Inferring Human Feelings and Desires for Human-Robot Trust Promotion

Xingzhi Guo[1], Yu-Cian Huang[1], Edwinn Gamborino[2], Shih-Huan Tseng[3], Li-Chen Fu[1,2(✉)], and Su-Ling Yeh[1,2]

[1] National Taiwan University, Taipei 10617, Taiwan
lichen@ntu.edu.tw
[2] NTU Research Center for AI and Advanced Robotics, Taipei 10617, Taiwan
[3] National Kaohsiung University of Science and Technology, Kaohsiung 82445, Taiwan
http://robotlab.csie.ntu.edu.tw/, http://ai.robo.ntu.edu.tw/

Abstract. Trust is a key component in developing successful interpersonal relationships. In this paper, we posit that the same is true for Human-Robot Interaction (HRI), since human trust toward robots can facilitate HRI in terms of comfort and usability. We investigated the ability of a socially assistive robot to promote trust in the social relationship with its user by inducing self-disclosure of the user's negative experiences and offering coping mechanisms to deal with these. To achieve this purpose, our system is equipped with deep learning techniques to detect the user's negative facial expressions, which in turn can be used as cues for the robot to proactively induce self-disclosure. Once triggered, using a conversational model, the robot engages the user to determine the cause of their negative mood. Then, it infers the user's internal feelings by applying Markov Chain Monte Carlo (MCMC) inference over a Bayesian Network on the user's utterance. Combining the information gathered from the concept inferencing process and the self-disclosure content, the system is able to estimate a set of desires from the Bayesian Network. Experiments show that our proposed work can correctly infer the user's feelings and desires from their utterances, as well as generate an appropriate response, resulting in the improvement of human's trust toward the robot.

Keywords: Human-Robot trust · Social robot companion · Bayesian network · Reinforcement learning · Commonsense knowledge graph

This research was supported in part by the Joint Research Center for AI Technology and All Vista Healthcare under Ministry of Science and Technology of Taiwan (MOST grants 107-2218-E-002-009, 107-2634-F-002-019 and 108-2634-F-002-016) and Center for Artificial Intelligence & Advanced Robotics, National Taiwan University.

P.-L. P. Rau (Ed.): HCII 2019, LNCS 11576, pp. 365–375, 2019.
https://doi.org/10.1007/978-3-030-22577-3_26

1 Introduction

In recent years, service robots have become ubiquitous in several aspects of our daily lives. Furthermore, they are expected to have long-term social interactions with their users. In these social tasks, one of the robot's pro-social factors—trust—plays an extremely important role. As a corner stone of Human-Robot Interaction, trust between humans and robots has been explored frequently by researchers from different disciplines (psychology and computer science). Researchers [12] believe that (1) a trustful relationship between humans and robots can prevent misusage or overuse of the robot, and that (2) Human-Robot trust can enhance human's reliance on robots.

According to one of the most accepted definitions in the literature [15], trust is the willingness to expose one's own vulnerabilities. Generally speaking, the vulnerabilities of people are related to negative life experiences. If a robot companion is able to properly induce a person to self-disclose said vulnerabilities and initiate a meaningful interaction, a chance for human-robot trust promotion can be created. Previous research in the fields of robotics [3,8,16] and psychology [4,9] suggest that for a social robot to be able to deal with a person's vulnerabilities and promote trust, it should display the following features in interaction: (1) empathy, (2) goodwill, and (3) awareness of personal preference. Empathy, a feature of social interaction present in humans and other animals, is the ability to understand and internalize the experience of others into oneself through means of verbal and non-verbal communication. Goodwill is defined as a generally friendly, helpful or cooperative attitude, which can be reflected in an agent's behavior. Finally, awareness of personal preference refers to the realization that different individuals, due to a variety of unknown factors, differ in their preferred objects, persons, environments, etc.

Studies on human vulnerability as well as the three factors mentioned above inspired the development of this work. In order to address the problem of human-robot trust promotion, we designed an interactive conversation system for a robot companion to induce a person to self-disclose their vulnerabilities and infer the user's emotional feelings and desires given what has happened to them through a commonsense knowledge graph. By adopting these two factors, the robot companion is able to appropriately handle its user's vulnerabilities by generating an appropriate response and ultimately promote trust through social interaction.

The remainder of this paper is organized as follows: Sect. 2 provides a survey on the related literature that inspired the present work. Section 3 further details the contributions of this work, which surround the development of the interaction system for Human-Robot Trust promotion: The vision module (Sect. 3.2) must be able to capture the user's facial expression and discriminate when the mood of the user is negative in real time, using this information as a cue to engage the user. Once in interaction, in order to induce self-disclosure, we built an inference model and a causal commonsense knowledge base based on ConceptNet [17]; with this information on hand, the robot has the capability of understanding human's common sense and therefore, the internal desires and causes for their negative

outlook (Sects. 3.3 and 3.4). In Sect. 4 we present an experiment designed to evaluate the user experience when interacting with our system as well as the self-reported perceived trust towards the robot. In Sect. 5, we finalize the document with a few closing remarks.

2 Related Works

In order to establish a firm theoretical background for the proposed idea, we review the literature from the research fields of Social Psychology and Human-Robot Interaction related to trust. In fact, psychologists have investigated interpersonal trust and presented several different trust models, as shown in Sect. 2.1. For HRI researchers, these interpersonal trust models involve factors of cognition, emotion and individual preference.

2.1 Interpersonal Trust and Human-Robot Trust

Interpersonal trust refers to trust between two or more individuals, a common phenomenon in our daily lives. One foundation of the interpersonal trust is cognition. Baier [1] suggested trust is accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one". In other words, the formation of cognition-based trust is taking into account the central elements: partners' competence, responsibility and goodwill. Besides cognition-based trust, Lewis and Wiegert [11] claimed that affective foundations for trust also exist, consisting of the emotional bonds between individuals. In addition, Lewicki *et al.* [10] also proposed an evolutionary interpersonal trust model. In this proposed model, the highest level of trust can be built only when a partner can fully understand another's value and preference and, therefore, take actions in favor of their partners. Thus, our papers explore the factors of empathy, goodwill and personal preference in the formation of Human-Robot Trust.

Lee [7] investigated the effects of robot's nonverbal behaviors on Human's trust during social interactions. Their experiments manipulated robot's gestures during Human-Robot Interaction (HRI), and afterwards, a trust game was conducted in order to measure the person's trust level towards the robot with different gestures during the prior interactions. The results of the experiments showed three positive gestural cues of robot for developing trust as follows: leaning-forward, having-arms-in-lap and open-arms.

Martelaro et al. [13] investigated human's trust and sense of companionship in HRI by manipulating robot's vulnerability and expressivity. The vulnerability of the robot was displayed via a personalized conversation. The expressivity was displayed by multimodal interaction (*i.e.* verbal and non-verbal behaviors). Their results showed that participants reported more trust and feelings of companionship with a vulnerable robot, and reported disclosing more of their internal feelings and vulnerabilities with an expressive robot when compared to a non-expressive robot.

Mota et al. [14] presented a pilot study about how people judge trustworthiness of a robot during social Human-Robot Interaction. They examined this phenomenon using 'Trust Game', a common scenario in behavioral economics. Qualitative results suggested that participants may follow a human-robot trust model which is quite similar to the interpersonal trust model. In addition, they also found that people try to interact socially with robots, but due to lack of common social cues, they draw from prior interpersonal social experience, or create new experiences by actively exploring the robot's behaviors.

In summary, most of the existing works exploring human-robot trust focused on cognition-based trust by manipulating the robot's performance during the tasks.

2.2 Emotional Feeling Inference

Among the crucial factors for trust promotion explored earlier (*i.e.* empathy, goodwill and personal preference), empathy is the only factor that involves building an affective bond with one's peers. From the definition of Paivio *et al.* [2], empathy is strongly related to one's ability to understand the emotional feelings of others. For example, a person can easily understand that losing money may cause feelings of sadness, anger and disappointment. Furthermore, a person with a strong sense of empathy may then display an appropriate response, based on the inferred emotional feelings of their partner. In other words, the inference of emotional feelings is the first step toward displaying empathy.

There are some existing works addressing the emotion inference tasks from the perspective of Natural Language Processing (NLP). Most of the research in this field considers this task as a multi-class classification problem where the classes are the human's emotions and the observed input data is the human's utterance (*i.e.* the text content of the speech).

3 Human Emotional Feeling and Desire Inference System

As mentioned earlier in this paper, the scope of this project is to build this system into a social robot companion that is able to promote trust with its user by interaction. Towards meeting this goal, in this paper we describe in detail a method by which a computer interface is able to detect a negative mood in the user based on their facial expression and then, using natural language, communicate with the user to expose these vulnerabilities and inferring the internal feelings and desire that cause their emotional distress. In the following subsections, the visual module and desire and feeling inference module are discussed.

3.1 Visual Module

Advanced service robots must integrate capabilities to detect human's presence in their vicinity and interpret human facial expressions. Therefore, facial expression recognition is a crucial capability for social robots, especially in the context of HRI.

Deep Temporal Appearance-geometry network (DTAGN) [5] is a deep learning model for human's facial expression recognition, which combines two deep networks: the deep temporal appearance network (DTAN) and the deep temporal geometry network (DTGN). As shown in Fig. 1, the DTAN, CNN-based model, is used to extract the temporal appearance feature for facial expression recognition. Meanwhile, the DTGN, a fully-connected DNN, is used to capture geometrical information about the motion of the facial landmark points. These two models are integrated in order to boost the performance of the facial expression recognition. In this paper, we apply the same architecture of DTAGN model and train it on Radboud Faces Database [6].

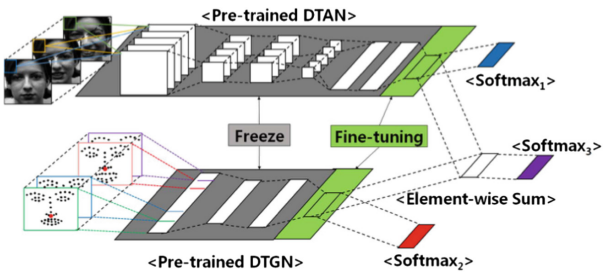


Fig. 1. The joint fine-tune architecture in face expression recognition.

3.2 Causal Commonsense Knowledge Graph

In order to display empathy and goodwill, the robot should be able to infer human's feelings and desires. In our system, we build a causal commonsense knowledge module, based on ConceptNet [17]. ConceptNet is the largest public commonsense knowledge base. It is composed of: (1) nodes: a natural language words/phrases; (2) edges: the relationships between two nodes; (3) weights: the level of intensity of each edge. The following sub-section will describe the feeling inference and desire inference modules separately.

Human Feeling Knowledge Graph. While ConceptNet contains a large variety of edges describing several types of relations between concepts, in this paper we set up two criteria to construct the human feeling knowledge graph: (1) The type of edge should be one of the causal relations shown in Table 1. (2) The end node should be an adjective phrase. Because of the characteristics of ConceptNet, most of the end nodes in a causal relation and containing adjective phrases is designed to describe human's feelings.

Human Desire Knowledge Graph. Similar to previous Human Feeling Knowledge Graph, the Human Desire Knowledge Graph can be constructed by applying two necessary criteria to screen certain nodes in the ConceptNet

Table 1. The causal relations in the ConceptNet and the examples of sentence patterns.

Relation	Sentence pattern
Causes	The effect of *VP* is *NP VP*
MotivatedByGoal	You would *VP* because you want *VP*
Desires	*NP* wants to *VP*
CausesDesire	*NP* makes you want to *VP*

and adding the valid node pairs to the knowledge graph. The first criterion is exactly the same as the first criterion in Human Feeling Knowledge Graph as described in the previous sub-section. For the second criterion, the ending node should be a verb phrase. This is because the objective of the desire inference module is to predict a probable action the user may want to do based on what they previously said.

3.3 Probabilistic Graphical Model for Inferring Feeling and Desire

We used Bayesian Networks to formulate the inference task as a maximum posterior problem, and applied Gibbs Sampling approximation algorithm to infer the posterior probability given the observed evidence. In our case, the observed evidence is the keywords extracted from the user's disclosure content, which are then mapped to the nodes in the pre-built commonsense knowledge graph as observed nodes.

Model Construction. In order to construct the Bayesian Network from the knowledge graph to model the dependencies among human's feelings, we retain the topological structure of the knowledge graph, and derive the conditional probability tables based on the raw weights of each edge.

We define the following concepts to help introduce our way to build the Bayesian network as shown in Fig. 2.

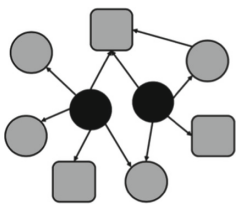


Fig. 2. An example of a constructed Bayesian network. Black nodes are seed nodes, square nodes represent feelings and circle nodes represent desires.

Seed: Seed is an observed node from which the Bayesian network starts to grow. The observations are all from the human's utterance. All the child nodes will be pulled out from the human's causal knowledge graph and become the structure of the resulting Bayesian network. There are two types of nodes in a causal knowledge graph — feelings and desires. Moreover, the Bayesian network could have more than one seed, depending on the content richness of the user's self-disclosure. All the child nodes of the seed nodes will be queried from the knowledge graph to construct the Bayesian network.
Width: Width refers to the maximum number of child nodes that are queried from the knowledge graph of each seed node. The definition of width can prevent useless data from the knowledge graph entering the Bayesian network. The candidate child nodes are first sorted by the corresponding weight, then only the top-N nodes, where N = width, can be added into the Bayesian network.
Depth: Depth is defined as the longest distance from the root, seed node in this case, to the corresponding queried child nodes. In this paper, depth is always set to be one under the assumption that human's desire or feeling is directly related to life events.

Conditional Probability Table. By querying the seed nodes and completing the topological structure of the Bayesian Network, we build the conditional probability table for each nodes based on the weights from the knowledge graph. The conditional probability table is set individually for each node in the Bayesian network as the CPT Builder algorithm shown in Fig. 3.

Function CPT_Builder(G) returns a complete conditional probability table $CPT(X)$
Local variables: X, all the nodes in the graph G.
W, the weights of the directional edges in the graph G.

for each x_i in X **do**
set $C = parents(x_i)$, the parent nodes of x_i
if $C = Null$ **then continue**
for each a_{ij} in the binary combinations in the length of C **do**

$$\text{set the value CPT}(x_i) = \begin{cases} sigmoid\left(\sum_c a_{ij} w_{c_j \to x_i}\right), x_i = True \\ 1 - sigmoid\left(\sum_c a_{ij} w_{c_j \to x_i}\right), x_i = False \end{cases}$$

where $w_{c_j \to x_i}$ is the weight from the parent node x_j to node x_i
return CPT(X)

Fig. 3. The algorithm for building the conditional probability table (CPT).

The algorithm assigns the conditional probability to each node x_i, except the seed nodes, by carrying out the calculation $CPT(x_i)$. The conditional probability of x_i is defined by a sigmoid based function, whose value range is [0.5, 1]. The lower-bound of the CPT function is set to be 0.5 since the existence of the causal relations means at least someone think the edge is valid; Moreover, the weights are all positive, therefore the result of sigmoid function can not be a negative value. The upper bound of the function is 1 since that is the natural characteristic of the sigmoid function.

Model Inference. The goal of the inference task is to find the feelings and desires in the user's mind given the observations.

As mentioned before, the evidence nodes are the observed life events in the human's self-disclosure. We use Gibbs Sampling approximation inference algorithm to infer the full joint probability of all nodes in the Bayesian Network, then the sampling results can be converted into the conditional probability by fixing the states of the observed nodes.

In this way, the posterior probability for all the nodes can be obtained. The larger the probability, the more likely it is the human's feeling or desire. Therefore, the unobserved nodes, feelings and desires, are ranked according to their values of posterior probability respectively, resulting in two ranked list for further usage.

4 Experiments

In order to evaluate the modules in the proposed system, an HRI experiment was designed to measure the system performance. Different from the explicit and direct evaluation for classification tasks with a testing or validation dataset, the proposed modules are in a human-in-loop system, involving human's subjective judgement; Thus, the evaluation for the proposed system is based on the ratings from human participants, using questionnaires or interviews.

Followed by this experimental paradigm, an HRI experiment is designed to measure the performance of Human Feeling Inference Module, Human Desire Inference Module and the overall Trust Promotion system. Participants were asked to interact with the robots with different configurations, as shown in Table 4-3, for two days, 30 mins per day and 12 interaction sessions in total. Also, Participants were asked to fill out questionnaires before and after the experiments based on their experience and observations in order to do the a temporal evaluation of the system.

In the following sections, the materials (Stimuli scenarios and questionnaires) used in the experiment are described. Afterwards, the results and discussions of each module and the overall system are presented.

4.1 Feeling Inference Evaluation

For this experiment, participants were asked to rate the performance of the system to infer human feelings. Participants were 23 males and 4 females ranging

from 22 to 48 years old (average = 25.0, stdev = 5.7). During an experiment session, a story describing a virtual character's negative life experience is presented to the participant. The purpose of the storytelling is to let the participant become immersed into the environment of the story, and substitute themselves into the role of the main character of the story.

The four stimuli scenarios were related to (1) Study Pressure; (2) Personal Affair; (3) Working Pressure and (4) Loneliness. The corresponding story descriptions were shown to the participants in order to encourage the feeling of character substitution.

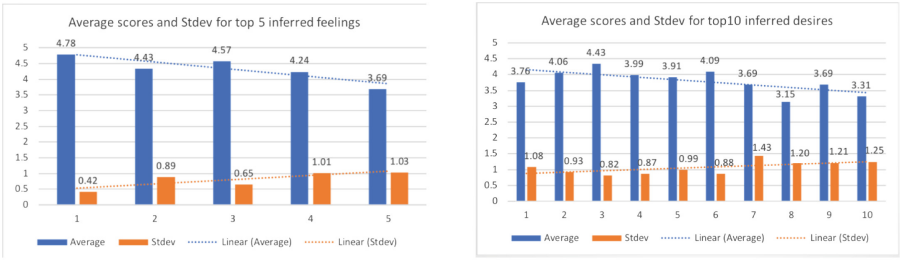
Afterwards, the robot started to induce the participant's self-disclosure. The participant would reply with the description about what happened to them (in the role of the virtual character) in natural language. Then, the human feeling inference module calculated the probability of each node in the constructed human's feeling knowledge graph. The inferred nodes were sorted with respect to the inferred probability. The top-5 nodes were shown to the participants, and rated by them using a Five-Point Likert scale. The score reflects the extent as to which the participant judged the inferred feelings as reasonable or correct given the scenario.

As show in Fig. 4(a), the overall average score for the inferred human's feeling was 4.25 with a standard deviation of 0.94. This score is significantly higher than average 3-pt (neither agree nor disagree), which can be interpreted as the participants agreeing that the inferred feelings are reasonable. Therefore, we can conclude that the system is able to leverage its human feeling inference module to display its empathy—that is, to show its ability to understand human's emotional feelings given what recently happened to them.

4.2 Desire Inference Evaluation

The participants and procedure are similar to the previous experiment but we only show the top-10 nodes to the participants for rating in a Five-Point Likert scale.

The overall average score for the inferred human's desire is 3.76 while the standard deviation is 1.4. Compared with the score of the previously inferred human's feelings, the overall average score of inferred desire is much lower and has greater variation. This can be interpreted as the result of individual's difference. While the personal variance exists, the inferred desires are generally reasonable since the average score for each inferred desire is greater than average, as shown in Fig. 4(b). While these results may vary according to the individual preferences of each person, we believe this effect may be mitigated by learning the individual preference of a user through continuous, long-term interaction.



(a) (b)

Fig. 4. (a) The rating statistics about the inferred feelings; (b) The rating statistics about the inferred desires.

5 Conclusions

We proposed an interactive human-robot trust promotion system, which endows a robot companion with the ability to correctly understand human's feelings and desires from their self-disclosure, showing empathy, goodwill and awareness of personal preferences. Our experimental results on human-robot interaction show that the proposed system outperform lesser system configurations and can promote the human's trust toward the robot companion.

References

1. Baier, A.: Trust and antitrust. Ethics **96**(2), 231–260 (1986)
2. Paivio, S.C., Laurent, C.: Empathy and emotion regulation: reprocessing memories of childhood abuse. J. Clin. Psychol. **57**, 213–226 (2001)
3. Cramer, H.S.M., Goddijn, J., Wielinga, B.J., Evers, V.: Effects of (in)accurate empathy and situational valence on attitudes towards robots. In: Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction (HRI), pp. 141–142 (2010)
4. Feng, J., Lazar, J., Preece, J.: Empathy and online interpersonal trust: a fragile relationship. Behav. Inf. Technol. **23**(2), 97–106 (2004)
5. Jung, H., Lee, S., Park, S., Lee, I., Ahn, C., Kim, J.: Deep temporal appearance-geometry network for facial expression recognition (2015)
6. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., van Knippenberg, A.: Presentation and validation of the Radboud Faces Database. Cogn. Emot. **24**(8), 1377–1388 (2010)
7. Lee, J.J., et al.: Computationally modeling interpersonal trust. Front. Psychol. **4**, 893 (2013)
8. Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., Paiva, A.: The influence of empathy in human-robot relations. Int. J. Hum Comput Stud. **71**(3), 250–260 (2013)
9. Lewicki, R., Wiethoff, C.: Trust, trust development, and trust repair. In: The Handbook of Conflict Resolution: Theory and Practice, pp. 86–107, January 2000

10. Lewicki, R.J., Tomlinson, E.C., Gillespie, N.: Models of interpersonal trust development: theoretical approaches, empirical evidence, and future directions. J. Manag. **32**(6), 991–1022 (2006)
11. Lewis, J.D., Weigert, A.: Trust as a social reality. Soc. Forces **63**(4), 967–985 (1985)
12. Lewis, M., Sycara, K., Walker, P.: The role of trust in human-robot interaction. In: Abbass, H.A., Scholz, J., Reid, D.J. (eds.) Foundations of Trusted Autonomy. SSDC, vol. 117, pp. 135–159. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-64816-3_8
13. Martelaro, N., et al.: Tell me more: designing HRI to encourage more trust, disclosure, and companionship. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction (2016)
14. Mota, R.C.R., et al.: Playing the 'trust game' with robots: social strategies and experiences. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (2016)
15. Nienaber, A.M., Hofeditz, M., Romeike, P.D.: Vulnerability and trust in leader-follower relationships. Pers. Rev. **44**(4), 567–591 (2015)
16. Oestreicher, L., Eklundh, K.S.: User expectations on human-robot co-operation. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication, pp. 91–96 (2006)
17. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4444–4451 (2017)