Joint Virtual Reality Conference of EuroVR - EGVE - VEC (2010)
T. Kuhlen, S. Coquillart, and V. Interrante (Editors)

Understanding Objects and Actions - A VR Experiment

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

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

Abstract

The human capability to interpret actions and to recognize objects is still far ahead of that of any technical system. Thus, a deeper understanding of how humans are able to interpret human (inter)actions lies at the core of building better artificial cognitive systems. Here, we present results from a first series of perceptual experiments that show how humans are able to infer scenario classes, as well as individual actions and objects from computer animations of everyday situations. The animations were created from a unique corpus of real-life recordings made in the European project POETICON using motion-capture technology and advanced VR programming that allowed for full control over all aspects of the finally rendered data.

Categories and Subject Descriptors (according to ACM CCS): J.4 [Social and Behav. Sciences]: Psychology—

1. Introduction

Reproducing an act by the interplay of perception and action, and using fine natural language for communicating the intentionality behind the act is what Aristotle called 'Poetics'. POETICON is an EU-funded research project that explores exactly this '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), 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 such 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*. To illustrate, imagine a computer animation in which two avatars are interacting in a kitchen environment handling different, clearly visible objects. If the two people were, for example, preparing a drink, surely everyone would be able to infer this from a few key interactions and manipulations of tell-tale objects. However, would people still be able to infer that a drink was being prepared 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? In effect, we here ask a question about the effectivity of pantomime—however, instead of highly trained, over-exaggerated actions that pantomimes usually perform, the VR experiments described in the following allow us a much more controlled approach.

2. Recordings and Animations

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). In the con-

text of this work, each scene was recorded in a natural kitchen/dining-room setting using motion capture of bodies and objects. The scripts included dialogue, actions and facial expressions. After the actors learned the script and practiced the scene several times, the recordings were started. Each scene was recorded with 4 different pairs of actors, each pair performing each scene 3 times to gather additional recording data. The recorded scenes are between 2 and 7 minutes long (depending on the scene and the pair of actors). 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). The position of the 2 persons was tracked with the Vicon motion capture system, using 2 helmets with tracking markers. In addition, for each scene, several key objects were identified and 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.

3. Experiment: Design and Analysis

In our experiment, the animations were shown to three groups of 10 participants 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 1). The conditions were switched interactively within the Virtools environment. 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.

As shown in Figure 2a), people were clearly able to recognize all 6 scenes (recognition rate: 75-100%) in Condition 1. At the other extreme, when no objects were visible (Condition 3, no objects), however, the first 3 scenes (cleaning, preparing a salad and setting the table) were still recognized (recognition rate: 60-80%), but the other 3 scenes were not (recognition rate: 0-10%). Thus, 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. Interestingly, for Condition 2, in which only bounding boxes were present, we observed a significant improvement in recognition rate compared to Condition 3 for the parcel and sangria scene. More detailed analyses will need to be done

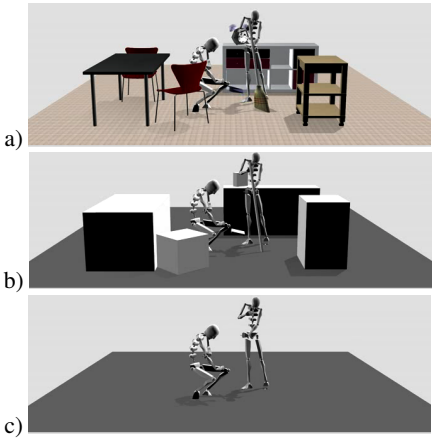

Figure 1: *Screenshot from the same animation frame of Conditions 1-3: a) high-res, b) low-res, and c) no-objects.*

to determine whether this effect is due to the manipulated objects, or perhaps to the environmental objects. We also observed a difference in *how* people described the scenes. As a first analysis for the text descriptions, we separated the verbs into 'actions with objects' (e.g. cleaning, taking, sweeping...) and 'body movements' (e.g. walking, looking, talking...). As Figure 2b) clearly shows, with less information in the animations, more 'body movements' and fewer 'actions with objects' were used.

The analyses reported here present only a brief look into the work that has been and will be done on the ability of the human to interpret complex (inter)actions. Through the use of state-of-the-art VR technology in this and future experiments, we hope to be able to shed further light on this fundamental cognitive capability.

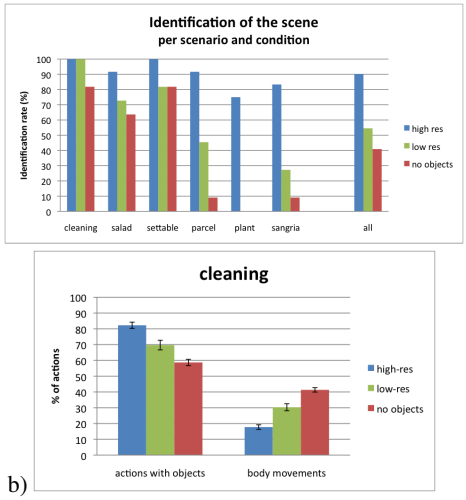

Figure 2: *a) Percentage of correctly recognized titles per scenario, b) Percentage of verbs describing actions and body movements for the cleaning scenario*