Sarvadevabhatla, R. K., Ng-Thow-Hing, V., Okita, S.Y. (2010). Extended DurationHuman-Robot Interaction: Tools and Analysis. Proceedings of the 19th IEEE InternationalSymposium on Robot and Human Interactive Communication (RO-MAN). September 12-15, Viareggio, Italy.

Extended duration human-robot interaction: tools and analysis

Ravi Kiran Sarvadevabhatla
Honda Research Institute USA Inc.
Mountain View,CA USA
rsarvadevabhatla@honda-ri.com

Victor Ng-Thow-Hing
Honda Research Institute USA Inc.
Mountain View,CA USA
vng@honda-ri.com

Sandra Okita
Columbia University
New York,NY USA
okita@tc.columbia.edu

Abstract— **Extended human-robot interactions possess unique aspects which are not exhibited in short-term interactions spanning a few minutes or extremely long-term spanning days. In order to comprehensively monitor such interactions, we need special recording mechanisms which ensure the interaction is captured at multiple spatio-temporal scales, viewpoints and modalities(audio, video, physio). To minimize cognitive burden, we need tools which can automate the process of annotating and analyzing the resulting data. In addition, we also require these tools to be able to provide a unified, multi-scale view of the data and help discover patterns in the interaction process. In this paper, we describe recording and analysis tools which are helping us analyze extended human-robot interactions with children as subjects. We also provide some experimental results which highlight the utility of such tools.**

I. INTRODUCTION

Human-robot interaction can be studied from either the perspective of the human participant or that of the robot. In the former case, the robot's behavior can elicit a response from the person, either externally observable or internally felt. Similarly from the robot's perspective, the discernible actions of the human partner can trigger behaviors in the robot in response to those stimuli. Internal state changes can also occur. The great challenge of human-robot interaction is that together, the robot and human (or humans) form a co-dependent relationship mutually influencing their responses in a continuous cause and effect pattern. One cannot consider each party in isolation when developing models for interaction. When working with humanoid robots, studying the interaction becomes more sophisticated as the appearance of the robot can raise expectations about the richness of communication and social protocols that need to be observed.

We have observed in our previous work in humanoid robots [2] that a person's impression of the robot can evolve or change during the course of the interaction session itself. As subjects(children in this case) attempted to communicate with the robot and observed various behaviors, their attitudes changed and consequently, behavioral responses. The cumulative behavior of a robot over an extended amount of time can begin to influence a person's attitude toward the robot. This is a phenomenon that cannot be easily observed with short exchanges such as when a person makes a quick query to a robot. On the other end of the spectrum, very long-term human-robot interaction over the course of days or weeks can be influenced by many other factors not directly attributable to the robot. For example, there may be events in person's daily life that could affect their mood and affect any consistency being sought in the study.

For this reason, we chose to focus our studies on *extended interaction* sequences where a person may interact with a robot in a continuous, uninterrupted task ranging from several minutes to about an hour in length. The length of an extended sequence is long enough to observe multiple turn-taking exchanges and patterns of behavior in both the robot and human. At the same time, the scope of interaction is typically restricted to a particular task domain. In such a setting, some delayed responses may occur for events that happened prior to several interaction exchanges. For example, fear can arise as a response which may not be exhibited immediately but finds an outlet after it builds up beyond a certain threshold.

A. Requirements

The time scale of events of interest can vary significantly. Changes of head pose or eye gaze can occur with a second as is common in many *micro-behaviors* [7]. On the other hand, the peripersonal space between the robot and person may vary slowly over a period of many minutes. Therefore, it is important that the monitoring and analysis of the interaction is conducted using tools which can handle multiple spatio-temporal scales so that nothing is missed.

The monitoring and recording of interaction should employ multiple, synchronized sensor modalities. The data obtained thereby provides multiple sources of evidence for analysis. Also, the interaction should be captured from multiple viewpoints to observe nuances that might be missed from a single, fixed viewpoint. For example, separate cameras are needed to record a person's facial expressions and an overhead view of the interaction setting. These multiple video streams need to be time-synchronized to produce a consistent, integrated visual perspective of the interaction.

By their very nature, recordings of extended interaction produce a very large amount of data. Usually, analysis involves manual annotation(coding) for events of interest in the recorded data(e.g. audio, video) of the interaction session. Typical annotations are done frame-by-frame for video and short segments of time for audio. Often several passes over the same data have to be made by an individual or with multiple coders to counteract human subjectivity. Given the data rates at which recording is done by today's state-of-art tools, the annotation process can be extremely labor-intensive and error prone when sessions last close to an hour. The problem is exacerbated when multiple video streams from different viewpoints are being recorded. As some phenomena can occur over several time frames, if the observer is not explicitly looking at the right time scale or the right view, crucial interaction cues can be missed.

Therefore, tools that can help eliminate the arduous task of coding micro-behaviors should be utilized. This includes applying state-of-the-art computer vision algorithms to automate the detection and documentation of micro-behavior occurrences as much as possible. In order to gain confidence that no false positives or false negatives occur, comparisons of the performance of computer algorithms with human judges should be made.

One contribution of the paper is a description of the tools developed to meet the aforementioned requirements so that they facilitate analysis of extended interactions. In addition, we describe some of our experiences using these tools and applying the mentioned analysis methods. To begin with, a brief overview of the measurement and analysis tools is presented below.

B. Measurement Tools

In Section III-A.1, we describe a scaleable system for recording synchronized multiple viewpoints during an interactive session. In addition to automatic behaviors, studies often have the requirement to model carefully scripted scenarios or offer the investigator manual controls to create repeatable conditions during interaction. Section III-A.2 describes our Wizard-of-Oz tool that allows a combination of manual and automated behaviors with auto-logging of robot behavior events. To obtain a direct measure of physiological arousal, we use skin conductance sensors. The associated data stream can be synchronized with the other audio and video data streams during post-processing.

C. Analysis Tools

Once all the data is recorded, the enormous amount of data needs to be examined and explored for possible patterns. The SAMA system described in Section III-B.1, illustrates how the multi-view camera data can be processed to obtain head pose and gaze-related annotations. Our other tool, MOVE-IT allows various data viewers to have their layouts customized and information exchanged to produce tools to allow linked exploration of data across a common timeline (Section III-B.2).

For the remainder of the paper, Section II discusses related work. Section III describes in further detail our suite of tools we use both for measurement and analysis of session data. Section III-B.2 describes preliminary experiments to assess the efficacy of analysis tools and associated results. A discussion on our experiences using these tools follows in Section V. We end by mentioning some recommendations for extending both the tools and analysis methods in Section VI.

II. RELATED WORK

A survey of human studies for HRI was done in [9] while studying the use of large sample sizes and multiple evaluation methods. The work also provides recommendations for planning, designing and conducting studies in HRI.

Most studies of human-robot interaction employ either a single camera or at most two cameras for studying the interaction. A study of detecting user engagement with a robot companion is described in [10] wherein they use 3 video cameras while [9] mentions a system similar to the one presented in this paper albeit with lesser number(4) of cameras.

Physiological sensors which can measure heart-beat, skin conductance etc. [16] [17] have been quite popular since they can provide a direct measure of subject's arousal, see [12] for an example. In [13], an unconventional, comfort-level indicator device is described which can be used by subject to indicate degree of discomfort with current state of interaction. The authors argue that deriving a high-level concept such as comfort from rich physiological data is not straightforward. They further mention as an alternative that subjects are very familiar with assessing their own subjective comfort level and may be able to communicate the same better using their indicating device. However, they concede no advantage gained from using this device.

Annotation of interaction data is usually manual with a large variation in the time durations. A comprehensive survey of multi-modal annotation tools is done in [14], which includes the free tool we have used(Anvil). The need for an automatic recognition and analysis system has been acknowledged by many researchers [7] [8]. In particular, [8] describes an extremely large, distributed system for collecting data on hospital activities and automatic processing of the 25 Terabytes of resulting data. However, the system needed day-to-day manual coding by 4 people for priming the automated analysis. In [15], a multi-modal approach to analyze human-robot interaction is presented while describing a tool named Interaction Debugger for data presentation, annotation and analysis. By combining the monitoring and analysis tool, they adopt a unified approach to data being recorded and hint at resulting advantages for real-time modification of robot's behavior. However, there could be issues of cognitive load arising from GUI window placement and sheer amount of data being presented via the visualization tool. The benefits of matching interface displays and controls to human mental models include reductions in mental transformations of information, faster learning and reduced cognitive load [11] – a factor which inspired the design of our Wizard-of-Oz and MOVE-IT interfaces. In the interest of focus and space, we shall not present the numerous references to various Wizard-of-Oz systems and robot control interfaces.

III. TOOLS AND METHODOLOGY

A. Measurement tools

1) Distributed camera system: In order to capture the complete range of behavior, 7 cameras were arranged in the

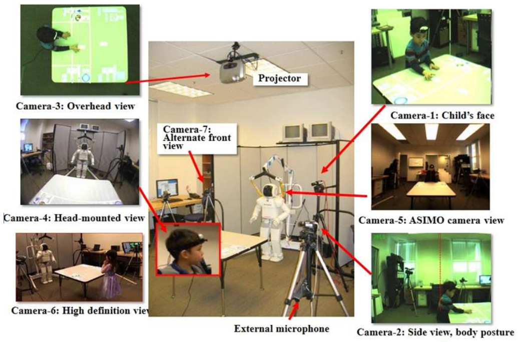

Camera-3: Overhead view
Projector
Camera-7: Alternate front view
Camera-1: Child's face
Camera-4: Head-mounted view
Camera-5: ASIMO camera view
Camera-6: High definition vie
External microphone
Camera-2: Side view, body posture

Fig. 1. Distributed recording system

observation room to capture several micro-behaviors (refer to Figure 1 for camera placements). The room itself had windows on one wall that allowed parents to observe the experiments from outside the room.

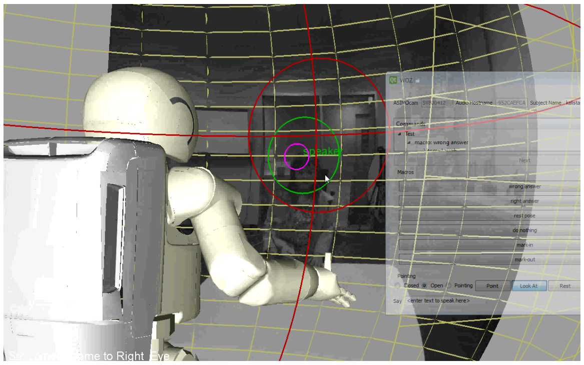
Camera-1 faces the child to capture facial expressions. Camera-2 provides a side view to determine the degree of body lean posture relative to a vertical reference line. The degree of lean can indicate level of interest in the activity. Camera-3 provides an overhead view of the table and child which is helpful for observing the choices the child makes during the task as well as any hesitation therein. Camera-4 features a head-mounted camera that allows us to estimate the gaze direction of the child as well as what the object of attention is in her vision. Camera-5 is taken from humanoids own cameras. This is valuable for recording what is directly observable by humanoids own sensors and consequently by any vision detection algorithms created. Camera-6 is a Sony high-definition DVCAM camera providing a wide field of view of both humanoid and the child face to face to observe whole body movements. Finally, Camera-7 is a Sony Handycam providing another view of the face from a different angle. Since facial expressions can give us a strong indication of the emotional state of the child, two viewpoints were established for the face since children often change their face orientation frequently. Cameras 1-3 are 640×480 JAI/Pulnix Gigabit Ethernet machine vision cameras, model TMC-6740GE. All three cameras are connected via Gigabit Ethernet cable directly to a single server which was able to directly digitize all video onto its hard drive at a rate of 15 fps. For Camera 4, we used a small miniature camera, measuring $2.5cm \times 2.5cm$ by Korea Technology and Communications Co., Ltd., model KPC-VSN500NH, providing 768×494 resolution, equipped with Swann fish-eye 150 degree lens to approximate the wide field of view in human vision. To calibrate the head mounted camera, we instructed the children to look at specific targets and adjusted the camera so that the target was in the center of the image. One potential problem is independent eye-gaze shifts from head direction, however [1] show that for table-top tasks, head motions correlated well with coded eye positions.

Audio was captured separately using a wide-array receiver microphone as well as lapel microphones attached to the child, of which the latter provide clear distinct utterances. The wide-array receiver microphone was synchronized with Cameras 1-5 and all data samples were timestamped to the same clock, eliminating the need for manual synchronization and digitizing. This procedure saved us countless hours of manual processing as we collected over 5 TB(Terabyte) of audio and video data for analysis.

2) Wizard-Of-Oz: The Wizard-of-Oz (WoZ) technique refers to the process of controlling a robot using surreptitious means of concealing the human operator so that the person interacting with the robot is unaware that it is under human control and believes it is acting autonomously. The method is a useful prototyping tool for evaluating perception and behavior algorithms prior to investing the effort to implement them. In the context of long-term interaction, there is a higher chance that the robot will encounter situations it cannot handle, and WoZ control interface can aid in helping the robot get over potential technical problems with its autonomous algorithms.

We have developed a WoZ control interface in our software framework called MOVE-IT (Monitoring, Operating, Visualizing, Editing Integration Tool) [6]. The framework allows various interactive elements to be combined to create a customized interface that is suitable for the particular task the robot will be used for in a study. For example, in Figure 2, our WoZ interface is used to interact with a little girl. The GUI portion of our interface features a script-based interface where an interactive script can be authored containing sequences of robot commands such as playable motion sequences and dialog. However, it can also contain conditional commands that allow a variety of responses to be specified for any interaction event in the script. There is also an array of buttons that can be customized to produce responses on-demand to react instantly to events that are not part of the script. Finally, a text prompt allows arbitrary dialog to be generated by the human operator. A keyboard manufactured out of silicone is used to prevent subjects from hearing the tell-tale typing noises that might betray the illusion of the WoZ. Although this functionality is useful, it also takes up a considerable amount of screen real estate. To alleviate this problem, the interface has adjustable transparency so that the operator is still free to see the area underneath which is devoted to visualization.

The visualization part of our WoZ interface allows the operator to observe multi-modal phenomenon, ranging from the current joint configuration of our robot to the location of sound sources in the room. The video from the robot's cameras are streamed and displayed on a panoramic surface in front of the robot model so the operator can see the robot's viewpoint, allowing remote (and therefore hidden) operation. This interface can constrain the operator to the robot's sensor limitations, which is useful as any computer algorithm would be subjected to the same constraints. The video display is also interactive in that the operator can click on any point of the display and have the robot look



ASIMOcam SV800412
Audio Hostname -952CAEFCA
Subject Name kalista

Fig. 2. Wizard-of-Oz control interface

or point in that location. We have found this essential for producing attentive behaviors. Augmented information can be displayed on the display that may be computed by the robot's vision algorithms. For example, in our humanoid model, our attention system identifies and labels the most likely speaker by combining sound localization and face detection algorithms.

For the human operator, WoZ control can be a very physically and mentally exhaustive process, as a single operator needs to be responsible for a host of verbal and non-verbal behaviors. To alleviate this, multiple configurations of our WoZ software can be run simultaneously so that control tasks can be split between more than one operator. For example, in our studies, we have one operator control dialog and the scripted interaction while another focuses on nonverbal pointing and looking. The combined effort of both operator provides a more lively robot than possible with a single operator. Although we use instant messaging software to communicate silently between operators, practice sessions are useful for developing better coordinated behavior. Finally, all robot commands generated through the WoZ are time stamped and logged so that no manual annotation of humanoid's behavior is required, allowing the researcher to only focus on coding the human behaviors.

B. Analysis tools

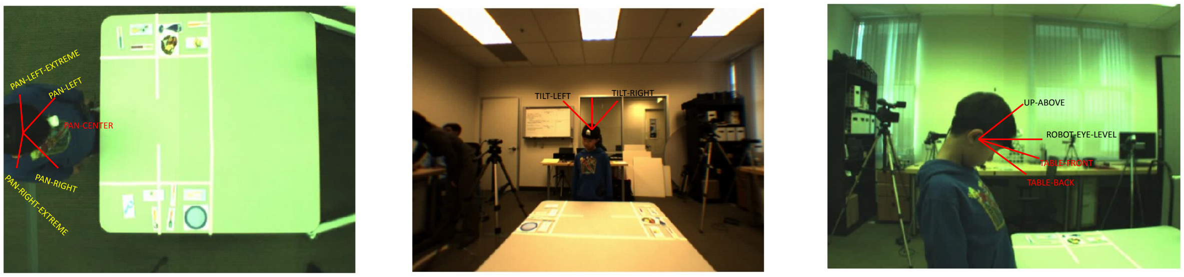
1) SAMA+Anvil: The basic design goal of using SAMA(Subject Automated Monitoring and Analysis) along with Anvil[1] is to analyze the sensor data (in particular, camera information) to provide clues to trends in the human-humanoid interaction. For instance, we may wish to know instances when the subject turned away from the humanoid or instances when humanoid and subject were speaking simultaneously (speech barge-in). SAMA analyzes the multi-view video data collected during the recording phase and outputs a semantic annotation tag set for each time-slice. For each time instance, it simultaneously processes the corresponding frame from each of the 5 cameras. For each frame, various pose-related face properties such as head-roll, tilt, pan are estimated. The relative position of the camera viewpoints (front, profile, side etc.) is also known. The face properties from all the viewpoints along with viewpoint information are combined and processed using a rule base, which determines the final semantic tag set for that time instant.

We now describe some details of how the tag set is produced for each set of input frames from the different cameras(see Figure 3). To determine the general direction(referred to as View-Zone) in which a subject's face is oriented, the face detection confidences from all ground camera(humanoid,PULNIX) are considered. If more than one camera view provides a good face detection confidence beyond a certain threshold, the viewing direction is considered to be between the cameras corresponding to the top-most two face detection confidences. If there is only camera for which confidence exceeds the aforementioned threshold, then the subject is considered to be directly looking towards that camera. To determine whether the subject is looking down, at the table or upwards, the head-roll value is thresholded with four thresholds – TABLE-BACK,TABLE-FRONT,ROBOT-EYE-LEVEL,UP-ABOVE(Figure 3(c)). To determine whether the subject's head is tilted, two thresholds – TILT-LEFT, TILT-RIGHT are used on head-tilt value(Figure 3(b)). To determine which way the subject's head is turned with respect to a vertical axis, the head-pan value is thresholded using five thresholds – PAN-RIGHT-EXTREME,PAN-RIGHT,PAN-CENTER,PAN-LEFT,PAN-LEFT-EXTREME(Figure 3(a)). Particular combinations of values within a tag set can be associated with intuitive human-robot interaction configurations. For example, View-zone=HUMANOID-CAM,head-roll=ROBOT-EYE-LEVEL,head-pan=PAN-CENTER may indicate that the child is looking directly at the humanoid. Yet another setting can indicate looking down at the table or looking up at the ceiling etc. Such configurations can be used to initiate configuration-specific behavioral responses in the humanoid's interaction model. In this way, the entire set of videos associated with an interaction episode can be coded with information on the current gaze location of the subject. The analysis from SAMA provides useful cues for where to focus on by indicating low-level gaze cues and their transitions. For this purpose, we use Anvil [4](see Figure 4). By combining the analysis from SAMA on video data with data from other sensors (audio, physio), we get an opportunity to examine hitherto unobserved long-range relationships between interaction elements. By combining information from multiple view-points, SAMA can provide an accuracy in tagging beyond what would be possible from a single view-point.

Refer also to Section III-B.2 for a quantitative assessment of the SAMA tool.

2) MOVE-IT: MOVE-IT (Monitoring, Operating, Visualizing, Editing Integration Tool) [6] is our software framework for combining interactive visual elements together to create cohesive applications. In Section III-A.2, we describe how

[1] Anvil is a free video annotation tool which offers provides multi-layered annotation based on a user-defined coding scheme. Refer to [4] for more details



(a) Threshold settings for head-pan

(b) Threshold settings for head-tilt

(c) Threshold settings for head-up-down

Fig. 3. Threshold settings used to obtain semantic tag sets in SAMA

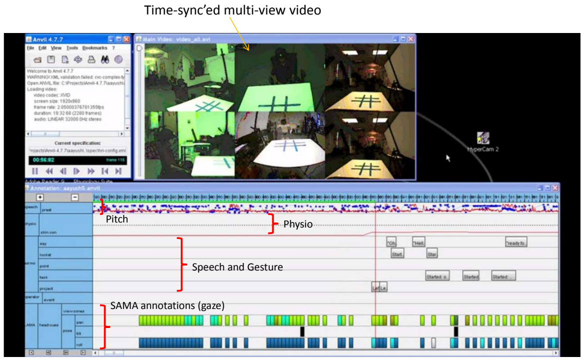


Fig. 4. Screenshot of ANVIL showing the time-synchronized multiple viewpoint video at the top and the speech, physio annotations and the automatically generated SAMA annotations.

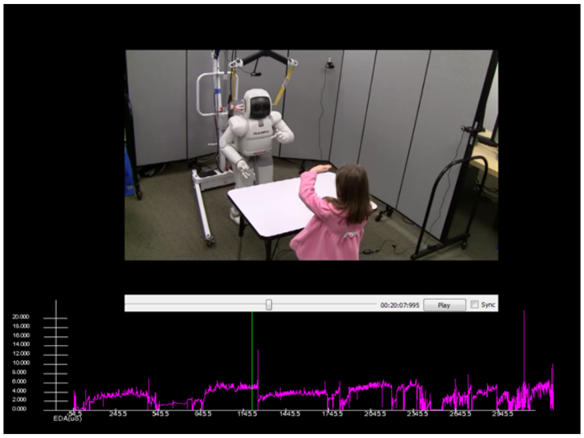


Fig. 5. Synchronized video and skin conductance streams.

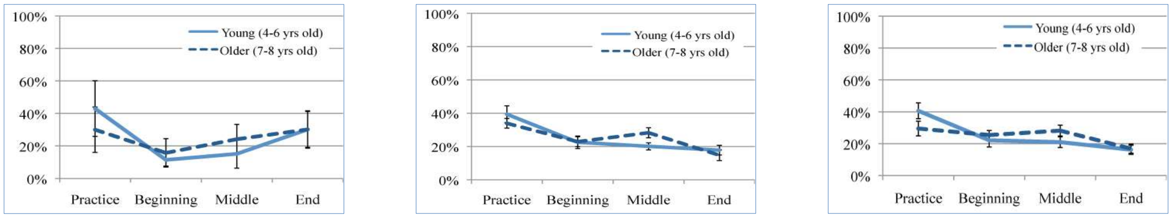
MOVE-IT was used to create a WoZ interface for the robot. Here, we used MOVE-IT to create a multi-modal analysis tool to obtain synchronized access to pre-recorded video and physiological skin conductance (which measures arousal) data streams. MOVE-IT provides the common workspace or canvas to place interactive elements like an audio-visual media player and a data plotter. The interface is visually simple and uncluttered, as only the interactive elements one needs for analysis are visible, without the clutter of unnecessary GUI elements. Each element has built-in functionality and behavior, but can also communicate events to each other by connecting the signals of one element to a function of the other. This allows the elements to have synchronized or dependent behavior on each other.

We needed to synchronize the video playback with the physiological data streams. By connecting the time bar of the media player to the time marker of the data plotter, we can highlight where in the plot the video corresponds to. More importantly, a command can be given to synchronize a point in a video to a point in the plot sequence. Alternatively, clicking on the data plot would cause the media player to jump to the corresponding frame. The plotter can also zoom in on a more narrow time range facilitating multi-scale time analysis from very short-term micro-behaviors to observing long term phenomenon.

Anvil itself is not well-suited for large, long-term video due to memory management issues, since it was originally intended to annotate short-term interactions. The media player in MOVE-IT, in contrast can handle very large high-quality video streams that can be several Gigabytes in size. It may be more informative to visualize multi-modal information such as simulataneous encoding of location and volume in a 2-D image display rather than as multiple time series of data channels. For high-dimensional data with a high degree of correlation such as joint angles of a robot, visualizing the robot directly as in MOVE-IT is preferable than viewing the time series of all joints individually.

IV. APPLICATIONS OF SAMA

We describe some working examples in which SAMA can be applied for data analysis. Video data of 10 test subjects (children between ages 4-to-8) was used to evaluate SAMA's capabilities. For the purposes of analysis, each video is divided into four sequential portions – Practice, Beginning, Middle and End – corresponding to the phases in the interaction session. SAMA's multi-view camera information and the multi-scale annotation can be applied as an assessment tool at three different levels.



(a) Camera position view with child gazing towards humanoid

(b) Head movement Roll where child's eye is at level of humanoid

(c) Head movement Pan, where child is centered facing toward humanoid

Fig. 6. The graphs show the percentage of time spent looking at humanoid at different sections in the interaction. Practice refers to the training practice session in the beginning. The actual session is divided evenly into beginning, middle and end.

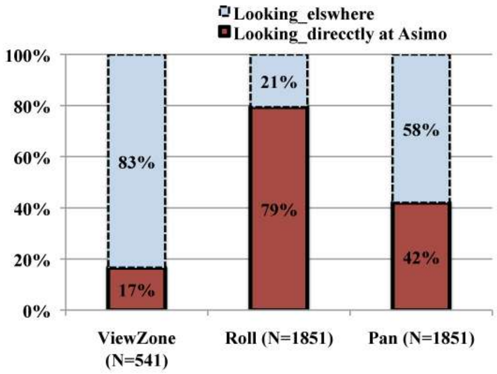


Fig. 7. Comparing different properties(x-axis) as measured by SAMA for incidents where child looked at humanoid. N against each property denotes number of incidents recorded for that property.

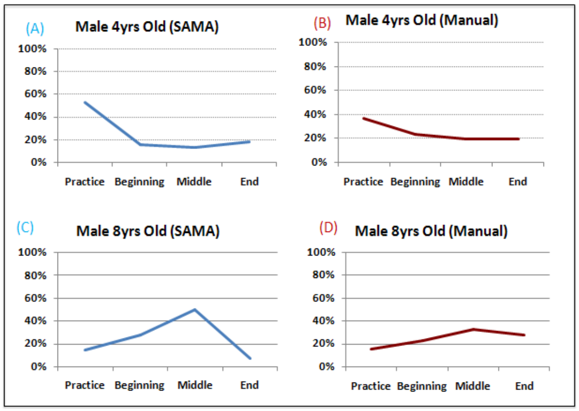


Fig. 8. SAMA compared to coding incidents by hand (manually).

At a general level, SAMA can analyze different pose-related face properties (head-roll, tilt and pan) and camera view positions. For example, in Figure 6, we consider one camera position (a) facing center toward humanoid, and two different head movements: (b) head movement Roll for incidents where the child's eye level is at humanoid's and (c) head movement Pan for incidents where the child is faced center toward humanoid. As can be seen from the graphs in Figure 6, even though the graphs differ slightly between the pose-related face properties and the camera view position, all three sources show similar attention patterns in children as the session progresses.

At an informative level, SAMA can analyze different camera view-points (frontal, profile, side) and the pose-related face properties (head-roll and pan) to help determine the most informative data source for analysis. One example (See Figure 7) is when we wish to determine which of the properties was useful to analyze for the situation of direct eye-contact with humanoid. In this case, SAMA provides the relative ratio between looking directly at the humanoid versus looking elsewhere for various properties (ViewZone, Roll, and Pan in Figure 7). The property with the best ratio value (Roll in this example) can be then used after performing such analysis.

At the application level, SAMA can assist in manual annotation by potentially speeding up the often time consuming and labor intensive task. In Figure 8, a comparison was made between SAMA and manual annotation to see (a) whether SAMA can generate a similar data pattern as manual annotation and (b) examine the differences in the number of incidents recorded. Two video clips where children interacted with the humanoid were examined (a 4-year old and 8-year old). To ensure reliability of the coding for the manual annotation, three human coders separately scored the video clips, with 95% agreement. As seen in the graphs from Figure 8, SAMA was able to generate similar patterns overall to the manual annotation.

V. DISCUSSION

In our experience studying extended human-robot interaction, there are three important requirements that must be addressed: simultaneous observation of human and robot, multi-modal data recording and analysis, and the ability to study phenomena at different time scales. This was evident in one of our previous studies involving humanoid teaching children how to set a table [2], where immediate physical details such as humanoid's voice and motions had a noticeable effect on learning. Longer term phenomena, such as the structure of the lesson (authoritative versus interactive), also affected learning. The tools we are developing were designed to meet these needs for studying extended human-robot interaction.

A. Simultaneous observation

In our operation of the WoZ interface, it was important not only to see the human participant through humanoid's camera "eyes", but to also see humanoid's current joint configuration. This provides feedback to the operator that the robot is obeying its commands, but can identify potentially dangerous situations if a child gets too close to the robot while it is moving. In one case, we were puzzled why children kept on trying to give picture cards to humanoid. However, once we observed what humanoid was doing via WoZ, it became apparent that some of humanoid's pointing motions were being interpreted by the children as the robot reaching out to grab something. This helps us design humanoid's behavior to be less ambiguous. By seeing a computer-graphics model of humanoid tied to its actual physical configuration, the operator gets an idea of what the child is seeing. In our first trial studies, our WoZ operator would dutifully click on the video display to look at different areas of the screen. Because she saw the camera display move around in response to her commands, she thought the robot would appear attentive. However, when we showed video of the entire interaction it became apparent that humanoid's head motions were very slight and unnoticeable. The solution was to click on views at wider distances apart to create more head motion as well as using pointing while looking which creates noticeable arm movement. In our multiple-WoZ scenarios, a live visualization of humanoid allowed the dialog operator to watch the behavior of the looking/pointing operator, providing better communication and allowing one operator to feed and react off the performance of the other.

The multiple camera views were also important in this respect. Having cameras closely focused on the face, allowed enough high resolution detail to be available to observe facial expressions, while other cameras could capture the full scene between the human participant and the robot. The head-mounted camera was useful for identifying what children were looking at, whether it being humanoid or other distractions in the environment (like the parents or researchers). This prompted us to re-design subsequent experiments where the parents were not visible and researchers were hidden from view. The results were extended interaction sequences where the children were less inclined to rely on other humans for help, and interacted more directly with the robot.

B. Multi-modal Recording and Analysis

Our current measuring system records video, audio and physiological data. Being able to synchronize the information and visualize how their simultaneous signals in an intuitive way was achieved with the MOVE-IT and Anvil tools. For developing more robust perception algorithms, these modalities can be obtained to create stronger confidence of state estimates of the environment. For example, we combined the sound sources with faces to identify speakers. On the analysis side, studying multi-modal cues helps us identify potential triggers and responses, each of which can occur in a different modality. For example, a robot speaking (audio) can trigger a child to look at the robot (visual).

In our current camera system, we did not note their relative locations to each other. If we had done that, we could localize the cameras in the environment and potentially retrieve more 3-dimensional information of the scene being viewed. Alternatively, depth or stereo cameras can be used, but current designs can produce noisy or low-resolution data. However, it remains unclear what kind of useful information 3-D knowledge can provide. Obtaining distance measures between the person and robot can easily be obtained from a top-view camera.

Because of its automatic and systematic nature, SAMA records about 2-3 times more incidents than manual annotation. However, since SAMA is found to generate a similar pattern as the manual version(Figure 8), it can speed up the data analysis. For example, a researcher can look at the SAMA generated patterns to eyeball potential segments in the video clip (e.g., more incidents found in the Middle section than the Beginning). As a work in progress, the next challenge will be to use SAMA with a larger data set.

C. Different Time Scales

Our original camera system captured all video streams onto a central camera server. However, the frame rates were not high enough to capture extremely short phenomena such as quick gestures or microexpressions, which are brief involuntary facial expressions [5]. By de-centralizing the capturing to the local machines the camera were attached to, we not only achieved faster frame rates, but produced a scaleable system that allowed us to add additional cameras easily. We had to make sure to synchronize all computers to a common time server so that the videos can be later synchronized when re-assembled as a mosaic.

Being able to see the entire timeline of interaction and zoom in on specific segments were useful for quickly identifying interesting events. In the case of the physiological data viewed in Figure 5, we could identify specific triggers to unusual physiological arousal activity such as humanoid suddenly talking after a long period of silence. At the longer timescales, we could notice a pattern of alternating high and low activity which coincided with the times when the child was engaging with the robot and stopping to listen to instructions from a computer.

For data with large sampling rates(physio) or dimensionality (video, robot joint angles), it may be simply impractical to view the entire interaction at one glance. One solution is more automated analysis of the data, which is what we resorted to with the SAMA tool for video. However, other data mining techniques should be applied not only to the high dimensional data within one modality, but across multiple, simultaneous modalities. We are exploring ways of combining rich visualization with automatic methods for highlighting useful incidents across time.

VI. CONCLUSION AND FUTURE WORK

We have presented a suite of tools and methods that we have developed for the purposes of studying extended human-robot interaction. On the measurement side, multiple camera systems, physiological measures, and a customizable WoZ interface provides researchers with a granular view of the interaction data and viewpoints to capture aspects of interaction at multiple time scales and sensor modalities.

Although we have not used SAMA in an online real-time fashion for the study, it is easy to do since the processing is on a per-frame basis. Such a mechanism would provide real-time gaze-related information to humanoid or a Wizard-Of-Oz operator. This generic method, in turn can be used for situations such as interaction repair or generating timely responses, thanks to the ease of integration that the existing communication framework [3] offers.

For analyzing the large amounts of data, automatic logging of robot events and automated analysis of camera data help minimize the amount of manual effort for coding and annotating the data produced. Moving forward, we will continue developing more intelligent tools for multi-modal analysis at different time scales. The main goal will be not to replace the analyst, but to assist the analyst in finding interesting details rather than deal with cognitive burden issues that arise with large volumes of data. Another direction would be to help the analyst handle the complexity of the environment, including keeping track of people and their social roles during group interaction.

For robot designers, being able to pinpoint what works and what does not is very useful for improving the overall behavior of the robot to producing engaging human-robot interaction. We are now able to capture an abundance of data that records the sessions. Using this knowledge to extract out important lessons and building intelligent behavior models for interaction will complete and validate this effort.

REFERENCES

[1] H. Yoshida and L. B. Smith, Whats in View for Toddlers? Using a head camera to study visual experience. *Infancy*, Volume 13, Issue 3. pp. 229-248, 2008

[2] S. Okita and V. Ng-Thow-Hing and R. K. Sarvadevabhatla, Learning Together: ASIMO Developing an Interactive Learning Partnership with Children, *18th IEEE International Symposium on Robot and Human Interactive Communication(RO-MAN 09)*, 2009, Toyama, Japan.

[3] V. Ng-Thow-Hing and K. Thórisson and R. K. Sarvadevabhatla and J. Wormer and T. List, Cognitive Map Architecture: Facilitation of human-robot interaction in humanoid robots, *IEEE Robotics and Automation Magazine*, vol. 16, no. 1, 2009, pp. 55-66

[4] M. Kipp, Multimedia Annotation, Querying and Analysis in ANVIL. *Multimedia Information Extraction*, Chapter 19, MIT Press, 2009

[5] E. A. Haggard and K. S. Isaacs, Micro-momentary facial expressions as indicators of ego mechanisms in psychotherapy, *Methods of Research in Psychotherapy*, Edited by L. Gottschalk and H. Auerbach, pp. 154-165, New York: Appleton-Century Crofts, 1966.

[6] V. Ng-Thow-Hing, MOVE-IT: a Monitoring, Operating, Visualizing, and Editing Integration Tool, *In submission*, 2010.

[7] K. Dautenhahn and I. Werry, A Quantitative Technique for Analysing Robot-Human Interactions, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002.

[8] S. Stevens and D. Chen and H. Wactlar and A. Hauptmann and M. Christel and A.J. Bharucha, Automatic Collection, Analysis, Access and Archiving of Psycho/Social Behavior by Individuals and Groups, *Capture, Archival and Retrieval of Personal Experiences (CARPE'06)*, Santa Barbara, CA, USA, pp. 27-34, 2006.

[9] C. L. Bethel and R. R. Murphy, Use of Large Sample Sizes and Multiple Evaluation Methods in Human-Robot Interaction Experimentation, *AAAI Spring Symposia*, 2009

[10] G.Castellano and A. Pereira and L. Iolanda and A. Paiva and P.W. McOwan, Detecting user engagement with a robot companion using task and social interaction-based features, *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*, Cambridge MA, USA, pp. 119-126,2009.

[11] J. Macedo and D. Kaber and M. Endsley and P. Powanusorn and S. Myung, The effects of automated compensation for incongruent axes on teleoperator performance, *Human Factors*, vol 40, pp. 541-553, 1999.

[12] D. Kulic and E. Croft, Anxiety Detection for Human Robot Interaction, *IEEE International Conference on Intelligent Robots and Systems*, pp. 389-394, 2005.

[13] K. Koay and M. L. Walters and K. Dautenhahn, Methodological Issues Using a Comfort Level Device in Human-Robot Interactions , *IEEE International Workshop on Robot and Human Interactive Communication(ROMAN)*, pp. 359-364, 2005

[14] K. Rohlfing and D. Loehr and S. Duncan and A.Brown and A. Franklin and I. Kimbara and J. Milde and F.Parrill and T.Rose and T.Schmidt and H. Sloetjes and T. Alexandra and S. Wellinghof, Comparison of multimodal annotation tools, *Gesprchforschung - Online-Zeitschrift zur Verbalen Interaktion*, vol. 7, pp. 99-123, 2006.

[15] T. Kooijmans and T. Kanda and C. Bartneck and H. Ishiguro and N. Hagita, Interaction debugging: an integral approach to analyze human-robot interaction. *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction(HRI)*, pp. 64-71, 2006

[16] `http://www.thoughttechnology.com/`

[17] `http://www.affectiva.com/`