Understanding Objects and Actions – A VR Experiment

C. Wallraven[1,2], M.Schultze[1], B. Mohler[1], E. Volkova[1], I. Alexandrova[1], A. Vatakis[3], K. Pastra[3]

[1]Max-Planck Institute for Biological Cybernetics, Germany, [2]Department of Brain and Cognitive Engineering, Korea University, Korea,
[3]Institute for Language and Speech Processing, Greece

Contact address: wallraven@korea.ac.kr

Introduction

- POETICON is an EU-funded research project that explores the 'poetics of everyday life', i.e., the synthesis of **sensorimotor representations and natural language in everyday human interaction**. POETICON views the human as a cognitive system as consisting of a set of different languages (the spoken, the motor, the vision language and so on) and aims to develop tools for parsing, generating and translating among them. Through inter-disciplinary research, it contributes to the exploration of what integration in human cognition is and how it can be reproduced by intelligent agents.
- One of the main goals of POETICON is to **provide a large, detailed corpus of recordings of human actions** (such as movements and facial expressions), **human-object interactions** (such as picking up an object or learning a novel object by exploring it), and **human-human interactions** (such as preparing a dinner, or cleaning the kitchen) in every-day contexts.
- What sets our work apart from previous, related efforts is the care taken to provide measured **ground-truth data by means of high-tech recording equipment** such as motion capture of human body movements and objects together with synchronized high-definition camera footage. The data recorded within the project is not only useful for modeling human (inter)actions through computational analysis, but also for novel, perceptual experiments within the context of action understanding.
- **Here, we present results from a perceptual experiment on the POETICON corpus that investigates peoples' ability to interpret the contents of an everyday scenario depending on the amount of information that is provided visually.**
- People watch a computer animation in which two avatars are interacting in a kitchen environment handling different, clearly visible objects. Are they able to recognize the scene? And would they still recognize it when the key objects are only represented as bounding boxes? What about when no objects at all are present? Will the actions alone be enough to uniquely determine the scenario?

Conclusions

- **People are able to recognize the different scenes if they see the avatars and the objects.**
- **Some scenes were easily interpretable from actions alone (even quite complex ones such as making a salad), whereas others were dramatically affected by the loss of context object information.**
- **More detailed analyses will need to be done to determine whether this effect is due to the manipulated objects, or perhaps to the environmental objects.**
- **With less information in the animations, more 'body movements' and fewer 'actions with objects' were used. When people had more information they described the 'high-level' actions instead of just 'body movements' of the avatars.**

Recordings and animation

- First, we selected 6 different scenes placed in a kitchen/dining-room scenario.
  - **cleaning the kitchen**
  - **preparing a Greek salad**
  - **setting the table**
  - **changing the pot of a plant**
  - **preparing Sangria**
  - **sending a Parcel**

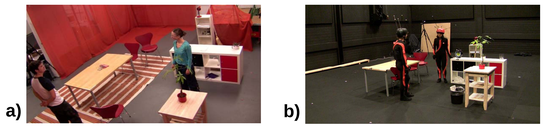

Fig.1: Setup :a) Natural recording, b) High-tech recording

- Each scene was recorded with 4 different pairs of actors in a natural kitchen/dining-room setting (see Figure 1) using motion capture of bodies and objects. The scripts included dialogue, actions and facial expressions.
- All scenes were recorded with 2 synchronized high-definition camcorders. The movement of the 2 persons was captured with **2 Moven motion capture suits** (Xsens technologies) and their position was tracked with the **Vicon motion capture system**, using 2 helmets with tracking markers. In addition, for each scene, several key objects were also tracked with the Vicon motion capture system.
- From the motion capture data, **animations** were created using 3DS Max. These animations include the two persons, realistic 3D models of the furniture (kitchen-table/ cupboard, table, service table and 2 chairs), as well as realistic 3D models of the Vicon-tracked objects. The motion capture data from the Moven suits was first imported into 3DS Max and positional and rotational drift was corrected manually using the Vicon data and the movie from one of the overview cameras as a reference. The objects were animated using the Vicon data and—where applicable—in addition attached to the hands of the manipulating individual to anchor the animation.
- These animations were then imported into Virtools to provide further flexibility in interactively manipulating the content of the animation for our experiments.

Stimuli and Setup

- In our experiment, the animations were shown to three groups of 16 participants each in three different conditions:
- **Condition 1: avatars and high-res objects**
- **Condition 2: avatars and low-res objects (boxes)**
- **Condition 3: only avatars, no objects (see Figure 2).**
- Participants watched the animations two times and were then asked to give a **title** to the scene, as well as to **describe** the actions of the two people and the used objects in the form of a script.

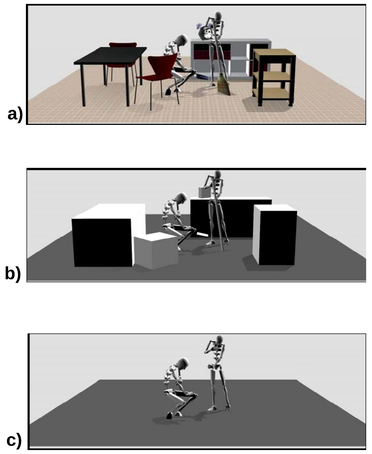

Fig.2:Screenshot from the same animation frame of Conditions 1-3: a) high-res, b) low-res, c) no-objects.

Results

- People were clearly able to recognize all 6 scenes (recognition rate: 75-100%) in Condition 1.
- When no objects were visible (Condition 3, no objects), the first 3 scenes (cleaning, preparing a salad and setting the table) were still recognized (recognition rate: 70-80%), but the other 3 scenes were not (recognition rate: 0-10%).
- In Condition 2, in which only bounding boxes were present, we observed a significant improvement in recognition rate compared to Condition 3 for the parcel scene.

(see Figure 3)

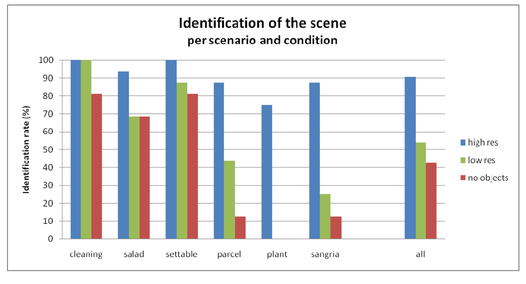

Fig.3: Recognition rate per scenario and condition

- Additionally the total number of verbs and nouns and the number of different verbs and nouns in the description were analyzed.
- There was no clear difference in the number of verbs and nouns between the 3 conditions.
- In Condition 2 more verbs were used.
- The total number of nouns was higher in the conditions with objects.

(see Figure 4)

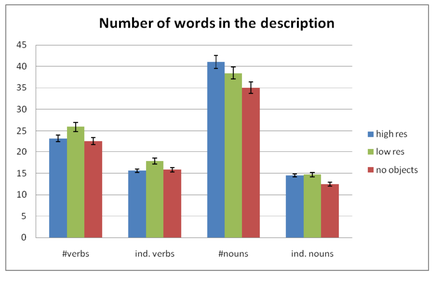

Fig.4: Number of verbs and nouns in the descriptions

- We also observed a difference in how people described the scenes.
- We separated the verbs into 'actions with objects' (e.g. cleaning, taking, sweeping...) and 'body movements' (e.g. walking, looking, talking...).
- As Figure 5) clearly shows, with less information in the animations, more 'body movements' and fewer 'actions with objects' were used.

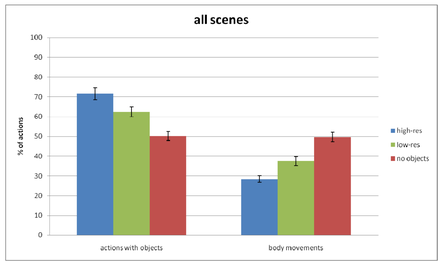

Fig.5: Number of'actions with objects' and 'body movements' per condition (in percent)