Building a Schema for the Description of HRI Experiments

Minsu Jang[1], Cheonshu Park[1], Jaehong Kim[1]

***Abstract*— We present an intermediary work on building a schema for describing HRI experiments in a formal way. A set of properties that can describe characteristics and metrics of experiments are extracted, and a simple data description schema based on the properties is introduced with an exemplary sample table-like descriptions. We plan to formalize the schema into an ontology so that researchers can reference for designing new HRI experiments and to come up with a set of standardized experiment processes and elicit benchmark measures.**

I. INTRODUCTION

Standardizing evaluation metrics and experimental scenarios for human-robot interaction is a seemingly impossible task to undertake as the styles and steps of interaction as well as evaluation criteria are widely different across various application areas. The first step of tackling the hard problem is to collect as much information as possible from previous works done in the field, analyze them, and build a formal database that can easily be queried for reference.

We present our preliminary work on building a formal description framework for the representation of the multiple aspects of HRI experiments. In section II, we enumerate the first batch of properties that can well characterize experiments. Then in section III, a set of sample descriptions in the tabular form is presented. Finally, future work is suggested in the last section.

II. A SCHEME FOR DESCRIBING HRI EXPERIMENTS

HRI experiments can be described by a set of properties that might be used to classify them into a number of categories. The categories can be referenced later for designing new HRI experiments and reuse related metrics. We succinctly enumerate and describe some of the essential and frequent properties in the following subsections.

A. Scenario Properties

An HRI experiment can basically be characterized and categorized by the properties of the domain scenario or task.

1) Application Area: A list of keywords that define specific domain is specified. Some examples include tele-presence, entertainment, guidance, medical, education, elderly-care, health-care, physical assistant, manipulation etc. Yanco et al. suggested a similar category called *TASK* with several example values e.g. *urban search and rescue*, *walking aid for the blind* and *delivery* [1].

[1]Minsu Jang, Cheonshu Park and Jaehong Kim are with Human-Robot Interaction Research Section, Electronics and Telecommunications Research Institute, Gajeong-ro 218, Daejeon-si, South Korea {minsu,bettle,jhkim504}@etri.re.kr

2) No of Interacting Partners: The number of participants a robot handles in an interaction session is an important indicator of cognitive capability. The interaction can be classified as one-to-one or multi-party. For the latter case, an integer value indicating the maximum number of participants can additionally be specified. Yanco et al. introduced a more general property called *HUMAN-ROBOT-RATIO* that denotes a non-reduced fraction of the number of humans over that of robots [1]. In our scheme, we consider more socially situated interactions where one robot interacts with one or more human users.

B. System Properties

Robots and accompanied system components may have different capabilities and be controlled in various ways.

1) Level of Autonomy: A robot system can be operated in different levels of autonomy. The most common method is called the WoZ(Wizard-of-Oz). It employs human operators to control robot systems. Fully automated operation is at the other extreme, and variable autonomy in between. This property is usually specified by a number from 1 to 10, in which the smaller value indicates the lower level of autonomy.

2) Interaction Modality: Robots interact in multi-modality in most scenarios. But in some cases limited modalities are employed e.g. either verbal only or non-verbal only. The value of this property is specified by a set of modalities.

3) Robot Platform: The name of robot platforms employed in an experiment are specified. Additional information such as a link to the specification of the platform can be accompanied.

C. Demographic Information

Demographic data is important as it often suggest the level of objectivity and reliability of experiments.

1) No of Participants: The number of participants is a major indicator of the reliability of the experimental results.

2) Age Distribution: This is usually specified by the mean age and the standard deviation of age distribution.

3) Gender Distribution: This indicates the gender specificity of the results.

D. Metrics

A number of metrics for evaluating robot performances have been introduced in HRI literatures, and a comprehensive review of them exists [2]. Some metrics are general enough to be widely employed, but many others have been invented for specific task-oriented experiments. Most popular examples of the former are the metrics defined in the Godspeed questionnaire such as anthropomorphism, animacy and likeability [3], while those for the latter include task efficiency, interaction fluency etc. Steinfeld et al. divided HRI metrics into two categories of *task* and *common*, and suggested a lengthy list of metrics with extensive references [2].

In this subsection, we introduce a scheme for specifying properties of a metric. Specific metrics collected from previous works are listed in Table I.

1) ID: A universally uniquely identifiable name is assigned to each metric e.g. uniform resource identifier.

2) Name: The literal name of the metric.

3) Measuring Method: Two representative methods of measuring a metric in HRI experiments are by user surveys or by automated or manual observations and quantification.

4) Subjectivity: A metric is either subjective or objective. The former includes those measured by survey, and the latter by observation.

5) Instrumentation: Some metrics are designed elaborately and validated in a number of studies that they are established as pseudo-standards. Some examples include Godspeed questionnaire [3] and PARADISE framework [4].

E. Experiment Annotation Properties

Several annotations can be accompanied to an experiment description for better understanding of its various contexts.

1) ID: A universally uniquely identifiable name is assigned to each experiment e.g. uniform resource identifier.

2) Year: The year of the experimentation.

3) Datetime: The date and time of the experimentation.

4) Duration: The duration of the experimentation.

5) Organization: The list of organizations involved in the experimentation.

6) DOI: A digital object identifier of the reference.

III. DESCRIBING HRI EXPERIMENTS

We have so far surveyed 20+ recent HRI papers that include user studies, and built a tabular description using a subset of the properties introduced in the previous section. The first step of building descriptions is to collect metrics and their properties. Table I shows a list of descriptions of subjective and objective metrics. Then, HRI experiments are described by specifying the properties and their values, as shown in Table II.

By building experiment descriptions, a researcher can effectively search for viable evaluation metrics to employ for her or his experiments. By querying the description database, one can easily extract major metrics that are most widely adopted for HRI system evaluation. Also, by the analysis of the meaning of each metric, we can identify the metrics that are specified in a different words but have the same semantics, which might allow us to standardize the terminologies for specifying various aspects of HRI experiments including evaluation criteria and metrics.

The final product of our work shall be a formal description language or an ontology that can be used to specify HRI experiments, which can be referenced and queried by human users as well as machines. We plan to design our description scheme in W3C's OWL Web Ontology Language [5].

IV. FUTURE WORK

a) Expanding the experiment database: Incorporating as many research papers as possible into the description database would be the most crucial task in the process of building a widely acceptable description framework for HRI experiments. We are trying to review research papers presented or published in the HRI-related conferences and journals in recent 5 years.

b) Expanding the description scheme: The current scheme has several limitations, one of which is that it does not differentiate between *experiment* and *experimentation*. Suppose that $EX01$ describes an experiment to evaluate the social presence of a tele-presence robot with or without a non-verbal expression capability. If the description includes sufficient information to replicate the experiment, multitudes of researchers might be able to conduct $EX01$ and report results, which might be a great step forward to standardized experiments. Defining an *experimentation* as *an act of conducting an experiment*, there might be many experimentations for an experiment. We plan to elaborate our design so that an experimentation and its results can be described independent of experiment.

Another crucial part of this work is to incorporate metrics and taxonomies studied in the seminal previous works [2], [1].

ACKNOWLEDGMENT

This work was supported by the Robot R&D Program (10041659) funded by the Ministry of Trade, Industry and Energy (MOTIE, South Korea).

REFERENCES

[1] H. A. Yanco, J. L. Drury, *et al.*, "Classifying human-robot interaction: an updated taxonomy." in *SMC (3)*, 2004, pp. 2841–2846.

[2] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for human-robot interaction," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 2006, pp. 33–40.

[3] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," in *Metrics for HRI Workshop, Technical Report*, vol. 471. Citeseer, 2008, pp. 37–44.

[4] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Paradise: A framework for evaluating spoken dialogue agents," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 271–280.

[5] D. L. McGuinness, F. Van Harmelen, *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[6] S. O. Adalgeirsson and C. Breazeal, "Mebot: a robotic platform for socially embodied presence," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 15–22.

[7] J. Fasola and M. J. Matarić, "Using socially assistive human–robot interaction to motivate physical exercise for older adults," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2512–2526, 2012.

[8] J. Hall, T. Tritton, A. Rowe, A. Pipe, C. Melhuish, and U. Leonards, "Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge," *Robotics and Autonomous Systems*, vol. 62, no. 3, pp. 392–399, 2014.

TABLE I

Descriptions of HRI Evaluation Metrics

ID	Name	Method	Subjectivity	Relevant Instrumentation	Reference
SM01	Social Presence	Survey	Subjective	NA	[6],[7]
SM02	Trust	Survey	Subjective	Grotz's Individualized Trust Scale	[6]
SM03	Cooperation	Survey	Subjective	Takayama questions	[6]
SM04	Engagement	Survey	Subjective	The Temple Presence Inventory	[6],[8]
SM05	Enjoyment	Survey	Subjective	NA	[9],[7]
SM06	Understandability	Survey	Subjective	NA	[9]
SM07	Convenience of Conversation	Survey	Subjective	NA	[9]
SM08	Compliance	Survey	Subjective	NA	[10]
SM09	Competency	Survey	Subjective	NA	[10],[11]
SM10	Persuasiveness	Survey	Subjective	NA	[10]
SM11	Sociability	Survey	Subjective	NA	[10]
SM12	Trustworthiness	Survey	Subjective	NA	[10]
SM13	Usefulness	Survey	Subjective	NA	[7]
SM14	Companionship	Survey	Subjective	NA	[7]
SM15	Capabilities as an Exercise Coach	Survey	Subjective	NA	[7]
SM16	Anthropomorphism	Survey	Subjective	GODSPEED Questionnaire	[12]
SM17	Animacy	Survey	Subjective	GODSPEED Questionnaire	[12]
SM18	Likeability	Survey	Subjective	GODSPEED Questionnaire	[9],[11],[12],[8]
SM19	Perceived Intelligence	Survey	Subjective	GODSPEED Questionnaire	[12],[7]
SM20	Perceived Safety	Survey	Subjective	GODSPEED Questionnaire	[12]
SM21	Overall Impression (Pos,Neg,Neutral)	Survey	Subjective	NA	[13]
SM22	Intention to Use	Survey	Subjective	NA	[13]
SM23	Level of Robot Understanding	Survey	Subjective	NA	[8]
SM24	Naturalness of Robot Behaviors	Survey	Subjective	NA	[11]
SM25	Perceived Agreeableness	Survey	Subjective	NA	[14]
SM26	Perceived Similarity	Survey	Subjective	NA	[14]
OM01	Mean Interaction Time Per Session	Observation	Objective	NA	[15],[7]
OM02	The Duration of Subject's Gaze Toward Robot	Observation	Objective	NA	[15]
OM03	The Number of Subject's Looks Toward Robot	Observation	Objective	NA	[15]
OM04	The Level of Subject's Knowledge Gain	Survey	Objective	Evaluation by Quiz	[15]
OM05	Task Success Rate	Observation	Objective	NA	[16]
OM06	Average Time for Task Completion	Observation	Objective	NA	[7]
OM07	Robot's Feedback Percentage	Observation	Objective	NA	[7]
OM08	Task Success	Observation	Objective	PARADISE Framework	[12]
OM09	Dialog Quality	Observation	Objective	PARADISE Framework	[12]
OM10	Dialog Efficiency	Observation	Objective	PARADISE Framework	[12]
OM11	Information Recall Correctness	Observation	Objective	Evaluation by Quiz	[11]
OM12	No of Interactions Per Day	Observation	Objective	NA	[17]
OM13	Accuracy of Friendship Estimation	Observation	Objective	NA	[17]
OM14	No of Subject's Decision Changes Caused by Robot	Observation	Objective	NA	[14]

TABLE II

An HRI Experiment Database (ID: The ID of an Experiment, App. Area: Application Area, LoA: Level of Autonomy, NoP: No of Participants, AD: Age Distribution, GD: Gender Distribution(M:Male,F:Female,U:Unknown))

ID	DOI	Year	App. Area	LoA	Modality	Platform	NoP	AD[$\mu(\sigma)$]	GD	Metrics
EX01	[6]	2010	Telepresence	1	Non-Verbal	MeBot	48	23.21(8.92)	24F/18M	SM01,02,03,04
EX02	[9]	2012	Guidance	10	Mixed	Furhat	86	35	39F/46M/1U	SM05,06,07,18
EX03	[10]	2013	Guidance	10	Verbal	Mindstorm	48	23.69(7.83)	24F/24M	SM08,09,10,11,12
EX04	[15]	2013	Child-care	1	Mixed	Nao	20	9.5	NA	OM01
EX05	[15]	2013	Child-care	1	Mixed	Nao	10	9.5	NA	OM02,OM03
EX06	[16]	2013	Guidance	10	Verbal	Mindstorm	20	NA(18∼65)	11F/9M	OM05
EX07	[7]	2012	Elderly-care	10	Mixed	Bandit	13	83(77∼92)	12F/1M	SM01,13,14,15,OM06,07,01
EX08	[7]	2012	Elderly-care	10	Mixed	Bandit	24	77(68∼89)	19F/5M	SM13,19,OM06,07,01
EX09	[12]	2012	Serving	10	Verbal	iCat	31	27.9(21∼50)	9F/22M	SM16,17,18,19,20,OM08,09,10
EX10	[13]	2013	Guidance	5	Mixed	Robovie	40	NA	NA	SM21,22
EX11	[8]	2013	Conversation	10	Non-Verbal	Birt	63	20.4(18∼32)	48F/15M	SM04,18
EX12	[11]	2013	Conversation	10	Mixed	Wakamaru	32	24.9(18∼61)	NA	SM09,18,24,OM11
EX13	[17]	2012	Education	5	Mixed	Robovie	NA	NA	NA	OM12,13
EX14	[14]	2012	Conversation	1	Mixed	Robosapien	40	NA	20F/20M	SM25,26,OM14

[9] S. Al Moubayed, J. Beskow, B. Granström, J. Gustafson, N. Mirning, G. Skantze, and M. Tscheligi, “Furhat goes to robotville: a large-scale multiparty human-robot interaction data collection in a public space,” in *Proc of LREC Workshop on Multimodal Corpora*, 2012.

[10] S. Andrist, E. Spannan, and B. Mutlu, “Rhetorical robots: making robots more effective speakers using linguistic cues of expertise,” in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 341–348.

[11] C.-M. Huang and B. Mutlu, “The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior,” *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 80–102, 2013.

[12] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. Petrick, “Two people walk into a bar: Dynamic multi-party social interaction with a robot agent,” in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 3–10.

[13] D. F. Glas, S. Satake, F. Ferreri, T. Kanda, N. Hagita, and H. Ishiguro, “The network robot system: enabling social human-robot interaction in public spaces,” *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 5–32, 2012.

[14] L. Takayama, V. Groom, and C. Nass, “I’m sorry, dave: i’m afraid i won’t do that: social aspects of human-agent conflict,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 2099–2108.

[15] T. Belpaeme, P. E. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, *et al.*, “Multimodal child-robot interaction: Building social bonds,” *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[16] C. Breazeal, J. Gray, and M. Berlin, “An embodied cognition approach to mindreading skills for socially intelligent robots,” *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 656–680, 2009.

[17] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro, “A two-month field trial in an elementary school for long-term human–robot interaction,” *Robotics, IEEE Transactions on*, vol. 23, no. 5, pp. 962–971, 2007.