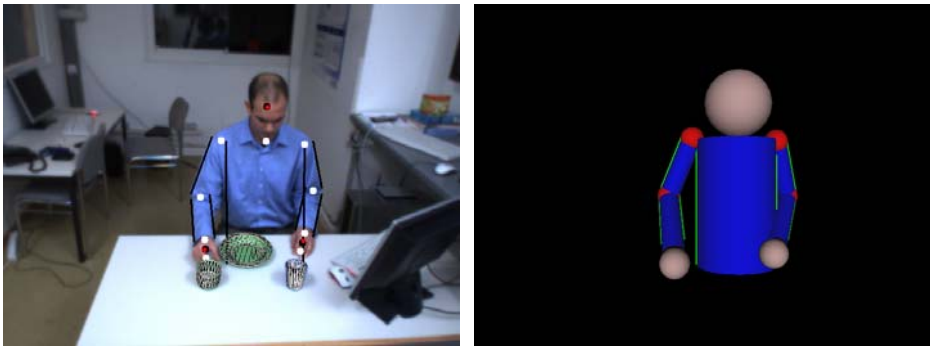


Figure 1 Left: 2D projection of person and object tracking result. Right: 3D visualization of the computed pose.

The stereo sequence was processed online i.e. in real-time at a frame rate of approximately 10 Hz, while at the same time storing the stereo sequence to hard disk for later processing for visualization purposes. Without this overhead and without any visualizations, the system achieves a frame rate of about 20 Hz on a 3 GHz CPU.

The video **MotionCaptureARMAR.avi** can be divided into six parts:

1. From the input stereo sequence, the left camera image is shown.
2. The skin color segmentation result for the left camera image is shown.
3. The shirt color segmentation result for the left camera image is shown.

Note: The two image pairs consisting of the results of skin and shirt color segmentation are the only input to the particle filter framework.

4. The contours of the estimated human model configurations are projected into the left camera image.
5. A 3D visualization of the estimated human model configurations is shown.
6. The estimated human model configurations are projected into the left camera image using 3D primitives. The result of object recognition and 6D pose estimation is projected into the same image. Please note that inaccuracies of the computed object pose are only due to the low effective resolution of the observed objects.

References

P. Azad, A. Ude, T. Asfour, and R. Dillmann (2007) Stereo-based Markerless Human Motion Capture for Humanoid Robot Systems. *IEEE International Conference on Robotics and Automation (ICRA)*, Rome, Italy.

J. Moren, A. Ude, A. Koene, and G. Cheng (2008) Biologically based top-down attention modulation for humanoid interactions. *International Journal of Humanoid Robotics,* 5(1):1-22 (to appear).

D. Omrčen, A. Ude, K. Welke, T. Asfour, and R. Dillmann (2007) Sensorimotor processes for learning object representations. In *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids 2007),* Pittsburgh, USA.