The POETICON enacted scenario corpus - a tool for human and computational experiments on action understanding

Christian Wallraven, Michael Schultze, Betty Mohler, Argiro Vatakis, Katerina Pastra

***Abstract*— A good data corpus lies at the heart of progress in both perceptual/cognitive science and in computer vision. While there are a few datasets that deal with simple actions, creating a realistic corpus for complex, long action sequences that contains also human-human interactions has so far not been attempted to our knowledge. Here, we introduce such a corpus for (inter)action understanding that contains six everyday scenarios taking place in a kitchen / living-room setting. Each scenario was acted out several times by different pairs of actors and contains simple object interactions as well as spoken dialogue. In addition, each scenario was first recorded with several HD cameras and also with motion-capturing of the actors and several key objects. Having access to the motion capture data allows not only for kinematic analyses, but also allows for the production of realistic animations where all aspects of the scenario can be fully controlled. We also 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. These results can serve as a benchmark for future computational approaches that begin to take on complex action understanding.**

I. Introduction

How do we recognize actions? How do we understand that someone is engaged in a complex task, such as preparing dinner? Making inferences about complex actions and scenarios from visual input alone is a seemingly easy and trivial task for the human brain - the amount of data and detail that humans need to process to arrive at these interpretations, however, is far from trivial. A deeper understanding of how humans are able to interpret human (inter)actions not only informs the perceptual and cognitive sciences, but it also lies at the core of building better artificial cognitive systems for action understanding. Action understanding and modeling have a long history in computer vision and computer graphics. However, the question for any of these systems is: how do we best evaluate their performance? Here, we introduce a new resource for both human and computational experiments that can serve as a benchmark in both fields: the POETICON enacted scenario corpus[1].

Reproducing an act by the interplay of perception and action, and using natural language for communicating the intentionality behind the act is what Aristotle termed '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. One of the main goals of POETICON is to provide a large, detailed corpus of recordings of human actions (movements and facial expressions), human-object interactions (picking up an object), and human-human interactions (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 comparative 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[2].

In the first part of the paper, we present the technical details behind the POETICON corpus and describe its contents and structure. In the second part of the paper, we present results from an initial 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 this idea, imagine a computer animation with avatars based on two persons interacting in a kitchen environment handling different, clearly visible objects. If the two persons 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 we 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? The analysis of the human data provided by this experiment yields a potentially important benchmark for computational experiments on action understanding that try to parse longer, more complex events into a structured series of sub-actions and interactions.

II. Related work and motivation

Humans are capable of inferring an incredible amount of detail about the actions of people, as well as their interactions. As far as action understanding is concerned, much

This work was supported by the EU-project POETICON ICT-215843.

C. Wallraven is with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea. Contact: wallraven@korea.ac.kr

C. Wallraven, M. Schultze, B. Mohler are with the Department of Human Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, Tübingen, Germany

A. Vatakis and K. Pastra are with the Institute for Language and Speech Processing, Athens, Greece

[1]see http://poeticoncorpus.kyb.mpg.de.

[2]For an initial description in the context of *linguistic* analysis, see [12].

research has been devoted to understanding how humans process very sparse motion displays, also known as point-light figures, invented by Johansson in the late 1970s [7]. These displays are created by attaching a light source to the joints of a person. The movement of these points is not only enough to provide strong cues about the person's sex [18], and mood [14], but also can be used to reliably predict actions [3]. The exact nature of the information that is processed by humans in this context remains under dispute [2], but it is clear that a great deal of information is provided by the motion of relatively few joints of the human body.

Accordingly, many databases are available that deal with action recognition. For human actions, in perception research this subject matter is also often found in the context of "biological motion", with databases of motion capture data including [9], [10]. Another well-known database is the CMU Graphics Lab Motion Capture Database (available at `mocap.cs.cmu.edu`), which contains motion capture data of various actions in categories such as locomotions, pantomime, and expressions. Whereas it contains a lot of data on locomotion patterns, longer activities as well as human interactions are very much underrepresented and in addition only very loosely organized. As far as human-human interactions are concerned, a recent motion-capture database [10] contains 20 elementary interactions (such as point to the ceiling, I am angry, pick up, etc.) each performed by one male and one female couple. In the context of computer vision, well-known annotated databases containing *video data* of human actions (as opposed to either motion capture data, or video data of actors in motion capture suits) include the IXMAS dataset for recordings in a controlled environment [19] and the Hollywood dataset for a larger, more varying dataset [11]. For a good overview of recent work in computer vision concerning recognition of actions from video, also see [13].

As far as action understanding on a larger scale is concerned, research on human perception has shown that most longer events are usually broken down into coherent sub-actions. For common everyday behaviors, these sub-actions usually correspond to the goals underlying the actor's actions [8]. Still snapshots from those boundaries are usually remembered better (for both adults and even for young infants [15]). Recent results from neuroscience have shown that the synchrony in the interpretative act of understanding and being immersed, for example, a suspense movie even extends to the brain itself by activating the same brain regions across individuals time-locked to plot events in the movie [5]. A prominent theory in this context is the Event Segmentation Theory [8]: this theory is based on a predictive model in which the observer creates a representation of the current action that is happening currently and uses it to predict future events. As soon as prediction errors occur, the current event models are updated which results in a transient increase in processing and therefore helps to encode the whole sequence of events into long-term memory. This is a compositional model in the sense that repeating patterns in the input can be used to improve prediction. It is also hierarchical in the sense that events and actions are represented at different granularities and levels (see also [1] for a review of different hierarchical, computational models used for modeling human action understanding).

As stated earlier, relatively few, well-structured datasets are available that contain video and/or motion capture data from extended scenarios, or longer action sequences (that is, more than 30 seconds). Perhaps the closest data source that is currently available is the kitchen dataset, which is part of the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database [6]. The dataset is still under development and currently contains data from 25 subjects each of whom was asked to prepare five different recipes in the kitchen. It contains video data from five stationary and one mobile camera, audio data, some inertial motion data from the subjects hand, as well as motion capture data of the full body including the hand. RFID tags are used to identify some nearby objects that are handled by the person.

Another, very similar dataset - also in terms of the setting chosen - is the TMU kitchen dataset [17], which contains recordings of a few (up to four) subjects, who are setting a table according to a pre-defined layout. It contains multi-view camera data, RFID data from a few objects, as well as kinematic data from a marker-less body tracking software. The table-setting task was done in a few different ways (inefficient robot-like, as well as efficient human-like strategies for setting the table), and some additional recordings on simple actions are available, making this corpus a good resource for learning simpler, more constrained human actions (although it faces the difficulty that only a few subjects are available, with no repetitions of the full scenario).

With our corpus, we have aimed at a broader range of everyday scenarios, which are, in addition, available *both* as a natural video recording and as a comparative, kinematic recording. In the comparative recordings, we collected motion capture data (from both people and a few key objects), as well as video and audio data. In the natural recordings, the same actors act out the script in a realistic surrounding without any disturbing marker sets, cameras, cables, etc. This data therefore provides an excellent testbed for computer vision algorithms that need to work in realistic environments. Additionally, the natural recordings were recorded from multiple view angles. Finally, in order to provide additional data for training and testing, we also recorded all scenarios *three times* with each pair of actors in both recording settings. Taken together, our data therefore is ideal for testing the generalizability of computational approaches to action understanding in both kinematic and realistic video data.

III. The POETICON corpus

In the following, we describe the every-day scenarios and the recording settings and equipment of the corpus.

A. Scenarios

First, we selected 6 different scenario that can take place in a typical kitchen/dining-room setting. The scenarios were: cleaning the kitchen, preparing a Greek salad, setting the

table, changing the pot of a plant, preparing Sangria, and sending a Parcel. Note that all of these events contained many sub-actions that actually need to happen in order for the whole event to unfold and to become meaningful. We carefully wrote the scripts such that they included spoken dialogue, several interactions with key objects, body movements across the recording volume, as well as several small situations that might elicit facial expressions (although the latter are not the focus here).

An excerpt from the cleaning scenario, for example, reads:

The floor is a bit dirty, especially now that you have changed the pot for the new plant. Before preparing dinner, you need to clean up a bit.
Person A: The plant sure looks good, but now the place is a bit dirty. We cannot have dinner like that!
Person B: Ok, lets tidy up quickly. What shall I do?
Person A: Get a dry cloth and sweep the dirt off this chair! I will sweep the floor.
Person B: SurePhew! I did not expect such a mess, when you said we would change the pot of the plant.!
Person B gets a cloth and cleans a chair from dirt that had been dropped while changing pots. Meanwhile, Person A picks up the broom and swipes the floor quickly. Person B brings the dustpan and holds it firmly in front of the broom, so that Person A pushes the dirt from the floor onto the dustpan. Person B picks it up and disposes of the dirt in the trashcan, but realizes that the trashcan is full (Person B then empties bottles and cans).

Both natural and comparative recordings took place in a kitchen / dining-room setup that contained a large table, four chairs, a high-table, and a sideboard with several compartments as the main pieces of furniture. The main objects that were handled depended on the scenario, and included objects such as a large alarm clock, a plant, a dustpan, a broom, etc. In addition, kitchen implements like cutlery and dishes/cups were also part of several scenarios.

After the actors learned the script and practiced the scenario several times, the recordings were started. Each scenario was recorded with 4 different pairs of actors. In addition, each actor pair performed the whole string of events three times to gather additional recording data and to provide information about intra-individual variance in acting out the scenarios. The resulting recordings are between 2 and 7 minutes long (depending on the scene and the pair of actors).

B. Natural recordings

The natural recordings took place in the kitchen setup and were recorded using 5 high-definition camcorders (Canon HF100) with resolution of 1960x1400 pixels, 2 of which carried a wide-angle lens (DHG 0,75x Wide Angle Converter 52 mm). The cameras also recorded sound with stereo microphones. The 5 cameras were placed to afford different views onto the action and interaction spaces in the room and are shown in Figure 1 n1)-n5). The two wide-angle cameras n1, n2 were used to record overviews of the whole scene and were placed at opposite corners of the room. The other three cameras focused on the kitchen table, the counter-top,

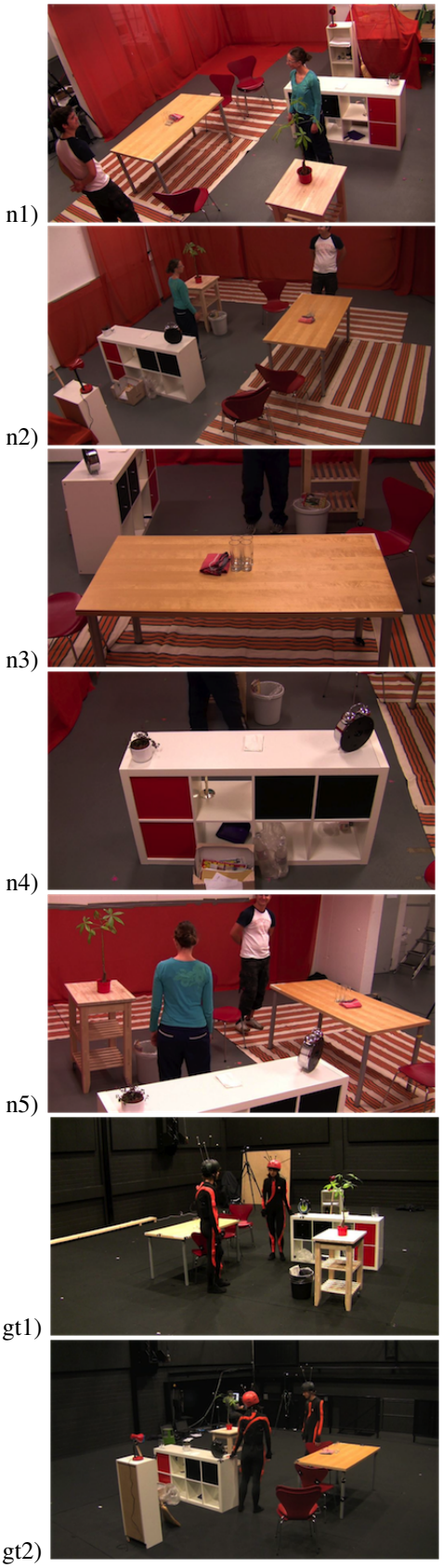

Fig. 1. Screenshots from the five camera perspectives of the same frame for the natural recording condition (n1-n5) and the two wide-angle perspectives of the comparative condition (gt1-gt2)

and the interaction space in which many of the person-to-person interactions took place. The videos recorded with the high-definition camcorders were cut and then exported into QuickTime-movie format (.mov, highest quality settings). Synchronization was achieved by cutting according to a starting signal on the audio tracks of the cameras.

C. Comparative, kinematic recordings

All scenarios for the comparative, kinematic recordings were recorded in the same kitchen setting. We first set up 2 synchronized, wide-angle HD camcorders placed at opposite ends of the setup similarly to the natural recording setting. Most importantly, however, the movement of the 2 persons was captured with 2 Moven motion capture suits (Xsens technologies) that yield high-definition data about a person's articulated movement based on inertial motion sensors. The big advantage of the motion suits is that they are not affected by occlusion as the sensors are mounted on the body directly. This was a critical feature for our scenarios, as we tracked the movement of two people in a confined space at the same time. As the inertial sensors are prone to drift, the position of the 2 persons in the room was tracked with the Vicon motion capture system, using 2 helmets with tracking markers. In addition, for each scene, several pieces of the furniture and key objects were fitted with markers and also tracked with the Vicon motion capture system. All kinematic recordings were resampled to the 60Hz base rate of the Moven suits.

All actors went through a short calibration and test phase in order to provide reliable motion capture data. In addition, all actors first did the natural recordings to become more familiar with the scenarios for the comparative recordings.

The Moven data was first post-processed to provide optimal reconstruction of each actor's actions (including setting foot contact points, re-starting the calculation of the inverse kinematics at difficult body postures, etc.). This data was then exported as a standard bvh-file format. The Vicon data was cleaned up and checked for lost markers and then annotated to provide information about the various rigid objects in the scene (the helmets of the two actors, furniture, and key objects). Synchronization between the camera videos and the kinematic recordings of both Moven and Vicon data was achieved by cutting the data according to both the audio track, and the start of the actions for the kinematic recordings. Kinematic recordings from Moven data and Vicon data were put into correspondence by hand-written software, as the Moven-suits suffer from position- and rotation-drift. We therefore read in the Vicon data of the tracked helmets of the two actors and slaved the relative kinematics to the coordinate system provided by the Vicon data - we are currently developing a better integration method of the two modalities to increase the quality further.

D. Consistency

As a detailed kinematic analysis is out of the scope of this paper, here we present results from a simple, but effective occupancy grid analysis that serves to show that people are consistent across the three repetitions, and that data is consistent within the same scenario type across actor pairs.

For this analysis, we took the kinematic data of the Vicon tracked helmet of one person, which roughly provides an indication of where the person stood in the room, and counted how many times this marker entered one of 100x100 equally spaced cells in the room. This occupancy grid allows us to effectively compare kinematic data without employing more involved methods such as time series warping.

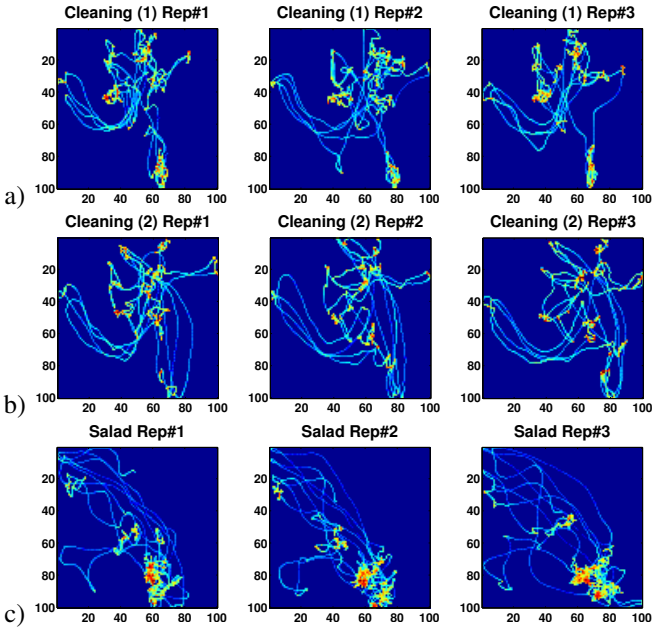

Fig. 2. Spatial occupancy grids for all three repetitions of the cleaning scenario for actor pairs 1 (a) and 2 (b), and for the salad making scenario for actor pair 1. Note the similarities within repetitions and within scenarios.

Examples of these occupancy grids are shown in Figure 2 (data is color coded with blue meaning little occupancy and red meaning high; to make the plot better visible, a logarithmic color scale was used). Figure 2a shows the grids for all three repetitions of the same actor pair for a cleaning scenario - in addition to clearly showing that different time was spent at different positions in the room, all three grids look very similar at first glance. Similarly, Figure 2b shows the grids for the repetitions of another actor pair - again all three grids look similar. This is also true for an example from the salad making scenario (Figure 2c). However, the occupancy grid for this scenario looks markedly different from the other two rows. In addition to the consistency within actors, there seems to be also consistency across actors within the same scenario - in other words, the occupancy grids might be used as a very coarse identification signal for what scenario it might be.

Figure 3 shows the correlation matrix obtained by correlating the data contained in the occupancy grid across 3x3 cleaning scenarios and 1 salad-making scenario (this data is a zoom of the total correlation matrix, the pattern for the other scenarios is similar, however). This correlation matrix confirms the data from Figure 2a,b in that data across *repetitions* contains a similar variance as data across *actors*.

The data of the first 12x12 block shows no clear sign of 4 separate 3x3 block structures which would clearly separate within-actor variation from across-actor variation.

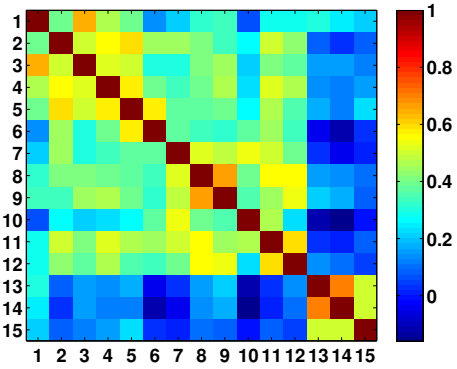

Fig. 3. Correlation of the spatial occupancy grid data from all 3 repetitions of 4 actor pairs for the cleaning scenario (1-12) and 3 repetitions of one actor pair for the salad scenario (13-15). Note the clear block structure separating the scenarios.

In conclusion, this straightforward analysis shows that our dataset is both consistent across repetitions and also across actor pairs - this analysis, however, could be very much improved by incorporating proper kinematic analysis and sequence aligning to characterize the scenarios further (see, for example, [4]).

IV. Animations

Given that detailed motion capture data is available, it becomes possible to create computer animated versions of the every-day scenarios. Such animations provide full control over all parameters and therefore open up novel possibilities for perceptual and also computational experiments.

For our purposes, we created animated scenarios from the motion-capture data of one actor pair for all six scenarios using 3DS Max. For this, we first created coarse skeletal models of the two actors fitted to their body size with the standard "biped" animatable character skeleton provided in 3DS Max (note, that the appearance of the characters can be easily changed as required). In addition, we created realistic 3D models of the furniture (counter-top, table, service table and 2 chairs), as well as realistic 3D models of the Vicon-tracked key objects for each scenario.

The motion capture data from the Moven suits was 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. This provided a much better integration of the two kinematic datasets than the automatic drift correction mentioned above. The key objects were animated using the Vicon data and—where applicable—in addition attached to the hands of the manipulating individual to better anchor the animation.

These animations were then imported into a real-time animation environment (Virtools) to provide further flexibility in interactively manipulating the content of the animation for our experiments. Examples of possible on-line manipulations include appearance and inclusion of objects, properties of body kinematics, as well as general rendering parameters (lighting, viewpoint, etc.). An example of an animation frame is shown in Figure 4a - the other two images will be explained below.

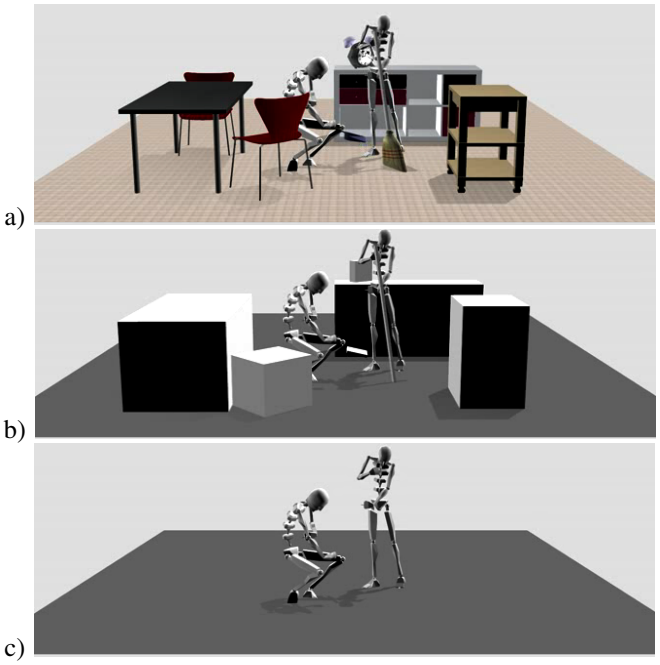

Fig. 4. Screenshot from the same animation frame of Conditions 1-3: a) high-res, b) low-res, and c) no-objects.

V. Experiment on Action Understanding

We know that humans can correctly identify human actions happening at short time-scales already from sparse visual information such as point-light displays: these include locomotion behaviors (walking, running, limping), simple object manipulations (picking up, carrying), as well as basic human-human communicative gestures (pointing, waving, greeting). Little research, however, has been conducted on how humans integrate information over longer time-periods to make sense out of extended activities, such as the ones recorded in our every-day scenarios. In addition, having access to controlled animation data makes it possible to create novel experimental paradigms for action understanding. To demonstrate this, here we report results from a first experiment that tests how well people can interpret a scenario given *different levels of visual information*.

A. Experimental design

One of the common tasks to test how well people process events is simply to collect verbal or written descriptions of the events. These descriptions are then analyzed in terms of their linguistic properties (such as number of words, number of verbs, etc.) as well as their semantic content (such as what is described at which level of granularity, etc.). In our experiment, participants therefore were able to see 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. We chose

to use this more memory-intensive task, as we wanted to tap into the stored representation of the scenario, rather than into the immediately formed one, which would be accessible by having participants provide a live voice-over *during* the playback of the animation.

The different levels of visual information were achieved by three conditions (example screenshots of the three conditions for the same animation frame are shown in Figure 4):

1) Condition 1: avatars and high-res objects
2) Condition 2: avatars and low-res objects (as bounding boxes)
3) Condition 3: only avatars, no objects

Condition 1 constitutes the baseline condition - people have presumably full access to all necessary information about the actions and objects that are needed to interpret the scene. This information, however, is purely visual - even though the actors of course provided detailed dialogue, our experiment only tests how well people can understand the events based on body movements and object interactions. Condition 3 is akin to asking how well people can interpret pantomime - an actor is miming to perform actions, which are, however, devoid of any object that is being acted upon. One of the advantages of computer animations is that this condition is actually derived from the real set of actions which contain the objects - this guarantees that the only information differing between the full animation and the pantomime condition is truly the absence of objects. This condition is interesting, as it tests humans' ability to infer complex events from actions and body movements alone - note also, that our animations did not include any facial animation, thereby eliminating a usually rather active channel in pantomimes. Finally, the animations which only show objects as bounding boxes (Condition 2) constitute an intermediate case - in many cases, having just a rough idea of the size of the object might be enough to disambiguate actions and to imply the relevant objects. In addition, relative object location is also key to identifying an action (especially object location in relation to the human body and the movement effector in particular). If size and relative location were robust enough cues, we would expect no difference between these animations and the animations containing the clearly identifiable objects.

The experiment was run as a between-group design with 48 participants randomly assigned to the three conditions, such that 16 participants saw each animation style. Participants were compensated at standard rates for their participation. Before the experiment started, participants were informed about the task and the exact experimental procedure. Each of the six animations was shown twice (the order of the animations was random for each participant, the different conditions in the three groups were switched directly in the Virtools environment). Participants had to watch the first repetition closely paying attention to the actions and potentially to (implied) objects in the scene. They then had time to write down a script-like summary of the scenario into a text-editor. After this, the animation was repeated and participants were allowed to correct their description. After each scenario, participants were allowed to take a brief break. The whole experiment took around 1.5 hours.

B. Global interpretation ability

The most straightforward analysis is to ask whether participants were able to correctly categorize the title of the scenarios, which provides a global measure of event understanding. For this, we rated the titles given to the scenarios for each participant as either correct or incorrect - the overall data is shown as the three rightmost bars in Figure 5. The data clearly shows that Condition 1 with all objects fares best, followed by Condition 2, and Condition 3. This results demonstrates that it is clearly possible to categorize scenarios based on visual information alone, even in the absence of objects. Note also, that the task was not a forced-choice task in which participants had to chose between one of the six categories - rather, it was a much more difficult free naming task, in which participants had to come up with a suitable title. In this context, the 42% recognition accuracy for the pantomime condition (Condition 3) is, indeed, impressive. However, as Figure 5 also shows, the data varies considerably depending on the condition *and the scenario* - we therefore need to interpret this data interaction more closely.

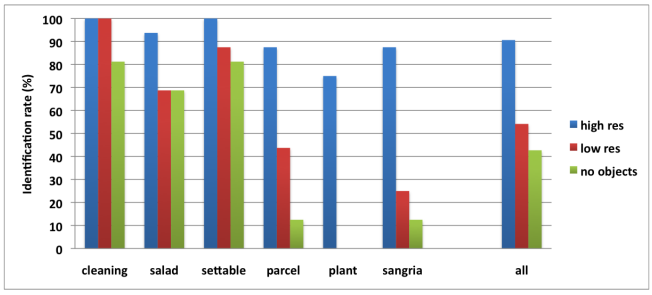

Fig. 5. Percentage of correctly recognized titles per scenario

Participants were clearly able to recognize all 6 scenes (varying between 75-100%) in Condition 1 with an average accuracy of 91%. At the other extreme, in Condition 3, when no objects were visible, the first 3 scenes (cleaning, preparing a salad and setting the table) were still recognized (recognition rates: 81%, 69%, 81%, respectively), but the remaining 3 scenes were not (recognition rate: 13%, 0%, 10% resp.). 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 (potting a plant). Interestingly, for Condition 2, in which only bounding boxes were present, we observed a significant improvement in recognition rate compared to Condition 3 for two of the little recognized scenes (parcel and sangria).

In summary, despite the fact that participants had no information as to what scenarios they were going to see, recognition results were surprisingly good for the conditions with little or reduced visual information. For two out of three scenes, indicating object sizes of key objects did seem to provide a valuable cue to the scenario category.

C. Analysis of action type

We also observed a difference in *how* people described the scenes. For this, the texts were subjected to a standard, automatic computational linguistic procedure that automatically extracts verbs from the texts. As a first analysis for the text descriptions, we separated these verbs into 'actions with objects' (e.g. cleaning, taking, sweeping...) and 'body movements' (e.g. walking, looking, talking...). As Figure 6 clearly shows, with less information in the animations, more emphasis was put on describing 'body movements' rather than 'actions with objects'. This trend was observed for all six scenarios. The *number of total verbs* in the descriptions, however, did not change across conditions, indicating that participants produced descriptions containing roughly equal numbers of actions.

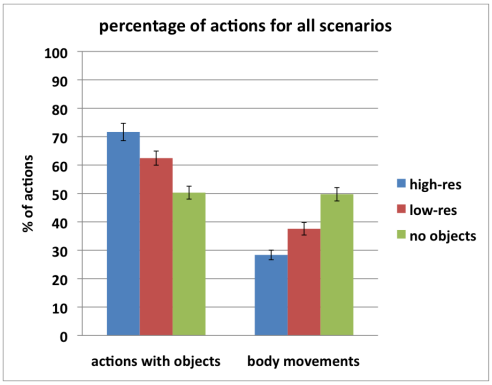

Fig. 6. Percentage of verbs describing actions and body movements

D. Recognition of individual actions

At the finest level of granularity, we analyzed how well participants were able to recognize each individual action / object sub-action within the scenarios. For this, we first determined a rough "ground-truth" script against which to evaluate the responses of the participants. An example of such a script is shown in Table I for the cleaning animation. We then matched the individual scripts against this table and determined which action/object pair was correctly identified. In the following, we will focus on the data for one scenario - the cleaning scenario.

First of all, we evaluated the intra-class correlation - we had 16 participants for each of the three conditions and it might be that their performance was actually too varying to provide a consistent picture. Based on [16], we used a well-established measure for rater reliability, more specifically ICC(2,'average'), which measures how well participants fit with the average performance. This measure is between 0 (no consistency) and 1 (full consistency) - the values for the three conditions were 0.88, 0.90, 0.90, respectively, indicating a relatively good agreement in terms of the interpretation performance in the three conditions.

On average, recognition accuracy for the cleaning scenario was 34.2% in Condition 1, 26.7% in Condition 2, and 18.5% in Condition 3 - a result, that one might have expected given our previous findings. What is more interesting, however, is to look into the data for individual actions as shown in Figure 7: looking at the blue bars for Condition 1, it seems that only a few actions were uniquely determinable overall (but see below). Several actions involving the broom were equally recognizable across all three conditions, whereas actions with the dustpan and the clock were less so. One reason might be that actions related to the broom are less context-dependent - the sweeping motion is instantly recognizable, for example.

Again, we want to stress that our results were obtained with free-form text - the performance levels we observed are therefore quite impressive in all conditions. It should also be noted that the interpretation at this fine-grained level - necessarily - has many potential sources of noise. First, a low recognition rate does not mean that the action/object pair was not recognized at all - it could simply be that participants did not remember the action in their script, or that they did not deem it important for the overall event structure. Second, not all objects that were named in the ground-truth script for the scenario were actually part of the animation (see Table I), such that it might be no surprise that the action involving the lamp would not be recognized well. Finally, since there were two people involved, sometimes actions happened that might be less well visible from the viewpoint from which the animations were rendered. We plan to address these questions in follow-up experiments in more detail.

TABLE I

TIMELINE OF ACTIONS AND OBJECTS FOR THE CLEANING ANIMATION. OBJECTS IN BOLD ARE PART OF ANIMATIONS IN CONDITIONS 1 AND 2.

time	action/object
0:03	A and B talk
0:13	A takes a rag from the **sideboard**
0:16	A wipes off a **chair**
0:13	B gets a **broom**
0:18	B uses the **broom**
0:29	A gets a **dustpan** from the **sideboard**
0:30	B swipes the dirt onto the **dustpan**
0:34	B returns the **broom**
0:37	A puts the **dustpan** onto the trolley
0:38	A takes a bottle and paper from the dustbin
0:46	A returns them to behind the **sideboard**
0:57	A empties the **dustpan** into the dustbin
1:02	A returns the **dustpan** to behind the **sideboard**
0:56	B gets a garbage bag and puts into the dustbin
1:12	B polishes/dries glasses at the table
1:14	A returns the **dustpan** into the **sideboard**
1:22	A walks behind the **sideboard**
1:24	A wipes off the lamp
1:29	A switches the lamp on and off
1:34	A gets a light bulb from **sideboard** and replaces lamp's bulb
1:37	B goes over to the **sideboard**
1:39	B wipes off the **sideboard** and the **alarm clock**
1:46	B lifts the **alarm clock** and puts it onto the **sideboard**
1:54	B gets batteries from **sideboard** and replaces **clock's battery**
2:00	A takes the old light bulb to the dustbin
2:06	A goes over to the **sideboard**
2:07	A takes the **clock**
2:08	A sets the **clock**
2:06	B wipes off the **sideboard**
2:16	A and B talk

VI. CONCLUSION

This paper has presented a novel corpus for action understanding that has several distinguishing features: First,

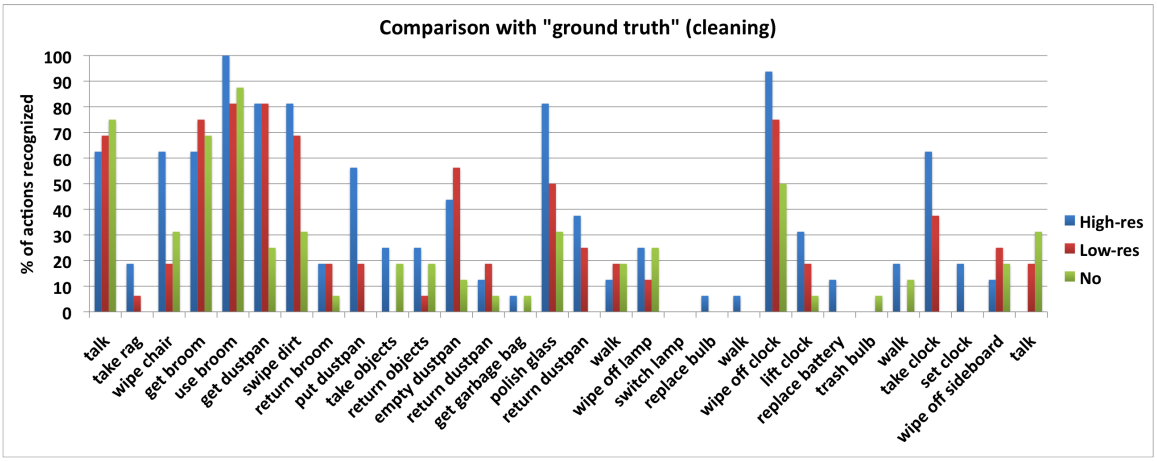

Fig. 7. Accuracy of participants for recognition of individual actions/objects. The x-axis description follows Table I

the corpus shows natural, yet script-controlled long event recordings at all interaction levels (human-human, human-object) in well-defined scenarios. Second, data from several individuals was recorded to provide further data on intra-individual variance. An occupancy grid analysis has shown that the data is consistent across both variations - further kinematic analyses are needed to address this issue fully. Third, the corpus also contains both natural as well as matched high-tech recordings which makes it suitable both for cognitive experiments and computational modeling. Finally, we provide multimodal data (multiple camera angles, audio, 3D kinematic data, 3D tracking data of focus objects) from different sensors that can support analysis across a large number of dimensions from 2D analysis of video streams up to complex models of 3D articulation.

In addition, we have presented results from a first experiment on human action understanding that has used data from the corpus. We have shown that humans are not only capable of inferring the overall, global scenario category, but also that we are rather adept at extracting detailed event structure - even in the absence of detailed visual information. Our experiments show that size information is often enough to infer the relevant objects in the context of the actions and vice versa. The analyses reported here of course 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. Finally, it will be interesting to see how well computer vision algorithms (based on both kinematic and, perhaps even more challenging, on only visual information) will be able to interpret our scenarios. Given that the scenarios seem consistent, the data can be easily split into training and testing sets to try and infer both global and local event structure.

REFERENCES

[1] M. M. Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, 12(5):201–208, Jan 2008.

[2] A. Casile and M. Giese. Critical features for the recognition of biological motion. *J Vision*, 5(4):348–360, Jan 2005.

[3] W. Dittrich. Action categories and the perception of biological motion. *Perception*, 22(1):15–22, Jan 1993.

[4] G. Guerra-Filho and Y. Aloimonos. The syntax of human actions and interactions. *Journal of Neurolinguistics*, in press, Jan 2010.

[5] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, Jan 2004.

[6] F. Hodgins and J. Macey. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. *CMU-RI-TR-08-22*, Jan 2009.

[7] G. Johansson. Visual perception of biological motion and a model for its analysis. *Percept Psychophys*, 14(2):201–211, Jan 1973.

[8] C. A. Kurby and J. M. Zacks. Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2):72–79, Jan 2008.

[9] Y. Ma, H. Paterson, and F. Pollick. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behav Res Methods*, 38(1):134–141, Jan 2006.

[10] V. Manera, B. Schouten, C. Becchio, B. G. Bara, and K. Verfaillie. Inferring intentions from biological motion: a stimulus set of point-light communicative interactions. *Behav Res Methods*, 42(1):168–78, Feb 2010.

[11] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition*, Jan 2009.

[12] K. Pastra, C. Wallraven, M. Schultze, and A. Vatakis. The poeticon corpus: Capturing language use and sensorimotor experience in everyday interaction. *Proceedings of the 7th Conf. on International Language Resources and Evaluation (LREC'10)*, 2010.

[13] R. Poppe. A survey on vision-based human action recognition. *Image Vision Computing*, 28(6):976–990, Jan 2010.

[14] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese. Critical features for the perception of emotion from gait. *J Vision*, 9(6):15.1–32, Jan 2009.

[15] M. Saylor, D. Baldwin, J. Baird, and J. LaBounty. Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development*, 8(1):113–128, Jan 2007.

[16] P. Shrout and J. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 86(2):420–428, Jan 1979.

[17] M. Tenorth, J. Bandouch, and M. Beetz. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS). In conjunction with ICCV2009*, 2009.

[18] N. F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *J Vision*, 2(5):371–387, Jan 2002.

[19] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, Jan 2006.