SSR 2018

PROCEEDINGS OF THE INTERNATIONAL PHD CONFERENCE ON SAFE AND SOCIAL ROBOTS

29th-30th September

2018, Madrid

SECURE
Safety Enables Cooperation in Uncertain Robotic Environments

European Commission

SOCRATES
Social Cognitive Robotics in The European Society

Foreword

The International PhD Conference on Safe and Social Robotics, SSR 2018, which was held on September 29/30 2018 in Madrid, was an official meeting between researchers from the EU MSCA training networks SECURE, SOCRATES as well as external researchers. The goal was to bring together ESRs to foster discussion on future directions within the field of HRI with a special focus on social and safe robotics. 27 early stage researchers have come together to discuss their latest work and share ideas about future experiments and projects. In addition, 16 experienced researchers and PIs have joined in, as well as two invited speakers, to provide experienced views and evaluate the work of the fellows. The conference was organised by the fellows for the fellows under the guidance of both consortia to also provide training in the area of academic meeting organisation.

Behind the conference organisation, there were two driving forces, the SECURE and SOCRATES projects. SECURE is a Marie Skłodowska-Curie Action funded by the European Commission under grant agreement No 642667 with the goal of training researchers specializing in safe human-robot interaction. Fellows are taught to tackle complex interaction environments that occur between humans and robots with the goal of providing new and innovative solutions over a wide range of scenarios and robot platforms.

SOCRATES is a Marie Skłodowska-Curie Action funded by the European Commission under grant agreement No 721619. with the common theme of interaction quality. Researchers are trained to design user-centered systems with a key focus on the topics of emotion, intention, adaptivity, design, and acceptance. The project also maintains a heavy focus on intersectoral collaboration between academia, caregivers, business, and robot manufacturers to ensure that designed systems fit the needs of a developing society.

SSR 2018 has been planned to bring these two research projects closer together as well as disseminate the work together to a wider audience. The conference itself was designed as a training exercise for the fellows from both projects. Papers submitted by fellows were peer reviewed by other fellows and external researchers in order to procure high quality feedback and train scientific writing. As this was a training exercise all papers from fellows were accepted assuming they met a sufficient quality. External researchers were invited to submit as well and the papers were peer reviewed and accepted as good contributions that would help further the discussion at the conference. For this reason, the proceedings are divided into two separate sections containing the 26 papers from fellows in their respective project sections and the 2 accepted external research papers in an own section, as they were subject to a standard review process with four reviewers for each paper.

Presentations at the conference took the form of poster sessions with spotlight talks for the project fellows, and longer oral presentations for accepted external researchers with the goal of maintaining an informal discussion friendly environment. We would like to thank all of the fellows, supervisors, and reviewers who have contributed to make this conference possible. We hope that through this conference all people who have attended left with a broadened view on research in this important field of research and many new connections and ideas that will help to accelerate research in the field of safe and productive HRI.

~SSR 2018 Organizers

Table Of Contents

SOCRATES Submissions

Organization

External Researchers

Transparency for Social Robots

Tanja Heuer[1] and Ina Schiering[1] and Reinhardt Gerndt[1]

***Abstract*— This paper investigates user acceptance and privacy concerns of social robots. Users want a transparent view about processing of personal information. Additionally, they want to be able to intervene. It needs to be possible, to modify default settings. To make users aware of potential risks and concerns it is necessary to involve users during the whole development process and a possible solution for transparency and intervenability may be a privacy dashboard for robots. This privacy enhancing technology provides insight into data processing and sensor use. Additionally, it is necessary to involve users during the development process to sharpen their awareness regarding this issues.**

I. INTRODUCTION

Natural human-machine interaction and social robotics are an emerging field. The first social robots as e.g. the Zenbo[1] already entered the smart home. They are able to control other smart devices at home, tell about the weather, news, appointments, support music streaming and send notifications to family members in case of emergency. To provide this wide range of functionalities, typically robots are employed with a wide range of sensors as cameras and microphones, are using supporting cloud services and connected social media platforms. Hence, social robots collect, process and transfer a huge amount of personal information. Because of the natural interaction with the social robot, which is perceived as a companion by users, this data transfer and processing is not transparent [1]. Also users typically have not the possibility to intervene or do not know how.

According to the Charter of Fundamental Rights of the EU, (Art. 7,8) "everyone has the right to respect for his or her private and family life, home and communications" and "everyone has the right to the protection of personal data concerning him or her". At the moment, these rights of users are not respected by most social robots. The case of Amazon Echo earlier this year gives an example where personal information was sent to someone else without (official) permission [2]. In addition, the General Data Protection Regulation (GDPR) of the EU (2016/679) [3] strengthens these rights in Europe and demand *data protection by design and default*.

In the context of a survey investigating acceptance of social robots and associated privacy concerns, an important aspect are user attitudes towards transparency and intervenability. These two requirements are part of the privacy protection goals [4], which are a common to model privacy requirements. Privacy protection goals are based on the security related goals *confidentiality, integrity, availability* and are augmented by the privacy related goals *transparency, intervenability and unlinkability*. Based on the results of the study, we consider to involve different already existing privacy tools and technologies into the development process of social robots like the privacy protection goals, the seven types of privacy and privacy dashboards to allow transparency.

[1]Ostfalia University of Applied Sciences, Faculty of Computer Science, Wolfenbuettel, Germany, {ta.heuer,i.schiering,r.gerndt}@ostfalia.de

[1]https://zenbo.asus.com/

II. RELATED WORK

A common technology to visualize important information are (privacy) dashboards, which are gladly used by different software applications. This dashboards allow users having an insight view and control about the processing of personal data. They ensure transparency and therefore are an important methodology [5]. An important prototype to investigate usability of privacy dashboards is Data Track [6], visualizing also implications from connected cloud services. With a focus on usability engineering, Raschke et al. [7] presented the idea of a GDPR compliant privacy dashboard. A privacy dashboard for FirefoxOS was proposed by Piekarska et al. [8]. Within a user study, it was investigated how participants make use of the privacy dashboard and what priorities they have. In this context also the Firefox add-on Lightbeam[2], which reveals relations between third party sites on the web is important to note. Additionally, Xu et al. [9] created a smartphone app which summarizes the use of sensors by different applications. The Google Dashboard was investigated [10] with the focus on user acceptance.

Privacy dashboards for smart home applications and smart buildings were developed, to guarantee a user-controlled access [11], [13]. Figure 1(left) shows an example for a smart meter context. In contrast to approaches as data track which try to visualize relations and implications by using a network structure, these privacy dashboards are merely list based. Bier et al. [12] investigate in a user study the interface PrivacyInsight (see Figure1(right)) which is structured similar to smartphone apps compared to a network based and a list based approach. Concepts for ex post transparency including privacy dashboards were furthermore investigated in a broad survey by Murmann et al. [14].

III. METHODOLOGY

In a survey conducted in 2018 during two events, the RoboCup 2018 in Montreal *(group 1)* and in contrast a music festival in Germany *(group 2)*, volunteers were asked about their priorities concerning features, usage and privacy concerns in the context of social robots. A thorough investigation of this survey is beyond the scope of this paper as

[2]https://addons.mozilla.org/en-US/firefox/addon/lightbeam/

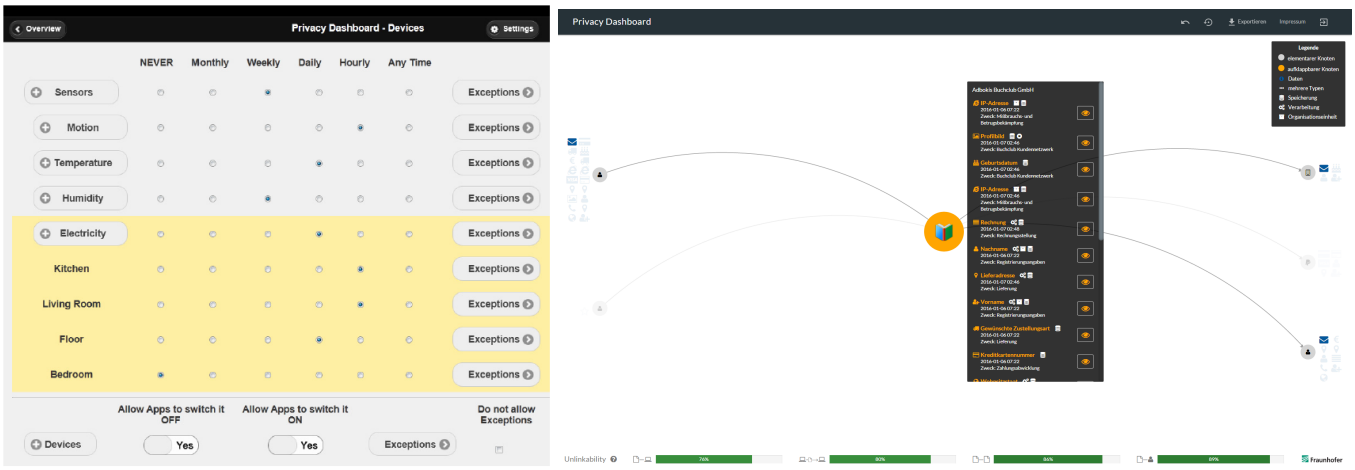


Fig. 1: **(left)** Privacy dashboard of a smart metering system [11], **(right)** PrivacyInsight user interface [12]

ongoing work. The central idea of the two focus groups that were investigated is to query young adults and to compare participants with a strong background in robotics as in *group 1* with a young adults with a standard technical background and no specific experience in robotics as in *group 2*.

The questionnaire used in the study was divided into three sections, the participants were asked about potential functionalities of a robot, interest in using social robots, potential privacy concerns and in the last part demographical information and technical background knowledge. Typically, participants answered the questionnaire in approximately 15 minutes.

In this short paper, the focus is on the investigation of one specific aspect of the survey, namely opinions of potential users towards transparency and intervenability of social robots. Participants were asked to choose their level of disagreement/agreement with the following statements in the range of -2 (strongly disagree) and 2 (strongly agree).

IV. Participants

In total, 73 people participated in the survey, consisting of 35 in *group 1* and 38 in *group 2*. 23 were female, 43 male and 7 decided not to disclose their sex. *Group 1* consists of 7 female, 25 male and 3 non-disclosed participants, *group 2* is divided into 16 female, 18 male and 4 non-disclosed participants. Hence, only 20% of *group 1* is female, whereas almost 40% female participants are in *group 2* . 90% of both groups are aged between 18 and 34.

V. Transparency and Intervenability

Participants of both groups stated that they have substantial privacy concerns in the context of social robots. They expressed a high interest in transparency regarding personal data and the possibility to intervene during the use of robots. As it can be seen in Figure 2 and 3, more than 60% of both groups strongly agreed (rating of +2) with most of the statements. Most of participants of *group 2* do not want to be able to turn on and off certain features of a robot. In comparison, *group 2* showed a general higher agreement on the statements (see Figure 4).

As an unexpected fact, participants of *group 2* showed a higher interest in transparency and intervenability than *group 1*. This effect may have different causes. Whereas *group 1* has a broad experience with robot technologies, *group 2* may be more critical towards the use of robots in their daily life. Robots are not a widespread technology for the general use yet and therefore the attitude towards robots at home is merely skeptical. Nevertheless, the results unambiguously showed, that the participants are aware of the lack of transparency and intervenability and that they want to have access to processed information.

VI. Discussion

In the interpretation of these results, it is important to consider that results of privacy surveys typically differ substantially from actual user behavior. Coopamootoo and Gro [15] investigated in the context of social networks in an empirical study the concepts of privacy and sharing attitude in contrast to the privacy and sharing behavior of users. Additionally, depending on the benefit users feel free to share their personal information [16], [17] and do not think about the risks.

Because of this two facts, it is necessary to involve (potential) users in the whole development process of robots. Because non-technical participants stated, that they are not interested in switch on and off certain features, it is necessary to take them into account. One the one hand, this allows a sensitization of users regarding features, accompanying sensors and it's risks. Users have the ability to shape essential features and the design of the robot. Thereby, they get an insight view into the operating principle of robots. On the other hand, a privacy respecting robot can be developed [18]. To use certain features, special sensors and personal information is needed. Involving the user can lead to a different implementation or to different levels of functionality. As one example, there is a vacuum cleaning robot. Depending on the needs and requirements of the user, there are different types:

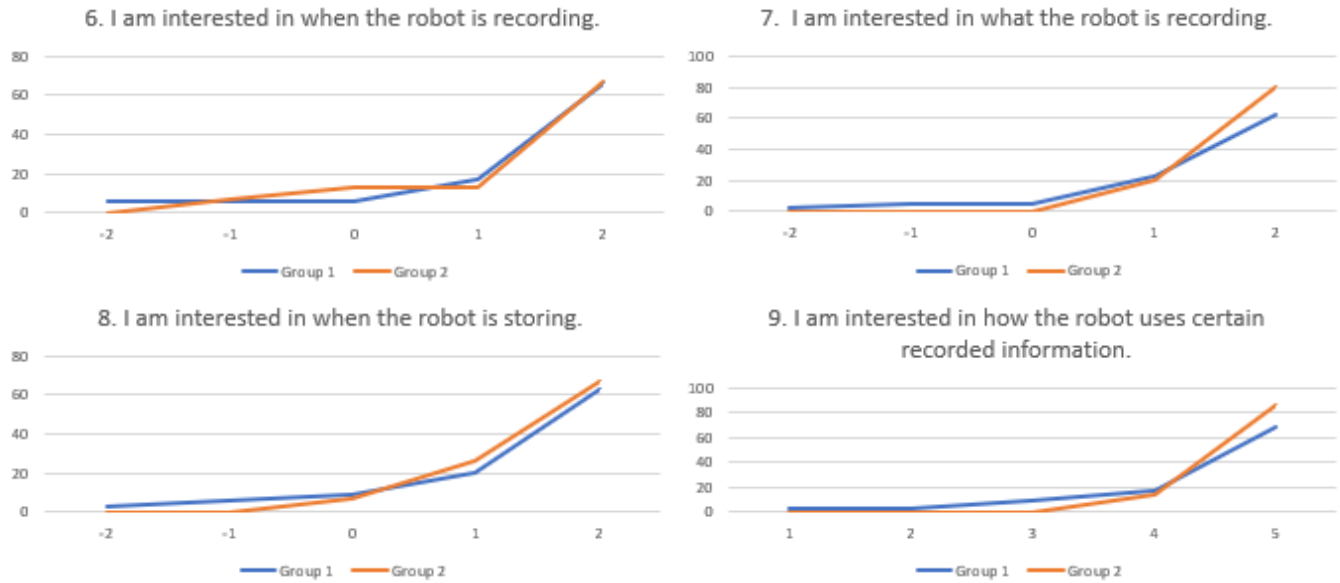


Fig. 2: Transparency for processed personal information

1) *Easy cleaning*: The robot drives around in the room or apartment and when it thinks it is finished, it stops cleaning the floor.
2) *Smart cleaning*: The robot has a laser range scanner and a camera. It creates a map of the room or apartment and drives through the room in an intelligent way, calculated by an algorithm until it has finished.
3) *Supervised cleaning*: The robot creates a map, cleans in an intelligent way. Additionally, the robot can be controlled via other smart home devices and an existing application for the mobile phone informs the owner about cleaning status, where the robot already drove and where it did not get.

Participatory design strategies, which involve the user into the development process, can figure out different needs and gradations regarding the functionalities. Additionally the users see, what is not possible without certain sensors and information and what is. At the moment, most of the existing smart home devices and robots needs to be connected to the internet all the time to allow full functionalities. But it should be possible, to refuse the provision of certain personal information or to disconnect sensors. Instead of complete non-availability, users should be able to decide on their own if they want to have features with only limited and restricted functional capabilities. Robots are able to collect text, videos, images, audio, location, etc. It is important, to get an overview of features and depending data types. Furthermore, it needs to be clear, how the personal data is processed and who has access to it. The purpose of processed data needs to be revealed.

This criteria and it's consequences on the use of the robot can be designed as a privacy dashboard. As shown in Fig. 1(left) for a smart home system, all existing sensors are listed and for every single room/purpose the user can decide on it's own what to allow, when and how often. This needs to be also possible for robots, ideally without a full loss of functionality. Additionally, this should include e.g. restrictions to enter bedrooms, video recording in the bathroom and policies for personal conversations (location-, time-, and situation-dependent). Because users need to be more careful and sensitized about their private life, it is necessary to ask about priorities, preferences and concerns [6], [8], [19]. To allay possible fears of using the dashboard, it needs to be *understandable, easy to use and clearly designed*, that also users without major technical background knowledge are able to use it. They should have co-determination in default privacy settings of the dashboard. This includes predefined privacy settings to protect the users private informations. Because of the complexity of such a dashboard, elements of serious games would be interesting to investigate. This also allows to test the sharing behavior of the user.

All in all, these first conceptual ideas needs to be further investigated. The privacy dashboard for social robots is a step forward to protect life and personal information of the user in their homes in a world full of smart technologies and connected things.

ACKNOWLEDGMENT

This work was supported by the Ministry for Science and Culture of Lower Saxony as part of the program "Gendered Configurations of Humans and Machines (KoMMa.G)".

REFERENCES

[1] M. M. De Graaf, S. B. Allouch, and T. Klamer, "Sharing a life with harvey: Exploring the acceptance of and relationship-building with a social robot," *Computers in human behavior*, vol. 43, pp. 1–14, 2015.
[2] The New York Times, Niraj Chokshi. Is alexa listening? amazon echo sent out recording of couples conversation - may 25, 2018 (visited: 3th september 2018). [Online]. Available: https://www.nytimes.com/2018/05/25/business/amazon-alexa-conversation-shared-echo.html

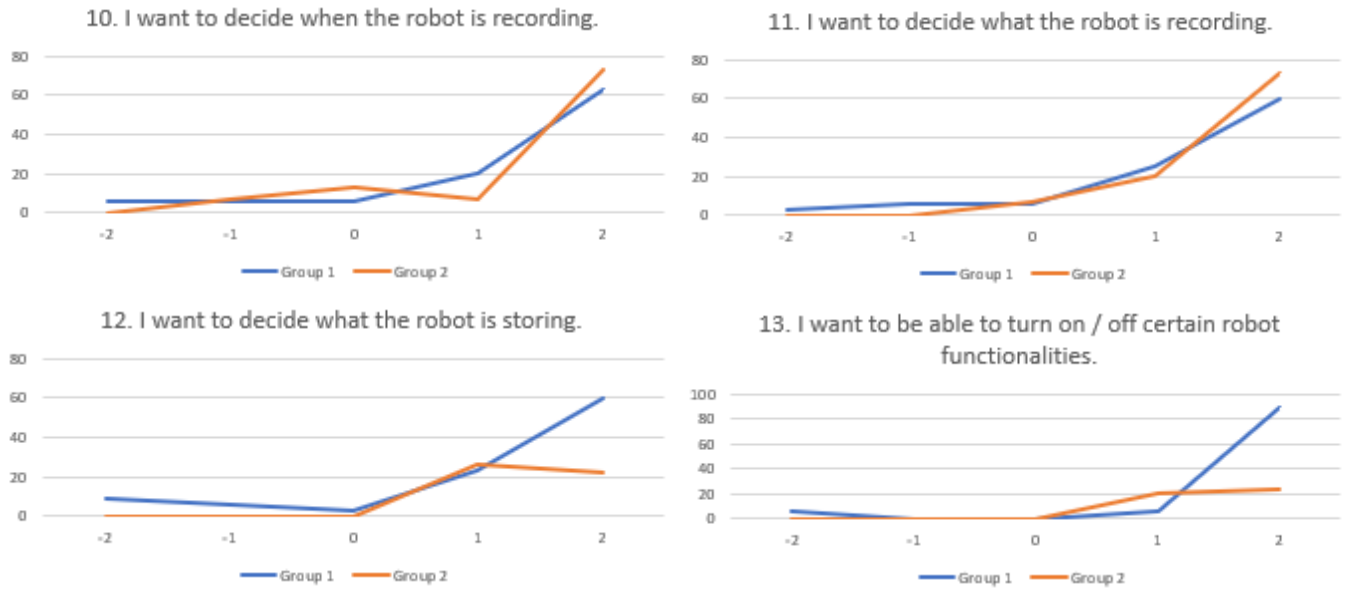


Fig. 3: Intervenability for processed personal information

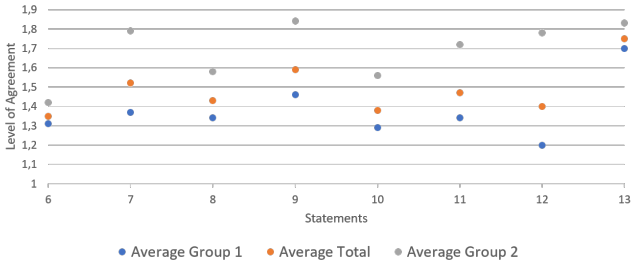


Fig. 4: Average of the Statements

[3] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," pp. 1 – 88.

[4] M. Hansen, M. Jensen, and M. Rost, "Protection goals for privacy engineering," in *Security and Privacy Workshops (SPW), 2015 IEEE*. IEEE, 2015, pp. 159–166.

[5] J. Siljee, "Privacy transparency patterns," *Proceedings of the 20th European Conference on Pattern Languages of Programs - EuroPLoP '15*, pp. 1–11, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2855321.2855374

[6] J. Angulo, S. Fischer-Hübner, T. Pulls, and E. Wästlund, "Usable transparency with the data track: a tool for visualizing data disclosures," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015, pp. 1803–1808.

[7] P. Raschke, K. Axel, O. Drozd, and S. Kirrane, "Designing a GDPR-compliant and Usable Privacy Dashboard," pp. 1–13, 2017.

[8] M. Piekarska, Y. Zhou, D. Strohmeier, and A. Raake, "Because we care: Privacy Dashboard on Firefox OS," *arXiv preprint arXiv:1506.04105*, 2015.

[9] Z. Xu and S. Zhu, "Semadroid: A privacy-aware sensor management framework for smartphones," in *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*. ACM, 2015, pp. 61–72.

[10] C. Zimmermann, J. Cabinakova, and G. Müller, "An Empirical Analysis of Privacy Dashboard Acceptance: The Google Case," *ECIS 2016 Proceedings*, p. 18, 2016. [Online]. Available: http://aisel.aisnet.org/ecis2016_rp Recommended

[11] P. Ebinger, J. L. H. Ramos, P. Kikiras, M. Lischka, and A. Wiesmaier, "Privacy in smart metering ecosystems," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7823 LNCS, pp. 120–131, 2013.

[12] C. Bier, K. Kühne, and J. Beyerer, "Privacyinsight: the next generation privacy dashboard," in *Annual Privacy Forum*. Springer, 2016, pp. 135–152.

[13] A. Leonardi, H. Ziekow, M. Strohbach, and P. Kikiras, "Dealing with Data Quality in Smart Home EnvironmentsLessons Learned from a Smart Grid Pilot," *Journal of Sensor and Actuator Networks*, vol. 5, no. 1, p. 5, 2016. [Online]. Available: http://www.mdpi.com/2224-2708/5/1/5

[14] P. Murmann and S. Fischer-Hübner, "Tools for achieving usable ex post transparency: a survey," *IEEE Access*, 2017.

[15] K. P. Coopamootoo and T. Groß, "Why privacy is all but forgotten," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 4, pp. 97–118, 2017.

[16] A. P. Felt, S. Egelman, and D. Wagner, "I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns," in *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*. ACM, 2012, pp. 33–44.

[17] J. Chen, A. Bauman, and M. Allman-farinelli, "A Study to Determine the Most Popular Lifestyle Smartphone Applications and Willingness of the Public to Share Their Personal Data for Health Research 1," vol. 22, no. 8, pp. 655–665, 2016.

[18] T. Heuer, I. Schiering, and R. Gerndt, "Privacy by design for social robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(published soon)*. IEEE, 2018.

[19] M. Van Kleek, I. Liccardi, R. Binns, J. Zhao, D. J. Weitzner, and N. Shadbolt, "Better the devil you know: Exposing the data sharing practices of smartphone apps," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 5208–5220.

Toward safe separation distance monitoring from RGB-D sensors in human-robot interaction

Petr Švarný, Zdenek Straka, and Matej Hoffmann

***Abstract*— The interaction of humans and robots in less constrained environments gains a lot of attention lately and safety of such interaction is of utmost importance. Two ways of risk assessment are prescribed by recent safety standards: (i) power and force limiting and (ii) speed and separation monitoring. Unlike typical solutions in industry that are restricted to mere safety zone monitoring, we present a framework that realizes separation distance monitoring between a robot and a human operator in a detailed, yet versatile, transparent, and tunable fashion. The separation distance is assessed pair-wise for all keypoints on the robot and the human body and as such can be selectively modified to account for specific conditions. The operation of this framework is illustrated on a Nao humanoid robot interacting with a human partner perceived by a RealSense RGB-D sensor and employing the OpenPose human skeleton estimation algorithm.**

I. Introduction

As robots are leaving safety fences and begin to share their workspace with humans, they need to dynamically adapt to interactions with people and guarantee safety at every moment. There has been a rapid development in this regard in the last decade with the introduction of new safety standards [1], [2] and a fast growing market of so-called "collaborative robots". Haddadin and Croft [3] provide a recent survey of all the aspects of physical Human-Robot Interaction (pHRI). There are two ways of satisfying the safety requirements for pHRI: (i) *Power and Force Limiting* and (ii) *Speed and Separation Monitoring (SSM)* [2]. In the former case, physical contacts with a moving robot are allowed but need to be within human body part specific limits on force, pressure, and energy. This is addressed by interaction control methods for this *post-impact* phase (see the survey [4]). Safe collaborative operation according to SSM demands that a *protective separation distance*, S_p, is maintained between the operator and robot at all times. When the distance decreases below S_p, the robot stops [2]. In industry, S_p is typically safeguarded using light curtains or safety-rated scanners.

In this work, we present a framework that combines state of the art solutions and realizes separation monitoring between a robot and a human operator in a detailed, yet versatile, transparent, and tunable fashion. The separation distance is assessed pair-wise for all keypoints on the robot and the human body and as such can be selectively modified to account for various interaction scenarios. The operation of this framework is illustrated on a Nao humanoid robot interacting in real-time with a human partner who is perceived by a RGB-D sensor.

The authors are with the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague (e-mail: petr.svarny@fel.cvut.cz; zdenek.straka@fel.cvut.cz; matej.hoffmann@fel.cvut.cz).

II. Related work

A functional solution for safe pHRI according to SSM will necessarily involve: (i) sensing of the human operator's as well as robot's positions and speeds, (ii) a suitable representation of the corresponding separation distances and (iii) appropriate responses of the machine.

Tracking the spatial location of the robot's keypoints is relatively easy thanks to forward kinematics and joint encoder values. The perception of human operator's location is more difficult. Zone scanners used in industry report the intrusion of an object into a predefined zone—a solution that is safe but very inflexible and essentially prevents most collaborative activities. Two key technologies have appeared recently that facilitate progress in this area: (i) compact and affordable RGB-D sensors (like Kinect) and (ii) convolutional neural networks for human keypoint extraction from camera images [5], [6]. These technologies together—albeit currently not safety-rated—make it possible to perceive the positions of individual body parts of any operator in the collaborative workspace in real time.

Once the robot and human positions are obtained, their relative distances need to be evaluated (see Flacco et al. [7] for a comparison of approaches). The robot and human body parts can be represented as spheres [8], capsules [9] or meshes [10] and they can be different for the robot and the human [11].

The approach is often "robot-centered" in the sense that the collision primitives are centered on the robot body and possibly dynamically shaped based on the current robot velocity [12], [13]. Even the biologically inspired approach to "peripersonal space" representation [10], [11], [14], [15] is robot-centered: the safety margin is generated by a distributed array of receptive fields surrounding the electronic skin of the iCub humanoid robot. Finally, there is a large body of work dealing with motion planning and control in dynamic environments. Most recent and most related to our approach are [9], [11], [16].

We propose a separation distance representation that treats robot and human keypoints equally and uses Euclidean distance in Cartesian space to evaluate all safety thresholds. In accordance with [2], velocities, reaction times, and uncertainties can all flow into the desired thresholds. Unique to our approach, the representation is maximally transparent with the easy incorporation of important features. In opposition to

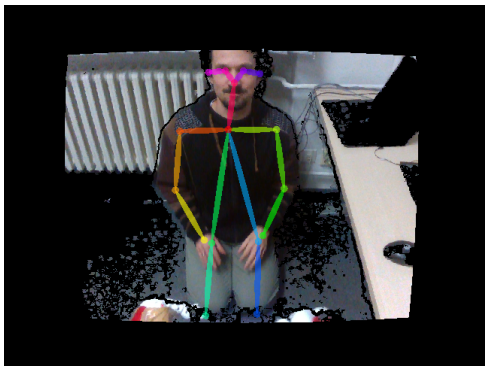


Fig. 1: Color aligned with depth stream with the rendered human keypoints from OpenPose.

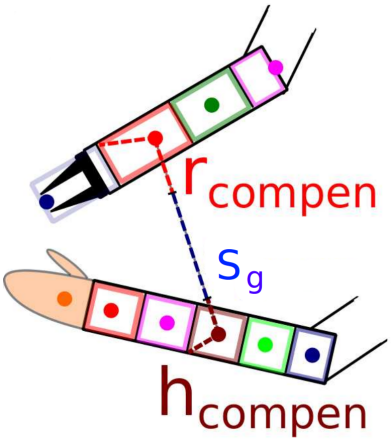


Fig. 2: Separation distance calculation between robot and human keypoints.

machine learning heavy approaches, our framework allows simple risk assessment and it is straightforwardly transferable between robotic platforms.

III. Materials and Methods

ft Human keypoints are perceived in the environment while robot keypoints are extracted from the model and current joint values. The relative distances are assessed and fed into the robot controller to generate appropriate responses.

A. Human keypoint 3D estimation

A server collects two streams from a RealSense SR300 camera: a color image aligned to the depth image (CAD) and a point cloud stream (PCS), also depth image aligned. We use Intel RealSense SDK with PyRealSense. The CAD image is sent to OpenPose [5] by PyOpenPose to estimate the human keypoints (see Fig. 1). The pixel coordinates of keypoints are paired with those from PCS. All our image operations use OpenCV3 [17].

The keypoints are transformed into the Nao's frame of reference by affine transforms. The rotation and translation for them are gained from a pre-experiment calibration.

B. Nao robot keypoints

A Nao humanoid robot (V3+) with keypoints on the left end-effector, forearm, and elbow was used to demonstrate the framework. We used forward kinematics with current joint encoder values as input to get the 3D position of these keypoints.

C. Separation distance representation

The *protective separation distance* S_p [2] needs to be maintained between any human and robot part such that the human will never collide with a moving machine. Its value will be determined based on reaction times etc. as in [2]. We extend S_p as a baseline with additional terms.

First, we want to account for "modulation" on the part of the human to grant larger distance from specific body parts (e.g. head) and on the part of the robot when carrying a sharp tool. Adding these distance offsets $\mathbf{r}_s$, $\mathbf{h}_s$ gives rise to a *guaranteed minimal separation distance* S_g.

Second, as only distances between keypoints will be evaluated, but separation distance between any body parts needs to be maintained, we add compensation coefficients, $\mathbf{h}_{compen}$ and $\mathbf{r}_{compen}$ (see Section III-D below). This is the *keypoint separation distance* S_d—the quantity that will be monitored between any keypoint pairs.

Therefore S_d is in the form of a matrix of separation distances between two given keypoints i, j ($S_d^{i,j}$) (see Section IV).

$$\begin{aligned} S_g^{ij} &= h_s^i + S_p + r_s^j \\ S_d^{ij} &= h_{compen}^i + S_g^{ij} + r_{compen}^j \end{aligned}$$

D. Keypoint compensation coefficients

Using a discrete distribution of keypoints allows fast calculation, but does not take the full volume of the bodies into account. The compensation coefficients r_{compen} and h_{compen} allow us to guarantee S_g even with a discrete keypoint distribution.

These coefficients are calculated in two steps. First, every part of the body is assigned to its nearest keypoint. Then the maximal distance over all of its assigned volume is selected as the compensation coefficient for the keypoint (see Fig. 2)—thereby always guaranteeing S_g.

E. Robot control

We used PyNaoqi to control the Nao. The Nao was moving his hands back and forth periodically in front of his chest. The robot stopped when an $S_d^{i,j}$ threshold was exceeded. The robot resumed operation upon "obstruction" removal. In addition, we defined a reduced speed distance: when $S_{d(reduced)}^{i,j}$ for any keypoint pair was exceeded, the robot reduced its speed to half.

F. HRI setup

The Nao robot was sitting in a fixed position with respect to the camera that captured the robot's workspace (see Fig. 1). Our setup is safe because of the Nao robot's size and power. In a real setting with a potentially dangerous

machine and safety-rated modes, S_p would be determined from [2]. In our case, the threshold was chosen arbitrarily.

The compensation values accounting for keypoint density (Section III-D) were determined by measuring the distances between keypoints (Table **??** and I). Only upper body keypoints were taken into consideration for the human operator. We call the set of keypoints of the nose, neck, eyes, and ears as the human head. In both, human and robot cases, the compensation coefficients were symmetrical and thus we list keypoint pairs only once.

End effector	**Wrist**	**Elbow**
0.06m	0.05m	0.06m

Nose	**Neck**	**Eye**	**Ear**	**Shoulder**
0.10m	0.25m	0.10m	0.10m	0.15m
Elbow	**Wrist**	**Hip**	**Knee**	**Ankle**
0.15m	0.15m	0.00m	0.00m	0.00m

TABLE I: Human compensation values $\mathbf{h}_{compen}$

IV. Results

We conducted three scenarios: (A) basic separation matrix, (B) specific separation values for the head of the human, (C) emulation of a sharp tool in the robot's hand.[1] Distances between all human and robot keypoints were evaluated simultaneously online. However, for clarity, we present only the interaction of the robot end-effector with two human keypoints (the right wrist and the nose) in the plots below. The baseline protective separation distance was set to $S_p = 0.05m$ and the reduced speed regime $S_{p(reduced)} = 0.20m$.

A. Basic scenario

In the basic experiment, we monitored the distance between the human wrist and robot end-effector – see Fig. 3. The relevant separation matrices are in the Table II.

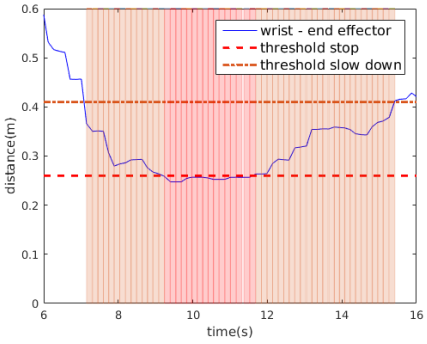


Fig. 3: Basic Scenario: presented are Nao end-effector and human wrist keypoint distances and thresholds (S_d and $S_{d(reduced)}$).

Crossing the threshold into the warning regime is detected by the robot around $t = 7s$ as shown by the orange shaded area. The robot enters reduced speed mode at this point.

[1]The video is available at `https://youtu.be/3DZyuuQlqPo`.

$S_{d(reduced)}$

Robot \ Human	**Nose**	**Wrist**
End effector	0.36m	0.41m

S_d

Robot \ Human	**Nose**	**Wrist**
End effector	0.21m	0.26m

TABLE II: Basic scenario: Separation matrix for keypoint pairs from Fig. 3.

Similarly, the next crossing is marked by red shading and the robot stops. The removal of the wrist from the safety zones resumes the robot's operations.

B. Head and body discrimination

The $\mathbf{h}_s$ for the head keypoints was enlarged by $0.15m$. This lead to the robot's higher sensitivity to situations when the human operator approached the robot with his head, as shown in Fig. 4.

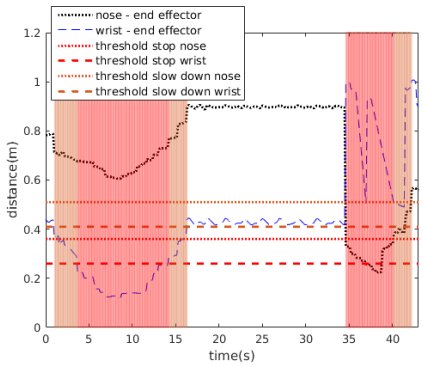


Fig. 4: Head and body discrimination: A higher separation threshold for the human head region.

$S_{d(reduced)}$

Robot \ Human	**Nose**	**Wrist**
End effector	***0.51***m	0.41m

S_d

Robot \ Human	**Nose**	**Wrist**
End effector	***0.36***m	0.26m

TABLE III: Head and body discrimination: Separation matrix for keypoint pairs from Fig. 4. Emphasis is on values altered w.r.t. to first scenario.

In the first half of the experiment, we see the reaction of the robot to the wrist keypoint. Later, we see that the robot reacts to the nose keypoint at a greater distance than to the wrist. Notice the different reactions of the robot (shown by the different shading) for similar distances of the two keypoints.

C. Dangerous tool usage

The left arm end-effector $\mathbf{r}_s$ was increased by $0.1m$ to simulate a possibly dangerous tool (see Fig. 5). The stopping and warning thresholds are now $0.1m$ farther away from the robot end-effector. This increase is added to the original

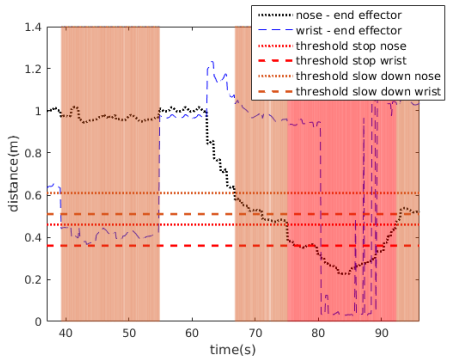


Fig. 5: Dangerous tool usage: Increased safety margin around robot end-effector.

functionality from the previous scenario, thus the robot reacts with greater sensitivity to the approach of the operator's nose keypoint as opposed to the proximity of the operator's wrist keypoint.

	$S_{d(reduced)}$	
Robot \ Human	**Nose**	**Wrist**
End effector	***0.61***m	***0.51***m

	S_d	
Robot \ Human	**Nose**	**Wrist**
End effector	***0.46***m	***0.36***m

TABLE IV: Dangerous tool usage: Separation matrix for keypoint pairs from Fig. 5.

V. Discussion and Conclusion

We presented a framework that realizes separation monitoring between a robot and a human operator. Distances are simply represented in Cartesian space in Euclidean norm and human and robot keypoints are treated equally. The separation distance is assessed pair-wise for all keypoints on the robot and human body and as such can be selectively modified. Velocity is not part of our representation but velocities can be converted into distance increments relying on measured quantities or worst-case constants per [2]. The framework was illustrated on a Nao humanoid robot interacting with an operator monitored by an RGB-D sensor.

RGB-D sensors are currently not safety-rated. However, their reliability can be improved [18], [19]. OpenPose itself also provides confidence values with every keypoint estimated. These enhancements and the transfer to a real-life industrial scenario with performance evaluation constitute our future work.

Nevertheless, safety-rated devices similar to those for zone monitoring that would provide 3D object coordinates and possibly human keypoints are needed. Other alternatives exist [20] next to RGB-D sensors. The availability of such technology would expand the possibilities of human-robot collaboration in the SSM regime.

Acknowledgment

Matej Hoffmann was supported by the Czech Science Foundation under Project GA17-15697Y. Petr Švarný and Zdenek Straka were supported by the Czech Technical University in Prague, grant No. SGS18/138/OHK3/2T/13.

References

[1] "ISO 10218 Robots and robotic devices – Safety requirements for industrial robots," International Organization for Standardization, Geneva, CH, Standard, 2011.

[2] "ISO/TS 15066 Robots and robotic devices – Collaborative robots," International Organization for Standardization, Geneva, CH, Standard, 2016.

[3] S. Haddadin and E. Croft, "Physical human-robot interaction," in *Springer Handbook of Robotics*, 2nd ed., B. Siciliano and O. Khatib, Eds. Springer, 2016, pp. 1835–1874.

[4] S. Haddadin, A. De Luca, and A. Albu-Schäffer, "Robot collisions: A survey on detection, isolation, and identification," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1292–1312, 2017.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, vol. 1, no. 2, 2017, p. 7.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.

[7] F. Flacco, T. Kroeger, A. De Luca, and O. Khatib, "A depth space approach for evaluating distance to objects," *Journal of Intelligent & Robotic Systems*, vol. 80, p. 7, 2015.

[8] F. Flacco, T. Kröger, A. De Luca, and O. Khatib, "A depth space approach to human-robot collision avoidance," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 338–345.

[9] C. Liu and M. Tomizuka, "Algorithmic safety measures for intelligent industrial co-robots," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3095–3102.

[10] M. P. Polverini, A. M. Zanchettin, and P. Rocco, "A computationally efficient safety assessment for collaborative robotics applications," *Robotics and Computer-Integrated Manufacturing*, vol. 46, pp. 25–37, 2017.

[11] A. M. Zanchettin, N. M. Ceriani, P. Rocco, H. Ding, and B. Matthias, "Safety in human-robot collaborative manufacturing environments: Metrics and control," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 882–893, 2016.

[12] B. Lacevic and P. Rocco, "Kinetostatic danger field-a novel safety assessment for human-robot interaction," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 2169–2174.

[13] V. Magnanimo, S. Walther, L. Tecchia, C. Natale, and T. Guhl, "Safeguarding a mobile manipulator using dynamic safety fields," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 2972–2977.

[14] A. Roncone, M. Hoffmann, U. Pattacini, L. Fadiga, and G. Metta, "Peripersonal space and margin of safety around the body: learning tactile-visual associations in a humanoid robot with artificial skin," *PLoS ONE*, vol. 11, no. 10, p. e0163713, 2016.

[15] D. H. P. Nguyen, M. Hoffmann, A. Roncone, U. Pattacini, and G. Metta, "Compact real-time avoidance on a humanoid robot for human-robot interaction," in *Human-Robot Interaction, HRI 18: 2018 ACM/IEEE International Conference on*. IEEE, March 58, 2018, p. 9.

[16] J. A. Marvel, "Performance metrics of speed and separation monitoring in shared workspaces," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 2, pp. 405–414, 2013.

[17] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

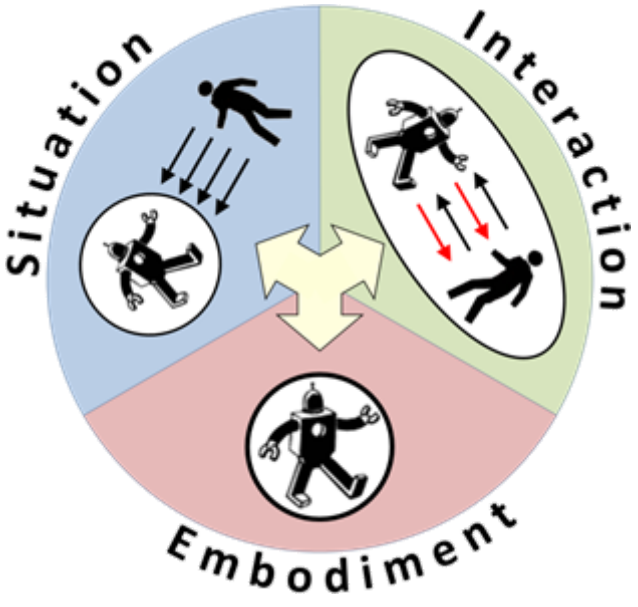
[18] F. Fabrizio and A. De Luca, "Real-time computation of distance to dynamic obstacles with multiple depth sensors," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 56–63, 2017.

[19] M. Ragaglia, A. M. Zanchettin, and P. Rocco, "Trajectory generation algorithm for safe human-robot collaboration based on multiple depth sensor measurements," *Mechatronics*, 2018.

[20] S. Savazzi, V. Rampa, F. Vicentini, and M. Giussani, "Device-free human sensing and localization in collaborative human–robot workspaces: A case study," *IEEE Sensors Journal*, vol. 16, no. 5, pp. 1253–1264, 2016.

SECURE

SECURE is a new Marie Skłodowska-Curie Action funded by the European Commission. It's aim is to train roboticists and research fellows on the cognitive and interaction level of robot safety. These fellows should then be able to cope with the new challenges for safety that come with the increased complexity in human work and living spaces. They also need to be familiar with safety concepts and solutions for a multitude of robotic platforms. Therefore, the SECURE network aims to train fellows on innovative scientific and technological requirements for safe human-robot interaction and will employ several of the currently best robot platforms in Europe.



The fellows are trained at six partner institutions in Europe and are supported by another five associated partners, ranging from large-scale international industrial partners to small enterprises, thus providing an optimal training environment for young researchers.

Measuring Habituation during Human-Robot Interaction

Grigorios D. Skaltsas

***Abstract*— Using measurements of physiological signals (eye-tracking, galvanic skin response, heart rate) and questionnaires during a series of human-robot interaction experiments, user stress metrics and habituation patterns are analyzed. The initial experimental results indicate that there seems to be a varying relation between human stress and robot speed as the human gets acquainted with the robot which seems to be also affected by the human perception of the task's success.**

I. Introduction

SECURE project is related to the security during the interaction between a human and a robot. Furthermore, the advertised position in the University of Hertfordshire was related to social robotics. Further to basic industrial safety standards, the psychological perception of safety seems to be a topic that many researchers are investigating from a variety of different approaches. However, proxemics and physiological sensing studies seem to dominate the psychological robot safety research domain [1].

Reading and evaluating the human's adaptation through biological signals could be the base for a performance optimization system targeting the minimization of human stress during the interaction [2-4]. It has been shown that safety is still perceived as low when the robot's trajectory planning and execution seems to be only avoiding collision [1]. Therefore, safety design has to include psychological factors which could be the adjustment of various parameters of the robot's motion, such as the speed profile in terms of acceleration, deceleration, maximum and minimum speed, proximity to the human or other objects, and also adjustments of behaviour based on robot's appearance [1].

Cultural and personal preferences identify which everyday human interaction characteristics are also important to implement in human-robot interaction (HRI). Methods employed, are commonly questionnaires, physiological metrics, and behavioural metrics [5].

*Research supported by SECURE project https://secure-robots.eu. funded by the EU Horizon 2020 research and innovation programme under grant agreement No 642667 (SECURE)

The notion of studying the user's habituation after a number of trials and identifying personal or generic trends in short or long term seems to not have been performed even in recent publications. The reason the habituation is studied in this project is so that more systems that adapt to the human as the human adapts to them. Then the robots can gradually increase their performance without being stressful to the humans whereas if the human has to adapt to an unknown system might result in low acceptance of the system or a rejection altogether especially if there is no prior knowledge about it.

For the research purposes, one study was carried out and two are yet to be completed where primarily galvanic skin response (GSR), heart rate (HR) and eye-tracking (ET) are analysed during sessions where the human is mostly passive whilst the robot is active.

II. Related Work

The following section shows some of the most related work sorted by data acquisition method. Some of the related studies combine more than one method. However, they are presented in the correspondent sections bellow based on the importance of the method used in the study and the critical points that highlight the usefulness of the method.

A. Galvanic Skin Response

GSR consists in reading the changes in human skin's conductivity when the sweat micro-glands respond to stressful situations. Dehais [6] used a motion planner based on Sisbot [7] for planning. A robot approached the human and handed an item. A training trial was performed with the users before the actual experiment, therefore the measured signals had already some adaptation effect. Kulic and Kroft used a predefined algorithm on a robotic manipulator that was fixed and the user was also sitting on a chair at a safe distance without being required to intervene to the task [8]. In their study it was demonstrated that by using a fuzzy interference controller, the user stress levels could be minimized in subsequent trials.

B. Eye-Tracking

ET offers physiological and subjective evaluation by correlating ET data with questionnaire responses

[6]. Overall in Human-Computer Interaction (HCI) as well as HRI, eye-tracking has been used to provide vision analytics [9] and offer an additional modality [10].

C. Heart Rate

Heart-rate in HRI has been used as a primary physiological response measure [11]. In another study two systems, one wearable and one laboratory high precision sensor were used to evaluate the user's response to some pictures shown and by immediately then filling a brief questionnaire [12]. Although in both experiments the HR measurements yielded measurable consistent results, it was not discussed how any possible habituation effects might in long term affect the measurements.

D. Questionnaires

Questionnaires have been widely used in HRI [13-15] as a method to collect user's feedback.

Joosse et al developed the BEHAVE II questionnaire that separates the responses based on attitude and behavior [16]. Morales et al tried to evaluate pleasantness of motion planning of an autonomous wheelchair via questionnaires [17].

[6] has combined physiological responses with questionnaires in an attempt to combine each other's results so that physiological responses will match the user's post experiment evaluation.

RoSAS questionnaire demonstrated that robot's appearance impacts its social evaluation [18].

Ragot et al performed a study where the participants had 15 seconds after every scene projected on a screen to self-assess in a 2 dimensional scale their arousal and valence [12]. The difference between this and previous studies is that the questionnaire was completed in small portions using simple numeric scales after each event so that the users could reflect more easily on how they felt and provide the "ground truth" tags for the recorded physiological data.

III. Approach

The objectives of the first experiment are to:

- Compare the findings of previous experiments in related studies verifying that the results are similar [6, 19, 20].
- Provide actual data on HRI sessions, where the human is passively participating, both from questionnaires and sensor readings.
- Explore the habituation patterns that might appear, create the proposed statistical model as a correlation between the sensor readings and the replies on the questionnaires.

A. Design of the Experiment

The experiment explored short-term habituation and had participants mostly being students and local residents from the nearby area that can access the university easily. The sample contained 29 participants (Male: 22, (Age: 34.5avg 10.7std) Female: 5 (28avg 4.9std)). Their knowledge on digital equipment was marked high on the average. The participants were split in four groups. All groups had to experience four distinct sessions.

For the habituation effects' study, all the sessions run sequentially with a small pause in between for a few minutes until the questionnaires are completed. The users had to evaluate their experience with the robot, combining it with the overall effectiveness of the task, whilst their physiological responses were recorded.

After the participants entered the lab, they read the participant information sheet and signed the consent form, the sensors were then fitted, calibrated and tested on each user on an individual basis at the beginning of the experiment. In order to obtain a base line for the GSR, a small resting period was introduced. The ET sensor had to be calibrated on an individual basis. In order to keep the base GSR updated, small pauses of a minute were introduced between the completion of the questionnaire and the next session.

In each session, the robot approached them from a distance of approximately 5 meters after coming out of an initial location where it would not be visible to the user. The robot during each session acted in a fully autonomous way, acting totally independent of any of the user's sensor measured feedback.

The structure of the sessions was based on the combination of two conditions. The first condition was the robot's speed and hence the perceived risk by the human of the robot crashing onto a wall or on the human upon approach. The speed choices were based on the robot's capabilities. The second was the delivery of an item that was on the robot but not securely attached to it, hence an extra risk perceived by the human as task failure, such as dropping the item at some point or seeing the item shaking during the transportation. For this experiment, the item chosen was a half full semitransparent water bottle. The user could see the shake of the water during its transportation by the robot. The combinations of these conditions create the following session scenarios:

- Fast speed carrying the bottle

- Slow speed carrying the bottle
- Fast speed without carrying the bottle
- Slow speed without carrying the bottle

To avoid bias, users were grouped as described earlier and set to participate in possible combinations of sequences of session scenarios as shown on table 1. The first two sessions for each group consist of the robot varying its speed alone. The last two sessions add the bottle carrying task combined with the variations of the speed. Adding the extra risk at the last two sessions of the experiment, compensates for the user's loss of interest and changing one condition each time helps compare the changes in the habituation pattern of each group in a controlled manner. The table, for clarity, is coded as follows:

- Condition: Fast (F), Slow (S)
- Carrying a bottle, Yes (B), No ( )

TABLE I. TABLE OF USER GROUPS

Group	session			
	1	**2**	**3**	**4**
1	F	S	F+B	S+B
2	F	S	S+B	F+B
3	S	F	F+B	S+B
4	S	F	S+B	F+B

Cumulative Robot's Speed and Task Pattern over each session.

The robot did not communicate to the user its movement intentions in any session. The users experienced the robot planning its movement spontaneously from by their visual perception of the robot's location and the engine's noise.

B. Platform and Sensory Choices

University of Hertfordshire's custom platform "sunflower", a service robot comprising of a mobile base, a waist link, and a tray. It is a medium sized robot built on a Pioneer 3DX base using two wheels on each side for its navigation. It has a static head, with non-functional large round 'eyes', mounted on a dynamixel based chain neck with 4DOF [21]. GSR and HR [22] and eye-tracker sensors were used for the physiological measurements [23].

C. Robot Trajectories and Speed Choices

The path of the robot (figure 1) was chosen so it would have maximum visual exposure to the user.

Also, it was combined with a maneuver that requires a sharp turn (top right corner) and the potential of a crash upon failure when it was still away from the user. The duration of the slow trajectory is approximately 48 seconds and 20 seconds for the fast one, giving enough time to the user's GSR to rise and drop approximately at the time when an event causing stress occurs. The user's curiosity should be heightened as to why the robot chooses this path to follow as opposed to a direct approach. The user should perceive the robot not as a completely human-like thinking entity but as a system with some way of reasoning that does not necessary act the way a human would. This, sebsequently, was revealed by the discussion with some of the participants and their questionnaire responses. It appears that it had an impact on the assessment of the task's efficiency later on in the questionnaire's section.

The speed is approximately 0.7 Km/hour in the fast mode and 0.47 Km/hour in the slow mode. In both fast and slow trajectories, the robot covers approximately 3.1 meters in the first straight segment and 5.75 meters in the second in which it approaches the participant. The turning lasts 3 seconds and including the stop and start of the robot on the turning spot is approximately 6 seconds. The safety distance is 30 to 50 cm from the participants' feet.

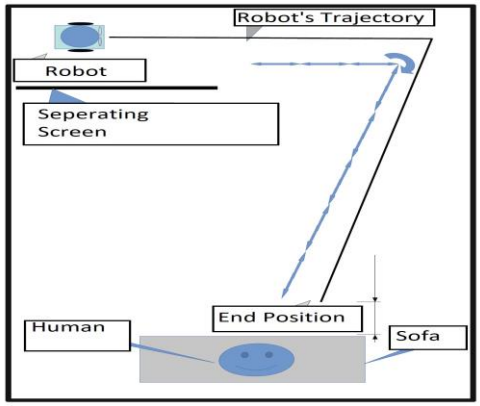


Figure 1 (Top View of Robot's Trajectory during each session)

D. Questionnaires

There are four questionnaires used for this experiment.

Once the first trial ended, the (1) "demographics sheet" asking age, gender, expertise with computers among others, the (2) "behind the wall" asking about the users experience whilst the robot was not visible and one copy of the (3) "main questionnaire" was handed out to the user whilst the sensors were still fitted.

The (3) "main questionnaire" was handed to the user after each trial. Hence it was completed four times for each user. It asked the user to evaluate the robot's performance. It also required the user to indicate on a

schematic showing the robot's trajectory during the trial, the parts where the robot was too fast or slow as well as where it could have failed the task. The (4) "general questionnaire" -which is handed out in the end- asking the user about his/her overall experience, as well as the (4) "demographics sheet" have the purpose to normalise the responses of the user.

IV. Primary Results and Points to be Addressed

Results are under analysis. From a qualitative point of view, there seem to be repeated patterns for most users' physiological responses in relation to specific events. User perception of the task's risks and complexity varies seemingly as the conditions vary in ways that the physiological responses do not always correspond to the questionnaire responses.

Emerging features such as stress signs due to specific event anticipation and their variance are currently being studied. For example, once the user has experienced the robot's trajectory for the first time, how long it takes before the turning point is reached and hence his/her GSR peaks anticipating the potential crash on the wall. Furthermore, how this changes when the speed changes or the task of carrying the bottle.

V. Future Work

There are two more experiments to be carried out. Their aim is to provide results that will clarify some points from the first experiment.

The second experiment is focused on a simpler unique movement with more repetitions in a higher speed. The third experiment will be focused on the user hearing the robot approaching. Study of stress and habituation of events such as low intensity touches of the robot to the seat are also under consideration.

The results of all the experiments' data might help drawing further conclusions on human stress and habituation during HRI and suggest methods of minimizing stress.

References

[1] P. A. Lasota, T. Fong, J. A. Shah, and others, "A survey of methods for safe human-robot interaction," *Foundations and Trends in Robotics*, vol. 5, no. 4, pp. 261-349, 2017.

[2] R. R. Fletcher, K. i. Amemori, M. Goodwin, and A. M. Graybiel, "Wearable wireless sensor platform for studying autonomic activity and social behavior in non-human primates," in *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 4046-4049.

[3] I. Daly *et al.*, "Towards human-computer music interaction: Evaluation of an affectively-driven music generator via galvanic skin response measures," in *Proc. 7th Computer Science and Electronic Engineering Conf. (CEEC)*, 2015, pp. 87-92.

[4] A. Sano and R. W. Picard, "Stress Recognition Using Wearable Sensors and Mobile Phones," in *Proc. Humaine Association Conf. Affective Computing and Intelligent Interaction*, 2013, pp. 671-676.

[5] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting Engagement in HRI: An Exploration of Social and Task-Based Context," in *Proc. Risk and Trust and 2012 Int 2012 Int. Conf. Privacy, Security Confernece Social Computing*, 2012, pp. 421-428.

[6] F. Dehais, E. A. Sisbot, R. Alami, and M. Causse, "Physiological and subjective evaluation of a human-robot object hand-over task," *Applied ergonomics*, vol. 42, no. 6, pp. 785-791, 2011.

[7] E. A. Sisbot, L. F. Marin-Urias, X. Broquère, D. Sidobre, and R. Alami, "Synthesizing Robot Motions Adapted to Human Presence," *International Journal of Social Robotics*, vol. 2, no. 3, p. 329, September 2010.

[8] D. Kulic and E. Croft, "Physiological and subjective responses to articulated robot motion," *Robotica*, vol. 25, no. 1, pp. 13-27, 2007.

[9] K. Kurzhals, M. Hlawatsch, C. Seeger, and D. Weiskopf, "Visual Analytics for Mobile Eye Tracking," vol. PP, no. 99, p. 1, 2016.

[10] P. Kasprowski, K. Harezlak, and M. Niezabitowski, "Eye movement tracking as a new promising modality for human computer interaction," in *Proc. 17th Int. Carpathian Control Conf. (ICCC)*, 2016, pp. 314-318.

[11] B. Kuehnlenz and K. Kuehnlenz, "Reduction of Heart Rate by Robot Trajectory Profiles in Cooperative HRI," in *Proc. ISR 2016: 47st Int. Symp. Robotics*, 2016, pp. 1-6.

[12] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors," presented at the International Conference on Applied Human Factors and Ergonomics, 2017.

[13] A. Chanseau, K. Dautenhahn, K. L. Koay, and M. Salem, "Who is in charge? Sense of control and robot anxiety in Human-Robot Interaction," in *Proc. 25th IEEE Int. Symp. Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 743-748.

[14] K. L. Koay, K. Dautenhahn, S. Woods, and M. L. Walters, "Empirical results from using a comfort level device in human-robot interaction studies," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 194-201.

[15] M. L. Walters, M. A. Oskoei, D. S. Syrdal, and K. Dautenhahn, "A long-term Human-Robot Proxemic study," in *2011 RO-MAN*, 2011, pp. 137-142.

[16] M. Joosse, A. Sardar, M. Lohse, and V. Evers, "BEHAVE-II: The revised set of measures to assess users' attitudinal and behavioral responses to a social robot," *International journal of social robotics*, vol. 5, no. 3, pp. 379-388, 2013.

[17] Y. Morales, A. Watanabe, F. Ferreri, J. Even, K. Shinozawa, and N. Hagita, "Passenger discomfort map for autonomous navigation in a robotic wheelchair," *Robotics and Autonomous Systems*, vol. 103, pp. 13 - 26, 2018.

[18] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The Robotic Social Attributes Scale (RoSAS): Development and Validation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, Vienna, Austria, 2017, pp. 254-262: ACM.

[19] T. Arai, R. Kato, and M. Fujita, "Assessment of operator stress induced by robot collaboration in assembly," *CIRP annals*, vol. 59, no. 1, pp. 5-8, 2010.

[20] D. Kulic and E. Croft, "Anxiety detection during human-robot interaction," presented at the Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on, 2005.

[21] lirec.eu. Available: http://lirec.eu/project

[22] shimmersensing.com. Available: https://www.shimmersensing.com/products/consensys-ecg-development-kits-update

[23] imotions.com. Available: https://imotions.com/asl-eye-tracking-glasses/

Investigating human perception of trust and social cues in robots for safe HRI in home environments

Alessandra Rossi[1] and Kerstin Dautenhahn[1,2] and Kheng Lee Koay[1] and Michael L. Walters[1]

Abstract— **Our aim is to create guidelines that allow humans to trust robots that are able to look after their well-being by adopting human-like behaviours. However, trust can change over time due to different factors, e.g. due to mechanical, programming or functional errors. It is therefore important for a domestic robot to have acceptable interactive behaviour when exhibiting and recovering from an error situation. As a first step, we investigated human users' perceptions of the severity of various categories of potential errors that are likely to be exhibited by a domestic robot. We conducted a questionnaire-based study, where participants rated 20 different scenarios in which a domestic robot made an error according to their severity. We clearly identified scenarios that were rated by participants as having limited consequences ('small' errors) and that were rated as having severe consequences ('big' errors). In order to define acceptable behaviours to recover the human trust, it is necessary to consider that errors can have different degrees of consequences and people's personalities and dispositions of trust may affect differently their perception of the robot. We used an interactive storyboard presenting ten different scenarios in which a robot performed different tasks, either correctly, or with small or big errors, under five different conditions. At the end of each experimental condition, participants were presented with an emergency scenario to evaluate their current trust in the robot. We conclude that there is correlation between the magnitude of an error performed by the robot and the corresponding loss of trust of the human in the robot. We also found a correlation both between individual personalities and characteristics of people and their perceptions of the robot and trust towards a robot.**

I. INTRODUCTION

In the not too distant future, autonomous robots will take part in peoples' daily living activities. In particular, humans will have to interact with them in domestic environments. This prospect will open two main challenges for consideration: Humans will need to accept the presence of the robot and they will also have to trust that their robotic companion will look after their well-being without compromising their safety. Trust determines human's acceptance of a robot as a companion and in their perception of the usefulness of imparted information and capabilities of a robot [1], [2]. Higher trust is associated with the perception of higher reliability [3]. Furthermore, other aspects such as the appearance, type, size, proximity, and behaviour of a particular robot will also affect user's perceptions of the robot [4], [5]. Syrdal et al. [6] showed that dog-inspired affective cues communicate a sense of affinity and relationship with humans. Martelaro et al. [7] established that trust, disclosure, and a sense of companionship are related to expressiveness and vulnerability. They showed how a sense of the robot's vulnerability, through facial expressions, colour and movements, increased perceived trust and companionship, and increased disclosure. Lohse et al. [8] demonstrated that robots with more extrovert personalities are perceived more positively by some users.

Robots are machines and they might exhibit occasional mechanical or functional errors. For example, the robot may turn off during a delicate task because its battery was fully discharged without warning, or a robot might unlock the front door to strangers who may be potential thieves. People might perceive errors differently according to the resultant consequences and the timing of when they happened. Indeed, the impact of 'big errors' or an accumulation of 'small errors' might be perceived differently.

Our works [9], [10], [11] analysed human users' perceptions of the severity of errors made by a robot and their impact on human users' trust. Such analysis was intended to categorise potential errors that are likely to be exhibited by a domestic robot according the participants' perceptions (i.e., which errors are considered having 'big' and 'small' consequences), and to identify how the timing and severity of these errors influence the participants' trust in robots. We analysed how human users' personalities and characteristics affect their trust towards robots. This is particularly relevant in designing guidelines for Human-Robot Interaction in home environments where the interaction is strictly connected to humans' dynamics.

Research Questions

This work has been carried out considering different assumptions to investigate the following research questions (R) and hypothesises (H):

R1 Which kind of erroneous behaviours impact a human's trust in a robot? **H1** We expect that there is a correlation between the magnitude of the error performed by the robot and the loss of trust of the human in the robot. We hypothesise that errors with severe consequences have more impact on humans' trust in robots.

R2 Does the impact on trust change if the error happens at the beginning or end of an interaction? **H2** We expect that there is a correlation between the timing in which the error is performed during the interaction and the loss of trust. Similar to Human-Human relationships [12], we believe that humans

[1]A. Rossi, K. Dautenhahn, K. L. Koay and M. L. Walters are with Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, Hatfield, UK [a.rossi, k.dautenhahn, k.l.koay, m.l.walters]@herts.ac.uk

[2]K. Dautenhahn is with the Departments of Electrical and Computer Engineering/Systems, Design Engineering, University of Waterloo, 200 University Ave. W. Waterloo, Ontario kerstin.dautenhahn@uwaterloo.ca

recover trust more completely and quickly after the violation of trust in a later stage of the Human-Robot relationship.

R3 Is it easier to recover/regain human trust when it is a big error that occurs either at the beginning or at the end of the interaction? Or is it easier to regain/recover it when a loss of trust is caused by a small error happening at either the ends of the interaction? **H3** We expect that there is a correlation between the time at which the error occurred and the magnitude of the error. We hypothesise that a big error has more impact on the loss of trust when it happens at the end of the interaction because the human users do not have time to recover from the loss of trust.

R4 Do personalities and characteristics of humans affect their perception of a robot? Do personalities and characteristics of humans affect their trust in a robot? **H4** We expect that there is a correlation between both the personalities and characteristics of people, their perception of the robot and their trust in a robot. As with Human-Human relationships [13], [14], [15], we hypothesise that people with stronger and more positive attitudes towards other humans are more likely to trust robots.

R5 Are the use of human social behaviours sufficient for humans to trust a robot to look after their well-being? **H5** We believe that social cues make robots more human-like, and better accepted by humans, then humans can be more inclined to rely on them.

R6 Can a human's trust in her robot change over time? **H6** We believe that trust could change if the initial conditions of trusting a robot change, e.g. the robot starts to show erratic behaviours.

II. Human perceptions of the severity of domestic robot errors

There are several definitions of trust, however there is a tendency [17] in adopting the following definition: "Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" [18, p. 51]. Trust is a complex feeling even between humans [16] and it can change during the course of interactions due to several factors [1].

Higher trust is associated with the perception of higher reliability [3]. Therefore, humans may perceive erroneous robot behaviours according to their expectations of a robot's proper functions [19]. However, robots can be faulty, due to mechanical or functional errors. For example, a robot might be too slow due to batteries running low. It might not be able to detect an obstacle and destroy a human user's favourite object, or the arm of the robot might cause a breakage during a delicate task. Each of these examples are robot errors, though their magnitude might be perceived differently according to the resultant consequences.

But which type of errors have more impact on human perceptions of robots? Factors may include severity and duration, the impact of isolated 'big errors', or an accumulation of 'small errors'. For example, Muir and Moray [31] argue that human perceptions of a machine are affected in a more severe and long-term way by an accumulation of 'small' errors rather than one single 'big' error. The embodiment of a robot may also have a major impact on the perception of it by humans [4].

What is perceived as a 'big error' and what is a 'small error'? People have individual differences, including age, gender, cultural and social habits, which may impact their perceptions of what are considered big or small errors. In order to study the differences in terms of the impact of errors on a human-robot interaction, first we have to establish what people consider subjectively to be 'small' or 'big' errors exhibited by a home companion robot. In this context, our first study was directed towards the classification of likely robot errors according to their perceived magnitude.

A. Method

This study has been organised as a within-subjects experiment. Each participant has been shown the same questions, rated using a 7-point Likert scale [1= small error and 7=big error].

B. Procedure

Participants were asked to imagine that they live with a robot companion in their home. However, the robot might make some mistakes. The participant has to complete a questionnaire rating the magnitude of the errors illustrated in different scenarios, e.g. "Your robot leaves your pet hamster outside the house in very cold weather". The questionnaire is composed of 20 questions, plus two optional in which the participant is free to add their own examples of errors not already included in the scenarios proposed.

C. Results

According to the resulting answers of 50 participants - (32 men, 18 women), 19 to 63 years old [mean 41, std 11.59]. All the questions with values < 4 are considered small errors, those with values > 4 are considered big errors and those with values $= 4$ are considered neutral errors. We identified 7 big errors, 6 small errors and 7 moderate errors. We did not find any significant differences between gender or age of the participants and their rating of the errors.

III. How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario

In order to enable safe Human-Robot Interaction in home environments, it is important to investigate how an interactive relationship can be established and preserved between human users and their robotic companions, along with the likelihood of robot errors occurring. In this context, this study investigated the impact of errors with different magnitudes and order of presentation on peoples' trust of robots.

A. Method

As part of a virtual, interactive storyboard, we observed and analysed participants' behaviours during interactions with a robot called Jace. We used a between-subject experimental design. Participants were asked to read a story and interact with the robot, using their mouse and keyboard,

whenever they were invited by the robot. In order to test our research questions, each experiment was executed under 5 different conditions: condition **C1**: 10 different tasks executed correctly by the robot; condition **C2**: 10 different tasks with 3 trivial errors at the beginning and at the end of the interaction; **C3**: 10 different tasks with 3 trivial errors at the beginning and 3 severe errors at the end of the interaction; **C4**: 10 different tasks with 3 severe errors at the beginning and 3 trivial errors at end of the interaction; and **C5**: 10 different tasks with 3 severe errors at the beginning and at the end of the interaction. All the conditions with errors were interspersed by the same 4 correct behaviours.

At the end of each condition, the participants were presented with a final task in which a fire started in their kitchen and they were presented with the following options 1) to trust the robot choosing the option "I trust Jace to deal with it."; 2) to not trust the robot choosing the option "I do not trust Jace. I will deal with it."; 3) to work with the robot, supervising the emergency, choosing the option "I want to extinguish it together with Jace."; 4) to not trust either the robot or themselves choosing the option "We will both leave and call the fire brigade.".

Finally, in order to analyse the interaction between the human participants and the robot, we asked the participants to answer two sets of different questions.

B. Procedure

Participants were asked to imagine that they lived with a robot as a companion in their home which helps them with everyday activities. They were tested using an interactive storyboard accessible through a web application.

We asked participants different questions at the beginning and end of the interaction:

Questionnaire 1 A pre-experimental questionnaire for 1) collecting demographic data (age, gender and country of residence), 2) the Ten Item Personality Inventory questionnaire about themselves (TIPI) [20], 3) 12 questions to rate their disposition to trust other humans [21] and 4) and to assess participants' experience and opinion with regard to robots.

Questionnaire 2 A post-experimental questionnaire including: 1) questions to confirm that participants were truly involved in the interactions and had noticed the robot's errors, 2) to collect participants' considerations about their feelings in terms of trust and appeasement (e.g."was the robot irritating/odd?" and "why did/did not you trust the robot?"), and their perceptions of the interactions (e.g. "did the scenario look realistic?") and 3) questions to collect the participants' evaluation of the magnitude of the errors presented during the interactions.

C. Results

We analysed responses from 200 participants (115 men, 85 women), aged 18 to 65 years old [avg. age 33.56, std. dev. 9.67]. Participants' country of residence was: 60% USA; 34% India; 6% European and other countries.

We asked participants four questions about the content of the scenarios to verify the level of their engagement with the story presented. Correct answers were received for 79.75% (max 92%, min. 71.5%). We analysed the responses of 154 participants, not including those who gave more than one wrong answer (thus identified as not paying very much attention to the study - which can be expected in an online survey) to the verification questions.

We observed that a majority of participants chose to deal with the emergency situation collaboratively, and a slightly smaller majority chose to trust the robot when tested with **C1**. Participants chose not to trust the robot when it made severe errors (**C5**), while they were more inclined to trust in teamwork when the robot made small errors (**C2** and **C3**). We also noticed that the number of participants who chose to trust the robot increased in **C3**. While this might indicate a tendency of participants to not trust the robot more when the severe errors were made by the robot at the beginning of the interaction, we did not find any statistically significant association.

We observed that the association of the choices of the participants for the emergency scenario and the experimental conditions is statistically significant ($\chi^2(12) = 32.91, p = 0.001$). The strength of relationship (Cramer's V) between the emergency choice and experimental conditions is moderate ($\phi_c = 0.26, p = 0.001$).

There is a correlation between the condition **C5** and the choice of the participants to not trust the robot (adjusted value > 1.96). We observed that participants' trust is affected more severely when the robot made errors with severe consequences. We did not find any significant dependency ($p > 0.3$) between the gender of the participants and their choices in trusting the robot to deal with the emergency. We did not find any statistically significant association for different age ranges of the participants and their emergency choices ($p > 0.12$). Therefore, we assume that these results can be generalised to a generic population independently of gender and age. Moreover, in order to test the association between participants' emergency choices and their country of residence, we used a Chi-Square Test. Since the majority of the countries of residence had only one participant, we applied the test only to India and USA. We observed that the association is not statistically significant ($\chi^2(3) = 4.138, p > 0.24$)).

We found a strong connection between the personality traits of agreeableness, conscientiousness and emotional stability, and their disposition of trust other people.

The majority of our participants did not have any previous experience of interaction with robots (79.97%, min=1, max=6, mean 1.64, std. dev. 1.27). Interestingly, from participants' responses we noticed that according to their experiences, extroverted participants tended to consider robots generally as a machine ($p = 0.007$) and agreeable participants as an assistant ($p = 0.007$), in contrast to their perceptions of the robot they interacted with in this study. In particular, extroverts perceived Jace as a friend ($p = 0.0019$) and a warm and attentive entity ($p = 0.0025$), while agreeable participants perceived Jace as a tool ($p = 0.0033$). We also found that extroverted participants would like to have Jace

as home companion ($p = 0.001, r = 0.269$) and believe it is reliable ($p = 0.002, F = 2.729$) and trustworthy in uncertain and unusual situations ($p(12) = 0.026, F = 2.025$).

Finally, we analysed participants' personalities and dispositions of trust with regard to their final choice of trusting the robot in an emergency scenario. We found that conscientiousness ($p(3) = 0.42, F = 2.803$) and agreeableness ($p(3) = 0.022, F = 3.320$) traits correlate with participants' propensity for trusting the robot, and participants' belief in benevolence of people also correlate with higher trust in Jace ($p = 0.014, F = 6.078$). Moreover, we observed that the errors made by the robot significantly affected participants' perception of the robot.

IV. CONCLUSIONS

Regarding the research question **R1**, our hypothesis **H1** suggested that there is a correlation between the severity of the error performed by the robot and humans not trusting the robot. Our study shows that the magnitude of the errors made by the robot, and humans not trusting the robot are correlated. In particular, participants' trust was affected more severely when the robot made errors having severe consequences. We also hypothesised in **H2** that the timing when the error is performed affects the trust towards robots (research question **R2**), and there is a correlation between the timing of when the error occurred and the magnitude of the error (research question **R3** and hypothesis **H3**). Our results marginally suggest also that there might be a tendency not to trust the robot when severe errors happen at the beginning of an interaction, but these differences were not statistically significant.

As indicated in Hypothesis **H4**, we found a correlation both between individual personalities and characteristics of people and their perception of the robot and trust towards a robot (research question **R4**).

We are currently investigating research questions **R5** and **R6**.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (Safety Enables Cooperation in Uncertain Robotic Environments - SECURE).

REFERENCES

[1] D. Cameron, J. M. Aitken, E. C. Collins, L. Boorman, A. Chua, S. Fernando, O. McAree, U. Martinez-Hernandez, and J. Law, "Framing factors: The importance of context and the individual in understanding trust in human-robot interaction," in *International Conference on Intelligent Robots and Systems*, 2015.

[2] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, 2011.

[3] J. M. Ross, "Moderators of trust and reliance across multiple decision aids (doctoral dissertation), university of central florida, orlando." 2008.

[4] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video–displayed agents," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 41–52, 2011.

[5] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, "Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2007, pp. 564–569.

[6] D. S. Syrdal, K. L. Koay, M. Gcsi, M. L. Walters, and K. Dautenhahn, "Video prototyping of dog-inspired non-verbal affective communication for an appearance constrained robot," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2010, pp. 632–637.

[7] N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds, "Tell me more designing hri to encourage more trust, disclosure, and companionship," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2016, pp. 181–188.

[8] M. Lohse, M. Hanheide, B. Wrede, M. L. Walters, K. L. Koay, D. S. Syrdal, A. Green, H. Hüttenrauch, K. Dautenhahn, G. Sagerer, and K. Severinson-Eklundh, "Evaluating extrovert and introvert behaviour of a domestic robot -a video study," in *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 2008, pp. 488–493.

[9] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "Human perceptions of the severity of domestic robot errors," in *Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds. Cham: Springer International Publishing, 2017, pp. 647–656.

[10] ——, "How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario," in *Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds. Cham: Springer International Publishing, 2017, pp. 42–52.

[11] R. Alessandra, D. Kerstin, K. Kheng Lee, and M. L. Walters, "The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario." vol. 9, 2018.

[12] O. Schilke, M. Reimann, and K. S. Cook, "Effect of relationship experience on trust recovery following a breach," *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15 236–15 241, 2013.

[13] M. P. Haselhuhn, M. E. Schweitzer, and A. M. Wood, "How implicit beliefs influence trust recovery," *Psychological Science*, vol. 5, pp. 645–648, 2010.

[14] T. Mooradian, B. Renzl, and K. Matzler, "Who trusts? personality, trust and knowledge sharing," *Management Learning*, vol. 37, no. 4, pp. 523–540, 2006.

[15] F. B. Tan and P. Sutherland, "Online consumer trust: A multi-dimensional model." vol. 2, 2004, pp. 40–58.

[16] R. M. Kramer and P. J. Carnevale, "Trust and intergroup negotiation," *Blackwell Handbook of Social Psychology: Intergroup Processes (eds R. Brown and S. L. Gaertner)*, pp. 431–450, 2003.

[17] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen, "User trust dynamics: An investigation driven by differences in system performance," vol. 126745. ACM, 2017, pp. 307–317.

[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[19] M. L. Walters, M. A. Oskoei, D. S. Syrdal, and K. Dautenhahn, "A long-term human-robot proxemic study," in *2011 RO-MAN*, July 2011, pp. 137–142.

[20] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr., "A very brief measure of the big five personality domains. journal of research in personality," pp. 504–528, 2003.

[21] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information Systems Research*, vol. 13, no. 3, pp. 334–359, 2001.

Virtual Reality as a Tool to Study Embodied Cognition

F. Foerster[1*], J. Goslin[1]
[1] *University of Plymouth, School of Psychology, PL4 8AA, United Kingdom*
**Contact: francois.foerster@plymouth.ac.uk, phone +44-74-26-94-11-53*

I. State of the Art

Contemporary cognitive and social robotics share important scientific questions with the fields of psychology and neuroscience [1, 2]. How does an artificial machine can learn to safely manipulate an object? While roboticists try to build machines performing as efficiently as humans, neuroscientists try to understand how the brain works and leads to intelligent behaviours. During the last 10 years, multiple novel tools to study human cognition appeared on the market, such as eye tracking and motion capture systems. Nowadays, modern Virtual Reality (VR) systems are intensively used in academia to investigate human behaviours. The most beneficial feature is that they provide both laboratory settings (well-controlled experiment) and near to ecological environment. However, VR can provide much more information than academics do collect, such as kinematic data and neural data when coupled with Electroencephalography (EEG). The aim of this extended abstract is to describe existing VR paradigms and our original VR setup coupled with EEG, used to study how humans build novel representations of objects and actions. The goal of our project is to provide a better understanding of the neural bases of novel objects and actions representations.

Recent advances in VR technology allow us to go beyond the initial perspectives. For instance, now VR is used to investigate how users' process their own space. Spatial cognition is important for psychologists as well as cognitive roboticists (eg. the importance of peripersonal space for a robotic arm safety [3]). Using an immersive virtual reality paradigm, Iachini et al. [4] investigated what are the distances necessary for a user to be comfortable interacting with a virtual avatar or a robot. Another VR study showed that motor affordances provided by everyday objects (eg. the handle of a cup) are processed only when the object is situated in the reachable space [5]. A similar VR environment was used [6] to establish how object knowledge is also accessed automatically upon viewing tools and other manipulable objects when they are within reach. This means that our implicit affordances perception and manipulation knowledge are modulated by the stimulus position in the space. The following research from the same team showed that neuronal μ rhythm (8-13 Hz) represents a neural signature of this affordance processing [7, 8]. To do so, the authors used goggles with projected stereoscopic images, which differ from our approach using full head-mounted displays.

Such investigations have been possible because VR now goes beyond passive viewing, and can be used to represent virtual tools that the user can manipulate by proxy through physical tracked controllers. This means that the user can manipulate and affect their virtual environment, a central tenet of embodied cognition. This approach allows us to examine some of the basic properties of the embodied approach through extract control methodological factors (eg. properties of the stimuli), that are robust, and repeatable. They also allow us to overcome significant logistic issues (eg. placing and removing an object manually at different distances from the participants thousands of times) that tend to make physical experimental studies unfeasible, or underpowered.

II. Method

As described briefly, VR is a modern tool to study how the human brain process objects and guide their manipulation. Our lab investigates these cognitive processes using a VR setup coupled with EEG recordings (Figure 1). Using two interfaced computers and a VR head-mounted display placed on the top of an EEG cap, the EEG recordings are synchronized with multiple events happening in the virtual environment. Our setup allows us to track the neural activity underlying the recognition of objects (ie. a stimulus onset) and motor control (ie. movement onset and grasp onset). Hence, the setup also provides the opportunity to investigate the emergence of novel object representations through the use of novel objects (Figure 2).

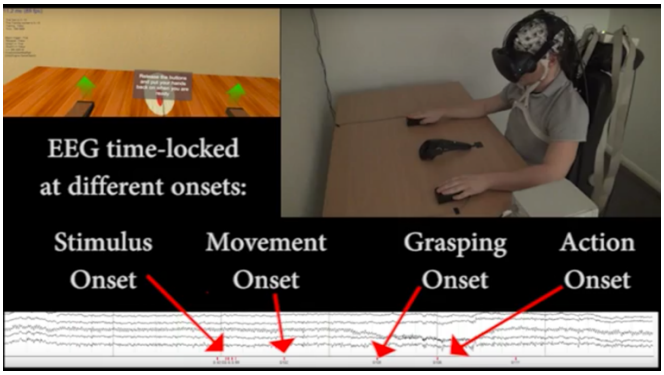


Figure 1 Representation of a virtual environment (top left) where the participant interacts with a controller (top right). Participant's EEG is synchronized to the key events of the experiment (bottom). Stimulus onset: the participant processed the apparition of an object. Movement onset: the participant released a hand from a button situated on the table. Grasping onset: the participant grasped the controller at a location A. Action onset: the participants placed the controller on a location B.

Figure 2 Example of 3D models created by the researchers that participants learn to manipulate.

Finally, as VR controllers are tracked in real-time by two cameras, the setup allows the researcher to track how participants manipulate them. For instance, the position, the rotation and the velocity of the controller can be tracked in the space during the transportation of the controller from a location A to a location B (Figure 3).

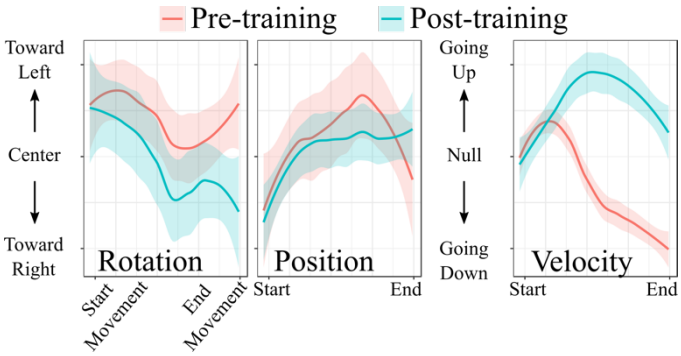


Figure 3 The task of the participant was to transport an object from a location A to a location B. In the middle of the experiment, participants perform another motor task (training). The tracked VR controller allows to record and compares the rotation, the position and the velocity (single axis presented) during the transport task (ie. hand movement) before and after the motor training. Here, performing a motor training influences the motor control of the transport task afterward, especially the velocity of the hand movement.

III. Conclusion

Coupling VR with EEG techniques allowed our team to investigate the neural activity underlying the recognition of novel tool and the selection of learnt tool use [9-11]. As investigated in cognitive robotics, we use this setup to understand how humans build representations of novel objects and actions. To conclude, the number of applications of VR goes beyond the primary goals expected twenty years ago. Most recent research in cognitive science and related fields couple VR with other well-known technologies, such as EEG techniques in order to overcome methodological limitations and therefore extend their scientific potentials.

References

[1] T. Schack and H. Ritter, "Representation and learning in motor action - Bridges between experimental research and cognitive robotics," *New Ideas Psychol.*, vol. 31, no. 3, pp. 258–269, 2013.

[2] R. De Kleijn, G. Kachergis, and B. Hommel, "Everyday robotic action: Lessons from human action control," *Front. Neurorobot.*, vol. 8, no. 13, pp. 1–9, 2014.

[3] D. H. P. Nguyen, M. Hoffmann, A. Roncone, U. Pattacini, and G. Metta, "Compact Real-time avoidance on a Humanoid Robot for Human-robot Interaction," *IEEE Int. Conf. Human-robot Interact.*, pp. 416–424, 2018.

[4] T. Iachini, Y. Coello, F. Frassinetti, and G. Ruggiero, "Body space in social interactions: A comparison of reaching and comfort distance in immersive virtual reality," *PLoS One*, vol. 9, no. 11, pp. 25–27, 2014.

[5] M. Costantini, E. Ambrosini, G. Tieri, C. Sinigaglia, and G. Committeri, "Where does an object trigger an action? An investigation about affordances in space," *Exp. Brain Res.*, vol. 207, no. 1–2, pp. 95–103, 2010.

[6] S. Kalénine, Y. Wamain, J. Decroix, and Y. Coello, "Conflict between object structural and functional affordances in peripersonal space," *Cognition*, vol. 155, pp. 1–7, 2016.

[7] Y. Wamain, F. Gabrielli, and Y. Coello, "EEG mu rhythm in virtual reality reveals that motor coding of visual objects in peripersonal space is task dependent," *Cortex*, vol. 74, pp. 20–30, 2016.

[8] Y. Wamain, A. Sahaï, J. Decroix, Y. Coello, and S. Kalénine, "Conflict between gesture representations extinguishes μ rhythm desynchronization during manipulable object perception: an EEG study," *Biol. Psychol.*, vol. 132, no. January, pp. 202–211, 2018.

[9] F. Foerster, J. Goslin, "Moving or using objects A difference of semantics or motor complexity?," (submitted), 2018.

[10] F. Foerster, J. Goslin, "Perceiving novel objects through their manipulative and functional properties: the independent contributions of the sensorimotor Mu and Beta rhythms," (under preparation), 2018.

[11] F. Foerster, J. Goslin, "Sensorimotor rhythms for the retrieval and selection of novel tool utilizations," (under preparation), 2018.

Progress Report: Language Learning for Safety during Human-Robot Interaction

Chandrakant Bothe, Sven Magg, Cornelius Weber and Stefan Wermter
Knowledge Technology, Department of Informatics
University of Hamburg
Hamburg, Germany
{bothe, magg, weber, wermter}@informatik.uni-hamburg.de

***Abstract*—In this work, we tackle the problem of developing a dialogue system for robots with multivariate behavioral adaptation as a preliminary step towards potentially learning safety concepts for safer human-robot interaction. With the concern of safety during the human-robot verbal interaction, the aim is to study and research different linguistic aspects. We found that language is very complex but comprehensive, and different linguistic features could be used to assess the behavior. We start with sentiment guided learning of the safety concepts. We also found that for the language interaction, we need a dialogue system which drives dialogue flow. In natural language understanding, dialogue act, which represents a functional type of utterance, plays a very important role in a dialogue system. We developed neural inference models to recognize and classify the dialogue acts. We also follow up with discourse analysis, which is one of the important processes in the development of dialogue systems. Results of research in this direction allow us to revisit the dialogue systems, develop and deploy on a robot to demonstrate a proof of concept.**

***Index Terms*—natural language processing, human-robot interaction, dialogue systems, dialogue acts, discourse analysis**

I. Introduction

In a conversation, humans use changes in a dialogue to predict safety-critical situations and use it to react accordingly. We propose to use these kinds of cues for safer human-robot interaction through early detection of dangers. In the section below, you will find the list as a research progress from learning the linguistic feature to developing a dialogue system for the robot which can adapt their behavior based on linguistic features. The features learned using learning approaches such as artificial neural networks and deep learning.

II. Approaches

A. Sentiment guided language learning

Sentiment can drive conversation based on their polarity. For example, being sentimentally positive in the language can bring positive utterances and vice versa. We attempt to model such a model to learn to estimate the sentiment of the next upcoming utterance based on a few preceding utterances [1]. Due to a low availability of sentiment annotated dialogue corpora, we use a sentiment classification for utterances, to learn sentiment changes within dialogues and ultimately predict the sentiment of upcoming utterances.

We show that training a recurrent neural network on context sequences of words, defined by two preceding utterances of each speaker with the sentiment class of the next utterance, leads to useful predictions of the sentiment class of the upcoming utterance. See the example in Figure 1 to relate the safety learning process using sentiment as a guiding cue. We also explore the emotion intensity detection by using character- and word-level recurrent neural network models [2].

B. Dialogue act recognition

Dialogue act represents a functional importance of an utterance. It is an aspect of natural language understanding where its recognition plays an important role in building the dialogue systems (DS). We develop several neural models to learn to recognize and classify the dialogue acts. For the recognition of dialogue act, the context within the dialogue is very important, hence, modelling the neural models the same way is crucial [3]. We develop a recurrent neural model which uses a character level language model feature for each utterance. This model surpasses some of the state-of-the-art results on the Switchboard Dialogue Act corpus.

However, we also attempt to answer the research question that how much context information is needed while recognizing the dialogue act of utterance. Hence, we develop a similar neural model with attention mechanism on the top, which computes the weights of the contribution of preceding utterances while recognizing the dialogue act of current utterance [4]. The architecture uses a bi-directional recurrent neural network with attention mechanism.

C. Discourse analysis

Discourse analysis can be performed by analyzing the dialogue act of sequence of utterances in a conversation.

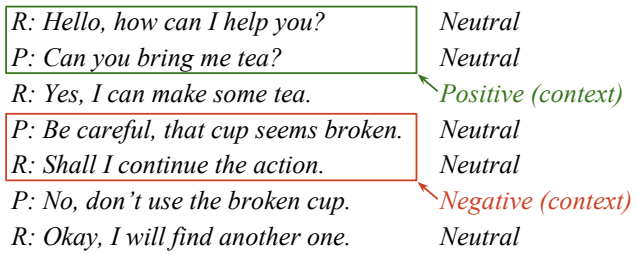


Fig. 1. Example for preparing the contexts

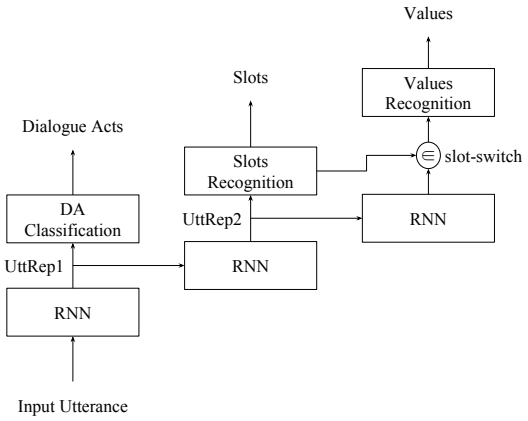


Fig. 2. Hierarchical recurrent neural networks for dialogue acts and slot-value pair recognition

Our live web-demo called Discourse-Wizard is available[1] for discourse analysis. The backend used for this live web-demo is similar to our previous work at the dialogue act recognition, and more details can be found in [5].

D. Dialogue systems (DS)

As a result, we aim to develop a dialogue system for the social robots which could take several linguistic features into account and infer accordingly. We developed a simple dialogue system which uses deep learning as a backend for spoken language understanding. As a first step, we develop a natural language interface for the simulated agent in AI2Thor environment [6]. The language understanding module is able to decode the input utterance into the symbolic representation using hierarchical recurrent neural networks as shown in Figure 2. For example, the utterance *"please move to the right"* can be decoded as $\{da : moveRobot, direction : right\}$ where *da* represents the dialogue act or intention, *direction* is a slot and *right* its value.

E. DS with politeness as a social cue

We developed the next part as a result of the dialogue system where we add another module like politeness detection, as shown in Figure 3. Dialogue act module is as same as spoken language understanding described previously. The response manager picks an appropriate response from the data file based on intention and the degree of politeness.

F. DS for robot adapting behavior based on politeness

As a proof of concept, we have developed and deployed our DS on the robot which adapts its behavior based on a degree of politeness. It is demonstrated with practical experiment as a part of the project during secondment[2]. DS communicated with the robot through state and motion managers for appropriate actions such as behavioral changes and navigation.

[1]https://secure-robots.eu/fellows/bothe/discourse-wizard-demo/ and full demo website at https://crbothe.github.io/discourse-wizard/

[2]We have accomplished this experiment during secondment in collaboration with the industrial partner SoftBank Robotics in Paris, France with their semi-humanoid robot Pepper.

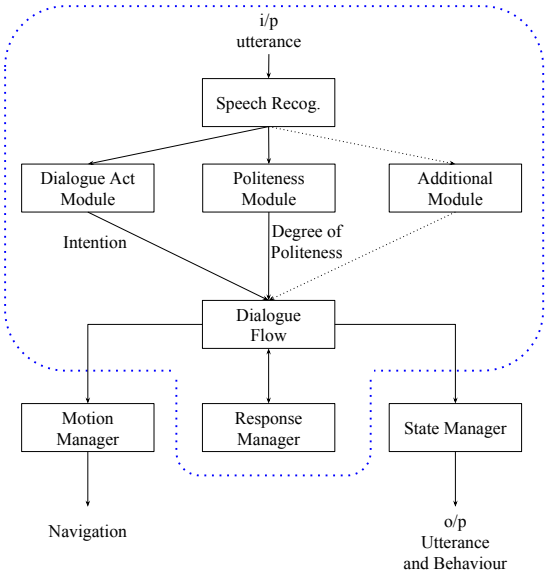


Fig. 3. Dialogue system with different modules

III. Conclusion

We discussed most of the stages of progress in our research in the direction of the language learning for safety during human-robot interaction. We gave the pointers to deal with the language processing for dialogue system development and integration of those with the robot that shall accordingly adapt its behavior.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 framework programme for research and innovation under the Marie Sklodowska-Curie Grant Agreement No. 642667 (SECURE).

References

[1] C. Bothe, S. Magg, C. Weber, and S. Wermter, "Dialogue-based neural learning to estimate the sentiment of a next upcoming utterance," in *Artificial Neural Networks and Machine Learning – ICANN 2017*, A. Lintas, S. Rovetta, P. F. Verschure, and A. E. Villa, Eds. Cham: Springer International Publishing, 2017, pp. 477–485.

[2] E. Lakomkin, C. Bothe, and S. Wermter, "GradAscent at EmoInt-2017: Character and Word Level Recurrent Neural Network Models for Tweet Emotion Intensity Detection," in *Proc. of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis at the Conference EMNLP*. ACL, 2017, pp. 169–174.

[3] C. Bothe, C. Weber, S. Magg, and S. Wermter, "A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks," in *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ERLA), 2018, pp. 1952–1957.

[4] C. Bothe, S. Magg, C. Weber, and S. Wermter, "Conversational Analysis using Utterance-level Attention-based Bidirectional Recurrent Neural Networks," in *Proc. of the International Conference INTERSPEECH 2018*, 2018.

[5] ——, "Discourse-Wizard: Discovering Deep Discourse Structure in your Conversation with RNNs," *preprint arXiv:1806.11420*, 2018.

[6] ——, "Natural Language Interface to Control an AI2Thor Simulator Agent using a Deep Learning Backend," in *[submitted to] Proc. of the International Conference IROS 2018: Workshop – Towards Intelligent Social Robots*, 2018.

Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks

Egor Lakomkin[1], Cornelius Weber[1], Sven Magg[1] and Stefan Wermter[1]

I. RESEARCH MOTIVATION

In the near future, the presence of robots in home environments will become more common, helping humans with daily tasks, for instance assisting elderly people. One important area where robots can be very helpful is identifying possible dangerous situations in home environments. Robots can observe the situation at the moment and try to evaluate if there is a potential threat. A robot can be taught to do this with machine learning methods. As an example, given a speech segment, it would be interesting to predict if a person is excited or neutral. In this paper, firstly I outline main questions and directions in my PhD research, then I present current achieved results followed by the related and future work sections.

II. RESEARCH DIRECTIONS AND ACHIEVED RESULTS

I identify three main directions in my research: 1) features and signal representations learning for speech emotion recognition (SER) task. 2) investigation of neural architectures which allow robust to an internal robot's and an environmental noise emotion recognition 3) research on the methods and approaches to incorporate information contained in modalities other than auditory to improve speech emotion recognition. For example, linguistic analysis of a spoken text or facial expression recognitions can help in difficult situations when analyzing only acoustic signal is not enough to infer an affective state of the speaker.

A. FEATURES FOR SPEECH EMOTION RECOGNITION

I evaluate several dual architectures which integrate representations of the automatic speech recognition (ASR) neural network: a fine-tuning and a progressive network. The fine-tuning architecture reuses features learnt by the recurrent layers of a speech recognition network and can use them directly for emotion classification by feeding them to a softmax classifier or can add additional hidden SER layers to tune ASR representations. Additionally, the ASR layers can be static for the whole training process or can be updated as well by allowing to backpropagate through them. The progressive architecture complements information from the ASR network with SER representations trained end-to-end. Our experiments on the IEMOCAP dataset show 10% relative improvements in the accuracy and F1-score over the baseline recurrent neural network which is trained end-to-end for emotion recognition. Results were published and presented at the IJCNLP 2017 conference [1].

[1]University of Hamburg, Department of Informatics, Knowledge Technology Institute. Vogt-Koelln-Strasse 30, 22527 Hamburg, Germany lakomkin @informatik.uni-hamburg.de

B. LEARNING EARLY EMOTION RECOGNITION

Acoustically expressed emotions can make communication with a robot more efficient. Detecting emotions like anger could provide a clue for the robot indicating unsafe/undesired situations. Recently, several deep neural network-based models have been proposed which establish new state-of-the-art results in affective state evaluation. These models typically start processing at the end of each utterance, which not only requires a mechanism to detect the end of an utterance but also makes it difficult to use them in a real-time communication scenario, e.g. human-robot interaction. We propose the EmoRL model that triggers an emotion classification as soon as it gains enough confidence while listening to a person speaking. As a result, we minimize the need for segmenting the audio signal for classification and achieve around 50% latency reduction as the audio signal is processed incrementally. The method is competitive with the accuracy of a strong baseline model, while allowing much earlier prediction. The results will be presented at the ICRA 2018 conference.

C. ROBUST ACOUSTIC EMOTION RECOGNITION

Many neural network-based architectures were proposed recently and pushed the performance to a new level. However, the applicability of such neural SER models trained only on in-domain data to noisy conditions is currently under-researched. In this work, we evaluate the robustness of state-of-the-art neural acoustic emotion recognition models in human-robot interaction scenarios. We hypothesize that a robot's ego noise, room conditions, and various acoustic events that can occur in a home environment can significantly affect the performance of a model. We conduct several experiments on the iCub robot platform and propose several novel ways to reduce the gap between the model's performance during training and testing in real-world conditions. Furthermore, we observe large improvements in the model performance on the robot and demonstrate the necessity of introducing several data augmentation techniques like overlaying background noise and loudness variations to improve the robustness of the neural approaches. The results were published at the IROS 2018 conference[1].

[1]Video https://www.youtube.com/watch?v=js_TCxl_wF4

D. SEMI-SUPERVISED EMOTION RECOGNITION

One of the issues in the area of affective computation is that the amount of annotated data is very limited. On the other hand, the number of ways that the same emotion can be expressed verbally is enormous due to variability between speakers. This is one of the factors that limits performance and generalization. We propose a simple method that extracts audio samples from movies using textual sentiment analysis. As a result, it is possible to automatically construct a larger dataset of audio samples with positive, negative emotional and neutral speech. We show that pretraining recurrent neural network on such a dataset yields better results on the challenging EmotiW corpus. This experiment shows a potential benefit of combining textual sentiment analysis with vocal information. The results were published and presented at the EACL 2017 conference.

III. RELATED WORK

Deep neural networks significantly boosted the performance of acoustic emotion recognition models. The majority of recent work focuses on learning to extract useful input representations and searching for neural architectures for emotion recognition, as neural approaches outperform traditional ones like support vector machines and decision trees [2].

Recurrent neural networks have an ability to model long-term context information and were successfully applied to emotion recognition [3], [4]. Convolutional neural networks can capture only a local context, but have an ability to model longer dependencies when their architecture was designed with a deep hierarchy [2]. Commonly, these methods train neural networks on pre-extracted features: MFCC coefficients, spectrograms and high-level information like formants, pitch, and voice probability. Alternatively, Trigeorgis et al. demonstrate a model that learns how to recognize the affective state of a person directly from the raw waveform [5]. Another explored direction is transfer learning: adapting audio representations trained initially for other auxiliary tasks, like gender and speaker identification [6] or speech recognition [1], [7].

Robustness to noise was a subject of several previous work. Attention mechanisms [3], [8] aim to identify useful regions for emotion classification automatically by assigning a low importance to irrelevant inputs, for example, non-speech or silence frames. Adding background noise during training improved the robustness of neural models in several acoustic classification tasks [9]. Different types of data augmentation methods were explored by Zhou et al. [10] to improve the performance of speech recognition. Supervised domain adaptation was proposed by Abdelwahab et al. [11] to mitigate the problem of training and testing mismatch conditions by tuning the model on the small set of test samples.

Our work on the robustness of the speech emotion recognition is close to Lane et al. [9] and our main difference is that our testing conditions are not synthetically constructed by overlaying clean samples with additive noise, but recorded on the iCub robot which adds a significant amount of ego-noise. We argue that distortions introduced by playing a sample through speakers, changing room conditions and distance from the speech source to the robot, reverberations, added external acoustic events and the robot's internal noise introduce non-linear deformations which are challenging for the neural network to deal with.

IV. FUTURE WORK

In future work, I plan to investigate further ways to enhance the data augmentation pipeline for a robust speech emotion recognition. For example, data-driven generative models, like generative adversarial networks, can produce realistic speech samples, which potentially can be useful during training. I plan to evaluate an option to enrich input representation with the information on the spoken text under noisy conditions as it appears to be difficult to analyze valence without it.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (SECURE)

REFERENCES

[1] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing Neural Speech Representations for Auditory Emotion Recognition," *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, vol. 1, pp. 423–430, 2017. [Online]. Available: http://www.aclweb.org/anthology/I17-1043

[2] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 8 2017.

[3] C.-W. Huang and S. Narayanan, "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition," *Proceedings of Interspeech*, pp. 1387–1391, 2016.

[4] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," *Interspeech*, 2015.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3 2016, pp. 5200–5204.

[6] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, E. M. Provost, and A. Arbor, "Progressive Neural Networks for Transfer Learning in Emotion Recognition," *Interspeech*, pp. 1098–1102, 2017.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "On the Correlation and Transferability of Features between Automatic Speech Recognition and Speech Emotion Recognition," *Interspeech*, pp. 3618–362, 2016.

[8] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," *Interspeech*, pp. 1263–1267, 2017.

[9] N. D. Lane, P. Georgiev, L. Qendro, and B. Labs, "DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments using Deep Learning," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 283–294, 2015.

[10] Y. Zhou, C. Xiong, and R. Socher, "Improved Regularization Techniques for End-to-End Speech Recognition," *CoRR*, vol. abs/1712.07108, 2017. [Online]. Available: http://arxiv.org/abs/1712.07108

[11] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4 2015, pp. 5058–5062. [Online]. Available: http://ieeexplore.ieee.org/document/7178934/

Progress Report: Language-modulated Actions using Deep Reinforcement Learning for Safer Human-Robot Interaction

Mohammad Ali Zamani[1], Sven Magg[1], Cornelius Weber[1] and Stefan Wermter[1]

***Abstract*— Spoken language can be an efficient and intuitive way to warn robots about threats. Guidance and warnings from a human can be used to inform and modulate a robot's actions. An open research question is how the instructions and warnings can be integrated in the planning of the robot to improve safety. Our goal is to address this problem by defining a Deep Reinforcement Learning (DRL) agent to determine the intention of a given spoken instruction, especially in a domestic task, and generate a high-level sequence of actions to fulfill the given instruction. The DRL agent will combine vision and language to create a multi-modal state representation of the environment. We will also focus on how warnings can be used to shape the DRL's reward, concentrating on the recognition of the emotional state of the human in an interaction with the robot. Finally, we will use language instructions to determine a safe operational space for the robot.**

I. INTRODUCTION

In the future, robots are expected to work as companions with humans in various areas including domestic scenarios such as care-giving. Human-robot interaction safety has not been well studied [1]. Even with well-engineered robots, it would be unrealistic to move robots directly from factories to home environments to perform complex tasks [2] [3] due to safety [4]. Moreover, robots also have to continuously adapt to new environments to avoid hazardous actions since using experts to program a robot for every environment is impossible. Hence, we need adaptive learning algorithms.

Spoken language can be considered one of the most effective communication channels to warn robots about threats. For example, robots may not notice an external threat or mis-planning that may harm a human or the robot itself. However, a human can warn or guide the robot by a verbal utterance toward a safer interaction. How robots react to safety warnings is not addressed exhaustively in the literature. The closest related research area is assigning tasks to robots by verbal instructions [5], [6], [7]. They follow rule-based methods to utilize spoken language instructions which can cover only a limited number of scenarios.

Our goal is to train a robot to safely perform complex tasks with the ability of processing environmental feedback, including guidance and warnings by a human, to shape a proper signal for updating its own policy. Therefore, our research is focused on three capabilities of the robot: generating high-level actions from verbal instructions, extracting reward from prosodic/sentiment features of the human speaker, and learning a safe workspace for the robot.

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (SECURE).

[1]Knowledge Technology, Department of Informatics, University of Hamburg, Germany. zamani, magg, weber, wermter @informatik.uni-hamburg.de

II. FOCUS AREAS

A. Mapping Spoken Instruction to a Sequence of Actions

We introduced a framework to obtain the intention of a given spoken instruction (e.g. "boil water") and generate the sequence of actions ("moveto kettle", "grasp kettle", ...) to fulfill the task [8], [9]. The intention detection was implemented with a 2 layer perceptron with 20% dropout and trained by the TellMeDave corpus to predict one of 10 predefined classes. Our model could achieve 89.57% accuracy in a 5-fold cross-validation.

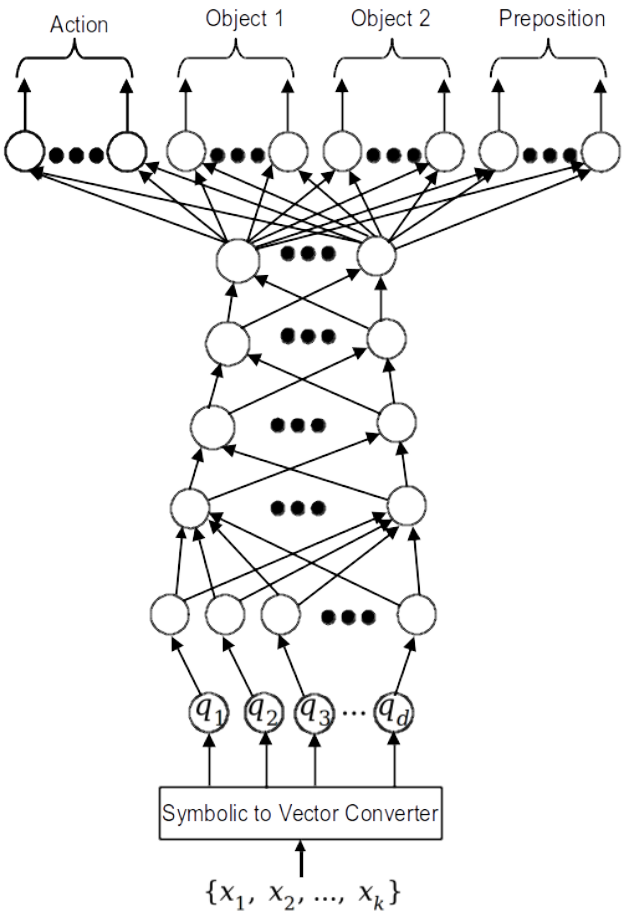


Fig. 1. The deep reinforcement learning architecture generates the sequence of actions. An MLP neural network is trained to approximate the action-value functions. The compositional linguistic state, $\chi(t)$, is presented to the network as a compositional vector which is a binary vector.

We developed a symbolics environment from the "Tell Me Dave" Corpus [10] to train the RL agent. The main contribution was to use a distributed symbolic state representation (e.g. {On Kettle Sink}, {Near Robot Sink}, ...) which reduced the learning time on given tasks. Our Reinforcement Learning was built based on the Deep Q-Network [11] architecture with modifications to support multiple Q functions and different types of value estimation. As shown in figure 1, there are four output groups in the network architecture. However, the actions in the corpus have different number of arguments (i.e. object1, object2, and preposition) from zero to three. Therefore, we masked the gradient based on the performed action.

In our case, the environment state was directly accessible through the simulation while this needs to be extracted in a real life scenario. Therefore, we will extend by encoding vision and instruction in a fused state similar to Shu et al. [12] in a more realistic simulator like AI2Thor [13] (see figure 2).

Fig. 2. The modular approach using intention detection and reinforcement learning trained for each objective to generate the sequence of actions [9].

B. Extracting Reward from the Human Speech

The robot needs to continuously process human speech to detect implicit interruptions or any change in the instruction. The robot is expected to be able to stop (both soft and emergency) with a minimum latency in an unsafe situation (see figure 4). We developed a reinforcement learning approach to optimize the accuracy and latency concurrently [14].

The model (see figure 3) was consist of recurrent neural network with Gated Recurrent Units [15] which learns a temporal representation from the extracted features of speech. The Emotion classification module (θ_c) used the GRU's output to determine emotion as angry or neutral. The action selection (θ_a) which is Monte Carlo Policy Gradient (or REINFORCE) [16] decides to either wait for the next speech frame or terminate the processing and read the emotion classification module. We also used the baseline estimation (θ_b) to estimate a baseline reward. Similar to [17], [18], this helps to lower the variance of the gradient signal.

As a result, our model achieved about 50% latency reduction with the same level of accuracy evaluated on the iCub recorded data in our lab. We also improved the robustness of emotion recognition by proposing data augmentation techniques like overlaying background noise [19].

As future work, emotion recognition will be used to filter warnings and to record this experience in the RL's memory for updating the agent's policy. We will use a pre-trained model in simulation to focus on learning new safety cases in the real scenario.

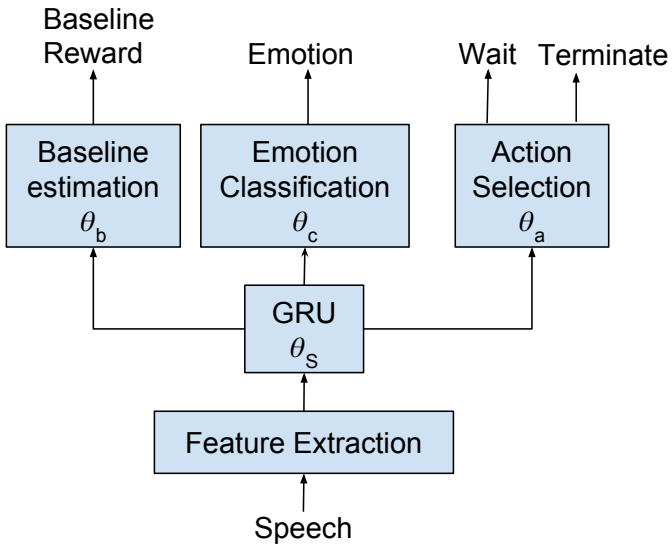


Fig. 3. The EmoRL model consists of 4 components: *Gated Recurrent Unit (GRU)*, *Emotion Classification (EC)*, *Action Selection (AS)* and *Baseline Reward Estimator (BRE)*. The GRU encodes the acoustic information of a speech signal which is used as a state representation. *EC* uses the state representation to evaluate the probability of the human speaker being in an angry state. *AS* and *BRE* determine the probability distribution over possible actions and the estimation of the baseline reward [14].

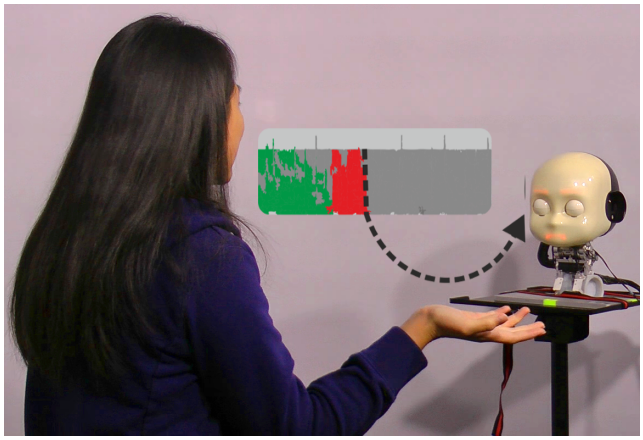


Fig. 4. Extracting reward from the human speaker. The robot analyzes continuously arriving acoustic input and only when it has enough information to evaluate the affective state of the speaker it will output the person's specific emotion. The robot is trained using reinforcement learning to make the dynamic decision: wait for more data or trigger a response [14].

C. Safe Human-Robot Collaboration in Manual Tasks

Safety becomes more important when humans work with robots collaboratively. For shaping such a collaborative scenario incrementally, as an initial step, we improved the learning of the Deep Deterministic Policy Gradient (DDPG) [20] in a reach-for-grasp task by introducing an adaptive (larger-than-life) augmented target [21]. Later, we used it to train a 2-DOF arm in an interactive scenario to reach multiple target points which improved the learning time by solving the

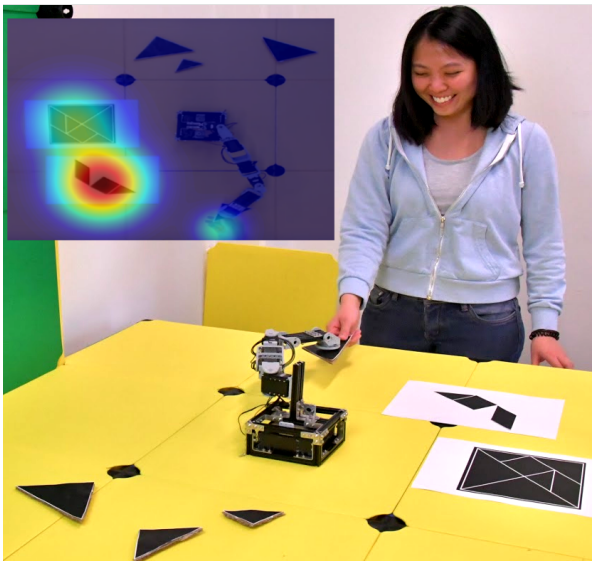


Fig. 5. A person is solving a tangram puzzle in collaboration with a robot arm. The robot arm is instructed to avoid the person's workspace while fetching puzzle pieces from the far end of the table. The top right image shows the top view overlaid with a spatial representation which can be learned by interaction with the user. The robot plans its motion incorporating the adaptive spatial constraints [23].

problem in simulation and deploying it on the robot when it gained enough confidence [22]. In a preliminary experiment (see figure 5), we demonstrated how spoken instructions can be mapped to a spatial representation of the robot's workspace which can be used as constraints for the path planner [23]. As a next step, we are also interested to learn grasping with verbally described spatial constraints in an end-to-end approach.

III. CONCLUSIONS

In this PhD project, spoken instructions are used in different areas and we focused on the high level action sequences for performing tasks in a domestic scenario. As a next step, we will concentrate on obtaining state representations in real life scenarios. In parallel, we proposed a model to detect angry emotions rapidly, which can be used as an implicit interruption to planning to lead to a safer human-robot interaction. As future work, we will focus on how the robot can learn from experience to immediately avoid the same behavior. We also investigate how teaching the operational space to the robot can be performed intuitively. We plan to extend this to a small kitchen scenario which can bring together all these ways of using spoken instructions/warnings to guide the robot towards safer interaction.

REFERENCES

[1] M. Vasic and A. Billard, "Safety issues in human-robot interactions," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 197–204.

[2] S. Schaal, "The new robotics–towards human-centered machines," *HFSP journal*, vol. 1, no. 2, pp. 115–126, 2007.

[3] S. Schaal and C. G. Atkeson, "Learning control in robotics," *IEEE Robotics & Automation Magazine*, vol. 17, no. 2, pp. 20–29, 2010.

[4] J. Peters and S. Schaal, "Learning to control in operational space," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 197–212, 2008.

[5] S. Lauria, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein, "Converting natural language route instructions into robot executable procedures," in *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*. IEEE, 2002, pp. 223–228.

[6] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein, "Mobile robot programming using natural language," *Robotics and Autonomous Systems*, vol. 38, no. 3, pp. 171–181, 2002.

[7] T. Nishizawa, K. Kishita, Y. Takano, Y. Fujita *et al.*, "Proposed system of unlocking potentially hazardous function of robot based on verbal communication," in *System Integration (SII), 2011 IEEE/SICE International Symposium on*. IEEE, 2011, pp. 1208–1213.

[8] M. A. Zamani, S. Magg, C. Weber, and S. Wermter, "Deep reinforcement learning using symbolic representation for performing spoken language instructions." *In 2nd Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics (BAILAR) on Robot and Human Interactive Communication (RO-MAN), 26th IEEE International Symposium on.*, 2017.

[9] ——, "Deep reinforcement learning using compositional representations for performing instructions." *Submitted to the Paladyn Journal of Behavioral Robotics*, 2018.

[10] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell Me Dave: Context-sensitive grounding of natural language to manipulation instructions," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 281–300, 2016.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[12] T. Shu, C. Xiong, and R. Socher, "Hierarchical and interpretable skill acquisition in multi-task reinforcement learning," *arXiv preprint arXiv:1712.07294*, 2017.

[13] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[14] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "Emorl: Continuous acoustic emotion classification using deep reinforcement learning," *accepted at the International Conference on Robotics and Automation (ICRA)*, 2018.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ICLR*, 2015.

[16] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[17] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[18] J. Gu, G. Neubig, K. Cho, and V. O. K. Li, "Learning to translate in real-time with neural machine translation," in *15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2017.

[19] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," *accepted to the International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[20] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, pp. 1–14, 2015.

[21] M. Kerzel, H. Beik Mohammadi, M. A. Zamani, and S. Wermter, "Accelerating deep continuous reinforcement learning through task simplification," *accepted at the International Joint Conference on Neural Networks (IJCNN)*, 2018.

[22] H. Beik Mohammadi, M. A. Zamani, M. Kerzel, and S. Wermter, "Online continuous deep reinforcement learning for a reach-to-grasp task in a mixed-reality environment," *to be submitted*, 2018.

[23] M. A. Zamani, H. Beik Mohammadi, M. Kerzel, S. Magg, and S. Wermter, "Learning Spatial Representation for Safe Human-Robot Collaboration in Joint Manual Tasks," *accepted in WORKMATE 2018: the WORKplace is better with intelligent, collaborative, robot MATEs on International Conference on Robotics and Automation (ICRA).*, 2018.

Learning peripersonal space in a humanoid robot and its application to safe human-robot interaction

Phuong D.H. Nguyen[1], Matej Hoffmann[2], Ugo Pattacini[1], Giorgio Metta[1]

***Abstract*— The paper presents research to develop the peripersonal space (PPS) representation in robots through a self-supervised learning procedure, which is motivated by the development of perception and motor skills in humans. This representation is constructed by the integration of multisensory data from robots' sensors (stereo cameras, artificial skin and proprioception), and serves as spatial perception of the space surrounding the robot body. A novel approach is proposed to develop this representation through the design of specific motor activities that will make use of, *e.g.* motor babbling and reaching-with-avoidance. We will also show how this representation aims to help the robot accomplish motor tasks in complex situations, such as Human-robot Interaction (HRI). Finally, we will describe the accomplishments and future steps to complete the proposed plan.**

I. INTRODUCTION

The abilities to adapt and act autonomously in an unstructured and human-oriented environment are necessarily vital for the next generation of robots, which aim to safely cooperate with humans. While this adaptability is natural and feasible for humans, it is still very complex and challenging for robots.

Many neuroscientific findings show that there are multisensory integration processes occurring in humans to represent the space close to the body that is termed peripersonal space (PPS) [1]. The PPS serve as a "safety margin" to facilitate objects manipulation [2], [3] and to ease a variety of human actions such as reaching and locomotion with obstacle avoidance [2], [4]. Notably, this is not the case for the far space away from the human body [5]. Moreover, this spatial representation is incrementally trained and adapted (i.e. expanded, shrunk, enhanced, etc.) through motor activities, as reported in [1], [4], [6], and more.

Those results suggest that by exploiting motor activities in exploratory tasks, agents can on the one hand develop their perception of the space around their bodies, and on the other hand use the spatial representation they have built to improve the quality of their motor skills.

The goal of this research is to construct a PPS representation for the upper body of a humanoid robot by leveraging on the repertoire of its motor actions, and then to use such enhanced spatial perception to finally refine the motor capabilities of the robot, especially in cluttered and dynamic environments. Specifically, the proposed research aims to contribute to the understanding and propose models and implementations pertaining to the following points:

- Mechanisms of development and learning of PPS representation from visual, tactile, and proprioceptive information;
- The interaction of motor skills (such as reaching capabilities) and multimodal perception;
- The utility of new adaptive PPS representations in control settings – in particular planning and reaching with simultaneous obstacle avoidance.

The developed models and algorithms will then be validated on the iCub humanoid robot for human-robot interaction in a cluttered environment.

II. RELATED WORKS

Computational models: Serino *et al.* [6] and Maggoso *et al.* [7], [8] analyzed two neural networks to deal with audiotactile and visuotactile stimuli, respectively. They both suggest bio-inspired networks for the PPS representation, and then assign the connection weights that model the neuronal plasticity. The models were only tested without a body and only in a simple static scenario, assuming body parts to be still. Moreover, they have not designed a training procedure, except for the tool-use case presented in [8].

Robotics models: Roncone *et al.* [9] proposed a model to investigate an integrated representation of visual and tactile sensors. The outcome is a visual collision predictor of objects being close to a robot's body, which is constructed by visuo-tactile contingency. This model can be used for a simple reaching/avoidance controller. However, they rely on a well-structured visual tracker for data collection and *a priori* knowledge of a robot kinematic model for frame transformation (between different sensory sources) rather than via autonomous learning.

Antonelli *et al.* [10] and Chinellato *et al.* [11] adopted radial basis function networks to construct the mapping (forward and inverse transformations) between stereo visual data and proprioceptive data by performing gazing and reaching activities. Their mapping requires markers for feature extraction with known disparity, and is apparently beneficial only for multi-sensory transformation and not as a spatial perception of the body's surroundings.

On the other hand, Contla [12] focused on the plastic nature of PPS to account for the modification the body undergoes, and on the impact of this plasticity on the confidence levels of reaching activities. The hypothesis is validated only in a simulated environment. Contla's work

[1]Phuong D.H. Nguyen, Ugo Pattacini, and Giorgio Metta are with iCub Facility, Istituto Italiano di Tecnologia, Genova 16163, Italy {phuong.nguyen, ugo.pattacini, giorgio.metta}@iit.it

[2]Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic {matej.hoffmann@fel.cvut.cz

is mainly concerned with the reachable space of the robot, whereas we focus on the PPS as "margin of safety" instead (see Section I).

The above review makes evident that the current research is very little regarded with building a model of the PPS through self-supervised learning as well as its exploitation to enhance the robot motor capabilities.

III. GENERAL APPROACH

To tackle the research questions, we propose a general approach for the project as follows:

- We evaluate and extend the PPS model of Roncone *et al.* [9] for the Human-robot Interaction (HRI) scenario, where the learned PPS representation serves as a collision predictor against the visually detected obstacles and as an aggregation of physically detected collisions (via tactile sensors). This guarantees the safety for robot's interaction with environments. Also, a robot control system for interaction scenarios needs designing with a master motion planner and a controller;
- We introduce a modulation of the PPS representation for adaptive robot behavior. The modulation can result in expanding or shrinking the "safety margin" depending for example on the properties of the relevant objects in the scene (e.g. fragile, threatening) or on the social context of the interaction. As a result, the robot will be able to interact with human partners in a shared workspace according to different internal states (e.g., relaxed vs. stressed);
- We finally propose a novel PPS model utilizing a neural network to integrate multi-sensory information from the stereo-vision, distributed skin and proprioception, which aims to seamlessly substitute the model of Roncone *et al.* in HRI architecture. The main purpose of the alternative model is to overcome the limitations of the available one (*i.e.* based on *a priori* robot's kinematic model, using visual tracker), and to enable the autonomous action-based learning procedure.

IV. EXPERIMENTS & RESULTS

In this section, we briefly describe our accomplishments in realizing the final aim of the project.

A. Motion planning algorithm for robotic manipulators in dynamic environment

In [13], we present a fast heuristic motion planning algorithm designed for a humanoid robot that employs the sampling-based RRT* algorithm directly in the Cartesian space and in a hierarchical fashion: (i) a collision-free path is planned for the end-effector; (ii) corresponding collision-free points for every via-point are searched for the robot elbow. The method is then validated in diverse scenarios, comprising *batch run-time measurements*, *tests for asymptotic optimality* and *benchmarks against state-of-the-art*.

The results demonstrate that our solution delivers real-time performance (generates path plans in a fraction of second on a standard PC) in the vast majority of cases in a significantly cluttered environment. Second, the results suggest that asymptotic optimality of the plans is preserved even for the additional control points. Third, a comparison with state-of-the-art algorithms on the same scenario shows that solutions cannot be found in reasonable time (less than 10s) when using other algorithms.

This method was applied to the iCub in real settings in the frame of the EU Project WYSIWYD [1] where our method guaranteed collision-free for robots' motion in a table top scenario.

B. Compact real-time avoidance of a humanoid robot for Human-robot Interaction

Taking inspiration from PPS representations in humans, we present a framework on the iCub humanoid that dynamically maintains such a protective safety zone, composed of the following main elements: (i) a visual human 2D keypoints estimation pipeline employing a deep learning based algorithm, extended into 3D using disparity; (ii) a distributed adaptive PPS representation around the robot's body parts, augmented from [9]; (iii) a visually reactive controller that incorporates all obstacles entering the robot's safety zone on the fly into the task (see [14]). The proposed solution is flexible and versatile since the safety zone around individual robot and human body parts can be selectively modulated (*e.g* stronger avoidance of the human head compared to rest of the body). Our system works in real time and is self-contained, with no external sensory equipment and use of onboard cameras only.

Pilot experiments in physical HRI scenario, *i.e.* reaching static target or following a trajectory with human experimenter interfering, demonstrate that an effective safety margin between the robot's and the human's body parts is kept during interaction.

C. Merging physical and social interaction for effective human-robot collaboration

We extended the work in [14] by designing a complete system in [15] (shown in Fig. 1) that merges elements of physical and social HRI, namely:

- A compact human-centered visual perception system for humanoid robots, which can detect human pose, and also recognize and track humans' manipulating objects;
- A simple symbolic "storage" of humans, objects, tools information to support social interaction, which contains the knowledge representations converted from perceived sensory representations of an environment;
- A visuo-tactile reactive controller that exploits the stereo-vision and the artificial skin of the iCub to allow the robot to safely react in both *pre*- and *post*-collision phases corresponding to visual and tactile stimuli respectively.

Through two interaction experiments (*i.e.* human-robot and robot-human object hand-over), we show that the complete system works in real-time controlling the robot's activities while guaranteeing safety for the human experimenter.

[1] wysiwyd.upf.edu

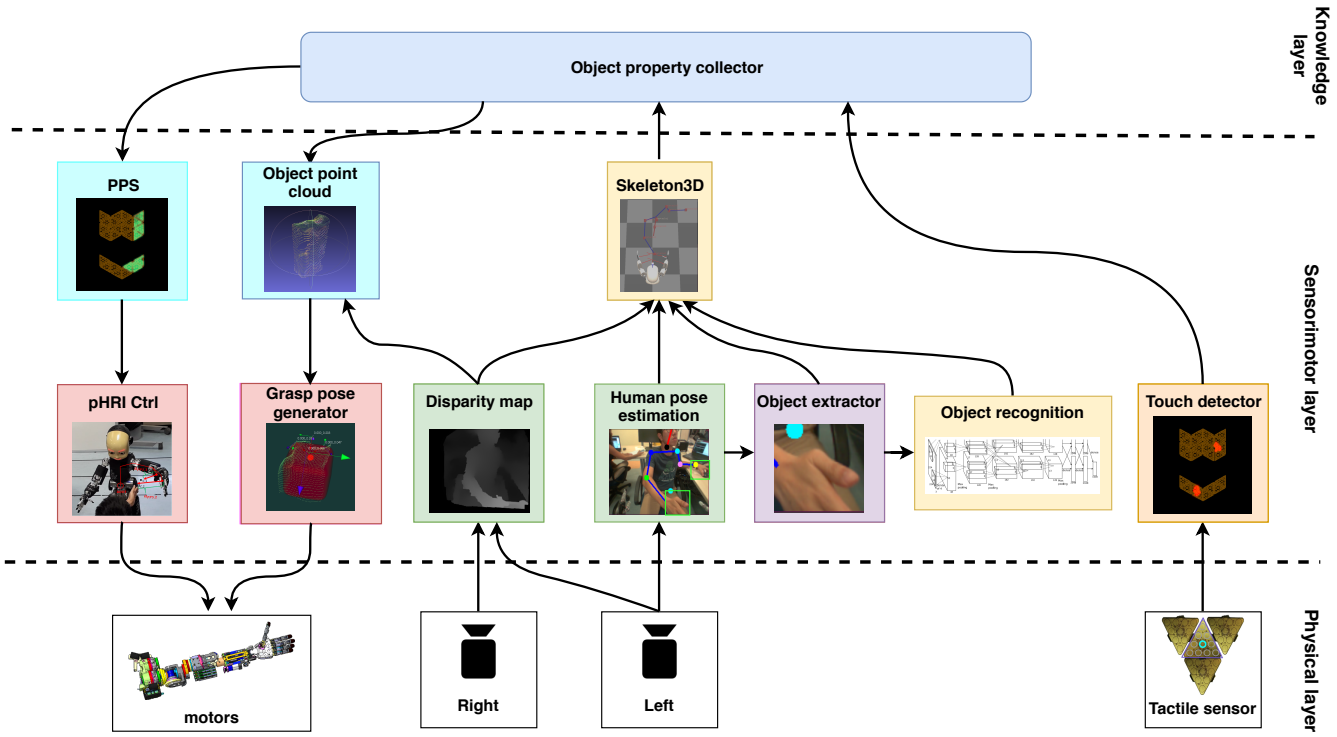


Fig. 1. Overview of the overall system comprising perception (right side) and action (left side) pathways.

The proposed visual perception system was also utilized to replace the wearable sensory suit for human tracking task in an ergonomic and reconfigurable Human-robot Collaboration [16]. The comparison results of tracking experiment (between our vision system and the wearable suit) prove the effectiveness and feasibility of our replacement for industrial application.

D. Learning visuomotor mapping in simulation and transferring to real world for robotics manipulation tasks

Recently, we design a framework to learn the visuomotor mapping in a single step [17] rather than considering the two problems (*i.e.* robot's kinematic modeling and visual-based pose estimation) independently and finding an offset mapping subsequently as in classical approach [18]. More specifically, we suggest to learn the mapping from an imprecise model in simulation using two components (as shown in Fig. 2): (i) A deep neural network (DNN) estimates the arm's joint configuration given images captured with the two eyes of the simulated robot and the corresponding head configuration. (ii) An image-to-image translation method bridges the domain gap to allow application of the DNN in the real world, since the image statistics between simulation and real world differ significantly.

In various experiments, we first show that the visuomotor predictor provides accurate joint estimates of the iCub's hand in simulation, and also can be used to obtain the systematic error of the robot's joint measurements on the physical iCub robot. We demonstrate that a calibrator can be designed to automatically compensate this error, and then validate that this enables accurate reaching of objects while circumventing manual fine-calibration of the robot.

V. CONCLUSIONS & FUTURE WORKS

In this paper, we have proposed a bio-inspired approach (i.e. learning via motor activities) to integrate the multisensory information (i.e. visual, tactile and proprioceptive) forming the spatial perception of surroundings for humanoid robots–peripersonal space representation, and to develop the sensorimotor competences from that enhanced perception. In addition, we have presented our achievements that consists in the design and realization of a *Multiple Cartesian point motion planning algorithm, Visuo-tactile control system for HRI* and *Visuomotor learning framework*, which were all successfully published ([13], [14], [17], [19], [20]) or submitted ([15], [16]).

The successive step will be concerned with extending the visuomotor mapping model [17] to additionally incorporate the tactile input. An action based learning process will also rely on our proposed *motor babbling* method [17], extended such that it can deal with a cluttered environment with randomly allocated obstacles. The simulation environment will be mainly exploited for data collection due to the safety, while domain adaptation methods like domain randomization, image-to-image translation will be used for bridging the reality gap. The resulting model will be used to estimate the spatial and temporal information of possible collisions of the robot's arm with visually detected objects, such that robot's collision-free motion planning can be generated. An advantage of the proposed method is that visual stimuli can be mapped directly into joint space in real-time, where well-established motion planning techniques such as Rapidly exploring Random Trees (RRT*) and Probabilistic RoadMap (PRM*) [21] can be applied.

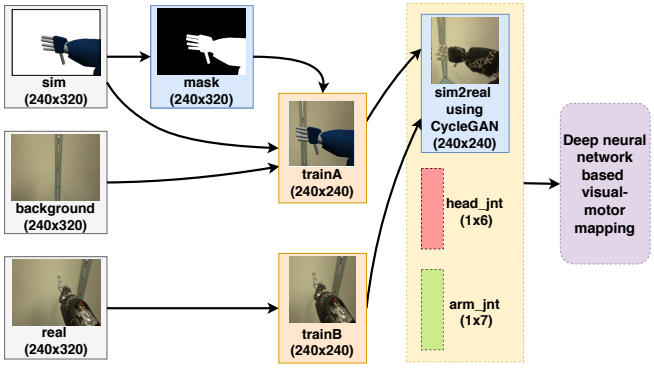


Fig. 2. Overview of the overall learning framework. Images obtained using a simulator are first being implanted with real background, and then CycleGAN [22] is used to synthesize realistically looking "sim2real" images. These are used as inputs to a deep neural network along with the head joints obtained from the simulator. The aim of the deep network is to estimate the arm joint configuration.

ACKNOWLEDGMENT

Phuong D.H. Nguyen was supported by a Marie Curie Early Stage Researcher Fellowship (H2020-MSCA-ITA, SECURE 642667). Matej Hoffmann was supported by the Czech Science Foundation under Project GA17-15697Y.

REFERENCES

[1] J. Cléry *et al.*, "Neuronal bases of peripersonal and extrapersonal spaces, their plasticity and their dynamics: Knowns and unknowns," *Neuropsychologia*, vol. 70, pp. 313–326, Apr. 2015.

[2] N. P. Holmes *et al.*, "The body schema and multisensory representation(s) of peripersonal space," *Cognitive Processing*, vol. 5, no. 2, pp. 94–105, Jun. 2004.

[3] C. Goerick *et al.*, "Peripersonal space and object recognition for humanoids," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, IEEE, 2005, pp. 387–392.

[4] E. Làdavas *et al.*, "Action-dependent plasticity in peripersonal space representations," *Cognitive Neuropsychology*, vol. 25, no. 7-8, pp. 1099–1113, Dec. 2008.

[5] A. Farne *et al.*, "Neuropsychological evidence of modular organization of the near peripersonal space," *Neurology*, vol. 65, no. 11, pp. 1754–1758, 2005.

[6] A. Serino *et al.*, "Extending peripersonal space representation without tool-use: Evidence from a combined behavioral-computational approach," *Frontiers in Behavioral Neuroscience*, vol. 9, Feb. 2015.

[7] E. Magosso *et al.*, "Visuotactile representation of peripersonal space: A neural network study," *Neural computation*, vol. 22, no. 1, pp. 190–243, 2010.

[8] E. Magosso *et al.*, "Neural bases of peri-hand space plasticity through tool-use: Insights from a combined computational–experimental approach," *Neuropsychologia*, vol. 48, no. 3, pp. 812–830, 2010.

[9] A. Roncone *et al.*, "Peripersonal Space and Margin of Safety around the Body: Learning Visuo-Tactile Associations in a Humanoid Robot with Artificial Skin," *PLOS ONE*, vol. 11, no. 10, e0163713, 2016.

[10] M. Antonelli *et al.*, "On-line learning of the visuomotor transformations on a humanoid robot," in *Intelligent Autonomous Systems 12*, Springer, 2013, pp. 853–861.

[11] E. Chinellato *et al.*, "Implicit Sensorimotor Mapping of the Peripersonal Space by Gazing and Reaching," *EEE Trans. Auton. Ment. Dev.*, vol. 3, no. 1, pp. 43–53, Mar. 2011.

[12] S. Ramírez Contla, "Peripersonal Space in the Humanoid Robot iCub," 2014.

[13] P. D. Nguyen *et al.*, "A fast heuristic Cartesian space motion planning algorithm for many-DoF robotic manipulators in dynamic environments," in *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on*, IEEE, 2016, pp. 884–891.

[14] D. H. P. Nguyen *et al.*, "Compact Real-time Avoidance on a Humanoid Robot for Human-robot Interaction," in *The 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2018, pp. 416–424.

[15] P. D. Nguyen *et al.*, "Merging physical and social interaction for effective human-robot collaboration," in *Humanoid Robots (Humanoids), 2018 IEEE-RAS 18th International Conference on*, (submitted), 2018.

[16] W. Kim *et al.*, "A Reconfigurable and Adaptive Human-Robot Collaboration Framework for Improving Worker Ergonomics and Productivity," *IEEE Rob. Autom. Mag.*, 2018, (under review).

[17] P. D. H. Nguyen *et al.*, "Transferring Visuomotor Learning from Simulation to the Real World for Robotics Manipulation Tasks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2018.

[18] J. Hollerbach *et al.*, "Model Identification," in *Springer Handbook of Robotics*, B. Siciliano *et al.*, Eds., 2008, pp. 321–344.

[19] T. Fischer *et al.*, "iCub-HRI: A Software Framework for Complex Human–Robot Interaction Scenarios on the iCub Humanoid Robot," *Frontiers in Robotics and AI*, vol. 5, p. 22, 2018.

[20] C. Moulin-Frier *et al.*, "DAC-h3: A Proactive Robot Cognitive Architecture to Acquire and Express Knowledge About the World and the Self," *IEEE Transactions on Cognitive and Developmental Systems*, vol. PP, no. 99, 2017.

[21] S. Karaman *et al.*, "Sampling-based algorithms for optimal motion planning," *International Journal of Robotics Research*, vol. 30, no. 7, pp. 846–894, 2011.

[22] J.-Y. Zhu *et al.*, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

Learning robust task priorities of optimization-based whole-body torque-controllers

Marie Charbonneau[1,2], Valerio Modugno[2,3], Francesco Nori[4], Giuseppe Oriolo[3], Daniele Pucci[1] and Serena Ivaldi[2]

***Abstract*— The ability for a humanoid robot to safely evolve within a human environment is currently an important topic of research. Generating robust whole-body movements is still an open challenge, especially in contexts where a robot may physically interact with people and objects. Generating complex whole-body movements for humanoid robots is now most often achieved with the use of multi-task whole-body controllers based on optimization or quadratic programming. To perform on a real robot, however, such controllers often require a human expert to tune or optimize the many parameters of the controller related to the tasks and to the specific robot. This problem can be tackled by automatically optimizing some parameters such as task priorities or task trajectories, while ensuring constraints satisfaction, through simulation. This approach however does not guarantee that the optimized parameters in simulation will be optimal also for the real robot. As a solution to help bridge this reality gap, the present paper focuses on optimizing task priorities in a robust way by looking for solutions which achieve desired tasks under a variety of conditions and perturbations. This approach, which can be referred to as domain randomization, can greatly facilitate the transfer of optimized solutions from simulation to a real robot. The proposed method is demonstrated using the humanoid robot iCub for a whole-body stepping task.**

I. Introduction

Applications involving humanoid robots have the potential to bring significant benefits to society. Nevertheless, the design of controllers for humanoid platforms is highly challenging, especially when robots are expected to physically interact with people or the environment.

A promising approach is to use whole-body torque-control methods, which decompose a desired complex behavior into several simple tasks, typically framed as a *stack-of-tasks* [1]. Such a framework requires the tasks to be hierarchized, either in a *strict* or a *soft* way. In strict prioritization strategies, a fixed task hierarchy is ensured by geometrical conditions (e.g. null space task projector) or by the use of optimization constraints [2], [3]. Conversely, soft task prioritization can be achieved by assigning each task a weight defining its relative importance [4]. However, in the case of complex problems, such as whole-body motion of a humanoid robot, the design and proper tuning of task priorities may not always be evident, making it tedious and time consuming.

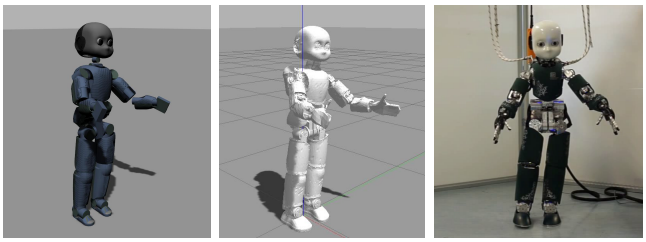


Fig. 1: Different robot models performing a whole-body motion with several tasks. We optimize task priorities for robustness, with the purpose to allow their transfer from the first model to the second, and eventually to the real robot.

A recent line of research seeks to tackle the issue of automatically learning whole-body task priorities [5], [6]. Since learning algorithms need a considerable number of iterations and use a random exploration which could harm hardware, they are usually applied in simulation. However, inherent differences between simulated and real robots can render an optimal solution untransferrable from one to the other. Closing this reality gap is the central focus of recent works in robotics [7] and related fields. One approach, Domain Randomization (DR) [8], consists in randomizing some aspects of the simulation to enrich the range of possible environments experienced by the learner. For example in [9], control policies are learned in simulation, given random friction and control delays, and results showed that the learned policies were also effective on the real robot. As a result, it appears that looking for solutions which are *robust*, in opposition to *optimal*, may allow to bridge the reality gap.

This work proposes a method to learn robust task priorities which achieve compliant and stable whole-body motions, while allowing to facilitate the transfer of results from simulation to reality by taking advantage of the DR approach. The effectiveness of the proposed method is demonstrated by optimizing parameters in simulation, and showing that it is possible to overcome issues stemming from large differences between the learning domain and the testing domain.

II. Methods

The method proposed for learning robust task priorities relies on two main parts: (i) an optimization-based whole-body torque-controller which tracks desired task trajectories

This project has received funding from the European Unions Horizon 2020 framework programme for research and innovation under the Marie Sklodowska-Curie Grant Agreement No.642667 (SECURE) and the ICT Grant Agreement No. 731540 (An.Dy).

[1]iCub Facility, Istituto Italiano di Tecnologia, Genova, Italy. name.surname@iit.it

[2]Inria Nancy - Grand Est, Team LARSEN Villers-lès-Nancy, France. name.surname@inria.fr

[3] Dipartimento di Ingegneria Informatica, Automatica e Gestionale, Sapienza Universita di Roma, Roma, Italy surname@diag.uniroma1.it

[4]Google DeepMind, London, UK. fnori@google.com

and sends joint torque commands to the robot, and (ii) an optimization method as described in [10], which poses no restrictions on the structure of the learning problem. Task priorities are then optimized at the end of an experiment (i.e. execution of a footstep): the fitness of the obtained trajectories is evaluated, allowing to update the task weights.

The controller assumes the modelling of the robot as described in [3], and the control input u to be composed of joint torques τ and contact forces F_C. A stack of tasks is defined with the objectives to stabilize the center of mass position X_{CoM}, stance and swing feet pose X_{stance} and X_{swing}, neck orientation X_{neck}, joint positions s, as well as to minimize joint torques τ. The torque-controller used in this paper was developed in previous works, and described in [11]. Here, the controller is used with the following optimization problem using soft task priorities:

$$u^* = \arg\min_u \ \frac{1}{2} \text{ cost} \tag{1a}$$

$$\text{subject to } Cu \leq b \tag{1b}$$

where the constraint (1b) ensures that the contact forces remain within the associated friction cones. The cost function (1a) is computed as the weighted sum of all task objectives:

$$\text{cost} = \sum_T w_T \left| \tilde{\ddot{X}}_T(u) \right|^2 + w_s \left| \tilde{\ddot{s}}(u) \right|^2 + w_\tau \left| \tau(u) \right|^2 \tag{2}$$

$\tilde{\ddot{X}}_T(u)$ and w_T are acceleration errors and weights associated to each Cartesian task T (CoM, stance, swing and neck), while w_s, w_τ are the weights of the postural task and joint torque regularization.

III. Experiments

A series of experiments were performed in order to validate empirically the hypothesis that the method described above is capable of optimizing task priorities, in such a way as to (i) allow the generation of robust whole-body motions, even when contacts due to physical interaction with the environment evolve in time and (ii) be able cope with imperfections in the robot model, disturbances, and noise.

Experiments were conducted in simulation using the open-source robot simulator Gazebo. They were performed with the iCub robot, using 23 DOF on legs, arms and torso, for whole-body torque control. The design of iCub has evolved over the years, which has a significant impact on the inertial properties of the robots. For instance, some models of iCub have tethered power supply, while others have battery packs installed on the back of the torso. This gives us a chance to test our method on different robot models.

The controller described in II was developed in Matlab/Simulink, allowing to control the motion of either a simulated or a real robot. It is applied here to the problem of performing a step, i.e. lifting the foot off the ground and placing it back on the ground.

The experimental procedure was divided into two main parts: (i) training task priorities with a first model of iCub, and (ii) validating the obtained task priorities with a different model of iCub.

1) Training with a first iCub model: First, task priorities were optimized on a simulated tethered iCub model, as shown in the left part of fig. 1, performing a whole-body movement (one step). The fitness function ϕ_{pr} was evaluated in 10 separate learning experiments, in order to optimize task priorities.

$$\phi_{pr} = \frac{1}{2}(\phi_p + \phi_r) \tag{3a}$$

$$\phi_p = -\frac{1}{P_{ZMP_{max}}} \sum_{t=0}^{t_{end}} \left| P_{ZMP} - O_{SP} \right|^2 \tag{3b}$$

$$\phi_r = -\frac{1}{X_{T_{max}}} \sum_{t=0}^{t_{end}} \sum_T \left| \tilde{X}_T \right|^2 - \frac{0.0001}{\tau_{max}} \sum_{t=0}^{t_{end}} \left| \tau \right|^2 \tag{3c}$$

This particular fitness function, ϕ_{pr}, favors robust solutions with ϕ_p by encouraging smaller excursions of the ZMP position P_{ZMP} with respect to the center of the support polygon O_{SP}. On the other hand, the term ϕ_r seeks to maximize performance on the Cartesian tasks with a minimal effort. In these equations, $X_{T_{max}}$, τ_{max} and $P_{ZMP_{max}}$ are normalization factors. In case the robot was unable to accomplish a full step, a penalty of -1.5 is added to ϕ_{pr}.

In addition, the robot was subjected to random sets of conditions during training, in order to achieve robustness through DR. For each learning iteration, the following conditions were randomized: Gaussian noise on input F/T sensor signals, swing foot, motion of the swing foot, displacement of the CoM, and a random number of random external wrenches applied to the chest. The external wrenches not only served to increase the robustness of the controller, but also to promote the soft behavior of the robot in case of physical interaction with people, while still keeping balance.

Having been verified to allow the first iCub model to successfully perform the desired stepping motion, the following hand-tuned task priorities were used as a starting point for the optmization:

$$w_{CoM} = 1 \tag{4a}$$

$$w_{stance} = 1 \tag{4b}$$

$$w_{swing} = 1 \tag{4c}$$

$$w_{neck} = 0.1 \tag{4d}$$

$$w_s = 0.001 \tag{4e}$$

$$w_\tau = 0.0001 \tag{4f}$$

Then, optimized task priorities were obtained by performing 200 learning iterations with applied to the control framework, with an exploration rate of 0.1. The optimization procedure was repeated for 10 separate trainings, allowing to verify the consistency of the method.

2) Testing with a second iCub model: In order to validate the robustness achieved with the optimized task priorities, while attempting to replicate conditions similar to performing experiments on the real robot, each one of the resulting 10 sets of optimized task priorities was tested on an iCub model with a battery pack on the back, as shown in the middle part of fig. 1. The robot was made to perform a sequence

TABLE I: Optimized task priorities: mean and standard deviation obtained from 10 different training experiments

weight	mean	std deviation
w_{CoM}	1	0
w_{stance}	0.9	1.3
w_{swing}	2.4	1.1
w_{neck}	0.6	1.2
w_s	1e−6	0
w_τ	1e−10	0

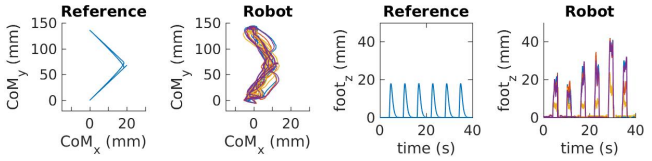


Fig. 2: Typical CoM and feet trajectories for 6 strides performed with the second iCub model. Each color denotes the use of a different set of optimized weights. The x, y and z axes correspond to the sagittal, frontal and vertical axes.

of whole-body movements (6 steps), under different noise conditions as those used for training. It was subjected to external wrenches on the chest, as well as Gaussian noise on the F/T sensor and joint velocity measurements.

IV. Results

The mean and standard deviation of the optimized task priorities, as obtained with the experiments explained above, are shown in table I.

These task priorities, when used with the controller described in Sec. II, allowed the first robot to perform one step, under the conditions used for training. They also successfully allowed the second robot model to perform 6 steps, under the noise conditions mentioned previously, with a success rate of 100%. In comparison, the starting task weights defined in 4 did not prove to be successful, showing that the optimized weights did improve the effectiveness of the controller.

The CoM and feet trajectories achieved with the optimized task priorities on the second robot model, illustrated in Fig. 2, show convergence of the robot motion. These results demonstrate that the optimized weights allow for a higher robustness of the controller.

V. Discussion and Conclusions

In summary, the proposed method can be used to generate robust task priorities for whole-body torque-control of humanoids. It was demonstrated by performing training on a first robot, then testing on a second model with different physical properties and working conditions.

A fitness function combining robustness and performance has shown to allow the obtention of sensible task priorities. In the achieved results, swing foot placement, crucial for stability at touchdown, is given high importance, while the neck orientation task a lesser one, allowing compliance to external perturbations (i.e. physical interactions with the environment, such as the impact of the foot on the ground). As for the postural task, its low priority allows it to be used as regularization (just as joint torques), instead of competing with Cartesian tasks.

Such a solution is interesting, as it may not have been *a priori* self-evident to an expert defining task priorities. Furthermore, the ranges over which sets of optimized weights were obtained show that although task priorities require proper tuning, the controller is not highly sensitive to a precise adjustment of task weights.

Finally, the proposed method has shown to achieve compliant and stable behaviors with a robot model different than the one used for learning, and subjected to diverse working conditions. The robustness achieved in this way is promising and could allow higher success when passing from simulation to real-world experiments. Upcoming work shall provide a more extensive analysis of the method, comparing results obtained with different fitness functions, as well as with and without domain randomization, in order to asses the contribution of fitness parameters and DR to the success of the method. Our approach shall also be tested with experiments on the real robot.

References

[1] M. A. Hopkins, D. W. Hong, and A. Leonessa, "Compliant locomotion using whole-body control and divergent component of motion tracking," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 5726–5733.

[2] L. Saab, O. E. Ramos, F. Keith, N. Mansard, P. Souères, and J. Y. Fourquet, "Dynamic whole-body motion generation under rigid contacts and other unilateral constraints," *IEEE Transactions on Robotics*, vol. 29, no. 2, pp. 346–362, April 2013.

[3] G. Nava, F. Romano, F. Nori, and D. Pucci, "Stability analysis and design of momentum-based controllers for humanoid robots," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 680–687.

[4] J. Salini, V. Padois, and P. Bidaud, "Synthesis of complex humanoid whole-body behavior: A focus on sequencing and tasks transitions," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 1283–1290.

[5] N. Dehio, R. F. Reinhart, and J. J. Steil, "Multiple task optimization with a mixture of controllers for motion generation," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2015, pp. 6416–6421.

[6] S. Ha and C. Liu, "Evolutionary optimization for parameterized whole-body dynamic motor skills," in *ICRA*, 2016.

[7] D. Clever, M. Harant, K. D. Mombaur, M. Naveau, O. Stasse, and D. Endres, "Cocomopl: A novel approach for humanoid walking generation combining optimal control, movement primitives and learning and its transfer to the real robot HRP-2," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 977–984, 2017.

[8] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 23–30.

[9] R. Antonova, S. Cruciani, C. Smith, and D. Kragic, "Reinforcement learning for pivoting task," *CoRR*, vol. abs/1703.00472, 2017.

[10] V. Modugno, G. Nava, D. Pucci, F. Nori, G. Oriolo, and S. Ivaldi, "Safe trajectory optimization for whole-body motion of humanoids," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, November 2017, pp. 763–770.

[11] S. Dafarra, G. Nava, M. Charbonneau, N. Guedelha, F. Andrade, S. Traversaro, L. Fiorio, F. Romano, F. Nori, G. Metta, and D. Pucci, "An online predictive kinematic planner for position and torque controlled walking of humanoid robots," 2018, submitted for publication.

Dense 3D Environment Reconstruction with an RGB-D Camera for Mobile Robot

Chih-Hsuan Chen

***Abstract*— In this paper, we present an environment reconstruction system to generate an indoor 3D map for mobile robots. Using an RGB-D sensor, the robot doesn't need the initial odometry. Furthermore, the system can be used for reconstruction of a 3D environment model by manually. We optimize our approach to reach 10Hz for the front-end and 1Hz for the back-end to fulfill the applications for the mobile robot. Our final goal is to develop the 3D SLAM systems which combine the Region Based Convolution Neural Network (MASK R-CNN) for creating a 2D semantic image and a 3D semantic map. The experimental results demonstrate some preliminary results for 3D reconstruction with an RGB-D camera for creating the point cloud map and the OctoMap for the mobile robot (Care-O-bot 4) in an indoor environment.**

I. Introduction

To explore an unknown indoor environment, a mobile robot needs to create a map and localize itself in the map simultaneously. This procedure, called simultaneous localization and mapping (SLAM), is challenging and difficult to deal with visual SLAM. The major challenge with visual SLAM is due to the uncertainty of measurements, varying light conditions, and noise from the sensor. The camera can be described the estimated poses of the robot from RGB-D data can create a 3D model of an indoor environment at the same time.

Mobile robots typically use wide range sensors such as 2D laser scanners for measuring the indoor environment with very high accuracy. The state-of-the-art laser-based SLAM (simultaneous localization and mapping) are known as [1] [2]. To estimate of a camera on the robot motion is known as visual odometry[3].

In this paper, we present an approach with the small improvement to build a 3D map and localize in the map simultaneously based on RGB-D data illustrated in Fig. 2. The 3D environment reconstruction system can be slipped into 3 parts, Front-End, Back-End, and mapping.

In the Front-End part, it can be divided into three subfunction, which includes feature extraction, feature matching and pose estimation. The features from RGB images around the corner and edge can be extracted. Once we collect all features, it can be applied these features key-points for matching with the pervious image. In our approach, we selected the Oriented FAST and Rotated BRIEF (ORB) feature extraction. Based on the features matching results, we can estimate the 3D poses of the any two corresponding frames using Efficient Perspective-n-Point (EPnP) [8]. Therefore, the robot is evaluated the transformation of each frames based these correspondences.

Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
chih-hsuan.chen@ipa.fraunhofer.de

Fig. 1: The RGB-D Camera mounted on the mobile robot Care-O-bot 4 [15]

As we mentioned the major problem in previous section, it will be accumulated the estimation error and cause the accumulating drift problem. In order to resolve this problem, we need to optimize the pose estimates between frames. In the Back-End part, the approach is applied General Graph Optimization (g2o) library which is open source framework for optimizing nonlinear error functions [14] to reduce the accumulating drift ,and the approach is applied loop closure for detecting the previous scene to provide optimizing loop[12]. For the mapping part, the point cloud map (PCL) and OctoMap[13] are utilized to express the 3D environment reconstruction. The preliminary results for 3D reconstruction are presented in Section IV. Finally, we conclude with some future works in Section V.

II. Related Work

A classical approach to visual SLAM, the Mono SLAM is the first real-time monocular visual SLAM system proposed by A.J. Davison [4]. MonoSLAM is using EKF filter as backend and using sparse feature extraction as frontend.

The Parallel Tracking and Mapping (PTAM) is proposed by Klein [5]. It achieves not only the tracking and mapping parallel, but it also introduces the nonlinear optimization instead of traditional optimization such as EKF filter or particle filter. After PTAM, many kinds of research in the field of visual SLAM are using nonlinear optimization as a backend.

The class of algorithms known as iterative closest point (ICP) [10], minimize the distance between two sets of point cloud, which can be generated from two raw scans. The ICP is applicable when we have in good initial guess, otherwise it is likely to be stuck into a local minimum.

ORB-SLAM [6] is known as backend and inherited from PTAM. Comparing with PTAM, there are several advantages for visual SLAM. It supports three types consisting of RGB-D cameras, stereo camera and Monocular camera. Instead of using Scale-invariant feature transform (SIFT) or speeded-up robust features (SURF) feature extraction, the frontend of ORB-SLAM is using ORB feature extraction. It could reduce the computation time and also improve consistency with rotation and zooming. It also uses loop closure to decrease the accumulating error from pose estimation.

III. 3D Reconstruction

The approach of 3D visual SLAM system consists of third main part, Front-End, Back-End, and mapping. The system architecture is illustrated in Fig. 2.

A. Front-End

Our implementation of the front-end consists of the third parts, feature extraction, feature matching and pose estimation. We are mainly using functions from OpenCV [7]. First, we extract ORB features which based on the FAST detector and the BRIEF descriptor proposed by Rublee et al [9], which can be determined landmarks by extracting descriptor vector from RGB image. Once we have the descriptors, feature matching will become a very critical port. Generally, it solves the data association for the landmarks by providing a measure for similarity in visual SLAM system. To match a pair of descriptors, we use Fast Library for Approximate Nearest Neighbors (FLANN) method [7] in case of large of matching point, it takes lower computation time than using brute-force matcher [7].

After we have feature matching results, we can utilize the widely known Random Sample Consensus (RANSAC) [8] for estimating ego-motion. Generally, the model is evaluated by measuring the error for each pose. Consequently, this separates the dataset into two subsets. The inliers can be fitting to the model and the outliers should be ignored. We also propose to use the keyframe to express the most representative frame for Back-End, loop closure, and mapping.

B. Back-End

The estimated ego-motion from front-end comes with an accumulating drift. The back-end of the SLAM system is dealing with the noise problem.

To minimize the error, the graph optimization bases on constraints between the nodes. We introduce the loop closure detection without making an assumption on the path. It is possible to check if the current frame matches with previous ones. The observation of a common point is seen in the past. It can trigger the new link between two poses that were separated. Once the graph has been initialized with the poses

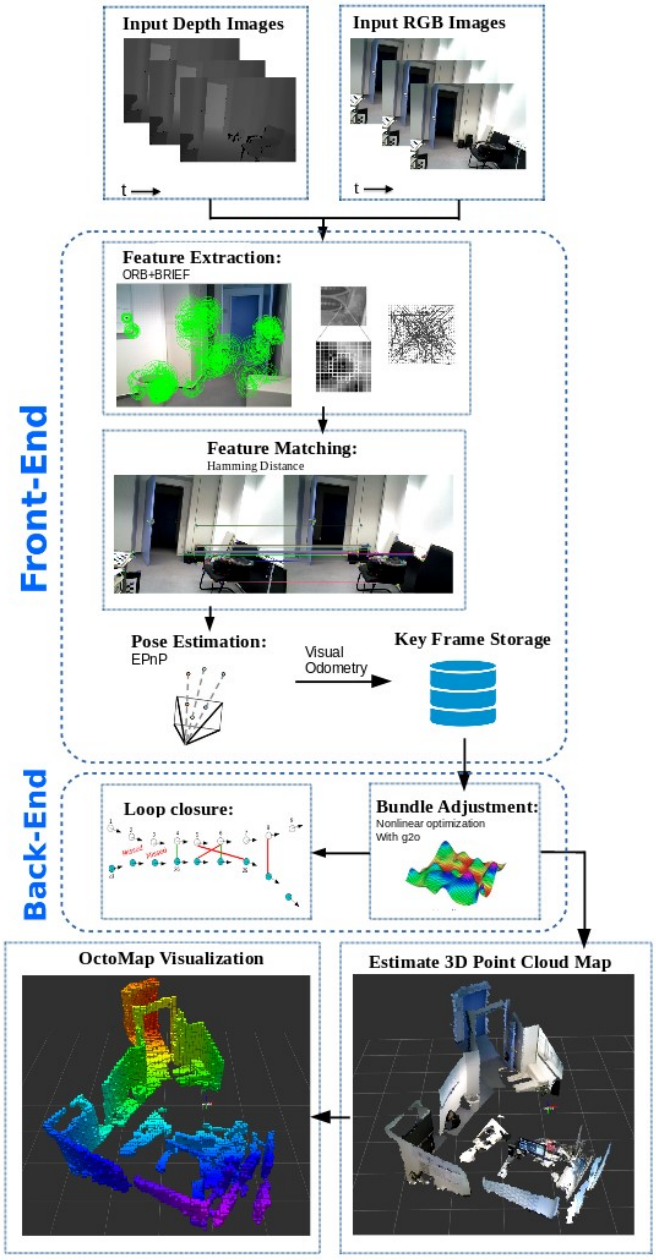
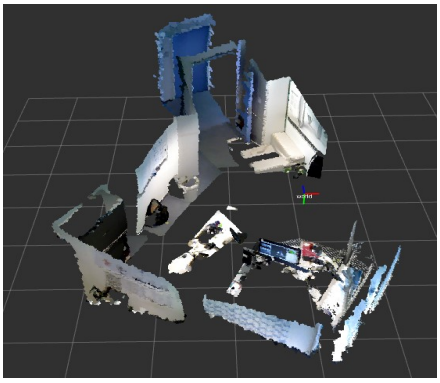


Fig. 2: System Architecture Diagram: Processing of the RGB-D data

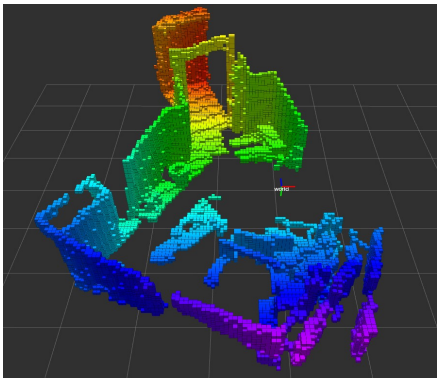
and the constraints from the loop closures, it can trigger the optimization. To resolve bundle adjustment, the VSLAM can be defined as a least squares optimization of an error function, and can be described by a graph model. We use g2o which is an open-source library for the optimization process and to minimize the error. This is the method chosen to solve the graph problem.

C. Mapping

At the end of 3D environment reconstruction, the overall map can be built from the sequence of data. We exploit two methods to represent the map, one is the 3D point cloud map, and other is the OctoMap which can be used to overcome the limitations of point cloud representation with reduced memory resource.



(a) Mapping after graph optimization



(b) Post-processing Octomap

Fig. 3: Map representation for our approach

We use the octree-based framework Octomap in an efficient tree structure that requires less memory consumption than PCL map, but the resolution of Map will also decrease. The figure illustrates how the RGB data and depth information can be used to compute the sequence of 3D transformations, and estimation of the robot poses. Subsequently, the system can be created both the Point-Cloud Map and OctoMap.

IV. Results

In this section, we show the preliminary experiential results illustrated in Fig.3. The data stream acquired from an RGB-D camera. Our approach computes the 6 DoF robot poses including trajectory and orientation and conduct a 3D map. The 3D reconstruction for indoor environment uses the Care-O-bot 4. The input data consisting of RGB and depth information for our approach is captured from an Asus Xtion camera, which is mounted on the robot. The map obtained from office and lab at Fraunhofer IPA. The preliminary results are running on XMG notebook with Intel Core i7-6820 4-cores. All software packages were developed using ROS indigo with Ubuntu 14.04, and OpenCV 2.

Fig 3(a) shows the result after graph optimization tracking created by point-cloud map with loop-closure. Thought the results are satisfying for small drift. Fig 3(b) shows the 3D Octomap after post-processing. The Octomap is valuable for exploration and robot navigation tasks.

V. Conclusion and Future Works

In this paper, we presented an approach for dealing with 3D environment reconstruction for mobile robot applications. We use a feature-based 3D SLAM approach with graph optimization to achieve the 3D reconstruction of an indoor environment. In the future, we are planning to combine the region based convolution neural network (MASK R-CNN) [11] for creating 2D semantic images and 3D semantic point cloud maps. The maps are allowed to provide more information for interacting with the indoor environment. Further results and experiments will be also integrated and tested with Care-O-bot 4.

Acknowledgment

The author has received funding from the European Union's Horizon 2020 framework programme for research and innovation under the Marie Sklodowska-Curie Grant Agreement No. 642667 (SECURE) and gratefully acknowledges the support.

References

[1] Hauke Strasdat, J. M. M. Montiel, and Andrew J. Davison. Scale Drift-Aware Large Scale Monocular SLAM. In Matsuoka, Yoky and Durrant-Whyte, Hugh F. and Neira, José, editor, Robotics: Science and Systems. The MIT Press, 2010.

[2] Kurt Konolige and Motilal Agrawal. FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. IEEE Transactions on Robotics, 24(5):1066–1077, 2008.

[3] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2004

[4] A.J Davsion and I. Reid, et. al:"Real-tiem single camera SLAM", IEEE Transactions on pattern Analysis and Machine Intelligence, vol. 20, no. 6 pp. 1052-1067, 2007

[5] G. Klein and D. Murray, "Parallel tracking and mapping for small at workspaces," in Mixed and Augmented Realty, 2007. ISMAR 2007. 6th IEEE and ACM international Symposium on , pp.225-234, IEEE, 2007.

[6] R. Mur-Artal, J. Montiel, and J.D tardos, "Orb-slam: a versatile and accurate monocular slam system, " arXiv:1502.00956, 2015.

[7] G. Bradski and A. Kaehler. Learning OpenCV: Computer Vision with
the OpenCV Library. O'Reilly Media, 2008.

[8] D. Nister, "Preemptive RANSAC for live structure and motion estimation," in Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV), 2003

[9] C. Wu, "SiftGPU: A GPU implementation of scale invariant feature
transform (SIFT)," http://cs.unc.edu/~ccwu/siftgpu, 2007.

[10] A. W. Fitzgibbon, “Robust registration of 2d and 3d point sets,” Image Vision Comput., vol. 21, no. 13-14, pp. 1145–1153, 2003.
[11] K. He, G. Gkioxari, P. Dollar, R. Girshick, “Mask R-CNN, ” IEEE International Conference on Computer Vision, 2017.
[12] F. Endres, J. Hess, J. Sturm, D. Cremers, W. Burgard, “3D Mapping with an RGB-D Camera, ” IEEE Transactions on Robotics, vol. 30, no. 1, 2014.
[13] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, “OctoMap: An efficient probabilistic 3D mapping framework based on octrees,” Autonomous Robots, 2013.
[14] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, W. Burgard, “g2o: A General Framework for Graph Optimization, ” IEEE International Conference on Robotics and Automation, pp. 3607–3613, 2011.
[15] Care-O-bot 4, https://www.care-o-bot.de/en/care-o-bot-4.html

Design of a robotic finger combining a linkage-based design and the push-pull cable technology*

Alexis Billier[1]

Abstract— **This paper will study the architecture of two robotic fingers. Both of these architectures consist of a linkage-based approach. The first architecture inspired by the work of the University of Laval consists of a three phalanges finger with 3 Degrees of Freedom (DoF) actuated by one motor. The second one is inspired by the results of DeTop, a research project funded by H2020, consists of two phalanges and two DoF, the last two phalanges, Intermediate and Distal are fused in one unique phalanx.**

I. INTRODUCTION

The development of Human-Robotic Interaction (HRI) shall pass by the development of robots that guarantee the safe behavior in physical interaction with humans and the external environment. These robots need to adapt to an open dynamic environment, to help and interact with humans workers, and to manipulate human designed tools. To achieve these objectives, the hands play an important role. They are the frontier between the robot and the external environment. As humans adapted the tools, the daily objects, to their hands, the easiest solution for using these objects is to mimic the human hand. Moreover, for a robot in direct contact with humans, the human-like aspect and behavior are required, to gain the trust and the approval of the humans as Siciliano explained[1].

A human finger is articulated by different tendons and muscles, as described by Schwarz[2], the tendons, usually, one for extension and one for the flexion, are connected to muscles to actuate the finger. In humanoids hand, this architecture is often an inspiration, especially in the cable-driven approach, for example, the hand of ICub[3]. The cables are replacing the tendons and the motors are replacing the muscles. The main problems are the space required and the number of actuators used to move the fingers. Another architecture mainly utilized is the link bar approach such as the DLR/HIT Hand II[4]. This architecture has the advantage to be more robust than the tendon driven architecture.

This paper proposes to combine both of these architectures. The finger consists in a bar linkage, but it is actuated by a cable. Another particularity is that a push/pull cable drives the finger.

To achieve these objectives two architectures will be studied, the first one inspired by the work of Laval [5] consists in a 3 DoF finger, the second one inspired by the work of DeTop, the SSSA-MyHand [6], consists of a 2 DoF cross-bar mechanism.

II. APPROACH

This section details the main objectives that should be fulfilled by the finger. The main one is to mimic the aspect of a human hand; the second is to be robust and safe for the HRI.

A. Hand anatomy

To achieve the first objective, it is important to study the anatomy of a human finger. It is composed of four bones:

- The distal phalanx
- The medial phalanx
- The proximal phalanx
- One additional bone is located in the palm: The metacarpal bone

The muscles that action the fingers are located in the forearm, and they can move the fingers thank the tendons. There are three tendons that move the fingers:

- Deep and superficial flexor tendons
- Extensor tendon

Figure 1 shows the anatomy of a human finger. The architecture is inspired by it. Both architectures mimic this schema. However, in the second architecture, the two last bones of the finger, the medial and the distal phalanx, are fused in one.

*This work was supported by the Innovative Training Network SECURE, funded by the Horizon 2020 Marie Skłodowska Curie Actions (MCSA) of the European Commission (H2020-MSCA-ITN-2014- 642667)

[1]A. Billier is with Danieli TelerobotLabs, Via Buccari 9, 16153 Genova, Italy a.billier@danieli.tlabs.it

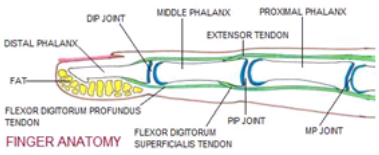


Fig. 1. Anatomy of the human finger

B. bar mechanism

For the finger architecture, the approach is to use a bar mechanism architecture. The main advantage of this method is the stiff transmission, thus the control of the finger is more precise. Usually, the motor that drives the finger is located in the palm of the hand, such as the DLR/HIT Hand II[4]. This configuration increases the weight of the hand, and, A.De Santis and al.[7] advises to limit the weight in moving member to increase the safety.

One solution is to keep this bar mechanism, but move the motor outside the palm, in the forearm; the advantage of the bar mechanism is conserved and the weight of the motor will be displaced in the forearm, increasing the safety.

C. push/pull cable

As the motors are located in the forearm, the movement must be transferred to the fingers. The chosen solution is to use a push/pull cable. In this way, only one motor and only one cable are used for both the extension and the flexion of one finger. The problem is the flexibility of the cable during the push phase. To limit it, the cable shall pass through a sheath.

The full design of the hand and palm, including the connection between the forearm and the finger, is under progress. The final design has not yet be chosen. However, a study was conducted about the configuration of two type of mechanisms bar for the finger: The Laval architecture and the Detop architecture.

III. Laval Architecture

A. Concept

The Laval hand [8] consists in a 4-phalanges finger, and three DoF. Figure 2 shows this architecture.

As explained by T.Laliberté and C.M. Gosselin in [9] the first parameters to choose are the lengths of the different phalanges, these lengths are chosen according to the existent lengths of the ICub conception, i.e. $l = 25.9mm; k = 22mm; j = 19mm$. Then the lengths of c_i are selected as the minimum possible, in our case $c_i = 5mm$. Then a ratio is selected: $R_i = a_i/c_i = 1.5$.

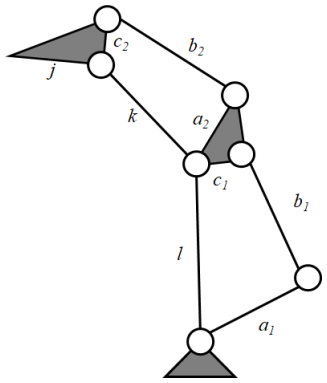


Fig. 2. The architecture of the laval design.

So $a_i = 7.5mm$. Another computation gives the lengths b_i: $b_1 = 25.78mm; b_2 = 21mm$..

The finger is actuated thanks to a crank system. The final architecture can be seen in figure 3.

The part number one is the metacarpal phalanx, the number two is the proximal, the number three is the medial, the number four is the distal, and the number five is the crank system.

B. Component

The main actuation components of the finger are made of bronze. This material allows a frictionless joint without using bearings. It gives the advantage of using smaller pins, thus allowing more space for the sensor. These metallic parts are covered by a 3D printed cover to give the shape of the finger. In the future, a tactile skin will cover the whole finger.

For the experiment, the actual prototype is actuated manually. The finger is attached to the palm and the cable will go through a sheath to be moved at the end. The cable is 1.5 mm diameter and the sheath is 1.8 mm diameter, this allows the movement of the cable through the sheath, and in the same time, it also limits the fold of the cable.

C. Experiment

As the prototype is not yet built, there is no physical experimentation. However, some digital simulations showed that the displacement of the cable between a full closure and a full opening is 26.8 mm. One of the main problems is the independence of the link in

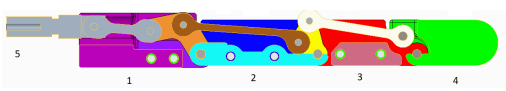


Fig. 3. The simulation of the finger, with a section in the middle.

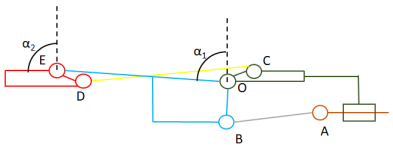


Fig. 4. The architecture of the DeTop design.

the finger. The only way to know the exact position of each link is to implement sensors in the finger. A type of sensors that can be used is magnetic sensors.

IV. DeTop Architecture

A. Concept

This finger consists of three phalanges and two DoF finger. Contrary to the previous architecture, the last two phalanges, middle and distal, are fused in one. Figure 4 shows this architecture.

The lengths of the phalanges are the same. One of the objectives of this finger was to make the pivot axis as close as possible to the internal surface of the finger. In this way, the contact surface varies less during closure and opening and thus permits the pose of a tactile skin.

For actuation, a crank system is also used. The final architecture can be seen in figure 5. The part number one is the metacarpal phalanx, the number two is proximal, the number three is the medial and distal, and the number four is the crank system.

B. Component

The main components are 3D printed for fast prototyping. Except the yellow part is in bronze. There is no need of sensor as the movements of the phalanges are linked. There is a relationship between the two joints, and the knowledge of the position of the crank gives the exact position of the finger.

As previously the actuation is done manually and uses the same system of cable and sheath.

C. Experiment

As the prototype is not ready yet, there is an only digital simulation of the finger. Yet it showed some relationship between the two main angles α_1 and α_2.

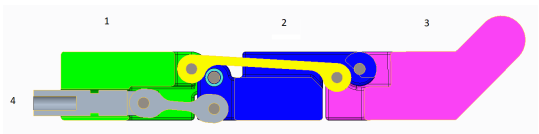


Fig. 5. The simulation of the finger, with a section in the middle.

The displacement of the cable between closure and opening is only 6mm.

V. Future work

In the future, some investigation will be done with both prototypes. One experimentation will be the force and the friction to move the finger. Another works need to be done is all the actuation and implantation of the hand. As the motor gives more power during the pull phase than the push phase and now the pull phase serves to open, the movement should be inverted as more force is required during the closure.

VI. Conclusions

This paper was an introduction to two different design of a linkage-based finger. The combination of the push/pull cable and the linkage-based mechanism is newly utilized. Future investigations are needed to develop it.

The Laval architecture has the advantage to be more realistic with 3 DoF. However, it needs sensors to know the exact position of the finger. The DeTop architecture is simpler and has the advantage to have a displacement of the cable four time smaller.

References

[1] Bruno Siciliano and Oussama Khatib. *Springer handbook of robotics*, chapter 15. Springer, 2008.

[2] Robert J Schwarz. The anatomy and mechanics of the human hand. *Artificial limbs*, 22, 1955.

[3] Steve Davis, Nikolaos G Tsagarakis, and Darwin G Caldwell. The initial design and manufacturing process of a low cost hand for the robot icub. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 40–45. IEEE, 2008.

[4] Hong Liu, Ke Wu, Peter Meusel, Nikolaus Seitz, Gerd Hirzinger, MH Jin, YW Liu, SW Fan, T Lan, and ZP Chen. Multisensory five-finger dexterous hand: The dlr/hit hand ii. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3692–3697. IEEE, 2008.

[5] Lionel Birglen, Thierry Laliberté, and Clément M Gosselin. *Underactuated robotic hands*, volume 40. Springer, 2007.

[6] Marco Controzzi, Francesco Clemente, Diego Barone, Alessio Ghionzoli, and Christian Cipriani. The sssa-myhand: a dexterous lightweight myoelectric hand prosthesis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5):459–468, 2017.

[7] Agostino De Santis, Bruno Siciliano, Alessandro De Luca, and Antonio Bicchi. An atlas of physical human–robot interaction. *Mechanism and Machine Theory*, 43(3):253–270, 2008.

[8] Clement M Gosselin and Thierry Laliberte. Underactuated mechanical finger with return actuation, June 9 1998. US Patent 5,762,390.

[9] Thierry Laliberté Clément M Gosselin. Development of a three-dof underactuated finger.

SOCRATES

The research in Social Robotics has a common theme of *Interaction Quality,* which is a concept for characterization of how a specific mode of interaction is *fit* for a given task, situation, and user. Interaction Quality often changes, for instance if an older adult gets tired and loses focus when interacting with a robot. Interaction Quality also depends on the robot's functionality and design. By slowing down the speed of the robot, Interaction Quality can be maintained. In general, Interaction Quality is a complex interplay between several performance measures and design parameters. In SOCRATES we address these issues from a range of perspectives in five research workpackages :

- *Emotion:* novel multi-modal methods to perceive human emotions from facial expressions, body motion, auditory and language cues
- *Intention:* new techniques to infer human goals and intention from natural language and video analysis
- *Adaptivity:* techniques to adapt a robot's behaviour to user needs
- *Design:* Novel design methods for hardware, interfaces, and safety
- *Acceptance:* Procedures for evaluation of user acceptance

Additional value and impact is generated by the unique multidisciplinary collaboration between academic disciplines that normally do not work together; computer science, cognitive science, biomechanics, ethics, social psychology, and social science. Intersectoral collaboration between academia, caregivers, business developers, and robot manufacturers will further strengthen novelty and impact by ensuring that relevant needs are addressed, and that research result are both economically and technically feasible.

An Adaptive Neural Approach Based on Ensemble and Multitask Learning for Affect Recognition

Henrique Siqueira

***Abstract*—In this paper, we evaluate the effect of Multitask Learning (MTL) in an ensemble with shared representations based on convolutional networks in the task of affect recognition from facial expressions. Our convolutional architecture is divided into three levels of hierarchy regarding MTL. The first level is conditioned to learn lower-level representations, which are shared with independent convolutional branches related to different tasks on the second level. While each independent branch is fostered to learn task-specific representations, the early shared layers are fostered to learn features that are relevant to multiple tasks due to the inductive transfer mechanism from MTL. The third level consists of an ensemble of convolutional branches responsible for learning higher-level representations and allowing re-training with unlabelled expressions. Our experiments show a slight improvement in recognition performance using MTL over Single Task Learning (STL) on the AffectNet dataset, but a significant reduction in training time. Finally, we discuss the potential use of MTL and hard constraints into the inference and re-training processes of the proposed approach to improve its generalization performance.**

***Index Terms*—Semi-supervised Learning, Multitask Learning, Ensemble Methods, Facial Expression Recognition**

I. Introduction

With the advance in health care, the modern society is enjoying longer lives. The long life expectancy accompanied by low birth rates dictate the growth of ageing populations in several countries, which already comprise over a tenth of the global population [1]. Besides physical health, psychological and sociological factors have a significant impact on well-being and good life quality in old age. Sociability, in particular, plays a crucial role against loneliness in advanced years, which is one of the main factors that lead older adults to experience feelings of depression and thoughts of mortality [2].

Studies from different areas including robotics, medicine and economics have suggested making use of social robots as home companions and social assistants in senior care facilities to address loneliness among older adults and to support their needs and independence [1]. In addition to their functional activities (e.g., dispensing of medication and providing reminders), such robots can establish social and affective relationships with older adults which reduce feelings of loneliness among older people and provide warm caregiving to them, as investigated by Pols and Moser [3].

A fundamental aspect of social robots is their affective capabilities; the ability to recognize, express or even have emotions, albeit having simulated ones [4]. Emotions are highly present in human interactions, by influencing our rational thinking and decision-making [5]. Sad facial expressions and a low tone of voice during a conversation, for instance, might encourage a friend to comfort you [6]. A social robot capable of identifying and using this emotional information for making decisions could enhance its social skills by initiating an interaction with a senior perceived as sad to support them with positive messages. As evidenced by Sabelli et al. [7] through an ethnographic study of a conversational agent in an elderly care center, such emotionally support not only improve engagement in interacting with a robot, but also reduce loneliness and positively regulates their feelings.

Despite the remarkable progress in the area of automatic emotion recognition (see Poria et al. [9] for a recent review of affective computing), most of the existing approaches are extensively trained using supervised learning techniques on a given dataset [10], [11], which frequently drop in recognition performance when trialled under different conditions than the one used for training [12], [13]. Taylor et al. [14] suggested that this drop in recognition performance may be caused by the inability of those approaches to account for individual differences, since the same emotional state can be expressed differently among individuals [5], [15]. Even the same person may present a high physiological variation for the same emotional state in different days [16]. Therefore, an emotion recognition system that could improve recognition performance over time with unlabelled expressions is beneficial to social robots as they could be able to enhance their emotional capabilities over interactions. This adaptive capability is especially needed for social robots in senior care facilities, since emotional expression variations in older adults may be even higher due to cognitive or physical issues [17].

As investigated in our previous work [8], an ensemble with shared representations can potentially be used as an adaptive emotion recognition system for social robots, where emotional expressions collected from human-robot interactions can be utilized for re-training the ensemble. Although re-training the system using the ensemble predictions from unlabelled expressions led to an improvement in recognition performance in the majority of cases, there were few cases where it degenerated the recognition capability. We hypothesize that providing more information about an emotional expression via Multitask Learning (MTL) might not only yield to better generalization performance, but might also make the re-training phase through ensemble predictions more efficient. MTL can be defined as an inductive transfer learning mechanism where multiple *related* tasks are trained in parallel using shared

The author is with Knowledge Technology, Department of Informatics, University of Hamburg, Vogt-Koelln-Str. 30, 22527 Hamburg, Germany siqueira@informatik.uni-hamburg.de

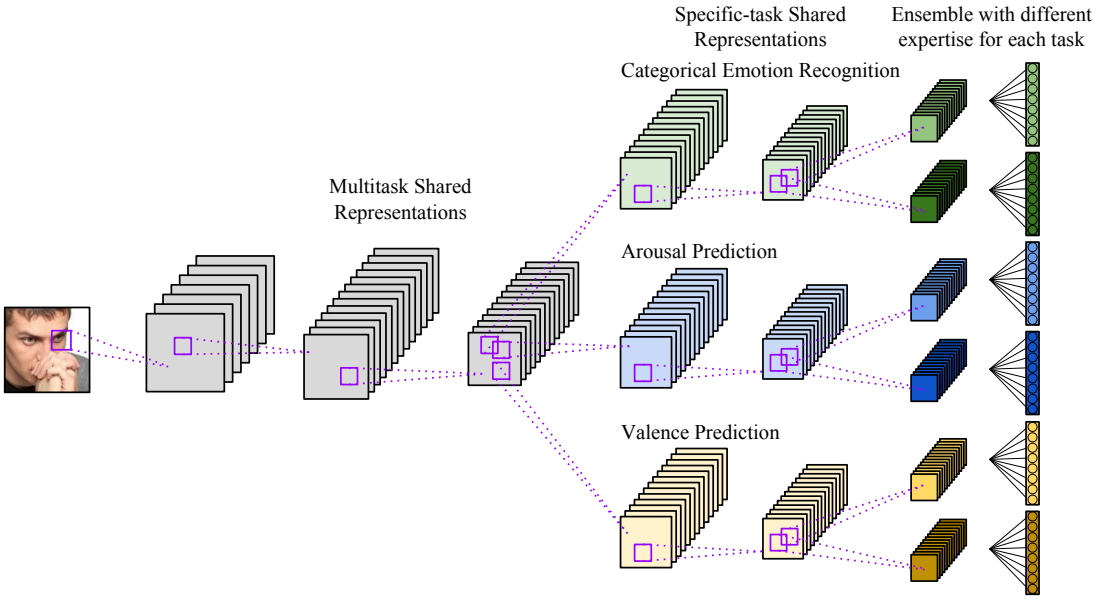


Fig. 1. Illustration of the proposed architecture for multitask learning. While the early layers in gray learn lower-level representations useful for multiple tasks, the three separate convolutional branches in green (top), blue (middle) and yellow (bottom) are employed to learn task-specific representations. On the right, the ensemble of convolutional branches is adopted as proposed by Siqueira et al. [8] for learning higher-level representations for each task.

representations [18]. Several studies have demonstrated the benefits of multitask learning on improving generalization performance and decreasing training time in contrast to Single Task Learning (STL) [14], [18], [19], where a machine learning method learns only one task at a time. Devries et al. [19] have demonstrated that facial expression recognition can be improved by training a convolutional neural network to detect facial landmarks as an auxiliary task in an MTL setting. Taylor et al. [14] employed MTL to account for individual differences for mood prediction by first clustering them regarding personality and gender, and subsequently, each cluster considered as a different prediction task, which resulted in an overall improvement on the generalization performance.

In this paper, we adapt our previous approach to employ multitask learning. Our approach consists of designing a convolutional architecture based upon three different levels of hierarchy regarding MTL. The first level is responsible for learning lower-level representations from the data. These representations are shared between multiple and independent branches in the second level, where each branch is constrained to learn features relevant to a particular task. In this work, we consider as related tasks the recognition of categorical emotional expressions (e.g., happy, sad and neutral), and the prediction of arousal and valence levels from the dimensional representations of emotion by Russell [20]. Lastly, the third level is an ensemble of convolutional branches with different expertise for each task, as described in Siqueira et al. [8]. In addition of presenting our preliminary analysis of the effect of MTL on the generalization performance on the AffectNet dataset [12], we discuss the potential benefits of using multiple information from the same input as hard constraint [21] in the inference and re-training processes of the proposed approach.

II. Approach

In the proposed approach illustrated in Figure 1, the early convolutional layers learn lower-level representations from the training data. They are conditioned to discovery features that are suitable to different and related tasks by the inductive transfer learning from multiple teach signals which are back-propagated from each task-related output to the shared representations, as defined by Caruana [18]: *"the multitask bias causes the inductive learner to prefer hypotheses that explain more than one task"*. These lower-level representations are shared between independent convolutional branches, each related to a specific task represented by different colors in Figure 1. The green convolutional branch, in the context of this paper, is fostered to learn important features to distinguish categorical emotions, where the blue and yellow branches to learn relevant features for predicting arousal and valence, respectively. In the highest level of the architecture, an ensemble of convolutional branches is employed as proposed in the work of Siqueira et al. [8]. The major goal for each branch in the ensemble is the development of higher-level representations from the training data that are different and complementary to other branches' expertise. If this assumption is satisfied, recognition performance might be improved by re-training the ensemble with their own predictions [8].

While multitask learning may improve the generalization capability of a model by fostering shared layers to learn features that are useful for different tasks, the different pieces of information gathered for each task from the same emotional expression might provide supplementary evidence for the correct classification of such expression. As an example, suppose that the convolutional branches responsible for the categorical emotion recognition classify a given expression, with a certain

degree of uncertainty, as happy or sad. Uncertainty cases may occur when some branches classify an image as belonging to a class A, while other branches classify the same image as belonging to a class B. By using prior knowledge about the task and different pieces of information from the same input, the valence prediction could have charged the same expression in our example as positive, and hence, the confidence for the categorical emotion recognition could have been increased towards the happy category. This strategy can be understood as imposing hard constraints in the inference and training processes, and this field of study is well explored in the book of Gori [21]. In spite of the potential benefits of imposing hard constraints into our approach, the experiments conducted for this paper are limited to the analysis of the effect of MTL on the recognition performance.

III. Preliminary Experiments

Fig. 2. Examples of the eight discrete categories from the AffectNet dataset [12] adopted in our experiments: Neutral (Ne), Happy (Ha), Sad (Sa), Surprise (Su), Disgust (Di), Fear (Fe), Anger (An) and Contempt (Co).

We evaluated the proposed approach on the AffectNet dataset [12], which consists of over a million face images collected by querying search engines with emotion-related keywords in six different languages. AffectNet is divided into the labelled training, unlabelled training, validation and test sets. Each set was manually annotated in terms of categorical and dimensional representations of emotion, except the unlabelled training set. In addition to the universal facial expressions proposed by Ekman [22] (see Figure 2), such as Happy (Ha), Sad (Sa), Surprise (Su), Fear (Fe), Disgust (Di), Anger (An) and Contempt (Co), the categorical representation of AffectNet also presents Neutral (Ne), None (No), Uncertain (Un) and Non-Face (NF) categories. For the dimensional representation, the dataset was annotated based on the circumplex model of affect proposed by Russell [20], where the *arousal* level indicates how excited or calm an event is, the *valence* level indicates how pleasant or unpleasant an event is. Continuous values ranging from -1 to 1 were assigned to emotional facial expressions, whereas -2 indicates images that belong to non-face and uncertain categories.

Our architecture is divided into three levels. The first level consists of three convolutional layers with 64, 128 and 256 filters. These lower-level representations are shared between three convolutional branches, one for each task: the classification of categorical emotions, and the prediction of arousal and valence levels. Each convolutional branch has one convolutional layer with 512 filters for learning features relevant to a specific task. Until this level, all of the convolutional layers are followed by batch normalization and max-pooling layers with a pool size of 2. The third and highest level is an ensemble of convolutional branches. For the categorical emotion recognition task, four branches compose the ensemble. Each branch in the ensemble is composed of one convolutional layer with 1024 filters, followed by the global average pooling layer, and the output layer with 8 neurons. To foster the development of different and complementary features for the same task in the ensemble, a different weighted loss function is assigned for each branch. This overall configuration is also adopted in the other two branches, which are responsible for predicting arousal and valence levels. However, their output layers have 41 neurons each, representing the discrete counterpart of the continuous emotional scales. This discretisation is necessary to assign a unique weighted loss function for each branch. As activation function, ReLU is adopted for all of the neurons, except the output layer where the softmax function is applied. During validation, we take the mean probability distribution from the ensemble.

TABLE I
Accuracy (%) and RMSE on AffectNet for categorical and dimensional representations of emotion.

Approaches	Categorical	Arousal	Valence	Params
MTL	50.32%	**0.37**	0.46	**50M**
STL	48.05%	0.39	0.47	**50M**
Mollahosseini et al. [12]	**58.00%**	0.41	**0.37**	180M

We adopt the single task learning counterparts of the proposed architecture as baselines. Thus, the network trained for categorical emotion recognition consists of five convolution layers with 64, 128, 256 and 512 filters, followed by an ensemble with four convolutional branches, each of which consisting of a convolutional layer with 1024 filters, an average pooling layer, and an output layer with 8 neurons. In addition to the comparisons with the baseline networks, we also compare our results with the approach proposed by Mollahosseini et al. [12] in the AffectNet paper. In their work, three different AlexNets [23] were re-trained on the AffectNet dataset, outperforming traditional classifiers and off-the-shelf facial expression recognition systems such as support vector machines and Microsoft Cognitive Services emotion API [1]. The faces are cropped using the facial coordinates provided by the dataset, and re-scaled to 96 x 96 pixels to reduce the computational cost. The pixel intensities from each image are normalized between 0 and 1. The networks were trained for 15 epochs using RMSProp with an initial learning rate of 0.001.

A. Initial Results and Discussion

Table I shows the accuracy for the categorical classification of emotions, the root-mean-square error (RMSE) for the predictions of arousal and valence levels, and the number of trainable parameters for each approach. MTL represents the proposed approach trained for multiple related tasks in parallel, whereas STL represents its counterpart but trained for one task at a time. Therefore, the results reported for STL are three different convolutional networks trained from scratch

[1]https://www.microsoft.com/cognitiveservices/enus/emotionapi

on AffectNet. This is also true for the results reported by Mollahosseini et al. [12]. Each AlexNet re-trained by them has roughly 60 million trainable parameters [23], resulting in 180 million parameters for the three networks.

Although the recognition performance of MTL and STL are similar, with the first reaching slight higher accuracy for categorical emotion classification, and lower RMSE for the arousal and valence predictions, the proposed approach can be trained t times faster than STL, being t the number of tasks to be learnt. The training time factor might be crucial for the application of the proposed approach for continual learning in robotic platforms, especially robots with limited computational resources. When compared with the methods proposed by Mollahosseini et al. [12], the proposed approach has achieved a substantial lower RMSE for arousal prediction, but has presented an inferior performance for categorical emotion classification and valence prediction. However, the adaptation of their methods for continual learning, where a robot should improve its recognition performance over time might be infeasible due to the high number of parameters.

IV. Conclusions and Future Work

We adapted our previous work on an ensemble with shared representation to account for multitask learning. MTL acts as an inductive transfer learning mechanism that frequently improves generalization performance by fostering shared representations to learn features that are useful for different tasks. Although the employment of multitask learning provided a small gain in recognition performance, it provided a significant reduction in training time since several tasks can be trained in parallel. This training time gain is an important factor for continual learning in social robots, since response time is fundamental to a natural interaction. Moreover, we discussed how different pieces of information from the same input regarding MTL could be used as hard constraints in the inference and training processes for improving generalization performance.

As future work, we will analyse the internal representations related to each level of hierarchy regarding MTL in the proposed approach. This analysis might explain the slight improvement on generalization performance obtained in our experiments. Furthermore, the potentiality of MLT and hard constraints for improving generalization performance discussed in this paper will be evaluated on AffectNet, including an analysis of the adaptive behaviour of the proposed approach on the unlabelled training set. In addition to a static dataset of emotions, the proposed approach will also be evaluated in a more naturalistic condition, where not only spatial but also temporal features are presented in the expression of the individual emotional state [24].

V. Acknowledgement

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 (SOCRATES).

References

[1] I. Pedersen, S. Reid, and K. Aspevig, "Developing social robots for aging populations: A literature review of recent academic sources," *Sociology Compass*, vol. 12, no. 6, p. e12585, 2018.

[2] A. Singh and N. Misra, "Loneliness, depression and sociability in old age," *Industrial psychiatry journal*, vol. 18, no. 1, p. 51, 2009.

[3] J. Pols and I. Moser, "Cold technologies versus warm care? on affective and social relations with and through care technologies," *ALTER*, vol. 3, no. 2, pp. 159–178, 2009.

[4] L.-F. Rodríguez and F. Ramos, "Computational models of emotions for autonomous agents: major challenges," *Artificial Intelligence Review*, vol. 43, no. 3, pp. 437–465, 2015.

[5] R. W. Picard, "Affective computing," MIT Media Laboratory, Perceptual Computing, Tech. Rep., 1997.

[6] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Robotics and Autonomous Systems*, vol. 58, no. 3, pp. 322–332, 2010.

[7] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: An ethnographic study," in *2011 6th ACM/IEEE International Conference on HRI*, March 2011, pp. 37–44.

[8] H. Siqueira, P. Barros, S. Magg, and S. Wermter, "An ensemble with shared representations based on convolutional networks for continually learning facial expressions," in *Accepted in Intelligent Robots and Systems (IROS) IEEE/RSJ International Conference. IEEE.*, 2018.

[9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[10] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, ser. ICCVW '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 19–27.

[11] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Nov 2015, pp. 582–587.

[12] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, pp. 1–1, 2017.

[13] H. Siqueira, P. Barros, S. Magg, C. Weber, and S. Wermter, "A sublayered hierarchical pyramidal neural architecture for facial expression recognition," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Apr 2018, pp. 1–6.

[14] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–1, 2017.

[15] A. Siqueira, Henrique Sutherland and, P. Barros, M. Kerzel, S. Magg, and S. Wermter, "Disambiguating affective stimulus associations for robot perception and dialogue," in *Accepted in IEEE-RAS 18th International Conference on Humanoid Robotics (Humanoids)*, 2018.

[16] R. W. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[17] S. Scheibe and L. L. Carstensen, "Emotional aging: Recent findings and future trends," *The Journals of Gerontology: Series B*, vol. 65, no. 2, pp. 135–144, 2010.

[18] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[19] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *2014 Canadian Conference on Computer and Robot Vision (CRV)*. IEEE, 2014, pp. 98–103.

[20] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[21] M. Gori, *Machine Learning: A Constraint-based Approach*. Morgan Kaufmann, 2017.

[22] P. Ekman, "The argument and evidence about universals in facial expressions," *Handbook of Social Psychophysiology*, pp. 143–164, 1989.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," *arXiv preprint arXiv:1803.05434*, 2018.

Toward Emotion Recognition From Early Fused Acoustic and Language Features Using Recursive Neural Networks

Alexander Sutherland

Abstract— **Recognising emotions from language is considered an important aspect of affective computing. However, the application of recognised emotions in an effective manner is often bound to the context where the emotion was detected, without the acquisition of information about the relation between spoken words and the recognised emotion. To apply recognised emotions to a broader context, knowledge about this dynamic must be accrued during the emotion classification process. In this paper, we outline a novel method of extracting these relations, using recursive neural networks to process the syntactic structure of speech in order to better understand how emotions are expressed and what spoken words they relate to.**

I. Introduction

Recognizing and responding to human emotions in HRI scenarios is regarded as important for acceptance of robots [4]. Acceptance and empathy is vital for long-term sustainable Human-Robot relations, as users are liable to reject or ignore robots they feel no connection with. To perform recognition, emotional expressions from multiple modalities, such as vision, audio, and language, are often combined to improve recognition accuracy and system robustness [3].

Current multimodal emotion recognition using speech reduces features of language to an abstract level for more convenient processing. While this simplifies processing, it makes the role of language structure more implicit than explicit and in this simplification, some information is lost. Structureless processing of language also requires the structure to be relearned as an emergent property to facilitate the understanding of relationships between words. While this is possible, the syntactic structure of language is already well defined and should not require relearning and running the risk of potential errors in syntactic understanding.

Through reintroduction of the structure of language into the categorical emotion recognition pipeline, we hope to visualize over the syntax graph how detected emotion expressions relate to acoustics and language, in contrast to what Socher et al. [2] did for sentiment on only language.

II. Recursive Neural Networks

Recursive Neural Networks, RvNNs, [1] are able to process structured data and therein learn feature patterns that occur in the structure of data. This allows RvNNs to use structure as a feature rather than having to relearn structure as an emergent property of the network. The fundamental difference between recursive and recurrent architectures is that recurrent neural networks, RNNs, are a special case of RvNNs that only handle a linear chain of input, whereas a RvNN may take an arbitrary number of inputs from a previous time-step. This is often described in the manner of a bottom-up analysis of a hierarchical tree structure, computing the values of parent nodes based on the nodes of their respective children.

Language has an explicit structure that can be exploited by RvNNs to provide a more nuanced understanding of how humans express sentiment based on the structure of language used [1], [2]. The benefit of using RvNNs for this is the increased granularity of class predictions over the syntax tree. Parent nodes are a product of their children and resulting classifications can be traced back to specific sub-branches within the syntax tree. An example is shown in the work of Socher et al. [2], where the authors show the negating word "not" influences the final sentiment classification. Our novel contribution will be the extension of this approach to incorporate acoustic data in a uni-modal and multi-modal fashion, with text, over categorical labels as opposed to sentiment labels. We expect that this will provide insight on how language and pronunciation influences emotion recognition through visualization over the syntax graph.

III. Proposed Method

In this position paper, a method of using RvNNs to process fused language and acoustic features will be outlined. An overview of how the system will process data can be seen in Figure 1. The composition function, g, will be the one used by Socher et al. [1], as it has shown promise when classifying sentiment. In Figure 1 we see that the network calculates intermediate probabilities in a bottom-up fashion, also allowing for predictions of individual nodes and parent nodes as the network works its way through the graph.

Predictions are attained through a projection layer that learns how to convert intermediate representations to a probability distribution over target emotions. Once the probability of every emotion class for every node is calculated this can be visualised in a tree by choosing the highest probable emotion. An example where this is useful is determining what phrases are responsible for emotional outcomes and motivating why different decisions were made based on occurring phrases and emotion predictions in each node.

Utterances with transcripts will be selected from the IEMOCAP dataset [8], a multimodal emotion recognition dataset. For each utterance, every word will be aligned with it's associated audio segment. Word features will be attained through pretrained word embeddings [6] and for audio

Knowledge Technology, Department of Informatics, University of Hamburg, Germany, sutherland@informatik.uni-hamburg.de

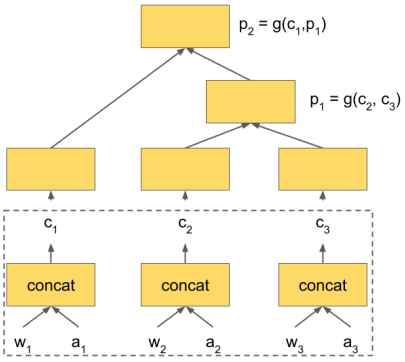


Fig. 1. This figure shows how a recursive neural network would process a three word sentence with paired audio and a specific syntactic structure. Here the word embeddings, w_i, are concatenated with extracted audio features, a_i, corresponding to each word. This results combined feature vector, c_i, that is then fed to the recursive network which calculate target probabilities, p_j, using the composition function g.

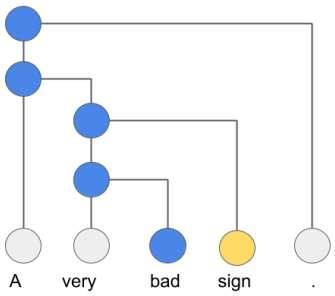


Fig. 2. RvNN output for categorical emotion recognition over the syntactic graph of a simple sentence. Different colours represent the highest predicted emotion in a particular node in the sentences syntax tree. Blue is sadness, yellow is happiness, and grey is emotionally neutral.

MFCC features [7] will be fed in sequence to a pre-trained LSTM to extract audio features for every word. Thereafter we will perform early fusion through concatenation of word and audio representations and feed the subsequent vectors to the RvNN based on syntactic structure to perform emotion recognition. Syntactic structure will be extracted through the use of standard NLP libraries available in Python.

IV. Expected Results

Expected results of applying a RvNN for this task will allow us to attain emotion predictions, using acoustic and language features, for every node in the syntactic graph of an input utterance. This will allow us to see which syntactic sub-trees contribute to emotion classifications. A possible additional outcome would be a higher emotion classification accuracy from acoustics alone and combined multimodally than with standard recurrent approaches.

Currently, we are able to show preliminary examples of using the text modality alone. In Figures 2 and 3 we see the results of applying a RvNN to the task of categorical emotion recognition, as opposed to sentiment classification. To do this we translate the labels of the Stanford Sentiment Treebank (SST) [2] to categorical labels, wherein positive labels are translated to happy, neutral to neutral, and negative labels to either angry or sad based on the classification of

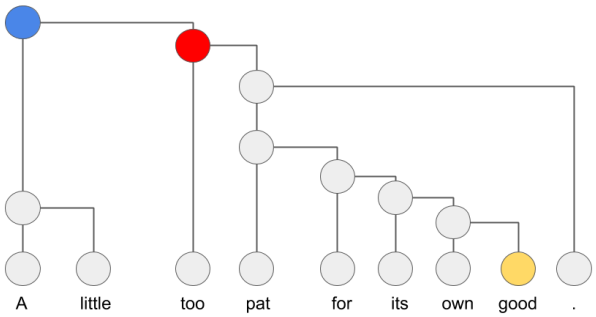


Fig. 3. An example of RvNN output showing that it is able to capture how certain sequences of words are able to shift the classification between negative categories over a syntax graph. Here the words “A little” shift the prediction from an overall angry classification (red) to a more sombre sad classification (blue).

an LSTM pretrained on the IEMOCAP dataset [8] for categorical emotion recognition. The SST is used for exemplary purposes and will be replaced by the IEMOCAP when the required syntax trees have been generated.

V. Conclusion

In this paper, we show preliminary work toward a novel method of processing language and audio features for improving visualisation and understanding of emotion recognition using recursive neural networks. We also visualised how categorical emotion predictions distribute themselves over the syntax graphs of two simple sentences.

Acknowledgements

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 (SOCRATES).

References

[1] Socher, R., Lin, C.C., Manning, C. and Ng, A.Y., 2011. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 129-136).

[2] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A. and Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

[3] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A. and Morency, L.P., 2017. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 873-883).

[4] Lim, A. and Okuno, H.G., 2015. A recipe for empathy. International Journal of Social Robotics, 7(1), pp.35-49.

[5] Tai, K.S., Socher, R. and Manning, C.D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

[6] Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[7] Logan, B., 2000, October. Mel frequency cepstral coefficients for music modeling. In ISMIR (Vol. 270, pp. 1-11).

[8] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), p.335.

Body motion properties as indicators of depression in elderly*

Vesna Poprcova

***Abstract*— Depression is a common mood disorder that is rapidly affecting lives of elderly worldwide. The detection of depression, however, is an issue because the common methods are subjective and depend of patient self-reports. Automated recognition may, therefore, be beneficial. This paper examines the possibility of using body motion properties as potential indicators of depression in elderly, and proposes an experimental method to assess the validity of such measures.**

I. Introduction

Population ageing is a worldwide trend and the proportion of elderly people is constantly increasing [2]. The growing burden of depression in elderly suggests that there is a need to develop automated depression detection which will help in effective care of patients suffering from depression. Automated depression detection can be used to support clinicians' decisions, to avoid false diagnosis as well as overcome subjective bias associated with self-reports.

Depression is a state of negative mood that may last for a long time and impact the individual's proper functioning [8]. In addition to effects on the person's thoughts, behaviors, feelings, and sense of well-being, depression has an impact to the motor system as well [15]. And indeed, past research has examined the effects of depression on body motion [5]. However, the current knowledge deals with young people. Yet, because aging has an impact on body motion [5], it would be beneficial to examine the effect of depression on body motion in elderly. Thus, the current paper focuses on examining the link between depression and body motion in elderly. It reviews the relevant literature and proposes an experimental methods to investigate the effect.

II. Related Work

Body motion is a central part of the human social communication [3]. It may be defined as the collection of signs such as posture, speed of movement, meaningful coordination of actions expressed by the human body [10].

Psychologists have shown that depressed individuals differ from non-depressed with regard to objectively quantified gross motor activity, body movements and motor reaction time [11]. For instance, sadness and depression are characterized by reduced walking speed, arm swing, and vertical head movements as well as slumped postures and larger lateral body sway [6].

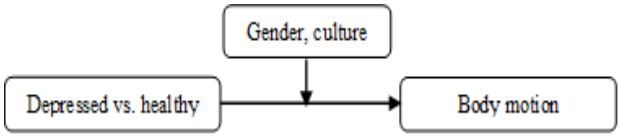
Past research examined visual indicators for depression, including body motion and periodical muscular movements [7]. Research have also used video data to recognize depression based on general movements [4, 9, 12], posture [9] and head pose [4, 12]. Results have shown that body expressions and head movements can be significant visual cues for depression detection [4].

Researchers dealt with the link between body motion and mood (rather than depression). One research, for example, found upper body postural features can predict people's mood [13]. Mood prediction has so far been tackled mostly through the direct mapping of video features to mood. Decision Tree classification and multimodal fusion of audio-visual and text features is performed by [16]. Head pose and movement features associated with the face are considered by [1] performing classification with Support Vector Machines (SVM) on depression recognition. They concluded that head movements of depressed people are different than that of normal person. Deep learning based approach is presented by [14].

Altogether, the literature demonstrates that automated detection of depression in general population, regardless of age, is an active research area. We propose an experimental method to examine the topic specifically in elderly.

III. Proposed Method

Research participants will be elderly people who will be recruited based on background information we will collect. We will make sure to recruited both depressed and non-depressed participant. Depression data will rely on self-report.



Participants will be asked to walk straight for several minutes in our lab. Movement will be recorded using a motion capture system (Qualysis, Sweden).

Relying on past research conducted in lab setting [5], we will use regression analysis to predict the depression in elderly by learning the relationship between body motion properties (walking, head movements, stability, posture) as features and depression scale. Human-verified examples will be provided to a regression algorithm which learns the mapping and novel video frames can then be interpreted by extrapolating from this learned mapping.

Context knowledge summarizes information about the environment, subject (gender, personality traits and culture), current activities and interactions. The current approaches largely do not take context knowledge into account but

* This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

The author is from the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev Beersheba 84105, Israel (email: vesna@post.bgu.ac.il).

analyzing the context when a movement is performed may lead to more robust recognition [5].

IV. Expected Results

We expect that the proposed method will define clearly which body motion properties can be successfully used as potential indicators to detect the scale of depression in elderly.

V. Conclusion

The expected results can contribute to the field of automated detection of depression in elderly. Such module for automated depression recognition can be integrated into a robot architecture and Human Robot Interaction (HRI) scenario. An additional key contribution of this work will be the design of a new database focused on elderly with the goal to validate experimentally the proposed model. The database will contain local elderly comprised of two groups: (a) diagnosed with depression and (b) elderly that are healthy with no clinical disorders.

Acknowledgment

The work described in this paper has been funded by the European Commission's Horizon 2020 research and innovation programme under the MSCA-ITN-2016 grant agreement No 721619 (SOCRATES project).

References

[1] Alghowinem., S., Goecke., R., Wagner., M., Parkerx., G., Breakspear., M..: Head pose and movement analysis as an indicator of depression. Affective Computing and Intelligent Interaction (ACII), 2013

[2] Bottazzi, D., Corradi, A., Montanari, R.: Context-aware middleware solutions for anytime and anywhere emergency assistance to elderly people. IEEE Commun. Mag., Apr 2006 vol. 44, no. 4, pp. 82–90

[3] Burgoon, J.: New Perspectives in Nonverbal Communication: Studies in Cultural Anthropology, Social Psychology, Linguistics, Literature and Semiotics. Journal of Language and Social Psychology, 1985

[4] Joshi, J., Goecke, R., Breakspear, M., Parker, G.: Can body expressions contribute to automatic depression analysis? International Conference on Automated Face and Gesture Recognition 2013

[5] Karg., M., Samadani., AA., Gorbet., R., Kuhnlenz., K., Hoey., J., Kulic., D.: Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. IEEE Transactions on Affective Computing, 2013, pp 341-359

[6] Michalak., J., Troje., N., Fischer., J., Vollmar., P., Heidenreich., T., Schulte., D.: Embodiment of sadness and depression - gait patterns associated with dysphoric mood. Psychosom Med, 2009, vol. 71, pp. 580–587

[7] Morales, M., Scherer, S., Levitan, R.: A Cross-modal Review of Indicators for Depression Detection Systems. Fourth Workshop on Computational Linguistics and Clinical Psychology, 2017

[8] Salmans, S.: Depression: questions you have - answers you need. People's Medical Society, 1995

[9] Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., Morency, L-P.: Automatic behavior descriptors for psychological disorder analysis. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2013, pp. 1–8

[10] Schindler, K., Van Gool, L., de Gelder, B.: Recognizing emotions expressed by body pose: a biologically inspired neural model. Neural Networks Volume 21 Issue 9 November 2008, pp 1238-1246

[11] Sobinand., C., Sackeim.,HA.: Psychomotor symptoms of depression. Am. J. Psychiatry, 1997, pp 4–17

[12] Stratou, G., Scherer, S., Gratch, J., Morency, L-P.: Automatic non-verbal behavior indicators of depression and ptsd: Exploring gender differences. Affective Computing and Intelligent Interaction (ACII), 2013, pp. 147–152

[13] Thrasher., M., Van der Zwaag., M. D., Bianchi-Berthouze., N., Westerink., J. H.: Mood recognition based on upper body posture and movement features. Affective Computing and Intelligent Interaction, 2014, pp. 377-386

[14] Tzirakis., P., Trigeorgis., G., Nicolau., MA., Schuller., BW., Zafeiriou., S.: End-to-End Multimodal Emotion Recognition using Deep Neural Networks. IEEE Journal of Selected Topics in Signal Processing, 2017, pp 1301-1309

[15] Tryon, W.: Activity Measurement in Psychology and Medicine. Applied Clinical Psychology Series. Springer, 1991

[16] Yang., L., Jiang., D., He., L., Pei., E., Cedric Oveneke., M., Sahli., H.: Decision Tree Based Depression Classification from Audio Video and Language Information. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016

Text-Based Inference of Object Affordances for Human-Robot Interaction

Michele Persiani[1] and Thomas Hellström[2]

***Abstract*—Affordances denote actions that can be performed in the presence of different objects. In this paper we present a model to generate names of possible affordances for a named object. We use a Conditional Variational Autoencoder as generative model and train it with sentences from a selected corpus. The model can be used in several ways in HRI, for instance by a service robot providing assistance to perform activities of daily living. The preliminary evaluation of the model shows good results compared to a benchmark method.**

***Index Terms*—Affordance, Intention recognition, Human-Robot-Interaction, Generative model, Autoencoder, Natural language processing**

I. Introduction

The term "affordance" was introduced by the American psychologist Gibson [1] to describe what an animal can do with a given object. It has since then been extensively utilized, interpreted, and re-defined (see [2] for an overview) in fields such as human-computer-interaction [3] and human-robot-interaction (HRI) [4]. We focus on its use within HRI, and use the term to denote actions that can be performed with a given object. As a simplified first approach we ignore the influence of different environments, and assume a one-to-many mapping G: *Objects* $\rightarrow$ *Affordances*. The object "door" may, for example, be used to perform the actions "open", "close", and "lock".

This paper presents ongoing work on how G may be learned from free-text corpora. The results show how it is possible to learn a generative model G that, given an object name, generates affordances according to a probability distribution that matches the used training data. Qualitatively results also indicate that the model manages to generalize, both to previously unseen objects and actions.

The paper is organized as follows. In Section II we give motivation for the work from an HRI perspective, followed by a brief review of earlier related work in Section III. The developed method is described in Section IV, and results from the evaluation are presented in Section V. The paper is finalized by conclusions in Section VI.

II. Affordances

Once learned, the function G can be used in several ways by a robot, for instance by a service robot providing assistance to perform activities of daily living. By visually identifying objects in the environment, or in the robot's verbal dialogue with the human, affordances can be inferred through G. The affordances may be used to infer the human's intention, which may guide the robot's behavior [5]. For example, if older adults want to talk to their distant children, a listening robot may infer that the adults wants to call them, and suggest making a phone call. G may also be used by a robot to decide how to act within a given context that affords certain actions. A service robot may, for example, suggest its user to read a book, if a physical book is visually detected. Affordances may also be useful for object disambiguation. When a human tells a robot to "pick it up!", the robot only has to consider objects with the "pick up" affordance in the current scene [4]. Inference of affordances may also be used to design robots that are understandable by humans, since mutually perceived affordances may contribute to explaining a robot's behavior [6], and thereby increase interaction quality [7].

[1,2] *Department of Computing Science*, *Umeå University*, Umeå, Sweden, michelep@cs.umu.se, thomash@cs.umu.se

III. Earlier work

Chao et al. [8] mine *semantic affordances* from a combination of crowdsourcing, images, and text. They show how in Natural Language Processing (NLP) objects and actions can be connected through the introduction of a latent space. Narashiman et al. [9] find links between object and action in text using deep reinforcement learning techniques [10]. Antanas et al. [11] relate affordances to the symbol grounding problem. By using image data, they map visual objects to utterances and actions, while through statistical methods they learn ontologies for affordances. Ruggeri and Di Caro [12] explain how affordance is a concept that sits in the middle between objectivity and subjectivity, and propose ontological views for their usage in Computer Science.

IV. Method

A generative model for the one-to-many mapping G : $Objects \rightarrow$ *Affordances* was trained with pairs <*object*, *action*>. These pairs were generated by *semantic role labeling* of sentences from a selected corpus. Objects and actions were represented by *wordvectors* throughout the process, which is illustrated in Fig 1 below.

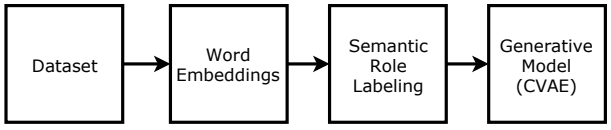


Fig. 1. Steps taken to obtain the generative model.

A. Corpus

As data source we used the *Yahoo! Answers Manner Questions* (YAMC) dataset[1] containing 142627 questions and corresponding answers. The corpus is a distillation of all questions gathered from the platform *Yahoo! Answers* during the year 2007. It is a small subset of the questions, selected for their linguistic properties such as good quality measured in terms of vocabulary and length.

B. Semantic Role Labeling

In NLP, semantic roles denote the semantic functions that words have in a given phrase [13]. For example, in the phrase "John looks in the mirror", the words "looks in" (denoted V) refer to the action being performed. "John" identifies the agent carrying out the action (denoted $A0$), and "the mirror" is the object (denoted $A1$) over which the action is performed.

Semantic role labeling [14] is the task of assigning semantic roles to words or groups of words in a sentence. A variety of tools exist for this task, with different conventions for the associated roles. As an example, for [15], the SEMAFOR parser [16] was used to infer human intention in verbal commands to a robot. In the current paper we used the parser in SENNA [17], which is a software tool distributed under a non-commercial license for academy[2].

After parsing the corpus using SENNA, phrases with semantic roles $A1$ and V being exactly one word each were selected. Each action V was lemmatized into the basic infinitive form since we were not interested in discriminating temporal or other variants of the verbs.

Finally, all pairs $A1$,V that appeared at least seven times were used to create data samples <*object, action*>. This number was found to be a good trade-off for filtering out spurious pairs.

A few examples of phrases and generated sample pairs <*object, action*> are shown in Table IV-B.

Phrase	**Sample pair**
Add flour.	<flour, add>
Crack the egg.	<egg, crack>
Set the mixer on two steps	<mixer, set>
Whip using the mixer	<mixer, use>
Open the oven.	<oven, open>
Enjoy the cake	<cake, enjoy>

Table IV-B Examples of object-action pairs generated from phrases in a recipe.

C. Word Embeddings

Word embeddings refer to a set of unsupervised methods that allow encoding of words as numeric vectors: *wordvectors*. In the numeric space, semantically and syntactically similar words are close if measured through cosine similarity. This is a desirable property for our generative model, as similar objects should show similar affordances.

GloVe [18] and Word2Vec [19] are common approaches to create word embeddings. We trained Word2Vec over YAMC to get embeddings for words that were most specific for our work. The selected dimensionality for the resulting wordvectors was 100. Qualitative evaluations of the resulting vectorial space showed how the computed embeddings were more suited to encode object-action pairs for common objects, than off-the-shelf embeddings.

[1]Obtained at *https://webscope.sandbox.yahoo.com/catalog.php?datatype=l*. Accessed July 3, 2018.

[2]See https://ronan.collobert.com/senna/. Accessed June 30, 2018.

D. Dataset

The words in each generated pair <*object, action*> were converted to wordvectors using the trained Word2Vec model, to provide numeric data to be used in the subsequent modeling. All data was divided into a training set comprising 15263 pairs, and a test set comprising 5088 pairs. Special care was taken to not include identical pairs in both training and test data sets. The data contained $N_O = 2628$ distinct object names and $N_A = 1167$ distinct action names that were collected into a dictionary.

E. Generative Model

We modelled the one-to-many mapping G: *Objects* $\rightarrow$ *Affordances*, using a *Conditional Variational Autoencoder* (CVAE) [20], illustrated in Fig 2.

A CVAE is a trainable generative model that learns a conditional probability distribution $\mathbf{p}(\mathbf{a}|\mathbf{o})$ while keeping a stochastic latent code in its hidden layers. They can be divided into two coupled layers: an encoder and a decoder. The encoder transforms the input distribution into a certain latent distribution $\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{o})$, while the decoder reconstructs the original vectors from its latent representation $\mathbf{z}$ together with the conditioning input $\mathbf{o}$, with output distribution equal to $\mathbf{p}_\varphi(\mathbf{a}'|\mathbf{z},\mathbf{o})$.

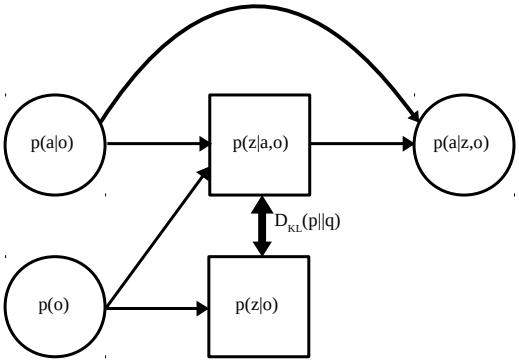


Fig. 2. CVAE with the addition of a parametric prior. Squares represent involved latent distributions.

The encoder's latent layer is regularized to be close to certain parametric prior $\mathbf{q}_\vartheta(\mathbf{z}|\mathbf{o})$. The lower-bound loss function for the CVAE is:

$$L_{CVAE} = \mathbb{E}[\log p_\varphi(a'|z,o)] - \lambda D_{KL}(q_\phi(z|a,o)||q_\vartheta(z|o)) \quad (1)$$

The first term accounts for how good the autoencoder reconstructs the input given its latent representation. The second term regularizes the hidden latent space to be close

to a certain posterior distribution. The factor λ balances how regularization is applied during learning. Starting from zero it is linearly grown up to one as the learning epochs advance. This technique addresses the *vanishing latent variable problem* and is referred to as KL annealing [21].

φ, ϕ, ϑ denotes the three disjoint sets of parameters of the components that are simultaneously involved in learning. More specifically, they represent set of weights for the three neural network composing the CVAE. The CVAE was trained using the training set generated as described above, and was implemented using the Keras [22] library for Python.

V. Evaluation

By inputting the name O of an object, and repeatedly sampling the CVAE, we obtain the same number of names for possible actions A. As described above, the sampling follows the estimated conditional probabilities $p(A|O)$. Hence, actions with high probability are output more frequently than actions with low probability. Since the CVAE outputs actions in numeric wordvector format, all output actions are "rounded" to the closest action word appearing in the dictionary. This is equivalent to a *K-NN* classification with $K = 1$, and the final output is the dictionary word belonging to the nearest neighboring wordvector. A few examples of the most probable generated actions for a given input objects is shown in Table V.

Input	Output
door	open, pull, put, loosen, grab, clean, leave, get, slide, shut
egg	hatch, poach, implant, lay, crack, peel, spin, whip, float, cook
wine	pour, add, mix, dry, rinse, melt, soak, get, use, drink
book	read, get, write, purchase, find, use, sell, print, buy, try
cat	declaw, deter, bathe, bath, spay, pet, scare, feed, attack
money	loan, inherit, double, owe, withdraw, save, waste, cost, earn, donate
knife	scrape, cut, brush, chop, use, roll, pull, remove, slide, rub
information	review, request, access, verify, present, obtain, identify, provide, submit, retain
body	trick, adapt, tone, adjust, recover, starve, cleanse, respond, flush, exercise
place	switch, prepare, hide, rent, start, own, guess, travel, avoid, suggest

Table V Examples of actions generated by the CVAE. For every input object we show the 10 most probable outputs, sorted from high to low probability.

Evaluation of generative models is in general seen as a difficult task [23]–[25], and one suggestion is that they should be evaluated directly with respect to the intended usage [23]. In that spirit we evaluated how often our model produced affordances that were correct in the sense that they matched test data.

Since the CVAE produces different results each time it is sampled, it was first sampled several times to estimate a probability distribution $p(A|O)$, to be compared with a similar estimation based on relative frequencies for the test data. A performance measure *Accuracy*, with values between 0 and 1, quantifies the similarity between these two distributions and was computed by the following algorithm.

1) $M \leftarrow 0$.

   For each of the N test samples $< O_j, A_j >, j = 1, \ldots, N$, repeat Steps 2-4:
2) Input object O_j to the CVAE and sample it 1000 times to estimate a probability distribution over possible actions for object O_j. Denote the set with the L most probable actions A_C.
3) As ground truth compute, for all actions A, $p(A|O_j) = N(A, O_j)/N(O_j)$, where $N(A, O_j)$ is the number of test data samples $< object, action >$ with $object = O_j$ and $action = A$, and $N(O_j)$ is the number of samples with $object = O_j$. For the resulting distribution, denote the set of the L most probable actions A_F.
4) If any action in A_C appears in A_F : $M \leftarrow M + 1$ (this corresponds to a notion of the CVAE output being "correct" for object O_j).
5) *Accuracy* $\leftarrow M/N$.

As benchmark method we generate actions using a random distribution, with $p(A|O_j) = \frac{1}{N_A}$ for all actions belonging to the training set, 0 otherwise.

We evaluated the benchmark in a similar fashion as described above, by replacing the CVAE for generation of the probability distribution in Step 2. Accuracy computed on the test set for CVAE and the benchmark are presented in Table V, for varying values of L.

L	CVAE	Random
1	0.099	0
5	0.552	0.09
10	0.781	0.322
15	0.862	0.498
20	0.903	0.773

Table V *Accuracy* for the CVAE and Random distributions, calculated as described above.

VI. Conclusions

We presented a novel generative model for text-based affordance generation by employing a Conditional Variational Autoencoder (CVAE). The presented preliminary results show that the model outperforms the benchmark method in generating possible actions for an input object.

For a given object, the action with highest probability to be generated by CVAE was most probable in test data 10% of the time. Given that the data set contained 1167 distinct action names, these results are quite satisfying. Considering more than just the action with highest probability, the accuracy for CVAE increases fast. For example, for $L = 5$, at least one correct action was output in 9% of the cases for the random distribution, and in 55% of all cases for CVAE.

As future work we will address several open questions:

- The used CVAE model is a complex architecture with several meta parameters and design choices. We will further investigate alternative designs, and use the performance measures to, possibly automatically, find optimal parameters.
- The relevance of using corpora like YAMC to generate affordances for HRI has to be investigated further. The difference between usage of language in human-robot

dialogue and in general corpora may affect accuracy, and alternative corpora could be considered.

- The generalization ability, i.e. performance for objects not present in the training data, will be investigated further. Successful generalization ability means that the method has true predictive power and does more than memorizing training data.
- Training with domain specific data will be investigated. As mentioned in the introduction, real object affordances typically depend on the environment, and better performance may be achieved by implicitly defining a specific domain (such as kitchen environments) and learn affordances with objects and actions relevant for that domain only.
- Alternative performance measures will be examined, for example based on a distance metric applied to the distributions in step 2 and 3 above. Explicit ways to assess the model's generalization ability will also be developed.
- Finally, envisioning robots as embodied agents, we will explore how affordances generation can be biased by taking into account specific perceptual and bodily abilities. The YAMC corpus expresses actions available to animals and humans, but can a robot, after a long day, just relax on the sofa?

ACKNOWLEDGEMENT

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

REFERENCES

[1] J. Gibson, *The theory of affordances in Perceiving, Acting, and Knowing. Towards an Ecological Psychology.* Hoboken, NJ: John Wiley & Sons Inc., 1977.

[2] M. Çakmak Mehmet R. Doğar, E. Uur, and E. Şahin, "Affordances as a framework for robot control," 2007.

[3] C. Schneider and J. Valacich, *Enhancing the Motivational Affordance of Human–Computer Interfaces in a Cross-Cultural Setting*. Heidelberg: Physica-Verlag HD, 2011, pp. 271–278. [Online]. Available: https://doi.org/10.1007/978379082632631

[4] T. E. Horton, A. Chakraborty, and R. St. Amant, "Affordances for robots: A brief survey," vol. 3, pp. 70–84, 12 2012.

[5] E. Bonchek Dokow and G. Kaminka, "Towards computational models of intention detection and intention prediction," vol. 28, 01 2013.

[6] T. Hellström and S. Bensch, "Understandable robots - what, why, and how," *Paladyn, Journal of Behavioral Robotics*, Accepted for publication.

[7] S. Bensch, A. Jevtić, and T. Hellström, "On interaction quality in human-robot interaction," in *International Conference on Agents and Artificial Intelligence (ICAART)*, 2017, pp. 182–189.

[8] Y.-W. Chao, Z. Wang, R. Mihalcea, and J. Deng, "Mining semantic affordances of visual object categories." in *CVPR*. IEEE Computer Society, 2015, pp. 4259–4267. [Online]. Available: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html

[9] K. Narasimhan, T. D. Kulkarni, and R. Barzilay, "Language understanding for text-based games using deep reinforcement learning," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 1–11. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1001.pdf

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[11] L. Antanas, O. A. Can, J. Davis, L. D. Raedt, A. Loutfi, A. Persson, A. Saffiotti, E. Ünal, D. Yuret, and P. Z. dos Martires, "Relational symbol grounding through affordance learning : An overview of the reground project," 2017.

[12] A. Ruggeri and L. D. Caro, "How affordances can rule the (computational) world," in *AIC@AI*IA*, 2013.

[13] X. Carreras and L. Márquez, "Introduction to the conll-2004 shared task: Semantic role labeling," 2004.

[14] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Comput. Linguist.*, vol. 28, no. 3, pp. 245–288, Sep. 2002. [Online]. Available: http://dx.doi.org/10.1162/089120102760275983

[15] A. Sutherland, S. Bensch, and T. Hellström, "Inferring robot actions from verbal commands using shallow semantic parsing," in *Proceedings of the 17th International Conference on Artificial Intelligence ICAI'15*, H. Arabnia, Ed., July 2015, pp. 28–34.

[16] D. Das, N. Schneider, D. Chen, and N. A. Smith, "Probabilistic frame-semantic parsing," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 948–956. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858136

[17] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Nov. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078186

[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *In EMNLP*, 2014.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1301.html

[20] C. Doersch, "Tutorial on variational autoencoders," 2016, cite arxiv:1606.05908. [Online]. Available: http://arxiv.org/abs/1606.05908

[21] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, 2016, pp. 10–21. [Online]. Available: http://aclweb.org/anthology/K/K16/K16-1002.pdf

[22] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[23] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," *CoRR*, vol. abs/1511.01844, 2015.

[24] D. Hendrycks and S. Basart, "A quantitative measure of generative adversarial network distributions," 2017.

[25] A. Kumar, A. Biswas, and S. Sanyal, "ecommercegan : A generative adversarial network for e-commerce," *CoRR*, vol. abs/1801.03244, 2018.

Learning Optical Flow For Action Classification

Çağatay Odabaşı[1]

***Abstract*— Action recognition is the task of assigning labels to human actions. It is particularly important for service robots because they need to react to human actions. For instance, if the user is cooking, the robot can offer to bring him some tools such as a pan or knife. In this work, the possibility of learning action recognition and optical flow extraction simultaneously using 3D Convolutional Neural Networks is analyzed. The preliminary results show that it is possible to learn two tasks together, but the proposed architecture needs further improvements.**

I. Introduction

People are performing a lot of different actions such as repairing something, walking, watching tv, eating, sleeping, taking medications, cooking, taking care of a baby. While performing them, they would need some external help. For instance, in Fig. 1, Lisha is taking care of a baby. So, she may need some help. In these situations, giving commands to robot would be too hard and the person may choose to not ask help from the robot. If the robot can understand these needs just by observing the people, it can offer some help without any explicit command. This would make user's life easier.

The action recognition is a task of labeling the data stream (video, image, skeleton, etc.) with an appropriate label such as walking, watching tv, etc. The main input source is a video stream which contains both spatial and temporal information. That's why the general approach is to exploit both domains to achieve high accuracy. To do this, the optical flow, which is an entity representing the displacement of certain part of the image, should be extracted from the video, because it contains rich temporal information content. The current focus is learning both tasks by using Convolutional Neural Networks (CNN) [11].

In machine learning, the loss function is an entity that assigns some cost to certain conditions. When the optimizer minimizes it, it is expected that the network will behave as requested. For example, if the loss function penalizes the misclassified actions, it is expected that the network will learn how to classify the actions correctly when the associated loss function is minimized.

In this work, the main aim is to investigate in learning optical flow and action recognition tasks simultaneously. This would allow us to train the action recognition network on smaller datasets. To do this, a new loss function including both action recognition and optical flow losses is generated. [20] proposed to use a similar cost function; however, they split the problem into two tasks. Instead of splitting the training procedure, the proposed network in this paper is trained by minimizing one joint loss function. The optical flow image is learned internally. So, the sole output of the network is the action recognition label.

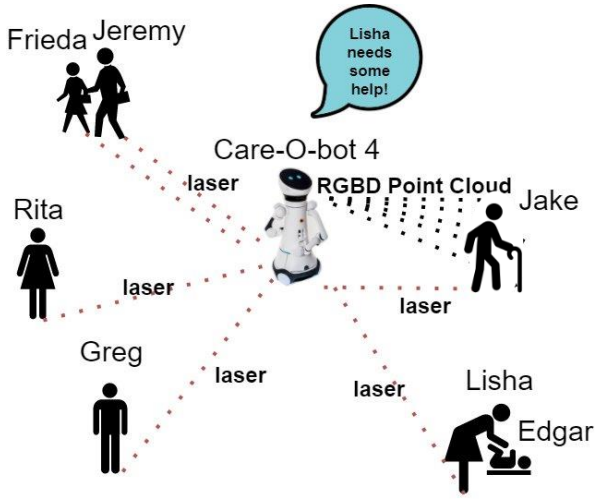


Fig. 1. Service Robots need to observe the people in the environment; so that, they can offer help.

The organization of this paper is as following. First, the related work is introduced to the reader in Section II. Second, the theory behind the proposed network is explained to the reader in Section III. Third, The implementation details are given in Section IV. Also, the preliminary results are presented and discussed in Section IV. Lastly, the paper is briefly concluded in Section V.

II. Related Work

Classical action recognition approaches [19], [18] are tracking temporal trajectories by using optical flow and then they are extracting spatial information such as HOG(Histogram of Oriented Gradient) [2] or spatiotemporal features such as HOF(Histogram of Oriented Flow) [10], MBH (Motion Boundary Histogram) [19] around these temporal trajectories.

The most common CNN architectures for action recognition are two streams networks for RGB and optical flow images [15], [1], [20] and 3D CNN which can convolve the video both in spatial and temporal domains [17], [3], [14]. Even though these approaches outperform on big datasets such as Kinetics [7], Youtube1M [6], they cannot reach the level that hand-crafted features based methods reached on relatively small datasets such as UCF101 [16] and HMDB51 [9]. One possible approach to this problem is fine-tuning. In

[1] The author is with Robot and Assistive Systems Department at Fraunhofer IPA, 70569 Stuttgart, Germany cagatay.odabasi@ipa.fraunhofer.de

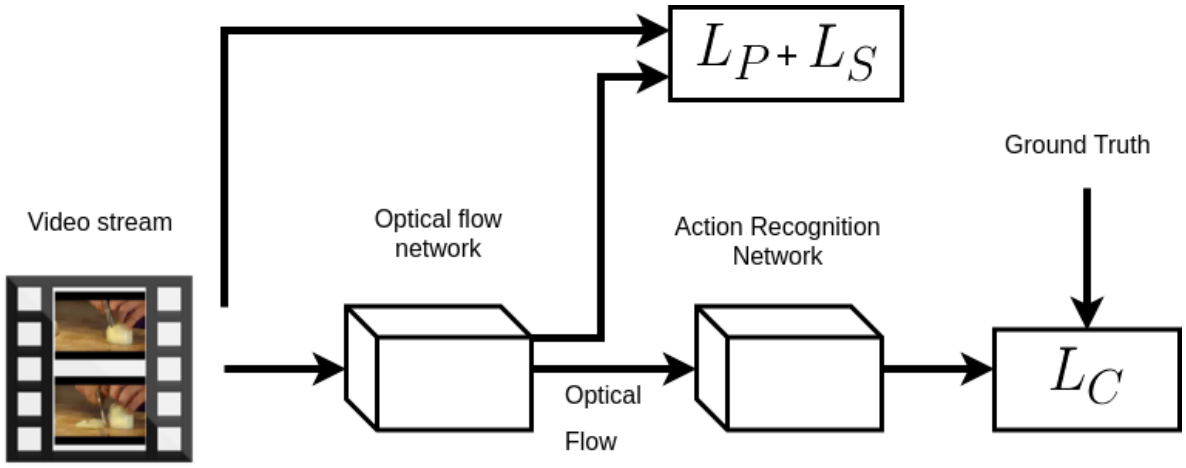


Fig. 2. The optical flow network accepts the video stream as input and its output is optical flow which is directed to the action recognition network. During the training, the optical flow loss and classification loss are combined and minimized together. The ground truth is true action recognition labels provided by dataset.

this approach, first, the network is trained on big datasets, then the classifier part of the network is retrained on a small dataset of interest. This approach would work if the datasets are similar to each other. If the small dataset differs dramatically from the big one, then there is no point of fine-tuning. Also, gathering a high amount of video data would be time-consuming. Another approach would be using optical flow as an input to the network which makes it possible to train on a small dataset [12].

For calculating optical flow efficiently inside the network, it's been proposed to use classical optical flow equations as a loss function [5], [20]; so that, the network can learn how to extract optical flow just by using input frames along with its own output without any supervision. They are basically warping the second frame by using the estimated optical flow and a differentiable warping function. Then, this warped image is extracted from the first image to calculate the photometric error.

In this work, the structure proposed in [20] is used; however, this work differs from them by trying to learn optical flow and action recognition simultaneously rather than two different tasks and also 3D convolutions are used rather than 2D, since 3D convolutions are capable of exploiting the spatial and temporal domains simultaneously by convolving them in both domains.

III. Methodology

A. Estimating Optical Flow

The part of the network that estimates the optical flow is called optical flow network. The rest is called action recognition network. The optical flow network is a 3D convolution based encoder-decoder network which is placed in front of the action recognition network as in Fig.2. The optical flow network takes three consecutive frames at a time and produces one optical flow for each set of input. Let's call these grayscale input frame at time step k as I_k. Clearly, the optical flow network is fed by I_{k+1}, I_k and I_{k-1}.

The loss function should be adjusted so that when it is minimized, the network should learn both optical flow and classification loss. Classification loss is the classical cross entropy loss function so let's refer to it as $\mathcal{L}_\mathcal{C}$. It penalizes misclassified actions.

Our optical flow loss function consists of photometric loss and smoothness loss. Smoothness loss is required for the aperture problem, so smoothness loss will force the system to learn just small motions.

The photometric loss is defined as:

$$\mathcal{L}_\mathcal{P} = \rho(W(I_k, O_k) - I_{k-1}) \tag{1}$$

where $\rho = (x^2 + \epsilon^2)^{1/p}$ is the Charbonnier cost, $W(I_k, O_k)$ is the warping function which warps the input image I_k by using optical flow O_k, so that, it will be identical to I_{k-1}. The implementation of this function is adapted from grid sampler of [4]. The warping function is sampling one pixel for each pixel position in the new image from the input image by using a flow field. This flow field indicates the displacement of each pixel of the input image.

The smoothness loss is defined as:

$$\mathcal{L}_\mathcal{S} = \rho(\nabla_x O_{x,k}) + \rho(\nabla_y O_{x,k}) + \rho(\nabla_x O_{y,k}) + \rho(\nabla_y O_{y,k}) \tag{2}$$

where ∇_x, ∇_y are the horizontal and vertical gradient operators applied to horizontal$O_{x,k}$ and vertical$O_{y,k}$ components of optical flow.

The general loss function can be written as:

$$\mathcal{L}_\mathcal{G} = \alpha_1 \mathcal{L}_\mathcal{C} + \alpha_2 \mathcal{L}_\mathcal{P} + \alpha_3 \mathcal{L}_\mathcal{S} \tag{3}$$

where $\alpha_{1,2,3}$ are manually selected constants which arranges the magnitudes of different losses. Therefore, the optimizer minimizes the joint loss $\mathcal{L}_\mathcal{G}$.

network can fit the training set completely, but it cannot get good results on validation dataset. To avoid this, we introduce Color Jitter on training videos. After each frame is

normalized which is a general method for most of the CNNs, the Gaussian noise is added to them as below:

B. Stacking the optical flow images

The proposed architecture is presented in Fig.2. As seen, the only modality that action recognition network uses is the output of the optical flow network. Since the action information is spread through the entire or a part of the video, it is not possible to make a good prediction with just one optical flow. Therefore, the optical flow outputs must be stacked.

In both training and testing mode, the network accepts a fixed number of sequential frames. These frames are divided into smaller groups where each group contains three frames. Each of these groups is sent to the optical flow network and as a result, it produces one optical flow image. These optical flow images are stacked in the time axis and are sent to the action recognition network to get the action recognition output.

IV. Results

A. Implementation Details

For the training part, 2 Nvidia GTX1080Ti with a batch size of 256 are used. Also, the image size is reduced to 56x56 to make the system memory efficient. The action recognition is done by using 3D-Resnet-18 proposed in [3]. A small 3D CNN network is added in front of it to infer the optical flow. It is kept as simple as possible due to computation power. As optimizer, we use Adam [8] presented in PyTorch framework [13], since Adam is easier to tune than Stochastic Gradient Descent (SGD) and its performance is comparable to SGD.

B. Evaluation

In this section, the preliminary results are presented. Therefore, a detailed evaluation should be carried out to assess the optical flow and action recognition performances. The aim is to maximize the classification accuracy of the system while minimizing the optical flow loss. That's why the loss and accuracy results of the model are presented. Note that the results are preliminary. The architecture is still in development. Current action recognition scores cannot reach the state-of-the-art level which is around 45% without fine-tuning. However, they prove that it is possible to learn two tasks simultaneously.

In Fig. 3, the training accuracy, training loss, validation accuracy, and validation loss are presented. The final values of the loss function on both sets are around 40. This means that the network can generalize well. However, the validation accuracy can reach up to 27%, although the training accuracy is around 90%. Therefore, we can conclude that the problem is in learning action recognition rather than optical flow. There could be several reasons for this. The most important one would be the stacking the optical flow images. The number of frames stacked would not be enough. On the other side, if the number of frames is increased, the computation cost and memory consumption increase dramatically.

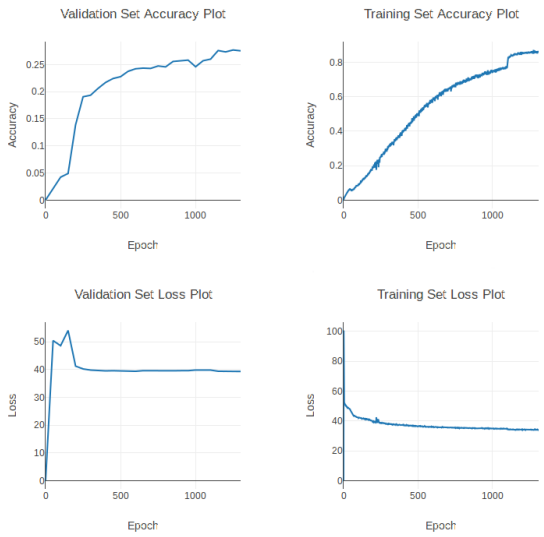


Fig. 3. The accuracy and loss results of action recognition network with optical flow part. The accuracy is normalized to [0,1]. So, the accuracy percentage can be calculated by multiplying the values with 100.

V. Conclusion

In this work, the possibility of learning unsupervised optical flow and action recognition tasks simultaneously is tested. To do this, a joint loss function is created which consists of both optical flow loss and the action recognition loss. Minimizing the overall loss function would allow the network to learn two tasks simultaneously.

The proposed method would allow us to learn with smaller datasets and our results show that the joint loss function can be minimized simultaneously. However, the action recognition performance is still too low, hence it needs some further analysis.

In future works, the loss function will be analyzed in detail. The findings will help us to optimize the architecture.

VI. Acknowledgement

This project is supported by SOCRATES project which is a MSCA-ITN-2016 Innovative Training Networks funded by EC under grant agreement No 721619. Also, I want to thank all of my colleagues in the Domestic and Personal Robotics Group at Fraunhofer IPA for their ideas and support.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*, volume 2, page 4, 2017.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[5] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016.

[6] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering 12*, pages 571–582. Springer, 2013.

[10] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[11] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, pages 319–345. Springer, 1999.

[12] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for action recognition. *arXiv preprint arXiv:1612.03052*, 2016.

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[14] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5534–5542. IEEE, 2017.

[15] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[17] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015.

[18] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[19] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[20] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G Hauptmann. Hidden two-stream convolutional networks for action recognition. *arXiv preprint arXiv:1704.00389*, 2017.

Natural Language Communication with Social Robots for Assisted Living*

Maitreyee Tewari[1] and Suna Bensch[2]

***Abstract*—We explore a new dialogue modelling approach for assistive social robots that could facilitate flexible conversation flows between a robot and a human. We propose to model topic change, clarification questions or misunderstandings during a dialogue, by introducing an expectation mechanism. Our approach formalizes the formation of a dialogue as a cooperation between two dialogue participants. We gain insight into the dialogue structure and how it could be shaped by several linguistic and pragmatic features. This a work in progress and a next immediate step is to implement and evaluate the model for conversations between a human and a robot.**

***Index Terms*—natural language communication, robots, assisted living, dialogue management, turn-taking, cooperation, human-robot interaction.**

I. Introduction

The demographic trend of an aging population is a challenge for the health care system in western countries. Social robots as assistive technology can support care-givers and enable older adults to live longer independently at home and improve quality of life [1]. For the integration of assistive social robots, it is important that they converse naturally with us.Therefore, such robots must interpret and react to human behaviour including gesturing, displaying emotions, and using natural language to conduct a dialogue (we look into only natural language aspects). A robot that is used in the context of elder care has to adapt to the varying and unpredictable nature of dialogues, such as sudden topic changes, misunderstandings, incomplete or inaccurate information (non-understanding), interruptions, humour and opposition. We introduce a new formal dialogue model that formalizes dialogue turns and sudden topic changes to allow flexible dialogue flows between a robot and a human and provides *insight* into the dialogue structure. We believe that assistive social robots should have robust and understandable dialogue management techniques, such that we can *interpret* the robot's behaviour during dialogues and modify it if necessary.

The formal model *co-operating distributed grammar systems with expectations* (CDGS_{exp} for short) is based on co-operating distributed grammar systems (CDGS) [2]. Such systems model cooperation among several agents that have a common goal. We consider a dialogue between a robot and a human as cooperation between two agents who have the common goal of conducting a successful dialogue. In the latter we refer to dialogue participants (human and robot) as agents. Expectations are anticipations of certain information that agents have when conducting a dialogue. For example, an agent A can expect that another agent B confirms agent A's request or answers agent A's question. We formalize expectations as internal control mechanism bounded by a given time frame. The time frame can be a measure of the number of turn takes during a dialogue or discrete time unit steps. The internal control mechanism enables flexible dialogue flows as it gives agent the possibility to not meet expectations immediately but, for example, change the current topic of conversation. CDGS_{exp} controls the dialogue flow according to the agent's expectations, to describe the agent's perspective during a dialogue, and to model the overall dialogue structure and its formation. In our approach we also shed light into several linguistic and pragmatic features that influence the dialogue structure.

*This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

[1]M.Tewari is with Department of Computing Science Umeå University Umeå, Sweden maittewa@cs.umu.se

[2]S.Bensch is with Department of Computing Science Umeå University Umeå, Sweden suna@cs.umu.se

II. Background

Dialogue management approaches are generally based on finite-state and data-driven methods [3]. For dialogue modeling, the data-driven approaches can easily become intractable because of the complexity of dilaogues (several agents contributing, different topics being discussed, giving turns and taking turns). On the other hand finite-state based approaches manually define how to conduct a dialogue and thus provide valuable insight into the dialogue structure, but manual definition of dialogue rules is time and labor costly. We are interested in developing a hybrid dialogue model [4], [5] that learns from data (the pragmatic and syntactic features) the dialogue structure and a formal model that allows us to add, delete or alter dialogue rules. As a first step, we focus on developing a suitable formal model for dialogues based on co-operating distributed grammar systems which are finite-state devices. A variant of CDGS is called *eco-grammar system* [6] and has been used to model dialogues for multi-agent systems. Their work was inspired by multi-agent protocol language, and provides flexible and adaptable reaction to unpredictable conversational space. In [7] the authors propose an extension, namely *reproductive eco-grammar system*, where the grammars follow a multi-agent protocol language to determine *which* social norms should be used to participate in a conversation. In [8] authors propose *conversational grammar systems*, which mimics natural language to define a formal model for dialogues. In [9] turn-taking behavior in dialogues is modelled with *CDGS with memories*. We extend CDGS with

Topics	Dialogue acts	SO	Agents/Utterances/Keywords	Nr.
GREET	OPENING		R: Hi Anna. How are you?	1
GREET	OPENING		A: Hi. Pretty good.	2
MEDICATION	REQUEST		R: Please make sure to take your pills.	3
JIM	QUESTION		A: Did you see Jim?	4
JIM	ANSWER		R: He was here this morning.	5
JIM	OFFER		R: Do you want me to call him?	6
JIM/HEALTH	FOLLOWUP		A: I want him to check my blood pressure.	7
JIM	OFFER		R: Ok. I'll let him know.	8
MEDICATION	REQUEST		R: Did you take your pills, Anna?	9
	AGREE		A: Right away.	10

Fig. 1. A fictional sample dialogue between a robot (R) and an older adult named *Anna* (A). The dialogue is analyzed based on several linguistic and pragmatics features, namely *topics*, *dialogue acts*, *sequence organization* (SO) and *keywords* (which are underlined).

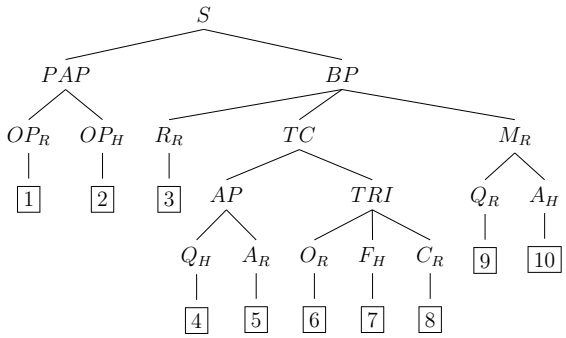


Fig. 2. The tree structure of the dialogue displayed in Figure 1. The leaf nodes are labelled with the numbers of the utterances in Figure 1.

an expectation mechanism and consider a certain set of linguistic and pragmatic features from which we can infer some aspects of the dialogue structure. In the first three models, eco-grammar systems were modified to provide flexibility to dialogues. This method could pose complexity in integrating it with the data-driven methods. For the latter model CDGS with memories, instead of memories we are attempting to build an internal control mechanism, more inclusive than memories. It would not only manage turn-takes, but also sudden topic changes, and other dialogue phenomena.

III. Methodology

A. Linguistic and pragmatic features for dialogue analysis

We consider dialogues as sequences of utterances, consisting of one or more sentences, aligned one after the other by participants through turn-takes. Consider the fictional dialogue in Figure 1 between a robot (R) and an older adult named *Anna* (A) in a health care facility. The dialogue is displayed in the fourth column "Agents/Utterances". We refer to the individual utterances with numbers which are displayed in the fifth column "Nr.", where an utterance can consist of one or more sentences. The dialogue starts with the two agents greeting each other (Utterances 1-2). Then the robot reminds Anna politely to take her pills (Utterance 3). Anna instead of answering the request (Utterance 3), changes the topic by asking whether the robot has seen Jim (Utterance 4). The robot answers Annas' question and offers to call Jim (Utterances 5-6). Anna then states that she wants Jim to check her blood pressure (Utterance 7) which is indirectly also an acceptance of the robot's offer to call Jim. The robot confirms that it will let Jim know that Anna wants to see him (Utterance 8). Then the robot reminds Anna again about her medicine (Utterance 9) which Anna promises to take right away (Utterance 10).

We analyze a dialogue considering the following linguistic and pragmatic features of utterances: topics, dialogue acts, sequence organization and keywords. We explain all four features briefly in the following.

The first column in Figure 1 shows some topics of the utterances. Topics determine the major constituent of an utterance. The second column shows the so-called dialogue acts [10] associated with each utterance. An utterance is a dialogue act if it has a communicative function, which specifies an activity performed in the dialogue such as asking a question, requesting information, accepting or rejecting a request or making a declaration. The third column in Figure 1 illustrates a possible sequence organization of the utterances in the dialogue. Sequence organization (SO for short) is empirically studied in *conversation analysis* [11]. Sequence organization describes how sequences of utterances can be ordered. In Figure 1 utterances forming a sequence with each other are connected by the displayed orange lines. For example, if two utterances occur consecutively (e.g. question-answer, greeting-greeting) then they can be described as *adjacency pair*. In Figure 1, Utterances 1 and 2 form an adjacency pair. Utterances do not have to be necessarily adjacent to each other, they can occur apart from each other in a dialogue and are then ordered as *First-Pair-Part* (FPP) and *Second-Pair-Part* (SPP). Furthermore, sequence of utterances can be categorized into three types of so-called *expansions*, namely base, insert and post [12]. In Figure 1, Utterance 3 is FPP_{base}) and is connected to the adjacency pair of Utterances 9-10 (which is the corresponding SPP, namely SPP_{base}). The topic change (i.e. *Did you see Jim?*) by Anna expands the so-called base sequence and introduces $\text{FPP}_{\text{insert}}$, which is followed by the robots response generating $\text{SPP}_{\text{insert}}$ (i.e. Utterances 4 and 5, respectively). Another feature that can influence how a dialogue is structured are frequently used words or phrases (i.e. keywords). In Figure 1 in the fourth column, keywords are underlined in blue or red (for domain specific words or phrases). Such keywords can facilitate dialog act or topic association and thus influence the structuring of the dialogue.

B. Inferring the tree structure of the dialogue

All the features elaborated so far (e.g. topics, dialogue acts, keywords, sequence organization) are organized into a tree structure which is illustrated in Figure 2. A tree structure serves the following two purposes:

1) To describe the overall dialogue structure based on the linguistic and pragmatic features, and
2) To extract rules for our model CDGS_{exp}.

The tree in Figure 2 illustrates that our example dialogue consists of two larger parts, namely an introduction into the dialogue represented by the subtree rooted at the node labelled

by PAP (e.g. greeting and asking about well-being) and a main part represented by the subtree rooted at the node labelled by BP. The topic change is represented by the subtree rooted at node TC. The leaves of the tree are labelled with the numbers of the individual utterances that can be found in Figure 1. The parent nodes (e.g. $OP_R, OP_H, R_R, Q_H, A_R$) are labels for the dialogue acts and one can restore the order in which they were uttered too. Note that the subtree with root label TC is a subtree that can only be formed by taking into account the topic change.

C. Formal background

In this section we provide the necessary definitions of co-operating distributed grammar systems (CDGSs). A CDGS consist of several so-called *components* that *work by taking turns* according to some *cooperation protocol*. The cooperation protocol defines when components are allowed to start and stop working. The components in a CDGS can be interpreted as agents working together with a *common aim* (e.g. to solve a problem).

Definition 1: A CDGS of degree n, with $n \geq 1$, is an $(n+3)$-tuple $G = (N, T, C_1, C_2, \ldots, C_n, S)$, where, N is a set of variables (called non-terminal symbols), T is a set of constants (called terminal symbols), S is the start symbol, for $1 \leq i \leq n$, C_i is a set of rules of the form $A \to \alpha$, where $A \in N$ and α is a string consisting of variables and/or constants (i.e. $N \cup T$). A rule $A \to \alpha$ means that a variable A can be replaced with the string α. The set of rules $C_1, C_2, \ldots, C_n$ are called *components*.

Example 1: Let $\hat{G} = (\{S, A, B\}, \{a, b, c, d\}, C_1, C_2, S)$ be a CDGS grammar, where

$$C_1 = \{S \to aA, B \to aA, A \to aA, A \to a\},$$

$$C_2 = \{S \to bB, A \to bB, B \to bB, B \to b\}.$$

Definition 2: Let $G = (N, T, C_1, C_2, \ldots, C_n, S)$ be a CDGS. For two strings x, y in $(N \cup T)$ and $1 \leq i \leq n$, we write $x \Longrightarrow_i y$ and say that y is derived in one derivation step from x by component C_i, if and only if $x = \gamma_1 A \gamma_2$ and $y = \gamma_1 \alpha \gamma_2$ for some $\gamma_1, \gamma_2 \in (N \cup T)$ and there exists a rule in C_i of the form $A \to \alpha$. A *derivation* (i.e. successive derivation steps) starts with the string S (i.e. the start symbol of G) and ends when a string w is obtained that consists only of terminal symbols.

The cooperation protocol for a CDGS can state that a component can make exactly k derivation steps, $\leq k$ steps, $\geq k$ steps, arbitrary many steps ($*$ cooperation protocol) or take the maximal number of derivation steps possible (t cooperation protocol).

Example 2: Let $\hat{G}$ have the cooperation protocol $= 2$, that is, each component must make exactly two derivation steps before the other component starts to work. The derivation starts with the start symbol S. Both components C_1 and C_2 can rewrite the start symbol S (by applying the rules $S \to aA$ or $S \to bB$, respectively). Let us assume that C_1 starts to work. The component C_1 has to make two derivation steps, $S \Longrightarrow_1 aA \Longrightarrow_1 aaA$, that is, first rewriting S by applying the rule $S \to aA$ and then rewriting A (in the string aA) by applying the rule $A \to aA$. Now component C_2 has to start rewriting and make two derivation steps. Let us assume the derivation $aaA \Longrightarrow_2 aabB \Longrightarrow_2 aabb$. That is, C_2 applied the rule $A \to bB$ to the string aaA generating the string $aabB$ and then applied the rule $B \to b$ to the string $aabB$ generating the string $aabb$. The string $aabb$ is a *terminal string* and consists only of terminal symbols and cannot be rewritten further.

This example illustrated how components generate strings by taking turns after two derivation steps according to the given cooperation protocol. Note that the components can generate many different terminal strings (e.g. $aabb$, bb, $aabbaabb$).

D. CDGS modeling expectations

In this section we provide the definitions of our new model CDGS_{exp} and apply it to the dialogue example in Figure 1 and show how the tree in Figure 2 is generated. We assume that a CDGS_{exp} works in $*$ cooperation protocol with the addition that an agent A starts working if the other agent B did not meet the expectation of agent A within a given time frame. An agent A stops working whenever it is ready to "hand the floor" to agent B. In a CDGS_{exp} a non-terminal symbol A on the right hand side of a rule may be extended with $\leq k$, where k is a positive integer, that is, $A[\leq k]$. The $\leq k$ in $A[\leq k]$ represents the time frame in which the other component is expected to rewrite the non-terminal A. The time frame measures the number of turn takes during a dialogue or of derivation steps. In the following example, we count the number of derivation steps each agent makes[1]. For example, if an agent C_1 applies a rule of the form $B \to aA[\leq 5]$, it represents that agent C_1 expects the other agent C_2 to rewrite symbol A within the next 5 derivation steps C_2 makes. If the other component does not rewrite the non-terminal that is expected to be rewritten within the given time frame, then the component that has the expectation starts working and applies a new rule with the same expectation. That is, if, for example C_1 applied the rule $B \to aA[\leq 5]$ and the component C_2 does not rewrite A within five steps, then component now C_1 applies a rule $A \to aA[\leq 2]$ and expects C_2 to rewrite the symbol A within its next two derivation steps. Let $\gamma_1 A \gamma_2$ be a string for some $\gamma_1, \gamma_2 \in (N \cup T)$ and let $r : A \to \alpha[\leq k]$ be a rule r in a component C_i. Then C_i derives y by applying r as follows: $\gamma_1 A \gamma_2 \Longrightarrow_i \gamma_1 \alpha \gamma_2 = y$. That is, $[\leq k]$ is not introduced into a string but only appears in C_i.

In our scenario where we consider assistive robots with conversational capabilities, this serves the purpose to give the older adult the freedom to react flexibly, and at the same time, ensure that the robot picks up a topic again if it's important and has not been answered by the older adult (see Utterance 3 and Utterance 9 in Figure 1).

The following example is simplified but should give the idea of how the tree in Figure 2 is generated in cooperation

[1]Note that we can just as easily count the number of turns each agent makes by defining a turn of an agent A as an application of a rule of the form $A \to a$, where $a \in T$ for a given CDGS_{exp} and component A.

between an agent C_1 (representing the robot) and an agent C_2 (representing the older adult). We assume that the CDGS_{exp} works in a leftmost derivation fashion, that is, it always rewrites the leftmost occurring symbol in a string. We can associate to each derivation a derivation tree.

Example 3: Let $\hat{G} = (N, T, C_1, C_2, S)$ be a CDGS_{exp}, where $N = \{S, PAP, BP, OP_R, OP_H, R_R, TC, M_R, AP, TRI, Q_R, A_H, Q_H, A_R, O_R, F_H, C_R\}$ (that is, all labels of the inner nodes in the tree in Figure 2), $T = \{\boxed{1}, \boxed{2}, \ldots, \boxed{10}\}$ (that is, all utterances given in Figure 1) and C_1 and C_2 contain the rules shown in Figure 3 (we number all rules for easier reference):

C_1	C_2
$\{r_1 : S \to PAP\ BP,$	$\{r_1 : OP_H \to \boxed{2},$
$r_2 : PAP \to OP_R\ OP_H[\leq 1],$	$r_2 : TC \to AP\ TRI,$
$r_3 : OP_R \to \boxed{1},$	$r_3 : AP \to Q_H\ A_R[\leq 1],$
$r_4 : BP \to R_R\ TC\ M_R[\leq 5],$	$r_4 : Q_H \to \boxed{4},$
$r_5 : R_R \to \boxed{3},$	$r_5 : F_H \to \boxed{7},$
$r_6 : A_R \to \boxed{5},$	$r_6 : A_H \to \boxed{10}\}$
$r_7 : TRI \to O_R\ F_H[\leq 2]C_R,$	
$r_8 : O_R \to \boxed{6},$	
$r_9 : C_R \to \boxed{8},$	
$r_{10} : M_R \to Q_R A_H[\leq 2],$	
$r_{11} : Q_R \to \boxed{9}\}$	

Fig. 3. The components C_1 and C_2 for the CDGS_{exp} $\hat{\hat{G}}$ in Example 3.

The robot initiates the dialogue which is represented by C_1 applying the rules r_1, r_2, r_3 in C_1, that is, $S \Longrightarrow_1 PAP\ BP \Longrightarrow_1 OP_R\ OP_H\ BP \Longrightarrow_1 \boxed{1}\ OP_H\ BP$. The symbol $\boxed{1}$ represents the Utterance 1 in Figure 1 by the robot. The component C_1 expects C_2 to rewrite the symbol OP_H within one derivation step. The component C_2 rewrites OP_H by applying rule r_1 in C_2, i.e. $\boxed{1}\ OP_H\ BP \Longrightarrow_2 \boxed{1}\ \boxed{2}\ BP$. The component C_1 applies the rules r_4, r_5 generating the string $\boxed{1}\ \boxed{2}\ \boxed{3}\ TC\ M_R$. The variable TC allows C_2 to change the topic. The derivation is continued in this fashion until we obtain the terminal string $\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}\ \boxed{7}\ \boxed{8}\ \boxed{9}\ \boxed{10}$, that represents the dialogue in Figure 1. and the tree in Figure 2

In our model, expectations are not restricted to only expecting certain dialogue acts (as in the above example) but can be topic changes too.

IV. Conclusions and future work

We proposed a new dialogue model for assistive social robots that can allow flexible conversation flows. Our expectation mechanism can allow that dialogue goals are met, but at the same time dialogues can be diverted (for now through sudden topic change) Our hybrid model has the following additional advantages compared to sole finite state approaches or data-driven approaches for dialogue models:

- We describe dialogue as a cooperation among agents instead of only capturing the machine's perspective.
- We gain insight into the structure of dialogues and in its formation.
- Our approach is extendable to several agents and can serve as models for human robot communication in which several robots and humans can communicate.

This paper reports work in progress and in the future we want to develop algorithms that learn how to map sets of features such as topics, dialogue acts, keywords, sequence organization into dialogue structures such as the one displayed in Figure 2. Once this is achieved, a CDGS_{exp} with expectations can be generated. We are interested in further investigating how our model can handle dialogue phenomena such as misunderstandings, non-understandings or opposition. Another of our tasks is an implementation of our formal model to test its validity and limitations.

References

[1] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: a survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229 – 252, 2015.

[2] E. Csuhaj-Varju, J. Kelemen, G. Paun, and J. Dassow, Eds., *Grammar Systems: A Grammatical Approach to Distribution and Cooperation*, 1st ed. Newark, NJ, USA: Gordon and Breach Science Publishers, Inc., 1994.

[3] C. Lee, S. Jung, K. Kim, D. Lee, and G. Lee, "Recent Approaches to Dialog Management for Spoken Dialog Systems," *Journal of Computing Science and Engineering*, vol. 4, pp. 1–22, 2010. [Online]. Available: http://www.dbpia.co.kr/Article/NODE01431934

[4] D. Traum, "Talking to virtual humans: Dialogue models and methodologies for embodied conversational agents," in *Modeling Communication with Robots and Virtual Humans*. Springer, 2008, pp. 296–309.

[5] P. Lison, "A hybrid approach to dialogue management based on probabilistic rules," *Computer Speech & Language*, vol. 34, no. 1, pp. 232–255, 2015.

[6] E. Csuhaj-Varjú, J. Kelemen, A. Kelemenová, and G. Păun, "Eco-grammar systems: A grammatical framework for studying lifelike interactions," *Artif. Life*, vol. 3, no. 1, pp. 1–28, Mar. 1997. [Online]. Available: http://dx.doi.org/10.1162/artl.1997.3.1

[7] G. Bel-Enguix, M. Grando, and M. Jiménez-López, "A grammatical framework for modelling multi-agent dialogues," *Agent Computing and Multi-Agent Systems*, pp. 10–21, 2006.

[8] G. Bel-Enguix and M. Jiménez-López, "Modelling dialogue as interaction," *International Journal of Speech Technology*, vol. 11, no. 3, pp. 209–221, 2008.

[9] S. Aydin, H. Jürgensen, and L. E. Robbins, "Dialogues as co-operating grammars," *Journal of Automata, Languages and Combinatorics*, vol. 6, no. 4, pp. 395–410, 2001.

[10] H. Bunt, "Dimensions in dialogue act annotation," *Proceedings of Language Resources and Evaluation Conference*, vol. 6, pp. 919–924, 2006.

[11] J. Sidnell and T. Stivers, *The handbook of conversation analysis*. John Wiley & Sons, 2012, vol. 121.

[12] S. Tanya, *Sequence Organization*. Wiley-Blackwell, 2012, ch. 10, pp. 191–209.

Towards measuring mental workload from facial expressions

Andre Potenza
Center for Applied Autonomous Sensor Systems (AASS), Örebro University
Email: andre.potenza@oru.se

***Abstract*— Multitasking is a common issue negatively impacting performance in robotic teleoperation and in particular, as we argue, in telepresence. Operating a telepresence robot typically involves engaging in a social interaction with other people who are collocated with the robot, while simultaneously having to control the robot, possibly resulting in an elevated mental workload. One way to mitigate this adverse effect is to have the telepresence robot execute certain tasks autonomously - when necessary. In this extended abstract, we discuss how mental workload measurements can contribute towards dynamically allocating tasks between user and robot, so that a high performance can be achieved, ideally throughout all ongoing tasks. To this end, we are proposing a method for estimating users' mental workload from facial expressions via learned models.**

INTRODUCTION

Remotely controlling a robot in a distant environment requires training with the control interface, and insufficient sensory input causes users to operate with limited information about the robot's surroundings. Moreover, this control task is often accompanied by additional activities, such as social interaction in the case of robotic telepresence [1]. The result is an increased mental workload and possibly diminished situation awareness, as users may have difficulties taking in and processing all relevant information that is available to them. If the operator's workload capacity is exceeded, this, in turn, can lead to reduced performance in one or all of the tasks being performed [2].

Indeed, some of the above-mentioned challenges can be mitigated by upgrading the teleoperated robot with capable sensors and efficient user interface design [3], by way of which a high level of situation awareness (SA) can be attained and retained with less effort [4]. However, the problem of multitasking persists regardless, and with the expectation of future telepresence robots providing enhanced actuation capabilities beyond navigation, users' workload is projected to increase even further. We argue that one potential solution to this issue could be found in mixed-initiative adjustable autonomy [5], in which the robot can decide to reallocate a subset of its functions, if deemed necessary.

A mixed-initiative adjustable autonomy system allows both human and robot to initiate a handover or takeover of functions or entire tasks. While the human user may trigger such a shift for any reason and at any point, the robot is required to have clear, predefined and measurable criteria to decide when and which task should be reallocated.

If a given task can be performed reasonably well by both agents, i.e., human operator and robot, in at least a subset of all possible situations (e.g., navigation), it is eligible for dynamic assignment between them. In an adjustable autonomy system, we identify two primary sets of criteria for determining how the combined total of the system's functionalities should be distributed:

1) Task-specific: If a task is eligible for automation, the robot needs to monitor its execution perpetually, regardless of which agent is currently in control of it. If the task performance falls below a preset threshold (for longer than a preset duration), its control authority may be shifted.
2) User-specific: Humans possess an intrinsic, yet variable capacity for processing information that is available to them. Human factors research involves a set of cognitive constructs that describe users' capacity to understand and process information available to them based on a multitude of cognitive constructs. Those constructs include, among others, situation awareness (SA) [6], [7], mental workload [8], stress and fatigue.

Obviously, these two classes are not entirely exhaustive, as the difficulty and criticality of tasks may vary in dynamic environments, with implications for the preference of the respective agent being in control. That notwithstanding, for the above-mentioned use case of teleoperated robots, a combination of these two classes is required. For instance, it would not make sense to assign all tasks to the human operator just because they are better at all of them, as it could cause mental overload and consequentially result in a loss of overall system performance. Hence, a key requirement of an effective mixed-initiative adjustable autonomy system is a reliable evaluation of users' mental states.

Here, we discuss two constructs from the human factors and ergonomics research and their suitability as metrics for monitoring of the operator state.

HUMAN FACTORS MEASUREMENT

As a cognitive construct, SA plays a vital role in automation - particularly when it comes to deciding on the appropriate level of automation (LOA) of a task or functionality. In fact, SA, together with mental workload, are typically in a complex interplay with the LOA and the impact on all three of these factors needs to be carefully considered when designing a system [9]. As a general rule of thumb, both workload and SA can be expected to decrease as the LOA is increased. While a low level of workload is in most cases desired, a low level of SA can be detrimental to system performance, as automation is often imperfect and expected

to fail in some situations. When this happens, a high degree of maintained SA allows users to assess the situation quickly and take appropriate measures to guide the system back to a nominal state.

While its relevance in system design is well established, its essence and the ways in which it is commonly measured [10] are very closely related to task performance rather than the operator's general mental state. As such, it does not add much information to our user-specific criteria class. Since measurement of the operator state should be unintrusive, implicit and objective, for this purpose it appears worthwhile to examine mental workload more closely.

Mental workload is a well-studied cognitive concept, researched in a variety of areas ranging from cognitive psychology to applied sciences such as user design. Yet, and even though almost everybody has an intuitive idea of what it denotes, there exists no single universally accepted definition of it in the literature [8], [2]. For most purposes, it could be described as the relative degree to which an individual's personal mental processing capacity is exhausted by the entirety of the mental processing that they are performing at a given time. Thus, if their capacity is exceeded and another task occupying the same mental resources is added, the performance in at least one of the currently performed tasks is expected to decline. In fact, it has been shown and is worth noting that tasks of disparate nature (e.g., spatial vs. verbal, visual vs. auditory) do not necessarily occupy the same attentional resources and may be performed simultaneously without interfering with one another [11].

In experimentation, a common way of recording subjects' workload is subjective self-reports at several points throughout an experiment. Arguably the most widely used tool for such reports is the NASA-TLX (Task Load Index) questionnaire [12], which allows subjects to rate perceived task difficulty and workload across multiple dimensions.

On the other hand, various physiological measures have been used to estimate workload objectively [13], ranging from slightly invasive (e.g., heart rate variability [14], skin conductance etc.) to more invasive (EEG [15]). While some of these methods have shown success, they lack practicability for casual users of telepresence robots. Since in any telepresence robot a camera is recording the operator's face by design, we intend to investigate the possibility of estimating users' workloads from facial expressions.

Proposed Method and Experiment

In recent years, deep neural networks have been applied to detect a variety of features from facial expressions, such as emotions [16], gender and age [17], arousal, etc.

In the proposed experiment we plan to have participants perform mental tasks of varying difficulty levels and intermittently report their experienced workload levels via the NASA-TLX [12]. Throughout the experiment, their faces are recorded with a regular RGB camera. From the collected time series data (video footage) and the reports serving as ground truth, we will train a recurrent convolutional neural network [18] whose purpose will be to classify the workload of individuals over time series segments.

Conclusion and future outlook

In teleoperation, automation should be partial and selective, dynamically adapting to the current user's condition, skills and needs. We have discussed the two broad means based on which performance can be evaluated - task-specific or user-specific. The latter type, if measured accurately enough, can function as a support for autonomous decision making in overall task allocation between human and robot. For mental workload, several different approaches exist as they can be either subjective or objective, as well as more or less invasive. In this paper, we argued for an approach that better suits the typical requirements found in telepresence robotics. This is what we aim to investigate in an upcoming user study.

Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

References

[1] M. Desai, K. M. Tsui, H. A. Yanco, and C. Uhlik, "Essential features of telepresence robots," in *Technologies for Practical Robot Applications (TePRA), 2011 IEEE Conference on*. IEEE, 2011, pp. 15–20.

[2] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, vol. 58, no. 1, pp. 1–17, 2015.

[3] A. Kristoffersson, S. Coradeschi, and A. Loutfi, "A review of mobile robotic telepresence," *Advances in Human-Computer Interaction*, vol. 2013, p. 3, 2013.

[4] M. R. Endsley, "Designing for situation awareness in complex systems," in *Proceedings of the Second International Workshop on symbiosis of humans, artifacts and environment*, 2001, pp. 1–14.

[5] B. Sellner, F. W. Heger, L. M. Hiatt, R. Simmons, and S. Singh, "Coordinated multiagent teams and sliding autonomy for large-scale assembly," *Proceedings of the IEEE*, vol. 94, no. 7, pp. 1425–1444, 2006.

[6] M. R. Endsley, "Measurement of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 65–84, 1995.

[7] ——, "Toward a theory of situation awareness in dynamic systems: Situation awareness," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[8] B. Cain, "A review of the mental workload literature," Defence Research And Development Toronto (Canada), Tech. Rep., 2007.

[9] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs," *Journal of Cognitive Engineering and Decision Making*, vol. 2, no. 2, pp. 140–160, 2008.

[10] P. Salmon, N. Stanton, G. Walker, and D. Green, "Situation awareness measurement: A review of applicability for c4i environments," *Applied ergonomics*, vol. 37, no. 2, pp. 225–238, 2006.

[11] C. D. Wickens, "Multiple resources and mental workload," *Human factors*, vol. 50, no. 3, pp. 449–455, 2008.

[12] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[13] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with computers*, vol. 21, no. 1-2, pp. 133–145, 2009.

[14] A. Hoover, A. Singh, S. Fishel-Brown, and E. Muth, "Real-time detection of workload changes using heart rate variability," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 333–341, 2012.

[15] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using eeg spectral power and erps in the n-back task," *Journal of neural engineering*, vol. 9, no. 4, p. 045008, 2012.

[16] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[17] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

Multimodal Robot Feedback While Learning a Novel Cognitive Exercise From a Human Teacher

Aleksandar Taranović[1]

***Abstract*— Socially-assistive robots could help their users in essential daily activities. However, teaching these tasks to a robot usually requires domain-specific robot programming, and hence substantial time investment. User-oriented methods for teaching robots can accelerate the learning process. The robot should disclose obtained knowledge and understanding of the new skill while learning it. Moreover, the robot should inform the teacher what additional instructions are necessary. This paper proposes adaptation of the robot feedback to a human teacher through the use of different robot modalities in the context of cognitive therapy.**

I. INTRODUCTION

Socially-assistive robotics has the potential to improve the quality of life for various groups of people [1]. For example, robots can provide cognitive therapy that slows down the cognitive decline of people with dementia. However, the interaction must feel natural and meet the expectations of the user. Otherwise, the robot does not serve its purpose. This paper outlines the use of robot modalities while a human is teaching it a new skill, and how to exploit the multimodal character of the interaction in this context. Moreover, the focus is on adapting the use of robot modalities when the robot provides feedback to lay users that are teaching it a new cognitive exercise.

People have different teaching styles [2]. Therefore, we should design robots that adapt to account for these differences. The robot should provide feedback to human teachers so that they can comprehend the robots understanding of the task. In this paper, the focus is on teaching robots new cognitive exercises because it is beneficial for people to perform diverse exercises since it minimizes their boredom due to repetition.

II. COGNITIVE THERAPY

Alzheimer's disease and other types of dementia affect an increasing number of people [4], and robots can assist in providing cognitive therapy [5]. One type of therapy are exercises that stimulate different parts of the brain. For instance, memory is trained using exercises of recalling objects from sequences (Fig. 1) that usually have minor differences among them. For example, the user needs to sort a set of objects in a predefined sequence. Performing the exercise multiple times with the same shapes can be boresome. Therefore, the exercises could engage more if different sets of shapes are used. However, the change of shapes requires reprogramming of the robot behavior, and that process commonly requires the involvement of a robotics expert. Similarly, if the rules of sorting the sequence are modified, it would require additional programming by the expert.

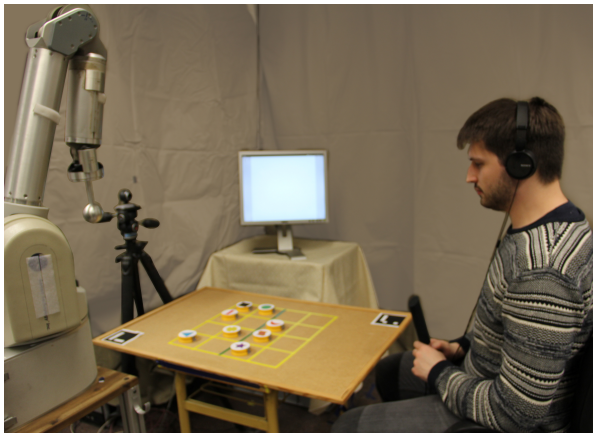


Fig. 1. Multimodal robotic system that supervises and assists its users in performing a memory exercise [3].

III. USER-CENTERED ROBOT TEACHING

Multiple algorithms have been developed for teaching robots new skills and for robots to interactively learn [6]–[9]. Amershi et al. [10] show that human teachers improve the efficiency of learning algorithms and also create systems that are more adapted to the needs of its users.

The robot should track the state of the environment and the user when it receives instructions. Moreover, it should create a policy based on those instructions. The robot should disclose its perception of the environment, provided instructions, and the learned policy. However, the manner of interaction influences the teacher, and the robot should be careful not to provide feedback that can misguide the teacher. Therefore, the robot must provide relevant information in the manner most suitable for the particular teacher. The robot should track and estimate which types of interaction provide optimal teaching experience that is defined as successful, fast and enjoyable for the human teacher. Teaching is successful if the robot learns the desired policy, and it is enjoyable if the human labels it as enjoyable in post-task assessments. Moreover, the robot can use emotion recognition algorithms to detect and track the enjoyment of the teacher during the teaching process.

Interaction can be unsuccessful because the teacher misunderstood feedback from the robot. This situation is detected

[1]Author is with the Institut de Robotica i Informatica Industrial, CSIC-UPC, Barcelona, Spain, ataranovic@iri.upc.edu

if robot's perspective about the task is unchanged after the repeated teacher's explanation. In this case, the robot should provide feedback in another manner, using different modalities. Moreover, the failure of the former feedback should influence modality choices in the future by labeling those modalities as less preferable.

The robot may misunderstand the instruction of the teacher. After additional explanations, the robot obtains the correct description of the tasks. Moreover, if possible, the robot should disclose how it came to the wrong conclusion. This can be accomplished by evaluating previous instructions, and if any of the instructions does not fit the new perception of the task, the robot should inform the teacher.

Robots interact using different modalities. Depending on the situation and the user, some are more appropriate than others [3]. If the robot detects strong ambient noise, it should not rely on verbal communication. Similarly, if the user has a hearing impairment the verbal interaction is not suitable. While robots can have diverse modalities, most common are speech, visual, and gestures.

Verbal communication is common among humans; however, the intricacies of languages are challenging for machines to understand, and hence act upon. Therefore, providing adequate feedback minimizes the ambiguity on both sides of the interaction. Several commercially available robots, including Pepper [1] and Baxter [2] , have a built-in display. Through this type of devices, the robot can provide rich visual information —e.g., a picture of a relevant object in a cognitive exercise. Gestures are another modality that can convey information, and they are especially suitable for providing spatial information because the robot can divert user attention towards an object of interest using pointing gestures. In some situations, a gesture augmented with a verbal command can provide a better quality of interaction than the unimodal interaction with only one of those modalities.

IV. TEACHING A COGNITIVE EXERCISE

In this paper, an example of a sequential memory exercise (Fig. 1) is used to illustrate the process of teaching the robot a cognitive exercise. The user is initially informed about a sequence of objects (e.g., the robot says object names), that need to be sorted in the same order. This exercise should stimulate the memory recall. In the example in Fig. 1 the objects are round tokens with different shapes printed on their top side. This robotic system can speak, perform gestures with its robotic arm, and show visual information on its display. A similar version of the aforementioned exercise requires that the user sorts the objects in the reversed order, by first selecting the last objects from the initial sequence, etc.

The teacher can explain the rules interactively. Depending on the exercise stage, the teacher can give some information, and ask the robot to perform a part of the exercise. Another way is performing one or more complete exercises correctly and then asking the robot to repeat. The robot generates a policy based on those demonstrations. While the robot is performing the exercise, the teacher gives feedback that guides the robot towards the correct policy.

If the robot is unsure as to which object the teacher is referring, it can point towards possible objects, or say a relative description. The choice of a modality the robot uses can depend on the modality that the teacher used. For example, if the ambiguity is because the robot was not capable of discerning towards which object the teacher was pointing, it can also use gestures. However, the robot could show the pictures of the possible objects, and ask the teacher to select the correct one. The current state of the environment and the users must be considered. For example, if the teacher is not looking towards the robotic arm, the robot may not want to influence the teacher to divert her or his gaze, hence the robot should avoid using this modality under the mentioned circumstances. Furthermore, the history of previous interactions should be a key factor when deciding which modality to use.

V. CONCLUSION AND FUTURE WORK

Modality adaptation is an important aspect of human-robot interaction. During robot teaching, the goal is to enable the human teachers to instruct the robot in a manner most natural to them while the robot provides them adequate feedback about its reasonings. This is important for having a high-quality learning process. In this paper, the adaptation of modality use when providing feedback is outlined.

In future work, user-centered robot teaching will be implemented and evaluated with user-studies with caregivers as teachers. The focus of the evaluation will be on the usefulness of the system and the quality of the interaction [11]. Furthermore, the teaching methods should also be evaluated on other use-cases that could improve the quality of life for older adults, in the line of the goals of the SOCRATES project.

ACKNOWLEDGEMENT

This work was supported by the SOCRATES project funded from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 and by the Spanish State Research Agency through the María de Maeztu Seal of Excellence to IRI (MDM-2016-0656).

[1] https://www.softbankrobotics.com/emea/en/robots/pepper

[2] https://www.rethinkrobotics.com/baxter/

REFERENCES

[1] D. Feil-Seifer and M. J. Matarić, "Defining socially assistive robotics," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, June 2005, pp. 465–468.

[2] F. Khan, B. Mutlu, and X. Zhu, "How do humans teach: On curriculum learning and teaching dimension," pp. 1449–1457, 2011.

[3] A. Taranović, A. Jevtić, and C. Torras, "Adaptive Modality Selection Algorithm in Robot-Assisted Cognitive Training," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, p. Forthcoming.

[4] B. Duthey, "Alzheimer Disease and other Dementias, Update on 2004," *World Health Organization*, no. February, pp. 1 – 77, 2013.

[5] A. Andriella, G. Alenyà, J. Hernández-Farigola, and C. Torras, "Deciding the different robot roles for patient cognitive training," *International Journal of Human Computer Studies*, vol. 117, no. February, pp. 20–29, 2018.

[6] J. E. Laird, K. Gluck, J. Anderson, K. D. Forbus, O. C. Jenkins, C. Lebiere, D. Salvucci, M. Scheutz, A. Thomaz, G. Trafton, R. E. Wray, S. Mohan, and J. R. Kirk, "Interactive task learning," *IEEE Intelligent Systems*, vol. 32, no. 4, pp. 6–21, 2017.

[7] M. Cakmak and A. L. Thomaz, "Designing robot learners that ask good questions," pp. 17–24, 2012.

[8] M. Cakmak, C. Chao, and A. L. Thomaz, "Designing interactions for robot active learners," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 108–118, June 2010.

[9] C. Chao, M. Cakmak, and A. L. Thomaz, "Towards grounding concepts for transfer in goal learning from demonstration," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2, Aug 2011, pp. 1–6.

[10] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the People: The Role of Humans in Interactive Machine Learning," *AI Magazine*, vol. 35, no. 4, p. 105, 2014.

[11] S. Bensch, A. Jevtić, and T. Hellström, "On interaction quality in human-robot interaction," in *Proceedings of the 9th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,*, INSTICC. SciTePress, 2017, pp. 182–189.

Are you playing with me? On the importance of Detecting and Recovering Disengagement in Mild Dementia Patients playing Brain-Training Exercises

Antonio Andriella[1]

***Abstract*— The ability to automatically detect disengagement in a human-robot interaction can improve the overall quality of interaction in term of acceptance and effectiveness. Several cases of study have been conducted on this topic but none of them seems to be exploring how such application could benefit older adults with mental impairments. In this positional paper, we evaluate the benefit of combining two of the four Observational Measurement of Engagement(OME) indicators and contextual information to detect disengagement in older adults affected from Mild Dementia and Alzheimer playing a brain-training exercise. Moreover, we empower a robotic system of a repertory of re-focus strategies in order to re-engage the patient once a disengagement is detected.**

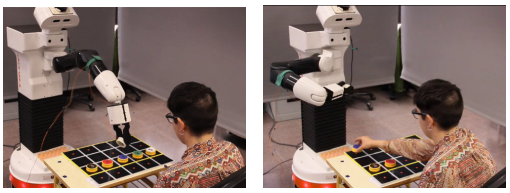
I. INTRODUCTION

In human-human interaction, engagement is defined as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" [1]. Engaging older persons with dementia in appropriate activities has been shown to yield beneficial effects such as increasing positive emotions, improving activities of daily living (ADL) and improving the quality of their life [2]. Cognitive training is based on a set of standard exercises designed to reflect particular cognitive functions; usually the therapist sets different range of difficulty levels within the standard set of tasks to suit the individuals level of capability. In this work, we propose the Syndrom KurzTest (SKT). The SKT is a short test for assessing cognitive impairment of memory and attention [3].

One perspective to explore engagement in HRI is to investigate the automatic prediction of engagement. The main idea is to predict disengagement behaviors in real-time, so the robot can provide recovering mechanisms to keep the user engaged and eventually re-engage him. To this end, several solutions have been proposed combining different features. Castellano *et al.* [4] focus their work on collecting task-related features and social interaction cues trying to address the issue related to robustness in real-world scenarios. Nakano *et al.* [5] propose an engagement estimation method that detects the users disengagement gaze patterns. Rich *et al.* [6] develop and implement a computational model for recognizing engagement between a human and a robot. Szafir *et al.* [7] design adaptive agents that monitor student's attention in real time using measurements from electroencephalography (EEG).

*This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 721619 for the SOCRATES project.

[1]A. Andriella is with Institut de Robòtica i Informàtica Industrial, CISC-UPC, C/Llorens i Artigas 4-6, 08028 Barcelona, Spain. aandriella@iri.upc.edu



(a) Robot provides help (b) User makes a move

Fig. 1: User is playing SKT with the robot's support.

The research on (dis)engagement detection has seen substantial advancements during the recent years. However there is still a considerable lack of experimental work focused on targeting persons affected by Mild Dementia and Alzheimer Disease. The study of how to automatically detect engagement is a necessary foundation for the development of non-pharmacological interventions for individuals with dementia, whether the interventions address depression or boredom. In the proposed scenario, we assume that the disengagement can be caused by the patient's lack of interest or negative attitude toward the task. The detection of disengagement of persons with dementia is expected to help such persons by increasing interest and his overall positive attitude.

In this positional paper, we attempt to fill the current gap by proposing a possible method aimed to potentially detect disengagement with older mentally impaired adults through a brain-training exercise. Our approach is based on the Observational Measurement of Engagement(OME) indicators, which were developed to specifically assess, within a certain subject, each level of engagement: attention, attitude, duration and refusal [8].

II. BRAIN-TRAINING EXERCISE SCENARIO

In this work we present a brain-training exercise based on a subset of the SKT. The goal of the test is to sort *n* tokens in ascending order on the board as quickly as possible and with the minimum amount of mistakes. An embodied robotic system, employed by a caregiver, is able to provide several levels of assistance on the base of the user performance and the state of the game combining different interaction modalities (speech and/or gesture). The assistance levels, as defined in [9] are:

- encouragement, in which the robot encourages the user to move a token;
- suggest subset, in which robot tells and points in the area of the solution;
- suggest solution, in which the robot tells and points the correct token;
- offer correct token, in which the robot gives the correct token to the user.

Figure 1 shows an example of interaction between the robot and the user.

III. SUGGESTED METHOD

A. OME Indicators and Contextual Information

We propose an extension to our previous work [10] adding a Disengagement Module which will be able to assess patient's disengagement based on OME indicators and contextual information. We decide to use only attention and attitude since we believe they are the most effective in this specific scenario. We define a stimulus as one of the assistance levels provided by the robot as defined in [9]. Each indicator is defined on a four-point scale: i) not attentive, ii) somewhat attentive, iii) attentive, and iv) very attentive. The specific outcome variables of the OME are defined as follows:

*a) **Attention**:* It is computed as the amount of time a participant looks at the robot and the board during the stimulus. The measurement starts as soon as the robot engages the user. To track the user's gaze we decide to use OpenFace [1]. We can define the percentage of time spent by the user focusing on the stimulus as:

$$attention = (T_b + T_r) * (100)/(Tot_{stim}) \quad (1)$$

where T_b is the time spent from the user on the board and T_r is the time spent by the user looking at the robot. Tot_{stim} is the total time of the stimulus (T_b+T_r<=Tot_{stim}). The outcome of this measure will be mapped on the four-point scale defined before.

*b) **Attitude**:* It is measured observing the user non-verbal expressions and it can be computed based on the concept of valence. To compute this value, we use Affectiva [2]. Here the valence metric likelihood is calculated based on a set of observed facial expressions [3]. We can define attitude as follows:

$$attitude = \arg\max_{i=1}^{4} \{perc\_attitude\_scale_i\} \quad (2)$$

where $perc\_attitude\_scale_i$ is the percentage of time the valence is on a defined point scale. At the end of an assistive action (stimulus) of the robot, one of the four point-scale is selected according to Eq. 2.

In this specific scenario is a primary aim of the robot to keep the user engaged in providing him with enough assistance in order to complete the game. Increasing the level of assistance could result in a loss of engagement by the patient since the task will be performed almost entirely by the robot. On the other hand, the selection of a lower level of interaction may result in insufficient assistance by the robot. This could mean the patient feeling frustration for not having achieved the goal or discouragement for having spent too much time to complete it.

We expect that the users engagement with the robot is both influenced by the task the user has to accomplish and the interaction with the robot. Moreover, we also expect that the different levels of assistance provided by the robot affect the different user's behaviors. For this reason, we include also the contextual information in the form of user performance as a parameter for deciding which action to perform in order to re-engage the user as soon as a disengagement is detected.

In particular we define the user performance in state s after an action of engagement e provided by the robot as:

$$user\_perf(e, s) = user\_move(e, s) * game\_diff(s) \quad (3)$$

where $game\_diff(s)$ is the current game difficulty in state s (computed based on the current state of the game and the user cognitive impairment) and $user\_move(e, s)$ is the outcome of the performed user move after an action of engagement e in state s. In other words, the harder is the game and the lesser is the level of assistance provided, the bigger will be the $user\_perf(e, s)$ value.

B. Disengagement Re-focus Stategies

The disengagement module for each state s computes a value defined as follows:

$$(dis)eng(s) = \alpha * attitude(s) + \beta * attention(s) \quad (4)$$

where α and β are weights for the two indicators. Those weights are important for analyzing the effect of each parameter separately and try to fine-tune the behavior of the developed module. Additionally, some studies point out how the effect related to ageing can affect the intentional display of facial emotions and the possibility of detecting unintended emotions [11]. So attitude, that is based on valence, can be ambiguous.

If a disengagement is detected, the robot based on the $user\_perf(e, s)$ value will evaluate which action to perform. To this end, it has been empowered with a repertoire of re-focus strategies in order to re-engage the user. The robot can:

- analyze the current user behavior and provide a more tailored support based on the current user performance (accuracy) and the time to perform the correct move (efficiency)
- re-engage the user providing the same assistance but with different modalities (using only speech or combining speech and gestures)
- alert the caregiver of the user's behavior asking him to intervene.

The module will provide at the end of the test session an accurate report of the user's total reaction time and number of mistakes.

[1] https://github.com/TadasBaltrusaitis/OpenFace
[2] https://www.affectiva.com/
[3] https://developer.affectiva.com/metrics/

IV. CONCLUSIONS

In this position paper, we evaluate the benefit to use a Disengagement Module in a brain-training exercise in order to detect people lack of attention and attitude toward the task. Combining user gaze direction, non-verbal features and contextual information, a robotic system will be able to evaluate the patient (dis)engagement and if it will deem it appropriate, it will try to re-engage it through a repertory of re-focus strategies.

As future work, we plan to validate the module in a real scenario. The main objective will be to evaluate the contribution of all OME indicators in the detection through questionnaires and videos analysis.

REFERENCES

[1] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005.

[2] K. K. Engelman, D. E. Altus, and R. M. Mathews, "Increasing engagement in daily activities by older adults with dementia," *Journal of Applied Behavior Analysis*, vol. 32, no. 1, pp. 107–110, 1999.

[3] J. E. Overall and R. Schaltenbrand, "The SKT neuropsychological test battery," *Journal of Geriatric Psychiatry Neurology*, vol. 5, no. 0891-9887, pp. 220–227, 1992.

[4] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. Mcowan, "Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features Interaction scenario," *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pp. 119–125, 2009.

[5] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Proceedings of the 15th International Conference on Intelligent User Interfaces - IUI '10*. New York, New York, USA: ACM Press, 2010, p. 139.

[6] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 375–382, 2010.

[7] D. Szafir and B. Mutlu, "Pay attention!: designing adaptive agents that monitor and improve user engagement," *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, pp. 11–20, 2012.

[8] J. Cohen-Mansfield, M. Dakheel-Ali, and M. S. Marx, "Engagement in persons with dementia: the concept and its measurement," vol. 6, no. 4, pp. 247–253, 2009.

[9] A. Andriella, G. Alenyà, J. Hernández-Farigola, and C. Torras, "Deciding the different robot roles for patient cognitive training," *International Journal of Human Computer Studies*, vol. 117, pp. 20–29, 3 2018.

[10] A. Andriella, G. Alenyà, and C. Torras, "Cognitive System Framework for Brain-Training Exercise based on Human-Robot Interaction," *Submitted for publication*, 2018.

[11] N. C. Ebner, Y. He, and M. K. Johnson, "Age and Emotion Affect How We Look at a Face: Visual Scan Patterns Differ for Own-Age versus Other-Age Emotional Faces," vol. 25, no. 6, pp. 983–997, 2012.

Increasing the Understanding between a Dining Table Robot Assistant and the User*

Samuel Olatunji, Tal Oron-Gilad, Yael Edan

Abstract— **This study is a preparatory stage of a larger study intended to increase the understanding between a dining table robot assistant and the user. The users are expected to be older adults who need assistance in their daily lives but the study begins with investigating the level of understanding with younger adults with the intention of comparing the interaction with older adults in further studies. The aim of the experiment is to identify the most appropriate mode of communication from the robot which will convey the state of the interaction between the user and the non-humanoid robot. The results of the present study reveal that voice feedback from the robot aids better understanding of the state of interaction compared to visual feedback in the absence of background noise while the visual feedback aids better understanding in the presence of noise. Even though most of the users had an opaque understanding of the interaction with the robot while using the voice feedback mode, the results point to the possibility of obtaining better understanding if both feedback modes are combined, to highlight the advantage of each modality, and the content of the information provided is improved. The study is the initial step towards a design framework for improving the understanding between a socially assistive robot (such as a table setting robot) and the user.**

I. Introduction

Socially assistive robots (SARs) are a possible solution to bridge the elder care gap [1], which is defined as the dearth of caregivers and healthcare professionals available to cater for older adults [2]. SARs can assist older adults in some activities of daily living such as meal setting [3]–[5]. This constitutes a form of human-robot interaction (HRI) where older adults are expected to interact with a robot serving as a dining table robot assistant. One of the challenges involved in this interaction which this study intends to address, is the mismatch commonly observed in the user's understanding of the state of the robot relative to the robot's actual state. This mismatch could lead to misuse – if the user over-relies on the robot, or disuse – if the user under-utilizes the robot [6]. In the sensitive setting of elder care, such consequences can significantly degrade the quality of user-robot interaction. The research addresses the following question: which information presentation mode from a non-humanoid table setting robot effectively communicates the state of the interaction to the user?

II. Literature Review

Optimal robot performance and user experience during human robot interaction (HRI) are important aspects that define quality of interaction [7]. Understanding the robot's state is a crucial link in the metrics of assessments which needs to be taken into consideration [8]. Understanding in the context of HRI can be described as the extent to which a human and a robot have adequate knowledge about each other's state to be able to successfully interact with each other [9]. Communicative actions could be sent from the user to the robot or vice versa in form of instructions or feedback [10]. These communicative actions when presented in the most comprehensible form promotes understanding which leads to a successful interaction of the user with the robot [8], [11]. It is a form of bidirectional presentation of information where the instructions could originate from the user or robot, encoded in a specific mode or multimode (such as visual, audial or gestural) and decoded through various mode recognition or perception techniques (such as GUI, speech or gesture recognition mechanisms) [12]. This bidirectional communication keeps both parties aware of the factors underlying each other's actions and allows them to correct erroneous factors that each may have [13]. Successful bidirectional communication between the robot and the human supports transparency of the interaction, team performance and trust in the automation [13]. The extent to which the human understands the robot's communicative action can be referred to as states of understanding as used by Clark and Schaefer [14] and further elaborated by Doran et al. [11] as presented in Table I.

TABLE I. States of Understanding

States of Understanding	Description
Opaque	Recipients perceive the inputs and outputs of a system without knowledge of how the input is mapped to the output.
Interpretable	Recipients perceive not just the inputs and outputs of a system but can also observe all the details that produced the output from the input. Understanding the details that map the input to the output usually requires the user to have background knowledge of the data and domain.
Comprehensible	Recipients perceive the inputs and outputs of a system and can also comprehend the meaning and relationship between the input and output. Symbols and words are often encoded in the system with a knowledge base that can help the user relate the input to the output.

Several studies have explored different modalities through which a robot may express its state to the user. These modalities include buzzers, light projections, motion [15], gestures, facial expressions, body language [16], speech [17] and augmented reality [18]. The choice of modality to use is strongly predicated on the several factors which particularly includes the type and capability of the robot [15], context of use and noise conditions [7], [11]. Noise has been observed in

* This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

All authors are from the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev Beersheba 84105, Israel (phone: +972-86461434; email: olatunji@post.bgu.ac.il, orontal@bgu.ac.il, yael@bgu.ac.il).

existing studies to corrupt the accurate interpretation and comprehension of information communicated to recipients [9]. This study hypothesizes that there will be an interaction between the mode of feedback and background noise. A second hypothesis is that visual feedback will influence a higher level of understanding in the presence of background noise, while voice feedback will do so in quiet environments. Extensive user studies are required to explicitly identify the most appropriate mode of communication that will promote understanding in the case of socially assistive robots that have no semblance of human morphological features such as the meal setting robotic arm used in this study.

III. METHODS

A. Overview

There are four groups in the study. The groups consisted of different combinations of feedback modes and noise. The feedback was provided by the robot to give the user information on the status of the interaction while the noise was simulated to depict typical noisy settings. Participants were asked to give voice commands to the robot to perform a pick and place task similar to what would be required in setting utensils and food items on a dining table. Objective and subjective measures were taken to assess the understanding the users had regarding the state of the interaction based on the feedback given by the robot. The overall experience with the robot was also assessed. The study took place at the intelligent robotics laboratory, Ben-Gurion University of the Negev, Israel.

B. Apparatus

The dining table robot assistant used was a robotic arm – KUKA iiwa (Intelligent Industrial Work Assistant) LBR (Lightweight Robot) with seven DOF (Degrees of Freedom). The KUKA enables fast development and integration of devices, using Robot Operating System (ROS) [19]. It is a lightweight robot for industrial applications that is designed for safe close cooperation between human and robot on highly sensitive tasks [20].

C. Participants

A convenience sample of sixteen people participated in the experiment (6 Females, 10 Males) aged 21-57 (mean 29.2 years). The intention is to experiment first with younger people who are more readily accessible and then proceed to use the lessons learned for the experiment involving older adults. There were 8 participants with Engineering background while the other 8 were from other disciplines. Each participant completed the study separately at different timeslots, so there was no contact between participants.

D. Experimental Design

The experiment was set as a between-participant factorial design with manipulations of feedback and noise conditions as independent variables. The feedback modes used were voice and visual feedback modes while the noise manipulation was a condition with the presence of an alarm noise in the background and without it, as illustrated in Table II. Participants were assigned randomly to one of the four groups. Each participant had either voice or visual feedback in the presence or absence of noise based on the group assigned. The visual feedback was in the form of a display on a screen situated near the robot displaying 'Good Work' on a green background when the robot sensed the voice command given by the participant and was moving as commanded. The display showed 'Not Done' on a red background when the robot was yet to carry out the commanded task or could not carry out the commanded task. The feedback information stayed on the screen till the next command was issued and the next feedback information related to the new command was displayed. The voice feedback gave the same information but in the form of a simulated human voice which was given repeatedly at specific intervals till the next command was issued. The noise effect was implemented in the form a repetitive rhythmic alarm sound in the background at approximately 55dB. The alarm was switched off in the groups without noise, and the sound level in the lab was maintained at approximately 35dB. The sound level of the voice feedback was at approximately 60dB such that the participants could hear the voice feedback well above the alarm noise.

TABLE II. EXPERIMENTAL GROUPS

		Alarm Noise	
		Present	***Absent***
Feedback	**Voice Feedback**	Group A	Group B
	Visual Feedback	Group C	Group D

E. Experimental task

Participants were assigned a task which consisted of two trials: The first trial was to give voice commands to lead the robot to pick and place pre-arranged fruits into a bowl while the second trial was to give voice commands to the robot to pick cups and arrange them in a predefined configuration (*Fig. 1*). The trials were counterbalanced between participants. It was designed using a Wizard-of-Oz technique where the users' commands were translated to the robot's motion in real time via the keypad of the robot by an experimenter.

F. Procedure

At the start of the experiment, the participants were asked to fill a consent form which described the experiment and what the participant was required to do. The participants were then asked to complete a pre-test questionnaire which included some demographic information, a Technology Adoption Propensity (TAP) index [21] and a Negative Attitude toward Robots Scale (NARS) [22]. The robot was then introduced to the participants as their table setting robot assistant who could carry out their commands to set items on the dining table. An instruction set of 8 commands was given to the participants to control the robot as described in Table III. Participants were asked to command the robot to accomplish the two trials described in the experimental design. Post-trial questionnaires were administered after each trial and a final questionnaire at the end of the experiment to assess the subjective experience with the robot assistant.

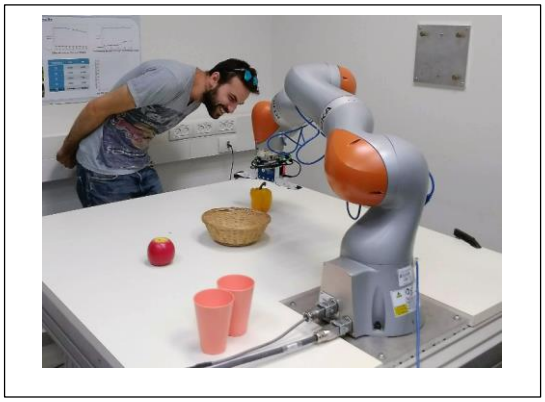


Fig. 1: Experimental setup using the KUKA robot

TABLE III. SET OF COMMANDS TO CONTROL THE ROBOT

Command	*Action of the robot*
Left	Moves towards the negative x axis
Right	Moves towards the positive x axis
Forward	Moves towards the positive y axis
Backward	Moves towards the negative y axis
Up	Moves towards the positive z axis
Down	Moves towards the negative z axis
Open	Opens the gripper
Close	Closes the gripper

IV. RESULTS AND DISCUSSION

The results obtained from the objective and subjective measures are presented in the following subsections.

A. *Demographics*

There was an equal distribution of participants within the 4 groups. The participants were mostly acquainted with the use of innovative technologies. On a scale of 1 (strongly disagree) to 5 (strongly agree), the TAP index reveals that most of the participants are optimistic about technology providing more control and flexibility in life (mean = 4.09, SD = 0.86). The NARS reveals that the participants do not have negative feelings about situations in which they interacted with a robot (mean = 2.14, SD = 1.2).

B. *Objective Measures*

The objective measures were the average time it took participants to complete the task (in seconds) and the average number of errors made by the robot while being commanded to pick and place the items. The independent variables for the experiment were the manipulations of the feedback mode and presence of noise. The dependent variable is the level of understanding the user has regarding the state of the interaction.

The average time it took participants to complete the task (consisting of both trials) in the experiment was 369 seconds (SD = 82 seconds). In the presence of the background noise, participants with visual feedback (group C) spent the shortest time on the tasks (mean = 327 seconds, SD = 16.84 seconds). In the absence of the background noise, participants with voice feedback (group A) spent shorter time on the tasks than participants with visual feedback (mean= 364 seconds, SD = 30.01 seconds). This is presented in Fig. 2. It is assumed that the longer it took the participants to complete the tasks, the less understanding they had regarding the interaction based on the feedback given by the robot.

Participants with the voice feedback in the absence of noise (Group B) experienced the least number of errors (mean =1, SD = 0.82) while participants with visual feedback in the presence of noise encountered the highest number of errors (mean = 2, SD = 1.63). The error values are indicated in Fig. 2 (in purple). It is assumed that less errors indicated to some extent that the participants had a good understanding of the interaction based on the feedback provided by the robot.

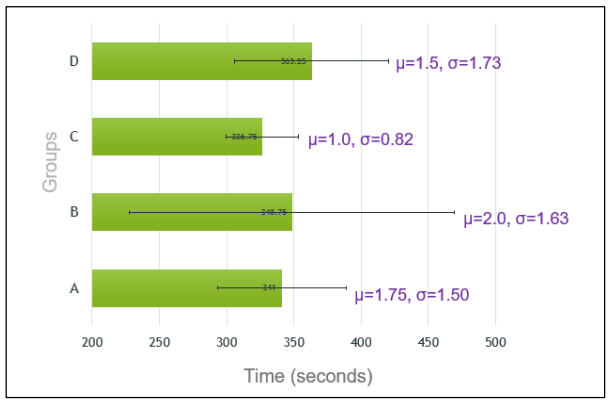


Fig. 2: Average time it took participants in each group to complete a task (bars represent SE), purple number represents average number of errors per task and SD.

B. *Subjective Measures*

The experience of interacting with the table setting robot is presented in *Fig. 3*. Only 2 (13%) of the participants considered the robot as understandable. These were in the groups with voice feedback. The subjective rating of the level of understanding the users in each of the groups have regarding the state of the interaction is presented in Fig. 4. The groups with voice feedback had more participants who understood the robot's feedback at an opaque level. There is a high possibility that their level of understanding was affected by the presence of noise since there were some participants with voice feedback in the absence of noise whose subjective ratings indicate a comprehensive understanding of the information the robot was communicating.

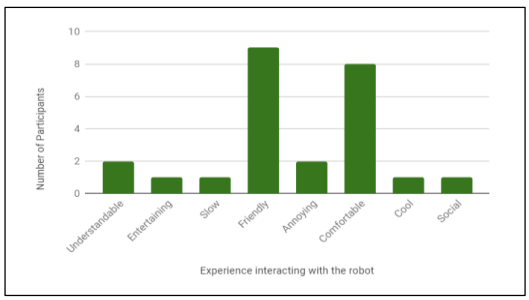


Fig. 3: Users' perception of the robot

Both the objective and subjective results reveal that visual feedback from the robot aided a better understanding of the state of interaction compared to voice feedback in the presence of background noise whereas participants experienced better understanding with the voice feedback when the noise was absent. Both feedback modes can therefore be combined to create an improved communication mode rather than utilizing voice feedback as the only communication mode.

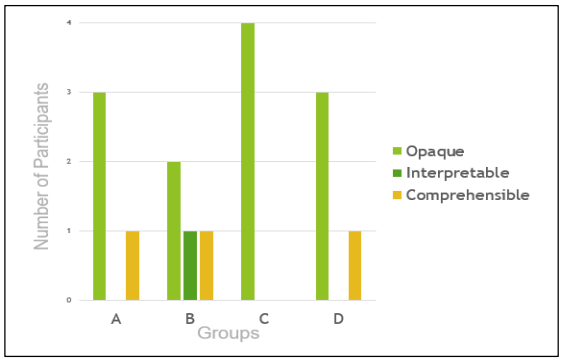


Fig. 4: Number of ratings of level of understanding for each group

The subjective experience of the participants revealed that 75% of the participants have just an opaque understanding of the interaction, despite their positive TAP scores and irrespective of the feedback mode being used. This therefore brings to the fore, the possibility that the content of information being displayed or spoken in words may have been insufficient to convey a comprehensive level of understanding of the information being presented by the robot. Three levels of information content could be displayed or voiced out which are connected with presenting what the robot is doing, the reason for the action(s) and consequence(s) of the action(s) [23]. In this study, only the state of the interaction (level 1) was displayed. Future work to improve the understanding will entail varying the content of the feedback to include the reason for the robot's actions (level 2) and the consequences of such actions (level 3). These studies will also be conducted with more participants to provide sufficient data for standard statistical significance tests.

V. CONCLUSION

The study revealed that the voice feedback mode used in the interaction between a table setting robot assistant and the user aided better understanding of the interaction state compared to the visual feedback in the absence of background noise. Visual feedback provided better understanding than voice feedback when noise is present. This gives insight for the next stage of the research which would include testing the combination of both feedback modality modes and varying the content of the information being provided to further improve the user's level of understanding.

REFERENCES

[1] D. Tang, B. Yusuf, J. Botzheim, N. Kubota, and C. S. Chan, "A novel multimodal communication framework using robot partner for aging population," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4540–4555, 2015.

[2] N. Super, "Who Will Be There to Care? The Growing Gap between Caregiver Supply and Demand," *Natl. Heal. Policy Forum*, 2002.

[3] D. McColl, W. Y. G. Louie, and G. Nejat, "Brian 2.1: A Socially assistive robot for the elderly and cognitively impaired," *IEEE Robot. Autom. Mag.*, vol. 20, no. 1, pp. 74–83, 2013.

[4] C. A. Smarr *et al.*, "Domestic Robots for Older Adults: Attitudes, Preferences, and Potential," *Int. J. Soc. Robot.*, vol. 6, no. 2, 2014.

[5] C. A. Smarr, A. Prakash, J. M. Beer, T. L. Mitzner, C. C. Kemp, and W. A. Rogers, "Older adults' preferences for and acceptance of robot assistance for everyday living tasks," *Proceedings of the Human Factors and Ergonomics Society*. pp. 153–157, 2012.

[6] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 39, no. 2, pp. 230–253, 1997.

[7] S. Bensch, A. Jevtić, and T. Hellström, "On Interaction Quality in Human-Robot Interaction," *Proc. 9th Int. Conf. Agents Artif. Intell. - Vol. 1 ICAART*, vol. 1, pp. 182–189, 2017.

[8] N. Balfe, S. Sharples, and J. R. Wilson, "Understanding Is Key: An Analysis of Factors Pertaining to Trust in a Real-World Automation System," *Hum. Factors*, no. 1983, 2018.

[9] T. Hellström and S. Bensch, "Understandable Robots -," *Paladyn, J. Behav. Robot*, 2018.

[10] N. Mirnig, S. Riegler, A. Weiss, and M. Tscheligi, "A case study on the effect of feedback on itinerary requests in human-robot interaction," *Ro-Man, 2011 Ieee*, pp. 343–349, 2011.

[11] D. Doran, S. Schulz, and T. R. Besold, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives," 2017.

[12] N. Mirnig, A. Weiss, and M. Tscheligi, "A communication structure for human-robot itinerary requests," *Human-Robot Interact. (HRI), 2011 6th ACM/IEEE Int. Conf.*, pp. 205–206, 2011.

[13] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, "Situation awareness-based agent transparency and human-autonomy teaming effectiveness," *Theor. Issues Ergon. Sci.*, vol. 19, no. 3, pp. 259–282, 2018.

[14] H. H. Clark and E. F. Schaefer, "Collaborating on contributions to conversations," *Lang. Cogn. Process.*, vol. 2, no. 1, pp. 19–41, Jan. 1987.

[15] K. Kobayashi and S. Yamada, "Making a Mobile Robot to Express its Mind by Motion Overlap," *Adv. Human-Robot Interact.*, no. December 2009, pp. 342–356, 2009.

[16] H. Knight and R. Simmons, "Layering Laban Effort Features on Robot Task Motions," *Proc. Tenth Annu. ACM/IEEE Int. Conf. Human-Robot Interact. Ext. Abstr. - HRI'15 Ext. Abstr.*, pp. 135–136, 2015.

[17] N. Mavridis, "A review of verbal and non-verbal human-robot interactive communication," *Rob. Auton. Syst.*, vol. 63, no. P1, pp. 22–35, 2015.

[18] J. Carff, M. Johnson, E. M. El-Sheikh, and J. E. Pratt, "Human-robot team navigation in visually complex environments," *2009 IEEE/RSJ Int. Conf. Intell. Robot. Syst. IROS 2009*, pp. 3043–3050, 2009.

[19] S. Mokaram *et al.*, "A ROS-integrated API for the KUKA LBR iiwa collaborative robot," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 15859–15864, Jul. 2017.

[20] M. A. K. Bahrin, M. F. Othman, N. H. N. Azli, and M. F. Talib, "Industry 4.0: A review on industrial automation and robotic," *J. Teknol.*, vol. 78, no. 6–13, pp. 137–143, 2016.

[21] M. Ratchford and M. Barnhart, "Development and validation of the technology adoption propensity (TAP) index," *J. Bus. Res.*, vol. 65, no. 8, pp. 1209–1215, 2012.

[22] D. S. Syrdal, K. Dautenhahn, K. Koay, and M. L. Walters, "The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study," *23rd Conv. Soc. Study Artif. Intell. Simul. Behav. AISB*, pp. 109–115, 2009.

[23] J. Y. C. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. J. Barnes, "Situation Awareness – Based Agent Transparency," no. April, pp. 1–29, 2014.

Service Assistant to Support the Elderly with Mobility Issues

Truong Giang Vo[1] Simone Kilian[2]

***Abstract*—The Care-O-bot 4 was developed at Fraunhofer IPA as a general purpose mobile service robot. In a continuous effort of exploiting its modularity and potential, we decided to develop new robots with Care-O-bot 4 as a basis. This paper details the process of developing the possible design of such a robot. In this particular case, the robot is envisoned as a mobile companion which is able to help the elderly with their mobility in domestic environment.**

***Index Terms*—elderly care, mobile robot, service robot, hardware design, concept design**

I. Introduction

In the last few decades, according to the nursing homes and elderly-care facilities, the average age of newly admitted residents has been increasing. Presently, the average age for new residents has reached 85 years. It is a known fact that the elderly would like to stay at their own homes as long as possible. However, living at home without constant care and observation might carry a certain risk of unattended emergency, especially for those who live alone. One of the common risks for the elderly is falling due to their frailness and difficulty in walking. For such a reason, we decided to develop a mobile service robot which can assist the elderly mobility at home, thus elevates the risk of falling. The robot should also be able to observe the elderly in domestic settings to alert responsible people in case of an emergency.

In recent years, there have been several researches, concepts as well as products focused on the mobility-assisted application. They were developed to cater different age groups in different settings. For example, Ottobock's Xeno is an reconfigurable electrical wheelchair for immobile patients [1], iBOT [2] is another example of electrical wheelchair, which claims to be able to move up and down staircases. Beside wheelchairs, there are also electrical rollators available in the market, such as Triumph Mobility's Rollz Motion[2] [3] and eMovements' ello [4]. From the traditional wheelchairs and rollators, these products evolved and became smarter, more complex and now have more functionality in general. However, they mostly cater for outdoor activities, and do not have observing function.

On the other hand, there are also researches which focused on the monitoring aspect of elderly care. For instance, Mobina [5] is a mobile communication platform which can detect emergency situation and contact the responsible personnel.

[1]Truong Giang Vo, Robot & Assistive Systems, Fraunhofer IPA, Germany, truong.giang.vo@ipa.fraunhofer.de

[2]Simone Kilian, University of Design Schwäbisch Gmünd, Germany, simone.n.kilian@outlook.de

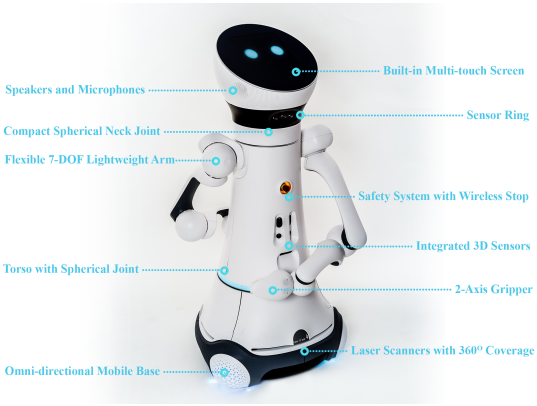


Fig. 1. Care-O-bot 4 developed by Fraunhofer IPA

Accompany project [6] is another example of using mobile service robot as a companion in domestic environment, where the robot, Care-O-bot 3 [7] in this particular case, can assume not only assistive but also preventive functions.

Contrary to the aforementioned products and researches, which either monitor or support the users, we envision the scenario where our robot could constantly monitor the well-being of the users, and assist their mobility only when they need it. In other words, we emphasize on letting the elderly to keep walking by themselves as much as possible, and the robot would always be by their sides and assist if necessary. Furthermore, we also place more importance into the observing capability of the robot, such that it can immediately response if incidents happen, especially in domestic settings, when the users are alone.

At Fraunhofer IPA, we had developed Care-O-bot 4 [8] (Fig. 1), a modular service mobile robot for general purpose. The base of the robot is a modular omni-directional mobile platform, which has 3 laser-scanners with 360° coverage. It also has its own computer with navigation and collision avoidance software modules integrated. As a continuous effort of exploiting the modularity of Care-O-bot 4, as well as expanding its applications in different scenarios, we intended to make use of its mobile platform as a basis, and adapt other parts to meet the requirements of this particular application of mobility assistance and monitoring.

II. Scenario

We created a persona, Agnes, to illustrate the application of our robot. Agnes is an 82 year old lady who is living independently in a small flat. She is mobile, but frail with a

risk of falling so she has been given a walking frame; however, she rarely uses it inside the apartment. Lately, some critical situations have occurred where Agnes lost her balance and almost fell. Additional assistance to support her at home was obviously required.

Agnes decided to rent a robot companion to support her 24 hours a day. As the robot is with her all the time, it is able to work with her at her own pace. To assist her mobility, the robot moves with her around her flat, and she is able to steady herself by holding on to it, sit on it if necessary or put objects onto the robot to get them transported. The robot is also able to detect emergency situations and trigger an alert either for her daughter or an emergency center (e.g. if she does not emerge from the bathroom within a set period of time or she has fallen).

III. User Study

In order to verify the scenario, and identify the actual needs and opinions of the elderly on the robot, we conducted a pre-development user study at one of the assisted-living facility in Göppingen, Germany. We interviewed 7 residents (6 female and 1 male) in the facility:

- The ages of the residents vary, from 70 to 98 years old.
- They live in their own apartment (25-60 sqm), which are similar to their home environments. The care-staff only come once in a while if needed.
- All of them are using either walking cane or rollator indoor.

According to the result of the interviews, we summarized several important key points regarding the functionality of our robot:

- The elderly have little experience with smart-phone and touchscreen in general. They would prefer to control the robot either by voice or by hand gestures.
- They do not want an autonomous wheelchair-like vehicle but still want some robust assistance while walking.
- They do not mind having a camera constantly monitoring their activities. It must be noted that we had explained to them that the camera leaves no personal footprint, as it only track the skeleton model of the users.

Additionally, we also interviewed 7 care-staff at the facility. According to their feedback, we decided that the robot should also have a medication schedule reminder to reduce their workload.

IV. Conceptual Design

A. Functionality

We decided to narrow down the target age group for this application. The robot would focus on helping the elderly, who are frail and prone to falling, but still maintain some degree of mobility by themselves. And in order to meet the requirements from the users, the robot should have the following functions:

1) A mobile platform so that the robot can follow the user inside the apartment. As mentioned earlier, the base of Care-O-bot 4 will be used, since it is modular and integrated with necessary hardware and software for autonomous navigation and collision avoidance.
2) The upper part of the robot must have handle to support partially the weight of the users when they lean on it.
3) The robot should have a seat for the users in case of need.
4) The robot must have a tray for object transportation within the apartment.
5) The robot must have one (or more) camera to monitor users' activities in case of emergency.
6) The robot should have voice/gesture recognition system, or other modes of communication (e.g. remote controller) for human-robot interaction.

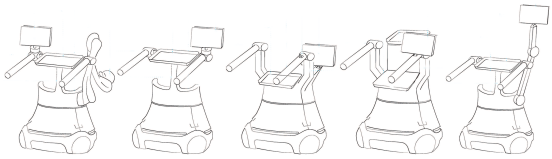


Fig. 2. Early concepts of the robot

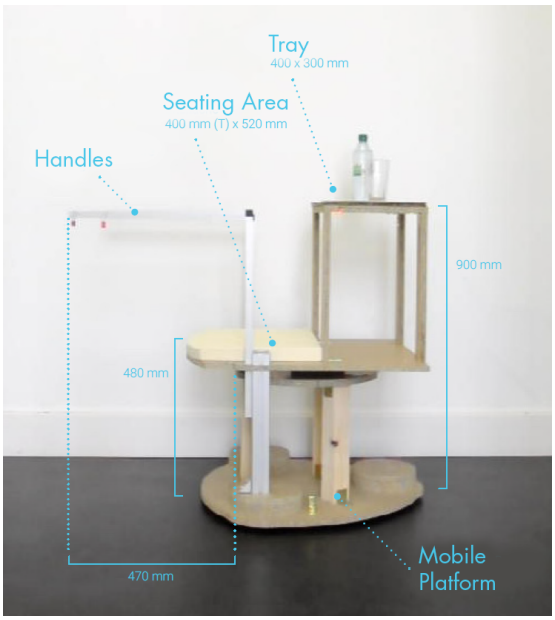


Fig. 3. Mock-up model of the design for ergonomic study

B. Ergonomic Study

In order to determine the precise size and shape of the robot, we built a mock-up model of the robot with adjustable seat and handles, as shown in Fig. 3, then conducted an ergonomic study with 5 participants (4 female and 1 male). Based on the feedbacks from the participants, we have several conclusions:

- Rollator-like handles are not suitable as they need to be long to ensure a comfortable distance between the user and the robot. Furthermore, long handles leads to large footprint and stability issues for the robot.
- On the other hand, the handles should be on all sides of the robot so that the users can access them regardless of

Fig. 4. Illustration of the final design

orientations. The height of the handles should also be adjustable to accommodate the users varying in sizes. However, the adjustment only need to be done once manually at the beginning of the service.
- The precise height of the seat is not an important factor, as users only use it for a short period of time while resting.
- The tray on the robot should be large enough to standard object such as a dinner plate (260mm in diameter).

We also conducted a static study to determine the stability of the robot. In particular, it is important for the handle to withstand partially the weight of the users without destabilizing the base. With the assumption that the average weight of a user is approximately 80kg, we modified the handles in such a way that it covers the entire robot, as shown in Fig. 4 . By doing so, we could limit the situations where a user puts his entire body weight on the robot. Overall, the form factor of the robot is as follows:

- Seating area: 480mm of height with the size of 350 x 550mm.
- Handle: adjustable height from 650 to 800mm.
- Tray: fixed height of 900mm.

C. Final Design

Fig. 4 shows the illustration of our robot. The upper part of the robot was kept simple while still maintain high degree of affordance. By doing so, the users would be able to operate the robot intuitively without the need of complicated instructions.

The cameras, with pan & tilt mechanism, were kept within the housing of the robot to monitor the users' activities. While doing so would reduce the field of view of the cameras, it is the trade-off we made to lessen the uneasiness of the users having the cameras visibly tracking them constantly.

There is a push-to-open drawer at the back of the robot. When it is time for the users to take medication, the robot will remind them via speakers. The users can then open the drawer and take the medication by themselves. There is also a tablet which can instruct the user which medication to take according to her subscription.

Finally, Fig. 6 shows some of the different ways the robot could be used in everyday activities.

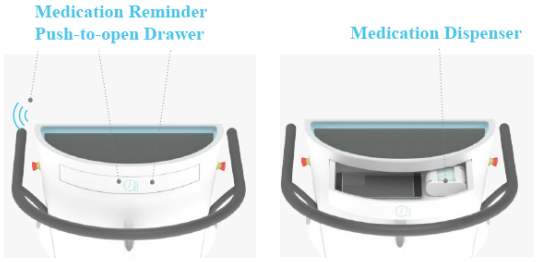


Fig. 5. Push-to-open drawer for tablet and medication

Fig. 6. Different usages of the robot

V. Conclusion & Future Work

In this paper, we described the process of developing a robot companion which can help the elderly with their mobility issues. It was developed with Care-O-bot 4 design as a basis. Currently, the robot is still a conceptual design. We are working on an operational prototype and test it in local elderly-care facilities, while keeping the commercialization potential in mind.

Acknowledgment

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 721619 for the SOCRATES project.

References

[1] Ottobock, Xeno, [Online]. Available: https://www.ottobock.com.hk/en/mobility-solutions/complex/power-wheelchairs/xeno/
[2] T. Nishiyama, F. Takiuchi, K. Ando, M. Arisawa, "The Functional Evaluation of Future Wheelchairs Contributing to Ecological Aid in Traveling", IEEE International Symposium on Environmentally Conscious Design and Inverse Manufacturing, 2005.
[3] Triumph Mobility, Rollz motion^2, [Online]. Available: http://triumphmobility.com/products/rollators/rollz-motion-2/
[4] eMovements, ello, [Online]. Available: https://www.ello-info.de/#ello
[5] Fraunhofer IPA, Mobina, [Online]. Available: https://www.aal.fraunhofer.de/de/projekte/mobina.html
[6] CORDIS, Acceptable robotiCs COMPanions for AgeiNg Years (ACCOMPANY) Project, [Online]. Available: https://cordis.europa.eu/project/rcn/100743_en.html
[7] Fraunhofer IPA, Care-O-bot 3, [Online]. Available: https://www.care-o-bot.de/de/care-o-bot-3.html
[8] Fraunhofer IPA, Care-O-bot 4, [Online]. Available: https://www.care-o-bot.de/de/care-o-bot-4.html

Analyzing Explicit(Speech) Modalities from Human-Human Interactions for building Context about a Robot-Assisted Dressing Task

Antonella Camilleri[1]

Abstract— **Robots that assist in the Activities of Daily Living (ADL), such as dressing, can support an aging population and lack of caregivers. These interactions are treacherous due to the implication of having end users like older adults who require great care and control on the overall interaction between the human and robot. The goal of collaborative interactions, such as ADLs, incorporates the success of the ADL task while taking into consideration the direct or indirect effect of the surrounding environment on the task and the user itself. Therefore a need to measure distractions or lack of collaboration between user and robot is vital. Collaboration in tasks like these evolve around the task itself and any measure of information from the interaction used to acknowledge progress needs to be carefully evaluated. Progress in task and state of interaction is context; and not being able to identify this, can be a sound indicator of distractions or lack of motivation to collaborate. Looking at human-human interactions for an assisted dressing task, speech utterances modality, together with other modalities, are used as a method of classification of the progress in the assisted task; by using LSTMs. A better classification of sequence state of task, due to speech utterances, would indicate that this explicit modality can be important to measure the level of collaboration and progress in such task.**

I. INTRODUCTION

Human-Robot Interaction (HRI) revolves around the robot's capability in practical tasks to allow an effective and successful human-robot collaboration. Assisting humans with daily tasks requires a robotic system which is capable of assessing the current state (*context*) of all entities in the interaction environment [1]. This current state needs to be based on the latest past, present or/and abrupt forthcoming interactions. Modeling or simulating interactions with humans is challenging. The complex and dynamic nature of the different states of interaction requires knowledge on states such as spatial, temporal and resources of the robotic system, the states of the human and objects interacting with the robot. All of these states form part of the environment surrounding the interaction which often holds information that humans use to infer about the required action. This ability enables the execution of tasks in a very practical and collaborative way. Furthermore, the collaboration between humans comes with the ability to distinguish between shared common grounds, ones own knowledge and collaborators' knowledge [2]. This knowledge shapes the required action making the interaction realistic and to a safer extent because the action is based on current states and not only on past modeled states.

Possibly, the obvious modality used to infer information about a collaborative task between humans is the symbolic representation of the explicit modality of speech utterances. Nonetheless, once the knowledge of how to carry out a task is known, speech utterances are observed to be limited in interactions between humans [3]. The hypothesis is, that due to the lack of speech utterance, such explicit modality is only related to extreme interaction states such as acknowledgment of progress in task or to correct actions. If this hypothesis is proved to be true than the lack or erroneous speech utterances can indicate different unknown states (*situations* that arise provided a specific *context*) in the interactions. These errors can be deduced as distractions from the collaborative tasks. Hence the objective of this paper is to examine if speech utterance in time are directly linked to affirming progress in the sequence of the task, meaning sequence prediction of the task is improved. The collaborative task examined is the assistive task of dressing an outer layer of jacket between humans.

[1]Antonella Camilleri is a PhD student at the University of the West of England, Bristol Robotics Laboratory, Coldharbour Lane, BS16 1QY, Bristol, United Kingdom. antonella.camilleri@uwe.ac.uk

II. MOTIVATION

A robot that provides support with dressing has to detect when the user is distracted and what the current states are in order to complete the task safely. These distractions can be instilled from noise or commotion in the environment or simply by a lack of attention from the user. Hence, knowing what lacks in explicit modalities when a user is distracted or not can allow the robot to adapt and perform the right action. The assisted task of dressing is complicated and depending on which part/sequence of task the robot is in, the current state and selection of actions can vary along progress of task. Consequently, establishing if an explicit speech utterance in time is related to the sequence of task is or not is imperative for a safe HRI.

III. EARLIER WORK

In the past, belief models for situation awareness have been implemented using Markov Logic provided that model spatio-temporal frames include epistemic information. Analyzing HRI requires methods that can handle multivariate time series inputs. One machine learning method used to train such data is the variation of recurrent networks called Long Short-Term Memory Units (LSTMs). This has been used to extract contextual features from multi-modal inputs in order to classify emotion or sentiment or action selection. Furthermore, in [4] skeleton data trained in a three-layer LSTM has been implemented to infer users interactive intent. Prediction of sequential tasks can be seen in [5] in which goal location of reaching motion is implemented and combined with LSTM to predict the next steps in sequence. The benefit

of using LSTM is the ability of creating long-term dependencies first by extracting features from each modality and then by looking at the relationship between the modalities. This method particularly holds the state of the neurons which is ideal for predicting sequences.

IV. PROPSED METHOD

The proposed method of implementation is LSTM. LSTMs are a variation of RNN with the ability to perceive previously in time and classify sequential input [6]. However, LSTMs are better because of the no vanishing gradient problem, but mostly because they have a forget gate with a purpose of linking distant occurrences to a final output [7] . Furthermore, LSTM preserve the error and can be back propagated through time and layers allowing recurrent nets to continue to learn over many time steps. This opens a channel to associate causes and effects remotely. This property in LSTM addresses the challenges of having delayed reward signal in realist environment interactions. Also, having a stacked LSTM architecture allows the hidden state of each level to operate at different timescale which is very likely to happen in this kind of interaction. This implies that user distractions or successfully completion of task can be predicting by observing the state of the memory cell representing this data input.

A. Experiments Procedure

The dataset used in this work was gathered during a dressing task [3] where 12 users were given assistance from another human posing as a robot. The users were wearing a motion tracking suit (Xsens) to record the spatio-temporal, position and orientation, of 23 points on the body. In the task, the users had to collaborate to put on a jacket several times. Each user put the garment on three times. Each dressing task took approximately 40s. For eye gaze tracking, the users were wearing a Tobi Pro Glasses which recorded eye gaze during the task. Video recordings were used to extract speech utterance in time. Speech was the modality used by the human getting dressed to provided instructions the other human performing the dressing task. Data processing of gaze with respect to shoulder and torso from the 23 point on the body were extracted and used for the prediction model. Additionally, the three main sequence steps (hand-elbow-shoulder) of the dressing task where encoded and presented as part of the output of the model to be predicted. Being able to obtain a higher categorical classification with utterances in the interaction would indicate that speech utterances frequency can be linked to progress in a collaborative task. Due to having discontinuities in speech utterances, a dual pipeline approach to the LSTM network was considered. A LSTM network was used to process word embeddings (explicit utterances) and another LSTM networks for extracting features from the implicit modalities. Provided that speech utterances are not continuous the two LSTMs models will be trained separately and combined on a concatenation layer. These combined features are fed through fully connected LSTM layers leading to an output layer with 3 outputs (hand-elbow-shoulder). These outputs would respectively represent the dressing up to the hand, elbow or shoulder (completed task).

V. EXPECTED RESULTS

The preliminary analysis of time difference between speech utterances and the progression between the three stages of the dressing task suggest that sequence classification of task state is likely to be improved. This indicates that explicit speech utterances can be related to the main state changes of this collaborative task and such modality can be used to extract context of progression to achieve the final goal in a collaborative task.

A better prediction of task sequence indicates that speech utterances are an explicit modality of interaction used to dictate progression or need of change in the interaction approach between human-human interactions.

VI. CONCLUSIONS

The current paper introduces an approach of examining the importance of explicit speech utterances in relation to progress in a collaborative task between humans. The preliminary results indicate that explicit speech utterances are important to measure collaboration and progress in the final task objectives.

As future work, we plan to finalize results and further evaluate the incorporation of similar explicit speech utterances as one of the modalities to assess the level of collaboration between a robot and a human in a real scenario. Specifically, the evaluation of collaboration will be based on alterations in the modality of speech when distraction are introduced in the environment of the interaction.

VII. ACKNOWLEDGEMENT

This work has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 721619 for the SOCRATES project.

REFERENCES

[1] P. Menezes, J. Quintas, and J. Dias, "The Role of Context Information in Human-Robot Interaction," *RoMan 2014 Workshop on Interactive Robots for aging and/or impaired people*, pp. 4–7, 2014.

[2] J. Quintas, G. S. Martins, L. Santos, P. Menezes, and J. Dias, "Toward a Context-Aware Human-Robot Interaction Framework Based on Cognitive Development," *Ieee Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–11, 2018.

[3] G. Chance, P. Caleb-Solly, A. Jevtic, and S. Dogramadzi, "What's up? Resolving interaction ambiguity through non-visual cues for a robotic dressing assistant," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 284–291, IEEE, 8 2017.

[4] K. Li, S. Sun, J. Wu, X. Zhao, and M. Tan, "Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-like Decision Mechanism," 2018.

[5] H. C. Ravichandar, A. Kumar, A. P. Dani, and K. R. Pattipati, "Learning and Predicting Sequential Tasks Using Recurrent Neural Networks and Multiple Model Filtering," pp. 331–337, 2016.

[6] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Tech. Rep. September, 2014.

[7] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, “Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks,” *PLoS ONE*, vol. 11, no. 1, 2016.

A First Step Towards Understanding the Effect of an Interactive Robot on User Experience in Motivational Interview

Neziha Akalin[1]

***Abstract*— This paper proposes a system where the robot gets facial valence and vocal arousal of the user during the interaction. The system maps the social cues onto two dimensional emotional scale. In experimental study, the robot will conduct motivational interview and become interactive in case of lower emotional states. The planned experimental results will show the effect of the socially interactive robot during the motivational interviewing.**

***Index Terms*— socially interactive robots, motivational interview, dimensional model of emotion.**

I. INTRODUCTION

To facilitate successful interaction between humans and robots, interaction needs to be natural and share similarities with human-human interaction. It is important for social robots to understand verbal and nonverbal social cues, which is a prerequisite for natural, safe and comfortable interaction and also helps in anticipating the needs and expectations of the user [8]. For socially interactive robots whose primary function is to interact with people, social interaction plays a crucial role [9]. It is important for these robots to encourage users pro-actively in social interaction.

The term "user experience" is a multifaceted concept and hard to define. Hartson et. al. [12] provided a definition as "the totality of the effect or effects felt by a user as a result of interaction with, and the usage context of, a system, device, or product, including the influence of usability, usefulness, and emotional impact during interaction and savoring memory after interaction". User experience comprises the users emotions, beliefs, preferences and perceptions that arise before, during, and after technology use [10]. The way a robot behaves during interaction with a human may affect their feeling of security, which is one of the dimensions of user experience [11]. Positive user experience with robots is necessary for achieving intended benefits [10]. Human-oriented perception, that is the capability of the robot to track human features (face, voice etc.) is also considered one of the aspects of user experience in human-robot interaction (HRI) [11].

Motivational interviewing (MI) is an emphatic and collaborative conversation style, which is goal oriented and designed for strengthening an individual's motivation toward a particular goal with a commitment to change [1]. Regular physical activity (PA) has been shown to reduce the risk of several chronic diseases [5]. MI could increase individuals physical activity (PA) [4].

[1]Department of Science and Technology, Örebro University, SE-701 82 Örebro, Sweden, neziha.akalin@oru.se

MI is usually delivered by a counselor in a face-to-face conversation. Social robots have the ability to engage participants for a motivational interview [2]. In the context of robot-based delivery mode, the counselor could be substituted by an embodied humanoid robot. There are very few studies employing a humanoid robot in MI [2], [3]. In [3], the authors employed a humanoid robot as a motivational agent in order to increase individuals' motivation towards physical activity. Their findings showed no benefit of MI on participant perceptions compared with traditional advice. They argue that the lack of positive effect of MI might be due to errors in speech recognition and incongruous nonverbal behaviors. In the more recent work [2], they also employed a humanoid robot that delivered a scripted motivational interview in physical activity. The results showed that many of the participants enjoyed the interaction and positively appraised the nonjudgmental aspect of the motivational interview with the robot.

The planned experimental procedure is inspired by [2] and followed the similar experimental design with the difference of an interactive robot. In our work, the robot takes into account the social cues (facial and vocal) and becomes interactive when the social cues are in the second and third quadrant of the dimensional emotional scale.

In the remainder of this paper, an overview of the proposed system is given in Section II. The method, experimental design and procedure are described in Section III, and the paper is concluded in Section IV.

II. PROPOSED SYSTEM

Emotions are modeled in two ways; discrete approach [13] and dimensional approach [14]. In the discrete approach, emotions are categorized into six basic emotions i.e. anger, disgust, fear, happiness, sadness and surprise [13]. In the dimensional approach, affective space described in dimensions. In Russell's valence-arousal scale, each emotional state can be placed on a 2D plane with horizontal axis (valence) and vertical axis (arousal) where valence ranges from unpleasant to pleasant and arousal ranges from calm to excited [14]. The four quadrants of valence-arousal space are as follows: high valence-high arousal (HVHA), low valence-high arousal (LVHA), low valence-low arousal (LVLA) and high valence-low arousal (HVLA). The emotions mapped to the second quadrant (LVHA) and third quadrant (LVLA) are negative emotions (angry, nervous, annoyed, sad, bored etc.) In dimensional emotion recognition, arousal is better predicted using audio cues whereas for valence, visual cues perform better [15].

In our proposed system, we adopted the aforementioned two-dimensional arousal and valence model [14]. The output emotion values in our system are provided by Affdex SDK [16] and openSMILE [17]. Affdex SDK [16] is a real-time facial expression recognition toolkit, trained on more than 5 million human faces to classify facial expressions and used in human-computer applications. openSMILE is an open-source audio analysis tool that is written in C++ and provides audio feature extraction in real-time [17]. The implemented system integrated these toolkits as ROS packages that can be used on any ROS-compatible robotic platform. Since arousal prediction is better from audio and valence prediction is better from vision, in the proposed system, we use facial valence and vocal arousal values. They are mapped onto two dimensional space. If the mapped emotions are in the second and third quadrant, the robot becomes interactive and expresses empathy which is one of the components of motivational interviewing. The proposed system is depicted in Figure 1.

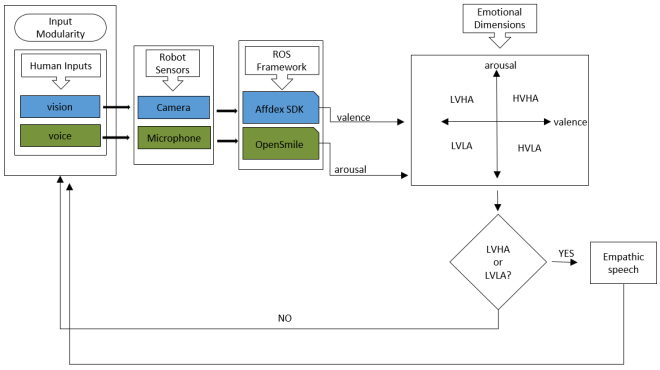


Fig. 1. The proposed system.

III. METHOD

A. Research Questions

The main focus of interest in this work is to test the impact of a socially interactive robot in motivational interviewing. We examine how a socially interactive robot that regards social cues can encourage participants to change their behavior and to provide a better user experience. We aim to evaluate the perceived effect of motivational interviewing in physical activity and the user experience. The research questions are as follows:

- Is the participants' perceived experience better when the robot is socially interactive?
- Can a socially interactive robot contribute to the effect of motivational interviewing?
- Can a socially interactive robot motivate participants to engage in physical activity?

B. The Experimental Procedure

The planned experiment takes place at Örebro University in PeisHome2 which is a living-room-like laboratory used for human-robot interaction experiments. Each session starts with informing the participant about the experiment. Thereafter, the participant is left alone with the robot (Pepper, humanoid robot) in the room and the motivational interview begins.

At the end of the interview, the participants fills out the Godspeed questionnaire [7] and the following Likert scale questionnaire in which each question has options ranging from "Strongly disagree" to "Strongly agree":

- This interview affected my motivation in a positive way.
- The robot helped me to recognize the need to change my behavior.
- I found the interview with the robot engaging.
- The content of each question was clear.
- The robot helped me to talk about changing my behavior.
- The robot helped me discuss the pros and cons of my behavior.
- I got frustrated during the interview.
- It was important listening to myself discussing my behavior.
- The robot acted as a partner in my behavior change.
- The robot helped me feel confident in my ability to change my behavior.
- I would use a robot like this in the future to keep me motivated.

This questionnaire is a modified version of Client Evaluation of Counseling [6] and the questionnaire developed in [2]. The questions in [2] were open-ended, we modified some of the questions as closed-ended. We are also planning to present the questions in [2] in the form of open-ended as optional questions.

IV. CONCLUSIONS AND FUTURE WORK

This position paper presents the general outline of the planned work. Thus far, we have implemented the system. There are a few studies using a robot in motivational interviewing, however no other study considers an interactive robot. The main challenges for our approach is timing of the robot speech since we do not have any reliable speech recognition in the current system. As future work, the proposed system and methodology remains to be evaluated with user studies. We will focus on exploring the effect of socially interactive robots that take into account facial and vocal cues and becomes interactive. The experimental results will show the effect of the socially interactive robot during the motivational interviewing.

ACKNOWLEDGMENT

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

REFERENCES

[1] Miller, William R., and Stephen Rollnick. Motivational interviewing: Helping people change. Guilford press, 2012.
[2] da Silva, Joana Galvo Gomes, et al. "Experiences of a Motivational Interview Delivered by a Robot: Qualitative Study." Journal of medical Internet research 20.5 (2018).

[3] Kanaoka, Toshikazu, and Bilge Mutlu. ”Designing a motivational agent for behavior change in physical activity.” Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. ACM, 2015.
[4] Martins, Renata K., and Daniel W. McNeil. ”Review of motivational interviewing in promoting health behaviors.” Clinical psychology review 29.4 (2009): 283-293.
[5] Warburton, Darren ER, Crystal Whitney Nicol, and Shannon SD Bredin. ”Health benefits of physical activity: the evidence.” Canadian medical association journal 174.6 (2006): 801-809.
[6] Madson, Michael B., et al. ”Measuring client experiences of motivational interviewing during a lifestyle intervention.” Measurement and Evaluation in Counseling and Development 48.2 (2015): 140-151.
[7] Bartneck, Christoph, et al. ”Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots.” International journal of social robotics 1.1 (2009): 71-81.
[8] Tapus, Adriana, et al. ”Perceiving the person and their interactions with the others for social roboticsa review.” Pattern Recognition Letters (2018).
[9] Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn. ”A survey of socially interactive robots.” Robotics and autonomous systems 42.3-4 (2003): 143-166.
[10] Alenljung, Beatrice, et al. ”User experience in social human-robot interaction.” International Journal of Ambient Computing and Intelligence (IJACI) 8.2 (2017): 12-31.
[11] Weiss, Astrid, et al. ”The USUS evaluation framework for human-robot interaction.” AISB2009: proceedings of the symposium on new frontiers in human-robot interaction. Vol. 4. 2009.
[12] Hartson, Rex, and Pardha S. Pyla. The UX Book: Process and guidelines for ensuring a quality user experience. Elsevier, 2012.
[13] Ekman, Paul, et al. ”Universals and cultural differences in the judgments of facial expressions of emotion.” Journal of personality and social psychology 53.4 (1987): 712.
[14] Russell, James A. ”A circumplex model of affect.” Journal of personality and social psychology 39.6 (1980): 1161.
[15] Nicolaou, Mihalis A., Hatice Gunes, and Maja Pantic. ”Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space.” IEEE Transactions on Affective Computing 2.2 (2011): 92-105.
[16] McDuff, Daniel, et al. ”AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit.” Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM, 2016.
[17] Eyben, Florian, Martin Wllmer, and Björn Schuller. ”Opensmile: the munich versatile and fast open-source audio feature extractor.” Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010.

The Effect of Affective Robot Behaviour on the Level of Attachment After One Interaction

Anouk van Maris[1]

Abstract— **Becoming emotionally attached to an assistive robot may have an impact on one's behaviour towards that robot. Therefore, it is important to investigate when attachment occurs and what strengthens it. This study investigated whether people can become attached to a robot after a single interaction, and whether the level of attachment differs according to the affective behaviour of the robot. No significant differences were found for the affective behaviour of the robot. This indicates that people do not become attached after a single interaction with a robot, and that affective behaviour does not influence attachment. However, non-significant differences and a low number of participants are reason for future research.**

I. INTRODUCTION

The number of older adults and their demand for care is growing, but the capacity to supply this demand is not [1]. Therefore, robots are being considered as a possible solution to meet the growing demand of care for older adults that cannot be met by the small number of caregivers. Before social robots can become useful additions to caregivers, the effects on older adults interacting with such a robot should be known. Becoming attached to a robot can provide benefits (e.g. alleviate loneliness and improve well being), but also disadvantages (e.g. increased dependence of the robot). The robot's affective behaviour may have an influence on this level of attachment, since affective behaviour results in a more natural interaction with the robot. However, the user may be deceived by this affective behaviour of the robot and raise false expectations of its abilities. Therefore, it is important to establish whether people become more closely attached to an affective robot with respect to a non-affective robot. This study aims to provide an impression to help planning a study regarding level of attachment of older adults to a social robot. It investigates whether there is a difference in level of attachment to a robot depending on the robot's affective behaviour after a single interaction.

II. BACKGROUND

The fact that people react to computers as social actors [2] is an indicator that they can become emotionally attached to machines and robots [3], [4]. If emotional attachment to a robot is high, the usability of this robot is perceived more positively and the intention to use it in the future is higher [5], [6], [7], resulting in a higher level of acceptance of this robot. Concerns of becoming emotionally attached to a machine or artificial agent (e.g. a too high level of dependency), have been raised at a theoretical level [4]. Therefore, the idea of becoming attached to a robot is not always welcomed. An example of this was found in a survey, where less than half of the participants thought it was acceptable for a child with Autism Spectrum Disorder to become attached to a robot [8]. Therefore, when and how attachment to robots occurs, and the consequences of this attachment, should be thoroughly investigated.

People have researched the level of attachment towards an assistive robot in the past, for example Weiss et al. [9] investigated whether adults and children could become emotionally attached to the robotic dog AIBO. They found that children became emotionally attached to the robot rapidly, where adults seemed to need a longer lasting interaction to form their first impression. However, as stated in the paper most adults observed the children that were interacting with the robot and did not interact with the robot themselves. This may have had an influence on the different outcomes for children and adults. Also, the number of adults participating in this experiment was far less than the number of children (18 versus 129) which may have had an influence as well. A different study that investigated attachment, which was performed by Sung et al. [5], found that people gave their Roomba vacuum cleaner a nickname and thought of it in terms of 'he' and 'she' instead of 'it'.

However, the studies mentioned above used non-anthropomorphic robots for their research. According to Weijers [10], it depends on the function and design of the robot whether it is perceived more like a machine or like a living thing. Also, people interact with social interfaces in the same way as they would with other humans [2]. This makes it likely that people become attached at a different level to a humanoid robot than the robots used in the studies mentioned before. Therefore, the study that will be discussed in this paper investigated the level of emotional attachment towards a humanoid robot. More specifically, it was investigated whether affective robot behaviour had an influence on the level of emotional attachment. It is expected, since affective robot behaviour results in a more natural interaction between a robot and its user, that affective behaviour results in people becoming more attached to the robot showing affective behaviour.

III. METHOD

In total 9 people (including 4 females) participated and completed the experiment (*min age* = 53, *max age* = 71, *M* = 61, *SD* = 4.8). Five participants interacted with a non-affective robot (2 female, 3 male), and four participants

[1]Anouk van Maris is a PhD student at the University of the West of England, Bristol Robotics Laboratory, Coldharbour Lane BS16 1QY, Bristol, United Kingdom anouk.vanmaris@uwe.ac.uk

interacted with the affective robot (2 female, 2 male). Participants were recruited through distribution of an email to university staff. Only people of age 50 and over were asked to participate, since a follow-up research to this study will involve older adults, and in a previous study age showed to have an influence on how people perceived the robot [11]. Fig. 1 shows the experimental setup. The experiment was run using Wizard-of-Oz, where behaviours are pre-programmed but can be run according to the responses of the participants. The wizard/experimenter was located behind the blue screen shown behind Pepper in Fig. 1, so they can hear the participants' responses but the participants could not see them operating the robot. As can be seen in Fig. 1, the robot used in this experiment is Pepper from Soft Bank Robotics [1].

The interaction involved a discussion regarding the seven wonders of the ancient [2] and modern [3] world. The robot would ask whether the participant could name some and would provide information on these wonders. If the participant could not name any more wonders, a list was shown on Pepper's tablet and it would ask what wonder the participant would like to discuss next. This would continue until all wonders were discussed. In the non-affective condition, the robot would not show affective behaviours during the interaction. In the affective condition, it would do so by for example saying a monument got destroyed in a fire showing sad behaviour or a monument still being mostly intact showing happy behaviour. The behaviours for showing these sad, happy and non-affective behaviours have been established in previous research [11]. Characteristics that were used to show the different emotions are head position and pitch of voice, among others. The interaction would last for approximately 20 minutes.

Before the start of the interaction, people were asked to fill in demographics and the Adult Attachment Scale [12] to determine their attachment style. After the interaction they had to fill in questionnaires regarding human and object attachment (adapted from [13] and [14]), together with the questions whether they thought the robot experienced emotions during the interaction and how often they would use the robot in the future if they had one for themselves. Questionnaires from both human and object attachment were used, as it depends on the robot's appearance and function whether it is perceived as an object or a living thing [10]. The human attachment questionnaire is divided in two categories: care and over-protection.

IV. RESULTS

All participants interacting with the non-affective robot reported they did not believe the robot experienced emotions during the interactions. All participants interacting with the affective robot reported that they did believe that the robot experienced emotions during the interaction. This indicates the implemented behaviours were perceived as intended.

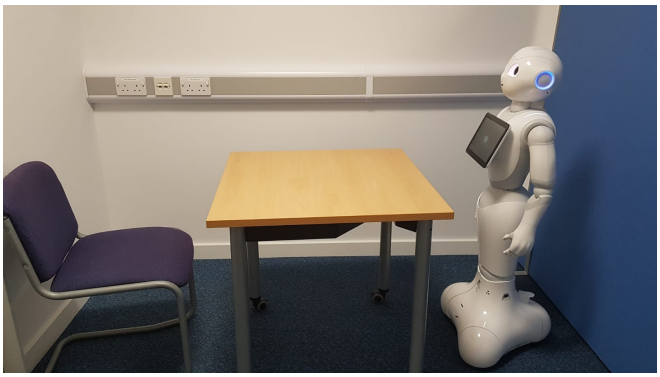


Fig. 1. Experimental set-up

The affective state of the robot did not have a significant influence on object attachment ($F(1,9) = 1.94$, $p = 0.21$). There was no significant effect of the robot's behaviour on person attachment found, neither for care ($F(1,9) = 0.027$, $p = 0.87$) nor for over-protection ($F(1,9) = 2.30$, $p = 0.17$). No correlations were found between participants' attachment style and object or person attachment. This would suggest that the robot's affective behaviour does not have an influence on people's attachment towards the robot after a single interaction.

None of the participants felt the robot cared much for them, since the care score of the human-attachment questionnaire was low for all participants. Some participants did feel the robot was over-protective. Although, as mentioned before, these results were not significant. This perceived over-protection occurred more for participants who interacted with the affective robot (3 out of 4) than the non-affective robot (1 out of 5). Participants interacting with the affective robot scored on average lower on the care-statements for the robot ($M = 15.0$, $SD = 4.2$ for the non-affective robot, $M = 10.8$, $SD = 5.0$ for the affective robot). These participants scored on average higher on the over-protection statements for the robot ($M = 7.8$, $SD = 2.5$ for the non-affective robot, $M = 10.0$, $SD = 1.6$ for the affective robot). Even though not significant, on average participants interacting with the affective robot scored a bit higher on object attachment ($M = 1.80$, $SD = 0.59$) than participants interacting with the non-affective robot ($M = 1.72$, $SD = 0.81$), which may suggest that people can become more attached to a robot showing affective behaviour. This result holds for the intention to use as well, where participants interacting with the affective robot scored a lower average ($M = 3.75$, $SD = 2.2$) than participants interacting with the non-affective robot ($M = 3.80$, $SD = 2.3$). A low average indicates higher intention to use. Even though not significant, the trend was found that people interacting with the affective robot would be more willing to use it in the future than participants interacting with the non-affective robot, as 1 out of 4 indicated they would not use the robot at all for the affective condition, where 3 out of 5 participants indicated this for the non-

[1] https://www.softbankrobotics.com/emea/en/robots/pepper

[2] https://www.wonders-of-the-world.net/Seven/List-of-the-seven-wonders-of-the-ancient-world.php

[3] https://www.wonders-of-the-world.net/Seven/List-of-the-seven-wonders-of-the-modern-world.php

affective robot.

V. DISCUSSION

The robot's affective behaviour did not have a significant influence on people's attachment. A potential cause can be the low number of participants. Another possible explanation for the absence of significant results is the nature of the interaction; that it was too informative and not personal enough for people to form an attachment. However, the interactive nature of the interaction was chosen so the interaction would remain the same for all participants, which would be harder to control when it would have been more personal.

Even though results were not significant, differences were found between conditions for human attachment, object attachment and intention to use in the future. The low scores for the care statements of the human-attachment questionnaire may be caused, as mentioned before, by the informative nature of the interaction, with too few personal additions. This may also have resulted in higher scores for over-protection, since participants might have felt they were not given enough freedom for a natural interaction with the robot.

Overall, attachment scores were low (average of 1.72 and 1.80 out of 5). This is similar to the result found by [9], which indicated that adults need more time than a single interaction to become attached to a robot.

Lastly, differences between the two conditions were found for willingness to use the robot in the future. However, since the number of participants was low this can also be caused by interpersonal differences. This will be investigated in future research.

VI. CONCLUSIONS

This study aimed to establish whether affective robot behaviour has an influence on a person's attachment towards that robot after a single session. Results show that behaviour does not have an influence. However, the small number of participants may have influenced these results, since non-significant differences between the two conditions were found. Therefore, this topic will be investigated in more depth in the future. Future work will investigate the effect of affective robot behaviour on older adults and their attachment towards the robot after several interactions spread over a long-term period, also taking into account habituation. It is expected that older adults will become more easily attached to the robot, since they may have fewer interactions on a daily basis compared to the participants in this study (university staff), which may influence their expectations of interacting with a robot.

ACKNOWLEDGMENT

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 721619 for the SOCRATES project

REFERENCES

[1] N. Unies, "World population prospects. the 2015 revision, new york, united nations," *Population Division*, 2015.
[2] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.
[3] M. R. Banks, L. M. Willoughby, and W. A. Banks, "Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs," *Journal of the American Medical Directors Association*, vol. 9, no. 3, pp. 173–177, 2008.
[4] J. P. Sullins, "Robots, love and sex: the ethics of building a love machine," *IEEE transactions on affective computing*, p. 1, 2012.
[5] J.-Y. Sung, L. Guo, R. E. Grinter, and H. I. Christensen, "my roomba is rambo: intimate home appliances," in *International Conference on Ubiquitous Computing*. Springer, 2007, pp. 145–162.
[6] P. Zhang and N. Li, "The importance of affective quality," *Communications of the ACM*, vol. 48, no. 9, pp. 105–108, 2005.
[7] C. Wilson, "Is it love or loneliness? exploring the impact of everyday digital technology use on the wellbeing of older adults," *Ageing & Society*, pp. 1–25, 2017.
[8] M. Coeckelbergh, C. Pop, R. Simut, A. Peca, S. Pintea, D. David, and B. Vanderborght, "A survey of expectations about the role of robots in robot-assisted therapy for children with asd: Ethical acceptability, trust, sociability, appearance, and attachment," *Science and engineering ethics*, vol. 22, no. 1, pp. 47–65, 2016.
[9] A. Weiss, D. Wurhofer, and M. Tscheligi, "i love this dogchildrens emotional attachment to the robotic dog aibo," *International Journal of Social Robotics*, vol. 1, no. 3, pp. 243–248, 2009.
[10] S. Weijers, "Exploring human-robot social relations," Master's thesis, University of Twente, 2013.
[11] A. van Maris, N. Zook, P. Caleb-Solly, M. Studley, A. Winfield, and S. Dogramadzi, "Ethical considerations of (contextually) affective robot behaviour," Forthcoming.
[12] N. L. Collins and S. J. Read, "Adult attachment, working models, and relationship quality in dating couples." *Journal of personality and social psychology*, vol. 58, no. 4, p. 644, 1990.
[13] A. Mackinnon, A. Henderson, R. Scott, and P. Duncan-Jones, "The parental bonding instrument (pbi): an epidemiological study in a general population sample," *Psychological Medicine*, vol. 19, no. 4, pp. 1023–1034, 1989.
[14] H. N. Schifferstein and E. P. Zwartkruis-Pelgrim, "Consumer-product attachment: Measurement and design implications," *International journal of design*, vol. 2, no. 3, 2008.

Emotion-Motion Interaction as a baseline for understanding non-verbal expression of computational empathy and user expectations*

Naomi Yvonne Mbelekani

***Abstract*—Most studies have explored user empathy towards robots, however, there is a lack of studies that explore emotional empathetic responses and reactions from a robotic arm. An argument is therefore presented on the use of emotions in robotic application, focusing on the ability to convey emotional information through motion and empathy, i.e. reading emotional information from motion. The study purpose was to describe robot motion as an expression of emotion (empathy). A small use study is presented where the effect of one factor of motion, namely speed, on the empathic perception in humans is investigated. An experiment was conducted to examine the reaction and response of participants regarding the expression of emotion through the movement of a robotic arm. Since it's a preliminary study, only four participants were recruited. Results reveal an interplay between emotion and motion as baselines for understanding non-verbal expressions of empathy by a robotic arm, as well as users' expectations towards the robot. The concept discussed is very relevant to advancing the quality of human-robot interaction (HRI).**

I. Introduction

Robots are becoming faster, cheaper, capable, and more flexible in performing different tasks and more interactive with humans [11]. As social robots are a trending field within HRI, there is a rapid need for them to be socio-emotionally intelligent (acquire social skills). As a result, affective computing consists of applying emotions to a robot, giving it the ability to recognize and express them, developing its ability to respond intelligently to human emotion, and enabling it to regulate and utilize its emotions [3]. This places emphases on enriched interaction patterns between humans and robots, by providing a prospect for assistance, companionship, and even therapy for those experiencing physical or mental distress [5]. However, this debate is usually lost when discussing nonverbal interaction of robotic arms and the expression of emotions and computational empathy. [8] argued that nonverbal information (motion, posture, gestures) is vital for social interaction. This are its communicative interface to the user, which serves likeability, increases user satisfaction, and perceived as trustworthy [4]. By exploring how emotion and motion interacts, we could come up with strategies to understanding non-verbal expressions of emotional empathy. Furthermore, we could even explore users' expectation regarding a robot that expresses its emotions through motion, therefore understand whether motion and emotion correlate. For that reason, the main objective of this study was to read information from motion and emotional empathy, as well as users' expectations regarding the interaction.

II. Motion in [E]motion: Move Physically and the Power to Move Emotionally

The relationship between motion (bodily movement) and emotion (feelings) may not be considered an etymological coincidence by some scholars. Most of us may not even have thought about this, but the roots for motion and emotion are virtually identical. The English word emotion comes from the Latin word *Movere* meaning to move and *exmovere* or *emovere* meaning to move out, hence to excite [10]. This derivation suggests a close link between emotion and body movements [10]. Ultimately, meaning there is a close relationship between the two variables, emotion and motion.

Current robots, whether humanoids or robotic arms are being designed to function as industrial aids, as well as as 'social partners' [7] and companions. Therefore, as much as their physical embodiment is considered to be important; their emotional embodiment should also be believed to be the same value. Robots as social agents and allies should be able to embody emotionally empathetic states when interacting with humans, no matter their physical embodiment. For example, robotic arm (KUKA), companions (AIST's PARO), household pets (Sony's AIBO), domestic cleaners (iRobot's Roomba), healthcare assistants (RIKEN Japan's Ri-Man), and educational aids (MIT's Kismet and Leo). Design of such robots depends on the interaction and social skills (Breazeal, 2002). In situations such as robotic arms, non-verbal emotional states of the robot have to be embodied, personified or exemplified by exploiting motion as robot body language. This can be in conjunction with the voice and screen semantic of the robot (if any), without excluding the tempo, pitch and pattern of interaction. For instance, what we refer to as '*emotion-motion interface'* (EMI), therefore, exploring the emotion-motion interaction on the robot.

A theory on body expressions called the Laban Movement Analysis (LMA; Laban, 1980) assumes two opposing forms of body movement: fighting form (active, prominent, brisk movements) and indulging form (unsteady weak movements), which reveal subjective inner attitudes or states [10]. This theory helps us understand how movement or motion expresses internal states. In one study, "change in the robot's motor behavior to match the user's speed invoked an *empathetic Chameleon Effect response* and improved the participants' overall perception of the robot" [2]. It is also argued that body movements' information provides sufficient guidance for people to perceive expression of emotion [7]. In another study that explored the meaning awarded to motion characteristics (for example: speed); it was revealed that perception of emotion such as: fast, jerky movements were linked with anger and happiness, while slow, smooth movements were associated with sadness [7].

Motions with strong velocity or speed tend to be perceived as anger or happiness, while motions with weak velocity tend to be perceived as sadness or tired [10]. Which means that fast speed or velocity does not necessary mean optimistic emotional experiences (e.g. happiness) and slow speed or velocity does not always mean pessimistic emotional experiences (e.g. sadness). Though, [7] argued that differences in kinematics of arm movements have helped differentiate between anger, joy, and sadness. However, we can still argue

*Research supported by SOCRATES, MSCA-ITN-2016–Innovative Training Networks, H2020. Naomi Yvonne Mbelekani (e-mail: ynmbelekani@gmail.com).

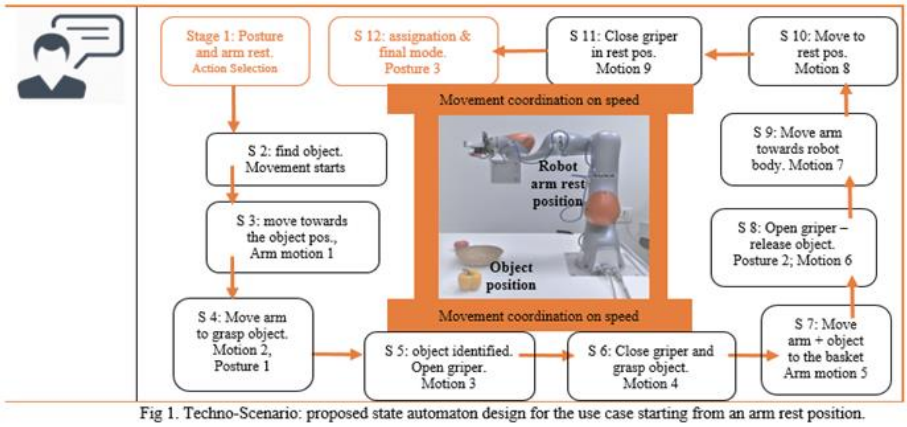


Fig 1. Techno-Scenario: proposed state automaton design for the use case starting from an arm rest position.

that emotions have a tendency to be affected by body motion, for example, people can be emotionally engaged when watching dance routines [7]. As a result, it can also be argued that in observing others move – we are moved ourselves, and while moving ourselves – we move others. Ultimately, it is important to understand the interface, the transition through it, and the meaning it holds for interaction moving forward.

A. *Gesture and posture in emotion communication*

Humans use imitation of gestures and postures as tools of communication, and it is regarded as important in enhancing quality of interaction in HRI [2]. Thus, imitation triggers some social interactions [3]. In one study where participants were asked to use a Nintendo Wii remote to mime gestures simultaneously with a robot, they noted feeling more comfortable while completing the task when the robot synchronized/mirrored their gesture speed [2]. Understanding how users perceive and give meaning to robot postures and gestures supports the design of robots that are able to socially and emotionally interact. Gestures have been identified as crucial to the design of robots [7]. Furthermore, robotic arms' body language (through gestures or postures) should be explored as a medium of conveying robot intentions [4]. We can hence argue that studies on robot postures and gestures are important because of the following reasons borrowed from [7]:

1) studying gesture interpretation is necessary to improve HRI especially for robots that have limited ability for vocal and facial expressivity;
2) previous studies in HRI have focused on how gestures are created without evaluating users' understanding of those gestures, so little is known about what factors affect gesture perception;
3) no previous work has investigated the characteristics of good designers and the role of expertise in gesture authorship.

This study explores non-verbal cues of dialogue and social behavior on a robot's bodily gesture and posture (approachable versus less approachable), and their subjective meanings. People tend to rely on facial expression as key indicator [7], hence it is important to exam whether in the absence of a face, robots can still convey emotional interaction using postures and gestures. Based on the study aim, the following question was explored: Can a robotic arm be considered as expressing emotion (empathy) based on degrees of motion response, gesture and posture? The following hypotheses were tested to address this question:

H 1: the robot's gesticulation motion mirrors emotional empathy based on its interactive speed.

H 2: the robot's postures are an exhibition of subjective meanings.

III. Study Methods

The study aimed to examine a robot's non-verbal expression of emotionally empathetic interaction through motion in a table setting scenario.

Participants: Four female individuals (age range: 19 to 27, M=20) were recruited randomly, whose native language consist of Arabic, English, Malayalam, and Russian. The highest education achieved were high school, bachelor, and masters. The study was conducted in a lab room (Mobile robotics lab) at the Ben-Gurion University of the Negev, Israel.

Apparatus: The robot KUKA LBR IIWA was used in this study, which comprises an interactive interface: one hand with multiple joints, seven actuated Degrees of Freedom (DOF) and a refined control system.

Design: The study was a within-subject experiment design. The independent variable tested was the speed of the robot, two levels: fast (100% full speed) and slow (50 % speed). The dependent variable tested was emotionally empathetic expression. The study measured subjectivity by questionnaires. Descriptive analysis of the study results was conducted afterwards.

Procedure: (a) Using Wizard of Oz, the scenario involved asking the participant to give voice commands to the robot in the form of direction (e.g. left-right, up-down, back-forward) on picking objects and placing them in a basket (see Fig 1.). They experienced two different motions: slow and fast mode. The robot was operated to adopt and personify a slow motion profile in contrast to fast motion. After the interaction, the participants were instructed to answer questions related to the interaction, e.g. "Pretend that you are in the scenario and you are feeling sad/happy, then describe how you would experience the interaction with the robot."

The predominant framework model in this study is, in essence, a model in which motion and emotion interacts which in turn predispose or motivate the robot towards explicit behaviors. For empathy is a contested concept and emotion is a broad phenomenon, the following are used as frameworks of understanding emotionally empathetic interaction in this study: compassion, friendly, understanding, intentional, relatable, considerate, and trustworthy. While interfering and annoying are used as negative emotional experiences.

(b) As the study was divided into two parts, the second part involved asking the participants to describe the meaning of the

robot posture, e.g., "what message do you think is the robot conveying to you?". They were provided with six postures, and to avoid bias by giving them a selection of emotions and them merely picking out what they think the researcher preferred, we asked them to think about it and give their own thought processed meanings. This required them to actually analyze the posture and give meaning based on their own understanding. This study used an image display methodology to acquire understanding.

Participants also filled out demographic information, the Technology Adoption Propensity questionnaire (TAP) [13], and the Negative Attitudes toward Robots Scale (NARS) survey [15]. This study attempts to design interaction using non-verbal programmable emotionally empathetic traits.

Measure: In the first part (a) of the study, we considered the subjective reality, thus how many times participants ticked an answer, which we then calculated to check for predilections. We then used descriptive analysis to understand the results achieved. In the second part (b) of the study, we paid attention to the distribution of subjective meaning on a robot's postures, based on participants' understanding.

IV. Preliminary Results and Discussion

The results achieved from the preliminary study, revealed the following notions. Based on the TAP, on a scale from 1-strongly disagrees to 5-strongly agree; 2 participants scored 4-agree and 2 participants scored 5-strongly agree towards the question: "technology gives me more control over my daily life". This revealed positive attitudes towards technology as participants thought it gives them a sense of control in their everyday lives. These results show confidence and a positive attitude towards the use of new technological devices. With regards to the NARS, on a scale from 1-strongly disagrees to 5-strongly agree; 3 participants scored 5-strongly disagree and one participant scored 1-strongly agreed on the question "I would feel uneasy if robots really had emotions." This revealed positive attitudes towards robots that display emotions and that they did not have negative thoughts about robots expressing emotions.

A. *Effects of motion and emotion valence*

Work on expressive robots for emotional interaction with humans is receiving increasing attention. Robots can engage in social interaction through socio-emotional intelligence [1], which enables the robot in sensing and interpreting various human emotions, moods and attitudes to guide its interaction. The processes used as frameworks for understanding emotion and empathy interaction in HRI are:

For the slow mode: The results on their experience interacting with the robot revealed that the interaction was perceived as understandable and friendly. When asked if you were feeling sad, how would you consider the robot's motion, as a result, concepts such as considerate, friendly and relatable separately received a less score; while trustworthy received a moderate score, and understandable was highly scored; whereas annoying received the least score. This tells us that slow motion has emotional significance, and may be considered in instances where a user may be feeling sad. In the scenario were the user may be feeling happy, a majority of the participants opted for friendly and trustworthy; while a small number scored considerate, understanding, and relatable. The results on the slow motion and its relation to emotional experience, revealed a positive attitude and emotional experience.

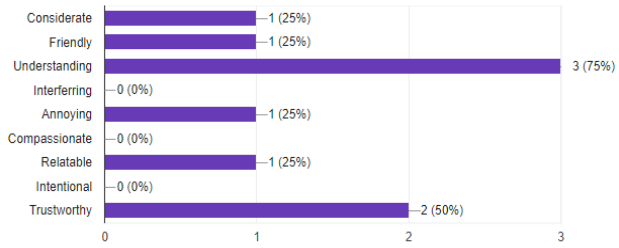


Fig 2. Slow mode: Feeling sad and the robot's motion:

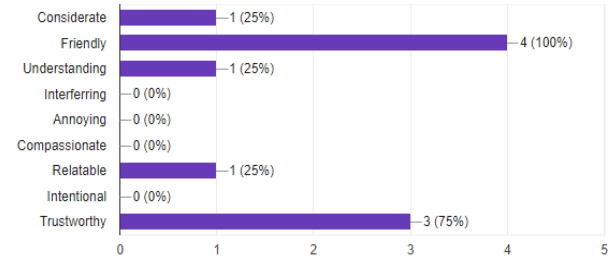


Fig 3. Slow mode: Feeling happy and the robot's motion:

For the fast mode: The results in this case revealed that a majority of the participants chose positive experiences while interacting with the robot. With regards to them feeling sad while interacting with the robot, a majority of the participants considered the robot's motion as considerate and friendly, while understandable, relatable and trustworthy received an average score; and the least score being interfering and annoying. In cases where the participants may be feeling happy, results revealed high scores for the robot as considerate, friendly and trustworthy; with the robot as understandable, compassionate and relatable receiving an average score; while interfering received the least score. Similar to the slow motion mode, the results show a positive emotional experience towards both motion modes of the robot.

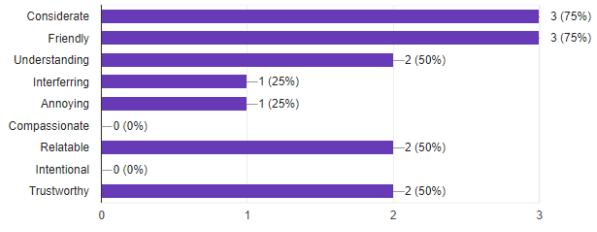


Fig 4. Fast mode: Feeling sad and the robot's motion

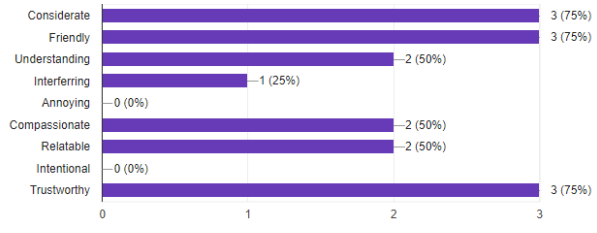


Fig 5. Fast mode: Feeling happy and the robot's motion

As a result, what we see from these results is that people's emotional understanding and experience through motion is varied and subjective. Due to the size of the sample, no clear predictions were made. However, whether fast or slow depending on individual differences, it is accurate to state that motion and emotion interact. Based on the idea that this study's sample size was small, as it is a preparatory study, the current results cannot be generalized to a wider population. Although, we expect to achieve more generalizable result based on a larger sample size on our forthcoming study.

B. Situational context on expressed user expectations

When asked about their expectation regarding non-verbal emotion, majority of the participants stated that they would like for motion of the robot to show emotional empathy, with one preferred a verbal interaction. The following results show when users would like for the robot to express emotional empathy, see Fig 2:

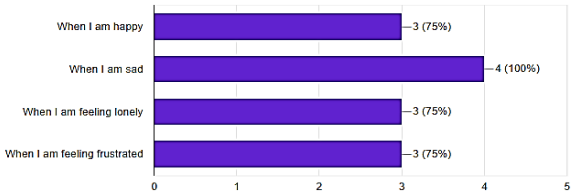


Fig 6. Expressed user expectation.

C. Posture characteristics and expressed meaning

With regards to robot gesture and posture, descriptions were given on the robot's posture (see Fig 3).

Fig 3: Expressed meaning on robot posture

Participants had different meanings on the different postures. This shows that robot postures and the meaning they convey are subjective. What one sees as friendly, another may see it as not friendly. As a result, we can argue that meaning is subjective and a response to emotion.

V. Conclusion

The study contributes to the field of emotion (empathy) expression and user expectations in HRI. This aims to look at certain properties of the robot, such as motion and emotion interaction. This study is important because adding emotion to motion or creating natural emotionally meaningful movement is one of the next and anticipated phases of robotics, thus proving valuable for robots. However, due to the small sample size, the results cannot be generalized to a larger population, but can be seen as a starting point. Furthermore, no clear correlation between motion and emotion was observed. Thus, whether one had an effect on the other cannot be claimed without further results.

Acknowledgment

We acknowledge the assistance provided by the participants and reviewers involved in making this study a success, as well as the funding provided by SOCRATES, an MSCA-ITN-2016 - Innovative Training Networks funded by EC under grant agreement No 721619.

References

[1] Breazeal, C. (2003). Emotion and sociable humanoid robots. 59: 119-155. DOI:10.1016/S1071-5819(03)00018-1

[2] Burns, R., Jeon, M., & Hyuk Park, C. (2018). Robotic Motion Learning Framework to Promote Social Engagement. *Applied Science, 8*: 241. DOI:10.3390/app8020241

[3] Cañamero, L., & Gaussier, P. (2004). Emotion understanding: Robots as tools and models. DOI:10.1093/acprof:oso/9780198528845.003.0009

[4] Chatterjee, S., Shriki, O., Shalev, I., & Oron Gilad, T. (2016, August 26-31). Postures of a Robot Arm- window to robot intentions? 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Columbia University, NY, USA

[5] Goodrich, M. A., & Schultz, A. C. (2007). Human-Robot Interaction: A survey. 1(3): 203-275. DOI:10.1561/1100000005

[6] Ioannidou, F., & Konstantikaki, V. (2008). Empathy and emotional intelligence: What is it really about? *International Journal of Caring Sciences, 1*(3):118-123. http://www.emotionalliteracyfoundation.org/research/Vol1_Issue3_03_Ioannidou.pdf

[7] Klein, B., Gaedt, L., & Cook, G. (2013). Emotional robots. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry, 26(2): 89-99. DOI:10.1024/1662-9647/a000085

[8] Li, J., & Chignell, M. (2011). Communication of emotion in social robots through simple head and arm movements. *International Journal Social Robot, 3*: 125-142. DOI 10.1007/s12369-010-0071-x

[9] Masuda, M., Kato, S., & Itoh, H. (2009). Emotion detection from body motion of human form robot based on Laban Movement Analysis. *PRIMA, LNAI 5925*: 322-334. https://link.springer.com/content/pdf/10.1007%2F978-3-642-11161-7_22.pdf

[10] Mccoll, D., & Nejat, G. (2013). Meal - time with a socially assistive robot and older adults at a long-term care facility, 2(1): 152-171. DOI:10.5898/JHRI.2.1.McColl

[11] Morita, J., Nagai, Y., & Moritsu, T. (2013). Relations between body motion and emotion: Analysis based on Laban Movement Analysis. https://mindmodeling.org/cogsci2013/papers/0202/paper0202.pdf

[12] Truschzinski, M., & Müller, N.H. (2014). An emotional model for social robots. ACM/IEEE international conference on human-robot interaction '14, Bielefeld, Germany.

[13] Ratchford, M, and Barnhart, M. (2012). Development and validation of the technology adoption propensity (TAP) index. *Journal of Business Research, 65*: 1209–1215. DOI:10.1016/j.jbusres.2011.07.001

[14] Smarr, C. A., Mitzner, T. L., Beer, J. M., Prakash, A., Chen, T. L., Kemp, C. C., & Rogers, W. A. (2014). Domestic robots for older adults: attitudes, preferences, and potential. *International Journal of Social Robotics, 6*(2): 229-247. DOI: 10.1007/s12369-013-0220-0

[15] Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). The Negative Attitudes Towards Robots Scale and reactions to robot behaviour in a live Human-Robot Interaction study. in Adaptive and Emergent Behaviour and Complex Systems: Procs of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009. SSAISB, pp. 109-115. URI: http://hdl.handle.net/2299/9641

[16] Yang-McCourt, I., & Bahli, B. (2014). Motion and emotion: An integrative approach of cognition and emotion in IS usage. *XXII Conférence Internationale de Management Stratégique*.

Organizing committee

Coordinators

Prof. Stefan Wermter
Prof. Thomas Hellström

Invited Speakers

Prof. Silvia Rossi
Prof. Dana Kulic

General and Financial Chairs

Dr. Sven Magg
Chandrakant Bothe
Egor Lakomkin
Henrique Siqueira
Alexander Sutherland
Mohammad Ali Zamani

Program Chairs

Alessandra Rossi
André Potenza
Anouk Van Maris
Chih-Hsuan CHen
Vesna Poprcova
Samuel Olatunji

Local Chairs

Francois Foerster
Alexis Billier
Marie Charbonneau
Aleksandar Taranović
Antonio Andriella
Naomi Yvonne Mbelekani

Publication Chairs

Phuong D.H. Nguyen
Neziha Akalin
Maitreyee Tewari
Cagatay Odabasi

Publication Chairs

Antonella Camilieri
Michele Persiani
Truong Giang Vo
Mohammad Thabet

Scientific Committee

We would like to thank our reviewers on the scientific committee for their excellent feedback.

Dr. Guillem Alenya
Dr. Neha Baranwal
Dr. Pablo Barros
Prof. Angelo Cangelosi
Nikhil Churamani
Dr. Vardit Fleischmann
Dr. Esteban Guerrero
Dr. Martin Heckmann
Shanee Honig
Dr. Aleksandar Jevtic
Prof. Lili Jiang
Dr. Mattias Kerzel
Dr. Andrey Kiselev
Luiza Mici
Prof. Kai-Florian Richter
Dr. Ola Ringdahl
Prof. Alessandro Saffiott
Dr. Avinash Singh
Dr. Hagai Tapiro
Dr. Anand Vazhapilli
Dr. Cornelius Weber