A Multi-party Multi-modal Dataset for Focus of Visual Attention in Human-human and Human-robot Interaction

Kalin Stefanov and Jonas Beskow
KTH Royal Institute of Technology
Stockholm, Sweden
kalins@kth.se and beskow@kth.se

Abstract

This papers describes a data collection setup and a newly recorded dataset. The main purpose of this dataset is to explore patterns in the focus of visual attention of humans under three different conditions - two humans involved in task-based interaction with a robot; same two humans involved in task-based interaction where the robot is replaced by a third human, and a free three-party human interaction. The dataset contains two parts - 6 sessions with duration of approximately 3 hours and 9 sessions with duration of approximately 4.5 hours. Both parts of the dataset are rich in modalities and recorded data streams - they include the streams of three Kinect v2 devices (color, depth, infrared, body and face data), three high quality audio streams, three high resolution GoPro video streams, touch data for the task-based interactions and the system state of the robot. In addition, the second part of the dataset introduces the data streams from three Tobii Pro Glasses 2 eye trackers. The language of all interactions is English and all data streams are spatially and temporally aligned.

Keywords: dataset, focus of visual attention, human-robot interaction, human-human interaction

1. Introduction

During recent years, we have witnessed the start of a revolution in personal robotics that is rapidly starting to become part of our everyday lives. The promise and potential of these systems is far-reaching; from general purpose co-worker robots that operate and collaborate with humans side-by-side, via household and service robots that are able to carry out or assist in daily chores to robotic tutors in schools that interact with humans in and around a shared environment. All of these scenarios require systems that are able to recognize and convey attention to objects in the environment. Human face-to-face interaction involves sophisticated mechanisms for conveying intentions, establishing common ground and acknowledging information transfer, often based on a combination of spoken and visual (nonverbal) signals. In order to be able to understand and model these processes and leverage their power in human-robot interaction, we need comprehensive multi-modal data collected in relevant settings. In this paper, we present a highly multi-modal interaction dataset that was recorded for this purpose. It involves multi-party human-robot interaction based around objects (virtual cards on a touch table), multi-party human-human interaction based around the same objects, as well as free multi-party human-human interaction without objects. The primary purpose of the dataset is to serve as a source for modeling visual attention patterns for robots interacting with humans, but the richness of the dataset also makes it useful for other studies on the dynamics of multi-party interaction.

So why a new dataset when there is already a wealth of multi-party interaction recordings that have been made? There are three main reasons for this,

- We want data for interaction in multiple settings, in particular with/without object manipulation;
- We want fully automated measurements (i.e. no need for manual labeling) in order to enable large-scale data-driven modeling, and to make it possible to obtain corresponding input in real-life human-robot applications (e.g. using only Kinect v2 sensor);
- We want all measurements to be spatially aligned, i.e. mapped to the same $3D$ space.

Of the datasets we are aware of, none of them satisfies all of these requirements - Johansson et al. (Johansson et al., 2013) describes a corpus of multi-party human-robot interaction (recorded with Kinect v1 device) involving object manipulation similar to our setup, but the dataset does not contain an all-human condition that would serve as a gold standard/training data for the robot's behavior. Oertel et al. (Oertel et al., 2014) and (Oertel et al., 2013) both describe multi-party human interaction corpora, but without object manipulation. Hung et al. (Hung and Chittaranjan, 2010) is a massively multi-party game corpus (8-12 participants) but involves no object manipulation. Nguyen et al. (Nguyen et al., 2014) has gaze data for dyadic interviews, but no object manipulation, while Carletta et al. (Carletta et al., 2006) features multi-party meetings with rich multi-modal annotation and objects, but offers no way to automate the measurements for real-life applications.

2. Dataset

2.1. Motivation

Nonverbal communication plays an important role in everyday face-to-face human interaction. Designing accurate computational models for nonverbal behavior recognition would be beneficial for a robot in order to carry out an interaction in more intelligent manner. The focus of visual attention is one such behavior and we wish to gain deeper understanding of how it is effected and explore patterns it exhibits under different interaction conditions. In particular we wish to gather data from multi-party interactions where a human is performing the same task as the robot would - in this case leading and taking part in simple knowledge-based games, with and without objects. This data will allow data-driven modeling of the robot's visual attention behavior. For comparison, we also want data from a corresponding multi-party human-robot setting (where the robot's behavior is rule-based). In order to address this task, the dataset is designed to contain three conditions,

- **human-human-robot** task-based interaction with a touch surface. The task is a collaborative problem solving scenario where two participants and a robot play a game. The game is an adaptation of the "Timeline" (Skantze et al., 2015) quiz game. The participants need to order a set of cards on the touch surface with the help of the robot. The robot is Furhat (Moubayed et al., 2013), back-projected human-like robot head. Figure 1 presents one of the questions in the game. In the rest of the text we will refer to this scenario as `Condition 1`;
- **human-human-human** task-based interaction with a touch surface. In this scenario the robot from `Condition 1` is replaced by a third human. In the rest of the text we will refer to this scenario as `Condition 2`;
- **human-human-human** task-based interaction without a touch surface. In this scenario three humans discuss a selected question in order to reach the correct answer. In the rest of the text we will refer to this scenario as `Condition 3`.

In total fifteen, 30 minute sessions were recorded, where approximately 10 minutes were spent on each of the three conditions, resulting in approximately 7.5 hours of data (approximately 2.5 hours per condition). Three humans took part in each recording session; two of the participants were new in every session, and one human taking the role of the robot in `Condition 2` and `Condition 3` that was the same in every session.

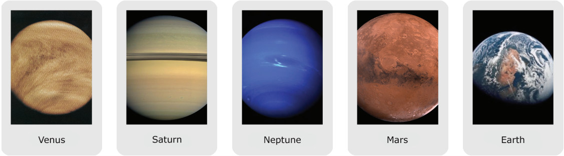

Figure 1: Example of one round of the quiz game. The task is to *"order the planets based on size, from smallest to largest"*.

2.2. Setup

All interactions occur around round table and the participants are seated. Although the chairs were not static, we tried to implicitly limit the space where the participants can place their chairs in order to form as close as possible an equilateral triangular pattern. This in turn results in spatial separation of the possible targets of visual attention. Figure 2 illustrates the spatial configuration of the setup. Table 1 summarizes the sensors used based on the condition.

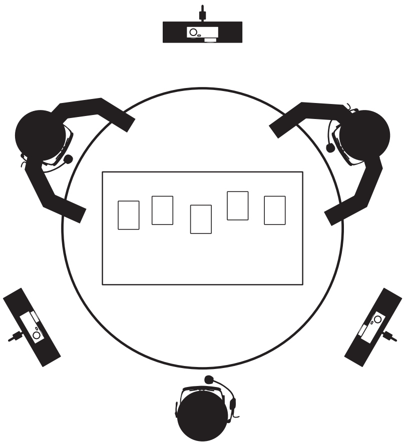

Figure 2: Spatial configuration of the setup and the location of different sensors used.

Condition 1	Condition 2	Condition 3
2 microphone 2 Kinect 2 GoPro 2 Tobii 1 touch surface 1 robot	3 microphone 3 Kinect 3 GoPro 3 Tobii 1 touch surface	3 microphone 3 Kinect 3 GoPro 3 Tobii

Table 1: The sensors used based on the condition.

The following data streams are included in the dataset,

- High quality audio is recorded using Shure dynamic headset microphones;
- High quality eye-gaze data[1] is recorded using Tobii Pro Glasses 2 eye trackers. High definition video from the front camera (first person view) of the eye trackers is also recorded;
- Color, depth, infrared, body (25 joints) and high definition face (pivot point, orientation and animation units) are recorded using Kinect v2 sensors;
- High resolution video is recorded using GoPro cameras;
- Game state (touch events, currently active objects, etc.) is recorded using Elo touch surface;
- Robot state (gaze targets, speech synthesis, speech recognition, etc.) is recorded using Furhat and IrisTK (Skantze and Al Moubayed, 2012).

2.3. Spatial and Temporal Alignment

We use frequency of 30 frames per second for temporal alignment of all data streams. Temporal alignment is done on several levels. First step is to synchronize all streams (color, depth, infrared, body and face) within each of

[1] Data from Tobii Pro Glasses 2 eye trackers is recorded only in the second part of the dataset.

the Kinect sensors. This is achieved by timestamps that are recorded for each frame in each of the streams. The second step is to temporally align all Kinect devices with each other. This is achieved by using visual information of hand claps which occur at the beginning and the end of each condition in each session. The hand claps visual information is further used to cut and align the GoPro video streams and the video streams from the front camera of the eye trackers (resulting in temporal alignment of the *3D* gaze data corresponding to each eye tracker's video streams). The hand clap is captured by all microphones and this information is used to cut and align the audio streams. Finally, visual information of the game state captured by all Kinect sensors is used to align the game state stream to the Kinect sensors' streams and a speech synthesis event time in the robot state is used to align the robot state stream with one of the audio streams. The end result of the procedure is all streams being cut based on the start and end of the three conditions in each session and temporally aligned with frequency of 30 data points per second.

Spatial alignment is done on several levels as well. First each Kinect, GoPro and Tobii Glasses color stream is calibrated with a calibration pattern to obtain the corresponding pinhole camera model intrinsic parameters. The second step is session specific calibration which occurred before the start of the recording of each of the sessions. We placed the same calibration pattern on the touch surface (where we defined the origin of the world coordinate system). This procedure produces the relative pose (extrinsic parameters) of all cameras (Kinect, GoPro and Tobii Glasses) in the world coordinate frame (center of the touch surface). The calculated transformation matrix for each device is then applied to the spatial data generated by that device resulting in all data streams transformed to the same world coordinate frame. Finally, the position of the robot head in `Condition 1` is estimated from the perspective of the Kinect that faces it and transformed to the world coordinate frame with the appropriate transformation matrix. The end result of the procedure is that all spatial data is transformed to one coordinate space which we defined in the middle of the touch surface. Figure 3 illustrates recovery of the spatial data (face pivot point and orientation) of one frame for all three Kinect devices after spatial alignment.

Figure 4 illustrates *2D* heat map and gaze plot of the estimated eye-gaze for the *human in the center* in `Condition 2` to a static image of the scene seen by the participant. The length of the mapped data is approximately 3 minutes and captures the gaze behavior during one round of the quiz game which involves the touch surface.

3. Initial Analysis

The objective of the presented analysis is to test a simple hypothesis - the touch surface in `Condition 2` will attract the focus of visual attention of the participants for considerably longer time than in `Condition 3`.

We defined the possible targets in the scene as `Human Center`, `Human Left`, `Human Right`, `Table`, and `Other`. Then, the target of visual attention of each participant is estimated on frame-by-frame basis by intersecting

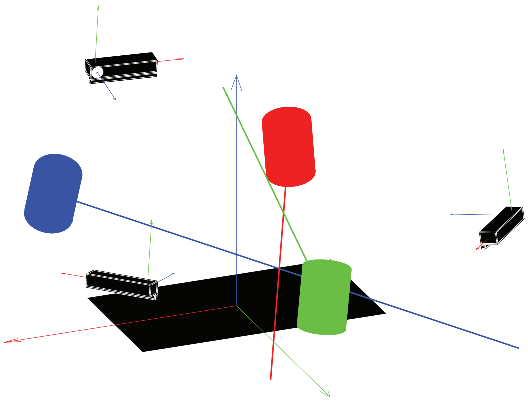

Figure 3: Reconstruction of one frame of the Kinect sensors' spatial data. The touch surface (the plane) is the origin of the world coordinate frame. The location and orientation of the faces recorded by each Kinect are transformed to that frame and drawn as cylinders and orientation vectors. The relative pose of the Kinect sensors is also drawn.

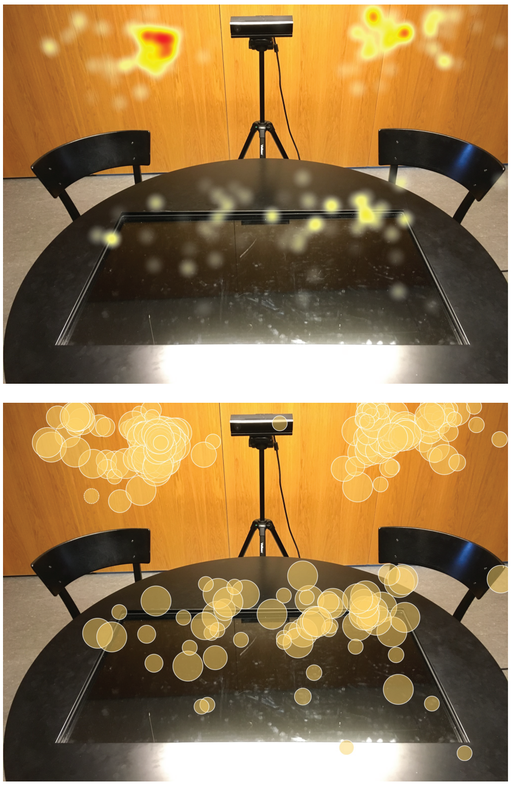

Figure 4: Heat map (top) and gaze plot (bottom) of the eye-gaze data for the *human in the center* in `Condition 2`. During the actual interaction there were two participants sitting in the chairs and playing a game on the touch surface.

the face orientation vector with the defined targets in the scene. In cases when the orientation vector was not intersecting any target we picked the closest one based on angular distance. Finally, if the angular distance to the closest target was bigger than 25° we labeled this instance as `Other`.

The results of this procedure for six sessions (approximately 33% of the total dataset) are illustrated in Figure 5. The results suggest that indeed the table in `Condition 2` attracts significantly more attention then in `Condition 3`. A paired-samples t-test was conducted to compare the number of frames each participant looks at the touch surface in `Condition 2` and `Condition 3`. There was a significant difference in the scores for `Condition 2` ($\mu = 9713.4$, $\sigma = 3018.6$) and `Condition 3` ($\mu = 5127.8$, $\sigma = 5044.6$); $t(11) = 2.5796$, $p = 0.0256$. It is important to observe that the location of the touch surface is where we would expect people to look when not engaged in the interaction.

4. Conclusions and Future Work

We have presented a rich dataset of multi-party interactions recorded using multiple depth sensors, eye trackers, high resolution cameras, close talking microphones and touch surface, all aligned temporally and in *3D* space.

The data will be used to build models of visual attention for task-based human-robot interaction. We have started to investigate frame-based predictive models for the candidate visual attention target of the robot based on the data generated by the two participants. This will be investigated further and presented in forthcoming contributions. These models will take advantage of the rich set of modalities present in the dataset (all of which can be captured in real-life human-robot interaction) - head, body and face motion parameters, touch events and speech (we plan to automatically transcribe the recorded speech), in order to predict realistic and human-like visual attention behavior, which will be implemented and evaluated on the Furhat robot. Future work will also involve analysis of the data, for example into the differences in interaction behavior of the participants and correlation between different modalities/data streams.

5. Acknowledgements

This work has been carried out with support from the ICT The Next Generation initiative. We would like to thank Gabriel Skantze and Martin Johansson for the support during the set up of Furhat and all participant who took part in the recordings.

6. Bibliographical References

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The ami meeting corpus: A pre-announcement. In *Proceedings of the 2nd International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39. Springer-Verlag.

Hung, H. and Chittaranjan, G. (2010). The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM'10, pages 879–882. ACM.

Johansson, M., Skantze, G., and Gustafson, J. (2013). Head pose patterns in multiparty human-robot team-building interactions. In *Proceedings of the International Conference on Social Robotics*, ICSR'13, pages 351–360. Springer.

Moubayed, S. A., Skantze, G., and Beskow, J. (2013). The furhat back-projected humanoid head-lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics*, 10(1).

Nguyen, L. S., Frauendorfer, D., Mast, M. S., and Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 16(4).

Oertel, C., Cummins, F., Edlund, J., Wagner, P., and Campbell, N. (2013). D64: a corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1–2).

Oertel, C., Funes Mora, K. A., Sheikhi, S., Odobez, J.-M., and Gustafson, J. (2014). Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, UM3I'14, pages 27–32. ACM.

Skantze, G. and Al Moubayed, S. (2012). Iristk: A statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI'12, pages 69–76. ACM.

Skantze, G., Johansson, M., and Beskow, J. (2015). A collaborative human-robot game as a test-bed for modelling multi-party, situated interaction. In *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, IVA'15, pages 348–351. Springer.

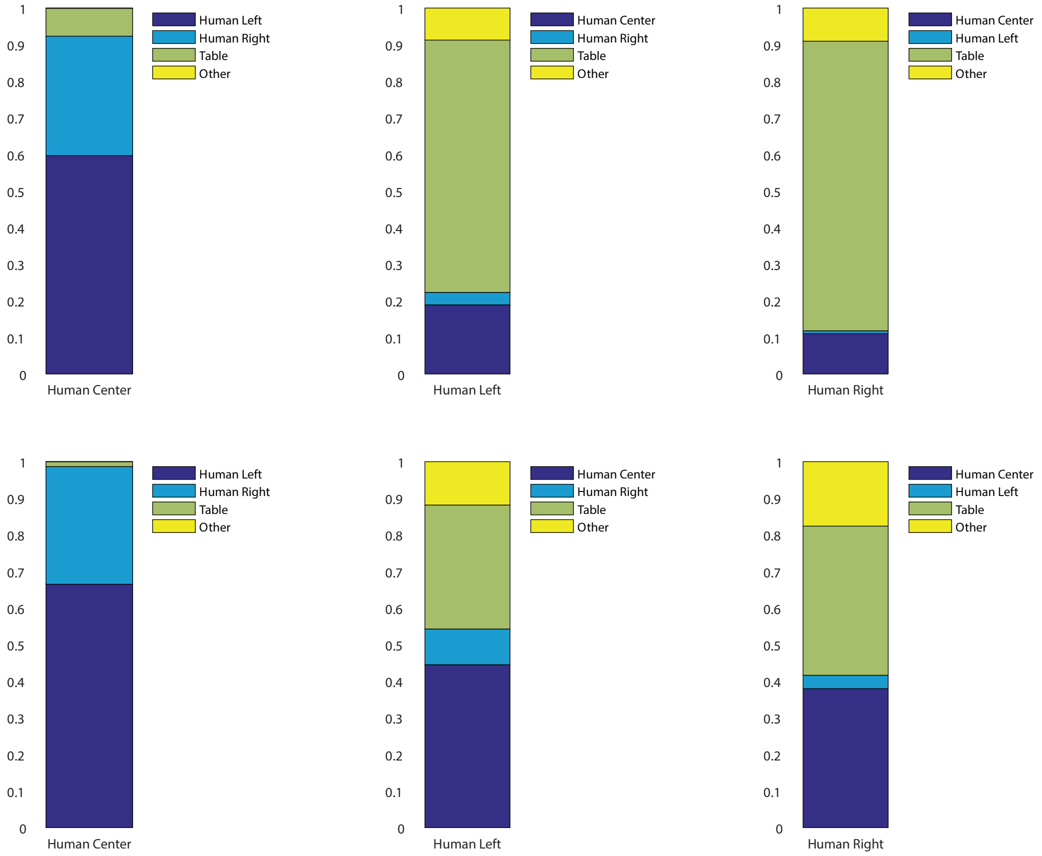

Figure 5: Targets distributions for the total length of six sessions; `Condition 2` (top) versus `Condition 3` (bottom). For example, the top left bar illustrates *"in* `Condition 2` *the human in the center looks at each of the other four targets certain fraction of the total length of all six interactions"*.