

 <p>PRESENCE A D A P T</p>	<p style="text-align: center;">ADAPT <i>IST-2001-37173</i> <i>Artificial Development Approach to Presence Technologies</i></p>
---	---

Deliverable Item 5.5 Validation of Multisensory Representations

Delivery Date: November 5, 2005

Classification: Public


Status: DRAFT

Responsible Person: Harri Valpola

Partners Contributed: ALL

Short Description: D5.1–D5.4 have described the development of a model for extracting coherent representations from multi-modal data. This deliverable reports initial experiments where the model is applied to multi-modal data recorded by the Babybot humanoid robot.

The status of the report is DRAFT because the analysis is not yet complete.

 <p>Information Society</p>	<p style="text-align: center;">Project funded by the European Community under the “Information Society Technologies” Programme (1998–2002)</p>
--	--

Contents

1	Introduction	2
2	Acquisition of the data	3
3	Development of visual features	3
4	Development of multisensory representations	4
5	Future research directions	4

1 Introduction

The Babybot humanoid robot platform has been used in the ADAPT project because it offers interesting opportunities for studying the development of multi-modal representations. The robot is equipped with colour stereo vision, touch sensors in the palm and fingers and proprioception in wrist and fingers. These senses are familiar to humans and we can therefore put ourselves in the position of the robot and imagine how the world looks and feels like to the robot.

On the one hand this setting is dangerous because humans are so proficient in learning multi-modal representations. Even a baby can do it. On the other hand, actually building a system like this offers interesting new insights about the mechanisms underlying our learning and perceptual abilities exactly because the task is surprisingly difficult.

One of the greatest challenges is that there is so much of data but so little of it is actually useful. Somewhere in the midst of all pixels there is a hand, for instance. We see it so easily that it has been a great embarrassment to machine learning community to admit that we still haven't built nothing close to the human visual system, let alone multi-modal integration.

The prevailing engineering approach has been to concentrate on the first things first: try to get vision and other individual senses working before building a multi-modal perceptual system. This seems very intuitive but unfortunately hasn't been very successful.

The approach in ADAPT has been somewhat different. The idea has been to build a multi-modal perceptual system that is tightly integrated. Borrowing ideas from the human perceptual development, we have developed a model of learning multi-modal representations that uses each modality to guide the development of each other modality.

The model and its motivation and background have been discussed in D5.1-D5.4 (although the model clearly took shape only in D5.3). This deliverable describes experiments with multi-modal data recorded by Babybot. At the moment, the analysis of the data is not yet complete but motivation, hypotheses and expected results are discussed.

2 Acquisition of the data

All data were recorded by Babybot using procedures and hardware described in D4.4. Three datasets have been used:

- 5000 log-polar colour images: The images don't form a continuous sequence but resulted in fixations to salient objects by the attention system (see D4.4).
- Grasping sequences under the control of a preprogrammed grasping module; includes all modalities.
- See-feel: objects placed in the visual field and felt by the hand (grasp reflex initiated by touching the palm with the object). Only vision, touch and hand configuration at the end of grasp can be expected to carry useful information. Eight different objects were used and ten samples of each were recorded.

The main reason for first using visual modality alone was to collect enough image statistics to be able to compress the huge amounts of visual data (76,608 pixels / sample) by unsupervised learning.

The grasping sequences are interesting because the robot perceives the results of its own movements, thus having the possibility to distinguish its own body from the environment. In addition, the ability to visually recognise the shape and position of a hand would be useful not only for visuomotor control of grasping but for interpreting the actions of others, too.

The third dataset may have the potential to develop visual shape features under the guidance of proprioceptive information about shape. With this data, this may well turn out to be impossible because the amount of the data is small compared to the difficulty of the problem. However, we expect that the experience gained with this data will help in designing future experiments exploring the issue more deeply.

3 Development of visual features

Unsupervised approach to learning has been quite successful in modelling the very first stages of human visual processing. Basically this means that human vision seems to make use of the statistical structure found in natural images. In our case, it is necessary to reduce the amount of visual data because the other modalities alone don't provide enough information to guide the development of useful features.

We have used principal and independent component analyses (PCA and ICA) to reduce the dimensionality of the visual data and to extract prominent features. These turned out to be mostly

- Edge features without colour selectivity and
- colour features without orientation selectivity.

Interestingly, this seems to agree quite well with what is known about the human visual system. The agreement may be partly due to the structure of the camera eye which has three colour sensitive receptors like human eyes, and partly because natural images have a statistics like that.

It should be emphasised that unsupervised learning was able to autonomously learn many interesting aspects about the data such as:

- Each pixel in the camera was sensitive to one colour (red, green or blue) but the edge features appeared to selectively discard the information about the wavelengths and only retain information about light intensity.
- The geometry of the camera pixels was learned autonomously. Due to the log-polar geometry of the cameras (see D4.4), edges appear to be curved. This was reflected in the developed edge features which correspond to curved edges on retina but straight edges in the outside world.
- The detectors were denser and with smaller receptive fields in the fovea, reflecting denser sampling.

Although the image dataset used to estimate the features did not have any temporal structure, the features can convey temporal information if their outputs are temporally filtered. Also, although ICA discovers a linear mapping, it is possible to include nonlinearities (such as the absolute value of the edge features) which make further processing stages useful (see D5.4 for more discussion).

4 Development of multisensory representations

By the time of writing, the analysis of the multi-modal representations is still underway. The plan is to study the development of visual features under the guidance of proprioceptive information using the model described in D5.3–D5.4.

Two sets of experiments have been planned:

- Using proprioceptive information about hand position and posture to guide the development of visual hand features.
- Using proprioceptive information about finger configuration to guide the development of visual shape features.

5 Future research directions

The development of the model for multi-modal integration and feature extraction is still in its early phases but it is already apparent that the approach is useful (with several real-world applications ranging from climate data analysis to mobile network signal detection) and has inspired new hypotheses about human perceptual learning.

A rather straight-forward and biologically motivated extension to the model will be to include nonlinear top-down prediction to guide learning. So far we have used linear predictions but it is nowadays thought that the apical dendrites integrate their inputs in a nonlinear fashion (see D5.4, Section 2.3, for discussion about the role of apical dendrites). This may open up new, simple and robust learning methods for nonlinear mappings because, unlike in normal supervised learning algorithms, it is not necessary to accurately predict the magnitude of feature activations. It is enough to find a correlate for the activations. This should obviate the need for prediction error computations and could instead rely on simple correlation-based techniques akin to Hebbian learning.

Another potentially very fruitful research direction that came up as a result of the ADAPT project is a new hypothesis about the role of attention in perceptual learning (see Valpola, 2004 and 2005, References in D5.4). It is well established through psychophysical research that attention has a major role in perceptual learning. However, the underlying mechanisms are poorly understood. Our model of perceptual learning relies on feedback information to guide the development of feature extraction. Anything that can modulate this information, should be able to modulate learning. Attention is clearly a process which modulates the flow of information and, moreover, it is strongly influenced by motivation and goals. This seems to put attention in a good position to mediate the guidance from motivation and goals to perceptual learning.