What Can You See? Identifying Cues on Internal States from the Kinematics of Natural Social Interactions

Madeleine Bartlett
CRNS, University of Plymouth
Plymouth, PL4 8AA, UK
madeleine.bartlett@
plymouth.ac.uk

C E R Edmunds
CRNS, University of Plymouth
Plymouth, PL4 8AA, UK
charlotte.edmunds@
plymouth.ac.uk

Séverin Lemaignan
CRNS, University of Plymouth
Plymouth, PL4 8AA, UK
severin.lemaignan@
plymouth.ac.uk
and
BRL, University of the West of England
Bristol, BS16 1QY
severin.lemaignan@brl.ac.uk

Tony Belpaeme
CRNS, University of Plymouth
Plymouth, PL4 8AA, UK
tony.belpaeme@
plymouth.ac.uk
and
ID Lab – imec
University of Ghent, Belgium
tony.belpaeme@ugent.be

Serge Thill
CRNS, University of Plymouth
Plymouth, PL4 8AA, UK
and
Interaction Lab
University of Skövde
541 28 Skövde, Sweden
serge.thill@
plymouth.ac.uk

Introduction

One goal of research on child-robot interactions is to enable robots to autonomously adapt to a child's behaviour in applications such as tutoring [4] and therapeutic settings [5], for example, adapting to a child's learning or therapeutic needs. This requires robots to track and interpret the internal states of human interaction partners. Studying how humans are able to infer the internal states of others can guide research aiming to endow robots with this skill. Researchers in the fields of psychology and Human-Robot Interaction (HRI) have identified that humans use information such as observed motor activity [7] and contextual information [3] to judge the internal states (e.g. intentions) of others. To design robots able to track the internal states of children it is necessary to first determine what internal-state cues are available from the different sources of information within a social scene, and thereby determine what data are sufficient for internal-state-reading in these scenarios. It is also important to consider the quality and availability of data.

Here we discuss the use of skeletal data which is often easily obtained and, when provided by tools such as OpenPose [9] which deals well with occlusions, of high quality. Specifically, we propose a methodology for identifying what humans gain from the kinematics of a child-child social interaction. The findings from studies based on this method-

ology could act as a baseline for what an artificial system can be expected to glean from such data.

Background

Studies examining the mirror neuron system (MNS) found in primates and humans indicate that humans use observed kinematics to make inferences about the observed actor [7, 3]. Broadly speaking, one can identify two types of theory which describe this process. The first type of theory proposes that recognition is a result of an observer mapping the observed kinematics onto their own motor system which allows them to simulate a representation of the intentions driving the observed action [7]. Importantly, this mechanism uses only kinematic information to infer intention. One problem with this account is that humans are able to deal with situations where the same action could be driven by different intentions (e.g. grasping a cup to drink, or to clean it) [3]. A second school of thought incorporates processing of contextual cues (e.g how dirty the cup is) into the MNS whereby identical actions driven by different intentions can be differentiated [2]. Evidence supporting the argument that contextual information influences intention-reading comes from Iacoboni et al. [3] who asked participants undergoing an fMRI scan to watch video clips of a reach-to-grasp action. The information available in the videos was manipulated with three conditions: (1) action embedded in context, (2) action without context, (3) context alone. These were nested within two further conditions such that the same action was driven by one of two intentions. Iacoboni et al. found that participants' neural activity was reliably different between the two intention conditions, and that the MNS was most active when the action was embedded in context. This suggests that intention recognition involves integrating both contextual and kinematic information.

The successful design and training of artificial internal-state-reading systems for child-robot interactions requires that a mapping between the inputs (e.g. a child's posture) and outputs (e.g. a child's internal state) is available. For this, it is important that we identify what internal-state information is available in the different data sources. This can be achieved by assessing what inferences humans are able to make from, for example, the kinematics and dynamics of a social scene (like on Fig. 1, right). One way to do this is by using point-light displays where the position and movements of an actors joints are denoted on an otherwise blank display. Studies using this method have already shown that humans are able to recognise features such as gender [1] and intention [8] from these types of stimuli. HRI researchers can use these findings to define what outputs an artificial system should be able to produce given kinematic data.

However, one key limitation of these studies is that the stimuli used are often artificially produced, e.g. by creating simulated motions in the point-light displays (e.g. [1]), or by filming actors performing the actions in an artificial setting, (e.g. [3, 8]). Whilst this allows researchers to demonstrate that internal-state information is available in kinematics, it does not provide us with insight into what humans can infer from the kinematics of real-world social interactions. Additionally, for child behaviour specifically, creating an artificial dataset may be more challenging, for example, due to variations in cognitive ability with age. Obtaining data from natural interactions is therefore potentially easier and more ecologically valid. The rest of this paper discusses a methodology aimed at identifying what internal-state information humans can glean from only the kinematic information available in a naturalistic child-child social interaction.

Figure 1: Original video clip vs. skeletal only data

Proposed Methodology

Predictions and Design: The proposed study aims to examine what information is available in the kinematics of a naturalistic child-child social interaction. To do this participants will either be shown the original or skeletal videos of real interactions (Fig. 1) and then asked questions about the videos. There will be two questioning conditions where participants are either asked only open-ended questions, or are also asked specific questions. Participants' responses following the original clips will be compared to those following the skeletal videos. Whilst we expect that participants will produce less detailed descriptions following skeletal compared to the original videos, we do expect participants to detect important features from the skeletal videos which would be useful to a robot system, such as the affective valence of the interaction, actions being performed, and the nature of the relationship between the agents.

The proposed study will have a 2 (open-ended/specific questions)$\times$2 (original/skeletal videos) design. Both conditions will be implemented between-subjects. Video presentation order will be fully random to control for ordering effects. Participants will be recruited from a crowd-sourcing platform.

Stimuli and Materials: To obtain naturalistic stimuli the proposed study will utilise videos of child-child pairs playing a game on a touch-screen table top from the PInSoRo dataset [6], made openly available by our group[1]. Short clips of child-child interactions approximately 30 seconds long will be extracted from the videos, each containing different social and interaction events (e.g. turn-taking, a disagreement). To isolate the kinematic information from contextual cues for the skeletal video condition, the OpenPose library [9] is used. It jointly detects human body, hand, and facial landmarks.

After each clip participants will be asked questions about the interaction. There will be two questioning conditions such that half of the participants are asked a single open-ended question following each video: "*Describe what you have just seen in the video*". This style of questioning reduces the risk of "leading questions", allowing us to explore what participants gain from the video without guidance. However, it is often difficult to analyse open responses and respondents may not provide enough detail to reflect their achieved level of insight on features-of-interest. To deal with these limitations half of the participants will be given the same open-ended question, then a series of specific questions which will guide respondents to discuss details of interest to the researcher in a quantifiable manner. The specific questions consist of multiple-choice and Likert scale questions such as "*What is the relationship between these characters: Friends/Neutral/Unfriendly?*" and "*Please rate how cooperative each character was: 1 = not cooperative at all, 10 = very cooperative*". Participants in the specific questioning condition will also be given a final open-ended question on each trial asking "*Did you notice anything else in the video?*".

[1] https://freeplay-sandbox.github.io

Conclusion

The proposed method aims to provide insight into what internal-state information humans are able to glean from kinematic data, with a focus on social situations. The findings of such a study have the potential to guide the design of artificial internal-state-reading systems by providing an expectation of what inferences/outputs the system should be able to draw from the data. Specifically, we plan to apply this knowledge to inform the design of an automatic classifier of social interactions. Whilst the study discussed focuses on kinematic data for internal-state reading in naturalistic interactions with children, this methodology could easily be adapted to examine the information available in a variety of data sources independently of other inputs. We argue that conducting this type of study is an important step when developing robot systems as it can help to streamline the process and provide more direct empirical support for the use of particular data types as inputs to the robot system. For example, by examining how humans recognise when a child is having difficulty with a task or activity, robot tutors could be made able to identify when assistance needs to be provided to a student during a lesson.

Acknowledgements

This work has been funded by the EU H2020 Marie Skłodowska-Curie Actions project DoRoThy (grant 657227) and the EU FP7 project DREAM project (www.dream2020.eu, grant no. 611391)

REFERENCES

1. Mather G and Murdoch L. 1994. Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. B* 258, 1353 (1994), 273–279.
2. Kilner JM, Friston KJ, and Frith CD. 2007. Predictive coding: an account of the mirror neuron system. *Cognitive processing* 8, 3 (2007), 159–166.
3. Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, and Mazziotta J C. 2005. Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3, 3 (2005), 0529–0535.
4. Baxter P, Ashurst E, Read R, Kennedy J, and Belpaeme T. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PloS one* 12, 5 (2017), e0178126.
5. Esteban PG, Baxter P, Belpaeme T, Billing E, Cai H, Cao HL, Coeckelbergh M, Costescu C, David D, De Beir A, and Fang Y. 2017. How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn, Journal of Behavioral Robotics.* 8, 1 (2017), 18–38.
6. Lemaignan S, Edmunds C, Senft E, and Belpaeme T. 2017. The Free-play Sandbox: a Methodology for the Evaluation of Social Robotics and a Dataset of Social Interactions. *arXiv preprint arXiv:1712.02421.* (2017).
7. Gallese V, Fadiga L, Fogassi L, and Rizzolatti G. 1996. Action recognition in the premotor cortex. *Brain* 119, 2 (1996), 593–609.
8. Manera V, Becchio C, Cavallo A, Sartori L, and Castiello U. 2011. Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Experimental Brain Research* 211, 3-4 (2011), 547–556.
9. Cao Z, Simon T, Wei SE, and Sheikh Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.