PhD Thesis :

Multimodal analysis and learning of human interactions by a companion robot: detecting and fulfilling user needs

Supervisors :

ENIB / IRL : Pr Cedric BUCHE (cedric.buche@cnrs.fr)
FLINDERS / IRL : Dr Paulo SANTOS (paulo.santos@flinders.edu.au)
ENIB / LAB-STICC : Dr Anne-Gwenn BOSSER (bosser@enib.fr) ?
NAVAL GROUP : Delphine KELLER ?

Context

Service robotics has been identified as having the largest potential market for years to come according to the international statistics about robotics[1], where the companion robot is probably the most promising application. In general terms, the gap that this project aims to solve is the autonomous learning of human interactions by a companion robot such that the robot is capable of identifying when the human needs help and how the robot should behave in this case.

In order to obtain an effective and natural Human-Robot Interaction (HRI), it is of utmost importance for the artificial agent to be able to understand and mimic human-human interaction. To this end, machine behaviour models can take direct inspiration from human social actions. In this context, speech, intonation, non-verbal gestures and emotions are some of the various parameters that compose the set of modalities defining this interaction. Robots must learn all of these aspects in order to have the closest possible robot-human interaction to human-human interaction.

The present proposal tackles an important part of this issue, where only non-verbal, visually-observable, behaviors are taken into account. We make the assumption that

[1] https://ifr.org/ifr-press-releases/news/service-robots-record-sales-worldwide-up-32

robot-human behaviours can be learnt from the observation of human-human interactions. A first challenge that motivates this research is the *multimodal analysis* of human interactions, that includes not only the automatic assimilation of the physical movements involved in non-verbal actions, but also the actors' intentions and object affordances facilitating the identification of situations in which the human needs help, and the inference the robot's course of action in this case. This is related to an age-old question in AI, which is how to infer from raw data high-level, abstract, concepts. In the present context, this manifests itself in learning intentions and affordances in human-human interactions from a robot sensor's standpoint. A second challenge in this project comes from the complexity involved in *faithfully reproducing the movements* of a human by a robot, as the latter and the former do not share the same body shapes. Thus, the effective reproduction of human actions by a robotics agent is one important scientific question this work aims to solve.

The objectives described in the present paragraph can be summarised as follows:

- **Objective 1:** the multimodal analysis of human interactions, that includes not only the automatic assimilation of the physical movements involved in non-verbal actions, but also the actors' intentions and object affordances.
- **Objective 2:** the effective reproduction of human actions by a robotics agent;

Statement of work

This work is organised into the following five work packages (WP), that should be developed within a total of 6 semesters (S1 -- S6).

(a) **WP1- Human trials [objective 1, duration S1, risk: low]-** Conduct experiments (human trials) involving common human-human interactions. This part of the project focuses on the socio-communicative behaviour that should accompany the monitoring of the task. One deliverable of this work package is the development of an experimental protocol that would be the most appropriate to get human subjects engaged in live interactions where the need of direct assistance is implicit (for example, the assembly of a complex IKEA piece of furniture). The challenge here includes multimodal scene analysis and the identification of the assistive actions, and when to apply them. In particular, we will investigate the coordination of gaze, head, torso, arm and hand movements of planned and perceived actions of the humans including the open monitoring of the companion robot's own movements via perception-action loops such as arm-gaze coupling. The low-level multimodal traces of the interactions will be semi-automatically enriched with higher-level annotations related to sub-tasks of the main goal (e.g. the furniture assembly), but that also involve emotional and mental states of the human users. During these interactions, a large amount of data will be collected from multiple sensors. At the lowest level the data will be restricted to videos of human-human interaction.

Additionally to human-human interaction, experiments will also use immersive teleoperation, i.e. “beaming” of human pilots, to control a robot in a robot-human interaction as data input to the learning process.

(b) **WP2 - Pre-processing [objective 1, duration S1, risk: low]** At a more abstract level, information from the user’s face (location of the eyes, nose, mouth, rotation of the head, direction of gaze, degree of opening of the eyes), the distance and the visitor's position in relation to the robot, and the different values of the robot joints will be used, as well as more sophisticated data such as facial expressions and age estimation will be grouped together to qualify the kind of assistance needed by the human user. To do so, pre processing will be done considering several possible solutions. OpenFace [Baltrušaitis et al., 2016] is an open source tool that can be used to detect the position and orientation of a face, estimate the direction of gaze as well as units of action in real time. [Cao et al., 2017] proposed a model capable of detecting 2D poses of several people in real time. OpenPose is a real-time multi-person system to jointly detect the human body, hand, facial, and foot key-points (in total 135 key-points) on a single image. One possibility to perceive movements is to represent the joints in 3D [Martinez et al., 2017]. The VNect model [Mehta et al., 2017] gives an estimate of a 3D pose in real time from a simple RGB camera (red, green, blue).

(c) **WP3 - Implementation [objective 1, duration S2- S3, risk: medium]:** Develop and implement autonomous socio-communicative behaviors into the robot cognitive architecture via statistical modeling of the multimodal behaviors monitored during the prior humans interactions.
Previous annotations will then be used to guide machine learning of diverse sensory-motor mapping tools, such as Semi-continuous Hidden Markov Models (HMM) or Dynamic Bayesian Networks (DBN), aiming at generating actions given the scene analysis and the incremental estimation of the task’s state. Within the project, we will explore machine learning from demonstration with user intentions, incremental statistical learning, parallel analysis and decoding of multimodal data. This architecture will be in charge of the perception-decision-action loop: it will implement the socio-communicative autonomous behaviors not only at the reactive level (gaze, head, arms, hands, torso movements) but also at the cognitive level with automatic turn management as well as reasoning. The cognitive level will provide a priori syntactic, semantic and pragmatic constraints to low-level sensory-motor statistical learning – i.e. admissible paths through all possible associations between observed and planned actions.

The model will be trained with the camera elements as input (plus addition of pre-processing) and the robot joints as output. The model will predict the next move based on the previous moves. Several solutions will be studied. The model of movement learning developed could be an Long Short-Term Memory neural network (LSTM). This type of network allows you to learn sequences, which is very useful especially in a real time system [Oh et al., 2015]. Temporal Deep

Belief Network (TDBN) has shown robust results in the generation of movements [Sukhbaatar et al., 2011]. The model could be the TDBN model [Lasson et al., 2017] if the database contains information about each joint for both the input and the output of the network. Regarding real-time, the models proposed by [Oh et al., 2015] could generate smoother movements. Network learning could use [Hanna and Stone, 2017] to check if the generated motions will not cause damage to the robot. The model proposed by [Huang and Khan, 2017] does not currently apply to the robot. However, the principle of generating one image according to another could perhaps be used in the case of the generation of movements in an interaction. Finally, [Hanna and Stone, 2017] proposed to simulate a physical robot to train a network with physical data.

(d) **WP4 - Evaluation [objective 2, duration S4-S5, risk: medium] Assess t**hese behaviors and the achieved social embodiment with acceptance measures and analysis of user attitudes. We will conduct experiments with a PEPPER humanoid robot. Two kinds of evaluations will be conducted: (a) the test of learnt behavioral models during real-time autonomous Human-Robot interactions, (b) a posteriori validation of socio-communicative profiles by post-hoc analysis of the pilot-user specific behavioral models via multidimensional scaling procedures. We will examine if the a priori social profiling of our test groups has an effective impact on the perceptual judgments of the subjects and the resulting objective behaviors. Behavioral data from the different conditions will be analyzed. We shall be particularly interested in expressive gestures (manual gestures and expressions), any gesture expressing feelings (e.g. enjoyment, anger, fear), states and cognitive processes (e.g. doubt, interest, to think). One objective is to confront human expertise with post-hoc statistical analysis of behavioral models that will be developed.

(e) **WP5 - Evaluation [objective 2, duration S5-S6, risk: low]** Assess these behaviors in the context of the Robocup@home competition which aims to develop the interaction skills between human and robot. As scoring and experiment platforms are organized by an international committee, the evaluation of this research on the competition would attest the scalability of the solutions developed in this project.

References

[Baltrušaitis et al., 2016] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface : an open source facial behavior analysis toolkit. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–10. IEEE.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In CVPR, volume 1, page 7.

[Martinez et al., 2017] Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In ICCV

[Mehta et al., 2017] Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.- P., Xu, W., Casas, D., and Theobalt, C. (2017). Vnect : Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG), 36(4) :44

[Hanna and Stone, 2017] Hanna, J. P. and Stone, P. (2017). Grounded action transformation for robot learning in simulation. In AAAI-17, pages 3834–3840

[Oh et al., 2015] Oh, J., Guo, X., Lee, H., Lewis, R. L., and Singh, S. (2015). Action-conditional video prediction using deep networks in atari games. In Advances in Neural Information Processing Systems, pages 2863–2871.

[Sukhbaatar et al., 2011] Sukhbaatar, S., Makino, T., Aihara, K., and Chikayama, T. (2011). Robust generation of dynamical patterns in human motion by a deep belief nets. In Asian Conference on Machine Learning, pages 231–246.

[Lasson et al., 2017] F. Lasson, M. Polceanu, C. Buche and P. De Loor
Temporal Deep Belief Network for Online Human Motion Recognition
30th International Florida Artificial Intelligence Research Society Conference (FLAIRS) pages 80-85, AAAI Press, 2017.

[Huang and Khan, 2017] Huang, Y. and Khan, S. M. (2017). Dyadgan : Generating facial expressions in dyadic interactions. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2259–2266. IEEE