Building Datasets for Training Robots' Social Intelligence

Minsu Jang · Do Hyung Kim · Jaeyeon Lee · Jaehong Kim
Human-Machine Interaction Research Group,
Intelligent Robotics Research Labs
Electronics and Telecommunications Research Institute,
South Korea
{minsu, dhkim008, leejy, jhkim504}@etri.re.kr

Abstract Project AIR(Artificial Intelligence for Robots) aims to build software modules that can mimic human social intelligence based on machine learning. In this paper, we provide specifications of various datasets we identified as being valuable for training social intelligence. Datasets for lip-reading, speech and affect recognition, human tracking, action recognition, gesture generation are being collected in the labs, test beds and living labs in the domain of elderly people's life care. We describe details of the datasets and how they can be used. We believe these datasets will contribute to the advancement of machine learning techniques for developing social robots and social intelligence.

Keywords Social Robots · Social Intelligence · Datasets

1 Introduction

Remarkable progress of machine learning technologies in recent years are making it possible to build robots with intelligence trained on various datasets. Image and speech recognition have become common skills for most robots. Difficult control problems like dexterous manipulation [1] and navigation in unknown area are being solved by machine learning such as deep reinforcement learning.

In our ongoing project, called AIR, "we aim to develop artificial intelligence technologies for human-care robots that can provide personalized socially assistive services for elderly people" [2]. Based on the idea that elder-care robots need to be able to conduct *empathic* interactions to be accepted as a companion and be used consistently for longer period of time, we identified two major technical areas of robot social intelligence: user profiling and social interaction.

User profiling has the goal of "*acquiring wider and deeper knowledge about human users*", and is composed of various skills for identifying and understanding many attributes of human users. Social interaction enables robots to conduct natural interactions with human users.

User profiling is composed of the following technical components:

- Human detection and tracking in the cluttered home environment using ego-centric robot vision
- Speech recognition specialized for older adults' vocal characteristics
- Human identification and attribute recognition
- Action recognition for daily life activities
- Emotion recognition from user's facial expressions and activity patterns

Social Interaction is composed of the following technical components:

- Non-verbal interaction behavior generation e.g. hand-shaking, hi-five, hugging etc.
- Co-speech gesture generation
- Turn-taking based on multi-modal interaction cues

Our research to develop these technical components is primarily based on machine learning techniques because recent deep learning-based techniques show promising performance in various machine intelligence tasks and they can make robots continually adapt to changing environments.

A major issue in applying machine learning techniques to develop robot social intelligence is the lack of training datasets. There are already a large number of datasets for machine learning and new datasets appear day by day nowadays, but they are mostly for identifying factual knowledge e.g. object classification, image segmentation, speech recognition etc. For training social intelligence, we need datasets to train more nuanced and subjective knowledge such as human attributes and their changes, emotions, moods, causality, non-specific daily actions, co-speech gestures etc. In table 1, we identified the properties of common recognition tasks and robots' social recognition tasks for social intelligence. To train social intelligence, we need datasets specialized for social recognition tasks.

<Table 1> Common recognition vs. social recognition

Common Recognition	Subject	Social Recognition
General Noun Existence (*There is a person.*)	Human	**Proper Noun** Identity, Attributes, Relations (*There is Mark who likes to wear red shirt and laughs a lot.*)
Verb Facts (*Appeared, Gone, Sat, Cleaning, Cooking*)	Action	**Adjective/Adverb** Quality (*Friendly, Unfamiliar, Indifferent, Engaged*)
Factual Sentence (*A person enters into the house., A person gets on the bus.*)	Context	**Causal Sentence** (She is happy to hear praise., He is going around because he's feeling lonely.

Domain specialization is another issue with datasets. We deal with human-care robots for elder-care. Datasets for user profiling should be collected for older adults. Also, recognition tasks should be performed by taking data from robots. For image recognition tasks, we need images collected from cameras that are mounted on mobile platforms. There are not many datasets that satisfy these conditions. Thus, building datasets for social intelligence has become a major and essential mission for project AIR.

2 Datasets for Robot Social Intelligence

In this section, we describe various datasets that are being built for training robot social intelligence.

3.1 Human Detection/Tracking

Human-care robots should be able to robustly detect and track older adults in the cluttered home environments. This dataset is composed of 13,000 images of older adults in various home environment settings. Images were collected from various sources including documentary videos and living labs.

Figure 1. A sample from human detection/tracking dataset

3.2 Speech Recognition

Due to specific vocal characteristics of older adults, ordinary speech recognizers do not perform well on older adults' speech data. To make speech recognizers specialized for older adults, a large-scale speech dataset has been built. Dataset is composed of 36,000 hours of speech data recorded by interviewing 3,600 older adults.

3.3 Human Attributes Recognition

Human attributes characterizes users as they can represent preferences, habits etc. Robots can make friendly comments or suggestions by observing user's appearances such as clothes and accessories. Human attribute dataset is composed of 10,000 images of older adults in various outfits. Dataset is annotated with multi-labels including 9 cloth classes with 31 attribute labels and 6 accessory classes with 2 attribute labels.

Figure 2. A sample human attribute data

3.4 Lip Reading

Lip reading is necessary for speech-based human-robot interaction for detecting user's speech activity while robot cannot perform speech recognition e.g. while robot is playing speech sound. Lip reading dataset consists of 12 words pronounced by 50 older adults, each word of which for 10 times.

Figure 3. Lip reading images and preprocessing results

3.5 Video Affect Recognition

Emotion can be expressed not only by facial expressions but also by action patterns and situational contexts. Video affect recognition tries to recognize emotional moods by observing user's actions. A dataset of video clips has been collected from TV shows in which older adults appear. 500 video clips were annotated with 5 sentences containing emotional descriptions. The recognizer trained on this dataset can produce sentential descriptions from a sequence of views.

Figure 4. Scenes from the video clip dataset

3.6 Action Detection/Recognition

For daily action recognition, we identified 55 common actions by observing daily activities of 53 older adults of higher than 70 years of age. Actions include eating with a folk, polishing vegetables, applying lipstick, hanging laundry etc., that are not common in previous action datasets. A large scale RGBD-S data has been collected from 9 living labs and a test bed. From living labs, 150 hours of video clips have been recorded and annotated. From the test bed, 50 older adults were recruited and videos were recorded from 32 different viewpoints, which results into 88,000 video clips of 55 actions.

Figure 5. Action data from living labs

3.7 Non-Verbal Interaction Behavior Generation

In conventional social robots, non-verbal interaction behaviors like hand-shaking and hugging are hand-crafted

by animators and replayed by robots when a specific event is recognized e.g. a user says 'let's handshake'. In this project, we wanted to make robots recognize interaction contexts by their own and generate interaction behaviors properly synchronized both in temporal and spatial configurations. For that purpose, we are applying end-to-end learning using sequence-to-sequence model. The training dataset is composed of 2,500 RGBDS clips of two older adults playing 10 interaction behaviors including bowing when someone enters, coming closer when someone calls, handshaking, hi-five, departing when an interaction partner waves hand etc. Three Kinect 2 cameras were employed to obtain non-occluded views of two interacting people.

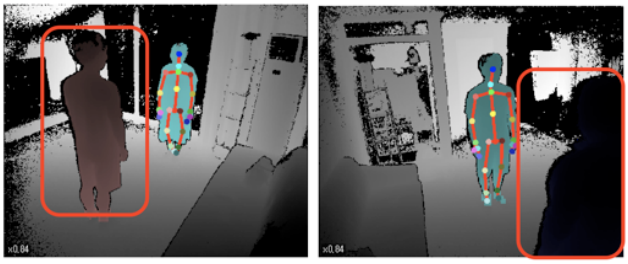

Figure 6. Two views from the non-verbal interaction behavior generation dataset

3.8 Co-Speech Gesture Generation

Co-speech gestures including iconic, metaphoric, deictic and beat gestures facilitate engaging and informative communication by allowing communicators to channel ideas in two separate modalities [3]. We believe properly generated co-speech gestures would elevate the effectiveness and acceptance level of social robots.

Our approach to co-speech gesture generation involves end-to-end learning from speech text to synchronized co-speech gestures. The first target domain is non-reciprocal scenarios like public speech. For training, we collected a set of 1,295 video clips of TED talks. Each clip has been annotated with synchronized transcripts and upper-body poses. Poses were extracted frame by frame automatically using OpenPose [4].

3.9 Multimodal Cue-based Turn-Taking

Spontaneous and contextually proper turn-taking is an essential social skill for natural interaction, but is very challenging to realize in a social robot. Humans interpret a number of social cues to estimate and predict floor management actions. Social cues include verbal signals including linguistic and non-linguistic features, gestures, body poses, facial expressions etc.

To train robots to predict turn-taking intentions of an interacting partner, we collected a video clips of two older adults casually conversing face-to-face. We recruited 102 people including 60 males and 42 females and recorded 51 interaction sessions. The dataset is consisted of 34 hours of video footages. Each clip is annotated with four floor management events e.g. take, release, wait and hold. Additional labels are also added to the dataset such as lip open/close state, gaze direction and positive/negative intentions.

3 Conclusion

In this paper, we described datasets designed and built for training social intelligence. Datasets specialized for the domain of social robots and elderly-care are not common and we believe that these datasets will contribute to the technical advancement in these domains. With this paper, we hope that discussions and efforts to identify and build datasets for training robot social intelligence be encouraged.

Acknowledgement This work was supported by the ICT R&D program of MSIP/IITP. [2017-0-00162, Development of Human-care Robot Technology for Aging Society]

References

1. OpenAI (2018). Learning Dexterous In-Hand Manipulation. CoRR, abs/1808.00177.
2. Minsu Jang, Jaehong Kim, and Jaeyeon Lee (2018). Project AIR: Developing Artificial Social Intelligence for Human-Care Robots, Workshop on Social Human-Robot Interaction of Human-Care Service Robots.
3. Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. Psychological bulletin, 137(2), 297.
4. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1611.08050