2021

Identifying Social Signals from Human Body Movements for Intelligent Technologies

Bartlett, Madeleine

UNIVERSITY OF
PLYMOUTH

Identifying Social Signals from Human Body Movements for Intelligent Technologies

by

MADELEINE BARTLETT

A thesis submitted to the University of Plymouth in
partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

April 2021

Copyright Statement

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment. This work has been carried out by Madeleine Bartlett under the supervision of Prof. Dr.Tony Belpaeme, Dr. Serge Thill, and Dr. Ian Howard. Parts of this work were funded by European Union FP7 projects DREAM (grant no.: 611391).

Word count for the main body of this thesis: 29,973

Parts of this thesis have been published by the author:

Bartlett, M., Edmunds, C., Belpaeme, T., Thill, S., & Lemaignan, S. (2019). What can you see? Identifying cues on internal states from the movements of natural social interactions. *Frontiers in Robotics and AI*, 6, 49.
https://doi.org/10.3389/frobt.2019.00049

Bartlett, M., Garcia, D. H., Thill, S., & Belpaeme, T. (2019). Recognizing Human Internal States: A Conceptor-Based Approach. *arXiv preprint arXiv:1909.04747*.

Bartlett, M. E., Costescu, C., Baxter, P., & Thill, S. (2020). Requirements for Robotic Interpretation of Social Signals "in the Wild": Insights from Diagnostic Criteria of Autism Spectrum Disorder. *Information*, 11(2), 81.
https://doi.org/10.3390/info11020081

Bartlett, M. E., Stewart, T. C., & Thill, S. (2021). Estimating levels of engagement for social human-robot interaction using Legendre memory units. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 362–366.
https://doi.org/10.1145/3434074.3447193

Collaborative work published but not included in this research:

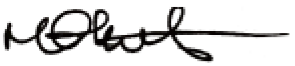
Senft, E., Lemaignan, S., Bartlett, M., Baxter, P., & Belpaeme, T. (2018a). Robots in the Classroom: Learning to Be a Good Tutor. In Proceedings of the

4th Workshop on Robots for Learning (R4L) - *Inclusive Learning, at HRI*

Cao, H. L., Esteban, P. G., Bartlett, M., Baxter, P., Belpaeme, T., Billing, E., ... & De Beir, A. (2019). Robot-enhanced therapy: development and validation of supervised autonomous robotic system for Autism Spectrum Disorders Therapy. *IEEE robotics & automation magazine*, 26(2), 49-58.
https://doi.org/10.1109/MRA.2019.2904121

Senft, E., Lemaignan, S., Baxter, P. E., Bartlett, M., & Belpaeme, T. (2019). Teaching robots social autonomy from in situ human guidance. *Science Robotics*, 4(35).
https://doi.org/10.1126/scirobotics.aat1186

Cao, H. L., Esteban, P. G., Bartlett, M., Baxter, P., Belpaeme, T., Billing, E., ... & De Beir, A. (2019). Developing and Validating a Supervised Autonomous Robotic System for Autism Spectrum Disorders Therapy.
https://doi.org/10.1109/MRA.2019.2904121

Signed:  **Date:** 18/11/2020

Acknowledgements

First and foremost, I would like to thank Serge Thill for his excellent supervisor skills. Thank you for your frank guidance, unending enthusiasm and constant support throughout this project. Thank you also to Tony Belpaeme for your encouragement and sharp eye for detail. And to Ian Howard, thank you for bringing a new perspective to the project and for your insights which helped to direct many of these studies. To all of you, thank you for giving me the courage and opportunity to explore, learn and grow in a field with which I was previously unfamiliar.

Special thanks also go to Séverin Lemaignan and Terry Stewart. To Séverin, for your immeasurable enthusiasm, creativity, and sharing of resources. A significant proportion of these experiments would not have happened without your scientific curiosity or data set. Thank you. To Terry, I want to thank you for your guidance and positivity which helped me to shape this PhD project into something I was truly excited to work on. Thank you for working with me and helping to make this project fun and interesting, even in the face of null results.

I would also like to thank Charlotte Edmunds for her help in collecting information from hundreds of papers for a meta-review, her patience as she helped me learn to code and passion for good science and even better statistics. I would like to thank Daniel Hernández for his many hours of hard work on the conceptor-based network. I also thank the team at the Bristol Robotics Laboratory for welcoming me during my visits and making my time there so fun that I just had to come back.

And finally, to all my family and friends, thank you for your unstinting support, welcome distractions, patience and feigned interest.

Abstract

Numerous Human-Computer Interaction (HCI) contexts require the identification of human internal states such as emotions, intentions, and states such as confusion and task engagement. Recognition of these states allows for artificial agents and interactive systems to provide appropriate responses to their human interaction partner. Whilst numerous solutions have been developed, many of these have been designed to classify internal states in a binary fashion, i.e. stating whether or not an internal state is present. One of the potential drawbacks of these approaches is that they provide a restricted, reductionist view of the internal states being experienced by a human user. As a result, an interactive agent which makes response decisions based on such a binary recognition system would be restricted in terms of the flexibility and appropriateness of its responses.

Thus, in many settings, internal state recognition systems would benefit from being able to recognize multiple different 'intensities' of an internal state. However, for most classical machine learning approaches, this requires that a recognition system be trained on examples from every intensity (e.g. high, medium and low intensity task engagement). Obtaining such a training data-set can be both time- and resource-intensive. This project set out to explore whether this data requirement could be reduced whilst still providing an artificial recognition system able to provide multiple classification labels. To this end, this project first identified a set of internal states that could be recognized from human behaviour information available in a pre-existing data set. These explorations revealed that states relating to task engagement could be identified, by human observers, from human movement and posture information.

A second set of studies was then dedicated to developing and testing different approaches to classifying three intensities of task engagement (high, intermediate and low) after training only on examples from the high and low task engagement data sets. The result of these studies was the development of an approach which incorporated the recently developed Legendre Memory Units, and was shown to produce an output which could be used to distinguish between all three task engagement intensities after being trained on only examples of high and low intensity task engagement. Thus this project presents the foundation work for internal state recognition systems which require less data whilst providing more classification labels.

Contents

Chapter 1

Introduction

1.1 Research Problem

The motivation behind this project was to investigate how artificial agents and systems might be made able to recognize human internal states based on observable human behaviours. Within the field of Human-Computer Interaction (HCI) there are a wide range of applications where internal state recognition is potentially beneficial. The utility of this ability is probably best illustrated by the field of Human-Robot Interaction (HRI). One of the core research problems facing HRI is that of developing autonomous systems which can interact with humans in an appropriate manner (Dautenhahn and Saunders, 2011). To perform autonomously in interactions with humans, a robot's behaviours have to 'make sense', both within the situational context and in regards to the human interaction partner's actions, behaviours and goals (Dautenhahn, 2007; Sciutti et al., 2018). In many interaction scenarios, achieving appropriate autonomous behaviour can be helped by enabling robots to recognize context-relevant human internal states. For example, when designing a robot to collaborate with a human on some multi-step construction task, it is useful if the robot is able to recognize their human partner's intentions, so that the robot can provide either complimentary or corrective behaviours (Akkaladevi et al., 2016; Palinko et al., 2016). Similarly, in more social settings, having a robot able to recognize a human's emotional state could provide the opportunity for 'empathetic' behaviours, such as sharing in a positive emotion (e.g. happiness), or pausing the interaction in response to a negative emotion (e.g. discomfort) (Cavallo et al., 2018).

Many solutions to the problem of internal state recognition have been presented. A large number of these are concerned specifically with emotional-state recognition, particularly identifying the six basic emotions (happy, sad, angry, surprise, fear, disgust) from facial expressions (Liu et al., 2017; Barros, Weber, and Wermter, 2015; Cohen et al., 2003; Bartlett et al., 2003). Liu et al.

(2017), for example, used facial expression images collected via a Kinect device to enable a robot to recognize the emotional states of happy, sad, angry, surprise, fear, disgust and neutral. This was achieved by implementing an Extreme Learning Machine classifier. Other approaches have utilized vocal cues (Hyun, Kim, and Kwak, 2006; Song, Han, and Wang, 2014) and physiological information (e.g. temperature) (Latif et al., 2015) as input to classify emotional states.

Solutions have also been developed for recognizing other human internal states. These include recognizing dominance and leadership (Beyan et al., 2016), task engagement (Rudovic et al., 2018; Sanghvi et al., 2011), social engagement (Kim et al., 2017) and experienced difficulty (Wendt et al., 2008). For example, Wendt et al. (2008) used heart rate and skin conductance as input for a classifier to recognize whether participants felt under- or over-challenged by a construction task. Despite this existing research, automated recognition of non-emotional internal states is comparatively under-researched. There is therefore a need for further exploration in this direction.

1.2 Defining 'Internal States'

The focus of this project is on the recognition of non-emotional internal states. Here a definition of what is meant by 'non-emotional internal states' (hereafter: internal states) is providing along with some examples of their importance to human-robot and human-computer interactions.

Primarily, internal states are herein defined as states which are experienced, but not considered purely emotional in nature. Whilst the six basic emotions are important for facilitating appropriate social interactions, there are other states which may be just as important in providing relevant social cues. These include states such as task engagement, boredom, friendliness, cooperation, confusion and discomfort. Many of these states fall under the definition of 'complex' emotions (i.e. any emotion that is an aggregate of two or more others (VandenBos, 2007)) which differ to basic emotions in how they are expressed and experienced. That is, their expression relies more on full body expression, than facial cues (Darwin and Prodger, 1998), and their experience is argued to involve more self-reflection than basic emotions (Lewis, 2008; Tracy, Robins, and Tangney, 2007). Other states differ from the basic emotions by being more cognitive than affective (emotional) in nature, such as task engagement, boredom and confusion, in that these states describe how someone experiences a task, event or problem (e.g. being bored by a

lecture, or confused by an instruction). Finally, states which are dependent on social contexts, such as dominance, cooperation and competition can also be considered as falling under this definition of 'internal states'.

Recognizing such states can be useful to a socially interactive agent in a range of contexts. For example, tutoring contexts where a robot or artificial tutor contributes to learning by interacting with a human participant engaging in an educational task. Here the human will experience different task-engagement states which could be useful for the tutor to recognize (e.g. bored, engaged). Similarly, in assisted-living contexts where artificial systems (e.g. smart devices such as the *Amazon Echo*) provide support to adults in the home, situations where interactions are required might include providing reminders for daily tasks. In these cases, having the artificial system able to recognize confusion (e.g. when the user feels they have forgotten to do something) would allow the system to appropriately (and autonomously) offer assistance. Alternatively, some robot applications involve a robot being situated in public areas and interacting with more than one person at a time. In such scenarios it may be useful for that robot to be able to recognize when a human is feeling distressed or when they are seeking assistance.

Given the value of recognizing these states to both social and functional interactions, the focus of this project is on exploring how such states might be made recognizable to an artificial agent or system. Moreover, this work explores how such states might be classified in such a way that more closely reflects their experience, and allows for more flexibility in responding. Consequently, this research project draws from several disciplines including Computer Science, Human-Robot Interaction (HRI) and Psychology. Thus, one of the goals of this project was to demonstrate how knowledge from Psychology can be used to inform research in Computer Science and HRI. The following sections present different psychological theories on how humans perceive and interpret the internal states of others. An overview of previous research where computational systems and robots have been designed to mimic some of these functions is also provided. The chapter concludes by highlighting the shortcomings of current techniques in dealing with a wide range of internal states, proposing an approach to overcoming these limitations and providing an outline of the rest of this Thesis.

1.3 Human Mind-Reading

First, theories of how humans are able to recognize the internal states of others are discussed. This skill is often referred to as 'mind-reading', 'theory of mind' or 'folk psychology'. A number of theories have been developed to explain this ability and many, if not all, posit that humans use observable behavioural cues as indicators of internal states (Gallese et al., 1996; Carruthers and Smith, 1996; Becchio et al., 2017).

For instance, the Simulation Theory posits that humans achieve insight into the internal states of others via an internal simulation (Shanton and Goldman, 2010). It is mainly supported by research concerning the presence of a mirror neuron system (MNS) in primates and humans (Gallese and Goldman, 1998). Mirror neurons are a type of visuomotor neuron in the brain which are active both during the performance of an action, and whilst the subject observes someone else performing that action (Rizzolatti and Craighero, 2004; Iacoboni and Dapretto, 2006). Thus, it is proposed that humans infer the mental states of others by mapping observed actions onto our own motor system, and thereby simulating a representation of the intentions and internal states driving those actions (Gallese et al., 1996). Alternatively, there is the Theory Theory, which posits that humans possess a collection of explanatory laws that relate internal states to behaviours (Gopnik and Wellman, 1994; Gopnik, 2003). This means that, when we observe an action or behaviour, we are able to apply these laws through a process of theoretical reasoning in order to identify the intentions or mental states which might be driving that action (Gopnik and Wellman, 1994; Carruthers and Smith, 1996).

Ignoring the mechanisms underlying theory of mind, both of these schools of thought propose that humans infer the internal states of others based on observable behavioural cues. This idea is also described by the "observability principle" which argues that humans are able to directly perceive the internal states of others via differences in observable actions/movements (Becchio et al., 2017). Support for this argument comes from a range of studies asking people to identify another person's internal state after isolating human motion and body postures from other cues. This is commonly achieved using point-light versions of video recordings of humans performing behaviours. Point-light videos generally consist of a series of dots representing joints and other important landmarks on the human body, presented against a blank

background. For example, Clarke et al. (2005) filmed pairs of actors performing a dialogue whilst portraying either fear, disgust or joy. They then presented participants with point-light versions of these videos and found that participants were able to identify the portrayed emotional states based solely on the movement information. Similarly, Atkinson et al. (2004) showed participants the point-light and original versions of videos of actors portraying anger, disgust, fear, happiness and sadness and asked participants to identify the emotion and rate its intensity. They found that participants viewing the point-light displays were still able to recognize the actor's emotion and the intensity of that emotion. Other studies have demonstrated that humans are able to recognize intentions (Manera et al., 2010; Manera et al., 2011) as well as emotions (Alaerts et al., 2011; Crane and Gross, 2007; Pollick et al., 2001) from just body movement information. In fact, in some cases it has been shown that body pose and movement information is more informative than other sources of information. For example, in the study conducted by Aviezer, Trope, and Todorov (2012) it was found that participants were better at identifying whether tennis players were experiencing an intense positive or intense negative emotional state from body pose information than from just facial expressions.

Ultimately what this suggests is that observable data available to artificial systems from human interaction partners (e.g. dialogue, vocal prosody, actions, facial expressions etc.) may be sufficient for recognizing human internal states. Findings from this type of research are frequently used to inform the development of artificial internal state recognition systems, and these approaches are discussed in Section 1.5.

1.4 Recognizing Internal States

Having established that internal states can potentially be recognized from observable cues, the first task in this project is to identify which observable behaviours might contain cues to the internal states with which we are concerned. This section is therefore dedicated to an exploration of what types of human behaviours might lend themselves to this task. A variety of human behaviours have been shown to be useful in allowing humans and artificial systems to recognize internal states. The below discussion focuses on some of the more widely researched behavioural modalities: facial expressions, vocal prosody and body movements and postures. It explores how both humans

and artificial systems have been shown to be able to use these data to identify the internal states of humans.

As part of this project, the question of recognizing internal states and covert behaviours from overt/observable behaviours was also explored in the context of diagnosing Autism Spectrum Disorder (see Appendix B) (Bartlett et al., 2020).

1.4.1 Facial Expressions

A rich pool of research has demonstrated that emotions can be recognized from facial expressions both by humans (Ekman and Friesen, 1971; Ekman, Friesen, and Ancoli, 1980) and artificial systems (Bartlett et al., 2003; Wimmer et al., 2008; Liu et al., 2017). However, research has also shown that other, non-emotional internal states can be identified from facial expressions. Whitehill et al. (2014), for instance, showed human raters video clips of people's faces whilst they were studying and asked them to rate how engaged these people were. Whitehill et al. (2014) found that human raters showed high levels of agreement when rating clips as showing either high or low engagement, and moderate agreement when rating the clips on a 4-point scale of engagement (none, low, moderate, high). Another study by Benedek et al. (2018) had participants view videos of humans either focusing their attention externally on a task, or internally on an imaginary task. This study found that participants were able to correctly identify whether attention was directed internally or externally based on the facial expressions of the people in the videos.

Artificial systems can also be made to recognize internal states from facial expressions. Grafsgaard et al. (2013), for instance, demonstrated that facial movements taken from videos of students interacting with tutors could be used by a classifier to accurately predict self-reported feelings of frustration and being rushed or hurried during the learning task. Similarly, Bosch et al. (2015) recorded students' facial expressions and head position whilst they completed a learning game on a computer. This data was successfully used to classify a range of internal states including boredom, confusion and engagement. Similar studies have further demonstrated that artificial systems can be trained to recognize internal states such as engagement (Hernandez et al., 2013; Thomas and Jayagopi, 2017) and frustration (McDaniel et al., 2007) from facial expression information.

1.4.2 Vocal Prosody

A second source of internal state information is vocal prosody - the intonation, stress and rhythm of speech. Humans have been shown to be able to recognize intention from prosody (Hellbernde and Sammler, 2016; Bryant and Barrett, 2007). For example, Hellbernde and Sammler (2016) showed that participants were able to recognize the intentions of criticism, doubt, naming, suggestion, warning, and wish from the prosodic features of single word and non-word utterances.

Artificial classifiers have been trained to distinguish between emotional states based on prosodic features (Litman and Forbes, 2003; Petrushin, 2000; Dai, Fell, and MacAuslan, 2008; Li and Zhao, 1998). Prosody has also been used to classify instances where a human experiences frustration during human computer interactions (Ang et al., 2002), how certain students feel during tutoring interactions (Liscombe, Hirschberg, and Venditti, 2005) and social attitude during conversation with a robot (Rosis et al., 2007).

1.4.3 Body Pose and Movement

Biological motion and posture behaviour, including gestures, walking and other movements humans make, have also been shown to communicate internal state information. This includes emotional states which can be recognized both by humans (Clarke et al., 2005; Pollick et al., 2001; Coulson, 2004) and artificial classifiers (Castellano, Villalba, and Camurri, 2007; Saha et al., 2014; Elfaramawy et al., 2017). Clarke et al. (2005) presented participants with point-light versions of videos of actors performing a dialogue whilst portraying an emotion (e.g. anger, joy and romantic love). Participants were able to recognize the emotional states anger, fear, joy, sadness, and love from these displays, suggesting that human movement alone is sufficient to recognize such states.

Outside of emotion recognition, human movements have been shown to be useful in the recognition of other internal states. For instance, in a study by Manera et al. (2011) participants were shown point-light videos of actors performing a reach-to-grasp action motivated by one of 3 socially-relevant intentions: (1) cooperation, (2) competition or (3) performing a personally-relevant action. Participants were able to identify the social intention based only on this movement information. A number of studies have also shown that socially relevant internal states and dispositions can be recognized from movement (Okada, Aran, and Gatica-Perez, 2015; Sanchez-Cortes et al., 2011;

Beyan et al., 2016; Sanghvi et al., 2011). Okada, Aran, and Gatica-Perez (2015) found that a classifier could recognize dominance and leadership based on movements participants made during group interactions. Similarly, Sanghvi et al. (2011) were able to use the postural behaviours of children to classify their engagement with a robotic game opponent.

1.5 State of the Art

Having established that a number of human behaviour modalities can be useful for recognizing internal states, the following section reviews the current state-of-the-art in internal state recognition for robots and artificial systems.

A variety of methods for classifying human internal states from observable behaviours have been developed. Due to the nature of the problem, most of these, if not all, draw in some way from Psychology in order to inform their approach or design. One particular group of methods draw on the theories surrounding the human 'Theory of Mind'. In particular, the Simulation Theory (Gallese and Goldman, 1998; Goldman et al., 2012) which posits that humans simulate observed actions of others in the motor regions of their own brain, and thus infer what intentions or internal states might drive those actions. Taking inspiration from this theory, some approaches for developing robots able to recognize the internal states of others involve using the robot's experiences of the goals which drive their own actions. For example, Kelley et al. (2008) used Hidden Markov Models (HMMs) to model five activities which were performed by a robot. The differences between these HMMs provided a mapping between observable actions and the driving intentions. Kelley et al. (2008) then demonstrated that the robot was able to correctly identify which of these five activities were being performed by a human actor, and the corresponding intentions.

Despite the success of this approach, it does face some limitations. In particular, it relies on providing a robot or artificial agent with the experience of the states it is to recognize in others. Whilst simulating action intention in artificial agents can be done simply by setting an explicit goal, simulating emotional states is a more complex task. This is largely because emotional states in humans involve an interplay between physiological responses (e.g. heart rate, hormone changes) and cognitive factors such as our appraisal of events in the environment (Moors, 2009). However, providing artificial agents with models of emotion has a number of potential benefits including

creating 'meaningful' rewards for reinforcement learning systems or providing artificial agents with mechanisms for adaptive behaviours (Cañamero, 2005). To illustrate, Hickton, Lewis, and Cañamero (2017) created a grounded affective system which utilizes 'hormone' responses to simulate fear. These 'hormones' alter the robot's state, i.e. by increasing movement speed and perceptual sensitivity in order to simulate a state of 'anxiety' and motivate different behaviours. Using this type of approach it could be possible to create a system which, having experienced this 'anxiety' state and the accompanying behaviours, could recognize this state in others by observing the associated overt behaviours (e.g. increased movement speed) and mapping them to its own experience.

Thus it is potentially possible to simulate complex/emotional internal states in artificial systems, and recognize internal states via a Simulation Theory approach. However, given that simulating such internal states is not a trivial task, a Theory Theory approach to internal state recognition is, at least currently, more straight-forward to implement. That is, rather than relying on the robot's experiences, we can imbue a robot or classifier with a set of causal laws linking observable behaviours to internal states. This approach characterises the majority of existing machine learning approaches to internal state recognition. For example, Foster, Gaschler, and Giuliani (2017) applied a rule-based classifier to the problem of having a robot recognize whether humans are experiencing an 'intention to engage' with the robot. Specifically, they used a robot-bartender scenario and designed the rules such that humans would be classified as intending to engage if they (1) stood close to the bar, and (2) turned their head towards the robot. This study found that, in an online user experiment comparing a number of methods, the rule-based method had the best overall performance in recognizing and responding to humans who intended to engage with the robot. This relatively simple approach demonstrates that artificial systems can be made to recognize human internal states from observable behaviours without relying on the ability to simulate those internal states. However, such a simplistic approach is only appropriate for more restrictive settings where a robot or classifier is only required to recognize a limited number of internal states. This is largely because each rule has to be hand coded which is not only arduous but also relies on definite knowledge of which behaviours communicate which internal states. In this particular example, the internal states being recognized are also fairly clearly communicated by human interaction partners - a person at a bar is going to actively try to attract the attention of the bartender if they

wish to be served. A great many scenarios, however, require the recognition of a wider range of internal states which are not necessarily as overtly expressed.

This has been achieved using more complex systems, often by drawing on knowledge of how humans recognize or express internal states. For instance, Daoudi et al. (2018) developed a new classification algorithm to distinguish between human reach-grasp-lift-place actions driven by different intentions. This research was motivated by findings that the movement kinematics of such actions are altered by the driving social intention (Quesque et al., 2013). The resultant classifier was able to use observable features of hand and arm movements, namely trajectories, to correctly identify whether the action was driven by a 'social' (give an object to another person) or 'personal' (keep the object for myself) intention. Thus, Daoudi et al. (2018) used evidence about how humans 'express' these motivations to inform the design of their classifier. Other research takes advantage of human 'expertise' in internal state recognition to justify the use of certain data sources for internal state recognition, and to establish a baseline against which to measure the success of a classifier. For instance, in the study conducted by Whitehill et al. (2014), 3 different classification approaches were trained to classify students' facial expressions in terms of engagement, and were compared to human raters who demonstrated a high level of agreement when rating how engaged the students were. Namely, they compared GentleBoost with Box Filter features, support vector machines (SVM) with Gabor features and multinomial logistic regression. Classifiers were given individual frames from videos of students studying to label. All three classification approaches were found to achieve a similar level of accuracy as human raters. This human expertise can be used to further 'streamline' the design process by identifying what behavioural features human raters find 'most useful' in interpreting internal states. This is illustrated by Sanghvi et al. (2011) who aimed to establish an approach to classifying children's engagement with a robotic game companion based on their body posture during the interaction. This study assessed the performance of a range of different classifiers trained with different feature sets. The features were selected based on feedback from human coders who not only rated the children's engagement levels in the videos but also provided their reasoning for their decisions, describing what aspects of the children's behaviours and postures led to their choice. By using behavioural cues which were useful to humans in internal state recognition, Sanghvi et al. (2011) streamlined their design process by identifying what features are most

likely to provide useful cues. To retain temporal information, the researchers used first, second and third derivatives of posture features, such as quantity of motion and body lean angle, over time as input for the classifiers. Sanghvi et al. (2011) found that the five 'best' classifiers, including an alternating decision tree, multi-class classifier and logistic regression, achieved accuracy scores of 79% or higher on the task of discriminating between 'engaged' and 'not engaged'.

These studies are a small sample from the rich pool of research that has dedicated itself to the automatic recognition of a variety of internal states. One of the main limitations shared by many of these approaches is that they tend to provide a restricted number of classification options. That is, many approaches are limited to simply stating whether or not an internal state is evident (Sanghvi et al., 2011; Wimmer et al., 2008; Daoudi et al., 2018). The main drawback of this approach is that it is reductionist; it can limit the amount of clarity a classifier can provide about someone's internal state. One practical repercussion of this is that it limits how flexible an interactive system can be in its responses. It should be noted that there are certain scenarios where a limited or binary approach is appropriate. For example in constrained contexts such as the bartender scenario presented in Foster, Gaschler, and Giuliani (2017). In this context the goal is to develop a robot able to recognize when someone wants to interact with them and order a drink. Thus the robot is required to make a binary decision about whether or not someone is wanting to interact, so having the robot recognize an intermediate intention (e.g. somewhat wants to interact/neutral) offers little potential benefit for the robot in terms of having it successfully perform its role as bartender.

On the other hand, there are some scenarios where being able to recognize multiple 'levels' of an internal state (e.g. not happy, somewhat happy, very happy) could enhance human-computer/human-robot interactions. For example, a tutor robot designed to recognize student confusion in a binary manner (e.g. whether or not a student is confused by a learning task) would be limited in terms of their possible responses. For instance, the robot could be made to provide an easier task when they detect that the student is confused. Whilst this could be appropriate when the student is extremely confused, it is not appropriate if they are only mildly confused and could complete the task with some additional hints or help. In contrast, a robot able to distinguish between mild and extreme confusion could provide different,

more appropriate responses to each state (e.g. providing hints when the student is mildly confused). Another context in which richer granularity could be required is in the safety systems of autonomous vehicles. In this application, whilst it might be preferable that the human 'driver' be constantly monitoring the vehicle in order to ensure it is performing as expected, it should also be acknowledged that humans are likely to perform other, secondary tasks instead of remaining vigilant. As a result, when the vehicle encounters scenarios which require that the driver take back control, it might be necessary to first alert the driver and bring their attention back to the driving task. In such scenarios, evidence has demonstrated that the amount of warning time needed for drivers to take-over control of the vehicle differs depending on whether the driver is distracted by a secondary task, and how distracted they are by that task (Mok et al., 2015; Mok et al., 2017). Thus having an automated vehicle able to recognize *how* distracted the driver is would allow it to provide appropriately timed warnings for initiating a control hand-over, thereby improving the safety of such systems. Whilst it is possible to achieve this richer granularity using categorical classification techniques, by introducing more target categories, this requires a data set which contains training examples from all of those categories. Such a data set can be difficult and time consuming to obtain given that the data must not only be collected but also labelled.

An alternative to categorical classification is regression. Regression methods, rather than providing a classification from a selection of discrete or categorical options, produce an output which is a value of a continuous variable. So, if we consider the current problem of recognizing internal states from observable human behaviours, a regression model would first require that the output variable be continuous. In the case of emotional internal states, a wealth of research indicates that many emotions can be described along a series of continuous dimensions such as valence and arousal (Fontaine et al., 2007; Mehrabian and Russell, 1974; Russell, 1980). As will be discussed in the next section, a number of other internal states might also be described or characterised by continuous dimensions. Importantly, a regression model would not require training on all of the possible values along such a continuous dimension. Instead, one need only provide enough training data to produce a model of how observable behaviours map onto, for example, valence, and the resultant model should be able to provide accurate predictions from previously untrained examples of other valence values. Some work has been done using this type of approach to predict human internal states based

on observable cues. For example, Nicolle et al. (2012) trained a regression framework to predict the four dimensions of valence, arousal, expectancy and power (which have been shown to describe the majority of emotional states (Fontaine et al., 2007)) based on head movements and facial expressions. However, whilst regressions do not require training on all potential outputs, it is generally accepted that a good representation of potential outputs is required in order to establish an accurate model (Maheswari, 2018, December 21). Consequently, this type of approach can still have substantial data requirements.

This section has highlighted the state-of-the-art of internal state recognition approaches. Whilst it is clear that there is a large variety of successful approaches, there are certainly some drawbacks, largely characterized by data requirements and/or limited classification options. This project is primarily concerned with training a system to provide multiple classification labels for a single internal state, which could potentially be used by artificial agents in order to provide more flexible and appropriate behavioural protocols. Importantly, the aim is to achieve this without requiring large amounts of data for training. The nature of a given internal state, and the way one defines these multiple labels, may lend itself to a solution which requires less training data. The following section focuses on answering the question of what representation of an internal state might be best suited to this task.

1.6 Representing Internal States

As outlined above, many existing internal state classification systems take an all-or-nothing approach whereby an internal state is classified as being either present or not (Sanghvi et al., 2011; Wimmer et al., 2008; Daoudi et al., 2018). However, if one were to use multiple classification options, including 'intermediate' states (e.g. no confusion, mild confusion, extreme confusion), one could achieve a finer-grained view of human behaviour as well as provide the opportunity for more flexible and appropriate responses from artificial agents. Achieving this, however, comes with its own difficulties. That is, classical categorical machine learning approaches require that the classifier is trained with examples from all classes. Similarly, regression approaches require training on a data set which provides a good representation of the possible outcome values. Collecting such a data set is very resource heavy, so providing a work-around is something worth trying.

Defining the problem as one of recognizing different intensities of internal states may lend itself to a solution. There is some evidence to suggest that the experience of different intensities of internal states and emotions is reflected in the intensity of their expression. For example, dominance and submission can be characterised by an energy component such that a dominant person is more energetic within an interaction than their submissive interaction partner (Burgoon, Johnson, and Koch, 1998). Furthermore, the intensity of facial expressions has been linked to the intensity of the experienced affective state (Hess, Blairy, and Kleck, 1997; Cacioppo et al., 1986). It seems reasonable, therefore, to expect that features of behaviour alter as a function of the intensity of an experienced internal state, at least in some cases. If this is the case, then it may be possible to train a classifier to recognize intermediate internal states without training. That is, assuming that the expression of an internal state varies as a function of the experienced intensity, if a classifier can be trained to recognize the extreme intensities of an internal state (e.g. no confusion vs. extreme confusion), then it may be possible to have that classifier produce an output to intermediate states which reflects the fact that these states are similar to, but also lying somewhere in-between, the two trained states. This can be achieved by using either discrete output variables or a continuous output variable.

The choice between these two types of output will often depend on the specific application for which the classifier is being designed. For example, consider the case of developing automated behaviour classification systems to augment the diagnosis of Autism Spectrum Disorder (ASD). The tools that currently exist to assist clinicians in making diagnostic decisions generally provide clinicians with a list of symptomatic behaviours which are rated in terms of a severity scale ranging from 1 to 4 (Lord et al., 2000). In this application, it may be preferable to design a classification system which mirrors existing diagnostic tools and labels behaviours in a similar, categorical way. In contrast, when developing a robot able to recognize and respond to human emotional states, given that emotional states can be described in terms of continuous dimensions of arousal and valence (Fontaine et al., 2007) it could be beneficial to have a system produce an output which translates to arousal and valence scores in order to capture a wide range of emotional states. The current project chose to focus on producing a categorical classification for two main reasons. First, most of the existing data sets are annotated in a categorical way, so producing a categorical output allows for the

outputs to be directly assessed against these ground-truth labels, without requiring that the data be re-annotated. Second, this project was initially part of the EU FP7 project DREAM[1], funded by the European Commission[2]. The goals of DREAM were to develop artificial systems and robots for use in the diagnosis of, and interventions for, ASD. For this project specifically, the focus was on developing an automated behaviour classification system which could augment the diagnostic process by providing objective quantifications of the severity of potentially diagnostic behaviours. As such it was felt that a categorical classifier would be most appropriate, as this would mirror the existing diagnostic tools. Once the DREAM project ended, the choice to produce a categorical classifier was maintained to allow comparison with the ground-truth labels. Thus the goal of this research was to produce a categorical classification system which could be used to identify multiple classes of an internal state after training only on the high and low intensity classes.

1.7 Research Question

This research project explored the following question:

How can an artificial system be made to identify human internal states in a way that requires less data, whilst providing more classification labels?

Specifically, we aimed to be able to recognize and label internal states in terms of their intensity in order to provide the opportunity for more flexible behaviours from artificial agents in human-computer interaction settings. This project was broken down into the following four questions:

1. What representation of internal states best reflects the experience of those states, and may lend itself to the problem of providing flexible and appropriate responses from artificial agents?

2. What internal states can be recognized from observable behaviours?

3. How successfully can such states be recognized by an artificial system using machine learning methods?

4. To what extent can a system recognize intermediate states after training on only the extremes?

[1] www.dream2020.eu
[2] grant number 611391

Success for this project is defined as the successful development of a system able to identify intermediate internal states from observable human behaviours after being trained only on the extremes of that state.

The first of these research questions has largely been answered by the literature review presented in this chapter. That is, the representation of internal states which best reflects their experience, and lends itself to flexible responses from artificial agents, is one where the state varies along a continuum of intensity.

1.8 The PInSoRo Dataset

All of the studies contained in this Thesis utilize the PInSoRo dataset (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017). I therefore provide details about the contents of this dataset here.

The PInSoRo data set was collected by filming children interacting either with another child, or with a Nao robot. The children were sat at an interactive touch-screen table (sand-tray) and were invited to interact and play games on the sand-tray in a free-play fashion (i.e. they were not provided

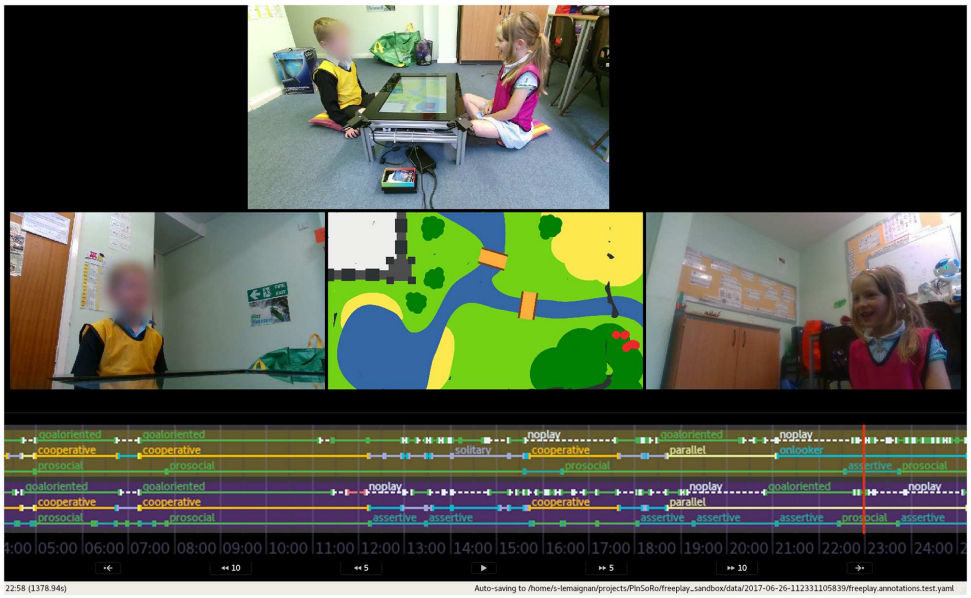


FIGURE 1.1: Screenshot of the annotation tool used to annotate the PInSoRo videos showing recordings from the two cameras recording the children's faces, the environment camera, and the recording of the sand-tray. Image taken from Lemaignan et al. (2018). Permission to reproduce this image has been granted under the Creative Commons Attribution License CC by 4.0.

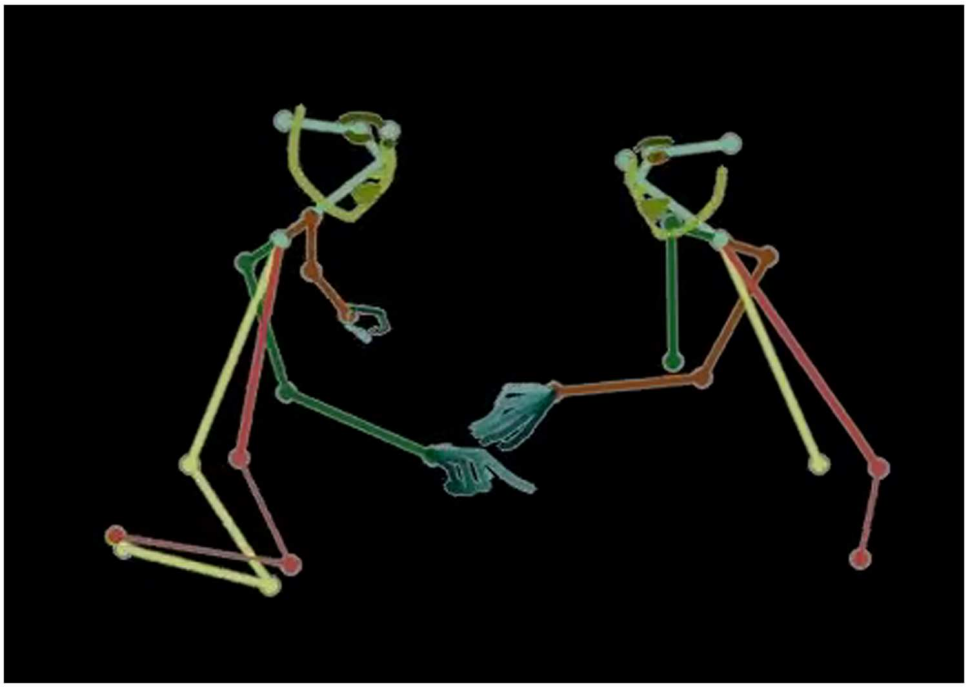


FIGURE 1.2: An image from the PInSoRo data set after post-processing using the OpenPose library to extract 2D skeletons, including facial landmarks and hand details. Image taken from Lemaignan et al. (2018). Permission to reproduce this image has been granted under the Creative Commons Attribution License CC by 4.0.

any rules or instructions by the experimenters). As can be seen in Figure 1.1, the children (or child and robot) were positioned so that they were facing each other, and several cameras were used to film different view-points of the interaction. Two cameras were attached to the table-top in order to record the faces of each child, and a third 'environment' camera was placed roughly 1.4m away from the table to provide a view of both children and the sand-tray. The children were allowed to interact for as long as they wanted (with an upper limit of 40 minutes). A total of 120 children were recorded with 30 children taking part in the child-robot condition, and 90 children in the child-child condition. As well as the videos of the interactions, the PInSoRo dataset also contains audio recordings of the interactions, and recordings of the children's (and robot's) activities on the sand-tray.

After the data was collected, the experimenters post-processed the data in a number of ways in order to generate additional data (Lemaignan et al., 2018). This included using the CMU OpenPose (Cao et al., 2017) to extract the xy coordinates of facial landmarks, action units, skeleton keypoints and gaze estimations for each child in each frame. The OpenPose library was also used to construct videos showing only the face and skeleton landmarks (see

Figure 1.2). Additionally, audio features were extracted, including prosodic, spectral and voice quality features. These data were then collated into the anonymous version of the data set along with annotations (if the video in question had been annotated). The videos were annotated with labels falling under three categories: social engagement, social attitude and task engagement. Five expert annotators were recruited for this task and (at the time of writing) roughly 75% of the data set has been annotated (Lemaignan, Edmunds, and Belpaeme, 2017).

The PInSoRo data set is openly available to researchers with the videos available on request, and the anonymous data set available for download from the data set web-page[3]. The studies reported in this Thesis utilized either the videos recorded by the environment camera (without audio) or the anonymous data set.

1.9 Thesis Contents

The remainder of this Thesis is structured as follows. Chapter 2 details the first study of this project which examined the second research question by exploring which internal states human observers were able to recognize from videos of children interacting. Participants were shown either the full visual scene or a processed version containing only movement and body posture information and were asked to provide ratings of which internal states they felt they could recognize from the videos. The results revealed that internal states relating to task engagement, such as boredom, could be recognized from both visual conditions. The published version of this study is presented in Appendix C.

Chapter 3 presents a validation study aimed at establishing whether the labels 'goal-oriented play', 'aimless play' and 'no play' available in the PInSoRo data set (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017) were reflective of 'high', 'intermediate' and 'low' task engagement states respectively. The results from this study showed that human raters did tend to rate the children in the 'goal-oriented play' videos as showing the highest task engagement, children in the 'no play' videos as showing the lowest, and children in the 'aimless play' videos as showing a level of task engagement which fell in the middle. The published version of this study is presented in Appendix D.

Chapter 4 details two classification experiments where videos of children (sampled from the PInSoRo data set) were classified in terms of engagement.

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

Two approaches were implemented, the first being a conceptor-based approach and the second being a delay network. Both approaches were trained using examples of high and low task engagement, and the delay network was then tested, not only on unseen samples from these classes, but also on the intermediate task engagement class. Results showed that, whilst performance on the trained classes was good, the methods used were not optimal for recognizing the third untrained class. The published version of the experiment using the conceptor-based approach is presented in Appendix D.

Consequently, a new approach was applied in Chapter 5. Namely, Legendre Memory Units (LMUs) were used as a pre-processing step for an MLP and a Logistic Regression. The results of this study demonstrated that, when LMU pre-processing was used, the outputs from these two systems after training on high and low engagement could be used to distinguish the intermediate engagement class without requiring training on that class. The version of this study which has been submitted for publication is presented in Appendix G.

Finally, Chapter 6 presents a summary of the works presented here and their main contributions.

The remainder of the Appendices consist of the following:

- Appendix B presents a journal paper discussing behavioural modalities and technologies which could be used when diagnosing Autism Spectrum Disorder (ASD). This discussion focuses on how measuring overt behaviours via technologies could provide insight into some of the covert behaviours associated with ASD.

- Appendix E presents a workshop paper discussing how one might represent behaviours typical of ASD in such a way that would allow a classifier to quantify those behaviours in a meaningful way for diagnostic purposes.

- Appendix F consists of a workshop paper detailing the methodology used in Chapter 2. The proposed methodology is presented as a first step to any internal state recognition research as a way of guiding the design of classification systems.

- Appendix G contains the journal paper detailing the study presented in Chapter 5 which has been presented as a poster.

- Appendix H also presents a journal paper which, at the time of submitting this Thesis, has been submitted for review. This paper is a

review of reporting practices pertaining to statistical power in papers published in the proceedings of the ACM/IEEE International Conference on Human-Robot Interaction.

1.10 Summary

In this Chapter we have discussed theories of how humans are able to recognize the internal states of others, and evidence regarding the behavioural modalities which might express these internal states. We have also discussed how this knowledge can be, and has been, used to inform the design of artificial internal state recognition systems. In particular, we have shown how the definition of internal states as varying in terms of intensity might lend itself to a novel solution to the problem of providing a non-binary identification of internal states. That is, by leveraging the assumption that the experience of internal states can be described along a continuum of intensity, one may be able to train a system to identify a range of 'intensities' without the need for training examples of every intensity.

Given our definition of 'internal states' and the evidence showing that body movements and posture are a rich source of 'non-emotional' internal state information for observing humans (Manera et al., 2011; Okada, Aran, and Gatica-Perez, 2015; Sanghvi et al., 2011) we chose to use body movement and posture, as well as some facial expression information, as the input for classification. The next step, then, was to establish what internal states could be recognized from this modality, and which states might be most readily recognized. The next Chapter describes the first study in this project, aiming to address these questions.

Chapter 2

Study 1 - What Internal State Information is Available in Human Motion?

This study was published in *Frontiers in Robotics and AI* (see Appendix C) (Bartlett et al., 2019b).

2.1 Introduction

Depending on the situation and task goals, artificial classifiers and social robotic agents can benefit from being able to recognize a range of different internal states and social dynamics. Tutor robots, for example, would benefit from being able to recognize task engagement. Assisted living systems might be improved by being able to recognize when a user is confused or distressed. A classroom robot designed to mediate child-child interaction would benefit from an ability to recognize when an interaction is becoming aggressive or hostile, or when one child is dominating an interaction. Research has demonstrated that a range of internal states such as these can be recognized by human observers from behavioural cues. These include emotions (Bartlett et al., 2003; Wimmer et al., 2008; Clarke et al., 2005; Pollick et al., 2001), intentions (Manera et al., 2011; Lewkowicz et al., 2013; Manera et al., 2010; Iacoboni et al., 2005), engagement (Sanghvi et al., 2011; Thomas and Jayagopi, 2017; Whitehill et al., 2014), confusion (Bosch et al., 2015) and pride (Tracy and Robins, 2008). Evidence has also demonstrated that interaction-dependent states or social dynamics can be recognized, such as dominance and leadership. This has been shown to be true of both humans and artificial recognition systems. For example, Beyan et al. (2016) recruited participants

in groups of four and asked them to complete a decision task. These interactions were filmed and the 3d positional data of facial landmarks was used as input to a classifier. This classifier was then able to identify participants who exhibited leadership behaviours based mainly on head pose and gaze direction information. Similarly, Sanchez-Cortes et al. (2011) had participants perform the same task as in Beyan et al. (2016) with the aim of training a classifier to recognize participants who exhibited/experienced states of dominance and competence. This study found that body movement behaviours were useful for recognizing dominance and leadership, and that head activity could be used to recognize competence.

The first concern of this project is with identifying which internal states and social dynamics might be recognizable from body movement and some facial expression information. In particular, the focus is on states which can be recognized from body movements produced in naturalistic interactions. The second concern is to be able to identify internal states which can be described in terms of intensity. Consequently, in this first study the aim was to establish a set of internal states which can be described in this way and that can be recognized from body movement, posture and facial expression information.

To this end, participants were presented with short video clips of social interactions between children. In order to examine which internal states can be seen from just the body movements of the children, some participants viewed the original video clips (full-scene condition), whilst others viewed pre-processed versions containing only movement, body posture and some facial expression information (movement-alone condition). Participants were then asked to rate the degree to which they felt certain internal states (e.g. boredom, frustration) or social dynamics (e.g. cooperation, dominance) were evident in the children's behaviours. States which can be recognized from movement information alone were then identified by comparing responses in each condition.

Hypotheses and predictions

Based on existing evidence that internal states and group dynamics can be identified from movement information (Sanchez-Cortes et al., 2011; Manera et al., 2011; Whitehill et al., 2014) the following hypotheses were proposed:

1. Participants will be able to draw internal state information from the movement-alone videos (Hypothesis 1).

2. There will be some internal states which are more readily recognized from movement-alone information than others (Hypothesis 2).

Specifically, for hypothesis 1 it is predicted that even in the movement-alone condition, the provided ratings will be sufficient to describe the internal states and social constructs identified in the observed interaction. This can be tested by training a classifier to identify clips based on the full-scene ratings, and assessing its performance when tested using the movement-alone ratings as input. Additionally, it is predicted that inter-rater agreement levels amongst participants will be above chance in both conditions (i.e. the same constructs will be robustly identified in the clips by the participants), but with higher levels of agreement in the full-scene condition. For the second hypothesis it is predicted that a classifier, when trained to identify the internal state labels assigned to each clip based on participants' ratings, will show better performance on some labels than others.

2.2 Method

2.2.1 Design and Participants

This study used a 2x1 between-subjects design comparing the effect of video type (full-scene vs. movement-alone) on ratings of how evident internal states and behaviours were in the videos. A total of 284 participants were recruited from Amazon's Mechanical Turk (MTurk) for this study. Of these, 85 participants were excluded due to providing incorrect responses to an attention check, and for completing the experiment too quickly. Demographic information regarding the remaining 199 participants is presented in Table 2.1. Participants were remunerated $1 (USD) for their participation upon completion of the experiment.

TABLE 2.1: Demographics of participants.

Condition	N	Mean Age (Range)	Gender (%M, %F)	% American	% English First Language
Movement-Alone	100	34.52 (22-70)	55%, 44%	75%	80%
Full-Scene	99	33.54 (19-72)	65%, 34%	69%	73%
Both	199	34.03 (19-72)	60%, 39%	72%	76%

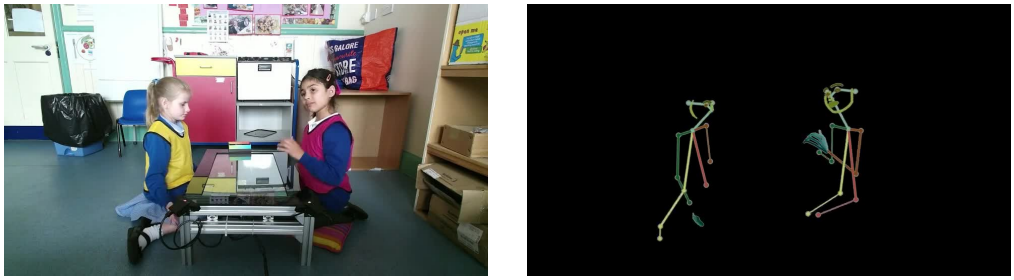


FIGURE 2.1: Captures of one of the twenty video-clips taken from the PInSoRo data set for this study. Left: version used for the *full-scene* condition showing the full visual scene. Right: version used for the *movement-alone* condition showing the 2D skeleton versions as extracted by OpenPose. Written consent for these images to be shared was obtained during collection.

2.2.2 Materials

Stimuli for this experiment were taken from the PInSoRo (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017) data set which is openly available to researchers[1]. This data set consists of videos (up to 40 minutes long) of child-child and child-robot pairs interacting whilst playing on a touch-screen table-top (sandtray). The children were allowed to engage in free-play (no defined task or goal) and were able to leave at any time. For this study only videos of child-child interactions were used. In order to provide a view of both children at the same time, videos filmed using a camera positioned roughly 1.4m away from the sandtray, with the sandtray in the centre of the camera's view, were selected. This allowed for each child to be viewed on either side of the frame (see Figure 2.1, left). From these videos, twenty 30-second clips (video only, no audio) were extracted for stimuli for this study.

The clip selection process involved two experimenters viewing full-scene versions of the videos and identifying notable 'events' or social dynamics. In particular, they were instructed to identify clips which depicted at least one of the constructs listed below. Due to the fact that no 'ground-truth' of the children's internal states was available, i.e. the children were not questioned about their experienced states during the collection of this data set, the labels used act as an estimation of what naïve observers might infer from the videos. Importantly, it should therefore be noted that neither these labels nor the inferences made by participants when responding to the questionnaire can be truly validated. The labels used were defined in terms of the children's behaviour as follows:

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

1. Boredom - at least one child was bored or not engaging with the task on the touch-screen (e.g. resting head in hand, interacting with touch-screen in slow/lazy manner)

2. Aggression - at least one child exhibited a physical aggressive action either towards the touch-screen or the other child (e.g. hitting the screen, pushing the other child's hand away)

3. Cooperation - the children were working together and/or communicating about how to perform a task (e.g. talking, joint attention (looking at the same object together), nodding)

4. Dominance - one child was bossy, performing most of the actions on the touch-screen or clearly in charge (e.g. pointing to touch-screen and talking at the other child, stopping the other child from using the touch-screen, being the only child to use the touch-screen)

5. Aimlessness - at least one child was interacting with the touch-screen in a non-goal-directed manner or without being very engaged in their task (e.g. sitting slightly away from touch-screen whilst still using it, slow/lazy movements on touch-screen, not always looking at what they're doing)

6. Fun - at least one child was having fun (e.g. laughing, smiling)

7. Excitement - at least one child behaved excitedly (e.g. more dynamic than just "having fun", hearty laughter, open smiling mouth, fast movements)

The experimenters first extracted and labelled clips independently, and then discussed their choices together in order to reach a consensus. Both children in each clip were taken into consideration such that if one child exhibited 'excitement' and the other 'boredom', both labels were applied to the clip (see Table 2.2).

The original versions of the selected clips made up the full-scene condition of this experiment. To construct the movement-alone versions, each clip was processed using the OpenPose library (Cao et al., 2017), an open source library which can be used extract the locations of joint points and other landmarks on the human body from a video feed and render them onto a black background to generate new videos (see Figure 2.1, right).

Participants providing their ratings about the children's behaviours, internal states and social dynamics via a questionnaire. The questionnaire was

designed by considering a selection of internal states and social constructs which, first, are related to the labels listed above, and second, might be desirable to have an artificial system (e.g. social robot) able to recognize. The resultant questionnaire consisted of 4 items concerning group dynamics, and 13 2-part items regarding possible internal states experienced by each child separately. In all cases, participants were asked to rate, on a 5-point scale ranging from 'Strongly Disagree' to 'Strongly Agree', how much they agreed with a statement that the children or a specific child was experiencing a given social dynamic or internal state. Each of the 13 pairs of questions were presented together such that participants were first asked about the child on the left, and then about the child on the right. Apart from this, the order of question presentation was fully randomized during the experiment (see Appendix A of Appendix C for the questions and response options).

TABLE 2.2: Labels that experimenters assigned to each clip during clip selection.

Clip	Label 1	Label 2	Label 3
01	Aggression		
02	Aggression	Excitement	Aimlessness
03	Excitement	Fun	
04	Cooperation		
05	Boredom	Aimlessness	
06	Cooperation		
07	Dominance		
08	Boredom		
09	Cooperation		
10	Cooperation	Dominance	
11	Cooperation	Dominance	
12	Aggression	Aimlessness	
13	Excited	Aggression	Aimless
14	Aggression	Fun	
15	Dominance		
16	Cooperation	Dominance	
17	Excitement	Aggression	
18	Aggression	Dominance	
19	Dominance		
20	Excitement		

2.2.3 Apparatus

The experiment script was written using the jsPsych library[2] and was remotely hosted from a private server. MTurk Workers were able to access the experiment through a link provided in an advert posted on the Amazon Mechanical-Turk website. Due to the online nature of this study, we were unable to control the physical set-up experienced by participants, nor the time and conditions under which the experiment was completed. A screenshot of how the questionnaire portion of the experiment was presented can be seen in Figure 2.2.

2.2.4 Procedure

For each video condition (full-scene and movement-alone) a separate experiment was posted. To ensure that participants only saw one condition, the experiments were posted one after the other and participants who had seen the first experiment were not given access to the second.

For both conditions the experiment proceeded as follows. Participants were first asked to provide their MTurk ID and presented with a welcome screen. This was followed by a consent form wherein participants were provided a short description of the experiment and information regarding their right to withdraw and contact details for the experimenters. Consent could be given by selecting one of two response buttons ("I do not consent" or "I do consent"). If the participant selected "I do not consent" the experiment was automatically closed and participants were returned to the MTurk advert page. If they instead selected "I do consent" they were provided with a "Continue" button which took participants to a series of four demographic questions (age, nationality, first language and gender). Following this participants were presented with the following detailed instructions:

"During this experiment you will be shown 4 30-second clips of children interacting. The children are sat either side of a touch-screen table-top on which they can play a game. Pay particular attention to the way the children interact. After each video you will be asked some questions about what you have watched."

This was presented for a minimum of 3500ms to ensure that it could not be inadvertently skipped. After the 3500ms had elapsed a "Continue" button would appear which took participants to the experimental trials.

Each participant was presented with 4 trials which each followed the same series of events. First a 30-second clip, randomly selected from the list

[2] https://www.jspsych.org/

Page 1 of 4.

How much do you agree with the following statements?

The children were competing with one another.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was sad.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was sad.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was aggressive.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was aggressive.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The children were cooperating with one another.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was excited.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was excited.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Continue

FIGURE 2.2: Screenshot of the online experimental setup (full-scene condition) showing the questionnaire, which was presented after the video clip in each trial. The image displayed at the top is a static snapshot of the clip. Written consent for the PInSoRo images to be shared was obtained during collection.

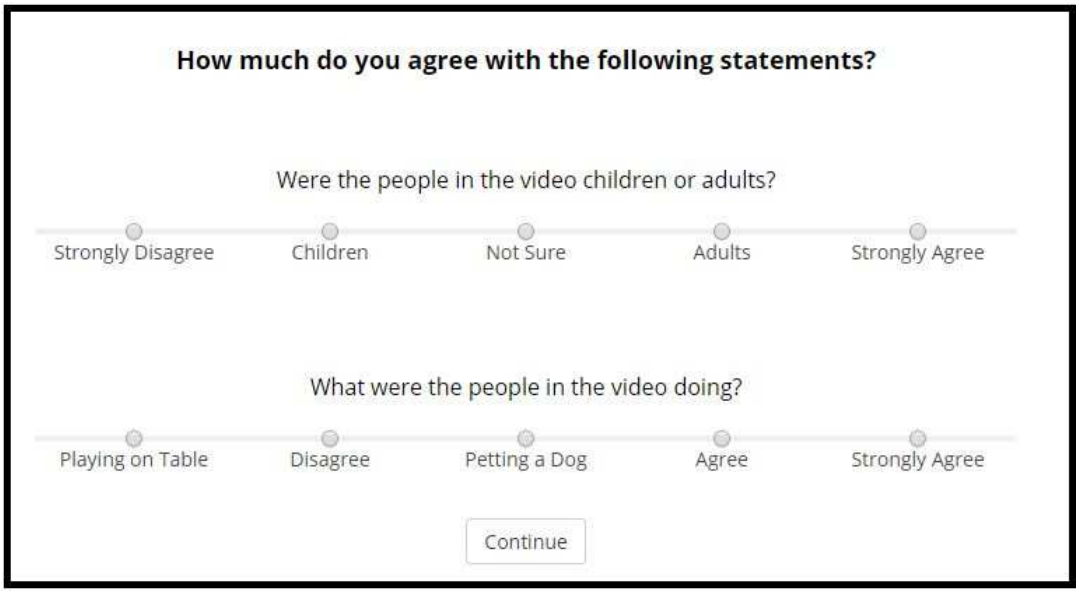


FIGURE 2.3: Screenshot of the attention check questions presented at the end of the experimental trials. These questions were presented in the same format as the main questionnaire, but with only one possible correct answer. Thus incorrect responses would indicate that a participant was not properly reading the questions. Incorrect responses to these questions resulted in the participant's data being excluded from the analysis

of 20, was presented. This was immediately followed by the 30-item questionnaire. In-between each trial participants were presented with a pre-trial screen instructed them that they must press any key in order to begin the next trial. After completing the fourth and final trial, participants were presented with an additional two questions which acted as an attention check. Responses to these questions were used to assess how attentive participants were and how diligently they had completed the experiment. They were therefore designed to be deceptive unless carefully read. That is, the questions and response options were presented in the same format as the questionnaire items, but with only two viable response options, and only one which was correct. For example, one of the questions read "Were the people in the video children or adults?" and had the response options of "Strongly Disagree", "Children", "Not Sure", "Adults" and "Strongly Agree" (see Figure 2.3). Participants who responded incorrectly were excluded from the analysis.

Once all of the experimental trials had been completed participants were shown a debrief page which thanked them for their participation, explained the purpose of the study and attention-check questions, and reiterated the

contact information for the experimenters in case of further questions or requests to withdraw from the study. Finally, participants were provided with a unique, randomly generated "survey code" and were instructed to return to the MTurk page and submit this code. Survey codes were later used by the experimenters to validate participation and authorize payment via the MTurk system. The experiment took between 20-30 minutes for each participant to complete.

2.3 Results

Data analyses were run using the Python `pandas` and `sklearn` toolkits in Jupyter Notebook. The analysis scripts can be found in the accompanying github repository (see Section 2.6 for details).

2.3.1 Inter-rater Agreement

The first step in the analysis was to examine whether participants in each condition gave similar ratings across all questions for each clip. This analysis was conducted to answer the question of whether there were any internal states or social constructs which were recognizable in both video conditions. To this end, inter-rater agreement scores were calculated across all 30 questions for each clip in each condition separately. A high agreement score for a clip would indicate that similar ratings were given, and therefore that similar states/behaviours were recognized by the participants viewing that clip.

The fact that there were unequal numbers of participants rating each clip means that Krippendorff's alpha (Hayes and Krippendorff, 2007) was the appropriate metric for inter-rater agreement. Alpha scores ranged from 0.058-0.463, i.e. from 'slight' to 'moderate' agreement according to the benchmarks provided by Landis and Koch (1977) (see Table 2.3).

A paired samples t-test was conducted to assess whether agreement scores differed across condition. This analysis showed that participants in the full-scene condition had significantly higher agreement scores ($M = 0.328$, $SD = 0.113$) than participants in the movement-alone condition ($M = 0.252$, $SD = 0.081$) (Paired Samples T-Test: $t(39) = 2.95$, $p = 0.008$, $d = 0.78$). Additionally, a t-test comparing agreement in the movement-alone condition to

TABLE 2.3: Table of inter-rater agreement scores for responses to each clip in each condition

Clip	Krippendorff's Alpha (3 d.p.)	
	Full-Scene (N)	Movement Alone (N)
1	0.446 (16)	0.186 (26)
2	0.181 (24)	0.270 (20)
3	0.393 (22)	0.369 (18)
4	0.444 (22)	0.262 (23)
5	0.328 (23)	0.283 (20)
6	0.463 (19)	0.359 (19)
7	0.091 (19)	0.236 (23)
8	0.339 (19)	0.312 (17)
9	0.097 (20)	0.058 (18)
10	0.396 (18)	0.086 (13)
11	0.280 (17)	0.234 (23)
12	0.368 (25)	0.298 (16)
13	0.334 (20)	0.189 (21)
14	0.310 (17)	0.309 (21)
15	0.422 (26)	0.242 (14)
16	0.192 (16)	0.272 (21)
17	0.273 (17)	0.183 (21)
18	0.334 (16)	0.331 (24)
19	0.415 (22)	0.304 (19)
20	0.451 (18)	0.250 (23)

chance (chance level Krippendorff's Alpha $= 0.0$) demonstrated that, despite the significantly lower agreement scores, agreement between participants in this condition was still significantly above chance (One Sample T-Test: $t(19) = 13.95$, $p =< 0.001$, $d = 3.12$).

The agreement within each condition suggests that participants in each condition did report recognizing similar states and social constructs in the clips. The greater agreement in the full-scene condition likely reflects the fact that, with full visual information there is less uncertainty about what internal states and social constructs are being observed than when the view is impoverished (i.e. just the movement information is visible). Having identified that participants within each condition did show a tendency to recognize the same internal states and social constructs in each clip, the next step is to examine whether there is any overlap in which states and constructs were

recognized in each condition.

2.3.2 Automatic labelling of internal states

Implementation of the classifiers described below was done primarily by Dr S. Lemaignan.

The following analyses examined whether there was any overlap in which internal states and social constructs were recognized by participants in each condition. This was investigated using supervised machine learning: would a classifier, when trained to label clips based on ratings from the full-scene condition, be able to label the clips equally well based on the ratings from the movement-alone condition? If so, this would suggest that the same information was reported by, and therefore recognized by and available to, participants in each video condition.

Pre-processing. The four group-dynamics ratings were excluded from this analysis. For the remaining questionnaire items, participant ratings were re-coded with values from 0 (*strongly disagree*) to 4 (*strongly agree*). Additionally, these scores were transformed so that results could be more readily interpreted in terms what behaviours and internal states characterized each clip, regardless of which child exhibited those behaviours. First, the absolute difference between scores for each child was calculated for each of the 13 constructs using the following equation:

$$diff_{construct} = abs(left_{construct} - right_{construct}) \tag{2.1}$$

This difference score is used to indicate the degree to which the children were rated as behaving in the same way, or experience the same internal state, for each construct.

The second score calculated was the sum for both children on each construct (shifted to fall in the range $[-2, +2]$):

$$sum_{construct} = (left_{construct} + right_{construct}) - 4 \tag{2.2}$$

This sum value indicates the strength of the rater's belief that a given construct was evident in the clip. These pre-processing steps resulted in 26 values for each clip: 13 difference scores and 13 sum scores.

Multi-label classification. In order to test whether participants reported recognizing the same constructs in each video condition, we used a classifier

TABLE 2.4: Classification results. Results are averaged over a 300-fold cross-validation. Values are given as percentages.

	Accuracy	Precision	Recall	F1-score
Full-scene	15.1	44.5	32.0	36.1
Chance	*3.7*	*27.3*	*14.0*	*17.4*
Movement-alone	15.8	41.6	32.7	36.3
Chance	*3.9*	*28.2*	*14.2*	*17.9*

TABLE 2.5: F1 scores for each independent label (Aggressive, Aimless, Bored, Cooperative, Dominant, Excited, Fun). Values are given as percentages.

	Agg	Aimless	Bored	Coop	Dominant	Excited	Fun
Full-scene	42.2	29.5	56.6	30.7	37.9	32.2	25.1
Chance	*18.8*	*17.3*	*11.7*	*18.2*	*20.0*	*18.6*	*11.4*
Movement Alone	43.7	19.4	58.5	29.6	43.4	31.2	27.5
Chance	*20.1*	*16.1*	*10.7*	*18.7*	*19.9*	*17.3*	*10.4*

to assess whether the ratings from each condition were sufficient for identifying the internal states or social constructs which had been used to initially label the clips. A classifier was trained in a supervised manner, using the 26 difference and sum scores as input, and the seven labels assigned during clip selection (Table 2.2) as target classification labels. Due to the fact that some clips had been assigned multiple labels, a multi-label classifier (Pieters and Wiering, 2017) was used, using 7-dimensional binary vectors wherein a zero value denoted that a label was not present in the clip, and a value of one indicated that it was.

First, four different classifiers were compared (Random Forest classifier, Extra-Tree classifier, Multi-Layer Perceptron classifier and k-Nearest Neighbour classifier), all of which were implemented using the Python `sklearn` toolkit. Hyper-parameters were optimized using a grid-search where applicable. This comparison showed that the k-Nearest Neighbour (kNN with $k = 3$) classifier provided the best overall performance and was therefore used for the following analyses.

Several metrics were calculated to assess the performance of the kNN including accuracy, precision, recall and F1 score (following the recommendations in Sorower (2010) and using the weighted implementations of the metrics available in the Python `sklearn` toolkit). Specifically, *accuracy* was

calculated as the percentage of instances where the predicted labels exactly matched with the actual labels (true positives). *Precision* was calculated as the ratio of true positives (tp) divided by the total number of predicted labels (true positives + false positives (fp)):

$$precision = \frac{tp}{tp + fp} \tag{2.3}$$

Recall was calculated as the ratio of true positives over the total number of labels that *should* have been found (true positives + false negatives (fn)):

$$recall = \frac{tp}{tp + fn} \tag{2.4}$$

The *F1 score* is the harmonic average of the precision and recall, and was calculated as:

$$F1score = \frac{2precision \cdot recall}{precision + recall} \tag{2.5}$$

Chance levels for each metric were also calculated by training the classifier with randomly generated labels (using the same distribution of labels as found in the real data set).

In the first stage of this analysis, the kNN classifier was trained with 80% of the full-scene ratings data, and tested on the remaining 20%. Second, the classifier was trained with 100% of the full-scene ratings data and tested on 100% of the movement-alone ratings. Results from these analyses are presented in tables 2.4 and 2.5. Table 2.4 shows that, whilst performance in both tests was poor to moderate (i.e. 15.8% accuracy for exact predictions of labels when tested on movement-alone ratings), performance was still markedly above chance. In fact, calculating permutation-based p-value using the procedure in Ojala and Garriga (2010) revealed that performance scores on both the full-scene testing data ($p = 0.02$) and the movement-alone testing data ($p = 0.01$) was significantly above chance.

Importantly, performance scores are very similar in each testing condition which indicates that, from the perspective of automatic data classification, the ratings data from the movement-alone condition contains roughly as much detail, and the same types of information, as the full-scene ratings. This, in turn, suggests that the movement-alone clips contain sufficient information for identifying at least some of the internal states and social constructs that can be recognized from the full-scene clips. In order to identify whether there were certain internal states or social constructs which were easier to recognize from the ratings data than others, the F1 scores for each

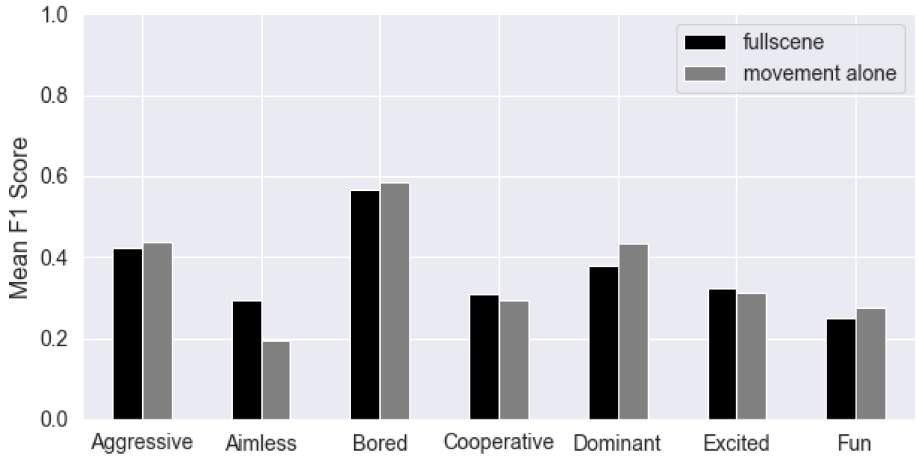


FIGURE 2.4: F1 scores describing the harmonic average of the precision and recall of the kNN for each of the 7 target labels when trained and tested with the ratings from the full-scene condition (black), and when trained with the ratings from the full-scene condition and tested with the ratings from the movement alone condition (gray).

label was calculated in each testing condition (see Table 2.5 and Figure 2.4). Based on these results there are certainly some internal state labels which appear to have been much easier for the classifier to identify than others. Namely the labels 'Bored' and 'Aggressive' out-performed all the other labels in both conditions (full-scene: 56.6% and 42.2%; movement-alone: 58.5% and 43.7% respectively). This suggests that these constructs may be as readily recognizable from the full visual scene as from an impoverished view containing only body movements and postures. In contrast, the F1-score for the label 'Aimless' dropped markedly when the movement-alone ratings were used as input (19.4%), compared to when the full-scene ratings were used (29.5%). This could indicate that aimlessness, whilst fairly well recognized when viewing the full visual scene, is much harder to identify from movement information alone. One possible explanation for this is that participants in the full-scene condition were able to use behavioural cues (e.g. eye gaze) which weren't available in the movement-alone clips. Thus it may be that there are behavioural cues which are particularly useful for recognizing aimlessness but that are unavailable in the movement-alone clips.

These analyses utilized the labels assigned by two of the experimenters during clip selection, which may not reflect the full range of internal states and social constructs which participants recognized in the clips. In order to investigate a broader range of constructs which may have been identifiable

in each video condition a factor analysis was conducted to identify latent constructs underlying the ratings data. This analysis was intended to identify more general constructs which participants may have used to understand the interactions, and which characterize a wider range of specific labels.

2.3.3 Factor Analysis

An Exploratory Factor Analysis (EFA) was conducted to identify more general constructs which describe how participants rated the videos. This analysis was motivated by the idea that, if similar latent constructs are found to underlie the ratings from each condition it would suggest that the same types of information were available to participants in each condition. Furthermore, identifying what these underlying constructs might be could provide an indication as to the 'classes' or 'types' of internal states (e.g. emotions) and social constructs (e.g. team dynamics such as dominance and leadership) that can be identified from movement information alone.

EFA. The appropriateness of an EFA was established by running a Kaiser-Meyer-Olkin (KMO) test which revealed that both the full-scene difference/sum scores (KMO = 0.88) and the movement-alone scores (KMO = 0.88) were suited for factor analysis. Additionally the Bartlett's test for sphericity was significant for both data sets (full-scene: $\chi^2 = 5219.979$, $p < 0.001$; movement-alone: $\chi^2 = 5447.747$, $p < 0.001$). These results indicate that it is appropriate to use an EFA on these data.

An EFA was carried out on the difference/sum scores from each video condition separately in order to examine what types of interaction information participants were able to draw from the full visual scene compared to the movement information alone. Specifically, the `factor_analyzer` Python module[3] was used to perform the EFA with *promax* rotation. Three factors were identified which explained 44% of the variance in the full-scene data, and 46% in the movement-alone data. Factor loadings for each of these three components in each video condition are reported in Table 2.6.

The similarity between the factors found in each condition was assessed using Pearson correlation tests. These tests revealed strong positive correlations between each pair of factors for Factor 1: $r = 0.94$, $p < 0.001$; for Factor 2: $r = 0.84$, $p < 0.001$; for Factor 3: $r = 0.81$, $p < 0.001$. These results support the hypothesis that the same latent constructs are evident in the ratings from

[3] https://github.com/EducationalTestingService/factor_analyzer

TABLE 2.6: Factor loadings for the three-factor solution using EFA, with factor loadings > 0.35.

	Factor 1: imbalance		**Factor 2: valence**		**Factor 3: engagement**	
	full-scene	*mov.-alone*	*full-scene*	*mov.-alone*	*full-scene*	*mov.-alone*
Diff Sad	0.41	0.52				
Sum Sad			0.72	0.53		0.49
Diff Happy	0.49	0.53				
Sum Happy				-0.51	-0.55	
Diff Angry	0.40	0.62				
Sum Angry			0.81	0.85		
Diff Excited	0.53	0.63				
Sum Excited					-0.71	
Diff Calm	0.45	0.63				
Sum Calm				-0.45		
Diff Friendly	0.69	0.56				
Sum Friendly				-0.60	-0.43	
Diff Aggressive	0.78	0.79				
Sum Aggressive			0.80	0.72	-0.36	
Diff Engaged		0.39			0.65	0.52
Sum Engaged					-0.64	-0.64
Diff Distracted					0.65	0.63
Sum Distracted			0.63			0.82
Diff Bored		0.44			0.61	0.54
Sum Bored			0.58		0.48	0.83
Diff Frustrated	0.53	0.61				
Sum Frustrated			0.70	0.69		
Diff Dominant	0.75	0.81				
Sum Dominant			0.53	0.52		
Diff Submissive	0.68	0.72				
Sum Submissive			0.54			

each condition. Thus it appears likely that participants in each condition relied upon the same general constructs when rating the clips.

Taking a closer look at the distribution of factor loadings allows us to interpret each latent construct. The first factor consists largely of difference scores for the emotion items as well as the team-work related items (dominant, submissive) and thus appears to describe how different the children's behaviours and internal states were during the interaction. This factor has therefore been labelled as *imbalance* as it seems to mostly describe the degree to which children were rated as exhibiting the same behaviours and internal states. For example, a high score on this factor would indicate that the children were rated as exhibiting very different states and behaviours, e.g. one child was rated as very happy, and the other as not happy at all.

TABLE 2.7: Classification results, including classification in EFA-space. Scores from the classification of clip labels copied from Table 2.4 for comparison. Values are given as percentages.

	Accuracy	**Precision**	**Recall**	**F1-score**
Full-scene, EFA	11.2	38.3	26.2	30.0
Full-scene, Labels	15.1	44.5	32.0	36.1
Chance	*3.8*	*28.1*	*14.2*	*17.8*
Movement-alone, EFA	11.7	35.1	27.0	30.3
Movement-alone, Labels	15.7	41.6	32.7	36.3
Chance	*3.9*	*28.3*	*14.2*	*17.9*

The second factor has positive correlations mostly with the sum items for negative emotions and behaviours (e.g. angry, sad and aggressive) in both conditions. Additionally, in the movement-alone condition, this factor also has strong negative correlations with the sum scores for positive items (e.g. happy, calm and friendly). It can, therefore, be interpreted as the *valence* of the interaction. To illustrate, a high score on this factor could indicate an interaction where both children were rated as being very sad or aggressive. Alternatively, in the case of an interaction which scored highly on the imbalance factor, a high score on the valence factor could indicate that one child was much more sad/angry than the other child was happy/friendly.

Finally, the third factor shows correlations mostly with items related to *task engagement*. Specifically, this factor has a strong negative correlation with *Sum Engaged*, and a strong positive correlation with *Sum Distracted* such that a high, positive value on this factor would indicate that, overall, the children were not very engaged with their task. At the same time, this factor is positively correlated with items related to the difference items; *Diff Engaged*, *Diff Distracted* and *Diff Bored*. Thus this factor also contains information about the degree to which the children exhibited the same task engagement behaviours. Consequently, a high positive value on this factor would indicate that, whilst overall the children were mostly rated as being distracted or bored, there was also a big difference between the children. For example, such a score could indicate that one child was extremely bored/distracted, whilst the other child was somewhat engaged in the play task.

Social expressiveness of the EFA-space embedding. As a final step in this analysis, the same classification methodology as described in Section 2.3.2 was applied to the EFA embedding of participants' ratings. This was done to examine whether these three factors, by themselves, would allow for

TABLE 2.8: F1 scores for each independent label, including after classification in the EFA-space. Scores from the classification of clip labels copied from Table 2.5 for comparison. Values are given as percentages.

	Agg	Aimless	Bored	Coop	Dominant	Excited	Fun
Fullscene, EFA	37.8	16.2	53.9	29.4	29.7	25.9	20.6
Fullscene, Labels	42.2	29.5	56.6	30.7	37.9	32.2	25.1
Chance	*19.1*	*16.5*	*11.7*	*19.0*	*19.6*	*17.4*	*11.0*
Movement alone, EFA	36.5	24.0	49.2	24.6	33.7	27.4	12.2
Movement alone, Labels	43.7	19.4	58.5	29.6	43.4	31.2	27.5
Chance	*19.8*	*16.4*	*10.7*	*18.9*	*19.9*	*17.9*	*10.5*

an effective and meaningful assessment of the ratings in order to describe the social interactions. To this end, the 26-dimensional ratings (difference and sum scores) were projected onto the 3-dimensional EFA space according to the following equations:

$$M^{EFA}_{fullscene} = M_{fullscene} \cdot \Lambda^{EFA}_{fullscene} \tag{2.6}$$

$$M^{EFA}_{movementalone} = M_{movementalone} \cdot \Lambda^{EFA}_{fullscene} \tag{2.7}$$

where $M_{fullscene}$ is the 396×26 matrix of participants' ratings, $M^{EFA}_{fullscene}$ is the 396×3 matrix of participants' ratings projected onto the EFA space, and $\Lambda^{EFA}_{fullscene}$ is the 26×3 matrix of the EFA factor loadings (Table 2.6). Both the full-scene and movement-alone clips were projected onto the same full-scene EFA space (i.e. the space constructed using the EFA factors generated from the full-scene ratings data).

A kNN classifier ($k = 3$) was then trained, following the same procedure as before, to predict each clip's position in the full-scene EFA space (i.e. its scores on each factor) based on the difference/sum ratings data. That is, the kNN classifier was first trained on 80% of the full-scene ratings and tested on the remaining 20%, and then trained on 100% of the full-scene ratings, and testing on all of the movement-alone ratings. Tables 2.7 and 2.8 show the results of this classification test. Whilst we do observe a drop of about 4-6% in performance, all of the performance scores are still above chance.

2.4 Discussion

This study set out to identify a set of internal states or social dynamics which could be identified from body posture and movement information by human observers. To this end, participants viewed clips of child-child pairs interacting in a free-play setting. Clips were selected based on whether at least one of the seven labels 'Aggression', 'Aimlessness', 'Boredom', 'Cooperation', 'Excitement' and 'Fun' could be used to describe the behaviour and perceived internal states of one or both of the children in that clip. Participants either viewed the full visual scene, or a pre-processed version showing only the body movements and postures of the children and were asked to rate how much they felt each child demonstrated experiencing a series of 17 constructs. The full-scene condition was used to approximate a 'ground-truth' of which internal states and social constructs could be interpreted/recognized from each video clip. Thus, the ratings from this condition could be compared to those from the movement-alone condition in order to establish whether the same states and constructs could be recognized from just the children's movements and postures.

This was done primarily by training a 3-kNN classifier to label the clips according to the seven original labels. The classifier was trained using the full-scene ratings and tested on the ratings from the movement-alone condition. Similar levels of performance were achieved by the classifier when tested with the ratings from each condition, suggesting that the movement and posture information was interpreted by participants in a similar way as the information from the full-visual scene. Combined with the inter-rater agreement scores, these results support the first hypothesis: participants will be able to draw internal state information from the movement-alone videos. That is, these results demonstrate that there was some similarity in how participants rated the clips in each condition, and thus that there are at least some internal states and social constructs which can be recognized from human movements and body postures with a similar degree of accuracy as from the full visual scene.

After establishing that movement and body posture information is sufficient for recognizing human internal states, this study also examined whether there were certain states which are more readily recognizable from these data than others. By calculating the F1-scores for each of the seven original labels, this study identified that the labels 'Boredom', 'Aggression' and 'Dominance' were most readily recognized regardless of video condition. In contrast, the

label 'Aimlessness' was much less successfully classified based on ratings from the movement-alone condition compared to when the full-scene ratings were used. These results suggest that endeavouring to train a classifier to recognize states such as Boredom or Aggression based on raw movement and posture data will likely be more successful than training a classifier to recognize Aimlessness. These results provide support for hypothesis 2 which posits that there will be some internal states which are more readily recognized from movement information than others. Additionally, the results of the EFA analysis suggest that the constructs of Imbalance, Valence and Engagement can be used to describe social interactions and can also be recognized from just movement and posture information.

2.4.1 Limitations

A number of potential limitations are associated with this work. The first to highlight is that the accuracy of the classifier, whilst above chance, was still relatively low. This may reflect the fact that the task of rating internal states from visual information is inherently difficult, and therefore the ratings used as input for the classifier may not have been the most optimal source of information. Additionally, the participants' ratings were likely a more noisy source of data than the video data, especially considering that there were multiple sets of ratings for each clip, which differed from each other in various ways. Despite this, the goal of this study was not to train a classifier to recognize internal states, but to identify which internal states could be most readily recognized by human observers from movement information alone. Thus the low accuracy of the kNN classifier is not overly concerning.

Second, participants did not have access to contextual information such as what game the children were playing, the state of the game and the pre-existing relationships between the children. The lack of such contextual cues would have made the task of rating the children's behaviour more challenging, and thus the ratings may not be as reliable or accurate as they could be. This limitation is particularly important to consider given that this study was motivated by the idea of creating artificial systems able to recognize human internal states. In most, if not all, applications of such systems or robots, it would be possible to provide an artificial system with at least some of these contextual details and have it factor them into its classification decision.

A third limitation which may have impacted the accuracy of the classifier, and the quality of responses, is that the questionnaire might not have been

optimal for this task. The questionnaire used was hand-crafted based on assumptions of what states were present in the interactions, and which might be useful for a social robot to be able to identify. It is therefore possible that it was not ideal for capturing a complete view of what participants were able to recognize from the clips. Consequently, future work would benefit from the development of a better, validated questionnaire for this type of research.

2.5 Conclusion

The results of this study demonstrate that it would be reasonable to expect a machine-learning algorithm to recognize certain human internal states and social constructs from human body movements and postures. Importantly, this study highlights states such as aggression and boredom, as well as the constructs of Imbalance, Valence and Engagement as likely to be more readily recognized from such data than others (e.g. aimlessness). In highlighting these states as more readily recognized, this study provides an answer to the second research question of this project: what internal states can be recognized from observable behaviours? Consequently, this study establishes a 'jumping off point' to guide the rest of the studies in this project, particularly decisions concerning which internal states to attempt to classify, and what types of data to use as input. Specifically, based on the EFA results and the finding that "Boredom" was most readily classified from participants' ratings, the remaining studies focus on the classification of task engagement from human movement and body posture information.

2.6 Open-Source Resources

The following github repositories contain scripts for the experiment and analysis.
https://github.com/maddybartlett/Thesis_Notebooks/tree/master/Chapter2_WhatCanYouSee,
https://github.com/maddybartlett/pinsoro-kinematics-study.

Chapter 3

Study 2 - Data-Set Validation

Parts of this study were presented and published as part of a workshop at the 2019 ACM/IEEE International Conference on Human Robot Interaction (see Appendix D) (Bartlett et al., 2019a).

3.1 Introduction

Based on the findings of the first study (Chapter 2; Bartlett et al., 2019b) that states relating to task engagement (e.g. boredom) are recognizable to humans from movement and posture information, the rest of this project focuses on classifying task engagement from observable human behaviours. Considering that the goal is to establish a method for classifying multiple levels of intensity of task engagement, it is necessary to establish a data set which contains examples of humans experiencing such states. We chose to continue using the PInSoRo data set, specifically the videos of child-child interactions. The videos had been annotated for a range of behaviours including whether each child was engaged in 'goal oriented', 'aimless' or 'no' play. The study reported in this chapter was designed to assess the assumption that these labels reflect 'high', 'intermediate' and 'low' levels of task engagement respectively. In order to examine this participants were presented with both the full-scene and movement alone versions of video clips which had been annotated with each label, and asked to rate one of the children's level of task engagement along a Likert scale.

Hypotheses and Predictions

Based on the assumption that the play labels reflect levels of task engagement, the following hypothesis was made:

1. Participants will rate children's engagement differently, depending on whether they were originally annotated as partaking in goal-oriented, aimless or no play.

It was predicted that participants' ratings of children's engagement would be highest for goal-oriented clips, lowest for no play clips, and that aimless clips will be rated lower than goal-oriented and higher than no play.

3.2 Method

3.2.1 Participants and Design

This study had a 2 (full-scene vs. movement-alone) x 3 (clip-type/annotation) design. Five participants (students and employees) were recruited from the University of Plymouth's School of Computing, Electronics and Mathematics on a volunteer basis. Demographic information was not collected. All participants took part in all six conditions.

3.2.2 Materials

For stimuli, forty-five video clips were extracted from the PInSoRo (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017) data set for this study. Clips were extracted based on the annotations for the 'purple child', positioned on the left of the frame in the video clips. A total of 15 'goal-oriented', 15 'aimless' and 15 'no play' clips were extracted. Selection was made semi-randomly whilst ensuring that the clips were of a reasonable length and that there were no anomalies within the clips (i.e. a third-party entering the frame). Clip lengths ranged from 12-30 seconds. After clips were selected, the movement-alone versions were constructed using OpenPose (Cao et al., 2017) as described in Chapter 2.

3.2.3 Apparatus

The experiment was written using the JSPsych library. For each participant, two separate experiment scripts were written, one for the full-scene clips, and one for the movement-alone clips. Clips were divided across experiment scripts such that each participant saw 9 examples of each clip-type, and each clip was rated by at least 3 participants. Each participant saw the same clips in each video condition.

The experiment was presented on a desktop computer. Participants were positioned a comfortable distance away from the screen where they could still reach the keyboard and mouse to provide responses. Only the experimenter was in the room with each participant during the experiment, positioned so that they were out of sight to the participant.

3.2.4 Procedure

For each participant the experiment was split across two days. Participants watched the full-scene clips on the first day and were then asked to return the next day when they would watch the movement-alone clips. Participants all received the following instructions before beginning the experiment:

You're about to watch several videos of children interacting with a touch-screen sand-tray. The children were able to either play a specific game on the sand-tray, or to do whatever they want. After each clip you will be asked to judge the child's level of task engagement.

Participants were then given the opportunity to ask the experimenter questions about what they would be doing and were instructed about their right to withdraw before beginning the experiment.

At the beginning of the experiment, the instructions were reiterated and participants were asked to provide consent. The consent form was presented within the experiment script and participants were given two options at the end of the form. If participants selected the "I consent" option, the experiment proceeded as normal. If participants selected "I do not consent" the experiment was terminated. Participants then viewed nine of each type of clip (a total of twenty-seven clips) presented in a random order. Following each clip, participants were presented with the question "*How engaged was the child with their task on the touch screen table-top?*". Participants rated the children's engagement using a 7-point Likert scale ranging from 1 = "Not at all Engaged" to 7 = "Highly Engaged". Once they submitted their rating they would continue on to the next clip.

At the end of the experiment on the first day, participants were given the opportunity to ask any questions they may have and were asked to return the next day to complete the second half. On the second day, the experiment proceeded in the same way except participants were shown the movement-alone clips instead of the full-scene clips. At the end of the second session participants were fully debriefed on the nature and purpose of the study and

were thanked for their participation. Each session took approximately 10-15 minutes to complete.

3.3 Results

All analyses were run in R Studio. The analysis scripts can be found in the accompanying github repository (see Section 3.5). The data were analysed in two main ways; examining inter-rater agreement, and comparing actual ratings.

3.3.1 Inter-Rater Agreement

Inter-rater agreement was examined in 2 different ways by calculating Krippendorff's alpha. The first analysis explored whether participants had provided similar responses for each clip-type regardless of video condition. This was done by calculating Krippendorff's alpha across all responses to each of the 3 clip-types. The alpha scores have been interpreted in terms of the benchmarks outlined by Landis and Koch (1977). Responses showed "fair" agreement for the goal-oriented play clips (Krippendorff's alpha = 0.269) and the no-play clips (Krippendorff's alpha = 0.267). Responses for aimless play clips showed "slight" agreement (Krippendorff's alpha = 0.171).

Agreement across each clip-type when viewing the full-scene clips compared to the movement-alone clips was then assessed. The results of this analysis can be seen in Table 3.1. Whilst agreement for the goal-oriented and no play clips remained fairly stable across video condition, the ratings for the aimless play clips show a marked drop in agreement in the movement alone condition (from 0.247 in the full-scene condition, to -0.022 in the movement-alone condition). One possible reason for this may be that participants relied more on cues available only in the full-visual scene for recognizing intermediate engagement. These could include spatial cues such as the child's position relative to the sand-tray, or facial expressions/gaze behaviours which are more difficult to interpret from the 2D skeleton figures. Additionally, it is likely that the childrens' behaviours when they exhibited goal-oriented and no play were more distinctive. That is, when the children were goal-oriented they were likely more focused on the sand-tray (looking down) and more expressive in their movements given that they were playing. In contrast, 'no play' behaviours likely involved less attention to the sand-tray (looking away) and less expressive behaviours as the child was not

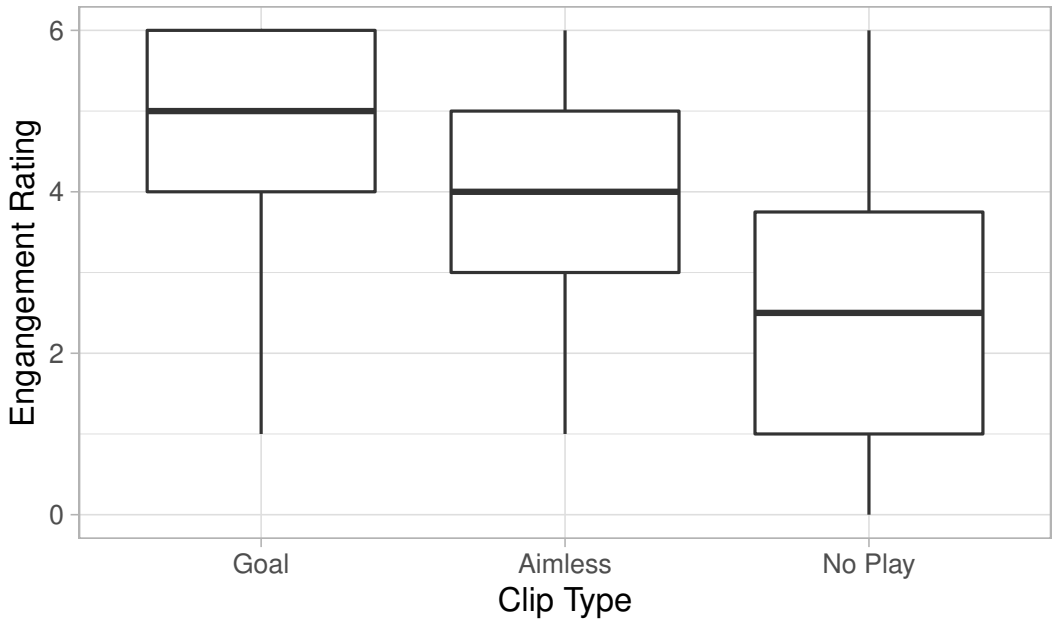


FIGURE 3.1: Boxplot of Engagement ratings for Goal-Oriented, Aimless and No Play clips. Ratings from both conditions were included in this plot resulting in at least 6 ratings for each clip, and at least 90 ratings for each clip-type.

TABLE 3.1: Table of inter-rater agreement scores for responses to each clip-type in each condition

Clip Type	Krippendorff's Alpha (3 d.p.)	
	Full Scene	Movement Alone
Goal Oriented Play	0.382 (fair)	0.368 (fair)
Aimless Play	0.247 (fair)	-0.022 (poor)
No Play	0.126 (slight)	0.202 (fair)

TABLE 3.2: Table of results for post hoc Tukey's Honest Significant Difference test.

Comparison	Difference	Significance (p adj)
Goal Oriented − Aimless	0.644	$p = 0.008$
Goal Oriented − No Play	2.378	$p < 0.001$
Aimless − No Play	1.733	$p < 0.001$

playing (sitting still). On the other hand, aimless play likely involved some attention to the sand-tray, and some playful behaviours but that were less energetic and expressive than goal-oriented play behaviours. Thus distinguishing between the extremes is much easier, whilst the amount of overlap between the extremes and the aimless behaviours potentially made this class harder for participants to label.

3.3.2 Ratings

The second set of analyses looked at the how participants rated each type of video. Overall mean engagement rating for goal-oriented videos was 4.81 ($SD = 1.25$), for aimless videos was 4.16 ($SD = 1.52$), and for no-play videos was 2.43 ($SD = 1.54$) (see Figure 3.1). An ANOVA revealed a significant main effect of clip-type on engagement ratings ($F(2, 267) = 64.99$, $p < 0.001$, $\eta_p^2 = .329$). Importantly, a *post hoc* Tukey test revealed significant differences between all conditions (Tukey's HSD: all differences >0.6, all p's <0.009; see Table 3.2).

3.4 Discussion & Conclusion

This study aimed to validate the assumption that the annotation labels regarding play style in the PInSoRo data set are analogous to different levels of task engagement. Participants viewed clips of child-child interactions where the left-hand child had been annotated as demonstrating goal-oriented, aimless or no play behaviour. They were asked to rate these clips in terms of how engaged they felt the child was with their play task. It was predicted that ratings of children's engagement would be highest for goal-oriented clips, lowest for no play clips, and that ratings for aimless clips would fall somewhere in-between.

It should be noted that there are a number of limitations with this study which prevent us from drawing strong conclusions. First, the order of condition was not counterbalanced; all participants saw the full-scene clips on day one, and the movement-alone clips on day two. Thus these results may have been influenced by order effects. Second, this study used a very small number of participants and no power analysis was conducted. This is largely because this study was only intended to assess whether the assumption that the task-engagement labels in the PInSoRo data-set could be translated in

this way. As such, this study was intended more as an exploration of semantics in order to provide support for interpreting the existing labels as high, intermediate and low engagement.

Additionally, the low agreement in ratings given for aimless play clips in the movement-alone condition are potentially concerning. It may suggest that there is not sufficient information in the movement-alone videos for recognizing intermediate task engagement. Alternatively, as discussed, it may be that distinguishing between the extreme behaviours and aimlessness was more difficult due to the overlapping behaviours (e.g. aimless and goal-oriented could both involve attention on the sand-tray, and aimless and no play could both involve less movements). Whilst it is possible that this might make the task of distinguishing between these behaviours more difficult for a classifier, it is also possible that a computational system will be able to identify and utilize objective, quantifiable differences that human observers do not.

Ultimately, the results do show support for the argument that the existing play labels reflect three intensities of task engagement; participants rated the clips such that goal-oriented clips showed the highest engagement scores, no-play clips showed the lowest, and aimless clips fell in-between these two extremes. The remaining studies in this project, therefore, continue to utilize the PInSoRo data set, exclusively using clips of the different play behaviours, which we henceforth refer to as task engagement.

3.5 Open-Source Resources

The following github repository contains scripts for the experiment and analysis as well as the final, anonymous data set.
https://github.com/maddybartlett/Thesis_Notebooks/tree/master/Chapter3_ValidatingDataset

Chapter 4

Study 3 - Classifying Internal States from Observable Behaviour

Parts of the work reported in this chapter were presented and published as part of a workshop at the 2019 ACM/IEEE International Conference on Human Robot Interaction (see Appendix D) (Bartlett et al., 2019a).

4.1 Introduction

The exploration of psychological studies on the experience of internal states provided in Chapter 1 has provided one possible answer to the first research question: *what representation of internal states best reflects the experience of those states, and may lend itself to the problem of providing flexible and appropriate responses from artificial agents?*. That is, representing internal states as varying along a continuum of intensity allows one to select multiple intensity 'levels' for an agent to respond to, thus providing the opportunity for more flexibility in an artificial agent's behavioural protocols. The first study in this project (Chapter 2) has also provided an answer to the second research question by identifying a selection of internal states which can be recognized by humans from movement and body posture information. Specifically, the state of boredom and the construct of task engagement were identified as readily recognizable by humans (Bartlett et al., 2019b). The next step in this project is to explore the third research question:

"How successfully can such states be recognized by an artificial system using machine learning methods?"

The review of the state-of-the-art in Chapter 1 revealed that a wide variety of machine-learning methods have been applied to the task of recognizing internal states. In contrast to existing categorical approaches which require that each class be trained, the final goal of this project is to be able to train a

system on the extremes of a state (high and low intensity) and then to have it generalize to estimate intermediate intensity states without training. The following studies therefore compare two types of approach, one which allows for interpolation between the trained extremes, and one which produces a continuous output (i.e. values ranging between -1 and +1) which can then be categorized into 'high', 'low' or 'intermediate'.

Additionally, given the promising results of the study presented in Chapter 2 (Bartlett et al., 2019b), the following studies focus on using body movement and posture information as input for recognizing internal states. Given that both internal states and body movements/postures are dynamic and change over time, and that it is therefore reasonable to expect that the unfolding of human movements over time carries more information than individual moments in time, only those techniques which are able to deal with temporal data are considered. Within the field of machine-learning a variety of approaches to classifying temporal data have been developed. These can be separated into two main classes, the first of which uses single values to describe a data signal over time, i.e. an 'average over time' value. For example, Sanghvi et al. (2011) used first, second and third derivatives of posture features over the course of each clip as input for their classifiers. Other methods involve having a classifier label individual frames (Whitehill et al., 2014; Bartlett et al., 2003). Bartlett et al. (2003) applied an SVM to the problem of classifying emotions from video sequences of people's faces. Their approach was to use seven binary classifiers (one for each classification label) and have each one classify every frame of a video sequence. This was done by giving each frame a score from 0-1 where 0 = emotion not present and 1 = emotion present. The classification decision was then made by selecting the classifier with the greatest overall score across all the frames contained in the video sequence.

Alternatively, a second set of approaches utilize some form of memory to deal with dynamic, temporal data. A number of these fall under the class of recurrent neural networks (RNNs). RNNs retain a memory of previous inputs through their internal hidden state (Poznyak, Oria, and Poznyak, 2018, Chapter 3). RNNs have been successfully applied to the recognition of human behaviour in a variety of contexts and for numerous purposes. For example, Tian, Moore, and Lai (2015) successfully applied a Long Short-Term Memory (LSTM) RNN to the problem of recognizing emotion in terms of arousal and valence from vocal cues. Similarly, Echo State Networks have been applied to emotion recognition from speech signals (Trentin, Scherer,

and Schwenker, 2015; Scherer et al., 2008). RNNs have also been successfully applied to predicting human intentions (Yan et al., 2019), recognizing emotion from videos of facial expressions (Zhang et al., 2018) and detecting conversational engagement from video and audio information (Lala et al., 2017).

Thus, methods such as RNNs which utilize 'memory', have proven successful at processing temporal data for classification of internal states. However, few of these approaches have been applied to the problem of estimating an untrained 'intermediate' state after training on two extremes along a continuum. One approach, however, which shows particular promise for this task is conceptors (Jaeger, 2014a) which have been shown to be useful for generating novel 'intermediate' dynamic patterns based on trained patterns (Jaeger, 2014b).

4.2 Approach 1 - Conceptors

Conceptors are neuro-computational mechanisms that can be used to characterize the state of a Recurrent Neural Network when it is driven by dynamical patterns (Jaeger, 2014a). For example, suppose an RNN is driven with the patterns (p) belonging to class P. Each pattern p_i will result in a different 'state' describing the network's activity. The states associated with a class of patterns will occupy a particular linear sub-space of the network's state space. Conceptors can be used to encode these sub-spaces and thus as maps or 'neural filters' associated with trained patterns (see Figure 4.1).

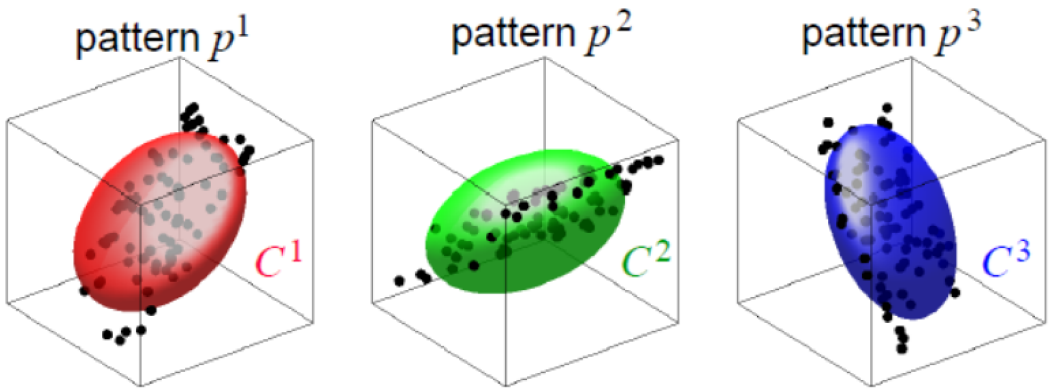


FIGURE 4.1: Illustration of the 'state' of an RNN when driven by three different patterns (p^1, p^2, p^3; black dots). These state clouds can then be characterized by Conceptors (C^1, C^2, C^3; colored ellipsoids). Image taken from Jaeger (2014a). Permission to reproduce this diagram has been granted by Prof. H. Jaeger.

Conceptors have been successfully applied to classification problems. For example, in Jaeger (2014b) a conceptor-based RNN is applied to the problem of identifying the speaker from voice recordings (i.e. the Japanese Vowel recognition task). In this study, after conceptors for each speaker were trained, testing involved feeding new patterns into the RNN and calculating the *positive* and *negative evidence* scores for each conceptor. These scores indicated how well the activity generated by a new pattern fits into the subspace characterized by each trained conceptor. A combined evidence score was then calculated, and the conceptor with the greatest combined score was chosen as the classification decision. The first study reported in this chapter applies this same approach to the problem of classifying levels of engagement based on human movement and posture information.

For this project the main motivation behind using conceptors was the fact that, once trained, conceptors can be combined together in order to create new conceptors which can be thought of as a state lying in-between the trained conceptors (Jaeger, 2014b). This has mainly been demonstrated in the use-case of pattern generation for the purpose of smoothing transitions between two patterns. For example, Jaeger (2017) report a study where a reservoir network was fed 15 human motion patterns including 'slow walk', 'fast walk', 'jog' and 'run'. Conceptors were trained for each of these patterns and were then fed back into the reservoir in order to generate the associated human motion pattern. New conceptors were generated by morphing two conceptors together, e.g. 'slow walk' and 'fast walk'. Smoother transitions between 'slow walk' and 'fast walk' could then be achieved by using first the 'slow walk' conceptor followed by the morphed conceptors to control the reservoir before finally feeding in the 'fast walk' conceptor. This morphing capability, if applied to classification rather than generation, is a promising solution to the problem of classifying intermediate states without requiring training.

The first part of this chapter describes a conceptor-based approach to classifying high and low task engagement from observable human movements. A conceptor-based network was trained on examples of the extreme levels of engagement along a continuum of intensity (i.e. high and low). The aim was to assess the performance of this approach with the eventual goal of combining the resultant conceptors to generate a third conceptor for classifying intermediate samples without training.

This work was conducted as part of the EU FP7 project DREAM[1], funded

[1] www.dream2020.eu

by the European Commission[2]. The goals of DREAM were to develop systems to support the use of socially interactive robots in the diagnosis of, and interventions for, Autism Spectrum Disorder. Consequently, one of the aims of this project was to develop a system which could be implemented in scenarios where a child would interact one-on-one with a social robot.

Hypotheses and predictions

This study was guided by a single hypothesis:

1. Conceptors trained on examples of high and low task engagement will be useful for distinguishing between test samples from these classes with an above-chance level of accuracy.

4.3 Method

4.3.1 Materials

The data set for this study was taken from the PInSoRo (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017) data set. For the purposes of this study, only the child-robot interactions were used. This was because the goal was to provide a system which could be implemented by the DREAM project wherein children would be interacting directly with a robotic system in a diagnostic setting. Specifically, data was extracted from the anonymous version of the PInSoRo data set which excludes the video streams (Lemaignan, Edmunds, and Belpaeme, 2017). This data set was constructed by pre-processing clips using the OpenPose library[3] (Cao et al., 2017) to extract skeletal and facial landmarks.

From this data set the pose, face and hands keypoints for each frame where the child had been annotated with the labels "goal-oriented play" (high engagement) and "no play" (low engagement) were extracted. Each 'frame' thus consisted of a 184-dimensional vector of the x and y coordinates for body, face and hand keypoints. A total of 354 'clips' were taken from this data set such that each clip was an $n \times 184$ matrix. A subset of "high" (62 clips) and "low" (115 clips) engagement clips made up the training data set. The remaining 177 clips made up the test data set.

[2] grant number 611391

[3] https://github.com/CMU-Perceptual-Computing-Lab/openpose/

4.3.2 Conceptor-Based Network

Implementation and evaluation of the classifier described below was done by Dr D. Hernández García.

Procedure

In order to create a conceptor-based network it was first necessary to compute 2 conceptors, one for each class. This was done by implementing an echo state network (ESN) with a single hidden layer reservoir. For each class the network was driven with all the training samples in each class one-by-one, according to the update equation described in Jaeger (2014a). For this procedure, each sample consisted of a single clip from the data set. From here, a conceptor for each class was computed from the state correlation matrix obtained from the ESN (for more details see Bartlett et al. (2019a) and Jaeger (2014a)).

Once a conceptor for each class had been computed, new samples from the test set were fed into the ESN. For each test sample a new state vector was generated, describing the state of the ESN whilst it was driven by this pattern. These vectors were compared to each of the trained conceptors, and a "positive evidence" score was calculated to describe the degree to which the new state vector could be characterized by each conceptor. Classification decisions were then made such that the conceptor with the highest "positive evidence" score was selected as the class to which the sample belonged.

4.4 Results

The results of testing the trained conceptors on previously unseen high and low engagement samples are shown in Figure 4.2 (right). Performance is above chance for both classes (high engagement: 60%, low engagement: 75%).

4.5 Discussion

The results of this study demonstrate that conceptors were successfully applied to the problem of distinguishing between high and low engagement states based on observable human pose information. Based on this and the studies using conceptor morphing in order to generate intermediate patterns (Jaeger, 2017), it is reasonable to expect that a new conceptor, generated by

combining the two trained conceptors, could potentially be used to recognize intermediate engagement states.

As detailed in Jaeger (2014b), new conceptors can be constructed either using logical AND, OR and NOT functions, or by mixing the two conceptors (C_1 and C_2) using a scaling factor (μ) as in the equation:

$$C_i = ((1 - \mu)C_1 + \mu C_2) \tag{4.1}$$

The use of boolean operations may not be appropriate for the task at hand as they over-simplify the problem. For example, if we were to use the boolean OR to construct an intermediate conceptor, the features used for classification decisions are restricted to those shared by both trained conceptors. Thus the resulting conceptor would be ignoring any features unique to the intermediate class. Alternatively, a new conceptor could be constructed by scaling between the two extremes using equation 4.1. This approach has proven particularly useful for smoothing the transitions between generated patterns (Jaeger, 2017) and is arguably more suitable for this project. Unfortunately, however, this could not be achieved within this project. Whilst an attempt was started, it could not be completed before the end of the collaboration with Dr D. Hernández García.

In parallel with this work, another approach was explored which was

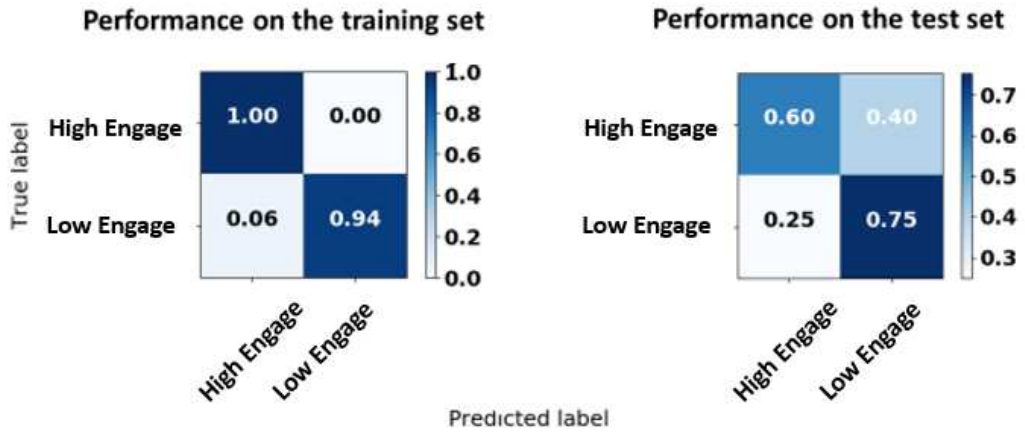


FIGURE 4.2: Confusion matrices showing classification performance of trained conceptors on training data (left) and test data (right).

inspired by recent work from Voelker and Eliasmith (2018). Voelker and Eliasmith (2018) present a biologically plausible dynamical spiking neural network, formulated in terms of the so-called Neural Engineering Framework (Eliasmith and Anderson, 2003) capable of exactly reproducing delayed time signals. In contrast with the rate-based approach of Conceptors, the ESN implemented by Voelker and Eliasmith (2018) uses spiking neurons. Spiking neurons can run in real-time on neuromorphic hardware and, at least in some cases, can be more energy efficient (Blouw et al., 2019). This is, therefore, a promising avenue to explore, particularly when we consider the possibility of implementing the classifier in a neuro-robotic platform.

4.6 Approach 2 - Delay Network

The approach of Voelker and Eliasmith (2018) allows one to create a spiking dynamical network which non-linearly encodes its input across a set delay interval. On one hand, thus, this approach is promising since the mathematical formulation leads us to expect high levels of accuracy and performance. However, it has only been demonstrated on very abstract, single-dimensional input patterns, whereas other approaches, for example, ESNs, have been shown to easily encode multi-dimensional inputs (such as the locations of various point-light markers on a human skeleton) (Mici, Hinaut, and Wermter, 2016; Bozhkov, Koprinkova-Hristova, and Georgieva, 2016). It is also not clear whether inputs that would be interesting in real-life conditions can be reduced to a smaller number of dimensions and still be meaningful for classification.

Hypotheses and predictions

The primary motive of this study was to examine whether the delay network could be trained on examples of the extremes of an internal state (i.e. 'high' vs. 'low' task engagement) and would then be able to recognize intermediate engagement as being in-between the two trained classes. Additionally, there are two secondary motives driving this study. The first is concerned with developing a classifier able to identify human internal states from human biological motion data. The second aim is to evaluate whether the approach of Voelker and Eliasmith (2018) can be applied to more realistic input patterns. This study is, therefore, largely exploratory in nature. However, the following hypotheses were proposed to structure this research:

1. The resultant classifier will be able to distinguish between high and low task engagement based on human movement information.

2. When presented with untrained samples of 'intermediate' engagement, the classifier will produce an output which is distinct from that produced for both trained patterns.

Specifically, for the second hypothesis it is predicted that the output produced for intermediate engagement can be characterised as being something in-between the outputs for the trained classes of high and low engagement.

4.7 Method

4.7.1 Materials

The data used in this study was again taken from the PInSoRo (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017) data set. In the study reported above using a conceptor-based approach we used the child-robot interactions. However, the present study considers the problem space of simply being able to recognize untrained states for application in a wider range of settings. As such, the aim of this study was to develop a more general approach and therefore it was appropriate to harness the possibility that children were more expressive in the child-child interactions. Consequently, the data was extracted from the child-child interactions within the anonymous version of the PInSoRo data set. From this data set each 'frame' was a 184-dimension vector consisting of the xy coordinates for the child's pose, facial features (including action units and gaze) and hand landmarks.

Within this data set, for annotation purposes, each child was labelled as either 'purple child' or 'yellow child' depending on the color of the vest they were given to wear. In the vast majority of videos, the 'purple child' was positioned on the left of the frame. To create the current data set, only the data from the 'purple child' in each annotated interaction was collected. All of the facial and skeletal data for the purple child in clips where they had been annotated with 'goal-oriented play' (high engagement), 'aimless play' (intermediate engagement) and 'no play' (low engagement) were collected. This gave a total of 248 clips (105, 52, 91 respectively).

The training set was constructed by taking 80% of the clips from each of the high and low engagement sets, reserving 20% for testing. All of the intermediate engagement clips were reserved for testing as the goal of this

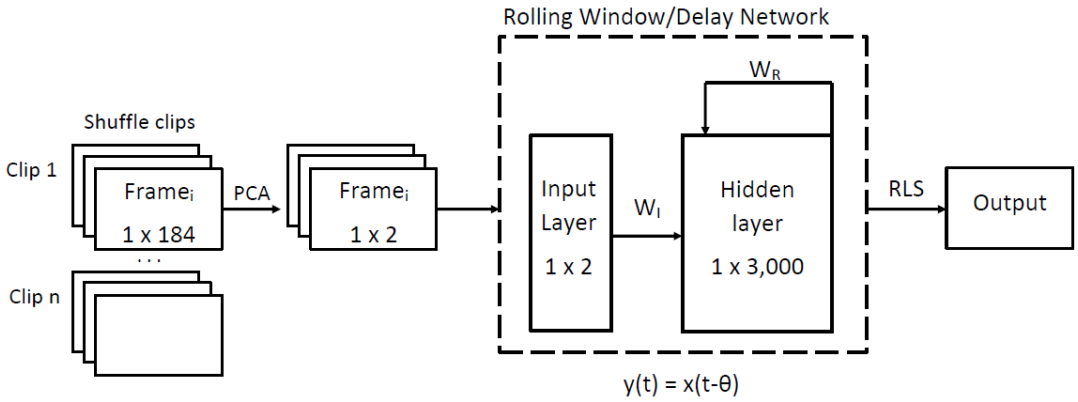


FIGURE 4.3: Schematic of Delay Network.

study was to see how well, after being trained on examples of high and low engagement, the classifier would perform on classifying the intermediate engagement clips as examples of an intermediate class.

4.7.2 Apparatus

This experiment used the same recurrent neural network as developed by Voelker and Eliasmith (2018), which can be thought of as an optimized reservoir. The network consists of a single hidden layer, a set of input weights, a set of output weights, and a set of recurrent weights. For the current study the delay network was implemented with 3,000 leaky-integrate-and-fire neurons in the hidden layer, and the decoder used the least-squares solver with L2 regularization (see Figure 4.3)

This approach differs from the conceptor-based method in how the connection weights are calculated. In ESNs (as used in the conceptor-based network) the weights are random, whereas the approach in Voelker and Eliasmith (2018) involves pre-computing the weights. This results in a network that is optimal for recording its own input over a period of time. That is, such a network can be used to approximate functions such as $y(t) = x(t - \theta)$, where θ (theta) is a scalar indicating how far into the past the network should remember. For this reason, this network is sometimes referred to as a delay network. This method works for any neuron model, including spiking leaky-integrate-and-fire neurons, as used here. The result is a a recurrent neural network where a rolling window is used in order to retain a memory of the history of the network's activity. The model is effectively a regression model where the classification problem is solved using linear least squares with regularization.

4.7.3 Procedure

Before feeding any data into the classifier, a Principle Component Analysis (PCA) model was constructed using the training data. An initial PCA analysis revealed that 2 components explained 64% of the variance, with component 1 explaining 40% of the variance, and component 2 explaining 24%. Additional components each explained <10% of the variance. Consequently, the constructed PCA model transformed the 184-dimensional data into 2 components. To understand the input to the classifier, the factor loadings for each of the PCA components were examined. A cut-off was applied such that only factors with a loading greater than +/- 0.08 were shown on each component. This revealed that component 1 was mostly correlated with the x coordinates of facial markers. Component 2, on the other hand, was mostly correlated with the y coordinates of facial markers. This is probably due to the fact that the children in the videos were mostly stationary, being in a seated position next to the sand-tray. Consequently, the majority of movement was likely in their facial expressions as they were talking and interacting with one another. Additionally, there may have been a lot of variation in the children's facial expressions between the two clip types (high and low engagement) used to construct the PCA, with children potentially being more talkative and expressive when highly engaged compared to when they demonstrated low task engagement. In contrast, the children's body's were fairly fixed in space given that the children were seated, and whilst there would have been some arm and hand movements, these probably showed less variance between clip types compared to facial expressions. Both training and testing data was transformed using this model.

The main parameter that required optimization was the theta value for the rolling window. This value can be thought of as the system's memory. As each frame of a clip is fed into the classifier, the rolling window retains a memory of the preceding frames. Consequently, the classifier does not classify based on individual frames, but takes into account the activity leading up to the current frame. Testing showed that a memory of 15-seconds ($\theta = 15$) produced the best classification results.

The final classifier was tested a total of 20 times. With each iteration (experiment) a new random sample of data was used for the 80% training and 20% testing sets. Additionally, for each experiment, a new set of weights was generated during the training phase and applied to the testing phases. The order of training and testing events was as follows: First, the classifier was trained on a random sample of 80% of the high-engagement data, and

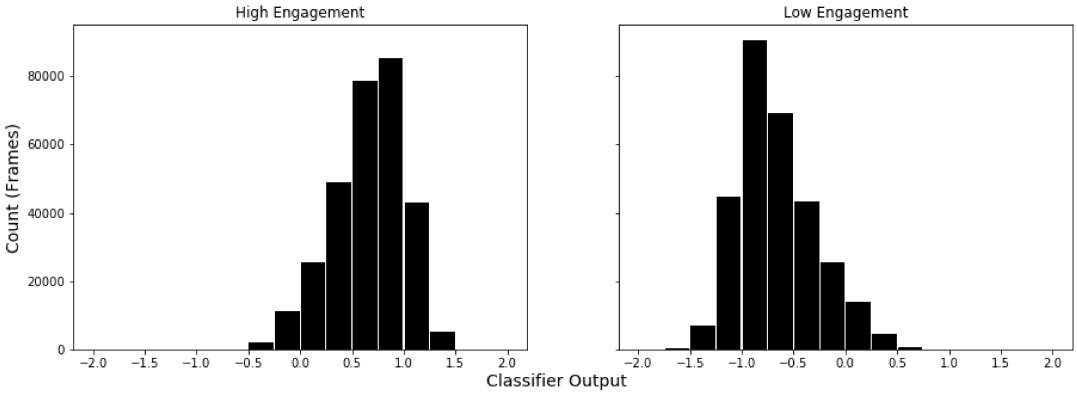


FIGURE 4.4: Distribution of the outputs given for each training frame from the high engagement class (left) and the low engagement class (right) across all 20 experiments.

80% of the low-engagement data. The first test phase involved testing on the remaining 20% of these data sets. Finally, the classifier was tested on all of the intermediate engagement data. The classification targets for each label were -1 (low engagement), 0 (intermediate engagement) and +1 (high engagement).

4.8 Results

Analyses were conducted using the Python `numpy`, `pandas` and `sklearn` toolkits in Jupyter Notebook. The analysis scripts can be found in the accompanying github repository (see Section 4.11 for details).

Before conducting any analyses, the outputs given for the training samples across all 20 experiments were plotted. These plots show that the distribution of outputs was generally centred around the target values (+1 for high engagement, -1 for low engagement) (see Figure 4.4).

Preprocessing: The output from the classifier was such that each frame was given a classification value between -1 and +1. As we are interested in producing a classifier which can provide accurate classifications over time, and not necessarily on individual frames, we applied a median filter to the data so that we could plot the most common classification within a given time frame within each clip. Before applying the filter we 'rounded' the raw classification output to its nearest target value. This process was done differently depending on whether just the outputs for the high and low engagement test data were being considered, or if the intermediate class was also included. That is, when examining just the high and low engagement test data, values

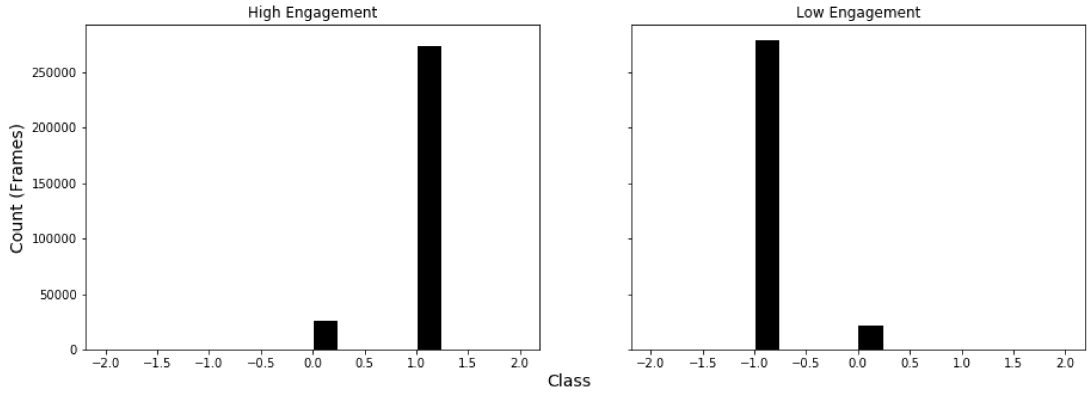


FIGURE 4.5: Distribution of output values given for the training frames after applying a median filter. Data taken from all 20 experiments.

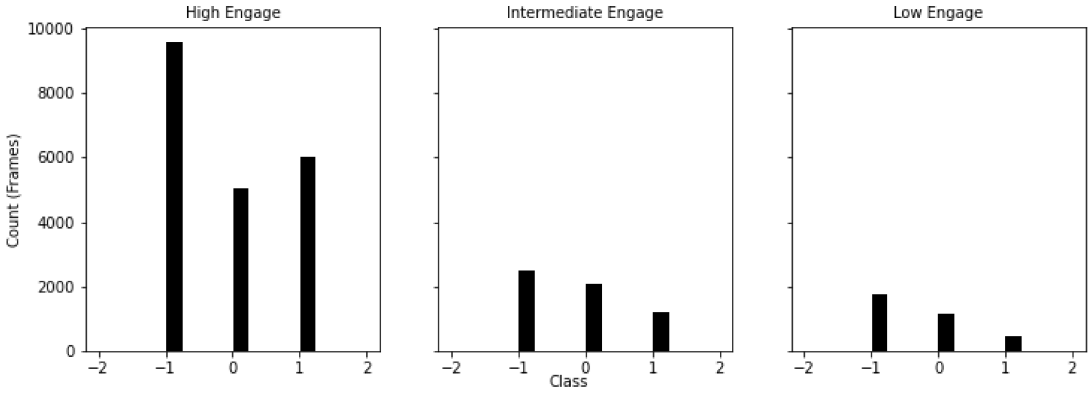


FIGURE 4.6: Distribution of output values given for the high, intermediate and low engagement testing frames after applying a median filter. Data taken from all 20 experiments.

< 0 were rounded to -1 (low engagement classification), and values > 0 were rounded to +1 (high engagement classification). In contrast, when including the intermediate class, values < -0.3 were rounded to -1, values < -0.3 and > 0.3 to 0 (intermediate engagement classification), and values > 0.3 to +1. The median filter was then applied such that the value for each frame was the median for a window of 99 frames.

Training Data: In order to assess the success of the training phase, the median filter was applied to the system's output during the training phase. Plotting the distribution of this output across all 20 experiments shows that the majority of clips in the training data sets were correctly classified (see Figure 4.5).

Testing Data: Examining performance on the testing data was split into

two sections. First, performance on just the trained classes (high and low engagement) was examined. As outlined above, this involved a larger threshold for values to be classified as either high or low such that any output value greater than 0 was rounded to +1, and values less than 0 were rounded to -1. The median filter was then applied and the distribution of outputs across all 20 experiments can be seen in Figure 4.6.

The percent of frames correctly classified in each experiment was then calculated and averaged across all 20 experiments showing that a mean of 56.13% ($SD = 12.91$) high engagement frames and 58.04% ($SD = 15.26$) low engagement frames were correctly classified in each experiment. This poor performance (no better than chance), in contrast to the high performance on samples from the training sets (see Figure 4.5) suggests that the classifier may have been overfitting to the training data. The large standard deviations, however, suggest that there may have been some experiments which performed well. Given that new weights were generated for each experiment, it may then be that within the 20 models there is at least one which produced an above-chance performance on all of the classes. Furthermore, what is of most interest is how the classifiers performed on the previously unseen intermediate engagement class. The next analysis therefore looks at performance on all three testing data sets (high, intermediate and low engagement) for each experiment separately.

In order to include the outputs from intermediate engagement samples the thresholds used for rounding the outputs to a classification label were altered as follows: values < -0.3 rounded to -1, values < -0.3 and > 0.3 to 0, and values > 0.3 to +1. The median filter was again applied to the outputs from high and low engagement samples in the testing data sets, along with the values for the intermediate engagement samples and the results were plotted. Plots showing the distribution of outputs across all 20 experiments can be seen in Figure 4.6 and the individual plots from all 20 experiments can be seen in Appendix A. From these plots we can see that there was no experiment in which the classifier was able to correctly classify the majority of samples from each class. That is, at least one class is confused for another.

Finally, averaging the percent of frames classified correctly showed that, 37.05% ($SD = 12.46\%$) of high engagement frames, 34.32% ($SD = 5.27\%$) of intermediate engagement frames and 42.68% ($SD = 13.11\%$) of low engagement frames were classified correctly following the application of the median filter. These results further suggest a case of overfitting to the training data, given that no model appears to have performed significantly above chance

on all three classes.

In an effort to eliminate the possibility that this poor performance was an artifact of the use of a median filter, the average output per clip was also calculated. The same performance metrics were examined (i.e. distribution of classification values for each clip, and average percent correct). Unfortunately, this analysis did not produce better classification performance.

4.9 Discussion

In this study the approach of Voelker and Eliasmith (2018) was applied to the problem of classifying human movement information into different internal states. Good performance was obtained when testing on the trained patterns - demonstrating that a classifier can recognize different internal states based on human movement information. However, performance on the testing samples, including the untrained intermediate class, was very poor. Thus, neither of the hypotheses put forward were supported. Furthermore, the pattern of good performance on training samples but poor performance on testing samples suggests that the model may have been overfitting to the training data. Whilst feature reduction (i.e. PCA) is often suggested as a method for preventing overfitting (Defernez and Kemsley, 1999; Liu, 2017; Kumar, 2019), in this case it may have been providing the wrong data; i.e. there may not have been enough information available in the PCA components for accurate discrimination. Indeed, some sources do suggest that using PCA can lead to poor results in regards to preventing or reducing overfitting (Rebala, Ravi, and Churiwala, 2019). In the current task, the classifier is trying to distinguish between 3 states which are closely related to one another (i.e. levels of engagement). In terms of behaviour, the differences between these states are therefore likely to be very small/subtle. That is, the children are in the same position (kneeling at the sand-tray), performing roughly the same task (interacting with the sand-tray or their companion sat on the other side of the sand-tray) in every clip. Therefore, the quantitative differences in how our subjects move in each state are likely to be much smaller than in cases where the activities being discriminated are much more distinct (e.g. the acts of following vs. passing someone as in Kelley et al. (2008)). Additionally, given that we do not train the classifier on one of our classes at all, we must consider that there is likely a need for more information from which to draw distinctions.

4.10 Conclusion

This chapter reports two approaches to the problem of classifying different levels of the human internal state task engagement based on observable human movement and posture behaviours. The first, a conceptor-based approach, was successfully trained to discriminate between high and low engagement. However, the step of constructing a new, untrained conceptor and testing it on the intermediate engagement class could not be carried out within the scope of this project. In contrast, the second delay-network approach was tested both on the trained high and low engagement classes, and on the untrained intermediate engagement class. The results of this experiment showed that the performance was effectively no better than chance.

4.11 Open-Source Resources

The repository containing the work for the Conceptor-based network can be found at: https://github.com/dhgarcia/conceptorsTest.

For the delay network, the following github repository contains scripts for the experiments and analysis. https://github.com/maddybartlett/Thesis_Notebooks/tree/master/Chapter4_DelayNetwork

Chapter 5

Study 4 - Estimating Untrained Intermediate States

Parts of this work have been published in the proceedings of the *2021 ACM/IEEE International Conference on Human Robot Interaction* (see Appendix G) (Bartlett, Stewart, and Thill, 2021).

5.1 Introduction

In the previous chapter a delay network was developed and applied to the problem of classifying an untrained intermediate class after training on two 'extreme' states along a continuum. The resultant network demonstrated poor performance on the trained classes as well as the untrained intermediate engagement class. Whilst the exact cause of this poor performance is unclear, a recent development presents a promising alternative.

Voelker, Kajić, and Eliasmith (2019) present a recurrent neural network which uses a novel architecture for dealing with temporal data - Legendre Memory Units (LMUs). LMUs produce an output which encodes both the input signal and information about the history of that input. If one wants to consider all inputs from the last θ seconds, one can use the LMU function to convert every 1 value in a d-dimension vector into q new values which characterises the input over the last θ seconds. Importantly, the LMU method improves on existing reservoir techniques in that their structure is derived from first principles in order to produce optimal reservoir-like behaviour. Simply put, Voelker, Kajić, and Eliasmith (2019) determined mathematically how an input should be transformed into a higher-dimensional output so that it best encodes the history of the input for the desired duration before constructing the RNN to do this transformation. This is in comparison with other methods, including ESNs such as that used in Chapter 4, which start with a general architecture and then explore different hyperparameters or

architectures until the desired behaviour is obtained. Further details of this approach can be found in Voelker, Kajić, and Eliasmith (2019).

Despite their short history, existing evidence demonstrates that LMUs can achieve state-of-the-art performance whilst being efficient to implement, with fewer parameters compared to other approaches, such as LSTMs and the recently proposed Non-saturating Recurrent Unit (Voelker, Kajić, and Eliasmith, 2019). As such, the method shows a lot of promise for dealing with temporally dependent tasks, whilst being well suited to the constraints of real-world applications (Blouw et al., 2020). This study, therefore, explores whether they offer a benefit for the kinds of applications seen in Human-Computer Interactions and social robotics, such as internal state recognition. Here the LMU method is used as a pre-processing step such that its output will be used as input to a system for the task of estimating engagement from dynamic patterns.

This study primarily investigates whether systems which incorporate the LMU pre-processing method will, after training on high and low task engagement, provide an output to intermediate task engagement which can be used to identify this class as being 'in-between' the two trained classes. More specifically, this is an investigation of whether LMU pre-processing will improve the system's performance, not only on the trained classes, but if it is also able to generalize to the untrained intermediate class. Three systems are compared on this task; a Nengo Deep-Learning Network (NDL), a Multi-Layer Perceptron (MLP) and logistic regression (LR).

Hypotheses and Predictions

Two hypotheses were put forward for this study:

1. The use of LMUs as a pre-processing step will change the performance of the systems.

2. The systems, after training on examples of high and low task engagement, will produce an output in response to examples of intermediate engagement which can be used to identify these samples as being related to, but different from, the extremes *without being trained on them*.

Specifically, for the first hypothesis, based on previous findings that methods incorporating LMUs outperform other machine learning methods (Voelker, Kajić, and Eliasmith, 2019; Wang et al., 2020), it is predicted that all three

systems will show an improved performance on data that has been pre-processed using LMUs, compared to the raw, unprocessed data. Furthermore, it is expected that this improved performance will apply when examining the outputs produced by the systems (NDL, LR and MLP) for both individual frames and full clips. For individual frames, improved performance is expected because, unlike the raw data, the LMU pre-processed frames will contain information about the history of the clip and it is assumed that, for this type of information, the dynamic unfolding of behaviour over time will contain more information about task engagement than individual snapshots in time. For full clips, an improvement is expected even when using simple, naïve metrics such as the predominant class of the frames contained in the clip. Measuring clip-wise performance is particularly relevant for naturalistic data sets, such as the PInSoRo data set, wherein although a clip might be labelled as a certain class, there is no guarantee that all the frames it contains are good exemplars of that class.

In regards to the second hypothesis, success is defined under the following predictions: (1) it will be possible to distinguish between system outputs produced when given random data, compared to engagement data, as input during testing, and (2) that sequences from the untrained, intermediate class will be distinguishable from the two trained classes. It should be noted that for this second prediction, some overlap between the task engagement classes is to be expected, particularly given the fact that we use naturalistic data as stimuli. To test these predictions the trained systems will be fed the testing clips from both trained classes as well as a new class of intermediate engagement (not previously seen by these systems) and randomly generated data. A k-nearest-neighbour classifier will then be used to distinguish between the outputs of each system based on 4 descriptive statistics (mean, standard deviation, skew and kurtosis). This analysis is also intended to examine whether any one of the six approaches best allows us to recognize the untrained class without sacrificing performance on the trained classes.

5.2 Method

5.2.1 Design

This study took a 3 (NDL vs. LR vs. MLP) x 2 (without vs. with LMUs) design examining the effect of system and pre-processing step on the performance (accuracy) of the system. This resulted in 6 conditions or approaches

being tested: (1) NDL, (2) LMU-NDL, (3) LR, (4) LMU-LR, (5) MLP, and (6) LMU-MLP.

5.2.2 Materials

The data used as input for this study was the same as that in Chapter 4. That is, data was extracted from the child-child interactions within the anonymous version of the PInSoRo data set (Lemaignan et al., 2018; Lemaignan, Edmunds, and Belpaeme, 2017). This gave 105 high engagement, 52 intermediate engagement and 91 low engagement clips wherein each frame was a 184-dimension vector consisting of the xy coordinates for body landmarks including joints, facial features and hands.

The training, testing and validation data sets were created by applying a 70/20/10 split respectively. For this experiment, the systems were only trained on examples from the high and low engagement sets, so the above split ratio was only applied to these two classes. Additionally, this split accounted for the number of clips in each set, rather than the number of frames. When creating the training data set, it was first necessary to account for the fact that the low engagement set had fewer clips than the high engagement set. As such, the equivalent of 70% of the low engagement clips was taken from both the high and low engagement sets to construct the training set. This same approach was used for constructing the testing and validation sets. Note that the validation set was only used for the NDL, but the high and low engagement data was still split 70/20 for the MLP and LR approaches. This resulted in a total of 126 clips from the low and high engagement sets being used for training (roughly 175,000 frames on average), and 36 clips for testing (roughly 50,000 frames). All 52 of the intermediate engagement clips (55,296 frames) were used for testing on untrained patterns.

In regards to the second hypothesis, the goal was to test whether systems trained to classify high and low engagement could also estimate intermediate engagement without being trained on any examples from this class. In order to establish that the systems were recognizing intermediate engagement samples as being *related to*, but different from the high and low engagement samples, it was necessary to verify that samples from the intermediate class were being treated as engagement data, and not simply as data which does not belong to the trained classes. This was done by testing the systems on random data, generated by creating arrays of random values in the same shape as the high engagement data (234507×184). The result was that there

were 105 random 'clips' in total, or 234,507 'frames'. A random selection of 18 'clips' from this random data set were used for testing in each experiment.

5.2.3 Apparatus

Legendre Memory Units

Before anything else, the raw data (high, intermediate and low engagement, and random data) was processed using the LMU method. The full architecture presented in Voelker, Kajić, and Eliasmith (2019) consisted of a linear dynamical memory (Legendre Delay Network, LDN) and a non-linear decoder. For this study, the non-linear decoder was replaced with the NDL, MLP and LR systems. As a result, the LMU method applied here involved feeding the input vectors (frames) into the LDN, the output from which would then be given as input to the three systems. The weight matrices connecting the input to the linear layer, and for the recurrent connection from the linear layer back to itself were pre-computed and fixed. Consequently, no training was required for this step. For further details on LMUs and the LDN see Voelker, Kajić, and Eliasmith (2019) and Voelker, Rasmussen, and Eliasmith (2020).

Setting Parameters: The parameters to be set for the LMU step were q and θ (theta). Specifically, θ values 1, 3, 5 and 7, and q values 2, 3 and 4 were tested. For each combination of these parameters, new high and low engagement data sets were generated and split such that 70% was used for training, and 20% for testing. The training data sets were then used as input to the MLP (with N-neurons = 200) to determine which combination of θ and q produced the best accuracy scores on average when tested with the unseen 20%. Each combination of θ and q values was tested 20 times. The results shown in Figure 5.1 demonstrate that $\theta = 3$ and $q = 4$ produced the highest accuracy scores on average.

An interesting feature of these results is that lower q values appear to increase the spread of accuracy scores. When q is 2, for example, there seems to be an increased probability that some networks will perform poorly whilst general performance (e.g. mean accuracy) remains similar.

Pre-processing: During the LMU pre-processing step, each clip was processed separately. Each clip was presented to the LMU network as a sequence of the 184 dimensional vectors which made up each frame. The output consisted of 736-dimensional vectors ($184 \times q$) such that each vector contained information about the current frame, as well as encoded information about

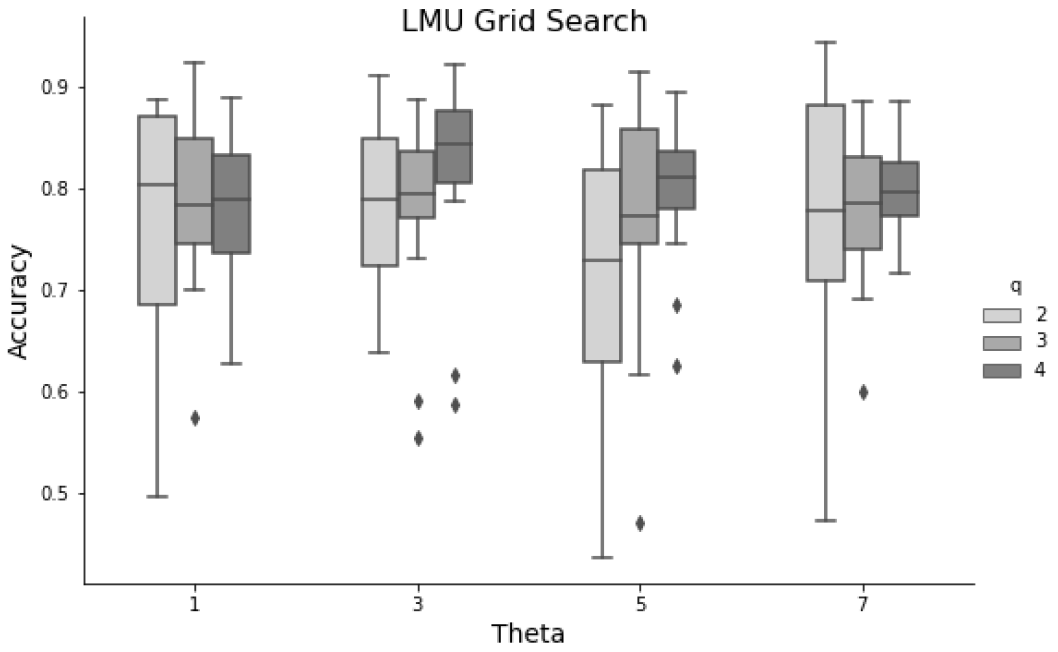


FIGURE 5.1: Box-plots showing accuracy for each LMU parameter combination in the grid search using MLP with LMU pre-processing. Each combination was tested 20 times.

the preceding 3 (θ) seconds. These outputs were saved as NumPy array files to be used as input for the classification systems (see Figures 5.3, 5.2 and 5.4).

Systems

For this study, 3 different systems were used and compared. Two of these were out-of-the-box methods. Namely a logistic regression (LR) and a Multi-Layer Perceptron (MLP) implemented using the `sklearn` toolkit in Jupyter. The LR used `sklearn`'s default settings and a maximum of 1,000 iterations (see Figure 5.3). When implementing `sklearn`'s MLP (with one hidden layer), the following parameters were used: the activation function was the rectified linear unit function (relu), and the weights were optimized using the stochastic gradient-based optimizer 'Adam' (Kingma and Ba, 2014) (see Figure 5.2).

The third approach, hereafter termed NDL, was constructed using the Neural Engineering Framework (NEF), specifically the NengoDL simulator (Rasmussen, 2019). This system was a feed-forward neural network with one hidden layer and ReLU activation functions (see Figure 5.4).

Setting Parameters: Separate grid-searches using the MLP and the NDL were conducted in order to establish the 'best' settings for a number of parameters. The input data for these grid-searches was the raw data without LMU pre-processing.

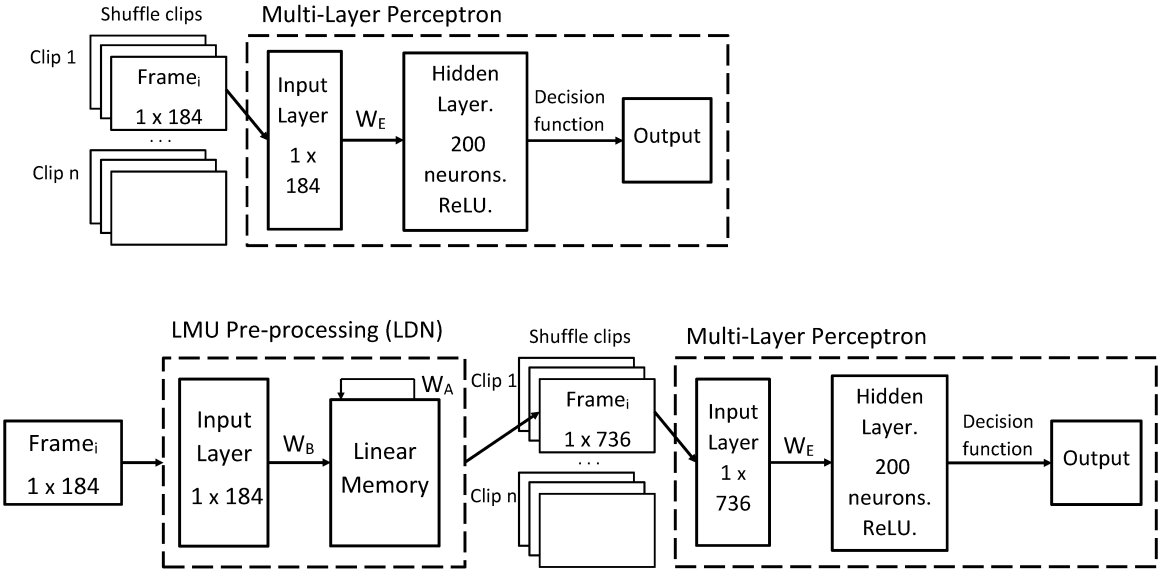


FIGURE 5.2: Schematic of MLP (top) and LMU-MLP (bottom) architectures.

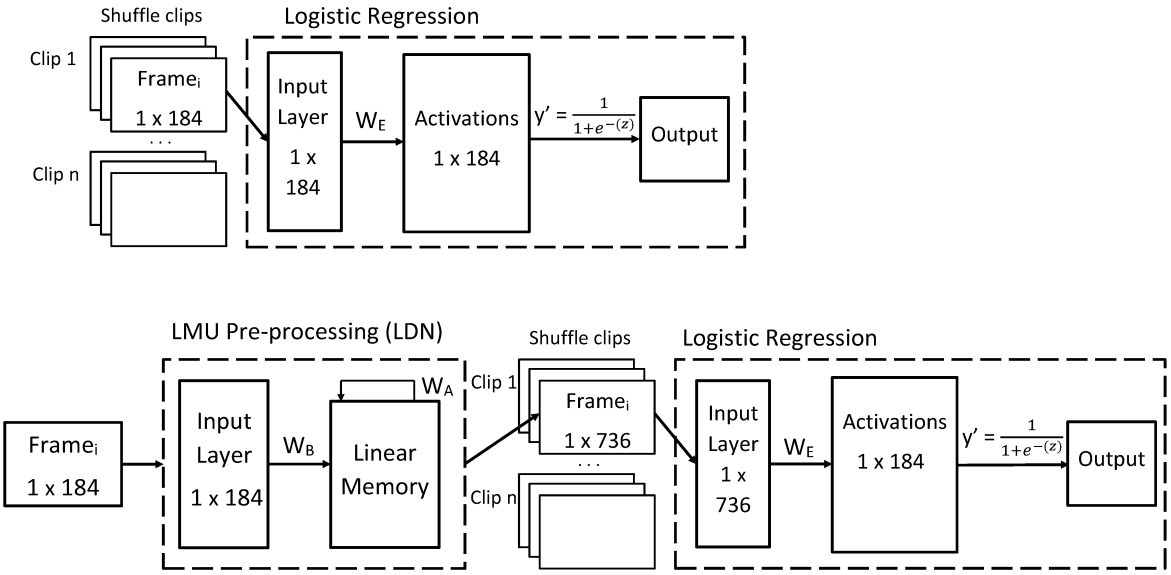


FIGURE 5.3: Schematic of LR (top) and LMU-LR (bottom) architectures.

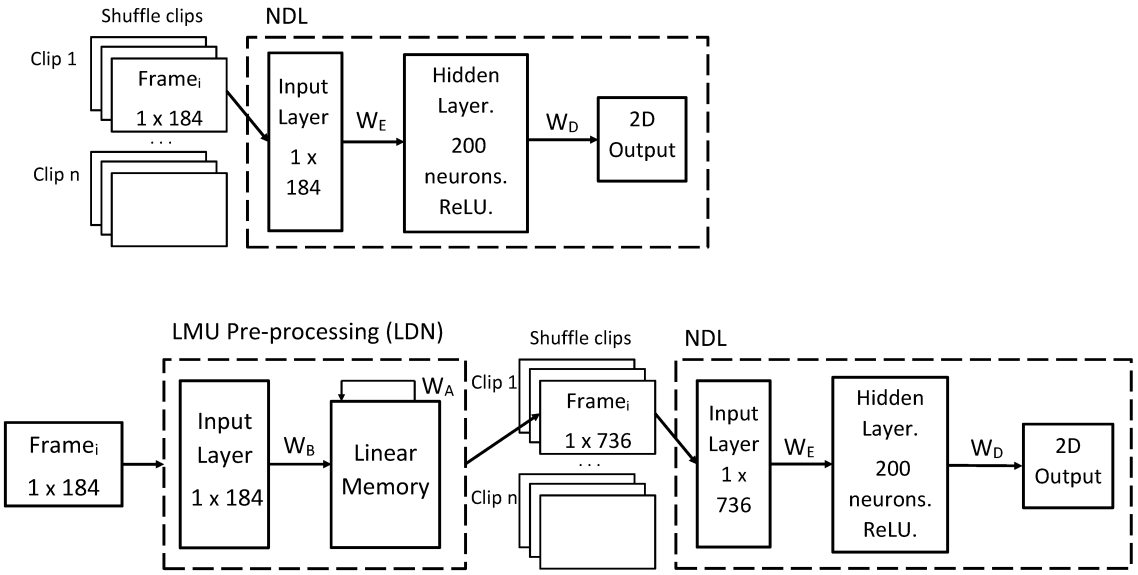


FIGURE 5.4: Schematic of NDL (top) and LMU-NDL (bottom) architectures.

For the MLP, different numbers of neurons in the hidden layer were compared. This grid-search compared 50, 100, 150 and 200 neurons, with each value being tested 20 times. Interaction plots showing each value for number of neurons against accuracy reveal that 200 neurons tended to produce the best accuracy scores (see Figure 5.5).

The grid-search for the NDL examined a range of learning rates (1e-01, 1e-03 and 1e-05). Interaction plots of learning rates against accuracy showed that the learning rate of 0.001 produced the highest accuracy. However, plotting the loss values during training revealed that the network did not reach a point where it had 'learnt' the two classes (i.e. the loss value was constantly decreasing and did not reach a point of stability). The network was then trained over 5,000 epochs, as opposed to the original 1,000, to see if the loss value would flatten out after further training. As can be seen in Figure 5.6 this was not the case. Consequently, it was decided that this network would not be included in the analysis due to how difficult it proved to be to train.

5.2.4 Procedure

Both of the remaining systems (MLP and LR) were trained and tested 20 times (each iteration being referred to as an 'experiment'). Regardless of the system being used, an experiment consisted of the following steps. First, the

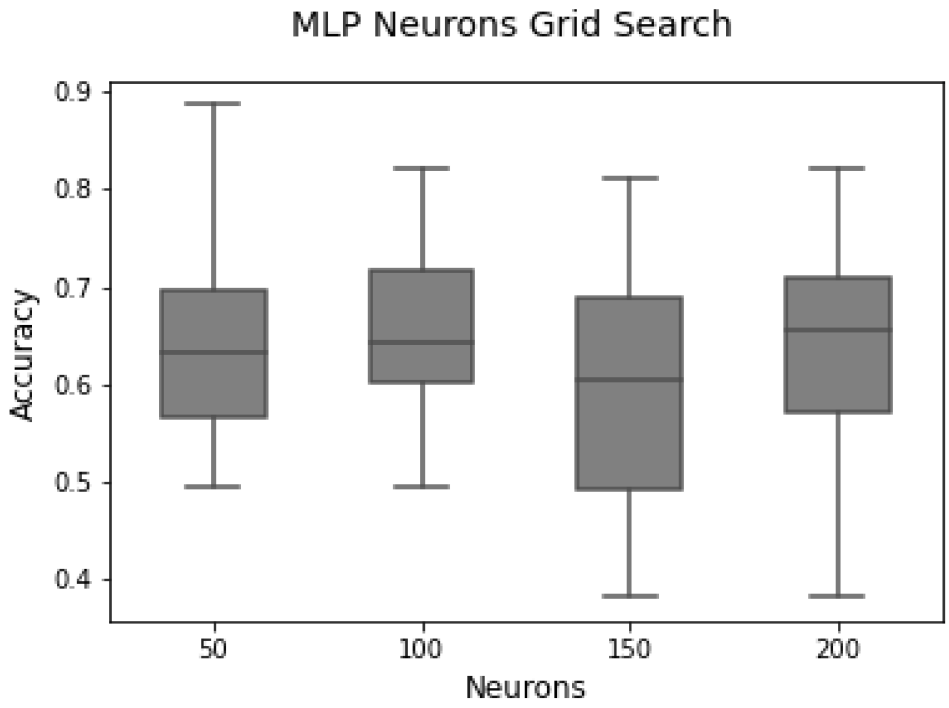


FIGURE 5.5: Box-plots of accuracy against each value for the number of neurons in the MLP hidden layer tested in the grid-search. Each value was tested 20 times.

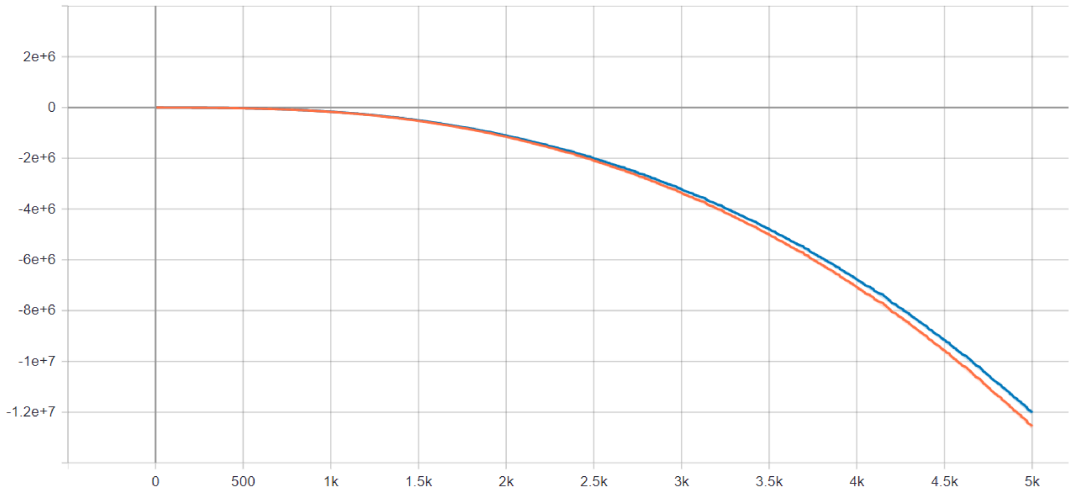


FIGURE 5.6: Plot of loss vs epoch when the NDL network was trained over 5,000 epochs with a learning rate of 0.001. Graph is taken from tensorboard and shows loss on the training data (blue) and loss on the validation data (orange).

high and low engagement data (either the pre-processed versions or the raw versions) were randomly shuffled and then split into the training and testing sets. The system was then trained on the training high and low engagement data and tested on the 20% testing high and low engagement data sets, as

well as on all of the intermediate engagement and random data.

As it was not possible to introduce a third classification option after training, the final system outputs (binary classification labels) could not be used. Consequently, for each system the outputs recorded for analysis were values produced before binarization into the two output classes. From the LR, the output used for analysis was the probability estimate denoting how probable it was that each sample was a member of each of the two trained classes (see Figure 5.3). From the MLP the final hidden layer's output (i.e. the output of the decision function) was recorded for analysis (see Figure 5.2).

All of the experiment and analyses scripts were run on a Lenovo Thinkpad L380 laptop running Windows 10. Each experiment using LR both with and without LMU pre-processed data took less than a minute. MLP using the raw data took roughly 20 minutes and MLP with LMU pre-processed data roughly 10 minutes.

5.3 Results

Analyses were conducted using the Python `numpy`, `pandas`, `SciPy` and `sklearn` toolkits in Jupyter Notebook. The analysis scripts have been made openly available (see Section 5.6 for details).

The following analysis has been split into two main sections which reflect the hypotheses. First, the effect of LMU pre-processing on system performance (accuracy) when tested on high and low engagement clips is evaluated. The second section of this analysis examines performance on the intermediate engagement and random classes with a view to establishing: (1) whether the untrained classes could be distinguished from the trained classes based on system output, and (2) whether there was a particular approach which produced the best overall performance on all 3 engagement classes.

5.3.1 Effect of LMUs on Performance on Trained Classes

In this section the results of training and testing using LMU pre-processed high and low engagement patterns are compared with training and testing using the original, raw patterns. It should be noted that the final outputs of the MLP and LR systems used here were binary classification decisions, wherein a sample was assigned to either the high or low engagement. However, because the problem at hand required a way for untrained classes to be estimated, it was necessary to obtain an output which could fall anywhere

within a range of values, and can therefore be considered as providing an estimation of state. This continuous output would then be translated so that it still reflected the categorical labels in the data. For the MLP approaches, the decision function was used as the output for analysis. The final output for the LR was the probability value that the sample belonged to the high engagement class. For both output metrics an output of >0.95 (rounded to 1) indicated a strong probability that the sample belonged to the high engagement class, and an output of <0.05 (rounded to 0) indicated a strong probability that it was from the low engagement class.

Frame-by-Frame Estimation

In order to assess performance of the MLP and LR on a frame-by-frame basis the distribution of estimation values given for test frames from both of the trained classes for all four architectures were plotted (see Fig 5.7). These plots reveal that the use of LMUs did improve performance on the trained classes. Specifically, for both LR and MLP, the use of LMUs seems to have facilitated an increase in the frequency of '0' or '1' classification decisions compared to the more spread out distribution of values when the data was not pre-processed. As a result of these plots it was decided that a 'correct' estimation for high and low engagement would be values of 1 (>0.95) and 0 (<0.05) respectively; the plots show that most of the estimation values fell into these bins, and there was therefore little benefit in broadening this threshold in terms of accuracy. The average percent of frames estimated correctly was then calculated revealing that LMU pre-processing did indeed improve performance. That is, for LR average percent correct rose from 21.93% (SD = 8.19) to 49.59% (SD = 14.96) for high engagement, and from 41.12% (SD = 15.4) to 76.00% (SD = 9.14) for low engagement. Similarly, for the MLP, correct estimation of high engagement samples increased from 60.97% (SD = 16.9) to 87.84% (SD = 11.38) and for low engagement from 49.12% (SD = 18.0) to 85.10% (SD = 85.1).

Clip-Wise Estimation

Whilst performance on individual frames shows promise, this does not necessarily mean that performance will be good on full clips. In particular, considering that the frames were extracted from clips of varying lengths, the systems may simply have learned to correctly identify only the frames from the lengthier clips. This next analysis therefore examines performance on

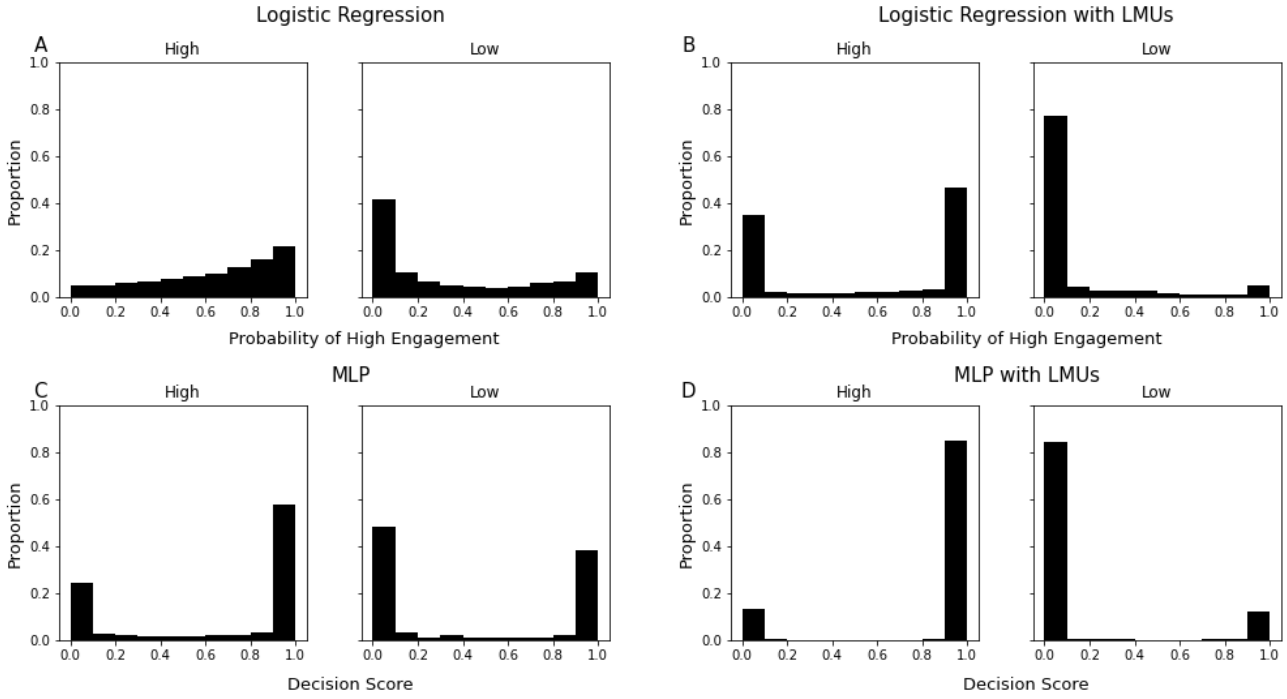


FIGURE 5.7: Histograms showing proportion of estimation values for each frame from the high and low engagement classes. Estimation value of 0 is treated as a low engagement classification, and 1 as a high engagement classification. Sub-figures display the results of classification using: (A) LR without LMU pre-processing (total frames: high = 794,184, low = 207,414), (B) LR with LMU pre-processing (total frames: high = 765,572, low = 207,405), (C) MLP without LMU pre-processing (total frames: high = 792,890, low = 226,601), (D) MLP with LMU pre-processing (total frames: high = 878,720, low = 221,186).

whole clips by calculating the average estimation value across all the frames in a clip. Figure 5.8 illustrates that performance on full clips was similarly improved by the addition of LMU pre-processing, and further that, for most systems, the majority of clips had an average estimation value falling into either the 1 or the 0 bin (i.e. >0.95 or <0.05). Calculating the average percentage of clips in each class identified correctly by each approach supports this conclusion (see Table 5.1). So whilst clip-wise analysis does demonstrate an overall drop in accuracy compared to the frame-by-frame analysis, performance of the MLP with LMU pre-processing is still well above chance for the two trained classes.

Effect of system type and LMU pre-processing

To verify whether the differences in performance described above are significant, a two-way ANOVA was performed, with system type (LR vs MLP)

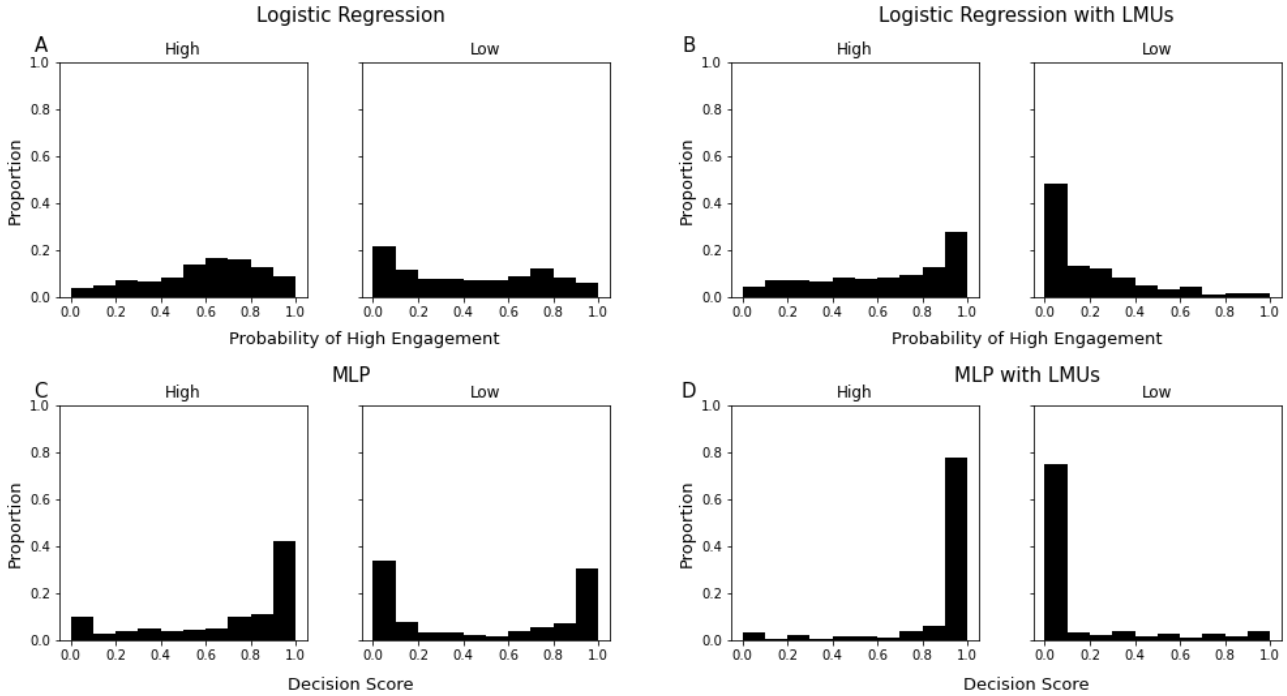


FIGURE 5.8: Histograms showing proportion of estimation values for each clip from the high and low engagement classes (360 clips in each plot). Classification value of 0 indicates a low engagement classification, and 1 is a high engagement classification. Sub-figures display the results of classification using: (A) LR without LMU pre-processing, (B) LR with LMU pre-processing, (C) MLP without LMU pre-processing, (D) MLP with LMU pre-processing.

and pre-processing (without vs. with LMUs) as independent variables and the percentage of correctly estimated high and low engagement clips as the dependent variable. Both assumptions of normality (Shapiro-Wilk test: $W = 0.983$, $p = 0.354$ and equal variances (Bartlett's test for sphericity: $\chi^2 = 0.990$, $p = 0.804$) were met.

The two-way ANOVA revealed a significant main effect of system such that LR (Mean = 26.74%, SD = 13.44) was significantly out-performed by MLP (Mean = 57.36%, SD = 20.57) (two-way ANOVA: $F(1,76) = 395.244$, $p < 0.001$, $\eta_p^2 = 0.443$). Additionally, and of key interest, the main effect of pre-processing step (with vs. without LMUs) showed that the use of LMUs (Mean = 57.36%, SD = 20.55) significantly improved performance compared to when LMUs were not used (Mean = 26.74%, SD = 13.47) (two-way ANOVA: $F(1,76) = 395.244$, $p < 0.001$, $\eta_p^2 = 0.443$). Finally, there was a significant interaction effect (two-way ANOVA: $F(1,76) = 25.040$, $p < 0.001$, $\eta_p^2 = 0.028$). Figure 5.9 suggests that the interaction effect was such that the effect of LMU pre-processing on performance accuracy was greater for the

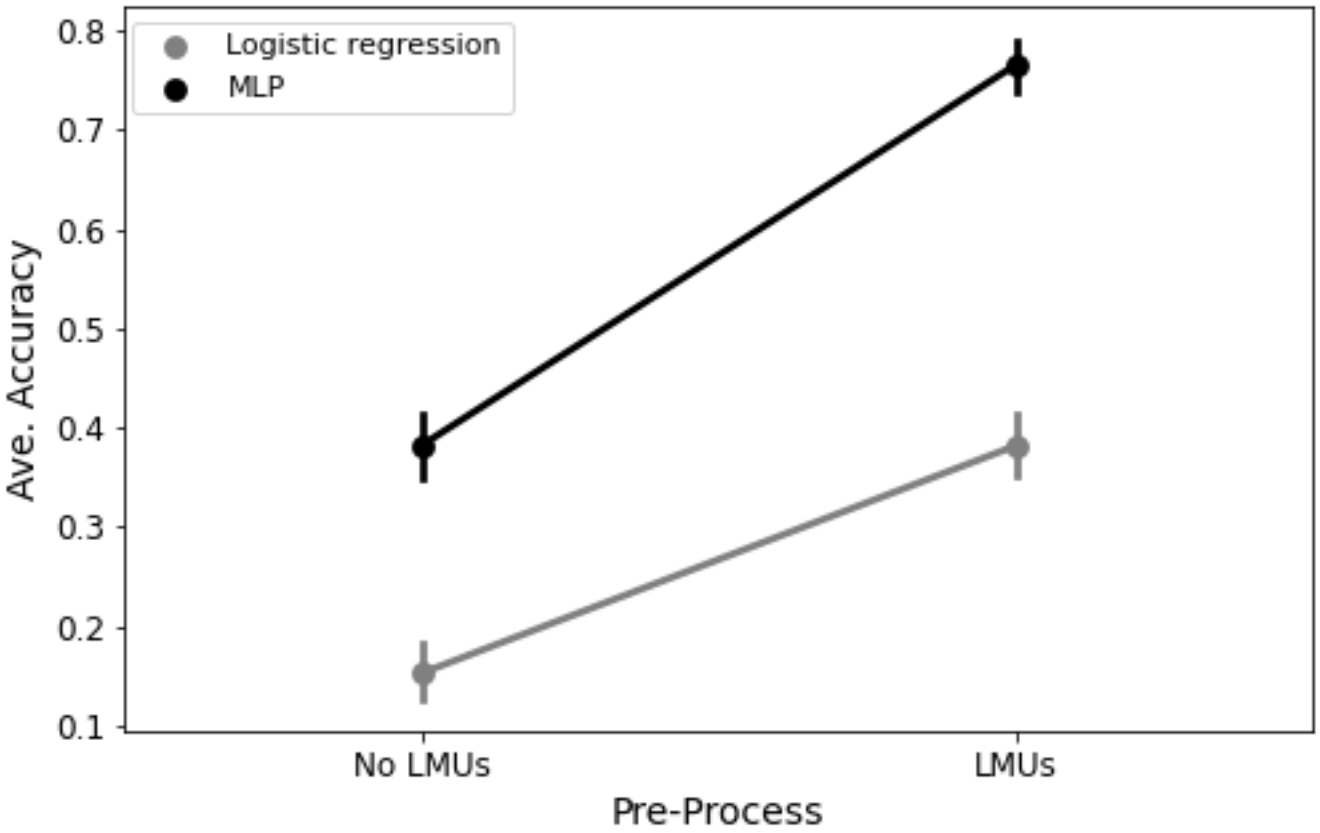


FIGURE 5.9: Interaction plot showing the interaction between classifier and pre-processing step on performance accuracy.

MLP than for the LR. That is, the difference in average performance for the MLP with vs. without LMU pre-processing (0.383) was significantly greater than for the LR (0.229). These results demonstrate that LMUs were an effective pre-processing step for facilitating improved performance on distinguishing between high and low engagement. Furthermore, it suggests that the best approach to use when separating classes on a clip-wise basis was the LMU-MLP.

5.3.2 Performance on Untrained Classes

Each approach was also tested on a third, unseen intermediate class of engagement. Here the intention was to see whether the LR and MLP would produce an output which could be used to identify this third class as being somewhere in-between the two trained classes. It is important, however, to also establish that the systems produced an output which identifies this class as still being related to the two trained classes, and not simply as something that does not belong to either. The systems were, therefore, also tested with random data as input. The output was analysed in a clip-wise manner.

The distributions of the average estimation value for each clip in all 20 experiments is presented in Figure 5.10. What can be observed is that the mean output values for intermediate engagement clips are generally more spread out between 0 and 1 than for the high and low engagement classes. Looking at the average percent of intermediate clips identified correctly reveals an interesting pattern such that the system with the best performance on the trained classes (MLP with LMU pre-processing) shows the worst performance on the intermediate class (see Table 5.1). However, this is likely because the less successful approaches tend to produce '0' and '1' estimation values less frequently for all classes, so the high accuracy on intermediate classes is likely an artefact of poor performance overall. Interestingly, the average outputs produced in response to random data are certainly distinct from all of the engagement classes, with a much greater tendency for an estimation value of or around 0, and much less spread.

To examine this further, the system's output for each frame was plotted along the timeline of 18 clips from each class in the first experiment of each approach. These plots are presented in Figure 5.11. The two effects which can most readily be seen from these plots are, first, that the outputs across the duration of each clip appear to differ markedly between classes for all four approaches, and second that the overall effect of LMU pre-processing was to stabilize and smooth the outputs of the systems. Of particular interest is that, where LMU pre-processing was used, the low engagement clips differ from the random clips in that the output in response to low engagement clips contains more instances where the output is non-zero. This illustrates a general difficulty with naturalistic data which is that their content is rarely

TABLE 5.1: Table of mean and standard deviation of percentages of clips that were estimated correctly by each approach in each experiment.

		LR	LMU-LR	MLP	LMU-MLP
High	*M*	8.89%	27.78%	42.22%	78.06%
	SD	(7.93)	(8.78)	(16.70)	(10.46)
Intermediate	*M*	87.50%	64.52%	43.85%	39.33%
	SD	(4.77)	(5.03)	(6.25)	(3.87)
Low	*M*	21.67%	48.61%	34.17%	75.00%
	SD	(10.96)	(9.92)	(15.04)	(10.17)

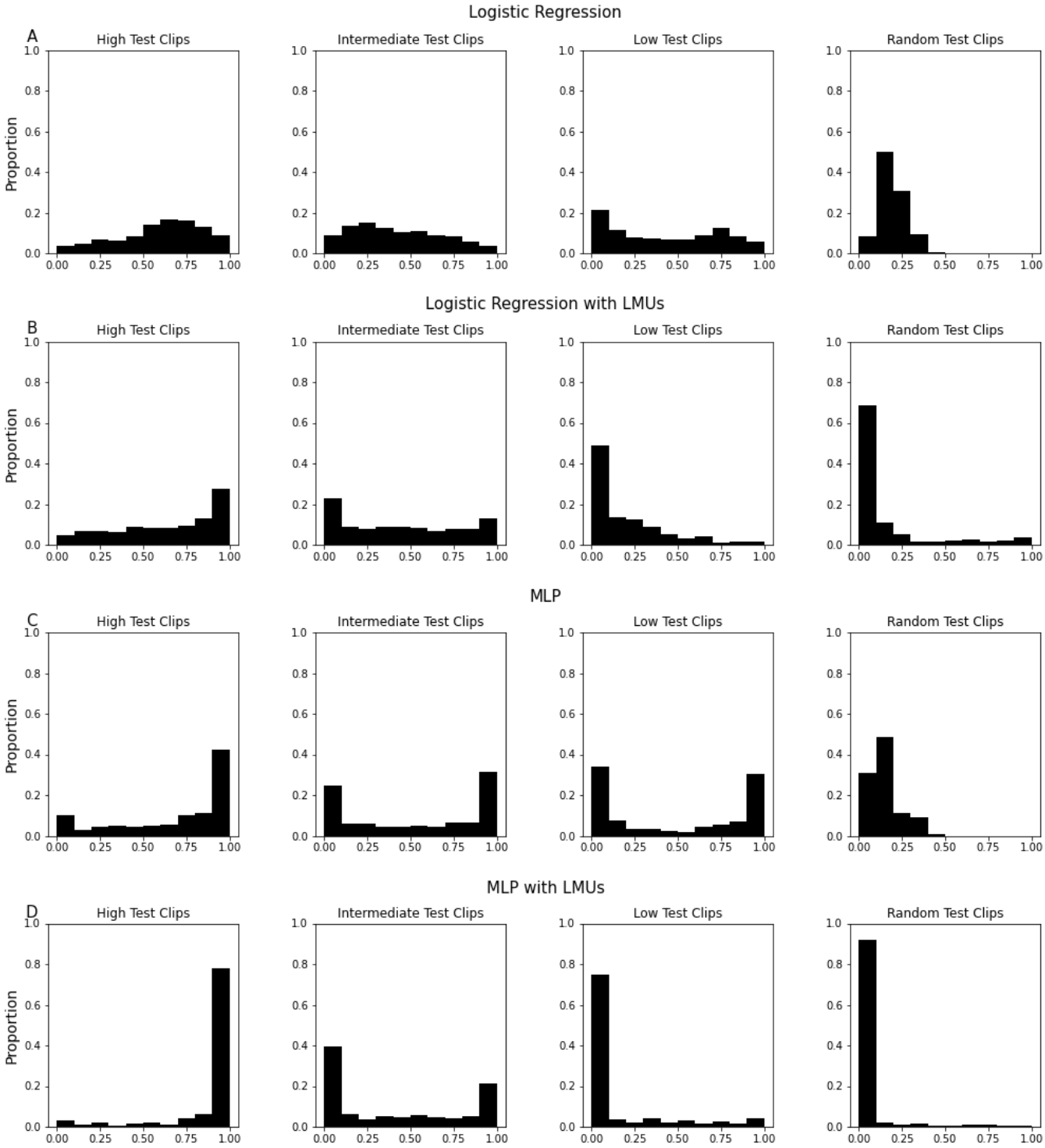


FIGURE 5.10: Histograms showing proportion of estimation values for each clip from all classes (total clips: high = 360, intermediate = 1040, low = 360, random = 360). Classification value of 0 indicates a low engagement classification, and 1 is a high engagement classification. Sub-figures display the results of classification using: (A) LR, (B) LMU-LR, (C) MLP, (D) LMU-MLP.

100% in-line with the class label given. However, these results also demonstrate that this fact can potentially be an advantage for this type of classification/estimation, something which will be explored further in the Discussion.

5.3.3 Separating the Classes

The next step, given the observation that the outputs for each class do appear to be at least somewhat distinct, is to establish whether the outputs for each clip do, in fact, contain enough information for distinguished between all the classes. To examine this, the descriptive statistics of mean, standard deviation, skew and kurtosis of the output given for each clip were calculated, and a simple k-Nearest Neighbours (kNN) classifier was used to test whether these were enough to differentiate between the classes. If successful, this would indicate that the information needed to distinguish between the classes is readily available in the outputs from each approach. This section therefore presents four approaches to classification: (1) LR-kNN, (2) LMU-LR-kNN, (3) MLP-kNN, and (4) LMU-MLP-kNN.

Random vs. Non-Random

This analysis was split into two tests. First it was examined whether the random clips could be distinguished from the engagement clips, regardless of intensity level. The intention here was to explore whether the random data occupied a different region of the four-dimensional space defined by the descriptive statistics, and are therefore not confused with examples of various levels of engagement, even if the system was not trained on some of those levels.

For this analysis a kNN (k=5) was given all 18 random clips from each experiment, along with 18 clips randomly selected from the high, low and intermediate test clips from the same experiment. For each of the four approaches the kNN was run 20 times - once for each experiment - and the results collated so that mean and standard deviation performance could be calculated. Average performance (percent correct) of each approach were as follows: for LR-kNN Mean = 0.929, SD = 0.027, for LMU-LR-kNN Mean = 0.782, SD = 0.099, for MLP-kNN Mean = 0.925, SD = 0.034, and for LMU-MLP-kNN Mean = 0.764, SD = 0.070. Confusion matrices showing the average percent of random and non-random clips identified correctly are presented in Figure 5.12. Overall, we observe good performance on this task but, interestingly, LMU pre-processing tends to result in less accurate performance.

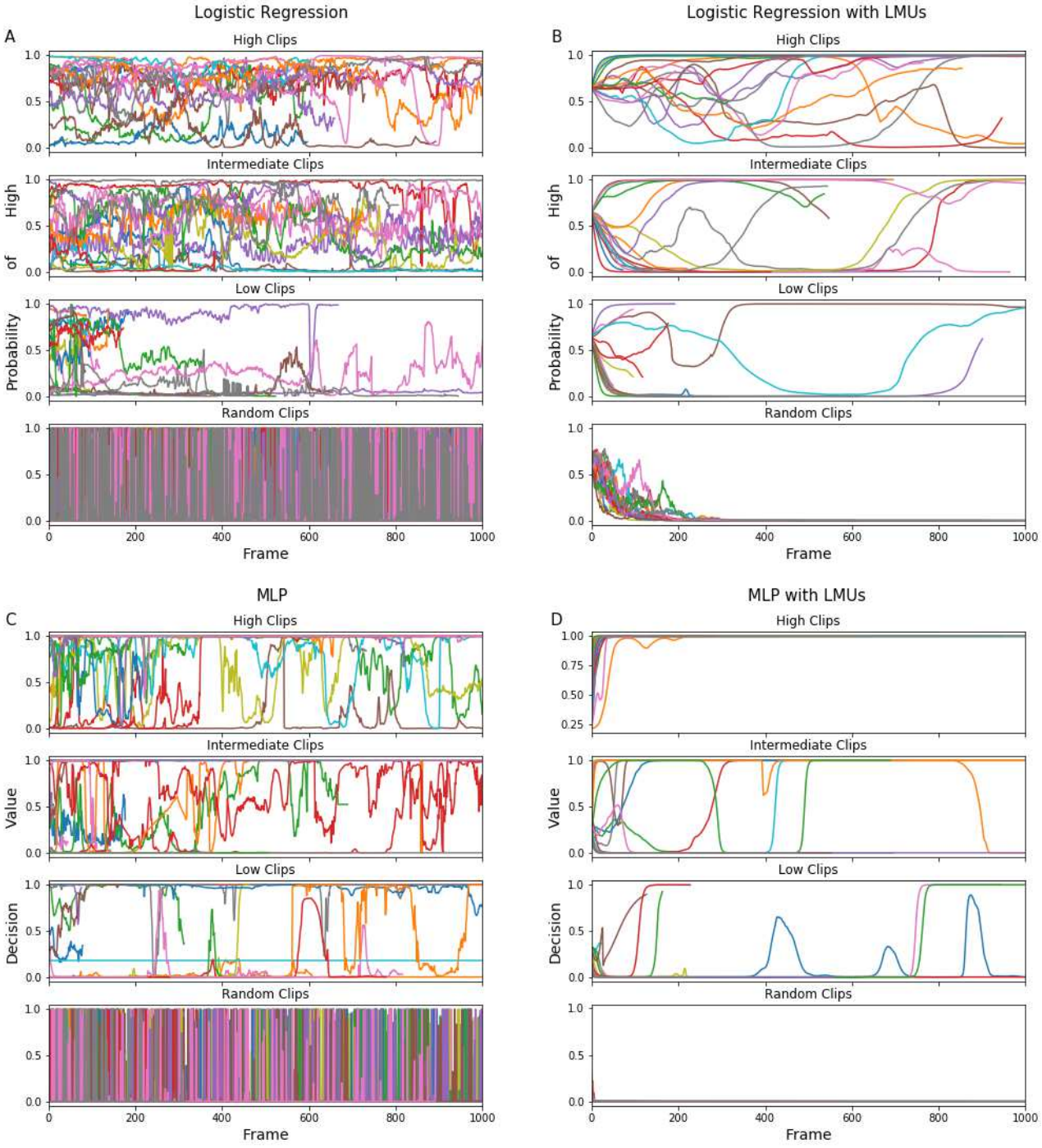


FIGURE 5.11: Plots of the estimation values for each frame of the first 18 clips of each type in the first experiment. Each coloured line represents a clip. (A) LR, (B) LMU-LR, (C) MLP, (D) LMU-MLP. Both the smoothing effect of LMU pre-processing and the different shapes of the timelines for the different classes can be observed.

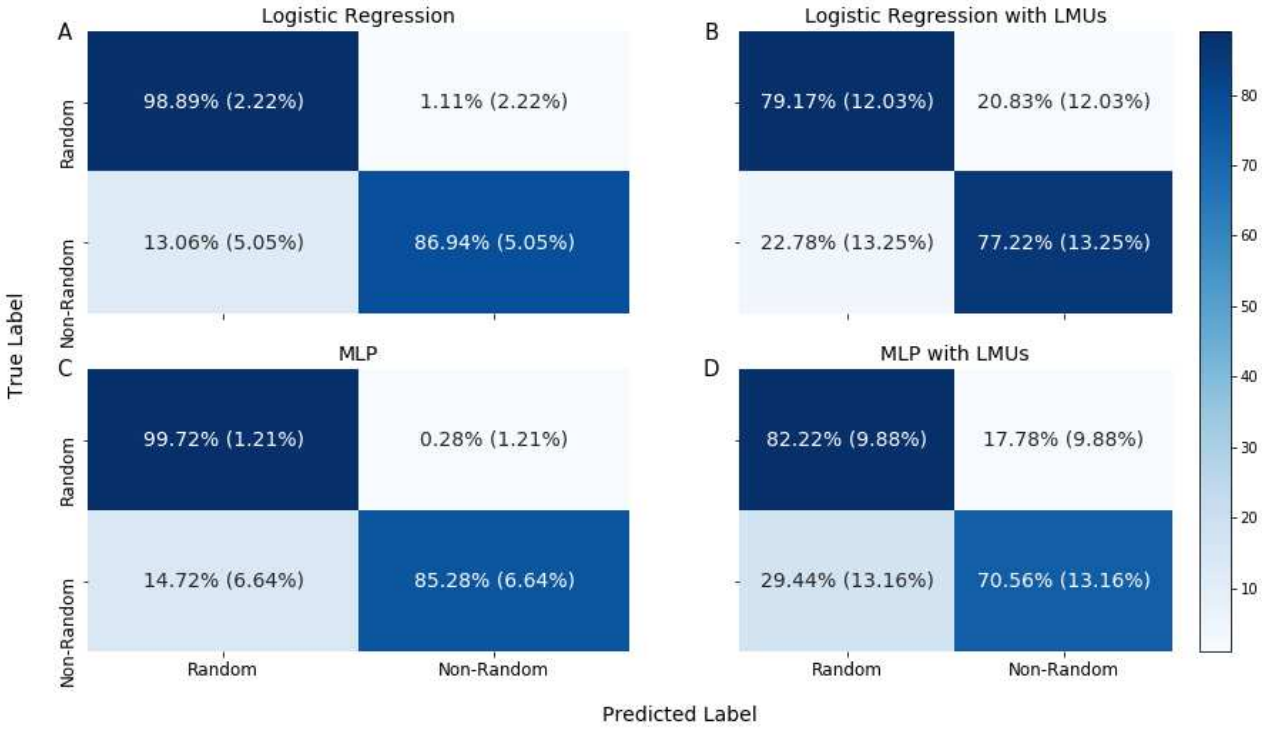


FIGURE 5.12: Average confusion matrices showing mean (and standard deviation) percent of random vs. non-random clips classified correctly by kNN for each approach. (A) LR-kNN, (B) LMU-LR-kNN, (C) MLP-kNN, (D) LMU-MLP-kNN.

High vs. Intermediate vs. Low Engagement

The second analysis looks at how well the three engagement classes could be dissociated based on the four simple descriptive statistics. Success would establish that a clip from an unseen intermediate class could indeed be distinguished from the trained classes even though the system was not trained on this class.

As with the previous analysis, a kNN (k=5) was used to see how useful the descriptive statistics were for separating high, intermediate and low engagement clips from one another. The kNN's were again run 20 times for each approach, once for each experiment. All 18 high engagement clips, and all 18 low engagement clips from each experiment were used as input, along with a random selection of 18 of the intermediate clips. Performance scores (percent correct) and confusion matrices were recorded so that averages could be calculated. The average confusion matrices showing mean percent of high, intermediate and low engagement clips identified correctly can be seen in Figure 5.13. These plots reveal that best overall performance, that is, good performance on all three classes, was achieved

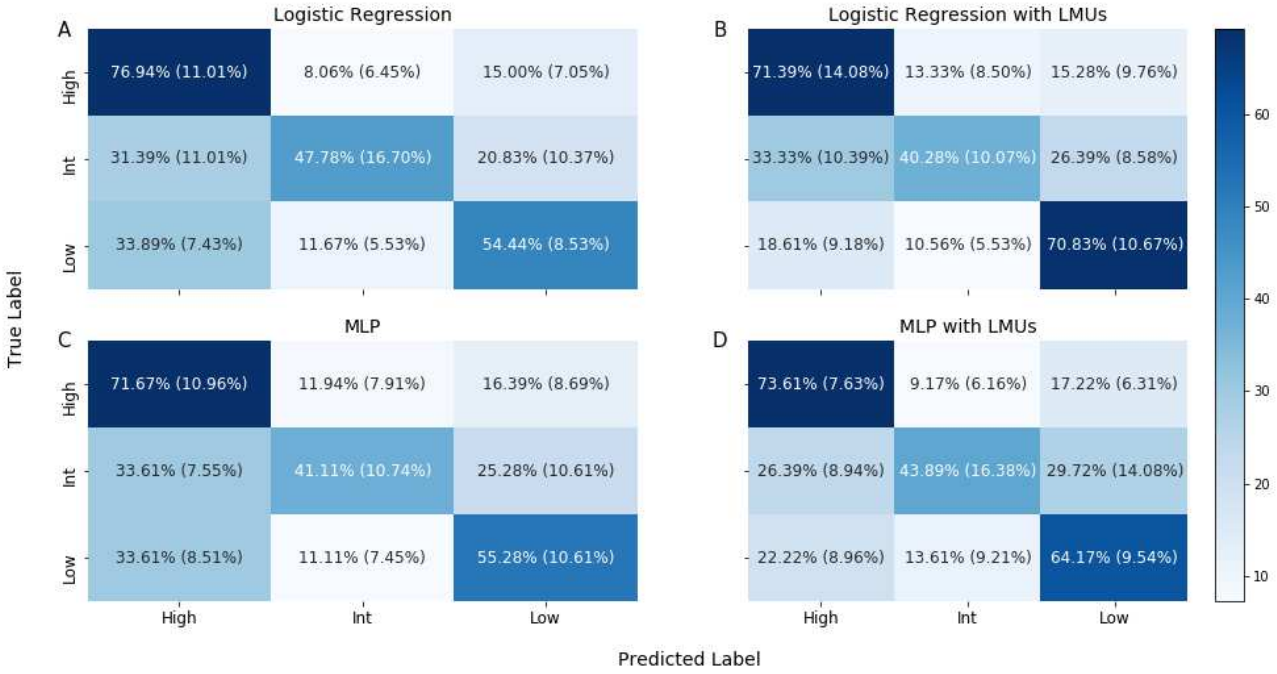


FIGURE 5.13: Average confusion matrices showing mean (and standard deviation) percent of high vs. intermediate vs. low engagement clips classified correctly by kNN for each approach. A: LR-kNN. B: LMU-LR-kNN. C: MLP-kNN. D: LMU-MLP-kNN.

with LMU pre-processing. Specifically, the output from both LR and MLP with LMU pre-processing resulted in good performance by the kNN both on the trained classes (high = 71.39%, low = 70.83% and high = 73.61%, low = 64.17% respectively) and on the untrained intermediate class (40.28% and 43.89% respectively). This conclusion is bolstered by the finding that overall mean performance without LMU pre-processing (LR-kNN: Mean = 0.597, SD = 0.071; MLP-kNN: Mean = 0.560, SD = 0.048) was lower than with LMU pre-processing (LMU-LR-kNN: Mean = 0.608, SD = 0.058; LMU-MLP-kNN: Mean = 0.606, SD = 0.058).

In order to explore the separability of these classes further it is useful to examine how each class clusters in the kNN space. To do this a 3-component PCA was performed on the descriptive statistics of the output data from each approach for the three engagement classes, and the same data was then projected back into the 3D PCA space (which can be more easily visualized than the original 4D space). Figure 5.14 shows that, for both MLP approaches, the clips appear distributed along a string in 3D space, with high and low engagement on either end, and intermediate engagement spanning the length of that line. To examine this further, the 3D strings produced from the data

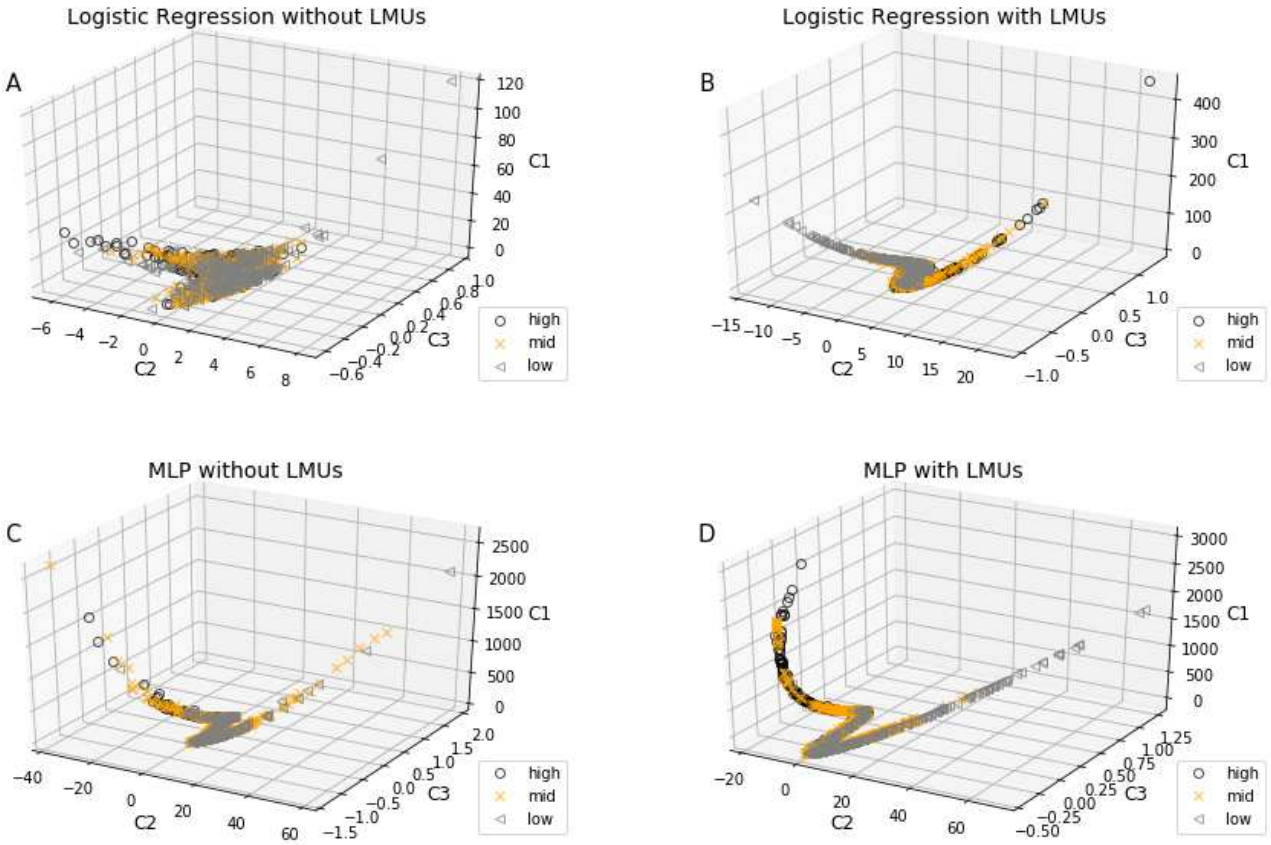


FIGURE 5.14: Engagement clips projected into 3D PCA space. (A) LR-kNN, (B) LMU-LR-kNN, (C) MLP-kNN, (D) LMU-MLP-kNN.

of the MLP approaches are flattened into 1D space, and density estimates for each class are computed in this new space. These transformations reveal that the addition of LMU pre-processing both produces a clearer separation between the two trained engagement classes, and narrows the distribution of the untrained class in a region that sits in-between the peaks of the trained classes (see Figure 5.15).

Finally, in order to establish whether the differences in kNN performance across the four approaches were significant, a two-way ANOVA was conducted with system (LR vs. MLP) and pre-processing step (with vs. without LMUs) as independent variables, and average performance scores (mean accuracy) as the dependent variable. A Shapiro-Wilk test showed that the residuals were normally distributed ($W = 0.987$, $p = 0.584$) and a Bartlett's test showed group variances to be equal ($\chi^2 = 2.916$, $p = 0.405$), indicating that this analysis was appropriate. The two-way ANOVA revealed that there was no significant main effect of system (LR: Mean = 0.603 SD = 0.066; MLP: Mean = 0.583 SD = 0.058) on kNN performance (2-way ANOVA: $F(1, 76) = 2.152$,

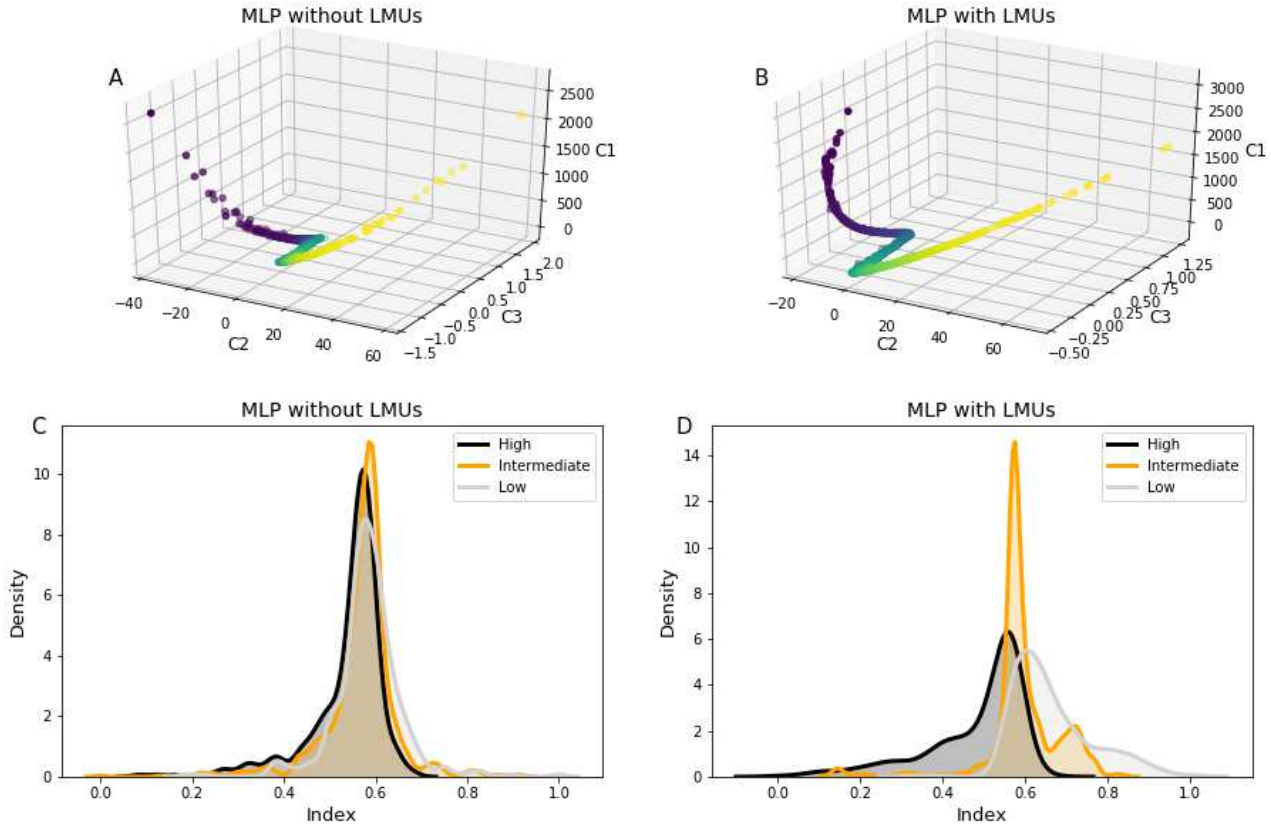


FIGURE 5.15: A & B: Order of clips along a string in 3D PCA space indicated by colour scale from dark to bright. C & D: Density plots of indexed engagement clips along this string. X axis indicates normalised distance from the starting point along the string computed as a cumulative sum of distances between each point and its preceding point.

$p = 0.147$, $\eta_p^2 = 0.026$). However, the effect of pre-processing step was significant, such that the kNN performed better when the data had been pre-processed with LMUs (Mean = 0.607 SD = 0.059) than when it had not (Mean = 0.579 SD = 0.064) (2-way ANOVA: $F(1,76) = 4.331$, $p = 0.040$, $\eta_p^2 = 0.052$). This result provides support for our first hypothesis that the use of LMUs as a pre-processing step would improve the system's performance.

5.4 Discussion

This study was designed to examine two main hypotheses. The first was that the use of the LMU method, presented in Voelker, Kajić, and Eliasmith (2019), as a pre-processing step would facilitate improved performance on systems tasked with classifying levels of task engagement from naturalistic human body movements and postures. This was tested by training a logistic

regression and an MLP on examples of high and low engagement and comparing performance when the input data was pre-processing using LMUs to when the raw data was used. Results showed that LMU pre-processing significantly improved performance such that a higher percentage of clips were classified correctly by a kNN based on the mean output score given for each clip. This provides support for the argument that LMUs are an effective alternative to reservoir computing methods for providing multiple classification labels when there is a paucity of data. Furthermore, it demonstrates that the LMU method, being efficient to implement whilst still achieving good performance on temporally dependent task, may be particularly beneficial in human-computer and human-robot interaction settings.

The second hypothesis was that the systems (MLP and LR), after training on high and low task engagement, would produce an output in response to intermediate task engagement which could be used to identify these samples as being related to, but different from, the trained extremes *without being trained on them*. Testing this hypothesis involved training all four approaches (LR, LMU-LR, MLP, LMU-MLP) with examples of high and low engagement, and then testing using unseen examples from the trained classes as well as two entirely unseen classes: intermediate task engagement and randomly generated data. After testing, the outputs of each approach were used as input for simple kNN classifiers in order to see whether (1) random data could be distinguished from engagement data, and (2) the three engagement classes could be distinguished from one another. The results of these analyses found support for this hypothesis; all mean accuracy scores were >0.75 when separating the random data from non-random data and all mean accuracy scores were >0.54 for distinguishing between the three engagement classes.

Whilst these results are promising, there are a number of interesting features in the results which are worth further exploration. For instance, whilst LMU pre-processing did improve kNN performance when discriminating between the three engagement classes, there is a notable decrease in performance accuracy when distinguishing between random and non-random data. It is likely that this is related to the smoothing effect of LMU pre-processing (as observed in Figure 5.11). That is, reducing the variation in the outputs to random data would have impacted the descriptive statistics which were used as input to the kNN. For example, standard deviation of the outputs to the LMU pre-processed random data were likely much smaller than for the raw random data. Despite this, there are, overall, clear benefits

to using the LMU pre-processing step, particularly in respect to the first hypothesis. Thus these results demonstrate a trade-off (a smoothing based on history which may in some cases lose relevant information) that the use of LMUs entails.

A second finding that should be noted is that, despite the fact that the outputs in response to LMU pre-processed random data was consistently a classification of low engagement, these two classes were not overly confused for one another by the kNNs. One likely explanation for this is that not every frame of the low engagement data, being from a naturalistic data set, would be a perfect example of low engagement. On the other hand, every 'frame' of the random data was an example of random. Interestingly, then, the imperfection of real data is potentially an advantage for this type of classification.

Third, it should be acknowledged that kNN performance was only marginally above chance for the intermediate engagement class. It should be noted that perfect performance on this class was not expected. This is partially because of the use of naturalistic data - it is highly unlikely that all the samples (frames) were 'perfect' examples of their class, and therefore some confusion is to be expected when classifying. Additionally, it is likely that the intermediate engagement class spans a rather large intensity space, with many samples being similar to samples in the trained high and low engagement classes. This argument is potentially supported by the PCA analysis, particularly the plots in Figure 5.14, if we assume that the string in 3D space reflects the intensity continuum. Furthermore, recall that the 'aimlessness' label was hardest to recognize from the movement and posture information for human participants in the study in Chapter 2 (Table 2.5). It was proposed that this might be because the movement and posture features were not sufficient for recognizing this behaviour. Clips labelled with 'aimless play' from the PInSoRo data set were used as the intermediate engagement class. Thus it could be that, whilst the general motor features are sufficient for recognizing high and low task engagement, additional input data, such as eye-gaze, might improve performance on this intermediate engagement class.

5.4.1 Avenues for Future Work

Arguably the most notable finding is that, when the data was pre-processed using LMUs, the outputs of both the logistic regression and MLP could successfully be used to identify a previously unseen class as being intermediate to the two trained extremes. This has a variety of potential repercussions.

First, this study shows that training a system to recognize multiple classes, where the classes are related to one another (i.e. intensities of an internal state) based on observable human behaviours does not necessarily require training on all of those classes or intensity 'levels'. That is, after training on frames from extremes of the task engagement intensity dimension (high and low) it was possible to use the system's output over whole clips to correctly identify whether unseen testing clips were random data, or belonged to one of three levels of engagement (high, intermediate or low). Additionally, this was possible using a simple kNN classifier with basic descriptive statistics of the clips as input. Thus, whilst the performances of each kNN classifier on the intermediate engagement clips were only slightly above chance, these results do show promise that more sophisticated methods may provide even better performance accuracy. So whilst these results are certainly encouraging, more work is needed to establish the reliability of this finding.

Another potential avenue for future work stems from the plots in Figure 5.15. These plots demonstrate that each of the three engagement classes spans a relatively large space even though all of the clips within each class are given the same label. Whilst it is a necessary feature of labelling naturalistic data sets into classes that differences within those classes (e.g. intensity between members of the same class) are lost, the distribution of clips in Figure 5.15 suggests that it might be possible to recover this information. In particular, figures 5.15A and 5.15B suggest an ordering of individual clips within classes. Consequently, one potentially interesting avenue for research would be to verify whether human raters would produce a similar ordering. For example, raters could be presented with pairs of clips and asked to select the 'most engaged' of the two. The resultant ordering could then be compared to the orderings obtained here in order to test the degree to which they capture an actual ordering along a continuum of engagement intensity. Furthermore, these labels could be used to train a more true regression model in order to develop a more accurate and precise model than the methods developed here.

Finally, a third route for future work relates to the human expression of internal states. The results of this study suggest that human movement and body posture can be used to place examples of task engagement along a dimension of intensity (see Figure 5.14). One potential line of questioning, then, is precisely which movements or postures communicate, first this state, and second the intensity of that state. For instance, is the intensity of the internal

state reflected in the intensity of human movements, or are different movements and postures associated with each level of intensity? Answering these questions would be useful not only for understanding human behaviour, and potentially shedding light on the human mind-reading ability, but also in providing more transparency to classification algorithms.

5.5 Conclusion

This chapter presents a study which aimed to answer the fourth and final question proposed in this project:

To what extent can a system recognize intermediate states after training on only the extremes?

In order to answer this question, three systems were compared on the task of classifying three classes of engagement, varying along a dimension of intensity, after training only on the two extreme classes (high vs. low). Of the three systems used, only the logistic regression and MLP proved trainable. This study also introduced the novel LMU method as a pre-processing step in order to transform the input (body position data for each frame of a video clip) such that each frame also contained information about the preceding frames.

The results demonstrate that LMU pre-processing provides an advantage for both logistic regression and MLP systems in this type of task. Furthermore, the LMU-MLP-kNN approach was identified as providing the best overall performance in estimating the untrained intermediate engagement examples without sacrificing performance on the two trained classes. Thus it appears that recognition of untrained classes after training on the extremes of a continuum is feasible, a finding which has important repercussions when designing a system to recognize multiple levels of an internal state, without the associated training data requirements. More work is needed to confirm these results as well as to develop this approach so that intermediate states can be more exactly placed along the intensity continuum.

5.6 Open-Source Resources

The following github repository contains scripts for the experiments, data sets used for analysis, and analysis scripts.

https://github.com/maddybartlett/Thesis_Notebooks/tree/master/Chapter5_LMUs

Chapter 6

General Discussion and Conclusion

This research project set out to explore ways in which artificial systems could be made to recognize non-emotional human internal states in a way which reflects the experience of those states, allows for more accurate classification, and could potentially be applied to the production of appropriate and flexible artificial-agent behaviours in human-computer interactions. Chapter 2 highlighted that states relating to task engagement can be recognized by human observers from just human movement and posture information with a similar degree of accuracy as from the full visual scene. Chapters 4 and 5 report on the evaluations of a number of methods for classifying or estimating task engagement from human movement information. In combination, these studies sought to evaluate the thesis of this research:

> By leveraging the assumption that the experience of internal states can be described along a continuum of intensity, one may be able to train a system to identify a range of 'intensities' without the necessity of labeled training examples from every range of states.

This Chapter discusses how the studies within this project provide answers to the research questions posed in Chapter 1, and thus how they address the over-arching thesis. The impact of this work is then outlined and avenues for future work highlighted.

6.1 Research Questions

In order to examine the thesis of this project a series of four research questions were put forward. Here we will revisit these questions and explain how the work presented in this thesis addresses them.

RQ1 **What representation of internal states best reflects the experience of those states, and may lend itself to the problem of providing flexible and appropriate responses from artificial agents?** Chapter 1 explored definitions and characterizations of human internal states, specifically non-emotional states, and highlighted that they can often be described in terms of 'intensity' (Hess, Blairy, and Kleck, 1997; Cacioppo et al., 1986; Burgoon, Johnson, and Koch, 1998). That is, at any time, a person's experience of a given internal state can vary in terms of how strongly or intensely the state is experienced. For example, it is more accurate to say that a person can experience low levels, medium levels or high levels of confusion, rather than simply stating that they are either confused or not confused. Thus, one representation of internal states which reflects how they are experienced is one which places states along a continuum of intensity. This is in contrast to most representations used in classification which treat internal states as being discrete, such that they are either present or not (Sanghvi et al., 2011; Foster, Gaschler, and Giuliani, 2017; Bosch et al., 2015).

RQ2 **What internal states can be recognized from observable behaviours?** Having established that internal states can be thought of as varying in terms of intensity, and that we wanted to create a system which could reflect this, it was next necessary to identify a selection of internal states which could be recognized from observable human behaviours. The literature review in Chapter 1 revealed that the modality of human movement and posture behaviour has already been demonstrated to be a rich source of such information for human observers (Manera et al., 2011; Okada, Aran, and Gatica-Perez, 2015; Sanghvi et al., 2011). Therefore, Chapter 2 examined which internal states, out of a selection, could be recognized by human observers from such information. The results of this study revealed that states related to task engagement, such as boredom, were recognizable even when the observer was viewing videos containing only movement and posture information. Furthermore, the study presented in Chapter 3 demonstrated that humans are able to interpret the 'intensity' of the task engagement state another person is experiencing. That is, participants viewing both the full-visual scene and the movement-alone versions of videos of children exhibiting task engagement behaviours showed agreement when rating the children in terms of how engaged they were with their task on a 7-point scale of 'Not at all' to 'Highly'.

RQ3 **How successfully can these states be recognized by an artificial system using machine learning methods?** This third research question was addressed by the studies presented in Chapters 4 and 5. In Chapter 4 a Conceptor-based network and a delay network were implemented and trained to identify high and low task engagement from movement data. Whilst the Conceptor-based network showed good performance when trained and tested on examples of high and low engagement, the results from the delay network showed that it was overfitting to the training data. It was proposed that this may have been due to the use of a PCA to reduce the number of input dimensions from 184 to 2. This process may have reduced the quantity of information available to the point of obscuring features which could be useful for differentiating between the classes. Thus, whilst the Conceptor-based network demonstrated that it was indeed possible to train a classifier to recognize states of engagement, there was arguably room for improvement in order to provide a classifier with more detailed input data. A very recent development presented a promising solution; the Legendre Memory Unit (LMU) (Voelker, Kajić, and Eliasmith, 2019).

In Chapter 5 the LMU approach was implemented as a pre-processing step to encode information both about the current frame, and the preceding 3-second's worth of frames. This LMU pre-processed data was then used as input for two separate systems, a logistic regression, and a Multi-Layer Perceptron. Both of these systems were trained on examples of high and low engagement and demonstrated good performance when tested on previously unseen samples from these classes. Furthermore, the LMU pre-processing successfully improved accuracy of both systems compared to when just the original raw data was used.

RQ4 **To what extent can a classifier recognize intermediate states after training only on the extremes?** The delay network presented in Chapter 4 was also tested on intermediate engagement patterns, but unfortunately showed very poor performance due, most likely, to the overfitting. However, the MLP and logistic regression approaches presented in Chapter 5 provided an output in response to the untrained intermediate engagement examples which was shown to be distinct from that produced in response to the two trained classes. Furthermore, the intermediate output was also distinguishable from that produced when

random data was used as input. Thus it was demonstrated that identifying an untrained, intermediate class after training on the extremes along a dimension of intensity is feasible.

6.2 Pushing the State-of-the-Art

6.2.1 Training Requirements

As discussed in Chapter 1 existing approaches to recognizing human internal states often use categorical classification methods and rely on the use of training examples for each class that is to be identified (Foster, Gaschler, and Giuliani, 2017; Wimmer et al., 2008; Daoudi et al., 2018; Whitehill et al., 2014). As a result of these data requirements, classifiers are often limited to very few classification labels. For instance, many approaches use an approach wherein the classifier's output simply states whether or not an internal state is present (Foster, Gaschler, and Giuliani, 2017; Wimmer et al., 2008; Daoudi et al., 2018). Other approaches which do incorporate different 'levels' of an internal state are limited to a binary approach such as in Whitehill et al., 2014 where classifiers were trained to recognize high and low task engagement based on facial expression information. In contrast, the LMU-MLP-kNN approach presented in Chapter 5 included three classification options - high, intermediate or low task engagement. Importantly, this was achieved without requiring training on samples from every class. That is, after data was pre-processed using the LMU method in order to ensure that each 'frame' contained information about the history of the 'clip', the MLP and LR were trained only on high and low engagement. The outputs (a continuous variable) produced by these systems in response to high, low and the untrained intermediate engagement testing data were then used by a kNN for classification. Consequently, one contribution of this work is demonstrating how categorical classification can be achieved without requiring training on every class.

6.2.2 Legendre Memory Units

Additionally, this project has taken a recently developed method - Legendre Memory Units - and applied it to a task unlike any that it had been used for before. LMUs, first introduced in Voelker, Kajić, and Eliasmith (2019), have been tested on non-dynamic classification problems such as the MNIST

digital classification task (Voelker, Rasmussen, and Eliasmith, 2020) and on dynamic problems such as forecasting aortic pressure for clinical purposes (Wang et al., 2020). However, this method has not (prior to the current research) been applied to the recognition of human internal states, or to the classification of data taken from video footage of human behaviour. Thus this project has both extended the potential use-cases for the LMU methodology, and has provided further evidence for its effectiveness in encoding continuous time-series data.

6.3 Future Work

Within this project, there are two main studies which present opportunities for future work. The first is that presented in Chapter 2, wherein it was found that human observers were able to recognize a range of different internal states and social constructs from videos containing only the movements made by children during interactions. Specifically, an EFA analysis revealed that there were three constructs underlying participants' ratings which were translated as Imbalance, Valence and Engagement (IVE). One potential route for future research, therefore, is to validate whether these three constructs are useful for summarising/describing social interactions. If this is the case, this would provide a basic framework for defining social interactions, which in turn could aid in the design of, for example, social robotics, by highlighting some simple concepts which are useful for such a robot to be able to recognize. Verifying this framework can potentially be done in two main steps. The first being to present humans with a range of social interactions and asking them to summarise these interactions in both quantitative and qualitative ways. Participant responses can then be assessed to see whether the IVE constructs emerge from that data. A second step, to explore the usefulness of these constructs, could effectively be a 'matching game', wherein participants are presented with a range of IVE descriptions (i.e. "the people are behaving in *similar* ways, being *positive* and *engaged* with their task"), and videos of social interactions, and asked to match them together.

Second, the results presented in Chapter 5 open up a number of directions for future research. Most immediately, these results should be validated by replication studies, and the final classification approach (LMU-MLP-kNN) assessed for its ability to generalize both to other internal states, and to other populations and contexts. Furthermore, there is certainly room for improvement when it comes to the classification performance. Whilst performance

of the LMU-MLP-kNN approach was above chance on the intermediate engagement class, a large number of samples were still confused for the two trained classes. It is possible that this was a result of not all of the samples being perfect members of their class. However, it may also be that this confusion was due to the limited amount of training data used. Whilst the approaches developed here were intended to overcome potential shortages in data, it cannot be denied that adding more training data would likely improve the classifier's ability to distinguish between classes, even if that training data still only consists of the two 'extreme' classes.

It should also be considered that applying this work to other contexts, internal states and data sets may require that the classification system be extended to include additional cues and data as input. The experiments reported here focused on a relatively structured interaction where the children were fairly stationary in space, with limited opportunities for movements and a common reference point (i.e. the sand-tray). Therefore, applying these methods to other interactions and contexts, specifically more dynamic interactions, will likely require that the system be provided with additional data. For example, in an interaction where children are sharing a toy, or are able to move around more it would likely be useful for the system to have information about the position of objects that the children are interacting with or moving around, as well as about the movements of the children.

Another potential avenue for improvement would be to use a data set specifically annotated for this task. Whilst Chapter 3 demonstrated that the annotation labels available in the PInSoRo data set could be mapped onto levels of engagement, it is likely that having annotators specifically label the data in terms of high, intermediate and low engagement would result in clearer class boundaries. Alternatively, the data could be labelled to better define the intensity continuum, thus providing the opportunity to use regression models. This could be done, for example, by presenting annotators with pairs of clips and asking them to select the 'most engaged' of the two. Eventually this could provide an 'ordering' of the clips which can then be used for regression models or to compare with the ordering of clips in PCA space presented in Figure 5.15. In terms of application, such a regression model could be used to provide specific 'scores' to describe the intensity levels of internal states, or one could introduce cut-off points along the intensity dimension to define as many 'intensity categories' as needed.

Another consideration that could be explored in future works is that of implementing this approach in real-time. A number of human-computer

interaction settings require real-time analysis of the internal states and behaviours of human participants in order to provide appropriate responses from an artificial agent. For example, the field of social robotics is geared towards developing robots which can interact socially with humans autonomously. In such settings, the artificial agent must be able to track and interpret their interaction partner's behaviour in real-time in order to make decisions about what behaviours it, the artificial agent, should perform next. As a simple illustration, when handing over an item to a human, a robot must be able to track the position of the human's hand in order to place their own within an appropriate proximity, and must recognize when the human has securely grasped the object before releasing it from their own grip. In terms of task engagement recognition, for example, it would be useful for a tutor robot to be able to recognize when a student is not engaged with their learning task so that they can appropriately offer encouragements and draw the student's attention back to that task. A series of studies by Blouw et al. (2020) has demonstrated that the LMU method shows promise for real-world applications so it seems that implementing the approach developed here in such a setting is, at least, feasible. Thus future work could explore how viable the methods developed here are for real-time application, and what improvements are needed for this to be successful.

6.4 Potential Applications

Within the field of Human-Computer Interaction there are a range of contexts which can potentially benefit from, or already require, an artificial system able to recognize human internal states. The types of internal states to be recognized, and the usefulness of being able to recognize different intensities of an internal state, differ from case to case. The following section discusses three examples of contexts where the methods developed in this project could prove particularly beneficial: security systems, social robotics and behavioural analysis for diagnosis.

6.4.1 Security Surveillance

One potential application for this technology is in analysing security surveillance footage. In recent years work has been done into developing technologies for automatically analysing behaviours in CCTV (closed-circuit television) footage in order to identify potentially anti-social or illegal activities

(Singh, Singh, and Gupta, 2020; Saveliev, Uzdiaev, and Dmitrii, 2019; Zulkifley et al., 2016; Ditsanthia, Pipanmaekaporn, and Kamonsantiroj, 2018). If the methods developed by this project are extended to internal states such as aggression then they could potentially be used by security technologies in order to identify individuals whose behaviour is aggressive or threatening. One of the difficulties faced when developing classifiers to be used in these settings is the amount of training data required. For example, Saveliev, Uzdiaev, and Dmitrii (2019) constructed a data set of 1086 videos to train and test their networks to recognize aggressive behaviours. Furthermore, much of the work focused on this kind of machine learning application have looked at the classification of only one or a few types of anti-social behaviour. To illustrate, the study by Saveliev, Uzdiaev, and Dmitrii (2019) looked at acts of physical aggression (fights, scuffles) and the manifestation of riots. In contrast, Singh, Singh, and Gupta (2020) used a range of 13 behaviours and anomalies (including abuse, burglary, fighting, shoplifting, road accidents, and vandalism) but required 128 hours of video as the data set. Due to the wide range of potential behaviours and instances which a comprehensive security system would need to recognize, a large data set is unavoidable. Whilst the methods presented in this Thesis do not negate these data requirements, they do potentially reduce the total amount of training data needed for recognizing behaviours which fall along a continuum of intensity. For example, one could potentially train a system on low-intensity and high-intensity physical aggression and then be able to identify intermediate-intensity aggression without training.

This is particularly useful in cases where the classification needs to be performed in real-time in order to alert a user to a potential disturbance, and where the level of intensity alters the required response. To illustrate, say the security system is set up in a shopping centre, if the classifier alerts someone to an instance low-intensity aggression, the user already has an indication of how many responders (i.e. security guards) might be required, which thus enables them to make decisions more quickly. In contrast, if the system did not provide an indication of intensity but simply alerted the user to an aggressive incident, the user would likely take longer to visually assess the situation before alerting security. So, whilst the classifier would not make decisions about the response needed itself, providing a label of intensity could allow whomever is monitoring the activity to more quickly judge what kind of response is required.

6.4.2 Social Robotics

Another context where the methods developed here could be particularly useful is the field of social robotics. As has been pointed out, one of the central goals of human-robot interaction research for social robotics is to develop robots which can interact autonomously with humans (Dautenhahn and Saunders, 2011). This requires that robots are granted a level of 'mind-reading' where they can recognize a human interaction partner's internal state and respond accordingly (Dautenhahn, 2007; Breazeal, Gray, and Berlin, 2009; Vernon, Thill, and Ziemke, 2016; Sciutti et al., 2018).

Whilst the reduced training data requirement is also a potential benefit for this application, a second, and arguably more valuable benefit, is the ability to provide multiple classification labels per state. Recognize multiple 'intensities' of an internal state provides the potential for more flexible and appropriate robot behaviours. That is, consider the most simplistic case where single response options are attached to each detected state. In the case where a robot is only able to recognize whether or not a human partner is experiencing a given state, this leaves the robot with a very limited behavioural repertoire, which will inevitably lead to inappropriate behaviours such as offering clarification when the human partner is not really that confused, or offering no clarification when the interaction partner is somewhat confused. However, if the robot is able to recognize multiple 'levels' of confusion, then the repertoire is much richer, and the appropriateness of the robots responses is likely to be more appropriate.

6.4.3 Behavioural Classification for Diagnosis

A third area where the methods presented in this project could be applied is that of behavioural diagnostics. A large portion of this project was funded by the project DREAM[1], funded by the European Commission[2]. DREAM aimed to develop systems to support therapists in the diagnosis and intervention of Autism Spectrum Disorder (ASD). Consequently, much of this project was guided and informed by these goals.

In particular, selecting human internal states as the focus for classification

[1] www.dream2020.eu
[2] grant number 611391

was partially influenced by the definition of ASD. The Diagnostic and Statistical Manual of Mental Disorders (DSM-V) (APA, 2013) defines Autism Spectrum Disorder (ASD) in terms of difficulties in two behavioural domains: social communication and interaction, and restricted or repetitive behaviours and interests (APA, 2013). Many of the individual behaviours which fall into these domains are covert and therefore rely on human expertise to interpret. For example, one diagnostic trait listed by the DSM-V is a failure to ask for comfort when needed (APA, 2013). In order to identify this trait the observer must be able to recognize whether the individual being assessed is experiencing a state of distress, and whether that distress is severe enough that one would expect them to seek comfort. Thus, having a behaviour classification system able to not only recognize a state of distress, but also able to provide a rating of the intensity of that state would clearly be useful in this setting. Such a system could be used to provide quantitative, objective measures of diagnostic behaviours in order to inform a therapist's decision when making a diagnosis. This idea of providing behaviour classification to aid in the diagnosis of ASD was explored in more depth in Bartlett et al. (2020) (see Appendix B).

6.5 Conclusion

The goal of this work was to develop an artificial system which could classify non-emotional human internal states from observable human behaviours. Based on the findings from an initial study exploring what internal states humans are able to recognize when observing others, it was decided that this work would focus on designing systems to identify task engagement from human body movements and posture information. A series of experiments explored a variety of machine learning approaches including a Conceptor-based approach, a delay network, an MLP and a logistic regression (LR). The final systems (LMU-MLP-kNN and LMU-LR-kNN) were successfully trained to classify high and low intensity task engagement based on movement and posture information extracted from videos of children interacting. Thus this work lends support to the Observability Principle (Becchio et al., 2017) by demonstrating that overt, observable human behaviours can act as cues for recognizing complex, non-emotional human internal states.

Additionally, this project aimed to examine whether the description of internal states as varying in terms of a continuous dimension, in this case an

'intensity' dimension (e.g. from low intensity task engagement to high intensity task engagement), would allow one to create a system able to identify multiple intensities of an internal state after training on only the extremes. That is, would a system trained on high and low intensity task engagement be able to estimate intermediate task engagement without training on that class. With most classical machine learning methods, creating a system to recognize these three intensities of task engagement would require that the system be trained on examples from all of the target classes. However, obtaining such data can be difficult, either due to a lack of available existing data sets, or due to the resources required for creating new data sets. This is particularly true for certain applications, such as developing systems for classifying potentially diagnostic behaviours. Collecting behavioural data in this context often requires recruiting from a limited (and potentially vulnerable) population and involving expert clinicians in the recruitment and collection process. Thus the current project was also an exploration of possible methods for reducing these data requirements in cases where the target internal state can be described in terms of an underlying continuous dimension. In line with this goal, the final systems developed in this project were able to achieve above chance performance on a wholly untrained intermediate intensity task engagement class, after being trained only on high and low intensity task engagement.

In this way, this project has provided support for the main thesis: *"By leveraging the assumption that the experience of internal states can be described along a continuum of intensity, one may be able to train a system to identify a range of 'intensities' without the necessity of labeled training examples from every range of states."* Whilst more work is needed to explore potential ways to improve on the methods presented here, and to test how well the methods can generalize to other populations and internal states, the work did succeed in developing a system which can provide multiple classification labels for a single, complex internal state, without requiring more training data.

Bibliography

Akkaladevi, S. C. et al. (2016). "Human Robot Collaboration to Reach a Common Goal in an Assembly Process." In: *STAIRS*, pp. 3–14.

Alaerts, K. et al. (2011). "Action and emotion recognition from point light displays: an investigation of gender differences." In: *PloS one* 6.6, e20989.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th ed. Washington, DC: Autor.

Ang, J. et al. (2002). "Prosody-based automatic detection of annoyance and frustration in human-computer dialog". In: *Seventh International Conference on Spoken Language Processing*.

Atkinson, A. P. et al. (2004). "Emotion perception from dynamic and static body expressions in point-light and full-light displays." In: *Perception* 33.6, pp. 717–746.

Aviezer, H., Trope, Y., and Todorov, A. (2012). "Body cues, not facial expressions, discriminate between intense positive and negative emotions". In: *Science* 338.6111, pp. 1225–1229.

Barros, P., Weber, C., and Wermter, S. (2015). "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)*. IEEE, pp. 582–587.

Bartlett, M. E., Belpaeme, T., and Thill, S. (2018). "Towards a full spectrum diagnosis of autistic behaviours using human robot interactions". In: *Proc. 1st Workshop on Social Robots in Therapy: Focusing on Autonomy and Ethical Challenges (SREC)*.

Bartlett, M. E., Stewart, T. C., and Thill, S. (2021). "Estimating levels of engagement for social human-robot interaction using Legendre memory units". In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 362–366.

Bartlett, M. E. et al. (2019a). "Recognizing Human Internal States: A Conceptor-Based Approach". In: *arXiv preprint arXiv:1909.04747*.

Bartlett, M. E. et al. (2019b). "What can you see? identifying cues on internal states from the movements of natural social interactions". In: *Frontiers in Robotics and AI* 6, p. 49.

Bartlett, M. E. et al. (2020). "Requirements for Robotic Interpretation of Social Signals "in the Wild": Insights from Diagnostic Criteria of Autism Spectrum Disorder". In: *Information* 11.2, p. 81.

Bartlett, M. S. et al. (2003). "Real time face detection and facial expression recognition: development and applications to human computer interaction." In: *2003 Conference on computer vision and pattern recognition workshop*. Vol. 5. IEEE, pp. 53–53.

Becchio, C. et al. (2017). "Seeing mental states: An experimental strategy for measuring the observability of other minds". In: *Physics of life reviews*.

Benedek, M. et al. (2018). "Are you with me? Probing the human capacity to recognize external/internal attention in others' faces". In: *Visual cognition* 26.7, pp. 511–517.

Beyan, C. et al. (2016). "Detecting emergent leader in a meeting environment using nonverbal visual features only". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, pp. 317–324.

Blouw, P. et al. (2019). "Benchmarking keyword spotting efficiency on neuromorphic hardware". In: *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*, pp. 1–8.

Blouw, P. et al. (2020). "Hardware Aware Training for Efficient Keyword Spotting on General Purpose and Specialized Hardware". In: *arXiv preprint arXiv:2009.04465*.

Bosch, N. et al. (2015). "Automatic detection of learning-centered affective states in the wild". In: *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 379–388.

Bozhkov, L., Koprinkova-Hristova, P., and Georgieva, P. (2016). "Learning to decode human emotions with Echo State Networks". In: *Neural Networks* 78, pp. 112–119.

Breazeal, C., Gray, J., and Berlin, M. (2009). "An embodied cognition approach to mindreading skills for socially intelligent robots". In: *The International Journal of Robotics Research* 28.5, pp. 656–680.

Bryant, G. A. and Barrett, H. C. (2007). "Recognizing intentions in infant-directed speech: Evidence for universals". In: *Psychological Science* 18.8, pp. 746–751.

Burgoon, J. K., Johnson, M. L., and Koch, P. T. (1998). "The nature and measurement of interpersonal dominance". In: *Communications Monographs* 65.4, pp. 308–335.

Cacioppo, J. T. et al. (1986). "Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions." In: *Journal of personality and social psychology* 50.2, p. 260.

Cañamero, L. (2005). "Emotion understanding from the perspective of autonomous robots research". In: *Neural networks* 18.4, pp. 445–455.

Cao, Z. et al. (2017). "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *CVPR*.

Carruthers, P. and Smith, P. K. (1996). *Theories of theories of mind*. Cambridge University Press.

Castellano, G., Villalba, S. D., and Camurri, A. (2007). "Recognising human emotions from body movement and gesture dynamics". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 71–82.

Cavallo, F. et al. (2018). "Emotion modelling for social robotics applications: a review". In: *Journal of Bionic Engineering* 15.2, pp. 185–203.

Clarke, T. J. et al. (2005). "The perception of emotion from body movement in point-light displays of interpersonal dialogue". In: *Perception* 34.10, pp. 1171–1180.

Cohen, I. et al. (2003). "Facial expression recognition from video sequences: temporal and static modeling". In: *Computer Vision and image understanding* 91.1-2, pp. 160–187.

Coulson, M. (2004). "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence". In: *Journal of nonverbal behavior* 28.2, pp. 117–139.

Crane, E. and Gross, M. (2007). "Motion capture and emotion: Affect detection in whole body movement." In: *In International Conference on Affective Computing and Intelligent Interaction.*, pp. 95–101.

Dai, K., Fell, H. J., and MacAuslan, J. (2008). "Recognizing emotion in speech using neural networks". In: *Telehealth and Assistive Technologies* 31, p. 38.

Daoudi, M. et al. (2018). "A new computational approach to identify human social intention in action". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 512–516.

Darwin, C. and Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Dautenhahn, K. (2007). "Socially intelligent robots: dimensions of human–robot interaction". In: *Philosophical transactions of the royal society B: Biological sciences* 362.1480, pp. 679–704.

Dautenhahn, K. and Saunders, J. (2011). *New Frontiers in Human Robot Interaction*. Vol. 2. John Benjamins Publishing.

Defernez, M. and Kemsley, E. K. (1999). "Avoiding overfitting in the analysis of high-dimensional data with artificial neural networks (ANNs)". In: *Analyst* 124.11, pp. 1675–1681.

Ditsanthia, E., Pipanmaekaporn, L., and Kamonsantiroj, S. (2018). "Video Representation Learning for CCTV-Based Violence Detection". In: *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*. IEEE, pp. 1–5.

Ekman, P. and Friesen, W. V. (1971). "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2, p. 124.

Ekman, P., Friesen, W. V., and Ancoli, S. (1980). "Facial signs of emotional experience." In: *Journal of personality and social psychology* 39.6, p. 1125.

Elfaramawy, N. et al. (2017). "Emotion recognition from body expressions with a neural network architecture". In: *Proceedings of the 5th International Conference on Human Agent Interaction*, pp. 143–149.

Eliasmith, C. and Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems.* Cambridge, MA: MIT Press.

Esteban, P. G. et al. (2018). "Proceedings of the Workshop on Social Robots in Therapy: Focusing on Autonomy and Ethical Challenges". In: *arXiv preprint arXiv:1812.07613*.

Fontaine, J. R. J. et al. (2007). "The world of emotions is not two-dimensional". In: *Psychological science* 18.12, pp. 1050–1057.

Foster, M. E., Gaschler, A., and Giuliani, M. (2017). "Automatically classifying user engagement for dynamic multi-party human–robot interaction". In: *International Journal of Social Robotics* 9.5, pp. 659–674.

Gallese, V. and Goldman, A. (1998). "Mirror neurons and the simulation theory of mind-reading." In: *Trends in Cognitive Sciences* 2.12, pp. 493–501.

Gallese, V. et al. (1996). "Action recognition in the premotor cortex." In: *Brain* 119.2, pp. 593–609.

Goldman, A. I. et al. (2012). "Theory of mind". In: *The Oxford handbook of philosophy of cognitive science*, pp. 402–424.

Gopnik, A. and Wellman, H. M. (1994). "10 the theory theory". In: *Mapping the mind: Domain specificity in cognition and culture*, p. 257.

Gopnik, Alison (2003). "The theory theory as an alternative to the innateness hypothesis". In: *Chomsky and his critics*, pp. 238–254.

Grafsgaard, J. et al. (2013). "Automatically recognizing facial expression: Predicting engagement and frustration". In: *Educational Data Mining 2013*.

Hayes, A. F. and Krippendorff, K. (2007). "Answering the call for a standard reliability measure for coding data". In: *Communication methods and measures* 1.1, pp. 77–89.

Hellbernde, N. and Sammler, D. (2016). "Prosody conveys speaker's intentions: Acoustic cues for speech act perception". In: *Journal of Memory and Language* 88, pp. 70–86.

Hernandez, J. et al. (2013). "Measuring the engagement level of TV viewers". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, pp. 1–7.

Hess, U., Blairy, S., and Kleck, R. E. (1997). "The intensity of emotional facial expressions and decoding accuracy". In: *Journal of Nonverbal Behavior* 21.4, pp. 241–257.

Hickton, L., Lewis, M., and Cañamero, L. (2017). "A flexible component-based robot control architecture for hormonal modulation of behaviour and affect". In: *Annual Conference Towards Autonomous Robotic Systems*. Springer, pp. 464–474.

HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (2021). Boulder, CO, USA: Association for Computing Machinery. ISBN: 9781450382908.

Hyun, K., Kim, E., and Kwak, Y. (2006). "Robust speech emotion recognition using log frequency power ratio". In: *2006 SICE-ICASE International Joint Conference*. IEEE, pp. 2586–2589.

Iacoboni, M. and Dapretto, M. (2006). "The mirror neuron system and the consequences of its dysfunction." In: *Nature Reviews Neuroscience.* 7.12, pp. 942–951.

Iacoboni, M. et al. (2005). "Grasping the intentions of others with one's own mirror neuron system". In: *PLoS Biology* 3.3, pp. 0529–0535.

Jaeger, H. (2014a). "Conceptors: an easy introduction." In: *arXiv preprint arXiv:1406.2671*.

— (2014b). "Controlling recurrent neural networks by conceptors". In: *arXiv preprint arXiv:1403.3369*.

— (2017). "Using conceptors to manage neural long-term memories for temporal patterns." In: *Journal of Machine Learning Research* 18.13, pp. 1–43.

Kelley, R. et al. (2008). "Understanding human intentions via hidden markov models in autonomous mobile robots." In: *Proceedings of the 3rd international conference on Human robot interaction - HRI '08*, pp. 367–374.

Kim, J. C. et al. (2017). "Audio-based emotion estimation for interactive robotic therapy for children with autism spectrum disorder". In: *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, pp. 39–44.

Kingma, D. P. and Ba, J. (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kumar, R. (2019). *Machine Learning Quick Reference: Quick and Essential Machine Learning Hacks for Training Smart Data Models*. Packt Publishing Limited.

Lala, D. et al. (2017). "Detection of social signals for recognizing engagement in human-robot interaction". In: *arXiv preprint arXiv:1709.10257*.

Landis, J. R. and Koch, G. G. (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

Latif, M. H. Abd et al. (2015). "Thermal imaging based affective state recognition". In: *2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*. IEEE, pp. 214–219.

Lemaignan, S., Edmunds, C. E. R., and Belpaeme, T. (2017). *The PInSoRo dataset*. DOI: 10.5281/zenodo.1043507.

Lemaignan, S. et al. (2018). "The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics." In: *PloS one* 13.10, e0205999.

Lewis, M. (2008). "Self-conscious emotions: Embarrassment, pride, shame, and guilt." In:

Lewkowicz, D. et al. (2013). "Reading motor intention through mental imagery." In: *Adaptive Behavior* 21.5, pp. 315–327.

Li, Y. and Zhao, Y. (1998). "Recognizing emotions in speech using short-term and long-term features". In: *Fifth International Conference on Spoken Language Processing*.

Liscombe, J., Hirschberg, J. B., and Venditti, J. J. (2005). "Detecting certainness in spoken tutorial dialogues". In:

Litman, D. and Forbes, K. (2003). "Recognizing emotions from student speech in tutoring dialogues". In: *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, pp. 25–30.

Liu, Y. H. (2017). *Python Machine Learning By Example*. Packt Publishing Ltd.

Liu, Z. et al. (2017). "A facial expression emotion recognition based human-robot interaction system". In:

Lord, C. et al. (2000). "The Autism Diagnostic Observation Schedule–Generic: A Standard Measure of Social and Communication Deficits Associated

with the Spectrum of Autism." In: *Journal of Autism and Developmental Disorders* 30.3.

Maheswari, J. P. (2018, December 21). *Breaking the curse of small datasets in Machine Learning: Part 1 [Web log post]*. https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d, Last accessed on 02-09-2020.

Manera, V. et al. (2010). "Inferring intentions from biological motion: a stimulus set of point-light communicative interactions." In: *Behavior research methods* 42.1, pp. 168–178.

Manera, V. et al. (2011). "Cooperation or competition? Discriminating between social intentions by observing prehensile movements". In: *Experimental Brain Research* 211.3-4, pp. 547–556.

McDaniel, B. et al. (2007). "Facial features for affective state detection in learning environments". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 29. 29.

Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Mici, L., Hinaut, X., and Wermter, S. (2016). "Activity recognition with echo state networks using 3D body joints and objects category". In:

Mok, B. et al. (2015). "Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles". In: *2015 IEEE 18th international conference on intelligent transportation systems*. IEEE, pp. 2458–2464.

Mok, B. et al. (2017). "Tunneled in: Drivers with active secondary tasks need more time to transition from automation". In: *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 2840–2844.

Moors, A. (2009). "Theories of emotion causation: A review". In: *Cognition and emotion* 23.4, pp. 625–662.

Nicolle, J. et al. (2012). "Robust continuous prediction of human emotions using multiscale dynamic cues". In: *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 501–508.

Ojala, Markus and Garriga, Gemma C (2010). "Permutation tests for studying classifier performance". In: *Journal of Machine Learning Research* 11.Jun, pp. 1833–1863.

Okada, S., Aran, O., and Gatica-Perez, D. (2015). "Personality trait classification via co-occurrent multiparty multimodal event discovery". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pp. 15–22.

Palinko, O. et al. (2016). "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 5048–5054.

Petrushin, V. A. (2000). "Emotion recognition in speech signal: experimental study, development, and application". In: *Sixth International Conference on Spoken Language Processing*.

Pieters, M. and Wiering, M. (2017). "Comparison of Machine Learning Techniques for Multi-label Genre Classification". In: *Benelux Conference on Artificial Intelligence*. Springer, pp. 131–144.

Pollick, F. E. et al. (2001). " Perceiving affect from arm movement." In: *Cognition* 82.2, B51–B61.

Poznyak, T., Oria, J. I. C., and Poznyak, A. (2018). *Ozonation and Biodegradation in Environmental Engineering: Dynamic Neural Network Approach*. Elsevier.

Quesque, F. et al. (2013). "Effects of social intention on movement kinematics in cooperative actions". In: *Frontiers in neurorobotics* 7, p. 14.

Rasmussen, D. (2019). "NengoDL: Combining deep learning and neuromorphic modelling methods". In: *Neuroinformatics* 17.4, pp. 611–628.

Rebala, G., Ravi, A., and Churiwala, S. (2019). *An introduction to machine learning*. Springer.

Rizzolatti, G. and Craighero, L. (2004). "The mirror-neuron system". In: *Annu. Rev. Neurosci.* 27, pp. 169–192.

Rosis, F. de et al. (2007). "'You are Sooo Cool, Valentina!'Recognizing Social Attitude in Speech-Based Dialogues with an ECA". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 179–190.

Rudovic, O. et al. (2018). "Personalized machine learning for robot perception of affect and engagement in autism therapy". In: *Science Robotics* 3, p. 19.

Russell, J. A. (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161.

Saha, S. et al. (2014). "A study on emotion recognition from body gestures using Kinect sensor". In: *2014 International Conference on Communication and Signal Processing*. IEEE, pp. 056–060.

Sanchez-Cortes, D. et al. (2011). "A nonverbal behavior approach to identify emergent leaders in small groups". In: *IEEE Transactions on Multimedia* 14.3, pp. 816–832.

Sanghvi, J. et al. (2011). "Automatic analysis of affective postures and body motion to detect engagement with a game companion". In: *Proceedings of*

the 6th international conference on Human-robot interaction. ACM, pp. 305–312.

Saveliev, A., Uzdiaev, M., and Dmitrii, M. (2019). "Aggressive Action Recognition Using 3D CNN Architectures". In: *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, pp. 890–895.

Scherer, S. et al. (2008). "Real-time emotion recognition from speech using echo state networks". In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, pp. 205–216.

Sciutti, A. et al. (2018). "Humanizing human-robot interaction: On the importance of mutual understanding". In: *IEEE Technology and Society Magazine* 37.1, pp. 22–29.

Shanton, K. and Goldman, A. (2010). "Simulation theory". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1.4, pp. 527–538.

Singh, V., Singh, S., and Gupta, P. (2020). "Real-Time Anomaly Recognition Through CCTV Using Neural Networks". In: *Procedia Computer Science* 173, pp. 254–263.

Song, K., Han, M., and Wang, S. (2014). "Speech signal-based emotion recognition and its application to entertainment robots". In: *Journal of the Chinese Institute of Engineers* 37.1, pp. 14–25.

Sorower, M. S. (2010). "A literature survey on algorithms for multi-label learning". In: *Oregon State University, Corvallis* 18.

Thomas, C. and Jayagopi, D. B. (2017). "Predicting student engagement in classrooms using facial behavioral cues". In: *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*, pp. 33–40.

Tian, L., Moore, J. D., and Lai, C. (2015). "Emotion recognition in spontaneous and acted dialogues". In: *2015 International conference on affective computing and intelligent interaction (ACII)*. IEEE, pp. 698–704.

Tracy, J. L. and Robins, R. W. (2008). "The nonverbal expression of pride: evidence for cross-cultural recognition." In: *Journal of personality and social psychology* 94.3, p. 516.

Tracy, J. L., Robins, R. W., and Tangney, J. P. (2007). *The self-conscious emotions: Theory and research*. Guilford Press.

Trentin, E., Scherer, S., and Schwenker, F. (2015). "Emotion recognition from speech signals via a probabilistic echo-state network". In: *Pattern Recognition Letters* 66, pp. 4–12.

VandenBos, G. R. (2007). *APA dictionary of psychology.* American Psychological Association.

Vernon, D., Thill, S., and Ziemke, T. (2016). "The role of intention in cognitive robotics". In: *Toward Robotic Socially Believable Behaving Systems-Volume I*. Springer, pp. 15–27.

Voelker, A., Kajić, I., and Eliasmith, C. (2019). "Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks". In: *Advances in Neural Information Processing Systems*, pp. 15544–15553.

Voelker, A., Rasmussen, D., and Eliasmith, C. (2020). "A spike in performance: Training hybrid-spiking neural networks with quantized activation functions". In: *arXiv preprint arXiv:2002.03553*.

Voelker, A. R. and Eliasmith, C. (2018). "Improving Spiking Dynamical Networks: Accurate Delays, Higher-Order Synapses, and Time Cells". In: *Neural Computation* 30.3, pp. 569–609. DOI: 10.1162/neco_a_01046.

Wang, R. et al. (2020). "Aortic Pressure Forecasting with Deep Sequence Learning". In: *arXiv preprint arXiv:2005.05502*.

Wendt, C. et al. (2008). "Physiology and HRI: Recognition of over-and underchallenge". In: *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, pp. 448–452.

Whitehill, J. et al. (2014). "The faces of engagement: Automatic recognition of student engagementfrom facial expressions". In: *IEEE Transactions on Affective Computing* 5.1, pp. 86–98.

Wimmer, M. et al. (2008). "Facial expression recognition for human-robot interaction–a prototype." In: *International Workshop on Robot Vision*. Springer, pp. 139–152.

Yan, L. et al. (2019). "Human-Robot Collaboration by Intention Recognition using Deep LSTM Neural Network". In: *2019 IEEE 8th International Conference on Fluid Power and Mechatronics (FPM)*. IEEE, pp. 1390–1396.

Zhang, T. et al. (2018). "Spatial–temporal recurrent neural network for emotion recognition". In: *IEEE transactions on cybernetics* 49.3, pp. 839–847.

Zulkifley, M. A. et al. (2016). "Kalman filter-based aggressive behaviour detection for indoor environment". In: *Information Science and Applications (ICISA) 2016*. Springer, pp. 829–837.

Appendix A

Distribution of classification values of testing clips from each of the 20 experiments using the Delay Network presented in Chapter 4

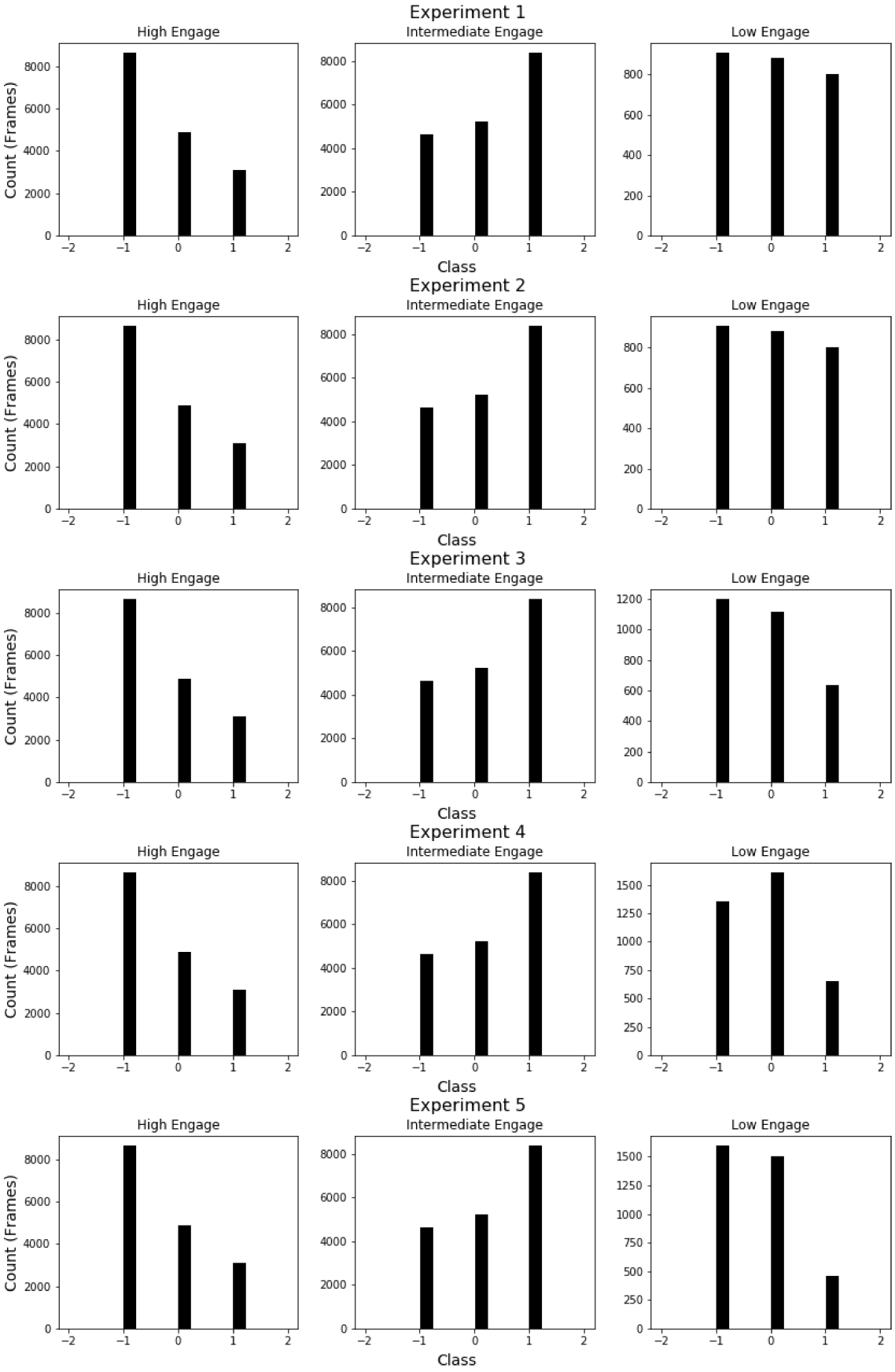


FIGURE A.1: Distribution of classification values of testing clips (experiments 1-5)

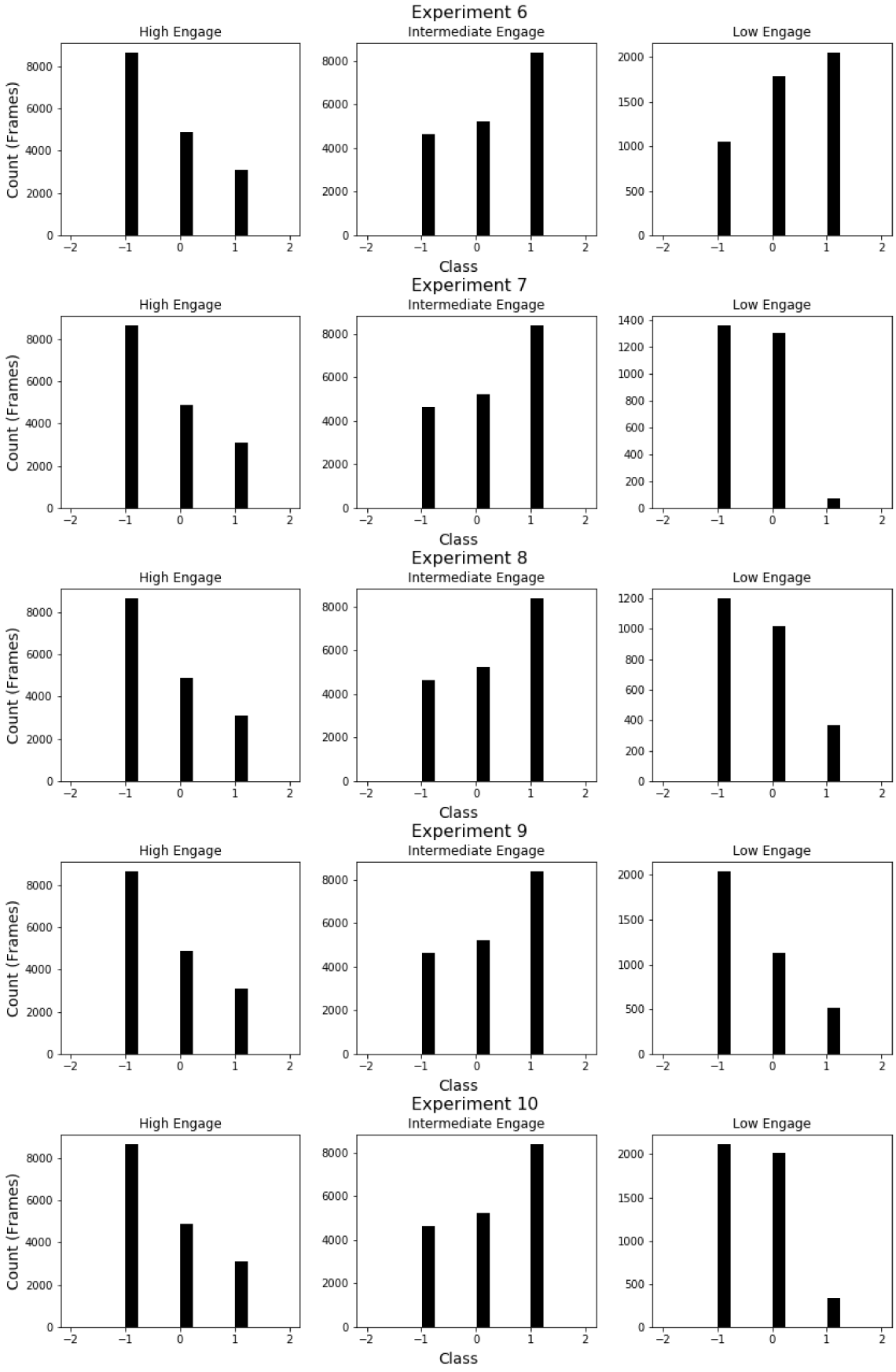


FIGURE A.2: Distribution of classification values of testing clips (experiments 6-10)

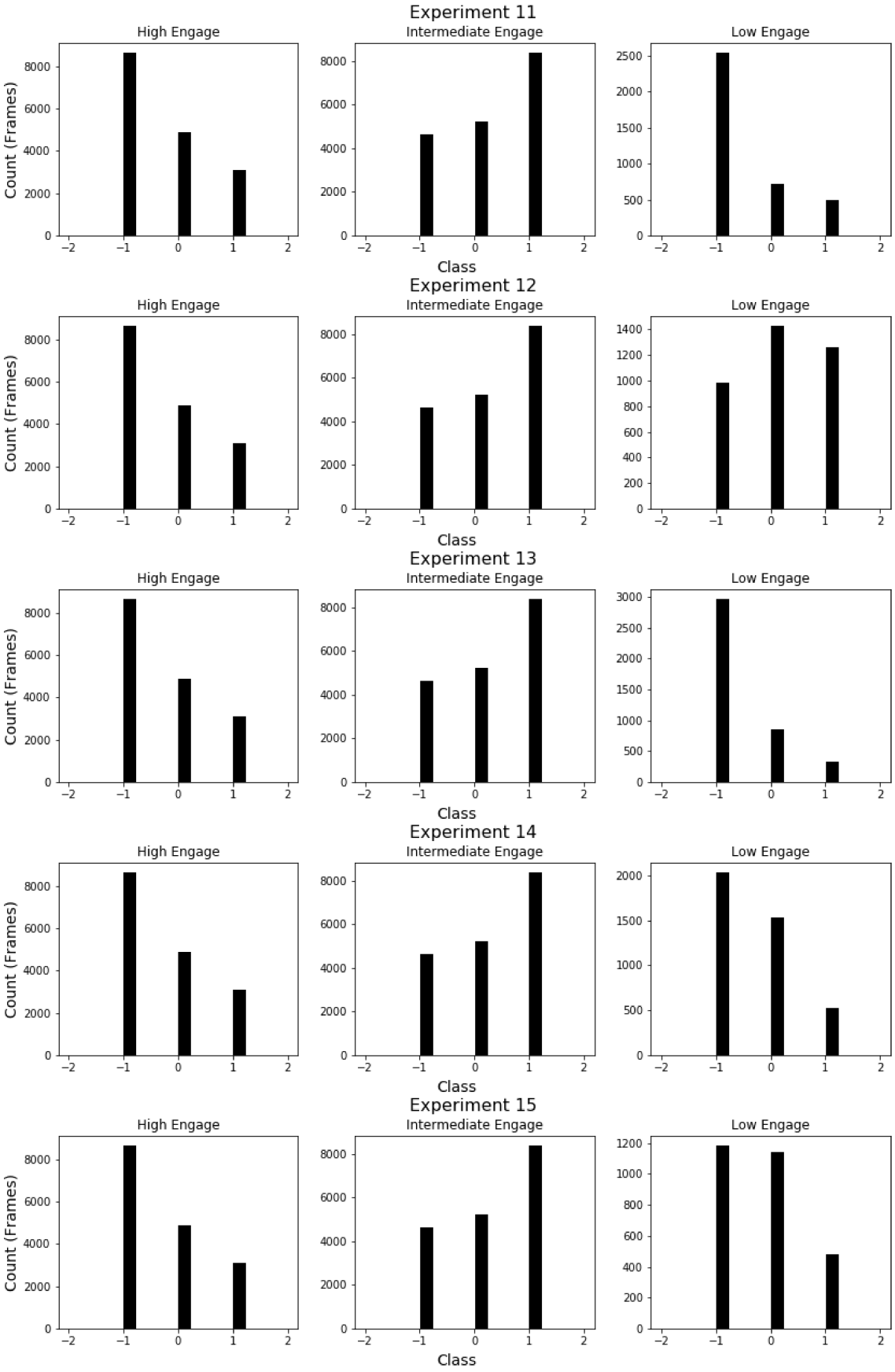


FIGURE A.3: Distribution of classification values of testing clips (experiments 11-15)

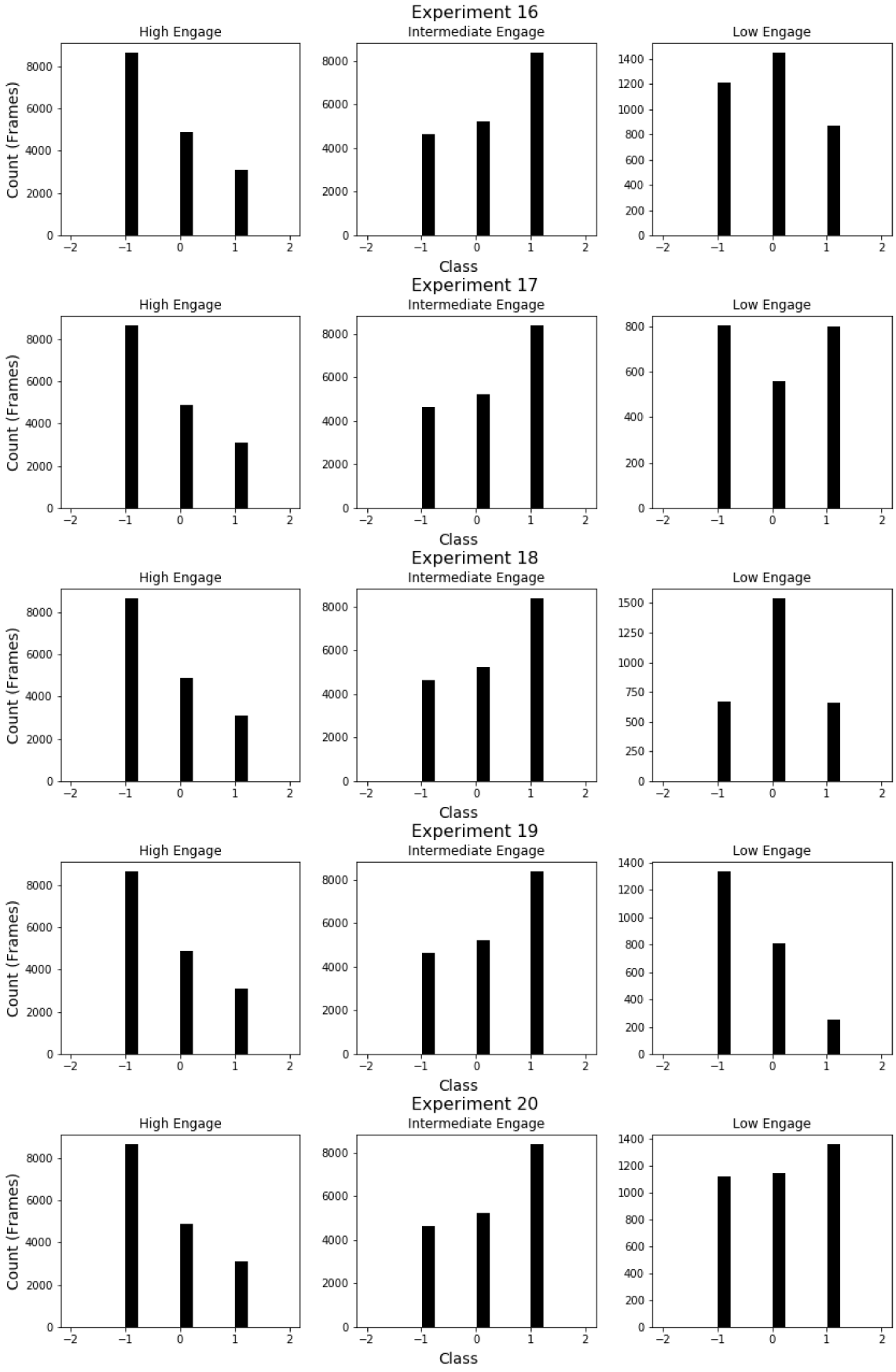


FIGURE A.4: Distribution of classification values of testing clips (experiments 16-20)

Appendix B

Information Article - Requirements for Robotic Interpretation of Social Signals ”in the Wild“: Insights from Diagnostic Criteria of Autism Spectrum Disorder

information

MDPI

Article

Requirements for Robotic Interpretation of Social Signals "in the Wild": Insights from Diagnostic Criteria of Autism Spectrum Disorder

Madeleine E. Bartlett [1,*], Cristina Costescu [2], Paul Baxter [3] and Serge Thill [4,5]

1 Centre for Robotics and Neural Systems, University of Plymouth, Plymouth PL4 8AA, UK
2 Department of Clinical Psychology and Psychotherapy, Babes-Bolyai University, Cluj-Napoca 400000, Romania; christina.costescu@gmail.com
3 Lincoln Centre for Autonomous Systems, School of Computer Science, University of Lincoln, Lincoln LN6 7TS, UK; pbaxter@lincoln.ac.uk
4 Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, 6525 XZ Nijmegen, The Netherlands; s.thill@donders.ru.nl
5 Interaction Lab, School of Informatics, University of Skövde, 541 28 Skövde, Sweden
* Correspondence: madeleine.bartlett@plymouth.ac.uk

Received: 12 December 2019; Accepted: 30 January 2020; Published: 1 February 2020

Abstract: The last few decades have seen widespread advances in technological means to characterise observable aspects of human behaviour such as gaze or posture. Among others, these developments have also led to significant advances in social robotics. At the same time, however, social robots are still largely evaluated in idealised or laboratory conditions, and it remains unclear whether the technological progress is sufficient to let such robots move "into the wild". In this paper, we characterise the problems that a social robot in the real world may face, and review the technological state of the art in terms of addressing these. We do this by considering what it would entail to automate the diagnosis of Autism Spectrum Disorder (ASD). Just as for social robotics, ASD diagnosis fundamentally requires the ability to characterise human behaviour from observable aspects. However, therapists provide clear criteria regarding what to look for. As such, ASD diagnosis is a situation that is both relevant to real-world social robotics and comes with clear metrics. Overall, we demonstrate that even with relatively clear therapist-provided criteria and current technological progress, the need to interpret *covert* behaviour cannot yet be fully addressed. Our discussions have clear implications for ASD diagnosis, but also for social robotics more generally. For ASD diagnosis, we provide a classification of criteria based on whether or not they depend on covert information and highlight present-day possibilities for supporting therapists in diagnosis through technological means. For social robotics, we highlight the fundamental role of covert behaviour, show that the current state-of-the-art is unable to characterise this, and emphasise that future research should tackle this explicitly in realistic settings.

Keywords: autism spectrum disorder; diagnosis; technology; behaviour

1. Introduction

Having robots engage socially with humans is a desirable goal for social robotics. It lowers the barrier to entry into interactions, as it allows the humans to engage and interact with the robot in a way similar to how they would interact with another human. This would remove the need for any specialist robotics knowledge or training for the human users, and thus substantially expands the application domains for social robots beyond the current largely restricted environments in which they are currently used. However, there remain a range of fundamental challenges to being able to

achieve this. Principal among these is that in order to behave appropriately, it is necessary for the robot to *understand* what its human interaction partner is doing (and indeed what they may do). Apart from current limitations in sensory detection technologies (which are improving), the problem remains that essentially the robot observer can only have information about observable (overt) behaviour, but has no access to the mental states (or covert aspects of behaviour) that led to these overt behaviours – this requires further inference. This fundamental challenge for social robotics is the topic of this contribution: we characterise the current state of the art with respect to this problem, synthesising advances across a range of technology disciplines, and highlighting where further technological advances can be most usefully made.

1.1. Recognising Human Internal States from Observable Kinematics in Social Robotics

The ability to infer the mental states of other agents is a fundamental component of social interaction. In humans, this ability is called "Theory of Mind". The exact mechanisms underlying it remain unclear; some hypotheses center around an ability to create folk-psychological models of other minds while others suggest that internal simulation mechanisms normally used to control one's own behaviour can be used to understand and predict the behaviours of others from observation [1,2]. In robotics, the latter, along with its connections to mirror neurons, has long inspired, for example, forms of imitation learning and action understanding that rely on the robot's own forward and inverse kinematic models [3–6].

That said, merely predicting the outcome of actions is not the same as understanding internal mental states from observable kinematics. The latter is seen as a pre-requisite for truly social robotics, yet remains a challenge [7]. While we will give a brief overview of relevant work in the sections below, much current work in social robotics does not address this directly but focuses on, among others, characterising end user requirements in specific applications [8] or studying the degree to which phenomena known from social sciences are applicable to human-robot interactions [9]. It is noteworthy that relatively little is actually required of the robots themselves in such studies, and a Wizard-of-Oz control paradigm is sufficient. Applications of social robots that do require the robot to possess at least some autonomous behaviour exist, for example in education [10] or robot-assisted therapy for disorders such as Autism Spectrum Disorder [11–13], but these are still relatively narrow domains within social robotics.

Overall, there is relatively little research that directly investigates the degree to which the state of the art currently allows social robots in the more general sense. At the same time, this is a timely question since, as we will discuss in this paper, technological progress in recent years does allow for relatively comprehensive observation of human agents in the environment and, together with advances in data analysis (for example, using deep networks) is at a point where it might be feasible to advance in this direction as well.

In this paper we evaluate this technological progress and the degree to which it fulfils the needs of social robots that would exist "in the wild", and not constrained to narrow domains. To perform such an evaluation requires a scenario that captures the essential requirements for social robotics. Here, we focus on the automation of the *diagnosis* of Autism Spectrum Disorder (ASD) for this purpose. We will detail this problem domain further below, but it is important to note that it is distinct from using robots in ASD *therapy*: indeed, diagnosis, in principle, does not even require a robot. On the other hand, diagnosing ASD does require the ability to observe social interactions and infer underlying mental states, which is the core requirement for social robots that we are interested in here. It is also a domain for which clear protocols, assessment criteria and so on exist. For our purposes, this is a crucial advantage over other social contexts because it provides us with the ability to evaluate the degree to which technology can meet these criteria. It is also worth noting that the automation of ASD diagnosis is in itself a relevant research topic; not to replace the clinical therapists involved, but to support them: as we will see below, the process is rather intensive but opportunities for alleviating the burden exist.

In the remainder of this introduction, we first describe ASD and diagnostic criteria. We then break these down into different categories, based on whether they focus just on the behaviour of the child or on the interaction itself, and whether they concern the assessment of overt or covert information. We then discuss the degree to which technological means can fulfil these requirements.

1.2. Diagnosing Autism Spectrum Disorder

ASD is characterised by the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [14] using two categories of behaviour: social communication difficulties and restricted or repetitive behaviour patterns. Since the identification of ASD [15,16], the literature has examined potential causes, intervention techniques and approaches to diagnosis. These investigations have revealed ASD to be a complex developmental disorder with high levels of heterogeneity within the clinical population in terms of symptom presentation and severity [17]. Furthermore, there are no biologically based tests for ASD [18]. As such, the diagnosis of ASD remains a very difficult task, relying on the interpretation of current and retrospective observations of an individual's behaviour, and of developmental aspects, by different specialists including psychologists, psychiatrists and speech therapists [19,20]. These observational judgements are then quantified according to standard protocols such as the Diagnosis Interview Revised (ADI-R) [21], the Childhood Autism Rating Scale (CARS) [22], and the Autism Diagnostic Observation Schedule Generic (ADOS-G) [23].

Despite the efforts made thus far to improve and standardise the diagnostic process (via the tools listed above), the variable nature of ASD and the emergence of symptoms in early childhood [24] amid ongoing developmental changes does cause difficulties for its identification and diagnosis [18]. While standardisation of the diagnostic process via tools such as those above has been effective in aiding clinicians in this task [20,25], there is room for improvement. In particular, surveys asking parents about the process of getting an ASD diagnosis for their child found that even though parents first seek a diagnosis when their child is aged 3.9 years (on average), a final diagnosis is not received until the child is 7.5 years. Consequently, one way in which the diagnostic process could be improved would be to reduce the time taken from when parents first seek a diagnosis to when a final diagnosis is received [26].

One way to address this would be to provide protocols which are easier to implement, and able to produce useful information without over-reliance on human expertise and thereby provide General (GPs) and other practitioners with the means to make more informed decisions about when to refer a patient for expert diagnosis. It is important to note that we do not propose to replace the assessments carried out by expert clinicians, but rather to make the process of accessing these assessments easier, cheaper and more efficient. We propose that technologies able to provide useful information about an individual's diagnostic status could contribute to achieving this goal.

Technical advances have long inspired research into how technologies can be applied to diagnostic scenarios, a method referred to in the medical field as Computed-Aided Diagnosis (CAD) [27]. These applications have various motivations including improving the objectivity of decision-making or measurement [28] and incorporating information into the diagnostic process that is more readily detected, measured or used by computers than humans alone [27]. While such techniques were applied to physiological maladies, the advent of technologies and methods for measuring human behaviours, e.g., via machine-perception-guided technologies, has created opportunities for augmenting the diagnosis of behavioural and psychological disorders such as ASD.

2. Observable Behavioural Cues

The first step in augmenting the diagnosis of ASD with technology is to identify whether there are any diagnostic markers that existing technologies can measure and quantify in a meaningful way. To do this we must first identify symptoms that have been sufficiently operationalised to provide objective definitions. Arguably, the DSM and existing diagnostic tools provide such definitions. Support for this claim comes from tests of the reliability and objectivity of these definitions via Inter-Rater Agreement

(IRA). Several studies have looked at IRA between clinicians on the items included in diagnostic tools. While evidence shows that IRA for observational judgements is typically low [29], recent studies examining IRA for clinicians' diagnostic evaluations using the ADI-R [30] and ADOS [31] tools (whose symptom definitions are based on those provided by the DSM) have demonstrated high levels of agreement for each behavioural marker outlined by each tool. These findings demonstrate that the DSM has successfully operationalised the diagnostic characteristics of ASD. As such, we believe that these definitions may provide enough information to propose quantifiable definitions that do not overly rely on human interpretation. If this is the case, it should be possible to apply computational and technological methods to their identification. Our discussion will revolve around which ASD behaviours can be considered overtly observable and can thus be identified with minimal or no reliance on human interpretation. In other words, we identify behaviours that can be tracked, measured and described by technological means.

The restrictive nature of diagnostic settings and the fact that many of the characteristics of ASD are defined by their persistence across time and different interactions (hereafter: "persistent behaviours"; e.g., "reduced sharing of interests" [14] would need to be present across multiple interactions) poses problems for temporally confined diagnostic sessions. To overcome these problems, many diagnostic tools require clinicians to observe and make judgements based on behaviours that are associated with these persistent behavioural traits (hereafter: "indicative behaviours"). For example, it was found that impairments in the perception of facial and body gestures is related to, and may be the foundation of, difficulties in social communication and intention understanding [32]. Similarly, abnormal visual processing of social information from faces [33] and impairments in visual engagement [34] have been linked with difficulties in understanding others' emotions. Evidence for such links allows diagnostic tools to use more common behaviours that do not need to be observed over time as indicators of ASD characteristics. Because persistent behaviours often require human interpretation, we argue that indicative behaviours are more appropriate as the targets for computational and technical measurement techniques. We will therefore be looking primarily at indicative behaviours, which are used by diagnostic tools and can be considered overt.

In terms of the behaviours defined by the DSM, Tables 1–3 below present an illustration of some of the considerations one must take into account when deciding the appropriateness of technologies for diagnostic purposes. In Tables 1 and 2 we identified whether each behaviour can be considered "covert" (i.e., requiring human interpretation to recognize). Those behaviours not marked as "covert" can be considered "overt". This judgement was made based on whether the behaviour can be clearly and unambiguously identified from observable behaviours alone, without having to incorporate information about the underlying intention or the appropriateness of the action. We also considered the locus of interactivity for each of the behaviours such that they are either "Interaction-Centred" (marked in Tables 1 and 2) or "Child-Centred" (not marked). Child-centred criteria are those for which only the behaviour of the child needs to be considered, for example, all the criteria under B4 (see Tables 1 and 2). Conversely, items such as all of A1 require the sensing of both interaction parties to provide an accurate assessment. These are therefore interaction-centred and impose additional challenges for automated methods; at a minimum, both the child and the therapist need to be detected and tracked by the sensory apparatus to capture the information necessary to characterise interaction-centred behaviours. It is important to note that we provide Tables 1 and 2 as a framework to illustrate the ideas presented in this review. Rather than being an authoritative classification of diagnostic criteria, we present it as a guide for future research, which should explore the viability of such applications of technology, the validity of the definitions it presents, and the development of technologies appropriate to augment the identification of each behaviour.

Table 1. Detailed breakdown of the behavioural cues for Category A that a therapist might use in ASD diagnosis based on DSM-5 criteria, and the corresponding required modalities.

	Required Modalities									**Class.**	
Behavioural cue	Gaze tracking	Speech detection	Speech analysis	Posture tracking	Gesture tracking	Facial expressions	Object tracking	Sound detection	Specific events	Covert behaviour	Interaction-centred
Category A **Persistent deficits in social communication and social interaction across contexts** **A1 Deficits in social-emotional reciprocity**											
1. One-sided conversations		✓									✓
2. Failure to offer comfort to others or to ask for it when needed			✓		✓				✓	✓	✓
3. Does not initiate conversation with peers		✓	✓	✓						✓	✓
4. Lack of showing, bringing, or pointing out objects of interest to other people				✓	✓		✓			✓	✓
5. Use of others as tools				✓	✓						✓
6. Failure to engage in simple social games				✓	✓					✓	✓
A2 Deficits in nonverbal communicative behaviours used for social interaction											
1. Impairments in social use of eye contact	✓										✓
2. Limited communication of own affect		✓	✓		✓	✓				✓	
3. Abnormalities in the use and understanding of emotion				✓	✓	✓				✓	✓
4. Impairment in the use of gestures					✓						
5. Abnormal volume, pitch, intonation, rate, rhythm, stress, prosody or volume in speech		✓									
6. Lack of coordinated verbal and nonverbal communication	✓	✓	✓		✓					✓	
A3 Deficits in nonverbal communicative behaviours used for social interaction											
1. Lacks understanding of the conventions of social interaction		✓			✓					✓	✓
2. Limited interaction with others in discussions and play	✓		✓	✓						✓	✓
3. Limited interests in talking with others			✓							✓	
4. Prefers solitary activities				✓	✓					✓	✓
5. Limited recognition of social emotions	✓	✓				✓					✓

Table 2. Detailed breakdown of the behavioural cues for Category B that a therapist might use in ASD diagnosis based on DSM-5 criteria, and the corresponding required modalities.

	Required Modalities										**Class.**
Behavioural cue	Gaze tracking	Speech detection	Speech analysis	Posture tracking	Gesture tracking	Facial expressions	Object tracking	Sound detection	Specific events	Covert behaviour	Interaction-centred
Category B **Restricted, repetitive patterns of behaviour, interests, or activities as manifested** **B1 Stereotyped or repetitive speech, motor movements, or use of objects**											
1. Repetitive hand movements					✓						
2. Stereotyped or complex whole body movements				✓							
3. Repetitive vocalisations such as repetitive guttural sounds, intonational noise making, unusual squealing repetitive humming		✓									
4. Perseverative or repetitive action / play / behaviour				✓	✓		✓				
5. Pedantic speech or unusually formal language			✓							✓	
B2 Excessive adherence to routines, ritualised patterns of verbal or nonverbal behaviour, or excessive resistance to change											
1. Overreactions to changes			✓		✓	✓				✓	✓
2. Unusual routines					✓		✓			✓	
3. Repetitive questioning about a particular topic			✓							✓	
4. Compulsions				✓	✓					✓	
B3 Highly restricted, fixated interests that are abnormal in intensity or focus											
1. Focused on the same few objects, topics or activities	✓	✓		✓	✓		✓			✓	
2. Verbal rituals		✓	✓							✓	
3. Excessive focus on irrelevant or non-functional parts of objects	✓		✓		✓		✓			✓	
B4 Hyper- or hypo-reactivity to sensory input or unusual interest in sensory aspects of environment											
1. Abnormal responses to sensory input				✓		✓			✓	✓	
2. Repetitively putting hands over ears				✓				✓			
3. Extreme interest or fascination with watching movement of other things				✓	✓		✓			✓	
4. Close visual inspection of objects				✓	✓		✓				

3. Automatic Quantification of Behaviour

Some of these behaviours, as described by the DSM, are not necessarily observable; however, they are associated with indicative behaviours. For the purposes of this review, we will present the case for the observability of both indicative and DSM defined behaviours. The following discussion reviews the challenges and opportunities associated with technologies that can be used to measure behavioural modalities associated with ASD symptoms. Examples of how these technologies have or could be applied are also discussed but it is important to note that not all applications or technologies will be discussed herein; rather it is a review of the behavioural modalities which have been addressed by technologies and are relevant for ASD diagnosis. Additionally, several technologies have already been applied to therapeutic settings [35,36], or to assist individuals with ASD in their daily lives [37,38] and may be mentioned in this paper but with the view to repurposing them for diagnostic scenarios.

Similarly, since we argue that the diagnostic requirements match onto general requirements for social robotics, there is also a substantial body of literature on identifying internal states (such as emotions) from observable behaviours in more general terms. Here, we briefly discuss such relevant work where applicable before moving on to the diagnostic requirements to highlight this connection.

Finally, this is primarily an overview of the challenges and opportunities available to researchers and clinicians in this field of research, rather than a review of all research pertaining to how technologies are relevant to individuals with ASD, as such there is a substantial pool of research which is not incorporated into this discussion.

Table 3. The number of times the behaviour modalities are identified in the behavioural cues listed in Tables 1 and 2, split according to whether the behavioural cues can be considered Overt or Covert and Child-Centred or Interaction-Centred. Highlighted (in grey) cells indicate where either overt/covert or child-centred/interaction-centred are more than double its counterpart. This is on the understanding (see text) that covert cues are more difficult to automatically characterise than overt cues, and that interaction-centred cues are more (practically) difficult to assess than child-centred cues.

Modality	Total Number	Interpretability of Behaviour		Locus of Interaction		A Cues	B Cues
		Overt	Covert	Child-Centred	Interaction-Centred		
1. Gaze tracking	6	1	5	3	3	4	2
2. Speech detection	10	4	6	6	4	7	3
3. Speech Analysis	11	0	11	7	4	6	5
4. Posture tracking	15	5	10	8	7	7	8
5. Gesture tracking	19	14	5	11	8	10	9
6. Facial expressions	5	1	4	2	3	3	2
7. Object tracking	7	2	5	6	1	1	6
8. Sound detection	1	1	0	1	0	0	1
9. Specific events	2	0	2	1	1	1	1
	Total	28	48	45	31		

3.1. Gaze Behaviour

3.1.1. Intention Recognition in Social Robotics

There is already a rich pool of research applying gaze-tracking techniques to the identification of socially relevant signals. For example, Nakano and Ishii [39] used gaze information, measured using a remote eye-tracking system, to estimate how engaged a user was in a conversation with a robotic agent. Similarly, Morency and colleagues [40] trained a robotic agent to recognize whether a human interaction partner was thinking about a response or waiting for the agent to respond based on gaze behaviour. As we will see, gaze tracking with ASD populations is largely used to identify atypical gaze behaviours, rather than to interpret internal states. However, based on these findings, gaze tracking might also be useful for identifying diagnostically relevant behavioural cues such as one-sided conversations (see Table 1). That is, application of a system such as that developed by Morency and colleagues [40] could provide a quantification of how frequently a child with ASD provides a turn-taking cue, and thereby a clearer understanding of how 'one-sided' their conversation is.

3.1.2. Requirements for ASD Diagnosis

Two aspects of gaze can be tracked using technologies: head direction (which overlaps with posture detection) and eye-gaze. Head direction tracking is relatively robust, and with several readily available algorithms, (e.g., [41]). Eye-gaze tracking, however, provides a much better indication of the orientation of visual attention. The usefulness of gaze tracking in the assessment of ASD symptoms is well established. We identify gaze tracking as a potential method for assessing six of the DSM

defined behaviours (see Table 3). Additionally, studies found associations between gaze behaviours and a variety of ASD symptoms, thus demonstrating the applicability of these technologies to ASD diagnosis. For example, the absence of preferential eye-contact with approaching adults is a predictor of the level of social disability [42], and children with ASD preferentially orient visually to non-social contingencies rather than to biological motion [43]. We will focus this discussion on two types or categories of gaze tracking technology: remote systems and wearables.

The term "remote systems" here refers to any non-invasive video-based camera or system, which can be positioned in an environment to track the eye movements of participants within its field of view. These systems are perhaps most useful for measuring interaction-centred behaviours where the full social scene must be taken into account, e.g., the position of objects of interest, or of other humans. For example, joint attention tasks can only be assessed by knowing the location and direction of gaze of the interaction partners, and the position of an object to which both partners should be attending. Joint attention in particular has been noted as an area where children with ASD demonstrate atypical gaze behaviours. For instance, Swanson and Siller [44] examined whether there were differences in the gaze behaviours of typically developing (TD) and ASD children during a joint attention task. They used a single remote system attached to a computer screen that displayed videos of an actor. Children's gaze behaviours were measured while they watched the video to see if they attended to the same areas of the screen as the actor. While Swanson and Siller did not find any differences between groups in global measures of gaze (e.g., overall looking time), they did detect differences in the microstructure of gaze behaviour (e.g., duration of first fixation). This not only demonstrates that gaze tracking is useful in the assessment of ASD behaviours, but also that using such technologies can allow us to identify behaviours which may not be identified by human observers.

Wearable gaze tracking systems range from head-mounted cameras to eye-tracking glasses and can be worn either by the child undergoing assessment or by a clinician or parent who is interacting with the child. Wearables allow the wearer more freedom of movement than remote systems and can be implemented outside of the diagnostic setting, allowing clinicians to gather diagnostic information about the child's daily life and at-home behaviours. Wearables are more appropriate for examining precisely what a child is looking at, i.e., investigations of attention orienting, in more naturalistic or dynamic settings. For example, Magrelli et al. [45] investigated how TD children and children with ASD orient their attention to social stimuli using a head-mounted eye-tracking device. This study specifically examined child behaviour during dyadic play interactions with an adult in environments that were familiar to the children. Magrelli et al. found that children with ASD looked at the adult's face less than TD children. This study demonstrates how wearable eye-tracking technologies could allow ASD diagnosis to include empirical, quantitative data about the child's behaviour during their every-day lives.

However, each of these techniques is associated with several challenges when applied to diagnostic settings and, therefore, opportunities for future development. For instance, the use of remote cameras requires some amount of restriction to the child's movements. To provide a full-frontal view of the face, single-camera techniques require the child to be relatively stationary and are ideally implemented to assess a child's behaviour during a task tailored to elicit differential eye-movements in ASD and TD children (as in [46]). Diagnostic settings however, often involve engaging children in several different tasks to assess a range of behaviours. Techniques such as switching between multiple cameras to find the optimal view seem, therefore, more appropriate to this setting. Wearables also offer a solution to this problem; however, the need for compact and comfortable technologies often results in some loss to the technology's accuracy [47].

3.2. Speech Behaviour

3.2.1. Intention Recognition in Social Robotics

It is well established that internal states and social signals can be recognized from features of speech. In particular, emotional states such as happiness, sadness, anger and fear were classified based on prosodic features of speech [48–50]. Similarly, prosodic features have been used to train classifiers to distinguish between positive, negative and neutral emotional states [51]. In terms of social signals, Hsiao et al. [52], for instance, demonstrated that turn-taking patterns and prosody features in speech could be used to classify high and low social engagement. This evidence clearly demonstrates that internal state information and social signals can be identified by classification systems based on speech and verbal behaviours.

3.2.2. Requirements for ASD Diagnosis

Speech processing has received increasing attention in recent years as commercial applications have come to the public. Solutions therefore exist that could be applied to automated analysis of speech during general, as well as diagnostic, interactions [53], although variability between speakers poses problems [54] that are particularly acute with child voices [55,56]. There are two broad types of speech properties that may be distinguished in the context of the diagnostic criteria: (1) detection of the presence/absence of speech (10 criteria; Table 3); and (2) the processing of the content of speech (comprised of detection of reportative speech, keyword recognition and understanding – 11 criteria; Table 3). The first of these can be addressed through the application of statistically-based signal processing techniques, for which there are a range of established solutions (e.g., [57,58]). Keyword recognition (which could also be used for repetition detection) lies in the area of speech recognition that is similarly well supported by a range of methods [57], including deep learning systems [59], although the complexity and noisiness of real-world contexts present further limitations. Speech understanding poses the most challenging level of analysis, with current technologies being limited to constrained settings until a greater level of context information can be incorporated [60]. In all of these cases, maximising the quality of the sound recordings using microphones (while minimising background noise, interference, etc) is clearly beneficial for maximising the performance of automated methods. In application to the diagnosis of ASD this may necessitate the deployment of multiple microphones, which introduces further issues of signal integration and sound source localisation, particularly with multiple speakers (e.g., the child and the clinician) present [61].

Children with ASD have difficulties both in generating and recognising vocal prosody and intonation [62], display a deficit in syllable production [63], and have substantially higher proportions of atypical vocalizations than TD children [64]. Differences in communication tend to be persistent, show little change over time, and may include monotonic intonation, deficits in the use of pitch and control of volume, in vocal quality, and use of aberrant stress patterns [16,65]. All these patterns can be observed around the age of 2, which has been proposed as the age at which a reliable diagnosis can be provided [66]. We identified a total of seventeen diagnostic behaviours as observable via speech behaviours (Tables 1 and 2). One of the main benefits of automated speech analysis for ASD diagnosis is that its use could speed up the assessment process in that clinicians would not be required to listen to and hand-code recordings of child speech. The second advantage we consider is that the use of technology allows for the assessment of child speech in their everyday lives and naturalistic interactions. For example, Warren et al. [67] used a digital language processor and language analysis software to record and analyse the conversational environments of children with ASD and TD children. The children wore the recording equipment in a pocket of their own clothing. They found that children with ASD engaged in fewer conversations and produced fewer vocalisations than TD children. Additionally, Warren et al. were able to examine what effect the language use and skills of the adults in the children's environments had on child speech. Their analysis of this data showed that the different language environments provided by adults (e.g., number of different words produced by adults,

frequency of responses to child utterances) may influence a child's linguistic development and thereby impose confounds into assessments of speech in children with ASD. While this technology can also be implemented within a classical diagnostic setting, this study demonstrates some of the benefits of technologies for gathering naturalistic data for assessment, which includes obtaining data that might otherwise be unavailable to clinicians (i.e., the child's language environment).

3.3. Posture and Gesture Behaviour

3.3.1. Intention Recognition in Social Robotics

Vision-based methods (using standard cameras/2D images) for human motion capture are well established [68], with face tracking being particularly developed. The recent advent of depth-based tracking and processing of detected skeletons in the scene (primarily using RGB-D data) resulted in additional well-established tools to facilitate various types of pose and behaviour analysis [69]. Depth-based methods can also be applied to hand-gesture characterisation [70], although sensory resolution constraints (e.g., hands and fingers being more difficult to detect) mean that image-based methods may currently remain more appropriate [71].

There is evidence demonstrating that emotional states (e.g., happiness, sadness, anger) [72–74] and internal states such as engagement [75] can be recognised from gesture and posture information collected through standard digital video devices. Similarly, body postures captured using the Microsoft Xbox Kinect device were successfully used to classify emotional states [76]. Outside of emotion recognition, other research showed that internal states and socially relevant dispositions or states can be recognised through pose and gestures. Okada et al. [77] were able to classify dominance and leadership based on gesture information. The main concern for using gesture and posture information during human-robot interactions "in the wild" is that fitting a robotic agent with a camera suitable for this purpose is not always straightforward. Current research generally relies on being able to use a camera system separate from any robotic agent, thus restricting the interaction environment. This is not to say that it is not achievable. Ramey and colleagues [78] for example, integrated the Kinect device into a social robot for tracking and recognising hand gestures. Similarly, Elfaramawy and colleagues [74] mounted a depth sensor onto a Nao robot to record movement data during an interaction with human users. This data was then used to classify whether the interaction partner was expressing the emotions anger, fear, happiness, sadness or surprise. These results demonstrate that internal state information and socially relevant information can be interpreted from gesture and posture behaviours.

3.3.2. Requirements for ASD Diagnosis

In terms of information directly relevant to the diagnostic criteria, methods of tracking and recognizing posture and gesture behaviours are typically targeted at the characterisation of individuals rather than groups of people, and so would be most appropriate for overt and child-centred behaviours, followed by overt and interaction-centred behaviours, provided both parties in the interaction are tracked. Twenty-four of the behaviours in Tables 1 and 2 are observable via posture and/or gesture behaviours.

Many of these behaviours are captured by research exploring deficits in motor-skills. The developmental trajectory of motor skills has been demonstrated to be predictive of the rate of language development [79,80], deficits in adaptive behaviour skills [81] and social communication skills [82]. Some studies conclude that between 80–90% of children with ASD show some degree of impairment in motor skills [83,84], and a recent meta-analysis concluded that motor deficits should be included in the core symptoms of ASD [85]. Furthermore, deficits in motor skills may affect fine and gross motor coordination, stereotyped movements and awkward patterns of object manipulation, lack of purposeful exploratory movements, and alterations of movement planning and execution [86–88]. Cook and colleagues [89] used a motion tracking system to explore whether individuals with ASD

demonstrated atypical kinematic profiles in arm movements compared to TD individuals. They found that individuals with ASD produced arm movements that were jerkier and proceeded with greater acceleration and velocity. Similarly, Anzulewicz et al. [90] used the sensors available in an iPad mini to measure the motor activity displayed by children with ASD as they played games on the device. Machine learning analysis of this data was used to identify whether there were differences between children with ASD and TD children, and found that children with ASD exhibited greater force of contact, different distributions of forces within gestures, and differences in gesture kinematics. Together these studies demonstrate not only that diagnostic information is available in behaviours which can be measured via motion sensing technologies, but also that these technologies are readily available in smart devices such as tablets and other touch screens.

Most demonstrations of technologies measuring atypical postures and gestures produced by individuals with ASD involve choreographed or specific motions and tasks (e.g., [89]). As such, more data of naturalistic gestures may be required before this technology can be fully implemented in diagnostic settings. The goal would be to provide data describing the differences between children with ASD and TD children in the kinds of gestures that are produced in social interactions and within the tasks involved in diagnostic assessments. However, with such a dataset, motion tracking technologies have a great potential for augmenting the diagnostic process by providing clinicians with information which is difficult to assess by human observers but which contains diagnostic identifiers.

3.4. Object and Sound Detection

Seven of the behaviours in Tables 1 and 2 also require object tracking and one requires sound detection. These modalities are considered separately from those in the paragraphs above since they are not directed specifically at a human agent. However, the same set of sensors may be deployed as for the other behavioural modalities, namely cameras (using 2D and depth images) and microphones.

Object tracking is particularly useful for assessments of joint attention, and in the ways children with ASD attend to and express their interest in objects. For example, Elison et al. [91] were able to categorise the behaviours of 12-month old children into distinct groups based on observed repetitive object manipulation behaviours. Furthermore, those children who demonstrate more repetitive object manipulation behaviours were more likely to be diagnosed with ASD at 24 months. Automating the measurement of these behaviours would require both gesture and object tracking but could reveal further identifiers for ASD or allow us to more precisely quantify the differences between groups on this type of task. Most demonstrations of automated object tracking in ASD contexts come in the form of robot-assisted therapies or diagnostic protocols. Petric et al. [92], for example, tested the efficacy of their autonomous robot protocol in carrying out four diagnostic tasks with children. In relation to object-tracking, these tasks involved the robot detecting whether the child was playing with a toy before attracting the child's attention (response to name), directing a child's attention to an object (joint attention), and to test whether a child would imitate actions using functional objects (functional imitation). The systems implemented in this study involved both the tracking of objects and the assessment of the child's behaviour with or towards that object in real time. While this application of object-tracking technologies is different to the application we propose in this review (i.e., we are not necessarily proposing the use of robots), this study does demonstrate how object tracking, alongside other methods like gesture tracking, can be used to assess child behaviour in real time during a clinical assessment to provide useful feedback.

There are a range of well-established methods and algorithms in the literature that are effective for object tracking based on visual data, with recent advances using deep learning methods (e.g., [93]). However, if manipulation is involved (as in items B1.4 and B3.1), then object occlusions may be problematic and so should be a focus of future developments. An additional challenge to this technique is that there is little empirical work quantifying differences between how children with ASD and TD children manipulate objects. Such work is essential before these techniques can be implemented in a

diagnostic setting because it would provide us with the identifiers, if there are any, which can be used to distinguish between children with and without ASD.

3.5. Facial Expressions

3.5.1. Intention Recognition in Social Robotics

Numerous technologies and approaches were developed to recognise and classify emotional facial expressions (EFEs). It has been demonstrated that emotional states can be recognised from facial expressions extracted from video data [94–98], (see also [99] for a survey of methods). Facial expressions have also proved useful for classifying engagement [100,101] showing that facial expressions are useful for identifying social signals beyond emotions.

3.5.2. Requirements for ASD Diagnosis

While the symptoms involving emotion expression have all been categorised as covert or "requiring human interpretation", technologies and techniques for identifying facial expressions, such as those described above, would be helpful in the assessment of how children communicate their own emotional states. However, this would be limited to examining the "strength" or frequency of emotional facial expressions rather than their appropriateness as this element requires human interpretation. Additionally, emotional expression analysis could aid in assessing how children detect and respond to the emotional expressions of others by combining such methods with gesture or eye tracking, or speech analysis. One study found that typically developed participants demonstrate different fixation and scanning patterns when observing faces expressing different emotions (e.g., more gazing at the mouth for happy and angry faces, and the eyes for sad faces) [102]. Additionally, another study found that children with ASD fixated on the mouth of happy and angry faces less than their TD peers [103]. If we take these findings together, they demonstrate a use-case for technologies which can be applied in naturalistic settings and are capable of simultaneously tracking the emotional expressions being communicated towards a child, and the child's gaze behaviours in viewing those expressions. This application would allow clinicians to include naturalistic data on emotion recognition capabilities in their diagnostic analysis. Alternatively, if this same method were applied in a controlled clinical setting, the use of automated emotion recognition would firstly help in validating whether an emotional expression was sufficient to communicate one emotion over another. Additionally, it would reduce the time needed to assess a child's gaze behaviours by automating the mapping between the occurrence of an emotional expression and the child's gaze behaviours in processing this expression, thus eliminating the need to manually code and map these events together.

Automated emotion classification from faces is typically based on the six basic emotions [104], and are associated with numerous limitations when applied to real-world situations (see [99,105] for reviews). However, given that during a diagnostic assessment, the clinician would act out the emotional expression (thus exaggerating the features), such methods may nevertheless be appropriate. Classification methods typically use Action Unit coding of facial expression features, with more recent attempts to incorporate other visual information, such as head behaviour [106]. Being a camera-based method, this characterisation of facial expression is subject to similar constraints as posture and gaze analysis.

4. Discussion and Conclusions

4.1. Limitations of Current Technology

In this paper, we discussed the state of the art of technological means to measure behavioural cues relevant to the diagnostic criteria for ASD. A consistent and reliable quantification of behaviour in the modalities identified that would go beyond the observational techniques currently employed has the potential to present clear advantages to clinicians in their evaluation of ASD symptoms.

It is apparent from our review that while there is definite scope for such automated quantification, there remain several limitations with current sensory technologies and their associated methods in this context. Some are due to practical constraints (e.g., the positioning and coverage of individual sensors), but the more problematic issues are typically related to diagnostic criteria involving a covert behavioural component, i.e., those behaviours that require some degree of interpretation in addition to the observation of the overt phenomena. Human assessors naturally bring their prior experience and extensive training into the diagnostic assessment process; for automated methods, this prior knowledge and experience must be codified for it to be applied. The problematic qualitative nature of such developed experience is an area in which the sensory interpretation methods discussed are currently lacking, for which deeper, more complex (perhaps even cognitive) models are required if they are to be sufficient to adequately augment human characterisation efforts.

Work in this direction must start on the more general level, outside of the confines of therapeutic settings. We have highlighted several existing works demonstrating how covert states/behaviours may be identified from overt behavioural cues at this level. A large body of work, for example, is devoted to the recognition of emotional states in a range of contexts. However, this is usually limited to the six 'basic' emotions [104] or to identifying the valence of emotion (positive, negative or neutral). As such, more work in this area is needed. In particular, further explorations of whether different, more complex covert states (e.g., frustration, distress, confusion) are shown in overt behaviours.

4.2. Classes of Behavioural Modalities in ASD Diagnosis

Seven behavioural modalities were described, which can be considered overt and therefore identifiable via technological means. Additionally, Tables 1 and 2 provide an initial framework for deciding which modalities are most appropriate for identifying and tracking these diagnostic behaviours. We propose this framework as a guideline for clinicians wishing to incorporate technological means of behaviour measurement into the diagnosis of ASD, as well as for researchers looking to develop and improve such technologies. In addressing the former goal, we have also identified behaviours we believe to be mostly, if not entirely, overtly observable. While covert behaviours do pose a challenge to technological measurement techniques, due to the requirement for human interpretation, our review identified some overt behaviours that were shown to be associated with, or indicative of, some of these behaviours. As such, the technologies and approaches we have discussed present an opportunity for clinicians to demonstrate support for their observations using quantifiable behaviours. For example, in assessing a child's ability to recognise emotional facial expressions, clinicians could both observe children's reactions to such expressions and measure the child's gaze patterns. This would not only provide empirical support for the clinician's conclusion, but may also assist in disambiguating a child's behaviour where there is uncertainty.

Alongside the overtness of each behaviour, we have also distinguished between behaviours that are expressed solely by the child being assessed (Child-Centred) and which are uniquely expressed within an interaction (Interaction-Centred). This distinction provides a framework for deciding which technologies or set-ups are most appropriate for measuring each behaviour, e.g., is a single camera more appropriate than multiple cameras (capturing the behaviour of all members of the interaction) for collecting visual data about a joint-attention assessment? Interaction-Centred behavioural cues do present complications in that they entail the tracking and characterisation of multiple individuals (minimally the child and the clinician) and their coordination, which is feasible, though posing additional challenges. Accounting for these considerations, it is noticeable that some of the modalities lend themselves more readily to immediate application than others, gesture tracking being the clearest example of this. Conversely, speech analysis remains a challenge, even assuming high performing speech recognition. Furthermore, we observe that 63% of behavioural cues across modalities require some degree of interpretation, and which would thus be currently difficult to automate.

4.3. Future Work

4.3.1. Diagnosis of ASD

Existing studies that deal with the use of technology in the diagnosis or treatment of ASD emphasise methodological differences in this broad field [107]. Our review suggests that more effort should be invested in developing technology-based applications that aim to benefit the diagnostic process for children with developmental disabilities, such as ASD or ADHD [108]. An additional, perhaps even greater, challenge in this field is not just to create effective technologies, but also to make them accessible for practitioners in terms of availability, ease of operation and cost. Technology-based tools have the potential to be an important resource in both assessment and treatment for individuals with ASD as they may be able to reduce the time and effort required by expert clinicians. As a result, diagnoses would become more accessible, consistent (through the application of standard recognition technologies for those overt aspects), and, potentially, more understandable. For instance, if a caregiver understands that a child's difficulty with recognising emotional facial expressions is related to the way the child attends to different facial features, the caregiver is able to apply this knowledge when providing the child with support during their daily lives, e.g., overtly directing the child's attention to relevant features during emotion-recognition games/exercises.

4.3.2. Social Robotics

As far as the field of social robotics is concerned, we have highlighted the need for algorithms that can infer covert, or internal states from observable kinematics. We have shown, in particular, that the main limitation is primarily on the algorithmic side and we recommend that more effort is put on addressing this directly. Indeed, we suggest (Section 4.1) that it may be necessary to integrate a more general cognitive aspect to this algorithmic processing. This provides a motivation for consideration of cognitive architectures in social robotics [109]: as we have highlighted in this paper, a robot controller that is merely responsive to observable behaviour is very unlikely to be sufficient for autonomous social interaction. As a means to further research in this direction, we have highlighted the overlap between the requirements of social robotics in general and ASD diagnosis in particular: as such, we argue that a system which can satisfactorily address the latter will also contain the technological developments required to advance the former.

4.4. Conclusion

Overall, this contribution highlighted that we are now at a point where it is feasible to incorporate novel, technology-based means into the diagnostic process for ASD. This opens up a new avenue of research, now ripe for exploring, focused on thorough evaluations of the benefits of, and further challenges in, technology-augmented diagnosis. With this paper, we hope to have provided the necessary starting points, highlighting for clinicians what is already possible, and for the developers of technology and psychology researchers, what the immediate obstacles are from a diagnostic point of view. The intent is to provide reliable and consistent quantitative data with which the diagnostic process can be improved, resulting in positive impacts for those children concerned. At the same time, it also highlights that further development of algorithms that can suitably assess covert states is a research avenue ready to be explored further in social robotics in general: with technological issues mostly solved and a good understanding of human-robot interactions from Wizard-of-Oz studies, this is the missing piece of the puzzle.

Author Contributions: Conceptualization, C.C., P.B. and S.T.; Formal analysis, M.E.B., C.C., P.B. and S.T.; Funding acquisition, C.C. and S.T.; Investigation, M.E.B., C.C., P.B. and S.T.; Methodology, M.E.B., C.C., P.B. and S.T.; Supervision, S.T.; Visualization, P.B. and S.T.; Writing—original draft, M.E.B., C.C. and P.B.; Writing— review & editing, P.B. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by European Commission FP7 grant number 611391.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Carruthers, P.; Smith, P.K. *Theories of Theories of Mind*; Cambridge University Press: Cambridge, UK, 1996.
2. Svensson, H.; Thill, S. Beyond bodily anticipation: internal simulations in social interaction. *Cognit. Syst. Res.* **2016**, *40*, 161–171. [CrossRef]
3. Demiris, Y.; Khadhouri, B. Hierarchical attentive multiple models for execution and recognition of actions. *Robot. Auton. Syst.* **2006**, *54*, 361–369. [CrossRef]
4. Demiris, Y. Prediction of intent in robotics and multi-agent systems. *Cognit. Process.* **2007**, *8*, 151–158. [CrossRef] [PubMed]
5. Haruno, M.; Wolpert, D.M.; Kawato, M. MOSAIC Model for Sensorimotor Learning and Control. *Neural Comput.* **2001**, *13*, 2201–2220. [CrossRef]
6. Metta, G.; Sandini, G.; Natale, L.; Craighero, L.; Fadiga, L. Understanding mirror neurons: A bio-robotic approach. *Interact. Stud.* **2006**, *7*, 197–232. [CrossRef]
7. Bartlett, M.E.; Edmunds, C.E.R.; Belpaeme, T.; Thill, S.; Lemaignan, S. What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions. *Front. Robot. AI* **2019**, *6*, 49. [CrossRef]
8. Bradwell, H.L.; Edwards, K.J.; Winnington, R.; Thill, S.; Jones, R.B. Companion robots for older people: importance of user-centred design demonstrated through observations and focus groups comparing preferences of older people and roboticists in South West England. *BMJ Open* **2019**. [CrossRef]
9. Vollmer, A.L.; Read, R.; Trippas, D.; Belpaeme, T. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Sci. Robot.* **2018**. [CrossRef]
10. Belpaeme, T.; Kennedy, J.; Ramachandran, A.; Scassellati, B.; Tanaka, F. Social robots for education: A review. *Sci. Robot.* **2018**. [CrossRef]
11. Cao, H.; Esteban, P.G.; Bartlett, M.; Baxter, P.; Belpaeme, T.; Billing, E.; Cai, H.; Coeckelbergh, M.; Costescu, C.; David, D.; et al. Robot-Enhanced Therapy: Development and Validation of Supervised Autonomous Robotic System for Autism Spectrum Disorders Therapy. *IEEE Robot. Autom. Mag.* **2019**, *26*, 49–58. [CrossRef]
12. Dautenhahn, K.; Werry, I. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmat. Cognit.* **2004**, *12*, 1–35. [CrossRef]
13. Scassellati, B.; Admoni, H.; Matarić, M. Robots for Use in Autism Research. *Annu. Rev. Biomed. Eng.* **2012**, *14*, 275–294. [CrossRef] [PubMed]
14. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*; American Psychiatric Publishing: Washington, DC, USA, 2013.
15. Kanner, L. Autistic disturbances of affective contact. *Nerv. Child* **1943**, *2*, 217–250.
16. Kanner, L. Follow-up study of eleven autistic children originally reported in 1943. *J. Autism Child. Schizophr.* **1971**, *1*, 119–145. [CrossRef]
17. Grzadzinski, R.; Huerta, M.; Lord, C. DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes. *Mol. Autism* **2013**, *4*, 12. [CrossRef]
18. Huerta, M.; Lord, C. Diagnostic Evaluation of Autism Spectrum Disorders. *Pediatr. Clin. N. Am.* **2012**, *59*, 103–111. [CrossRef]
19. Yates, K.; Le Couteur, A. Diagnosing autism. *Paediatr. Child Health* **2013**, *23*, 5–10. [CrossRef]
20. Rogers, C.L.; Goddard, L.; Hill, E.L.; Henry, L.A.; Crane, L. Experiences of diagnosing autism spectrum disorder: A survey of professionals in the United Kingdom. *Autism* **2016**, *20*, 820–831. [CrossRef]
21. Le Couteur, A.; Lord, C.; Rutter, M. *The Autism Diagnostic Interview-Revised (ADI-R)*; Western Psychological Services: Los Angeles, CA, USA, 2003.
22. Schopler, E.; Reichler, R.J.; DeVellis, R.F.; Daly, K. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *J. Autism Dev. Disord.* **1980**, *10*, 91–103. [CrossRef]
23. Lord, C.; Risi, S.; Lambrecht, L.; Edwin H. Cook, J.; Leventhal, B.L.; DiLavore, P.C.; Pickles, A.; Rutter, M. The Autism Diagnostic Observation Schedule–Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *J. Autism Dev. Disord.* **2000**, *30*, 205–223. [CrossRef]

24. Zwaigenbaum, L.; Bryson, S.; Garon, N. Early identification of autism spectrum disorders. *Behav. Brain Res.* **2013**, *251*, 133–146. [CrossRef] [PubMed]
25. Falkmer, T.; Anderson, K.; Falkmer, M.; Horlin, C. Diagnostic procedures in autism spectrum disorders: a systematic literature review. *Eur. Child Adolesc. Psychiatry* **2013**, *22*, 329–340. [CrossRef] [PubMed]
26. Crane, L.; Chester, J.W.; Goddard, L.; Henry, L.A.; Hill, E. Experiences of autism diagnosis: A survey of over 1000 parents in the United Kingdom. *Autism* **2015**, *20*, 153–162. [CrossRef]
27. Doi, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput. Med Imaging Graph.* **2007**, *31*, 198–211. [CrossRef] [PubMed]
28. Xiao, Y.; Zeng, J.; Niu, L.; Zeng, Q.; Wu, T.; Wang, C.; Zheng, R.; Zheng, H. Computer-Aided Diagnosis Based on Quantitative Elastographic Features with Supersonic Shear Wave Imaging. *Ultrasound Med. Biol.* **2014**, *40*, 275–286. [CrossRef]
29. Hallgren, K.A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor. Quant. Methods Psychol.* **2012**, *8*, 23–34. [CrossRef]
30. Zander, E.; Willfors, C.; Berggren, S.; Coco, C.; Holm, A.; Jifält, I.; Kosieradzki, R.; Linder, J.; Nordin, V.; Olafsdottir, K.; Bölte, S. The Interrater Reliability of the Autism Diagnostic Interview-Revised (ADI-R) in Clinical Settings. *Psychopathol.* **2017**, *50*, 219–227. [CrossRef]
31. Zander, E.; Willfors, C.; Berggren, S.; Choque-Olsson, N.; Coco, C.; Elmund, A.; Moretti, Å.H.; Holm, A.; Jifält, I.; Kosieradzki, R.; Linder, J.; Nordin, V.; Olafsdottir, K.; Poltrago, L.; Bölte, S. The objectivity of the Autism Diagnostic Observation Schedule (ADOS) in naturalistic clinical settings. *Eur. Child Adolesc. Psychiatry* **2015**, *25*, 769–780. [CrossRef]
32. O'Brien, J.; Spencer, J.; Girges, C.; Johnston, A.; Hill, H. Impaired Perception of Facial Motion in Autism Spectrum Disorder. *PLoS ONE* **2014**, *9*, e102173. [CrossRef]
33. Adolphs, R.; Sears, L.; Piven, J. Abnormal Processing of Social Information from Faces in Autism. *J. Cognit. Neurosci.* **2001**, *13*, 232–240. [CrossRef]
34. Sacrey, L.A.R.; Armstrong, V.L.; Bryson, S.E.; Zwaigenbaum, L. Impairments to visual disengagement in autism spectrum disorder: A review of experimental studies from infancy to adulthood. *Neurosci. Biobehav. Rev.* **2014**, *47*, 559–577. [CrossRef] [PubMed]
35. Scassellati, B.; Boccanfuso, L.; Huang, C.M.; Mademtzi, M.; Qin, M.; Salomons, N.; Ventola, P.; Shic, F. Improving social skills in children with ASD using a long-term, in-home social robot. *Sci. Robot.* **2018**, *3*, eaat7544. [CrossRef]
36. Washington, P.; Wall, D.; Voss, C.; Kline, A.; Haber, N.; Daniels, J.; Fazel, A.; De, T.; Feinstein, C.; Winograd, T. SuperpowerGlass: A Wearable Aid for the At-Home Therapy of Children with Autism. Available online: https://dl.acm.org/doi/pdf/10.1145/3130977 (accessed on 31 January 2020).
37. Gentry, T.; Kriner, R.; Sima, A.; McDonough, J.; Wehman, P. Reducing the Need for Personal Supports Among Workers with Autism Using an iPod Touch as an Assistive Technology: Delayed Randomized Control Trial. *J. Autism Dev. Disord.* **2014**, *45*, 669–684. [CrossRef] [PubMed]
38. Tang, Z.; Guo, J.; Miao, S.; Acharya, S.; Feng, J.H. Ambient Intelligence Based Context-Aware Assistive System to Improve Independence for People with Autism Spectrum Disorder. In Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA, 5–8 January 2016. [CrossRef]
39. Nakano, Y.; Ishii, R. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In Proceedings of the 15th international conference on Intelligent user interfaces, Hong Kong, China, 7–10 February 2010; pp. 139–148.
40. Morency, L.P.; Christoudias, C.M.; Darrell, T. Recognizing gaze aversion gestures in embodied conversational discourse. In Proceedings of the 8th international conference on Multimodal interfaces, Banff, AB, Canada, 2–4 November 2006; pp. 287–294.
41. Zhu, X.; Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012. [CrossRef]
42. Jones, W.; Carr, K.; Klin, A. Absence of Preferential Looking to the Eyes of Approaching Adults Predicts Level of Social Disability in 2-Year-Old Toddlers With Autism Spectrum Disorder. *Arch. Gen. Psychiatry* **2008**, *65*, 946. [CrossRef] [PubMed]
43. Klin, A.; Lin, D.J.; Gorrindo, P.; Ramsay, G.; Jones, W. Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature* **2009**, *459*, 257–261. [CrossRef]

44. Swanson, M.R.; Siller, M. Patterns of gaze behavior during an eye-tracking measure of joint attention in typically developing children and children with autism spectrum disorder. *Res. Autism Spectr. Disord.* **2013**, *7*, 1087–1096. [CrossRef]
45. Magrelli, S.; Jermann, P.; Noris, B.; Ansermet, F.; Hentsch, F.; Nadel, J.; Billard, A. Social orienting of children with autism to facial expressions and speech: a study with a wearable eye-tracker in naturalistic settings. *Front. Psychol.* **2013**, *4*. [CrossRef]
46. Frazier, T.W.; Klingemier, E.W.; Beukemann, M.; Speer, L.; Markowitz, L.; Parikh, S.; Wexberg, S.; Giuliano, K.; Schulte, E.; Delahunty, C.; Ahuja, V.; Eng, C.; Manos, M.J.; Hardan, A.Y.; Youngstrom, E.A.; Strauss, M.S. Development of an Objective Autism Risk Index Using Remote Eye Tracking. *J. Am. Acad. Child Adolesc. Psychiatry* **2016**, *55*, 301–309. [CrossRef]
47. Noris, B.; Nadel, J.; Barker, M.; Hadjikhani, N.; Billard, A. Investigating Gaze of Children with ASD in Naturalistic Settings. *PLoS ONE* **2012**, *7*, e44144. [CrossRef]
48. Petrushin, V.A. Emotion recognition in speech signal: experimental study, development, and application. In Proceedings of the Sixth International Conference on Spoken Language Processing, Beijing, China, 16–20 October 2000.
49. Dai, K.; Fell, H.J.; MacAuslan, J. Recognizing emotion in speech using neural networks. *Telehealth Assist. Technol.* **2008**, *31*, 38.
50. Li, Y.; Zhao, Y. Recognizing emotions in speech using short-term and long-term features. In Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998.
51. Litman, D.; Forbes, K. Recognizing emotions from student speech in tutoring dialogues. In Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, St Thomas, VI, USA, 30 November–4 December 2003; pp. 25–30.
52. Hsiao, J.C.y.; Jih, W.r.; Hsu, J.Y.j. Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. In Proceedings of the Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012.
53. Oller, D.K.; Niyogi, P.; Gray, S.; Richards, J.A.; Gilkerson, J.; Xu, D.; Yapanel, U.; Warren, S.F. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13354–13359. [CrossRef] [PubMed]
54. Benzeghiba, M.; Mori, R.D.; Deroo, O.; Dupont, S.; Erbes, T.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; Rose, R.; Tyagi, V.; Wellekens, C. Automatic speech recognition and speech variability: A review. *Speech Commun.* **2007**, *49*, 763–786. [CrossRef]
55. Gerosa, M.; Giuliani, D.; Brugnara, F. Acoustic Variability and Automatic Recognition of Children's Speech. *Speech Commun.* **2007**, *49*, 847–860. [CrossRef]
56. Kennedy, J.; Lemaignan, S.; Montassier, C.; Lavalade, P.; Irfan, B.; Papadopoulos, F.; Senft, E.; Belpaeme, T. Child Speech Recognition in Human-Robot Interaction. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017. [CrossRef]
57. Rabiner, L.R.; Schafer, R.W. Introduction to Digital Speech Processing. *Found. Trends Signal Process.* **2007**, *1*, 1–194. [CrossRef]
58. Ramirez, J.; M., J.; C., J. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In *Robust Speech Recognition and Understanding*; I-Tech: Vienna, Austria, 2007; Volume 6, pp. 1–22. [CrossRef]
59. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Rahman Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
60. Moore, R.K. Spoken language processing: Piecing together the puzzle. *Speech Commun.* **2007**, *49*, 418–435. [CrossRef]
61. Athanasopoulos, G.; Verhelst, W.; Sahli, H. Robust speaker localization for real-world robots. *Comput. Speech Lang.* **2015**, *34*, 129–153. [CrossRef]
62. Shriberg, L.D.; Paul, R.; McSweeny, J.L.; Klin, A.; Cohen, D.J.; Volkmar, F.R. Speech and Prosody Characteristics of Adolescents and Adults With High-Functioning Autism and Asperger Syndrome. *J. Speech Lang. Hear. Res.* **2001**, *44*, 1097–1115. [CrossRef]
63. Rapin, I.; Dunn, M. Language disorders in children with autism. *Semin. Pediatr. Neurol.* **1997**, *4*, 86–92. [CrossRef]

64. Sheinkopf, S.J.; Mundy, P.; Oller, D.K.; Steffens, M. Vocal Atypicalities of Preverbal Autistic Children. *J. Autism Dev. Disord.* **2000**, *30*, 345–354. [CrossRef]
65. Paul, R.; Augustyn, A.; Klin, A.; Volkmar, F.R. Perception and Production of Prosody by Speakers with Autism Spectrum Disorders. *J. Autism Dev. Disord.* **2005**, *35*, 205–220. [CrossRef] [PubMed]
66. Centres for Disease Control and Prevention. Autism Spectrum Disorder (ASD): Data and Statistics. Available online: https://www.cdc.gov/ncbddd/autism/data.html (accessed on 31 January 2020).
67. Warren, S.F.; Gilkerson, J.; Richards, J.A.; Oller, D.K.; Xu, D.; Yapanel, U.; Gray, S. What Automated Vocal Analysis Reveals About the Vocal Production and Language Learning Environment of Young Children with Autism. *J. Autism Dev. Disord.* **2009**, *40*, 555–569. [CrossRef] [PubMed]
68. Moeslund, T.B.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126. [CrossRef]
69. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334. [CrossRef] [PubMed]
70. Suarez, J.; Murphy, R.R. Hand gesture recognition with depth images: A review. In Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication, Paris, France, 9–13 September 2012. [CrossRef]
71. Mitra, S.; Acharya, T. Gesture Recognition: A Survey. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2007**, *37*, 311–324. [CrossRef]
72. Castellano, G.; Villalba, S.D.; Camurri, A. Recognising human emotions from body movement and gesture dynamics. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 12–14 September 2007.
73. Saha, S.; Datta, S.; Konar, A.; Janarthanan, R. A study on emotion recognition from body gestures using Kinect sensor. In Proceedings of the 2014 International Conference on Communication and Signal Processing, Bangkok, Thailan, 10–12 October 2014; pp. 56–60.
74. Elfaramawy, N.; Barros, P.; Parisi, G.I.; Wermter, S. Emotion recognition from body expressions with a neural network architecture. In Proceedings of the 5th International Conference on Human Agent Interaction, Bielefeld, Germany, 17–20 October; pp. 143–149.
75. Sanghvi, J.; Castellano, G.; Leite, I.; Pereira, A.; McOwan, P.W.; Paiva, A. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In Proceedings of the 6th International Conference on Human-Robot Interaction, Lausanne Switzerland, 6–9 March 2011; pp. 305–312.
76. Guerrero Rázuri, J.F.; Larsson, A.; Sundgren, D.; Bonet, I.; Moran, A. Recognition of emotions by the emotional feedback through behavioral human poses. *Int. J. Comput. Sci. Issues* **2015**, *12*, 7–17.
77. Okada, S.; Aran, O.; Gatica-Perez, D. Personality trait classification via co-occurrent multiparty multimodal event discovery. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November; pp. 15–22.
78. Ramey, A.; González-Pacheco, V.; Salichs, M.A. Integration of a low-cost RGB-D sensor in a social robot for gesture recognition. In Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Lausanne, Switzerland, 6–9 March 2011; pp. 229–230.
79. Bedford, R.; Pickles, A.; Lord, C. Early gross motor skills predict the subsequent development of language in children with autism spectrum disorder. *Autism Res.* **2015**, *9*, 993–1001. [CrossRef]
80. Leonard, H.C.; Bedford, R.; Pickles, A.; Hill, E.L. Predicting the rate of language development from early motor skills in at-risk infants who develop autism spectrum disorder. *Res. Autism Spectr. Disord.* **2015**, *13–14*, 15–24. [CrossRef]
81. MacDonald, M.; Lord, C.; Ulrich, D. The relationship of motor skills and adaptive behavior skills in young children with autism spectrum disorders. *Res. Autism Spectr. Disord.* **2013**, *7*, 1383–1390. [CrossRef]
82. Bradshaw, J.; Klaiman, C.; Gillespie, S.; Brane, N.; Lewis, M.; Saulnier, C. Walking Ability is Associated with Social Communication Skills in Infants at High Risk for Autism Spectrum Disorder. *Infancy* **2018**, *23*, 674–691. [CrossRef]
83. Ming, X.; Brimacombe, M.; Wagner, G.C. Prevalence of motor impairment in autism spectrum disorders. *Brain Dev.* **2007**, *29*, 565–570. [CrossRef] [PubMed]
84. Hilton, C.L.; Cumpata, K.; Klohr, C.; Gaetke, S.; Artner, A.; Johnson, H.; Dobbs, S. Effects of Exergaming on Executive Function and Motor Skills in Children With Autism Spectrum Disorder: A Pilot Study. *Am. J. Occup. Ther.* **2013**, *68*, 57–65. [CrossRef] [PubMed]

85. Fournier, K.A.; Hass, C.J.; Naik, S.K.; Lodha, N.; Cauraugh, J.H. Motor Coordination in Autism Spectrum Disorders: A Synthesis and Meta-Analysis. *J. Autism Dev. Disord.* **2010**, *40*, 1227–1240. [CrossRef] [PubMed]
86. Pierce, K.; Courchesne, E. Evidence for a cerebellar role in reduced exploration and stereotyped behavior in autism. *Biol. Psychiatry* **2001**, *49*, 655–664. [CrossRef]
87. Minshew, N.J.; Sung, K.; Jones, B.L.; Furman, J.M. Underdevelopment of the postural control system in autism. *Neurology* **2004**, *63*, 2056–2061. [CrossRef]
88. Rinehart, N.J.; Bradshaw, J.L.; Brereton, A.V.; Tonge, B.J. Movement Preparation in High-Functioning Autism and Asperger Disorder: A Serial Choice Reaction Time Task Involving Motor Reprogramming. *J. Autism Dev. Disord.* **2001**, *31*, 79–88. [CrossRef]
89. Cook, J.L.; Blakemore, S.J.; Press, C. Atypical basic movement kinematics in autism spectrum conditions. *Brain* **2013**, *136*, 2816–2824. [CrossRef]
90. Anzulewicz, A.; Sobota, K.; Delafield-Butt, J.T. Toward the Autism Motor Signature: Gesture patterns during smart tablet gameplay identify children with autism. *Sci. Rep.* **2016**, *6*, 31107. [CrossRef]
91. Elison, J.T.; Wolff, J.J.; Reznick, J.S.; Botteron, K.N.; Estes, A.M.; Gu, H.; Hazlett, H.C.; Meadows, A.J.; Paterson, S.J.; Zwaigenbaum, L.; Piven, J. Repetitive Behavior in 12-Month-Olds Later Classified With Autism Spectrum Disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **2014**, *53*, 1216–1224. [CrossRef]
92. Petric, F.; Hrvatinic, K.; Babic, A.; Malovan, L.; Miklic, D.; Kovacic, Z.; Cepanec, M.; Stosic, J.; Simlesa, S. Four tasks of a robot-assisted autism spectrum disorder diagnostic protocol: First clinical tests. In Proceedings of the IEEE Global Humanitarian Technology Conference (GHTC 2014), San Francisco, CA, USA, 10–13 October 2014. [CrossRef]
93. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the ACM International Conference on Multimedia, Glasgow, UK, 1–4 April 2014. [CrossRef]
94. Michel, P.; El Kaliouby, R. Real time facial expression recognition in video using support vector machines. In Proceedings of the 5th international conference on Multimodal interfaces, Vancouver, BC, Canada, 5–7 November 2003; pp. 258–264.
95. Bartlett, M.S.; Littlewort, G.; Frank, M.; Lainscsek, C.; Fasel, I.; Movellan, J. Recognizing facial expression: machine learning and application to spontaneous behavior. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 568–573.
96. Bargal, S.A.; Barsoum, E.; Ferrer, C.C.; Zhang, C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo Japan, 12–16 November 2016; pp. 433–436.
97. Littlewort, G.; Bartlett, M.S.; Fasel, I.R.; Chenu, J.; Kanda, T.; Ishiguro, H.; Movellan, J.R. Towards social robots: Automatic evaluation of human-robot interaction by facial expression classification. *Adv. Neural Inf. Process. Syst.* **2004**, pp. 1563–1570.
98. Zhang, L.; Jiang, M.; Farid, D.; Hossain, M.A. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Syst. Appl.* **2013**, *40*, 5160–5168. [CrossRef]
99. Corneanu, C.A.; Simon, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [CrossRef] [PubMed]
100. Nezami, O.M.; Dras, M.; Hamey, L.; Richards, D.; Wan, S.; Paris, C. Automatic Recognition of Student Engagement using Deep Learning and Facial Expression. *arXiv* **2018**, arXiv:1808.02324.
101. Liu, T.; Kappas, A. Predicting Engagement Breakdown in HRI Using Thin-slices of Facial Expressions. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
102. Eisenbarth, H.; Alpers, G.W. Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion* **2011**, *11*, 860–865. [CrossRef] [PubMed]
103. Åsberg Johnels, J.; Hovey, D.; Zürcher, N.; Hippolyte, L.; Lemonnier, E.; Gillberg, C.; Hadjikhani, N. Autism and emotional face-viewing. *Autism Res.* **2017**, *10*, 901–910. [CrossRef]
104. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]

105. Pantie, M.; Rothkrantz, L. Automatic analysis of facial expressions: the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445. [CrossRef]
106. Zeng, Z.; Pantic, M.; Roisman, G.; Huang, T. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 39–58. [CrossRef]
107. Grynszpan, O.; Weiss, P.L.T.; Perez-Diaz, F.; Gal, E. Innovative technology-based interventions for autism spectrum disorders: A meta-analysis. *Autism* **2013**, *18*, 346–361. [CrossRef]
108. Robe, A.; Dobrean, A.; Cristea, I.A.; Păsărelu, C.R.; Predescu, E. Attention-deficit/hyperactivity disorder and task-related heart rate variability: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **2019**, *99*, 11–22. [CrossRef]
109. Baxter, P.; Lemaignan, S.; Trafton, J.G. Cognitive Architectures for Social Human-Robot Interaction. In Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand, 7–10 March 2016; pp. 579–580.

Appendix C

Frontiers in Robotics and AI Article - What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 26 June 2019
doi: 10.3389/frobt.2019.00049

What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions

Madeleine E. Bartlett[1]*, *Charlotte E. R. Edmunds*[2], *Tony Belpaeme*[1,3], *Serge Thill*[4,5] and *Séverin Lemaignan*[6]

[1] Centre for Robotics and Neural Systems (CRNS), University of Plymouth, Plymouth, United Kingdom, [2] Warwick Business School, University of Warwick, Coventry, United Kingdom, [3] ID Lab—imec, University of Ghent, Ghent, Belgium, [4] Interaction Lab, School of Informatics, University of Skövde, Skövde, Sweden, [5] Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, Netherlands, [6] Bristol Robotics Lab, University of the West of England, Bristol, United Kingdom

OPEN ACCESS

Edited by:
Cigdem Beyan,
Istituto Italiano di Tecnologia, Italy

Reviewed by:
Radoslaw Niewiadomski,
University of Genoa, Italy
Atesh Koul,
Istituto Italiano di Tecnologia, Italy
Giulia Perugia,
Uppsala University, Sweden
Jordi Vallverdu,
Autonomous University of Barcelona, Spain

***Correspondence:**
Madeleine E. Bartlett
madeleine.bartlett@plymouth.ac.uk

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: 12 January 2019
Accepted: 06 June 2019
Published: 26 June 2019

Citation:
Bartlett ME, Edmunds CER, Belpaeme T, Thill S and Lemaignan S (2019) What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions. Front. Robot. AI 6:49. doi: 10.3389/frobt.2019.00049

In recent years, the field of Human-Robot Interaction (HRI) has seen an increasing demand for technologies that can recognize and adapt to human behaviors and internal states (e.g., emotions and intentions). Psychological research suggests that human movements are important for inferring internal states. There is, however, a need to better understand what kind of information can be extracted from movement data, particularly in unconstrained, natural interactions. The present study examines which internal states and social constructs humans identify from movement in naturalistic social interactions. Participants either viewed clips of the full scene or processed versions of it displaying 2D positional data. Then, they were asked to fill out questionnaires assessing their social perception of the viewed material. We analyzed whether the full scene clips were more informative than the 2D positional data clips. First, we calculated the inter-rater agreement between participants in both conditions. Then, we employed machine learning classifiers to predict the internal states of the individuals in the videos based on the ratings obtained. Although we found a higher inter-rater agreement for full scenes compared to positional data, the level of agreement in the latter case was still above chance, thus demonstrating that the internal states and social constructs under study were identifiable in both conditions. A factor analysis run on participants' responses showed that participants identified the constructs *interaction imbalance*, *interaction valence* and *engagement* regardless of video condition. The machine learning classifiers achieved a similar performance in both conditions, again supporting the idea that movement alone carries relevant information. Overall, our results suggest it is reasonable to expect a machine learning algorithm, and consequently a robot, to successfully decode and classify a range of internal states and social constructs using low-dimensional data (such as the movements and poses of observed individuals) as input.

Keywords: social psychology, human-robot interaction, machine learning, social interaction, recognition

1. INTRODUCTION

One of the main goals in the field of Human-Robot Interaction (HRI) is to create robots capable of recognizing and adapting to human interaction partners in an appropriate manner (Dautenhahn and Saunders, 2011). In human-human interactions, the appropriateness of our responses to others is often a result of our ability to recognize the internal states (e.g., intentions, dispositions) of our interaction partner (Domes et al., 2007). Here we focus on internal states and social constructs relevant to task engagement and social relations between interaction partners. For example, we consider states that can be thought of as dispositional judgments (e.g., friendliness), states which can be considered emotional and are embedded within a social context (e.g., aggression), and states relevant to task performance (e.g., boredom). These states are communicated through both verbal and non-verbal cues (Pollick et al., 2001; Manera et al., 2011). Endowing robots and behavior classification systems with a similar ability to recognize internal states based on non-verbal behaviors would allow for more appropriate, autonomous human-robot interactions (Breazeal et al., 2009; Vernon et al., 2016), and for classification systems to provide more detailed insights into human behavior, e.g., for security purposes (Gowsikhaa et al., 2014).

1.1. Internal State Recognition

HRI research exploring approaches to achieving on-line recognition of human internal states/behavior draws on our understanding of how humans themselves infer internal states and social constructs. For example, a rich history of research has led to the assumption that humans are able to infer the internal states of others by observing their actions and movements (Gallese and Goldman, 1998; Manera et al., 2011; Quesque et al., 2013) and facial expressions (Ekman and Friesen, 1971; Haidt and Keltner, 1999; Tracy and Robins, 2008). In their paper, Manera et al. (2011) claim that *"in some circumstances, the movement of a human body... is sufficient to make judgments... in relation to the actor's intention"* [p. 548]. The idea here is that our intentions or emotions influence differences in the movements we make and, as observers, we are able to pick up on these differences and use them to infer the internal state of the person performing the action (Pollick et al., 2001; Ansuini et al., 2014; Becchio et al., 2017). To examine this researchers have used point-light displays and other methods to isolate movement information from other sources of information. Point-light displays denote the position and movements of an actor's joints on an otherwise blank display. Studies using this type of stimulus have shown that humans are able to use observed movement to infer an actor's gender (Kozlowski and Cutting, 1977; Mather and Murdoch, 1994; Hufschmidt et al., 2015), intention (Manera et al., 2010; Quesque et al., 2013) and emotional state (Pollick et al., 2001; Alaerts et al., 2011).

Available evidence also suggests that internal states and social constructs which fall under our definition of being socially relevant, dispositional or related to task engagement/performance are recognizable from observable movement. Okada et al. (2015) found that observable movements and non-verbal audio information produced during spontaneous, naturalistic interactions were sufficient for classifying dispositions and social behaviors such as dominance and leadership. Similarly, Sanghvi et al. (2011) demonstrated that postural behaviors could be used to classify a child's engagement with a robotic opponent, with which the children are playing a game. Beyan et al. (2016) asked four unacquainted individuals to complete a group decision task. They found that a classifier, when fed the 3D positional data of the interaction, was able to identify leaders within the group based on head pose and gaze direction information. Sanchez-Cortes et al. (2011) applied a computational framework to the inference of leadership and related concepts (e.g., dominance, competence) from non-verbal behaviors in a group interaction. Interactions in this study took place between four previously unacquainted individuals whose interactions were spontaneous and minimally structured. Sanchez-Cortes and colleagues were able to identify which behaviors were most informative for the recognition of the different leadership concepts. For example, conversational turn-taking and body movement behaviors were found to be the most informative for inferring leadership, whereas head activity and vocal pitch were the most informative for inferring competence.

States which are socially relevant, dispositional or task related, (such as friendliness, dominance or engagement) are particularly relevant for HRI research where the aim is to provide a socially interactive agent. In such scenarios it is preferable to have an agent which can provide appropriate social behaviors and responses (Dautenhahn and Saunders, 2011). Whilst emotion and intention recognition are definitely important for generating appropriate autonomous social behaviors from a robot, some HRI scenarios would also benefit from an ability to recognize internal states as we have defined them here. For instance, a teaching robot, such as those developed by the L2TOR project (Belpaeme et al., 2015), would be better able to provide appropriately timed encouragements or prompts if able to recognize when a student is bored or not engaged with the learning task.

As a result, HRI researchers have begun exploring ways in which observed movement can be utilized by robots and artificial systems to enable automated interpretation of, and responding to, the internal states of humans (Schrempf and Hanebeck, 2005; Han and Kim, 2010). Whilst humans also use other cues such as tone of voice (Walker-Andrews, 1997), findings such as those described above suggest that movement information may be sufficient for recognizing some, if not all, human internal states.

1.2. Current Study

1.2.1. Motivation and Approach

To take advantage of this information for the purposes of internal state recognition it is important to first identify what internal state information is available in movements and body postures. This knowledge is particularly useful for streamlining the design process for a robot or classifier able to interpret such data. For example, if we want to design a system able to recognize when a human is bored, we first need to know what data is sufficient, if not optimal, for recognizing this state. Would the system need

to take multiple behaviors into account, e.g., movements and prosodic features, or would movement alone be enough? In the case of internal states such as emotions and intentions, previous research suggests that movement information is sufficient for gaining insight (e.g., Tracy and Robins, 2008; Manera et al., 2011; Quesque et al., 2013). Given that the aim of HRI research is to create systems and robots which can be deployed in the real world, it is also important to consider that a classifier must be able to deal with natural, spontaneous human behaviors. Consequently, it is important to explore whether (and which) internal states can be recognized from the movements produced in natural human interactions. A a growing pool of studies have examined this (e.g., Sanchez-Cortes et al., 2011; Sanghvi et al., 2011; Shaker and Shaker, 2014; Okada et al., 2015; Beyan et al., 2016; Okur et al., 2017; Kawamura et al., 2019). However, further research is needed to provide a better understanding of which internal states can be inferred from such movements.

We therefore propose that an exploration into how readily different types of internal states can be identified from naturalistic human behavior would be beneficial for the streamlining of future HRI research. That is, by identifying which internal states are best recognized from a particular behavioral modality (e.g., biological motion), future research can identify which data sources are most useful for a given recognition task.

This study takes the first steps in this direction by developing a method for determining which internal state information is reported as identifiable by humans when they observe people in natural interactions. Given the strength of evidence suggesting that movement information is useful for identifying emotional and other internal states or social constructs (e.g., Pollick et al., 2001; Gross et al., 2012; Quesque et al., 2013; Beyan et al., 2016), this modality is likely to be a rich source of internal state information. Further, by extending this work to naturalistic interactions, we will find which internal states are likely to be identified in more ecologically valid settings. The usefulness of these states to HRI, indicate that an exploration of which internal states, from a selection of several, are recognizable from human movements would be helpful in guiding future research and development. To address this, we aim to examine and compare how reliably humans report identifying a number of different internal states and social constructs from observable movements.

To summarize, the main aim of this study is to demonstrate a method for identifying: (1) whether the data source of choice (in this case observable movements) can be used by humans to infer internal states and social constructs, and (2) what internal states and social constructs are readable from the movements within the data set. To do so, we will present short video clips of social interactions (exhibiting seven different internal states and social constructs) to participants. These clips come from the PInSoRo (Lemaignan et al., 2017) data set made openly available by our group[1]. This data set consists of videos of child-child or child-robot interactions. Children were asked to play for as long as they wanted on a touch-screen table-top device. For this study, we will solely use the child-child interactions as these are more likely to involve spontaneous behaviors throughout the children's interactions with one another. Some participants will view short clips including the full visual scene (full-scene condition) and others clips containing only movement and body posture information (movement-alone condition). These clips will contain at least one noticeable internal state (for details of the selection process see the Method section). Following each clip, participants respond to a series of questions where they can describe the internal states (e.g., boredom, friendliness) or social constructs (e.g., cooperation, dominance) they identified in the children's behaviors. By comparing responses in each condition we expect to be able to identify constructs which are likely to be recognizable from movement information alone.

1.2.2. Hypotheses and Predictions

Based on previous findings that humans are able to recognize internal states such as emotions (Gross et al., 2012) and group dynamics such as leadership (Beyan et al., 2016) from human motion information, we expect the following:

1. Participants will report being able to draw internal state information from the movement-alone videos (Hypothesis 1). Specifically, we predict that even in the impoverished movement-alone condition, the provided ratings will be sufficient to describe the internal states and social constructs identified in the observed interaction. This can be tested by training a classifier on the full-scene ratings, and assessing its performance when tested on the movement-alone ratings.
2. However, given that participants in this condition are provided with fewer visual cues than those viewing the full-scene videos (e.g., lack of resolution for facial expressions) we expect a higher recognition error rate in the movement-alone condition compared to the full-scene condition (Hypothesis 2). If this is the case, we predict that inter-rater agreement levels amongst participants will be above chance in both conditions (i.e. the same constructs are robustly identified in the clips by the participants), but with higher levels of agreement in the full-scene condition.

2. METHOD

2.1. Design and Participants

This study examined the effect of video type (full-scene vs. movement-alone) on responses to questions about the nature of the interaction depicted in the videos. We used a between-subject design: participants saw either *full-scene* clips (**Figure 1**, left) or *movement-alone* clips (**Figure 1**, right). 284 participants were recruited from Amazon's Mechanical Turk (MTurk). A total of 85 participants were excluded from analysis due to incorrect answers to an attention check (discussed in Procedure), leaving 199 participants (see **Table 1** for demographics). All participants were remunerated $1 (USD) upon completion of the experiment.

2.2. Materials

The stimuli used for this experiment were extracted from the PInSoRo data set. This data set contains videos (up to 40 min long) of pairs of children interacting whilst playing on a touch-screen table-top. For the present study we extracted twenty 30 s clips from these videos. We wanted to provide participants with

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

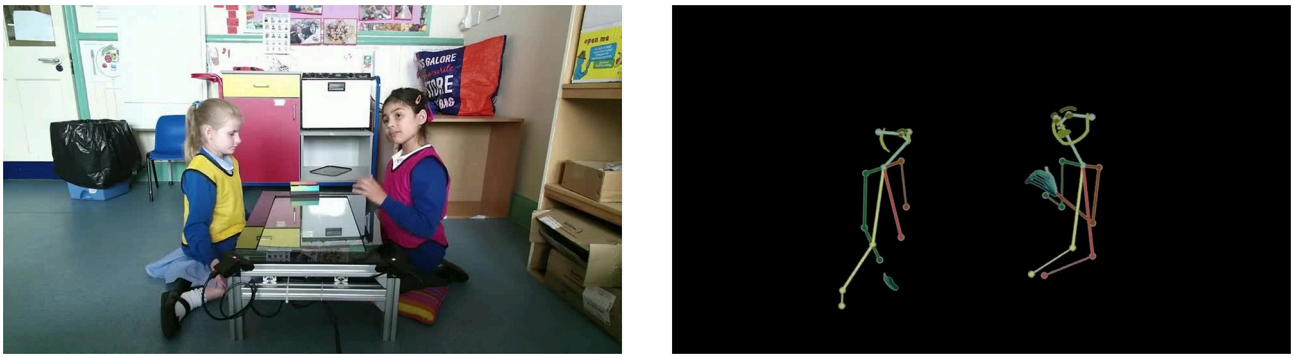


FIGURE 1 | Captures of one of the twenty video-clips, *full-scene* condition on the left, *movement-alone* condition on the right. Written consent for these images to be shared was obtained during collection.

TABLE 1 | Demographics of participants included in the analyses.

Condition	*N*	Mean Age (Range)	Gender (%M, %F)	% American	% English First Language
Movement-Alone	100	34.52 (22–70)	55%, 44%	75%	80%
Full-Scene	99	33.54 (19–72)	65%, 34%	69%	73%
Both	199	34.03 (19–72)	60%, 39%	72%	76%

clips which showed both children in the frame at the same time. We therefore selected our stimuli from videos filmed using a camera which had been positioned roughly 1.4m away from the touch-screen table-top, with the table-top in the center of the camera's view, thus allowing for each child to be viewed on either side of the frame (see **Figure 1**, left).

Two versions of the same clips were extracted: the *full-scene* clips were the raw video footage of the children playing, recorded from a static camera (**Figure 1**, left); the *movement-alone* clips were based on the exact same clips, but post-processed to extract skeletal and facial landmarks (using the OpenPose library[2]; Cao et al., 2017). Resulting landmarks were rendered on a black background, and connected to each other using colored lines, so that each child was depicted as a stick-man-style figure (**Figure 1**, right).

Clip selection was made based on whether a notable "event" or social dynamic occurred, defined as the labels listed in **Table 2**. This was done by watching the full-scene clips and working out what internal states and social constructs might be inferred from the children's movements. Specifically, two experimenters selected and labeled clips (by first independently extracting and annotating clips from the PInSoRo dataset, and second discussing to reach consensus) wherein at least one of the following seven concepts described the children's behavior or their interaction in the full-scene clips (see **Table 2**):

1. Boredom - at least one child was bored or not engaging with the task on the touch-screen (e.g., resting head in hand, interacting with touch-screen in slow/lazy manner).
2. Aggression - at least one child exhibited a physical aggressive action either toward the touch-screen or the other child (e.g., hitting the screen, pushing the other child's hand away).
3. Cooperation - the children were working together and/or communicating about how to perform a task [e.g., talking, joint attention (looking at the same object together), nodding].
4. Dominance - one child was bossy, performing most of the actions on the touch-screen or clearly in charge (e.g., pointing to touch-screen and talking at the other child, stopping the other child from using the touch-screen, being the only child to use the touch-screen).
5. Aimless play - at least one child was interacting with the touch-screen in a non-goal-directed manner or without being very engaged in their task (e.g., sitting slightly away from touch-screen whilst still using it, slow/lazy movements on touch-screen, not always looking at what they're doing).
6. Fun - at least one child was having fun (e.g., laughing, smiling).
7. Excitement - at least one child behaved excitedly (e.g., more dynamic than just "having fun," hearty laughter, open smiling mouth, fast movements).

It was decided that multiple labels could be applied to each clip for two reasons. First, the two children in each clip could have behaved in very different ways. Thus, if one child was bored and the other excited, the clip would be assigned both the Boredom and Excitement labels (see **Table 2**). Second, we recognized that a lot can happen in 30 s (the duration of the clips) resulting in changes in the internal states or social constructs which could be inferred from the children's behaviors. For example, an interaction might involve an excited child pushing the other away so they didn't have to share the touch-screen, causing the second child to sit and watch in a manner denoting boredom, this clip could be labeled with Excitement, Aggression and Bored. These labels were selected based on two considerations: (a) the events and internal states which appear available the dataset, and (b) events and internal states which would be useful to a robot which might observe or mediate

[2]https://github.com/CMU-Perceptual-Computing-Lab/openpose/

TABLE 2 | Labels that experimenters assigned to each clip during clip selection.

Clip	Label 1	Label 2	Label 3
01	Aggressive		
02	Aggressive	Excited	Aimless
03	Excited	Fun	
04	Cooperative		
05	Bored	Aimless	
06	Cooperative		
07	Dominance		
08	Bored		
09	Cooperative		
10	Cooperative	Dominance	
11	Cooperative	Dominance	
12	Aggressive	Aimless	
13	Excited	Aggressive	Aimless
14	Aggressive	Fun	
15	Dominance		
16	Cooperative	Dominance	
17	Excited	Aggressive	
18	Aggressive	Dominance	
19	Dominance		
20	Excited		

such an interaction. Recognizing boredom and aimless behavior would allow a robot to appropriately encourage a child to take part in a task. Recognizing when a child is being dominant or aggressive could provide a robot with cues to mediate and balance the interaction, or request assistance from a human adult (e.g., in the case of aggressive behavior). Recognizing excitement, fun and cooperation could be used to cue positive feedback from the robot, or to signal that the robot need not interject. The selection was made independently by two of the authors, using a consensus method to reach agreement. It is important to note that interactions in this data set were minimally controlled - pairs of children from the same school class were asked to play on a touch-screen table-top for as long as they wanted. Whilst structured play options were provided, they were not enforced. The selected clips were stored on a private server for the duration of the experiment.

Similarly to the selection of clip labels, the questions were constructed by the experimenters based on the types of internal states and social constructs we might want an artificial system to recognize within a scene. The open question was a single item which asked participants *"What did you notice about the interaction?."* The closed questions were a series of 4 unique questions concerning group dynamics, and 13 2-part questions wherein participants were asked the same question twice, once regarding the child on the left and once regarding the child on the right. Each of these 13 pairs were displayed one after the other. Otherwise, the order in which the questions were presented was random (see **Appendix A** for the questions and response options).

It is important to note that the ground-truth of what internal states the children were experiencing during their interactions is not available. As such, neither the labels used for clip selection and labeling, nor the inferences participants provide in their questionnaire responses can be truly validated. The labels were, therefore, also an attempt to work out what naive observers would infer from the videos.

2.3. Apparatus

The experiment was designed using the jsPsych library[3], and remotely hosted from a private server (**Figure 2** shows a screenshot of the experiment). The experiment was accessible via Amazon Mechanical Turk (MTurk) to MTurk Workers. An advert was posted on MTurk containing a link to the experiment. The remote/online nature of this study means that we had no control over the physical set-up experienced by the participants.

2.4. Procedure

The two video conditions were posted as separate experiments. To ensure that participants did not complete both conditions, the experiments were posted one at a time. Upon opening the experiment participants were asked to provide their MTurk ID and then shown a welcome screen. This was followed by a consent form where participants were asked to provide consent by selecting one of two response options ("I do not consent," or "I do consent"). If participants selected "I do not consent," the experiment would close. If they selected "I do consent" participants were able to press a "Continue" button and proceed to an instruction screen. This was followed by a series of 4 demographic questions (age, nationality, first language and gender). An instruction screen was then presented for a minimum of 3,500 ms, containing the following text:

> *"During this experiment you will be shown 4 30-second clips of children interacting. The children are sat either side of a touch-screen table-top on which they can play a game. Pay particular attention to the way the children interact. After each video you will be asked some questions about what you have watched."*

Participants could then press any button to continue on to the experimental trials.

All participants were asked to complete 4 trials and were presented with the same series of events within each trial. Each trial started with a 30 s clip selected randomly from the list of 20, which was immediately followed by the questions. Upon completion of the fourth trial, participants were shown an additional 2 questions which acted as an attention check (see **Figure 3**). Responses to these questions were used to assess how attentive participants were and how diligently they completed the experiment. Participants who responded incorrectly were excluded from analysis.

Participants then viewed a debrief page which thanked them, explained the purpose of the study and attention-check questions, and provided participants with contact information if they had further questions or desired to withdraw their data. Participants were then provided with a "survey code" which was randomly generated and were instructed that they had completed

[3]https://www.jspsych.org/

Page 1 of 4.

How much do you agree with the following statements?

The children were competing with one another.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was sad.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was sad.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was aggressive.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was aggressive.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The children were cooperating with one another.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the left was excited.

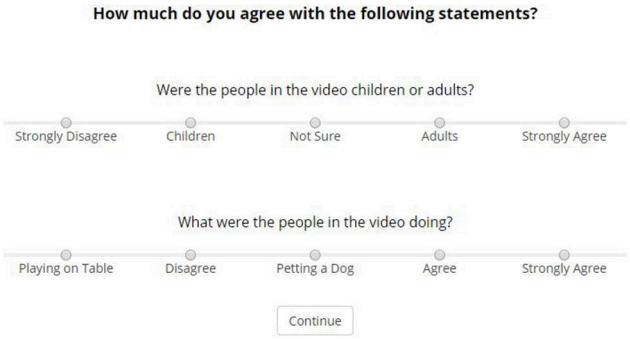
Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

The child on the right was excited.

Strongly Disagree | Disagree | Not Sure | Agree | Strongly Agree

Continue

FIGURE 2 | Screenshot of the online experimental setup showing the questionnaire, just after watching the video clip (here in the full-scene condition). The poster image displayed at the top is a static snapshot of the clip. Written consent for these images to be shared was obtained during collection.



How much do you agree with the following statements?

Were the people in the video children or adults?

Strongly Disagree | Children | Not Sure | Adults | Strongly Agree

What were the people in the video doing?

Playing on Table | Disagree | Petting a Dog | Agree | Strongly Agree

Continue

FIGURE 3 | Capture of attention check questions presented at the end of the questionnaire. Single correct answer provided. Questions and responses are presented in the same format as the rest of the questions in order to test whether participants read the questions.

the experiment and should now return to the MTurk page in order to submit their survey code. The survey codes participants submitted were later compared to those generated to validate participation and payment was authorized via the MTurk system. The experiment took between 20 and 30 min to complete.

The resulting data set is fully anonymous, and made publicly available at https://github.com/severin-lemaignan/pinsoro-kinematics-study/blob/master/fulldata.csv.

3. RESULTS

All data analyses were performed with the Python `pandas` and `sklearn` toolkits. The notebook used for this article, allowing for the replication of our results, is available online, see section 5.

The responses to the open questions revealed no insights beyond those addressed in the specific questions. Therefore, the analyses of these responses are not included in this report.

3.1. Inter-rater Agreement

To determine inter-rater agreement and reliability, we calculated agreement scores across all 30 questions for each clip in each condition separately. This analysis was performed to examine whether participants in each condition gave similar ratings across all questions when they had viewed the same clip. High agreement would indicate that participants had interpreted similar things from a given clip, e.g., participants might all have felt that the children in a clip were being friendly and cooperative, or aggressive and competitive. Whilst this analysis does not reveal exactly what participants interpreted from the videos, it does indicate whether they gave similar ratings, and therefore reported recognizing similar states/behaviors. Given that each clip was rated by a varying subset of participants, Krippendorff's alpha (Hayes and Krippendorff, 2007) was the most appropriate metric of rater agreement (see **Table 3** for number of raters and agreement per clip). The alpha scores ranged from 0.058 to 0.463 i.e., from "slight" to "moderate" agreement (Landis and Koch, 1977).

TABLE 3 | Table of inter-rater agreement scores for responses to each clip in each condition.

Clip	Krippendorff's Alpha (3 d.p.)	
	Full-Scene (N)	Movement Alone (N)
1	0.446 (16)	0.186 (26)
2	0.181 (24)	0.270 (20)
3	0.393 (22)	0.369 (18)
4	0.444 (22)	0.262 (23)
5	0.328 (23)	0.283 (20)
6	0.463 (19)	0.359 (19)
7	0.091 (19)	0.236 (23)
8	0.339 (19)	0.312 (17)
9	0.097 (20)	0.058 (18)
10	0.396 (18)	0.086 (13)
11	0.280 (17)	0.234 (23)
12	0.368 (25)	0.298 (16)
13	0.334 (20)	0.189 (21)
14	0.310 (17)	0.309 (21)
15	0.422 (26)	0.242 (14)
16	0.192 (16)	0.272 (21)
17	0.273 (17)	0.183 (21)
18	0.334 (16)	0.331 (24)
19	0.415 (22)	0.304 (19)
20	0.451 (18)	0.250 (23)

A t-test was conducted to assess whether the two conditions differed in their agreement scores across all 20 clips. This analysis revealed that participants in the full-scene condition showed significantly higher agreement ($M = 0.328$, $SD = 0.110$) than participants in the movement-alone condition ($M = 0.252$, $SD = 0.079$) (Paired Samples T-Test: $t_{(39)} = 2.95$, $p = 0.008$, $d = 0.78$). These analyses show that participants viewing the full-scene clips demonstrated higher levels of agreement in their ratings than those viewing the movement-alone clips. However, participants in the latter condition still showed some agreement compared to chance (chance level Krippendorff's Alpha = 0.0; One Sample T-Test: $t_{(19)} = 13.95$, $p =< 0.001$, $d = 3.12$), suggesting that some internal states and social constructs were recognizable within the movement information in both conditions.

3.2. Automatic Labeling of Internal States

The following analysis explored the question of whether the internal states and social constructs which were available to/inferred by humans when viewing the full visual scene was also available in the movement-alone condition.

We investigated this question using supervised machine learning: would a classifier, trained to label internal states and social constructs from the full-scene ratings, then label the social situations equally well from the movement-alone ratings? If so, this would suggest that the same interaction information was recognized by, and therefore available to, participants in each video condition.

Pre-processing Participants' ratings were coded from 0 (*strongly disagree*) to 4 (*strongly agree*), each construct being recorded as $left_{construct}$ and $right_{construct}$ (see **Appendix A**). Before the following analyses were run, the data from the right-left paired questions was transformed so that results could be more easily interpreted in terms of what behaviors were evident in the interactions, ignoring whether it was the child on the right or the left who was exhibiting this behavior. First, for each question we calculated the absolute difference $diff_{construct} = abs(left_{construct} - right_{construct})$ between the score for the left child and the right child. This score was calculated so that we could more easily see if the children were rated as behaving in the same way, or experiencing similar internal states. Examining the individual scores for each child would have meant that in order to see the dynamics between the children, each clip would have needed to be analyzed separately. Second, for each question we calculated the sum (shifted to the range $[-2, 2]$) $sum_{construct} = left_{construct} + right_{construct} - 4$ of the scores for both children. This score was calculated because the difference score does not contain information about the strength of the rater's belief that the behavior or internal state was evident in the clip. For example, we might have the same difference score for clips where raters believed that both children behaved aggressively and that neither child behaved aggressively. The sum score tells us the degree to which a state was identifiable in the clip.

Multi-label classification To test whether the same interaction information was reported in each video condition we examined whether the ratings from each condition were sufficient to identify the types of internal states or social constructs which were depicted in the videos.

The classifier was trained in a supervised manner, using the 30 ratings provided by the participants (questions from **Appendix A**, pre-processed as indicated above) as input, and the seven labels assigned to each clip during selection (**Table 2**) as the target classification classes. Because the clips could be assigned multiple labels (e.g., a given interaction can be *fun* and *cooperative* at the same time), we used a multi-label classifier (Pieters and Wiering, 2017), using 7-dimensional binary vectors (wherein a zero value denoted that a label was not present in the clip, and a value of one denoted that it was).

We compared the performances of four of classifier (random forest classifier, extra-tree classifier, multi-layer perceptron classifier and a k-Nearest Neighbor classifier, using implementations from the Python `sklearn` toolkit; hyper-parameters were optimized using a grid search where applicable), and eventually selected a k-Nearest Neighbor (with $k = 3$) classifier as providing the best overall classification performance.

Accuracy, precision, recall and F1 score were calculated to assess the performance of the classifier (following recommendations in Sorower (2010) and using the *weighted* implementations of the metrics available in the Python `sklearn` toolkit). Specifically, in the following, *Accuracy* reports the percentage of instances where the predicted labels match exactly with the actual labels; *Precision* is calculated as the ratio $\frac{tp}{tp+fp}$ of true positives divided by the total number of predicted labels (true positives + false positives); *Recall* is calculated as $\frac{tp}{tp+fn}$, i.e. the ratio true positives over the total number of labels that *should* have been found (true positives +

false negatives). Finally, the *F1 score* is the harmonic average of the precision and recall, calculated as $\frac{2precision \cdot recall}{precision+recall}$.

To see how well the classifier performed, we compared performance against chance. Chance levels for these metrics were calculated by training the classifier with randomly generated labels (using the same distribution of labels as found in the real data set), and then measuring the classifier's performance on the actual testing data set.

Results are shown in **Table 4**. In both testing conditions, performance is poor to moderate (for instance 15.8% accuracy for the exact predictions of correct labels in the movement-alone clips), but remain markedly above chance levels (following Ojala and Garriga (2010) permutation-based p-value for classification significance, we found $p = 0.02$ for the full-scene classification, and $p = 0.01$ for the movement-alone classification, ruling out with high probability the null hypothesis that the classification results are due to chance).

Importantly, we found that prediction scores are very similar when testing the classifier on the full-scene ratings or when testing on the movement-alone ratings. This indicates that, from the perspective of automatic data classification, participants who viewed the movement-alone videos were able to report similar details as participants in the full-scene condition. This suggests that the movement-alone videos contain sufficient information to identify different internal states and social constructs.

To identify whether there were particular internal states or social constructs which were easier to recognize than others, the F1 score for each label was calculated. These results are reported in **Table 5** and **Figure 4**. We can see that in both conditions the labels "Bored" and "Aggressive" have higher F1 scores than the other labels. Additionally, the F1 scores for these labels when classifying the full-scene ratings (Bored: 60.0%, Aggressive: 39.0%) are similar to the F1 scores when testing was done on the movement-alone ratings (Bored: 58.5%, Aggressive: 43.7%). This suggests that these constructs are as readily recognized when viewing the full visual scene as when viewing only body movements. In contrast, the F1 score for "Aimless" when testing on full-scene ratings is similar to the scores for most of the rest of the labels (30.3%) but drops to be much lower than any other label when testing was done on the movement-alone ratings (19.4%). This could be interpreted as showing that aimless play, whilst fairly well recognized from the ratings of full visual scene videos, is much harder to recognize from ratings produced when participants viewed only movement information.

This analysis relied on the labels assigned by some of the authors during clip selection. However, participants may have been able to recognize other internal states or social constructs not covered by these labels. In order to investigate possible latent constructs that participants in both conditions may have relied on, we next performed a factor analysis on the dataset.

TABLE 4 | Classification results. *Full-scene* results are obtained by training the classifier on 80% of the full-scene ratings, and testing on the remaining 20%; *Movement-alone* results are obtained by training the classifier on 100% of the full-scene data, and testing on the movement-only ratings.

	Accuracy	Precision	Recall	F1-measure
Full-scene	15.1	44.5	32.0	36.1
Chance	3.7	27.3	14.0	17.4
Movement-alone	15.8	41.6	32.7	36.3
Chance	3.9	28.2	14.2	17.9

Results are averaged over a 300-fold cross-validation. Values are given as percentages.

3.3. Factor Analysis

An Exploratory Factor Analysis (EFA) was performed to explore what types of information participants reported recognizing from the videos. If similar latent constructs are found to underlie participants responses in each condition, this would support the conclusion that participants reported identifying the same types of information in each type of video. Additionally, exploring what factors load into each construct would provide an indication of what these types of information are.

EFA Preliminary assessments revealed a Kaiser-Meyer-Olkin (KMO) statistic of 0.89 and the Bartlett's Test of Sphericity was significant, indicating that the data was suitable for performing an EFA. EFA was performed on the ratings data from each video condition separately to examine what types of interaction information participants were able to draw from the full visual scene compared to movement information alone. We used the

TABLE 5 | F1 scores for each independent label.

	Aggressive	Aimless	Bored	Cooperative	Dominant	Excited	Fun
Full-scene	42.2	29.5	56.6	30.7	37.9	32.2	25.1
Chance	18.8	17.3	11.7	18.2	20.0	18.6	11.4
Movement Alone	43.7	19.4	58.5	29.6	43.4	31.2	27.5
Chance	20.1	16.1	10.7	18.7	19.9	17.3	10.4

*See **Table 4** for the meaning of each row. Values are given as percentages.*

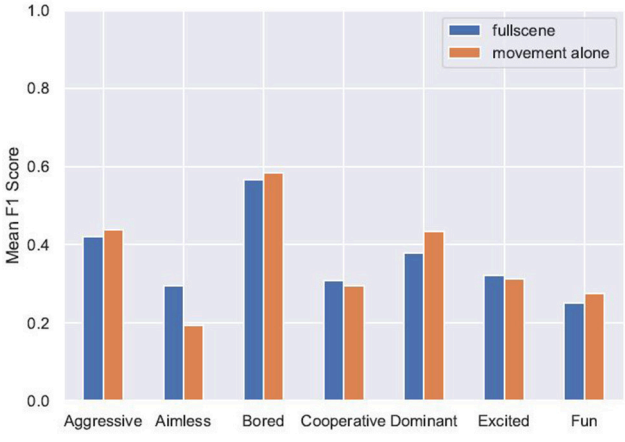


FIGURE 4 | F1 scores of individual label predictions in both conditions.

TABLE 6 | Factor loadings for the three-factor solution using EFA, with factor loadings > 0.35.

	Factor 1: imbalance		Factor 2: valence		Factor 3: engagement	
	Full-scene	*Mov.-alone*	*Full-scene*	*Mov.-alone*	*Full-scene*	*Mov.-alone*
Diff sad	0.41	0.52				
Sum sad			0.72	0.53		0.49
Diff happy	0.49	0.53				
Sum happy				–0.51	–0.55	
Diff angry	0.40	0.62				
Sum angry			0.81	0.85		
Diff excited	0.53	0.63				
Sum excited					–0.71	
Diff calm	0.45	0.63				
Sum calm				–0.45		
Diff friendly	0.69	0.56				
Sum friendly				–0.60	–0.43	
Diff aggressive	0.78	0.79				
Sum aggressive			0.80	0.72	–0.36	
Diff engaged		0.39			0.65	0.52
Sum engaged					–0.64	–0.64
Diff distracted					0.65	0.63
Sum distracted			0.63			0.82
Diff bored		0.44			0.61	0.54
Sum bored			0.58		0.48	0.83
Diff frustrated	0.53	0.61				
Sum frustrated			0.70	0.69		
Diff dominant	0.75	0.81				
Sum dominant			0.53	0.52		
Diff submissive	0.68	0.72				
Sum submissive			0.54			

`factor_analyzer` Python module[4] to perform the EFA, additionally using a *promax* rotation. Three factors were found to explain 44% of the variance in the full-scene ratings, and 46% in the movement-alone ratings. The factor loadings for each component can be seen in **Table 6**.

A Pearson correlation was conducted to examine the similarity of components found in the full-scene and movement-alone ratings. A strong positive correlation was found between each pair of components: for Factor 1: $r = 0.94, p < 0.001$; for Factor 2: $r = 0.84, p < 0.001$; for Factor 3: $r = 0.81, p < 0.001$. This supports the hypothesis that the same latent constructs are relied upon by the participants to rate social interactions, be it based on raw video footage (full-scene) or on a simplified, movement-only, stick-man-style representation (movement-alone).

By inspecting the distribution of factors loadings in **Table 6**, the latent constructs can be further interpreted. It appears that the first component is describing how different the children's behaviors and emotional states are, i.e. this factor describes an *imbalance* in the children,s social, behavioral, and emotional states. For instance, a high value on this scale would show that the children were reported as behaving very differently, e.g., if one child was highly engaged, the other was not very engaged at all.

The second component describes the overall *valence* of the interaction. A high value on this factor would indicate a negative, adversarial interaction where the children were rated as being sad, aggressive etc. Alternatively, a (lower) positive valence value might result from an interaction where one child was rated as being more sad or aggressive than the other child was happy. For both conditions this component has positive correlations with the *Sum* items for negative emotions and behaviors (e.g., Anger, Aggression). For the movement-alone condition, this component also has negative correlations with *Sum* items for positive emotions and behaviors (e.g., Happiness, Friendliness).

The third component is mostly describing the children's *engagement* with their task. In comparison to the other two components it contains more of a mix of *Sum* and *Difference* items, and therefore describes both how similar the children were in how engaged they were, and the overall level of engagement within the interaction. A high value on this third factor would show that the children were rated as showing different levels of engagement, but a strong indication of boredom within the interaction as a whole.

Social Expressiveness of the EFA-Space Embedding One may wonder whether these three factors alone would allow by themselves for an effective assessment of a social interaction, i.e. is the social "expressiveness" of our EFA factors as good as the original 26 factors? This can be investigated by re-applying the same classification methodology as used in section 3.2 to the EFA embedding of the participants' ratings.

To this end, the 26-dimensional participant ratings were projected onto the smaller, 3-dimensional, space spanned by the EFA factors (the *EFA-space*):

$$M^{EFA}_{fullscene} = M_{fullscene} \cdot \Lambda^{EFA}_{fullscene}$$
$$M^{EFA}_{movementalone} = M_{movementalone} \cdot \Lambda^{EFA}_{fullscene}$$

with $M_{fullscene}$ the 396×26 matrix of the participants' ratings, $M^{EFA}_{fullscene}$ the 396×3 matrix of the participants' ratings projected onto the EFA space, and $\Lambda^{EFA}_{fullscene}$ the 26×3 matrix of the EFA factor loadings (**Table 6**). Both the full-scene clips and the movement-alone clips where projected into the same space (spanned by the factors found during the full-scene EFA).

Then, we retrained the same classifier (a kNN with $k = 3$) as in section 3.2, and tried to predict social labels from EFA-projected ratings unseen at training time. **Tables 7**, **8** show the results. We observe a drop of about 4–6% in performance, but still above chance.

4. DISCUSSION AND CONCLUSION

Psychology literature has long established the importance of observing physical group behaviors to provide us with a unique window onto the agents' internal states, as well as the current state of the social interaction. Specifically, we have previous

[4]https://github.com/EducationalTestingService/factor_analyzer

TABLE 7 | Classification results, including classification in EFA-space. *EFA-space* means that the dimensionality of the training and testing data is reduced to 3 by projecting the ratings onto the 3-dimensional space spanned by the EFA factors; non-EFA values copied from **Table 4** for comparison.

	Accuracy	Precision	Recall	F1-measure
Full-scene, EFA	11.2	38.3	26.2	30.0
Full-scene	15.1	44.5	32.0	36.1
Chance	3.8	28.1	14.2	17.8
Movement-alone, EFA	11.7	35.1	27.0	30.3
Movement-alone	15.7	41.6	32.7	36.3
Chance	3.9	28.3	14.2	17.9

Values are given as percentages.

TABLE 8 | F1 scores for each independent label, including after classification in the EFA-space.

	Aggressive	Aimless	Bored	Cooperative	Dominant	Excited	Fun
Fullscene, EFA	37.8	16.2	53.9	29.4	29.7	25.9	20.6
Fullscene	42.2	29.5	56.6	30.7	37.9	32.2	25.1
Chance	19.1	16.5	11.7	19.0	19.6	17.4	11.0
Movement alone, EFA	36.5	24.0	49.2	24.6	33.7	27.4	12.2
Movement alone	43.7	19.4	58.5	29.6	43.4	31.2	27.5
Chance	19.8	16.4	10.7	18.9	19.9	17.9	10.5

*Non-EFA values copied from **Table 5** for comparison. Values are given as percentages.*

evidence of the role of *movements/actions* as an important social signal (Gallese and Goldman, 1998; Alaerts et al., 2011). The main contribution of this paper is to investigate the question of what different states are identified by observers of naturalistic interactions, looking at the (rather messy) social interactions occurring between children while playing together.

This study aimed to examine the kinds of information humans report recognizing from the movements of such naturalistic social interactions. We investigated the following question: is movement information alone (in our case, the moving skeletons of two children playing together, pictured on a uniform black background) sufficient for humans to successfully infer the internal states and social constructs experienced and present within a social interaction? Our methodology involved a between-subject, on-line study, where participants were asked to rate children's behaviors along 17 dimensions, having either watched the raw footage of short interaction videos, or only the skeletons and facial landmarks extracted from the same video clips. This resulted in about 800 unique human ratings, covering both conditions, across 20 different clips, selected for displaying a range of different internal states and social constructs.

We explored the ratings data set (which is publicly available, see the details in the following section) using two main data mining techniques. We first trained a classifier on the full-scene ratings with hand-crafted social labels to then attempt to automatically identify these social labels on the movement-alone ratings. Our results show that training our best performing classifier (a 3-kNN) on 80% of the full-scene ratings and testing on the remaining 20% results in a (cross-validated) precision of 46.2% and recall of 33.6%. We found very similar levels of precision and recall (respectively 41.6 and 32.7%) when testing on the movement-alone ratings: the assessment of the social interaction taking place between two children, made by naive observers watching a low-dimensional, movement-alone video-clip of the interaction, carries similar informational content regarding the internal states and social constructs as the original raw video footage. Based on this finding, we can tentatively conclude that whilst the movement alone videos contain fewer pieces of information, the pieces of information available are as meaningful as those in the full scene videos. Furthermore, we can assess that these pieces of information can be interpreted by human observers in a similar way as those in the full scene videos.

To better make sense of these results, we employed a second data mining technique (Exploratory Factor Analysis, EFA) to attempt to uncover underlying latent factors that would in effect embody stronger cognitive constructs, implicitly relied upon by the humans when assessing a social interaction. We ran independent EFAs on the ratings provided for the full-scene videos and those provided for the movement-alone clips.

To our surprise, the latent factors found by the EFA were strongly correlated between both conditions. In both condition, one factor was measuring the **behavioral imbalance** between the two children (i.e. how similar or dissimilar their behaviors were); a second factor reflected the **valence of the interaction**, from adversarial behaviors and negative emotions, to pro-social and positive behaviors and emotions; finally a third factor embodied **the level of engagement** of the children. These constructs may be indicative of the constructs humans use to interpret social interactions in general. Further research is needed to confirm whether or not this is the case. However, if it is it would provide further insights into how humans approach the interpretation and understanding of social interactions. That is, these three factors may represent the basic cognitive constructs humans use to understand social interactions. Consequently, HRI research could use these constructs as a basic framework for exploring human behavior for classification purposes.

Using the 3-dimensional subspace spanned by these three EFA factors, we have furthermore shown that 'summarizing' the internal states and social constructs inferred by the participants into the 3 latent constructs—imbalance, valence, engagement—only slightly degrades the ability of the classifier to predict the social labels associated with the interaction. This reinforces the hypothesis that these three constructs might play a foundational role in the human understanding of social interactions.

The results of both the classification analysis and EFA demonstrate that it is reasonable to expect a machine learning algorithm, and in consequence, a robot, to successfully decode and classify a range of internal states and social constructs using a low-dimensional data source (such as the movements and poses of observed individuals) as input. Specifically, whilst this study does not examine the ability to identify the correct internal states or social constructs, we have shown that, in a robust way, people agree in their reports of what they have seen both within and between conditions. As such, our study

shows that, even though assessing social interactions is difficult even for humans, using skeletons and facial landmarks only does not significantly degrade the assessment. Future studies aiming to train a robotic system would ideally utilize a training dataset where the internal states and social constructs have been verified (and therefore a ground-truth is available). This study provides the evidence to guide this type of work, for example by demonstrating that training a robot to recognize aggression from movement information is likely to be more successful than recognizing aimlessness.

4.1. Opportunities for Future Work

Given that this work is exploratory in nature, it presents a number of opportunities for future work. First, while above chance, the accuracy of the classifier is relatively low. This may reflect the inherent difficulty of rating internal states and social constructs for an external, naive observer (such as the raters recruited for this study). The literature on emotion recognition does show that humans are able to recognize emotional states from impoverished stimuli with a high level of accuracy [e.g., 44–59% in Alaerts et al. (2011), 59–88% in Gross et al. (2012)]. Similarly, research regarding the recognition of dispositions and social behaviors indicate that computational techniques can achieve a higher recognition accuracy than the current study. For example, Okada et al. (2015) achieved around 57% accuracy in classifying dominance. However, there is some evidence to suggest that humans may not be as accurate as computational classifiers in identifying internal states as we define them here. To demonstrate, Sanghvi et al. (2011) found that whilst human observers were able to recognize engagement to an average of 56% accuracy, their best classifier achieved an 82% level of accuracy. Whilst the accuracy scores presented here are much lower, the existing literature suggests that this may be a result of the fact that humans do seem to demonstrate some difficulty in recognizing these types of states. Additionally, it is important to remember that the classifier in this study labeled the clips using the ratings of all the left/right child questionnaire items, whereas previous research has tended to use the raw visual and/or audio information for classification by both computational systems (Okada et al., 2015) and human observers (Sanghvi et al., 2011). This high dimensional input may have had the effect of diluting the specificity and causing the classifier to use irrelevant or unhelpful inputs when making classification decisions. Additionally, the low classification accuracy may result from the fact that the questionnaire used in this study might not have been good enough. As such, future research would benefit from developing and optimizing the questionnaire.

Additionally, the present study does not explore precisely which movement characteristics were useful for participants in making inferences about the internal states of the children in the videos. In this study we employed a supervised classification technique to demonstrate that social interaction assessments based on full-scene or movement-only stimuli were of similar quality–most notably, our input were ratings of social interactions by human observers. This technique is *not* practically transferable to a robot, as robots would have to directly classify the raw stimuli (a video stream or skeletons), without having access to intermediate ratings of the agents' states. Creating such a classifier is an important next step in deciphering how humans recognize internal states, and therefore in deciding how a robot or classifier can be endowed with a similar skill, for which our present results provide a solid foundation.

The fact that the internal states experienced by the children in the videos could not be validated does present a further limitation for this study. A number of datasets demonstrating one or a subset of the internal states we are interested in are available. For example, the Tower Game Dataset consists of human-human pairs collaborating on a task, and has been annotated for joint attention and entrainment behaviors reflecting cooperation and collaboration (Salter et al., 2015). Similarly, the DAiSEE dataset contains videos of individuals watching videos in an e-learning setting and is annotated for the internal states of boredom, confusion, engagement, and frustration (Gupta et al., 2016). Other datasets include: the UE-HRI annotated for engagement (Ben-Youssef et al., 2017), the ELEA annotated for perceived leadership and dominance (Sanchez-Cortes et al., 2011) among others. Replicating this experiment using a validated dataset may provide stronger classification and inter-rater agreement results. However, few ecologically-valid datasets present the range and variety of internal states as are available in the PInSoRo dataset. As such, this present research represents an important first step in framing the research methodology for analysis of complex, real-life social interactions.

4.2. Conclusion

The aim of this study was to identify social constructs or human internal states which a socially interactive robot could be made to recognize. Analyzing the weighted precision scores for each classification label revealed that "Aggressive" and "Bored" were classified correctly more often in both conditions, whilst "Aimless" was classified correctly much less from the movement-alone ratings. This suggests that training a robot to recognize aimlessness based on movement information might not be as successful as training recognition of boredom. Practically speaking, this finding suggests that designing a tutor robot, such as those used by L2TOR (Belpaeme et al., 2015), to recognize when a child is bored by their task based on movement information would be more successful than having the robot recognize when a child is performing the task in an "aimless" or "non-goal-directed" manner. Such a robot could then appropriately offer encouragement or an alternative task.

Additionally, these findings suggest that exploring other data sources for recognizing human internal states may reveal that certain behavioral modalities may be more useful for recognizing different states. In this way, the method we have demonstrated here can be used to streamline research aimed at teaching robots [and other classification technologies, e.g., automatic classification of security footage (Gowsikhaa et al., 2014)] to recognize human internal states. By applying this method to different types of input data, research can identify the optimal behavioral modality for recognizing a particular human internal state.

These findings have significant impact for both social psychology and artificial intelligence. For social psychology,

it consolidates our understanding of implicit social communication, and confirms previous findings that humans are able to recognize socially relevant information from observed movements (Iacoboni et al., 2005; Alaerts et al., 2011; Quesque et al., 2013). For artificial intelligence, and in particular, for social robotics and human-robot interaction, it provides support for the intuition that low-dimensional (about 100 skeletal and facial points per agent vs. full video frames comprising of hundred of thousands of pixels), yet structured observations of social interactions might effectively encode complex internal states and social constructs. This provides promising support for fast and effective classification of social interactions, a critical requirement for developing socially-aware artificial agents and robots.

5. RESOURCES FOR REPLICATION

Following recommendations by Baxter et al. (2016), we briefly outline hereafter the details required to replicate our findings.

5.1. Study

The protocol and all questionnaires have been provided in the text. The code of the experiment is available at https://github.com/severin-lemaignan/pinsoro-kinematics-study/. Note that, due to data protection regulations, the children' video clips are not available publicly. However, upon signature of an ethical agreement, we can provide them to the interested researcher.

5.2. Data Analysis

The full recorded experimental dataset, as well as the complete data analysis script allowing for reproduction of the results and plots presented in the paper (using the Python *pandas* library) are open and available online, in the same Git repository. In particular, a iPython notebook with all the steps followed for our data analysis is available here: https://github.com/severin-lemaignan/pinsoro-kinematics-study/blob/master/analysis/analyses_notebook.ipynb.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Ethics guidelines of the University of Plymouth Ethics Committee, with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of the University of Plymouth.

AUTHOR CONTRIBUTIONS

MB, CE, and SL contributed to the design, running, analysis, and write-up of this study. ST and TB contributed to the supervision and funding of this study.

FUNDING

This work is part of the EU FP7 project DREAM project (www.dream2020.eu), funded by the European Commission (grant no. 611391). It has received additional funding by the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant no. 657227).

REFERENCES

Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S. P., and Wenderoth, N. (2011). Action and emotion recognition from point light displays: an investigation of gender differences. *PLoS ONE* 6:e20989. doi: 10.1371/journal.pone.0020989

Ansuini, C., Cavallo, A., Bertone, C., and Becchio, C. (2014). The visible face of intention: why kinematics matters. *Front. Psychol.* 5:815. doi: 10.3389/fpsyg.2014.00815

Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., and Belpaeme, T. (2016). "From characterising three years of HRI to methodology and reporting recommendations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation* (Christchurch: IEEE Press), 391–398.

Becchio, C., Koul, A., Ansuini, C., Bertone, C., and Cavallo, A. (2017). Seeing mental states: an experimental strategy for measuring the observability of other minds. *Phys. Life Rev.* 24, 67–80. doi: 10.1016/j.plrev.2017.10.002

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E. J., Kopp, S., et al. (2015). "L2TOR-second language tutoring using social robots," in *Proceedings of the ICSR 2015 WONDER Workshop* (Paris).

Ben-Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M., and Lim, A. (2017). "Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (New York, NY: ACM), 464–472. doi: 10.1145/3136755.3136814

Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., et al. (2016). "Detecting emergent leader in a meeting environment using nonverbal visual features only," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (New York, NY: ACM), 317–324. doi: 10.1145/2993148.2993175

Breazeal, C., Gray, J., and Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *Int. J. Robot. Res.* 28, 656–680. doi: 10.1177/0278364909102796

Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR* (Honolulu, HI).

Dautenhahn, K., and Saunders, J. (2011). *New Frontiers in Human Robot Interaction*, vol 2. Amsterdam: John Benjamins Publishing.

Domes, G., Heinrichs, M., Michel, A., Berger, C., and Herpertz, S. C. (2007). Oxytocin improves mind-reading in humans. *Biol. Psychiat.* 61, 731–733. doi: 10.1016/j.biopsych.2006.07.015

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* 17:124.

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* 2, 493–501.

Gowsikhaa, D., Abirami, S., and Baskaran, R. (2014). Automated human behavior analysis from surveillance videos: a survey. *Artif. Intell. Rev.* 42, 747–765. doi: 10.1007/s10462-012-9341-3

Gross, M. M., Crane, E. A., and Fredrickson, B. L. (2012). Effort-shape and kinematic assessment of bodily expression of emotion during gait. *Hum. Move. Sci.* 31, 202–221. doi: 10.1016/j.humov.2011.05.001

Gupta, A., D'Cunha, A., Awasthi, K., and Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv* arXiv:1609.01885.

Haidt, J., and Keltner, D. (1999). Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. *Cogn. Emot.* 13, 225–266.

Han, J.-H., and Kim, J.-H. (2010). "Human-robot interaction by reading human intention based on mirror-neuron system," in *2010 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (Tianjin: IEEE), 561–566.

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* 1, 77–89. doi: 10.1080/19312450709336664

Hufschmidt, C., Weege, B., Rder, S., Pisanski, K., Neave, N., and Fink, B. (2015). Physical strength and gender identification from dance movements. *Pers. Individ. Diff.* 76, 13–17. doi: 10.1016/j.paid.2014.11.045

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 Biol.* 3, 0529–0535. doi: 10.1371/journal.pbio.0030079

Kawamura, R., Toyoda, Y., and Niinuma, K. (2019). "Engagement estimation based on synchrony of head movements: application to actual e-learning scenarios," in *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (New York, NY: ACM), 25–26.

Kozlowski, L. T., and Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Percept. Psychophys.* 21, 575–580.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

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* arXiv:1712.02421.

Manera, V., Becchio, C., Cavallo, A., Sartori, L., and Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Exp. Brain Res.* 211, 547–556. doi: 10.1007/s00221-011-2649-4

Manera, V., Schouten, B., Becchio, C., Bara, B. G., and Verfaillie, K. (2010). Inferring intentions from biological motion: a stimulus set of point-light communicative interactions. *Behav. Res. Methods* 42, 168–178. doi: 10.3758/BRM.42.1.168

Mather, G., and Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. B* 258, 273–279.

Ojala, M., and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* 11, 1833–1863. doi: 10.1109/ICDM.2009.108

Okada, S., Aran, O., and Gatica-Perez, D. (2015). "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (New York, NY: ACM), 15–22.

Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., and Esme, A. A. (2017). "Behavioral engagement detection of students in the wild," in *International Conference on Artificial Intelligence in Education* (Wuhan: Springer), 250–261.

Pieters, M., and Wiering, M. (2017). "Comparison of machine learning techniques for multi-label genre classification," in *Benelux Conference on Artificial Intelligence* (Groningen: Springer), 131–144.

Pollick, F. E., Paterson, H. M., Bruderlin, A., and Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition* 82, B51–B61. doi: 10.1016/S0010-0277(01)00147-0

Quesque, F., Lewkowicz, D., Delevoye-Turrell, Y. N., and Coello, Y. (2013). Effects of social intention on movement kinematics in cooperative actions. *Front. Neurorobot.* 7:14. doi: 10.3389/fnbot.2013.00014

Salter, D. A., Tamrakar, A., Siddiquie, B., Amer, M. R., Divakaran, A., Lande, B., and Mehri, D. (2015). "The tower game dataset: a multimodal dataset for analyzing social interaction predicates," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an: IEEE), 656–662.

Sanchez-Cortes, D., Aran, O., Mast, M. S., and Gatica-Perez, D. (2011). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. Multi.* 14, 816–832. doi: 10.1109/TMM.2011.2181941

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of the 6th international conference on Human-robot interaction* (New York, NY: ACM), 305–312.

Schrempf, O. C., and Hanebeck, U. D. (2005). "A generic model for estimating user intentions in human-robot cooperation," in *ICINCO* (Barcelona), 251–256.

Shaker, N., and Shaker, M. (2014). "Towards understanding the nonverbal signatures of engagement in super mario bros," in *International Conference on User Modeling, Adaptation, and Personalization* (Aalborg: Springer), 423–434.

Sorower, M. S. (2010). *A Literature Survey on Algorithms for Multi-Label Learning.* Corvallis: Oregon State University.

Tracy, J. L., and Robins, R. W. (2008). The nonverbal expression of pride: evidence for cross-cultural recognition. *J. Personal. Soc. Psychol.* 94:516. doi: 10.1037/0022-3514.94.3.516

Vernon, D., Thill, S., and Ziemke, T. (2016). "The role of intention in cognitive robotics," in *Toward Robotic Socially Believable Behaving Systems-Volume I* (Springer), 15–27.

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: differentiation of multimodal information. *Psychol. Bull.* 121:437.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

A. APPENDIX

A.1. Questions

Open Question: "What did you notice about the interaction?"

Specific Questions: For all of the following questions participants were asked to report how much they agreed with each statement. Answers : *Strongly Disagree / Disagree / Not Sure / Agree / Strongly Agree*

1. "The children were competing with one another."
2. "The children were cooperating with one another."
3. "The children were playing separately."
4. "The children were playing together."

6-7 "The character on the left/right was sad."
8-9 "The character on the left/right was happy."
10-11 "The character on the left/right was angry."
12-13 "The character on the left/right was excited."
14-15 "The character on the left/right was calm."
16-17 "The character on the left/right was friendly."
17-18 "The character on the left/right was aggressive."
19-20 "The character on the left/right was engaged with what they were doing on the table."
21-22 "The character on the left/right was distracted from the table."
23-24 "The character on the left/right was bored."
25-26 "The character on the left/right was frustrated."
27-28 "The character on the left/right was dominant."
29-30 "The character on the left/right was submissive."

Appendix D

HRI 2019 Workshop Paper - Recognizing Human Internal States: A Conceptor-Based Approach

This paper was presented in a workshop at the *2019 IEEE/ACM International HRI Conference* and published under the Creative Commons Attribution License (Bartlett et al., 2019a).

Recognizing Human Internal States: A Conceptor-Based Approach

1st M E Bartlett
University of Plymouth
Plymouth, United Kingdom

2nd Daniel Hernández García
University of Plymouth
Plymouth, United Kingdom

3rd Serge Thill
University of Skövde
Skövde, Sweden
Donders Institute
Nijmegen, The Netherlands

4th Tony Belpaeme
University of Ghent
Ghent, Belgium
University of Plymouth
Plymouth, United Kingdom

***Abstract*—The past few decades has seen increased interest in the application of social robots to interventions for Autism Spectrum Disorder as behavioural coaches [4]. We consider that robots embedded in therapies could also provide quantitative diagnostic information by observing patient behaviours. The social nature of ASD symptoms means that, to achieve this, robots need to be able to recognize the internal states their human interaction partners are experiencing, e.g. states of confusion, engagement etc. Approaching this problem can be broken down into two questions: (1) what information, accessible to robots, can be used to recognize internal states, and (2) how can a system classify internal states such that it allows for sufficiently detailed diagnostic information? In this paper we discuss these two questions in depth and propose a novel, conceptor-based classifier. We report the initial results of this system in a proof-of-concept study and outline plans for future work.**

***Index Terms*—Internal States, Engagement, Conceptors, Socially Interactive Robots, Recognition**

I. Introduction

The development of socially interactive robots has inspired research into various applications for these tools. One application is in therapy and care, where robots can be used to provide daily support to patients, and as tools to augment interventions and provide quantitative data for clinicians [1]. We specifically consider the use of robots in interventions for children with Autism Spectrum Disorder (ASD). The Diagnostic and Statistical Manual of Mental Disorders (DSM-V) defines ASD as a neuro-developmental disorder characterized by persistent deficits in social communication and interaction, and restricted or repetitive behaviours and interests [2]. Diagnosing ASD involves the subjective interpretations by experts of observations of a child's behaviour made by clinicians and caregivers [3]. This subjectivity, and the clinical heterogeneity typical between ASD cases [4], means that the diagnostic process could be improved through the use of more quantitative, objective measures of child behaviour. This can be achieved using behaviour classification systems.

Developing an artificial system to recognize ASD symptoms is not a straight-forward task due to the social nature of ASD. This is because correct classification of social and interaction behaviour often requires the ability to infer the internal-states (e.g. intentions, emotions) of the observed individual. For example, identifying when a child fails to ask for comfort when needed (a symptom of ASD [2]) requires that the observer recognize that the child is experiencing a negative internal state. However, endowing robots with this skill would provide numerous benefits for ASD interventions. For instance, if an intervention involves regular interaction with a social robot, it would be useful to have the robot able to report quantitative diagnostic information. Firstly, clinicians could use this information to track their patient's progress through the intervention, or to support their initial diagnostic decision. Secondly, the robot itself could use internal-state and diagnostic information to autonomously decide on appropriate behaviours to perform.

In approaching the problem of developing artificial systems able to recognize human internal states, there are two key questions which must be addressed: (1) what internal state information is available in behaviours which can be assessed and quantified by artificial systems, and (2) how can these states be represented by a classification system to provide both detailed assessments and flexible behavioural responses from a social robot. The rest of this paper discusses possible answers to these questions in the context of quantifying the diagnostic behaviours of children with ASD. We present two studies carried out as a proof-of-concept to demonstrate that the internal state of task engagement could be classified based on observable human movement information, and that this classification could be done by a system able to represent internal states as points along a continuous dimension. The logic behind our choice of internal state and its desired representation is described, where relevant, in the introductions to each experiment.

II. Experiment 1

Whilst most ASD symptoms cannot be described as wholly overt, many have been linked with directly observable behaviours. For example, motor skills have been shown to be predictive of social communication skills for children with ASD [5]. Additionally, an increased tendency to orient towards non-social contingencies rather than biological motion is indicative of ASD [6]. These and other studies have linked movement and gaze behaviours to several ASD characteristics. Movement and gaze information can be measured or estimated by observing body movements or poses, which can be easily

made accessible to artificial systems, e.g. by converting the position of an individual's joints to coordinates in space. Consequently, we argue that such observable information can be useful for social robots designed to make inferences about human internal states pertaining to ASD symptoms.

Designing a system to recognize this kind of diagnostic information, however, is non-trivial. We would need to have identified how observable behaviours relate to symptomology, and define which symptoms we are best able to recognize and describe in terms of severity based on behavioural data. Given the complexity of obtaining and labelling such data, we chose to perform a proof-of-concept study demonstrating the feasibility of our approach using data from a non-clinical population. We therefore chose to examine whether the internal state of task engagement could be identified and classified into different classes, based on the 'intensity' of the experienced state. That is, we aimed to train a classifier to distinguish between 'high', 'intermediate' and 'low' task engagement based on the behaviour of typically developing children. Before a classifier could be implemented, however, we first needed to verify that the internal state of interest (task engagement) was recognizable from the movement information available in our data set.

For this study, the desired data set was defined as one which contained the movement information of humans experiencing, but not intentionally communicating, different levels of a non-emotional internal state. To ensure that the internal state was not being communicated we decided that the subject should not be interacting with another human. With these considerations in mind, the data set for this experiment was taken from the openly available PInSoRo data set [7][1]. This data set comprises videos of child-robot pairs interacting with each other and a touch-screen table-top (the sand-tray). We argue that these videos meet the requirements of showing humans experiencing internal states which could be described along a continuum (i.e. engagement with the touch-screen) which were not being actively communicated (i.e. due to the lack of a human interaction partner). The videos have been annotated for a number of behaviors including whether the child was engaged in "goal oriented", "aimless" or "no" play. We believe these annotations are analogous to "high", "intermediate" and "low" levels of task engagement respectively. A preliminary study was designed to validate this assumption.

A. Method

1) Participants: Five participants (students and employees) were recruited from the University of Plymouth's School of Computing, Electronics and Mathematics on a volunteer basis. Demographic information was not collected.

2) Materials: A total of forty-five video clips were extracted from the data set for this study. We selected fifteen clips with the annotation "goal-oriented play", fifteen with the annotation "aimless play" and fifteen with the annotation "no play". Clip lengths ranged from 12-30 seconds.

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

After clips were selected we extracted both the full visual scene versions and the movement-alone versions. The movement-alone versions were processed such that they depicted the children's joint-points, connected by coloured lines, against a black background. These videos act as visual representations of the data used as input for the conceptor-based system in that they depict only movement and pose information by showing the position of the child's body in each frame.

3) Procedure: For each participant the experiment was conducted over two days. Participants watched the full visual scene videos on the first day and were then asked to return the next day when they would watch the movement-alone videos. Participants all received the following instructions before beginning the experiment:

You're about to watch several videos of children interacting with a touch-screen table-top. The children were able to either play a specific game on the touch-screen, or to do whatever they want. After each clip you will be asked to judge the child's level of task engagement.

Participants were then given the opportunity to ask any questions they may have had and were instructed about their right to withdraw before beginning the experiment.

This study was created using JSPsych and presented on a desktop computer. Participants were positioned a comfortable distance away from the screen where they could still reach the keyboard and mouse to provide responses. At the beginning of the experiment, the instructions were reiterated. Participants were then presented with a consent form within the experiment script and given two response options. If participants selected the "I consent" option, the experiment proceeded as normal. If participants selected "I do not consent" the experiment was terminated. Participants then viewed nine of each type of clip (a total of twenty-seven clips) presented in a random order. Following each clip, participants were presented with the question "*How engaged was the child with their task on the touch screen table-top?*". This question was accompanied by a 7-point Likert scale ranging from 1 = "Not at all Engaged" to 7 = "Highly Engaged". Participants used this scale to report how engaged they thought the child in the clip had been and then continued on to the next clip.

At the end of the experiment on the first day, participants were given the opportunity to ask any questions they had and were asked to return the next day to complete the second half. On the second day, the experiment proceeded in the same way except participants were shown the movement-alone videos instead of the full visual scene videos. Each participant saw the same twenty-seven clips in both sessions. At the end of the second session participants were fully debriefed on the nature and purpose of the study and were thanked for their participation. Each session took approximately 10-15 minutes to complete.

B. Results

The following analyses were run using RStudio.

1) Inter-Rater Agreement: The data were analyzed in two main ways. We firstly examined inter-rater agreement by calculating Krippendorff's alpha for the responses. We initially checked whether participants gave similar responses for each of the three types of videos. To do this, Krippendorff's alpha was calculated for responses to all of the videos of each type. The alpha scores have been interpreted in terms of the benchmarks outlined by Landis and Koch [8]. Responses showed "fair" agreement for the goal-oriented (high engagement) clips (Krippendorff's alpha = 0.269) and the no-play (low engagement) clips (Krippendorff's alpha = 0.267). Responses for aimless (intermediate engagement) clips showed "slight" agreement (Krippendorff's alpha = 0.171). The low levels of agreement can partially be explained by the fact that there were very few raters (2-4) per clip. As such we did not expect perfect levels of agreement and argue that the levels obtained suggest a sufficient degree of similarity in participants' ratings.

We then examined whether participants had higher agreement when viewing the full visual scene clips compared to the movement-alone clips for each clip type. The results of this analysis are reported in Table 1. For the goal-oriented and no-play clips, participants tended to show similar levels of agreement in each condition. However, for the aimless clips, participants demonstrated poor agreement when viewing the movement-alone clips.

TABLE I
TABLE OF INTER-RATER AGREEMENT SCORES FOR RESPONSES TO EACH CLIP-TYPE IN EACH CONDITION

Clip Type	Krippendorff's Alpha (3 d.p.)	
	Full Scene	movement-alone
Goal Oriented	0.382 (fair)	0.368 (fair)
Aimless	0.247 (fair)	-0.022 (poor)
No Play	0.126 (slight)	0.202 (fair)

2) Ratings: The second set of analyses looked at the how participants rated each type of video. Overall mean rating was 4.81 (SD = 1.25) for goal-oriented clips, 4.16 (SD = 1.52) for aimless clips, and 2.43 (SD = 1.54) for no-play clips. An ANOVA revealed a significant main effect of clip-type on ratings ($F(2,267)=64.99$, $p<0.001$). Importantly, a *post hoc* Tukey test revealed significant differences between all conditions (Tukeys HSD: all differences >0.6, all ps <0.007; see Table 2).

TABLE II
TABLE OF RESULTS FOR POST HOC TUKEY'S HONEST SIGNIFICANT DIFFERENCE TEST.

Comparison	Difference	Significance (p adj)
Goal Oriented − Aimless	0.656	$p = 0.007$
Goal Oriented − No Play	2.348	$p < 0.001$
Aimless − No Play	1.722	$p < 0.001$

These results demonstrate that participants rated the clips in terms of engagement such that goal-oriented clips showed the highest levels of engagement, no-play clips showed the lowest levels, and aimless clips fell in-between these two extremes. Consequently, we feel our assumption that these annotations reflect different levels of engagement is sufficiently supported for these data to be used to train and test a conceptor-based classifier to recognize engagement based on observable behaviour. The remainder of this paper describes the design and initial tests of such a classifier.

III. EXPERIMENT 2

In addressing the second question of how to represent internal states, we consider that ASD diagnosis involves ranking behaviours in terms of severity [9]. In this way, behaviours important to ASD diagnosis can be thought of as lying along a continuum of severity. To emulate this we need a classification technique which can identify different 'levels' along a continuous dimension. This can be achieved using classical machine learning techniques by training a classifier on examples of each severity level. However, obtaining a large enough training data set for this would be very time-consuming and difficult, owing to the need to have expert commentators provide a severity label for each example. We therefore require a method which can learn several classification categories for each behaviour of interest, using a limited training data set. One approach which is suited to this task is conceptors [10].

Conceptors are neuro-computational mechanisms that can be used for learning a large number of dynamical patterns based on learned prototypical extremes [10]. This approach assumes that there is a continuum underlying the behavior. New patterns can be generated by combining and morphing the learned extremes. As such, we argue that conceptors may be appropriate for classifying human internal states. The second study described here tested this hypothesis by designing a conceptor-based system to recognize task engagement from observable human movements.

A. Method

1) Materials: The data set for this study was again taken from the PInSoRo data set. All of the clips annotated with the labels "goal-oriented play" (high engagement) and "no play" (low engagement) were extracted (total of 354 clips). Clips were preprocessed such that the xyz coordinates of the child's joints in each frame were taken as the input for the conceptor-based classifier. A subset of "high" (62 clips) and "low" (115 clips) engagement clips made up the training data set. The remaining 177 clips made up the test data set.

2) Conceptor-Based Classifier: The conceptor-based approach is based on a key dynamical phenomenon in Recurrent Neural Networks; "if a 'reservoir' is driven by a pattern, the entrained network states are confined to a linear subspace of network state space which is characteristic of the pattern" [10]. In this way the dynamics of a pattern (in our case an overt behavior for a classifiable activity like engagement) will occupy different regions of the state space, and can be encoded in a conceptor. A conceptor (C_j) acts as a map associated with a pattern (p_j). To build a conceptor-based classifier we computed J conceptors, one for each class in our classifier. To obtain the conceptors an echo state network (ESN) was

TABLE III
PREDICTING INTERNAL STATES WITH CONCEPTORS.

Algorithm: Conceptor-based classification.
Input: A test sample s belonging to one a class j.

1) Take a sample s from the test set.
2) Drive the reservoir with sample s to obtain a state vector $z = [x(1) \cdots x(n)]$, where n is the # of steps in s.
3) For each Conceptor C_j compute $h(j) = z^T C_j z$, a "positive evidence" quantity of z belonging to class j.
4) Collect each evidence $h(j)$ into a $j-$dimensional classification hypothesis vector $h^+ = \{h(1) \cdots h(j)\}$.
5) Classify s as belonging to class j from $j = argmax(h^+)$.
6) **END**

Output: Class sample s belongs to.

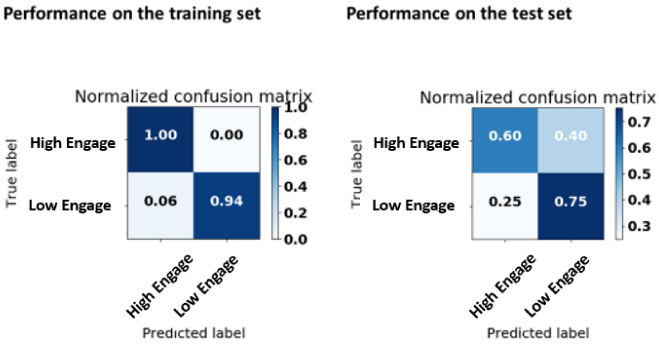


Fig. 1. Confusion matrices showing classification performance of trained conceptors on training data (left) and test data (right).

first created with an input layer of K input units and a hidden layer reservoir of N neurons. For each class the network will be driven, independently, with all training samples s_j^m in each class j, according to the ESN state update equation:

$$x(n+1) = \tanh(W \cdot x(n) + W^{in} \cdot p(n+1) + b) \qquad (1)$$

This yielded a set of network states $X_j = [x(1) \dots x(t)]$ where t is the number of time-steps in s_j from which a state correlation matrix $R_j = X_j X_j^T / M_j$ is obtained, where M_j is the total number of samples for class j. Next we computed conceptor C_j through the equation:

$$C(R, \alpha) = R(R + \alpha^{-2})^{-1} \qquad (2)$$

Where R is a correlation matrix and $\alpha \in (0, \infty)$ in an "apperture" parameter. For more see [10].

Once we computed a conceptor matrix for each class we were able to classify a new sample s from the test set by feeding it into the ESN reservoir to obtain a new state vector $z = [x(1) \dots x(n)]$. then, for each conceptor, the "positive evidence" quantity $z^T C_j z$ was computed. This led to a classification by deciding for $j = argmax(z^T C_j z)$ as the class j to which the sample s belongs. The procedure for the conceptor-based classifier is summarize in Table III-A2.

B. Results

The resultant conceptors were tested using previously unseen samples from the high and low engagement categories. The results of this test are shown in Figure 1. Performance is above chance for both classes (high engagement: 60%, low engagement: 75%).

IV. CONCLUSIONS

This study demonstrates that it is possible to train a conceptor-based system, on real non-periodic data, to classify between high and low engagement based on observable human behavior. The conceptor-based system successfully learned to recognize high and low engagement from observable human movement. Future work will construct new conceptors by linearly combining these learned conceptors. We will then test whether these new conceptors can be used to recognize intermediate levels of engagement identified in the PInSoRo data set.

If new conceptors can be generated, this method will show promise for use in providing diagnostic information for clinicians assessing children with ASD. The ability to interpolate between extremes along a continuum means that such a system could be trained on a smaller dataset, whilst still achieving a high level of detail through the generation of multiple intermediate classification categories.

ACKNOWLEDGMENT

This work is part of the EU FP7 project DREAM project (www.dream2020.eu, grant nr. 611391) and the H2020 L2TOR project (www.l2tor.eu, grant nr. 688014).

REFERENCES

[1] H. M. Van der Loos, D. J. Reinkensmeyer, AND E. Guglielmelli, "Rehabilitation and health care robotics." In Springer handbook of robotics, pp. 1685-1728, 2016.
[2] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders: DSM-5." Autor, Washington DC, 5th ed, 2013.
[3] C. L. Rogers, L. Goddard, E. L. Hill, L. A. Henry, and L. Crane, "Experiences of diagnosing autism spectrum disorder: a survey of professionals in the United Kingdom." Autism, vol. 20(7), pp. 820–831, 2016.
[4] B. Scassellati, H. Admoni, and M. Matari, "Robots for use in autism research." Annual review of biomedical engineering, vol. 14, pp. 275–294, 2012.
[5] J. Bradshaw, C. Klaiman, S. Gillespie, N. Brane, M. Lewis, and C. Saulnier, "Walking Ability is Associated with Social Communication Skills in Infants at High Risk for Autism Spectrum Disorder." Infancy, 2018.
[6] A. Klin, D. J. Lin, P. Gorrindo, G. Ramsay and W. Jones, "Two-year-olds with autism orient to non-social contingencies rather than biological motion." Nature, vol. 459(72440, pp. 257–261, 2009.
[7] S. Lemaignan, C. Edmunds, E. Senft, and T. Belpaeme, "The Free-play Sandbox: a Methodology for the Evaluation of Social Robotics and a data set of Social Interactions." arXiv preprint arXiv:1712.02421, 2017.
[8] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data." Biometrics, pp. 159–174, 1977.
[9] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles and M. Rutter, "The Autism Diagnostic Observation ScheduleGeneric: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism." Journal of Autism and Developmental Disorders, vol. 30(3), 2000.
[10] H. Jaeger, "Controlling recurrent neural networks by conceptors." arXiv preprint arXiv:1403.3369, 2014.

Appendix E

HRI 2018 Workshop Paper - Towards a Full Spectrum Diagnosis of Autistic Behaviours using Human Robot Interactions

This paper was presented in a workshop at the *2018 IEEE/ACM International HRI Conference* and published in their proceedings with a non-exclusive and irrevocable license to distribute the article (Esteban et al., 2018; Bartlett, Belpaeme, and Thill, 2018).

Towards a Full Spectrum Diagnosis of Autistic Behaviours using Human Robot Interactions

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

Tony Belpaeme
University of Plymouth
Plymouth PL4 8AA
tony.belpaeme@plymouth.ac.uk

Serge Thill
University of Plymouth
Plymouth PL4 8AA
serge.thill@plymouth.ac.uk

ABSTRACT

Autism Spectrum Disorder (ASD) is conceptualised by the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [1] as a spectrum, and diagnosis involves scoring behaviours in terms of a severity scale. Whilst the application of automated systems and socially interactive robots to ASD diagnosis would increase objectivity and standardisation, most of the existing systems classify behaviours in a binary fashion (ASD vs. non-ASD). To be useful in interventions, and to overcome ethical concerns regarding overly simplified diagnostic measures, a robot therefore needs to be able to classify target behaviours along a continuum, rather than in discrete groups. Here we discuss an approach toward this goal which has the potential to identify the full spectrum of observable ASD traits.

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is defined by the DMS-V in terms of two behavioural domains: social communication and interaction, and restricted or repetitive behaviours and interests [1]. Recent advances in our understanding have led to the re-conceptualisation of ASD as a spectrum. This concept refers to: (1) differences in presentation and severity within the clinical population, (2) the continuous distribution of "autistic traits" between the general and clinical populations, and (3) subgroups [6]. Diagnosis of ASD cannot, therefore, be thought of as a binary classification (e.g. non-ASD vs. ASD) but rather in terms of severity scales applied to multiple behaviours and traits. Diagnosis thus relies largely on subjective interpretations of various sources of information [2, 10], and children with ASD demonstrate high levels of clinical heterogeneity [4, 11]. The diagnostic standard of ASD could, therefore, be improved by more quantitative, objective measures of social response.

These benefits can be provided by introducing automated systems into the diagnostic process in the form of socially interactive robots [3], and systems to aid in the diagnosis of several behavioural and psychological disorders including ASD [7, 12] have been developed. However, in contrast with the diagnostic requirements, these systems usually approach behaviour classification in a binary fashion; individuals are classed as either *ASD* or *non-ASD* [12]. This lack of sensitivity to intermediate cases brings with it the ethical issues of overly simplified diagnostic measures, such as potentially classifying a large proportion of the behaviours which fall on the autism spectrum as non-ASD [7]. Here, we discuss an approach toward, and the benefits of, non-binary, automated classification of autistic behaviours embedded within human-robot interactions.

2 ROBOTS AS DIAGNOSTIC TOOLS FOR ASD

The prospect of introducing robots into interventions for ASD has become increasingly popular due to findings indicating that robots can promote motivation, engagement, and the occurrence of otherwise rare social behaviours in children with ASD [2, 14]. They have therefore been proposed as an effective tool for helping children develop and employ social skills, and to transfer these skills to interactions with humans [2, 13]. Whilst less attention has been given to the role of robots in ASD diagnosis [14], such an application of robot technology does offer unique benefits including: (1) standardisation of stimulus and recording methodology, and (2) increased repeatability [2, 8]. It has also been argued that a robot's ability to generate social prompts allows for the controlled elicitation and examination of social responses [2]. This is in-line with the goal of diagnostic instruments such as the Autism Diagnostic Observation Schedule (ADOS) [5], i.e. to elicit spontaneous behaviours in a standardised context. Furthermore, the finding that children with ASD interact more with technology than with humans [8] indicates that having a child interact with a robot during assessment may facilitate the production of a wider range of behaviours. This facilitation could, in turn, provide richer data for the purposes of diagnostic analysis [14].

On-line behaviour adaptation is important for autonomous robots in ASD interventions due to the high variability seen between children with ASD [3]. This process requires the system to track and classify the child's behaviour before appropriate responses can be selected. However, many systems which are used to classify behaviours in therapeutic settings are limited to simple, easily distinguished classes; they do not identify intermediate classes [12]. Wall and colleagues [12] used a subset (8 out of 29) of behaviours coded from ADOS to design a diagnostic algorithm which could differentiate between children with and without ASD. Whilst the algorithm could classify cases correctly, Wall and colleagues simplified the problem by removing the middle diagnostic classes, leaving only ASD and non-ASD. As a result, individuals who fall in the middle of the ASD spectrum were identified as non-ASD. Furthermore, an attempt to replicate these findings found that the algorithm was not robust enough to deal with a different dataset and a larger group of coded behaviours was required to identify individuals diagnosed as being in a mid-spectrum ASD class [7].

The spectrum nature of ASD means that to avoid under-identification and to allow the system to provide useful information for decisions about therapeutic approaches, classes of behaviour which do not fall at the extremes of the spectrum, e.g. High-Functioning Autism, should be identifiable. Contemporary approaches to non-binary classification are rare. Bone and colleagues [7] used a similar machine learning method to that of [12],

but incorporated all the behaviour codes from ADOS which made the classification system more robust and more accurate. Including the middle diagnostic classes did decrease the accuracy but it still remained high (i.e. 96% dropped to 82%). However, this approach is still labor intensive and time-consuming, and is designed to be run off-line using data collected by the clinician.

3 CLASSIFYING CONTINUOUS BEHAVIOURS USING CONCEPTORS

For a classification system to accurately identify the intermediate classes of ASD, it must be able to classify behavioural patterns ranging from "typical of the general population" to "severely atypical". This can be achieved using purely machine learning methods. However, this requires a large, representative data-set which is often difficult and time-consuming to obtain due to the need to annotate the training data-sets. We therefore require a methodology that can deal with the spectrum nature of ASD by representing behaviours over continuous dimensions, and which requires less data for learning than traditional machine learning methods. One approach is to use conceptors [9]; neuro-computational mechanisms that can be used for learning a large number of dynamical patterns. Conceptors can also be combined and morphed to generate new patterns based on learned prototypical extremes along a behavioural continuum, e.g. a system given the prototypes for "walking" and "running" can generate patterns for "jogging" [9]. This approach assumes that there is a continuum underlying the behaviour, which is well suited to the symptomology of ASD [1], as demonstrated by ADOS [5] which scores behaviours such as speech abnormalities on a scale of 0 ("no evidence of abnormality") to 3 ("markedly abnormal").

To represent the spectrum nature of ASD using conceptors, a recurrent neural network can be provided, for example, with the prototype patterns for typical and markedly abnormal speech behaviour. Relevant information from these input patterns are then represented as the internal state of the system. These internal states are then used for classification, rather than the inputs themselves. Conceptors can be computed to represent the state of each dimension of speech (volume, intonation, stress, etc.) within each pattern, and clustered to form groups. These groups represent the key components of the behavioural continuum which are described by the prototype patterns provided. Morphing of these patterns using linear mixes of the prototype conceptors allows the system to interpolate less extreme patterns into the representational continuum for the behaviour. When provided with inputs of behaviours which fall in the middle of this continuum, the system already has a representation of the internal state this input would provoke, and can classify that input according to the continuum, rather than into a discrete class.

4 DISCUSSION AND CONCLUSIONS

In this paper we have briefly discussed how conceptors could provide an alternative to machine learning methods of automated behaviour classification for ASD diagnosis. By representing behaviours as continuous, the proposed approach has the potential to identify a more complete spectrum of ASD behaviours, rather than just extreme behaviours. Implementing such a system within a socially interactive robot would also leverage those benefits, providing a control system able to more accurately assess child behaviour to inform response selection, as the robot would be able to appropriately select and perform social prompts to elicit behaviours from the child in a standardised and repeatable manner. This application accommodates the goals of diagnostic models, e.g. ADOS [5]. Our next steps are to develop such a system, based on data from the DREAM project [1] [13], to train the system and test its performance.

ACKNOWLEDGMENTS

This work is part of the EU FP7 project DREAM project (www.dream2020.eu), funded by the European Commission (grant no. 611391)

REFERENCES

[1] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed. ed.). Autor, Washington, DC.
[2] Scassellati B. 2005a. Using social robots to study abnormal social development. *5th International Workshop on Epigenetic Robotics: Modelling Cognitive Development in Robotic Systems, Lund University Cognitive Studies* (2005a), 11–14.
[3] Scassellati B. 2005b. Quantitative metrics of social response for autism diagnosis. *In Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop.* 2005, September 2005 (2005b), 585–590.
[4] Scassellati B, Admoni H, and Matarić M. 2012. Robots for Use in Autism Research. *Annual Review of Biomedical Engineering* 14, 1 (2012), 275–294.
[5] Lord C., Risi S., Lambrecht L., Cook Jr E. H., Leventhal B. L., DiLavore P. C., Pickles A., and Rutter M. 2000. The Autism Diagnostic Observation Schedulefi?!Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders* 30, 3 (2000).
[6] Lai M. C., Lombardo M. V., Hakrabarti B., and Baron-Cohen S. 2013. Subgrouping the Autism Spectrum: Reflections on DSM-5. *PLoS Biology* 11, 4 (2013), e1001544.
[7] Bone D., Goodwin M. S., Black M. P., Lee C. C., Audhkhasi K., and Narayanan S. 2015. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of Autism and Developmental Disorders* 45, 5 (2015), 1121–1136.
[8] Petric F. 2014. Robotic Autism Spectrum Disorder Diagnostic Protocol: Basis for Cognitive and Interactive Robotic Systems. (2014).
[9] Jaeger H. 2014. Conceptors: an easy introduction. (2014), 1–11. http://arxiv.org/abs/1406.2671
[10] Rogers C. L., Goddard L., Hill E. L., Henry L. A., and Crane L. 2016. Experiences of diagnosing autism spectrum disorder: a survey of professionals in the United Kingdom. *Autism* 20, 7 (2016), 820–831.
[11] Geschwind D. H.and Levitt P. 2007. Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology* 17, 1 (2007), 103–111.
[12] Wall D. P., Kosmicki J., Deluca T. F., Harstad E., and Fusaro V. A. 2012a. Use of machine learning to shorten observation-based screening and diagnosis of autism. . *Translational psychiatry* 2, 4 (2012a), e100.
[13] 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.
[14] Thill S, Pop CA, Belpaeme T, Ziemke T, and Vanderborght B. 2012. Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn* 3, 4 (2012), 209–217.

[1] http://www.dream2020.eu/

Appendix F

IDC 2018 Workshop Paper - What Can You See? Identifying Cues on Internal States from the Kinematics of Natural Social Interactions

This paper was presented in a workshop at the *2018 ACM Interaction Design and Children Conference*.

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 J M, Friston K J, and Frith C D. 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 P G, Baxter P, Belpaeme T, Billing E, Cai H, Cao H L, 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 Behavioural 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 S E, and Sheikh Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. in *CVPR*.

Appendix G

HRI 2021 Submission - Estimating levels of engagement for social Human-Robot Interaction using Legendre Memory Units

This paper was presented as a poster at the *2021 ACM/IEEE International Conference on Human Robot Interaction* and published in the proceedings companion (*HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* 2021; Bartlett, Stewart, and Thill, 2021).

Estimating Levels of Engagement for Social Human-Robot Interaction using Legendre Memory Units

Madeleine E. Bartlett
madeleine.bartlett@plymouth.ac.uk
CRNS, University of Plymouth
Plymouth, UK

Terrence C. Stewart
National Research Council of Canada, University of Waterloo Collaboration Centre
Waterloo, ON, Canada
terrence.stewart@nrc-cnrc.gc.ca

Serge Thill
Donders Institute for Brain, Cognition, and Behaviour, Radboud University
Nijmegen, The Netherlands
s.thill@donders.ru.nl

ABSTRACT

In this study, we examine whether the data requirements associated with training a system to recognize multiple 'levels' of an internal state can be reduced by training systems on the 'extremes' in a way that allows them to estimate "intermediate" classes as falling in-between the trained extremes. Specifically, this study explores whether a novel recurrent neural network, the Legendre Delay Network, added as a pre-processing step to a Multi-Layer Perception, produces an output which can be used to separate an untrained intermediate class of task engagement from the trained extreme classes. The results showed that identifying untrained classes after training on the extremes is feasible, particularly when using the Legendre Delay Network.

KEYWORDS

Task Engagement, Intensity, Legendre Memory Units, MLP, Activity Recognition

ACM Reference Format:
Madeleine E. Bartlett, Terrence C. Stewart, and Serge Thill. 2021. Estimating Levels of Engagement for Social Human-Robot Interaction using Legendre Memory Units. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447193

1 INTRODUCTION

In the field of Human-Robot Interaction (HRI) there are, generally speaking, a wide range of application contexts which require, or can be improved by, the identification of human internal states. For example, a large pool of research has been dedicated to designing social robots to interact with children in educational settings [2, 6, 10, 11]. In these scenarios it is especially important that the robot's behaviour is not confusing or distressing for the child, and that it facilitates rather than disrupts learning [3]. This largely relies on the robot's ability to accurately estimate some internal state that the child is experiencing, e.g. task engagement or confusion.

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3447193

In order to provide flexible and appropriate responses to detected internal states it is potentially useful to have a robot recognize multiple 'levels' or 'intensities' of an internal state. For example, if a robot is able to recognize whether their partner is 'a little confused' or 'extremely confused', it can be made to provide a quick tip in the former case, and to reiterate the full instructions in the latter. Whilst it is technically possible to use classical machine learning methods to estimate different intensities of an internal state, most approaches require training on samples from every class. Thus one would need a training data set which includes examples of each intensity 'level'. However, creating such a data set is both difficult and time-consuming. Alternatively, it may be possible to leverage the underlying continuum of intensity to train a system on only the extremes of a state and have it produce an output to intermediate states which can be used to identify them as being different from, but related to, the trained states.

Thus the first goal of this study was to explore whether a system, after training on examples of high and low task engagement, could produce an output to intermediate engagement which would identify it as being 'in-between' the trained states. A second goal of this study was concerned with constructing an architecture able to deal with dynamic temporal data. There is a wide variety of existing approaches which meet this requirement, including Long-Short Term Memory's (LSTMs) and Echo-State Networks. However, a recently developed approach has emerged which shows promise but is yet to be tested on HRI-specific tasks; the Legendre Delay Network (LDN) [4, 12]. The key advance here is that a traditional neural network is combined with a separate linear recurrent layer (the LDN) [12]. The distinguishing feature of the LDN is that its structure has been derived from first mathematical principles to produce optimal reservoir-like behaviour, overcoming the reservoir-design issues of previous networks. The LDN transforms a D-dimensional input at a given time t into a $q \times D$ dimensional vector such that the output at time t contains information about the input signal between times $t - \theta$ and t. This transformation is linear, so it is simple to compute, and its mathematical derivation shows that it is optimal [12]. Consequently, unlike other machine learning algorithms, there are no parameters (other than θ and q) to fine-tune. Existing evidence demonstrates that the LDN is efficient to implement whilst still achieving state-of-the-art performance on standard benchmark tasks [4, 12]. As such, the method shows significant promise for dealing with temporally dependent tasks, whilst being well suited to the constraints of real-world applications [5]. This study, therefore, explores whether the LDN method offers any benefit for the types of internal state recognition tasks often seen in HRI.

Specifically, we investigate whether a Multi-Layer Perceptron (MLP) will, after training on high and low task engagement, provide an output to intermediate task engagement which can be used to estimate this class as being 'in-between' the two trained classes. This serves as a baseline to address the core question of whether pre-processing using the LDN will improve the system's performance. Based on previous findings that methods incorporating the LDN outperform other methods [12, 13] the following hypotheses are tested in this study:

(1) The use of the LDN as a pre-processing step will improve performance of the MLP.
(2) The systems, when trained on examples of high and low task engagement alone will produce an output to intermediate task engagement which can be used to estimate this state as being related to, but different from, these extremes, *without being trained on them.*

2 METHOD

2.1 Design

This study took a 2x1 design such that it examined the effect pre-processing step (without vs. with LDN) on the performance (accuracy) of an MLP. This resulted in 2 conditions or approaches being tested: (1) MLP, (2) LDN-MLP.

2.2 Materials & Apparatus

2.2.1 Data. The data set for this study was taken from the anonymous version of the PInSoRo [8, 9] data set. This data set consists of the xy coordinates for facial and skeletal landmarks from each frame of videos depicting child-child pairs engaged in a free-play session. This data set also includes annotation labels describing the children's behaviours. Only the data concerning the child positioned on the left-hand side of the frame was used. We used the annotation labels of 'goal-oriented play', 'aimless play' and 'no play' to select 'clips' which have been demonstrated to be analogous to 'high', 'intermediate' and 'low' task engagement respectively [1]. The final data set consisted of 105 high engagement, 52 intermediate engagement and 91 low engagement 'clips' wherein each frame was a 184-dimension vector of the xy coordinates for facial, skeletal and hand landmarks.

To account for the fact that the low engagement set had the smallest number of clips, training testing and validation data sets were created by taking the equivalent of 70/20/10% (respectively) of the low engagement set from both the high and low engagement sets. This resulted in 126 clips (M $\approx$ 175,000 frames) from the high and low engagement sets being used for training, and 36 clips (M $\approx$ 50,000 frames) for testing.

All 52 of the intermediate engagement clips (55,296 frames) were reserved for testing. Additionally, a random data set was constructed; i.e. arrays of random values of the same shape as the high engagement data set (234507 $\times$ 184). A random selection of 18 'clips' from this random data set were used for testing in each experiment.

2.2.2 Legendre Delay Network. The LDN [12] was implemented as a pre-processing step. The full architecture presented in [12] consisted of a linear dynamical memory (the LDN) and a non-linear decoder. In the current study, the non-linear decoder was replaced with an MLP. As such, this step involved feeding the raw data into the LDN and then using the output as input for the MLP.

Setting Parameters: The parameters to be set for the LDN were q and θ (theta). θ values 1, 3, 5 and 7, and q values 2, 3 and 4 were tested. Each combination of these parameters was tested 20 times by constructing the LDN pre-processed data and then using it as input to the MLP. The average accuracy of the MLP in each 'experiment' was plotted to see which combination of values tended to result in the best accuracy across all 20 experiments. This revealed $\theta = 3$ and $q = 4$ as the 'best' values.

2.2.3 Multi-Layer Perceptron. This study used `sklearn`'s MLP, implemented with the following parameters: the activation function was the rectified linear unit function, and the weights were optimized using the stochastic gradient-based optimizer 'adam' [7]. Additionally, only one hidden layer was implemented.

Setting Parameters: To determine the number of hidden neurons $n = 200$ in the MLP, a grid-search was conducted using the raw data as input. This grid-search compared 50, 100, 150 and 200 neurons, with each value being tested 20 times. Plotting performance accuracy for each experiment showed that 200 neurons tended to produce the best accuracy scores.

2.3 Procedure

Both of the approaches (MLP and LDN-MLP) were run 20 times (i.e. 20 'experiments'). The first step was to pre-process the data using the LDN. Each clip was processed separately such that the LDN was presented with a sequence of the 184 dimensional vectors which made up each frame. The outputs consisted of 736 dimensional vectors ($184 \times q$) which contained information about the current frame, as well as encoded information about the preceding 3 (θ) seconds. This step was done once for each of the four data sets.

For both approaches, each experiment then consisted of the following steps. The high and low engagement data sets were shuffled and split into training and testing sets. The random and intermediate engagement data sets were also shuffled, and a random selection of 18 'clips' from the random set were extracted. The MLP was trained on the high and low engagement data, and tested on the testing high and low engagement data sets as well as all of the intermediate engagement and 18 'clips' from the random data.

All of the experiment and analyses scripts were run using Jupyter Notebook on a Lenovo Thinkpad L380 laptop running Windows 10. Each experiment using the MLP alone took roughly 20 minutes and MLP with LDN pre-processing took roughly 10 minutes.

3 RESULTS

Analyses were conducted using the Python `numpy`, `pandas`, `SciPy` and `sklearn` toolkits in Jupyter Notebook. The analysis scripts can be found in the accompanying github repository (see Section 6).

Note that the system output used for analysis was the activation of the final hidden layer (decision function) in the MLP. This allows for a continuous output rather than binary classification values, which is necessary to accommodate additional classes that the MLP was not actually trained on. Effectively, the system outputs are in the form of a scalar between 0 and 1 indicating an estimated likelihood that the input belonged to the high engagement class. A

'correct' estimation for high and low engagement are values >0.95 and <0.05 respectively.

3.1 Performance on Trained Classes

This section compares performance of each approach when tested on high and low engagement patterns.

Performance on whole clips was examined by calculating the average estimation value across all the frames in a clip. The percent of correct estimations are summarized in Table 1. It can be seen that performance was improved by pre-processing using the LDN.

Table 1: Table of mean and standard deviation of percentages of frames and clips that were classified correctly by each approach in each experiment.

	Clips		
	High	Intermediate	Low
MLP	42.22% (16.70)	43.85% (6.25)	34.17% (15.04
LDN-MLP	78.06% (10.46)	39.33% (3.87)	75.00% (10.17)

To verify whether these differences in performance are significant, a one-way ANOVA was performed, with pre-processing (none vs LDN) as the independent variable and the percentage of correctly estimated high and low engagement clips in each experiment as the dependent variable. Both the assumption of normality (Shapiro-Wilk test: $W = 0.972$, $p = 0.426$ and of equal variances (Bartlett's test for sphericity: $\chi^2 = 0.568$, $p = 0.451$) were met. A significant main effect of approach showed that performance without LDN pre-processing (Mean = 38.19%, SD = 7.48%) was significantly worse than when the LDN was used (Mean = 76.53%, SD = 6.27%) (one-way ANOVA: $F(1, 38) = 308.339$, $p < 0.001$, $\eta_p^2 = 0.890$). These results demonstrate that the LDN was an effective pre-processing step for facilitating improved performance on distinguishing between high and low engagement.

3.2 Estimating Untrained Classes

Each approach was also tested on a third, unseen intermediate class of engagement. 'Correct' estimation values for intermediate engagement were values between 0.05 and 0.95.

The average percent of correct estimation values reveals that the system which performed best on the trained classes (LDN-MLP) also shows the worst performance on the intermediate engagement class (see Table 1). However, the MLP alone demonstrated worse performance on the trained classes, which could indicate that it had a higher tendency to produce intermediate estimation values for all classes. Thus, the better accuracy on the intermediate class is likely an artefact of poor performance overall.

3.3 Separating the Classes

The next step is to establish whether the system outputs contain enough information to distinguish between each class and accurately categorize them. To test this, the descriptive statistics of mean, standard deviation, skew and kurtosis were calculated for each test clip and used as input to a very simple classifier that can only pick up on clearly separable classes; a k-Nearest Neighbours (kNN) classifier. If successful, this would indicate that the information needed to distinguish between the classes is available. This section therefore presents two approaches to classification: (1) MLP-kNN, (2) LDN-MLP-kNN.

3.3.1 Random vs. Non-Random. First, we examine whether random clips can be distinguished from the engagement clips regardless of intensity level. This would demonstrate that random classes indeed occupy a different region in the 4 dimensional space defined by the descriptive statistics of the clips and are therefore not confused with examples of various levels of engagement.

This analysis took all 18 random clips from each experiment with a random selection of 18 high, low and intermediate clips from the same experiment. For each approach the kNN was run 20 times - once for each experiment - and the results were collated so that mean and standard deviation performance accuracy could be calculated. Average performance (percent correct) of each approach were as follows: Mean = 93.33% SD = 2.22 for MLP-kNN, and Mean = 76.11% SD = 9.15 for LDN-MLP-kNN. The confusion matrices showing mean percent of random and non-random clips classified correctly can be seen in Table 2. Overall, we observe good performance on this task but, interestingly, LDN pre-processing reduces this.

3.3.2 High vs. Intermediate vs. Low Engagement. The second analysis examines whether a kNN classifier can distinguish between the three engagement classes based on the four descriptive statistics. The same analysis of a kNN with k=5 was used to assess how well the descriptive statistics from each approach could be used to separate the engagement classes (high, intermediate and low). Again, a kNN was run 20 times for each approach. For each experiment, all of the high and low engagement clips were used (18 in each class), and a random sample of 18 intermediate clips were extracted. The average confusion matrices showing mean percent of high, intermediate and low engagement clips which were identified correctly can be seen in Table 3. Best performance is achieved with LDN pre-processing with 47.78% (SD = 13.77) of the untrained intermediate engagement clips being classified correctly, and 73.89% (SD = 8.44) and 64.44% (SD = 10.45) of the high and low engagement clips classified correctly respectively.

Overall, we demonstrate above-chance performance both on distinguishing clips of engagement (even if their class was not trained) from random data and on distinguishing between the three classes of engagement (one of which was untrained). While the latter performance in particular can likely be improved, this supports our second hypothesis. Furthermore, whilst we do acknowledge that kNN performance is only marginally above chance for the intermediate state, it is important to recognize that we cannot, and do not, expect perfect performance on this task. This is partially because we are using a naturalistic data set, and therefore not all samples (frames) will be 'perfect' examples of their class. Additionally, the intermediate class probably spans a rather large intensity space, with some samples being very similar to either of the trained classes.

To establish whether there was a significant difference in how each kNN performed, a one-way ANOVA was conducted with pre-processing (without vs. with LDN) as the independent variable, and average performance scores (mean accuracy) as the dependent

Table 2: Average confusion matrices showing mean (and standard deviation) percent of random vs. non-random clips classified correctly by kNN for each approach.

MLP

Actual \ Predicted	Random	Non-random
Random	99.17% (2.65)	0.83% (2.65)
Non-random	12.50% (4.93)	87.50% (4.93)

MLP-LDN

Actual \ Predicted	Random	Non-random
Random	80.00% (13.19)	20.00% (13.19)
Non-random	27.78% (7.45)	72.22% (7.45)

Table 3: Average confusion matrices showing mean (and standard deviation) percent of high vs. mid vs. low clips classified correctly by kNN for each approach.

MLP

Actual \ Predicted	High	Int	Low
High	72.78% (9.28)	13.06% (7.71)	14.17% (6.91)
Int	34.72% (11.50)	38.33% (16.56)	26.94% (11.15)
Low	35.83% (8.51)	12.22% (8.35)	51.94% (8.66)

MLP-LDN

Actual \ Predicted	High	Int	Low
High	73.89% (8.44)	10.83% (6.68)	15.28% (6.54)
Int	25.56% (11.71)	47.78% (13.77)	26.67% (10.33)
Low	21.39% (8.66)	14.17% (9.04)	64.44% (10.45)

variable. Assumptions of normally distributed residuals ($W = 0.957$, $p = 0.129$) and equal group variances ($\chi^2 = 0.014$, $p = 0.907$) were met. A one-way ANOVA revealed that there was a significant main effect of pre-processing step such that the kNN performed better when the data had been pre-processed with the LDN (Mean = 62.04% SD = 4.23) than when it had not (Mean = 54.35% SD = 4.34) (1-way ANOVA: $F(1, 38) = 32.152$, $p < 0.001$, $\eta_p^2 = 0.458$). This result provides support for our first hypothesis.

4 DISCUSSION

The present study was driven by two hypotheses. The first was that LDN pre-processing would improve accuracy on the non-trivial task of identifying engagement levels from human movement information in video clips from naturalistic data. This was tested by comparing the performance of an MLP classifier on LDN pre-processed data to performance on the raw data. We found that when the LDN was used, performance was significantly improved such that a higher percentage of clips were estimated correctly based on the mean output score for each clip. This provides support for the argument that the LDN is effective in providing multiple classification labels when there is a paucity of data.

The second hypothesis was that the output of a system trained on the extremes of an internal state (high vs. low task engagement) could then generalize to estimate an intermediate level of task engagement as being related to, but different from, the trained states in a meaningful way. Both approaches were trained on examples of high and low engagement. We then checked whether simple kNN classifiers could be used to distinguish a) random data from engagement clips and b) the three classes of engagement. Again, we found support for this hypothesis – all mean accuracy scores >0.72 when separating the random data from non-random data and >0.47 for distinguishing between the three engagement classes – but interestingly, for the distinction of random data, LDN pre-processing deteriorated performance for both systems.

The likely explanation for this is that the smoother output for random data following LDN pre-processing impacts the descriptive statistics by, for example, leading to a much smaller standard deviation than when the LDN was not used. Overall, although we observed clear benefits to using the LDN, in particular with respect to our first hypothesis, this demonstrates the trade-off (a smoothing based on history which may in some cases lose relevant information) that its usage entails.

Arguably, the most notable finding is that, when the data was pre-processed using the LDN, the outputs of the MLP could successfully be used by a kNN to classify a previously unseen class as being intermediate after the MLP was trained only on the extrema. This has a variety of potential repercussions. First, it is a demonstration that human internal states can be classified based on observable human behaviours after training on a relatively small training set; an average of 72016 frames from a total of 126 clips (roughly 40 minutes of video data). Furthermore, we have shown that training a classifier to estimate multiple 'levels' (i.e. intensities) of an internal state based on observable human behaviours does not necessarily require training on all of those levels. We do note, however, that performance of each kNN classifier on intermediate engagement clips was only slightly above chance. So whilst these results are certainly encouraging, more work is needed to establish the reliability of this finding.

5 CONCLUSION

In conclusion, this study has demonstrated that, on trained classes, adding LDN pre-processing significantly improves performance to a level one might only expect when using deep networks. Furthermore, based on these findings it appears that identification of untrained classes after training on the extrema of a continuum is feasible. More work is needed to confirm these results, and to further develop this approach so that intermediate states of task engagement can be more exactly placed along the continuum of intensity.

6 ONLINE RESOURCES

Experiment and analysis scripts can be found in the github repository https://github.com/maddybartlett/LMU_MLP_TaskEng.

REFERENCES

[1] Madeleine Bartlett, Daniel Hernandez Garcia, Serge Thill, and Tony Belpaeme. 2019. Recognizing Human Internal States: A Conceptor-Based Approach. *arXiv preprint arXiv:1909.04747* (2019).

[2] Tony Belpaeme, James Kennedy, Paul Baxter, Paul Vogt, Emiel EJ Krahmer, Stefan Kopp, Kirsten Bergmann, Paul Leseman, Aylin C Küntay, Tilbe Göksun, et al. 2015. L2TOR-second language tutoring using social robots. In *Proceedings of the ICSR 2015 WONDER Workshop*.

[3] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[4] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith. 2019. Benchmarking keyword spotting efficiency on neuromorphic hardware. In *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*. 1–8.

[5] P. Blouw, G. Malik, B. Morcos, A. R. Voelker, and C. Eliasmith. 2020. Hardware Aware Training for Efficient Keyword Spotting on General Purpose and Specialized Hardware. *arXiv preprint arXiv:2009.04465* (2020).

[6] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. 2015. The cowriter project: Teaching a robot how to write. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 269–269.

[7] D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[8] Séverin Lemaignan, Charlotte Edmunds, and Tony Belpaeme. 2017. The PInSoRo dataset. https://doi.org/10.5281/zenodo.1043507

[9] Séverin Lemaignan, Charlotte ER Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one* 13, 10 (2018).

[10] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. 2017. Give me a break! Personalized timing strategies to promote learning in robot-child tutoring. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 146–155.

[11] Emmanuel Senft, Séverin Lemaignan, Paul E Baxter, Madeleine Bartlett, and Tony Belpaeme. 2019. Teaching robots social autonomy from in situ human guidance. *Science Robotics* 4, 35 (2019).

[12] Aaron Voelker, Ivana Kajić, and Chris Eliasmith. 2019. Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*. 15544–15553.

[13] R. Wang, E. Huang, U. Chandrasekaran, and R. Yu. 2020. Aortic Pressure Forecasting with Deep Sequence Learning. *arXiv preprint arXiv:2005.05502* (2020).

Appendix H

Transactions in HRI Submission 2020 - Have I got the power? Analysing and reporting statistical power in HRI

This paper has been submitted for publication in *ACM Transactions on Human-Robot Interaction* and is under revision at the time of submitting this Thesis.

Have I got the power? Analysing and reporting statistical power in HRI

MADELEINE E. BARTLETT, CRNS, University of Plymouth, United Kingdom
CHARLOTTE E. R. EDMUNDS, Queen Mary, University of London, United Kingdom
TONY BELPAEME, ID Lab – imec, University of Ghent, Belgium and CRNS, University of Plymouth, United Kingdom
SERGE THILL, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

The study of Human-Robot Interaction is enriched by knowledge and techniques from many disciplines, especially the social sciences. These diverse perspectives are a great strength, but can also result in HRI unwittingly inheriting problems. Specifically, in recent years the validity of many results in social science have been undermined by the interplay between reporting biases and methodological issues, with particular concern over the high rate of low-powered research studies. Discussions of these issues have led to guidelines for improvement, including recommendations for reporting and methodological practices such as the use of power calculations. Here we investigate whether research involving human participants in HRI might be susceptible to similar concerns. We examine reporting of power and effect size in papers published in the proceedings of the HRI conference in the years 2010-2012 and 2017-2019 (before and after major publications concerning issues around replication) but do not find any significant difference in reporting practices between these two time periods. Of concern is that it largely remains impossible to verify the power of a study. This leaves HRI research open to similar criticisms as the rest of social science. We conclude with simple recommendations for best practices in reporting and conducting HRI research.

CCS Concepts: • **General and reference** → **Surveys and overviews**; **Computing standards, RFCs and guidelines**; **Reference works**.

Additional Key Words and Phrases: Reporting practices, Reproducibility, Power, Methodology, Best practice

ACM Reference Format:
Madeleine E. Bartlett, Charlotte E. R. Edmunds, Tony Belpaeme, and Serge Thill. 2018. Have I got the power? Analysing and reporting statistical power in HRI. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 15 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Scientific knowledge accrues from fragments: we creep closer to true understanding by incrementally building on previous data and theory [Kuhn 1962]. The unique opportunity and strength of Human-Robot Interaction (HRI) research is that we combine information from many, disparate, non-overlapping fields. However, in this plethora of research

Authors' addresses: Madeleine E. Bartlett, madeleine.bartlett@plymouth.ac.uk, CRNS, University of Plymouth, Plymouth, United Kingdom, PL4 8AA; Charlotte E. R. Edmunds, Queen Mary, University of London, London, United Kingdom; Tony Belpaeme, ID Lab – imec, University of Ghent, Ghent, Belgium, B-9052, CRNS, University of Plymouth, Plymouth, United Kingdom, PL4 8AA; Serge Thill, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, 6525 HR, The Netherlands.

lies a hidden challenge. We must be informed by these fields, whilst maintaining scientific rigor and avoiding both methodological and theoretical pitfalls inherent in those fields.

Currently, there is an important methodological issue in many of the social sciences that inform HRI that needs careful consideration. In the so-called "replication crisis" [for an amusing summary, see Neuroskeptic 2012], many studies in Psychology [Klein et al. 2014; Open Science Collaboration and others 2015], Medicine [Begley and Ellis 2012; Ioannidis 2005, 2016], Economics [Camerer et al. 2016] and other fields [Schoenfeld and Ioannidis 2012], have failed to replicate. Scientific replication, in this sense, is the process of *exactly* repeating a prior study and producing the same results [Cacioppo et al. 2015]. Successful replications provide support for the existence of the previously reported effect. Failure to replicate, on the other hand, highlights potential limitations to the original study [e.g. Edmunds et al. 2018], limitations on the conditions under which the effect can be observed [e.g. Edmunds et al. 2019] or indicates that the original findings may have been the result of statistical quirks [Simmons et al. 2011].

That is not to say that social science has no value and that none of the research is reliable or valid. These are *potential* issues. After all, social science research is inherently more noisy than computer science. A computer program given the same input will output the same thing (almost) every time. However, people are likely to change their answers, at least slightly, just because they were bored, or thought about the question differently, or were daydreaming. Thus, even if a study was perfectly conducted, a social scientist would still look to replicate experiments across different samples of people. Indeed, that is what the alpha level tries to represent. The replication crisis, however, showed us that far more studies did not replicate than expected given our statistical assumptions.

Given the intermingled heritage of HRI and social science research, one might worry that HRI might also suffer from this higher than expected rate of failure to replicate [Baxter et al. 2016]. Indeed, some have begun to point out specific instances where key studies fail to replicate [Irfan et al. 2018] and a dedicated 'Reproducibility' track was added to the ACM/IEEE International Conference on Human-Robot Interaction (hereafter: HRI conference) in 2020.

We initially set out to examine whether there was evidence of a replication crisis in HRI, specifically in studies published in the proceedings of the HRI conference. We chose to focus on conferences rather than journal publications as being more representative of the latest research directions and venues where any change in approach would be noticeable first. Amongst the many conferences that contain elements of HRI, we chose to focus on HRI because it is one of a select number dedicated to human-centric research and has a higher impact (with, for example, an h-index of 28 as reported in [Lab [n.d.]] for the year 2018) than alternatives such as, for example, the IEEE International Conference on Robot & Human Interactive Communication (RO-MAN, with an h-index of 9 for the same year). As such, work published in HRI can be considered representative of work that influences future research, including through the methodological choices made.

Thus, this article will explore reporting practices surrounding power analyses and related information in papers published in the proceedings of the HRI conference during 2010-2012 and 2017-2019. First we discuss what pieces of information are important to report when considering replicability. This is followed by an exploration of reporting practices in the HRI conference. Finally, we provide a number of recommendations for improving reporting practices and an example of how to appropriately conduct a power analysis.

1.1 What to report?

Reporting practices are key to the replicability of science [National Academies of Sciences et al. 2019, Chapter 6]. It is generally expected that any research report should provide sufficient information for the reader to be able to repeat the study as originally run [Field 2016; Field and Hole 2002]. An insufficiently detailed method section will lead to

readers being unclear on what was done and therefore unable to fully understand the findings. Additionally, it means that researchers intending to replicate the study rely on the original researcher(s) being contactable, able and willing to provide details of the method and any materials used. Ensuring that other researchers are readily able to replicate our work allows a scientific field to regularly identify any studies which do not replicate, and thus facilitates action to prevent a replication crisis.

That being said, two pieces of information provide insight into the likelihood that a study could be replicated. First, is the **power** of a study. The power of a test is the probability that the test correctly rejects the null hypothesis when the specified alternative hypothesis is true [McCrum-Gardner 2010]. These calculations tell us, as readers, how probable it is that the observed effect could have been found, and can be an indicator for how likely it is the effect will be replicated.

Power analyses can be conducted post-hoc or a priori. Post-hoc power analyses are conducted after the study has been conducted. This approach has been used to estimate the power of a study after the fact, using the actual sample size and observed effect size. It is generally accepted, however, that these analyses are largely meaningless [Levine and Ensom 2001; O'Keefe 2007; Thomas 1997]. However, post-hoc power analyses can be conducted wherein we use the actual sample size and the population effect size (i.e. a predicted, rather than the observed, effect size) [O'Keefe 2007]. This approach allows us to establish how appropriate the study design was for examining hypotheses about the population effect [O'Keefe 2007]. We provide an example of this kind of post-hoc power analysis later in the paper. A-priori power analyses, on the other hand, are conducted during the design stage of a study, and it is recommended to use desired power and predicted effect size to determine the required number of participants [Dattalo 2008; Nayak 2010]. A-priori power calculations for determining sample size are particularly useful in ensuring the replicability and reliability of science. If the number of participants is too low for a given effect size, it is not possible to be certain that the observed experimental effect is not just noise [Nayak 2010]. In these cases, it is unlikely (though not impossible) that the effect found will be produced in a replication study. A-priori power analyses allow researchers to ensure that an appropriate sample size is obtained to prevent such cases. Furthermore, studies exploring potential causes of the replication crisis experienced in other fields revealed that many of the studies employing human participants which failed to replicate were underpowered, and that this could explain the low rate of replication success [Diener and Biswas-Diener [n.d.]; Open Science Collaboration and others 2015; Stanley et al. 2018; Świątkowski and Dompnier 2017]. Not reporting a-priori power not only restricts our ability to interpret how reliable research is, but also leaves a scientific field susceptible to a replication crisis by inhibiting its ability for self-reflection and assessment. Reporting power calculations is, therefore, extremely informative for both the researcher (i.e. in determining sample size) and the scientific community (i.e. how the study's findings can be interpreted and how likely the results are to replicate).

The second important piece of information is the **effect size**. This is partly because calculating a-priori power requires a predicted effect size which is usually obtained from previous, similar studies. In the context of preventing potential replication crises, effect sizes are particularly useful in meta-analyses looking to summarise an effect across multiple replications because they can be averaged to give a better idea of the 'true' size of an effect [Coe 2002]. So, similar to power, reporting effect sizes reduces a field's risk of unwittingly falling into a replication crisis, by allowing that field to keep track of the reliability and replicability of its findings. Additionally, effect size is also valuable in the interpretation of a study's findings. Effect size quantifies the difference between groups/conditions [Coe 2002]. This value is therefore as informative as, if not more-so than, statistical significance; whilst significance can tell us whether or not an effect was observed, it does not tell us the size of that effect. As an illustrative example, consider a simple experiment where participants are testing whether a pill will improve their performance on an upcoming test.

Key Terms and Concepts

Alternative hypothesis
A statement that a difference or effect is not due to chance, suggesting a relationship between variables.

Null hypothesis
A statement that there is no actual relationship between variables, and that any observed effect is due to chance.

Significance (p-value)
Assuming the null hypothesis is true, the p-value denotes the probability that we would get results as large as the one observed. A smaller p-value indicates that there is stronger evidence in favour of the alternative hypothesis. Significance is indicated by this value being lower than a predefined cut-off (most commonly 0.05 or 5%).

Type I and II errors
Type I and II errors are concerned with either rejecting or accepting the null hypothesis.

A Type I error occurs when we reject the null hypothesis when it's true. It's otherwise referred to as a false positive and is captured by the significance level (p-value) of a test.

A Type II error, on the other hand, is when we accept the null hypothesis when it's actually false (i.e. the alternative hypothesis is true). It's therefore referred to as a false negative.

Power
The power of a test is the probability of not making a Type II error. In other words, it measures the ability of a test to correctly reject the null hypothesis.

The most commonly accepted minimum level of power is 80%. If a test has 80% power it means that the test has an 80% chance of detecting a difference of a given effect size, if such a difference exists. Power is linked to the sample size of a study in that a larger sample size will increase power.

Effect size
Effect size quantifies the difference between groups. It is therefore thought of as indicating the effectiveness of a treatment or experimental condition.

Exact Replication
An exact replication is an attempt to exactly recreate a study by using the original methodology, and recreating the conditions. Exact replications aim to determine whether the original findings are true by testing whether the same results can be found again, under the same conditions.

Conceptual Replication
A conceptual replication tests the same hypotheses as the original study, but using different methods. The aim of a conceptual replication is to test the truth of the theory behind the original findings, and to determine the conditions under which these findings will occur.

The researcher hypothesises that orange pills will improve performance more than green pills. They find a significant difference: participants who took orange pills scored 1 point higher on a test than those that took green pills. However, whether this was an *important* difference depends on what the test was out of. If the test had 10,000 questions then this would be equivalent to an improvement of 0.0001%, which would probably not be worth further research. Whereas, if the test was out of 5, this would be a 20% improvement, and thus probably an interesting difference. The effect size is one way that researchers attempt to quantify this insight.

Thus, if effect sizes are not reported, it is then not possible to evaluate the relevance and importance of the reported findings [Coe 2002]. If these unknown effect sizes are small, and therefore the findings are less reliable, it may be that future research is based on unreliable findings. In these cases, it is more likely that the reported effect would not replicate. Reporting effect sizes is therefore not only important for allowing other researchers to conduct power analyses, but also in enabling the scientific community to fully understand the results of a study.

Consequently, the aim of the current work was to directly evaluate the extent to which research accepted into the HRI conference stumbles into the methodological pitfall of under-reporting key information which can impact replication. We are interested in whether publications around the replication crisis might have influenced reporting practices in the HRI conference. Whilst we acknowledge that increased awareness does not facilitate change, it is reasonable to expect that it may have encouraged action. Therefore, given the high publicity of the replication crisis in the years 2012-2017, we propose a single, simple hypothesis: that there would be a difference in the frequency at which replication-relevant information (power analyses and effect sizes) are reported between the years 2010-2012 and 2017-2019. Specifically we predict an increase in reporting frequencies, following publications on the replication crisis and reporting recommendations.

2 METHOD

2.1 Data Collection

We adhered to the PRISMA guidelines [Moher et al. 2009] for the collection and extraction of data. We considered papers accepted to the HRI conference in the years 2010-2012 and 2017-2019 inclusive, excluding the alt.HRI papers. Searching Web of Science and the ACM Digital Library resulted in 199 papers that met these criteria. The primary keywords employed were *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* and search results were refined by selecting the publications for the years 2010-2012, 2017-2018 (date last searched: 18/04/2019). An additional 50 papers from 2019 were retrieved directly from the conference proceedings, producing 249 papers total. These were then reviewed to see if they met any of the exclusion criteria. Studies were excluded if a) the paper did not include data from human participants, b) the paper primarily described the development of a new technology/system, c) the paper involved the observation of human behaviour which was not measured or discussed quantitatively, d) the paper assessed the usability of a single technology or system, e) the paper was a meta-analysis or review, f) the paper described the design of a methodology or measurement tool. The number of papers that met each criterion are presented in Table 1. Note that papers could meet multiple exclusion criteria so the numbers in Table 1 do not reflect the total number of papers excluded. A total of 78 papers were excluded, leaving 171 papers to be examined below.

2.2 Data Extraction

The information we extracted from each paper was informed by the literature examining replicability in other fields, principally Psychology and other social sciences [e.g., Bakker et al. 2012; Stanley et al. 2018; Świątkowski and Dompnier

2017], suggesting that a central contributor to low levels of replications in Psychology is low statistical power. Power depends on both sample and effect size [Asendorpf et al. 2013; Munafò et al. 2017]. Thus for each study, we recorded details of any power analyses and the reported effect size for each test in each study. In addition, we collected the sample size and descriptive information such as whether the study was a replication (either exact or conceptual), the robots included in the study, the level of robot autonomy, and the source of the human sample. Finally, we also noted whether each study was pre-registered, whether statistical assumptions were checked, whether code and data were open source and whether they used Bayesian statistics. These latter techniques have been recommended by some as helping reduce p-hacking [Dienes 2011; Munafò et al. 2017; Nosek et al. 2018].

These data were made anonymous and collated onto a spreadsheet which is publicly available (see Section 6 for relevant links).

3 RESULTS

The analysis was conducted using Python 3.7 in Jupyter Notebook [Kluyver et al. 2016]. The notebook and data are publicly available (see section 6 for relevant links). As noted above, here we review the features of 171 papers presented at the HRI conference in the years 2010-2012 and 2017-2019.

3.1 Descriptive statistics

In this section, we briefly describe some surface features of the dataset. This dataset consisted of 58 papers from 2010-2012 and 113 papers from 2017-2019.

The three most frequently used robots has changed over time: Robovie, Simon and Keepon robots were most frequently used in 2010-2012, whereas Nao, Furhat and Tega were most popular in 2017-2019. Further, there has been an increase in the percentage of studies using autonomous robots over time (29.31% in 2010-2012, 37.17% in 2017-2019), as opposed to Wizard-of-Oz (29.31% to 19.47%).

The sample of human participants remained similar. Although a large proportion of studies utilised undergraduate populations for their samples (2010-2012: 37.93%, 2017-2019: 38.94%), the majority included adults from the general population in their sample (2010-2012: 44.83%, 2017-2019: 43.36%).

Finally, only three studies out of 171 were explicitly reported as replications. Although this appears similar to the replication rate reported in psychology (HRI: 1.75%, Psychology: 1.06%; [Makel et al. 2012]), we note an important caveat: our definition was rather liberal in that we included any study that re-used a previously published method. In contrast, the rate for psychology was based exclusively on studies that specifically used the word “replication” in

Table 1. Number of papers which met each exclusion criteria.

Exclusion Criteria	N Papers
No Human Participants	14
Primarily Technology Design	37
Observational Study	11
User Study	27
Review	3
Method Design	2

Table 2. Number of tests with effect size reported overall and in each year-group

Year-Group	Not Reported	Reported	% Reported
Overall	1845	458	19.89%
2010-2012	670	131	16.35%
2017-2019	1175	327	21.77%

their paper *and* included a replication. In addition, of the three studies, all were conceptual replications. Thus, we are probably overestimating the number of true replications compared to the estimate by [Makel et al. 2012].

3.2 Power Analyses

We first look at how many studies reported a power analysis. Of the 171 papers included in the analysis, only 6 were recorded as having reported performing a power analysis (2010-2012: 1, 2017-2019: 5). Of these, 2 reported calculating an estimate for sample size where $power = 0.8$, 1 where $power = 0.9$ and 1 where $power = 0.95$. The fifth reported calculating a post-hoc sample size estimate where $power = 0.8$. The final paper which reported conducting a power analysis did not report any values and did not use it to inform their sample size. The main take-away from this is that fewer than 4% of studies reported conducting a power analysis to derive their sample size.

This has serious implications for the validity and interpretability of HRI research. Failure to report power may indicate that researchers are not using power to determine the number of participants they need in their study. In order to draw valid conclusions, the sample size needs to be sufficient for detecting an effect if one is present. If the sample is *too small* the chance of a type II error (false negative) increases [Nayak 2010]. On the other hand, sample sizes that are *too large* are associated with a substantial increase in power, which can lead to an exaggerated tendency to find statistically significant results [Faber and Fonseca 2014]. Thus, the use of power calculators, such as G*Power [Faul et al. 2007], is vital to determine a sample size that produces meaningful data.

If power is calculated but not reported, this is less problematic, but still concerning. Hiding the results of a power analysis may mislead other researchers as they cannot determine how much faith to put into the data, undermining our ability to build new knowledge. The simple remedy is to report power, something which is becoming a standard requirement for publication in leading social science journals [e.g., Association for Psychological Science [n.d.]].

Given this need for future studies to report a-priori power, it is important to consider whether the data needed for these calculations is available. Namely, future power analyses will require predicted effect sizes. Whilst it is possible to use accepted effect size cut-offs (e.g. Cohen's effect size index [Cohen 2013]), it is generally recommended to use effect sizes from existing similar studies as estimates [McCrum-Gardner 2010; Schäfer and Schwarz 2019]. This, naturally, requires that previous studies report their effect sizes. We therefore now look at whether studies reported effect sizes to discover whether it would be possible for future studies to conduct power analyses.

3.3 Effect Size

Here we look at all the tests conducted across all studies where effect size should be reported. We found that, across all years, there was a total of 2,303 tests where effect size should be reported. Of these, only 459 tests (19.93%) reported effect size (see Table 2).

We looked to see whether there was a significant difference in the proportion of tests where effect size was reported in each year-group. A chi-square test revealed that the proportion of tests where effect size was reported is significantly

Table 3. Number of studies where effect size was reported at least once overall and in each year-group

Year-Group	Not Reported	Reported	% Reported
Overall	122	49	28.65%
2010-2012	46	12	20.69%
2017-2019	76	37	32.74%

greater in 2017-2019 (21.84%) than in 2010-2012 (16.35%) ($\chi^2(1) = 9.283, p = 0.002, \phi = 0.063$). However, this could be a result of studies in 2017-2019 reporting more tests. We therefore also examined the number of studies where effect size was reported at least once.

The number of studies where effect size was reported at least once for each year-group is presented in Table 3. To assess whether the increase in the percentage of studies reporting effect size at least once from 2010-2012 (20.69%) to 2017-2019 (32.74%) was significant, we again performed a chi-square test. This revealed that the difference was not significant ($\chi^2(1) = 2.166, p = 0.141, \phi = 0.113$). This result suggests that the significant difference in the number of tests where effect size was reported is likely due to an increase in the number of tests being reported, rather than a general increase in reporting of effect sizes within the field.

In both year-groups the majority of studies did not include any report of effect sizes for any of their tests. The lack of reporting presents a concerning issue, in particular when we consider the need for more HRI papers to be replicated. One of the first steps in replicating a study, and indeed to carry out any study, is to calculate the number of participants required to achieve a pre-determined power level. These calculations require a target power level, the significance level, and an effect size for the effect being explored. Whilst power level (usually at least 80%) and significance level (usually $\alpha = 0.05$) can be determined by the researchers, effect size is usually retrieved from an existing, similar study [McCrum-Gardner 2010]. In the case of the majority of papers considered here, however, this information is not available to be retrieved, and therefore future studies will find it difficult to calculate power.

Additionally, given that effect size describes the size of the difference between two experimental conditions [Coe 2002], this result also means that the majority of results reported in our sample cannot be accurately or correctly interpreted. For example, many of the studies are reporting the effect of some manipulation on human behaviour, and whilst we may know how significant that effect is, without the effect size we cannot know how effective the manipulation was. It's all well and good if making a robot yellow rather than green improves human willingness to work with the robot, but it's only really worth taking this into consideration if the effect is medium to large.

It is possible to calculate effect size in a post-hoc fashion based on the group means and standard deviations for a test. We therefore looked to see whether this information was available for us to calculate effect sizes ourselves. We found that 49% of studies reported group means at least once, and 40% reported standard deviations, markedly more than reported effect sizes. However, this would still only provide us with a view of less than half of the studies if we chose to calculate post-hoc effect sizes. We therefore felt that it was more meaningful to focus this discussion on reporting practices, than to attempt to calculate effect size ourselves in a post-hoc manner.

4 DISCUSSION

This work was originally intended to be an exploration of whether a replication crisis might exist within the HRI literature. However, when we discovered that the rate of reporting information relevant to such a discussion appeared to be low, we chose instead to examine reporting practices in more depth. We hope this exploration will therefore

act as a preliminary step towards reviewing the replicability of HRI research, by highlighting necessary changes, and providing guidance on how these changes can be implemented.

In this paper, we reflected on trends in reporting practices in the field of HRI. We examined publications in the HRI conference before and after a period of key publications elucidating the features of the "replication crisis" [e.g. Asendorpf et al. 2013; Baxter et al. 2016; Bosco et al. 2016]. Specifically, we were looking for trends in the reporting of power and effect size in the hope that we would see significant improvements over time. Worryingly, only 4% of papers reported conducting a power analysis to determine the appropriate number of participants required. This means that 96% of studies have not shown whether or not their study is adequately powered. Additionally, whilst a greater number of studies reported effect size for at least one test (28.65%) the majority did not. Furthermore, there was no significant increase in the number of studies reporting effect sizes in the years 2010-2012 compared to 2017-2019. Thus our hypothesis that awareness of the 'replication crisis' might encourage action which may be evidenced in the reporting of power and effect size was not supported.

As a considerable chunk of literature indicates that poor reporting practices and low power go hand-in-hand with a replication crisis [Cumming 2014; Open Science Collaboration and others 2015], one might wonder whether this indicates that the HRI conference is currently in the midst of its own crisis. Unfortunately, the work presented here shows that the state of HRI statistical reporting is so poor we cannot even state definitively whether replication is something that we should or should not be concerned about. Without knowing power or effect size it is difficult to assess how likely it is that a study will replicate. Furthermore, at least within the HRI conference itself, there are not sufficient replication studies to draw any conclusion on the probability of a crisis. Arguably this is more concerning than knowing that HRI is experiencing a replication crisis; it is difficult, if not impossible, to grow and develop as a field, without being able to reflect on where improvement is needed.

4.1 Summary of Best Practices

The low rate of reporting power analyses and effect sizes highlights that HRI suffers from considerable short-comings in both experimental methodology and reporting practices. However, the recent investigations into research practices in social science have resulted in a plethora of recommendations to improve reporting and replicability [Asendorpf et al. 2013; Bosco et al. 2016]. For example, Baxter and colleagues [Baxter et al. 2016] suggested using Bayesian statistical models for analysis, sharing of datasets, results and analysis scripts, and an overall increase in the number of studies being replicated. On the other hand, Asendorpf and colleagues [Asendorpf et al. 2013] recommended that studies should ensure the adequacy of their statistical analyses. One aspect of analysis which is important in this regard is correcting for multiple comparisons. Finally, Nosek et al., [Nosek et al. 2018] reported that pre-registration of study methods can facilitate replication[1]. Given these (perhaps overwhelming) suggestions, we provide the following straightforward recommendations for future papers for those interested in improving their experiments and how they are reported.

4.1.1 Power and sample size. First, and perhaps most importantly, researchers need to use power analyses to establish an appropriate sample size for their studies. Numerous tools exist for calculating sample size, including G*Power [Faul et al. 2007], PASS [Kaysville UT: NCSS. 2018], SAS [SAS Institute 2004] and the Power and Sample Size website [LLC [n.d.]]. For more detailed information on the importance of power analyses, and how to conduct them, see

[1] Although we do not investigate these aspects in detail here, we did collect initial data. For example, we found that no study in our 2010-2012 sample reported pre-registration of materials or methods, the provision of open-source data, results or analysis scripts or use of Bayesian analysis, and only a handful of studies (1, 2 and 1 respectively) did so in our sample from 2017-2019. This preliminary analysis shows that there remains plenty of room for improvement in reporting these aspects as well.

Examples of calculating Power

Independent Samples T-test

Let us propose a study investigating the effects of a new robot behaviour, compared to an established one, on acceptance from human users. We plan to do this using a between-subjects design (2 separate groups) and to use a questionnaire to measure acceptance. We will therefore be using an independent t-test to test the difference between groups.
Previous, similar studies have demonstrated an effect size of 0.89 between groups. We also assume that we want to use an alpha level of 0.05, and for our study to have a power level of 80%. Using the G*Power software (Figure 1) we can calculate that 21 participants would be needed in each group in order to detect an effect if one exists.

2-way ANOVA

Now let's imagine that we want to look at the effect of user gender and robot behaviour (new vs. old) on acceptance. We again use a between-subjects design which gives us 4 groups. If we can't find an effect size from a previous study, we can use Cohen's suggested medium effect size value which is 0.25. As before, we want an alpha level of 0.05 and 80% power. G*Power gives us an estimate of 128 participants in total, so 32 participants per group (see Figure2).

Post-hoc Power - Independent Samples T-test

Finally, assume that we have now conducted a study investigating the effects of a new robot behaviour vs. an established behaviour on acceptance from human users (the study we proposed in the first example). We didn't conduct a power analysis before running the experiment, so we choose to run a post-hoc analysis in order to establish how appropriate our design was. We recruited 60 participants in total, 30 in each group, and found an effect size of 0.71. By averaging the effect sizes from related studies we obtain an estimate of the population effect size. Our resultant population effect size is 0.5. Using these values (sample size = 30, effect size = 0.5, alpha level = 0.05) we can calculate the power of our study for detecting this effect. G*Power shows us that our study was underpowered for detecting an effect of 0.5 with a power level of 48% (see Figure 3).

[McCrum-Gardner 2010]. Here we provide two examples of how one might use G*Power to calculate sample size for a HRI study.

4.1.2 Reporting practices. Second, there is a clear and pressing need for changes in reporting practices. To some extent, this can be addressed by individual researchers, authors and labs. This kind of knowledge transfer will likely be incredibly important given the diversity of measures, methods and statistical approaches used in HRI research. For some broad advice on this topic, we recommend [American Psychological Association [n.d.]; Field 2017; Field and Hole 2002; Schimel 2012]).

More specifically, we also provide an outline appropriate for most Null-Hypothesis Significance Testing analyses were p-values are reported. At the beginning of a results section it is useful to provide a brief description of any tests that were carried out. Before reporting test results one should describe how the test relates to the hypotheses, what assumption checks were carried out and any steps taken to deal with violations (e.g. use of non-parametric tests). In reporting actual tests, we suggest the following structure:

(1) What was being tested

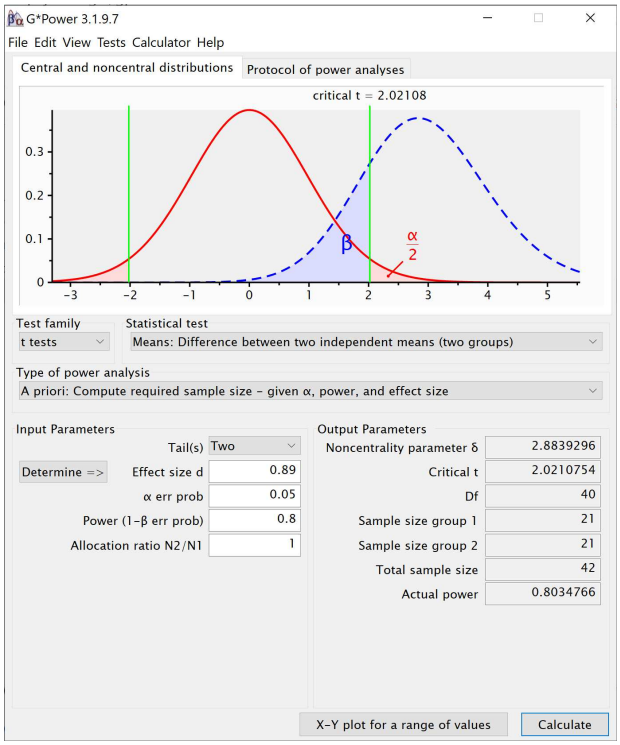


Fig. 1. Screenshot of G*Power - calculating required sample size for test comparing 2 independent groups.

(2) What test was used
(3) A sentence describing the finding (including means and standard deviations, or confidence intervals, for each experimental condition)
(4) The numerical test results (test statistic, degrees of freedom, p-value, effect size)
(5) A brief explanation of what this means in relation to the experiment and hypotheses

For example, say we are reporting on a test comparing the effect of practice in Virtual Reality (VR) on the ability to fly a drone through a maze, we would present it as follows: "To compare the effect of practice in VR on time taken to navigate the maze we conducted a one-way ANOVA with practice condition (none vs. VR) as the independent variables, and time taken (measured in minutes) as the outcome variable. We found that participants who practiced had significantly shorter completion times (Mean = 26.3, Standard deviation = 10.5) than participants who did not practice (Mean = 65.7, Standard deviation = 11.6) (One-Way ANOVA: $F(1, 58) = 3.97$, $p = 0.04$, $\eta_p^2 = .064$)[2]. This indicates that practice in VR had a positive impact on participants' performance on the test."

For a more detailed overview we suggest Professor Andy Field's 'Writing Up Research' document available in "How to design and report experiments" and as an online PDF [Field 2016, 2017].

4.1.3 Training. The issue of adequate reporting of statistical analyses is much larger than simply stating "please do X, Y and Z". Given the multidisciplinary nature of HRI as a field, it is important for us to recognize that researchers come from a variety of backgrounds. So whilst we encourage researchers to employ and report power analyses and effect size

[2]Numerical values used in this example are arbitrary and nonsensical.

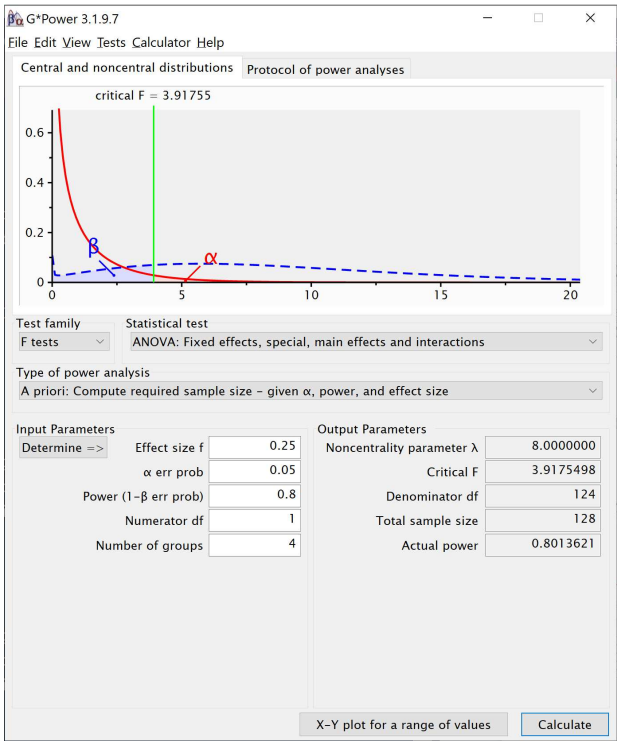


Fig. 2. Screenshot of G*Power - calculating required sample size for test comparing 4 independent groups.

calculations, we also recognize that training in these methods is also required. We therefore also suggest that training programs be designed for providing, at the very least, statistical and methodological skills from across the contributing fields. For example, the HRI conference might want to organise a session on this in an upcoming year. Additionally, a number of online, open-source resources for statistical training are available, including those provided on Coursera [Coursera [n.d.]] and Datacamp [Datacamp [n.d.]]. Another incredibly useful resource is the Discovering Statistics website [Field 2017] which provides video tutorials on how to run statistical tests, and instructional guidelines on how to report them.

4.1.4 Reporting standards. Along with requirements for reporting statistical tests, it is also useful for any scientific field to develop a standard style of reporting. This aids in systematic reviews or meta-analyses, as any reviewer will be able to easily locate information of interest from large samples of papers. However, it also assists the paper's authors in that, by following a standardized structure it is easier to identify whether any information might be missing from the report, and therefore restricting the ability of other researchers to conduct replications.

As well as taking steps individually to improve our statistical and reporting practices, journals and conferences also have a responsibility to ensure that authors report adequate statistical and methodological detail. Additional reporting requirements from Journals and Conferences could be introduced to mirror the fact that, for example, many APA journals (e.g. The Journal of Experimental Psychology) have made a-priori power analyses a requirement for paper submission. Information such as power, effect size, assumption checks etc. are easy to check. The HRI conference

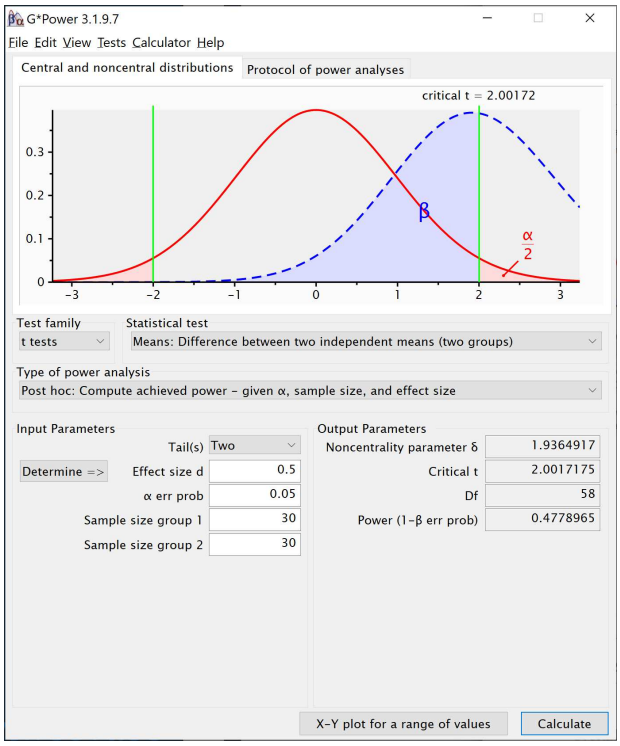


Fig. 3. Screenshot of G*Power - calculating post-hoc power.

has made a significant step in this direction by introducing the 'Reproducibility in Human-Robot Interaction' track. However, more can certainly be done to promote validation over novelty within scientific research.

5 CONCLUSION

We set out to examine the rates at which power and effect sizes were reported in papers presented at the HRI conference before and after awareness of the replication crisis. We did not find an overall improvement in reporting practices. This lack of available information prevents exact (or direct) replication to the extent that it is not even possible to say whether or not HRI has a replication crisis as observed in the social sciences. This should be deeply concerning for anyone who has a vested interested in HRI research. To help remedy this situation, we have provided simple recommendations that we hope the field will find helpful to follow and will spark further discussion on how improvements can be made.

6 OPEN-SOURCE RESOURCES

The data set and analysis script can be found at https://github.com/maddybartlett/HRI_Reporting_Practices.

REFERENCES

American Psychological Association. [n.d.]. *Publication Manual of the American Psychological Association.* https://apastyle.apa.org/manual/new-7th-edition

J. B. Asendorpf, M. Conner, F. De Fruyt, J. De Houwer, J. J. A. Denissen, K. Fiedler, S. Fiedler, D. C. Funder, R. Kliegl, B. A. Nosek, et al. 2013. Recommendations for increasing replicability in psychology. *European Journal of Personality* 27, 2 (2013), 108–119.

Association for Psychological Science. [n.d.]. *Submission Guidelines.* https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#CRIT

M. Bakker, A. van Dijk, and J. M. Wicherts. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science* 7, 6 (2012), 543–554.

P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 391–398.

G. C. Begley and L. M. Ellis. 2012. Raise standards for preclinical cancer research. *Nature* 483, 7391 (2012), 531–533. https://doi.org/10.1038/483531a

F. Bosco, J. M. Carp, J. G. Field, H. IJzerman, M. Lewis, M. Munafo, B. A. Nosek, J. M. Prenoveau, J. R. Spies, and R. Giner-Sorolla. 2016. Maximizing the Reproducibility of Your Research. (2016).

J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. L. Olds, and H. Dean. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. (2015).

C. F. Camerer, A. Dreber, E. Forsell, T. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016), 1433–1436. https://doi.org/10.1126/science.aaf0918 arXiv:https://science.sciencemag.org/content/351/6280/1433.full.pdf

R. Coe. 2002. It's the effect size, stupid. In *Paper presented at the British Educational Research Association annual conference*, Vol. 12. 14.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Coursera. [n.d.]. Online Courses Credentials From Top Educators. Join for Free. https://www.coursera.org/

G. Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

Datacamp. [n.d.]. Learn R, Python Data Science Online. https://www.datacamp.com/

P. Dattalo. 2008. Sample-size determination in quantitative social work research.

E. Diener and R. Biswas-Diener. [n.d.]. *The Replication Crisis in Psychology*. http://noba.to/q4cvydeh

Z. Dienes. 2011. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* 6, 3 (2011), 274–290. https://doi.org/10.1177/1745691611406920

C. E. R. Edmunds, F. Milton, and A. J. Wills. 2018. Due Process in Dual Process: Model-Recovery Simulations of Decision-Bound Strategy Analysis in Category Learning. *Cognitive Science* 42 (2018), 833–860. https://doi.org/10.1111/cogs.12607

C. E. R. Edmunds, A. J. Wills, and F. Milton. 2019. Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology* 72, 2 (2019), 151–167. https://doi.org/10.1080/17470218.2017.1370477

J. Faber and L. M. Fonseca. 2014. How sample size influences research outcomes. *Dental press journal of orthodontics* 19, 4 (2014), 27–29.

F. Faul, E. Erdfelder, A. Lang, and A. Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

A. Field. 2016. Writing Up Research [PDF]. http://www.discoveringstatistics.com/docs/writinglabreports.pdf

A. Field. 2017. Discovering Statistics. https://www.discoveringstatistics.com/

A. Field and G. Hole. 2002. *How to design and report experiments*. Sage.

J. P. A. Ioannidis. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Jama* 294, 2 (2005), 218–228.

J. P. A. Ioannidis. 2016. Why Most Clinical Research Is Not Useful. *PLOS Medicine* 13, 6 (06 2016), 1–10. https://doi.org/10.1371/journal.pmed.1002049

B. Irfan, J. Kennedy, S. Lemaignan, F. Papadopoulos, E. Senft, and T. Belpaeme. 2018. Social psychology and human-robot interaction: An uneasy marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 13–20.

Kaysville UT: NCSS. 2018. *Power Analysis and Sample Size Software*. https://www.ncss.com/software/pass/

R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams Jr, Š. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, et al. 2014. Investigating variation in replicability. *Social psychology* (2014).

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, et al. 2016. Jupyter Notebooks-a publishing format for reproducible computational workflows.. In *ELPUB*. 87–90.

T. S. Kuhn. 1962. The structure of scientific revolutions. *Chicago and London* (1962).

Scimago Lab. [n.d.]. https://www.scimagojr.com/journalrank.php?category=1709&area=1700&year=2018. Accessed 18, 2020.

Marc Levine and Mary HH Ensom. 2001. Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 21, 4 (2001), 405–409.

HyLown Consulting LLC. [n.d.]. Power and Sample size. http://powerandsamplesize.com/

M. C. Makel, J. A. Plucker, and B. Hegarty. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7, 6 (2012), 537–542.

E. McCrum-Gardner. 2010. Sample size and power calculations made simple. *International Journal of Therapy and Rehabilitation* 17, 1 (2010), 10–14.

D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* 151, 4 (2009), 264–269.

M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. Du Sert, U. Simonsohn, E. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 1–9.

Engineering National Academies of Sciences, Medicine, et al. 2019. *Reproducibility and replicability in science*. National Academies Press.

B. K. Nayak. 2010. Understanding the relevance of sample size calculation. *Indian journal of ophthalmology* 58, 6 (2010), 469.

Neuroskeptic. 2012. The nine circles of scientific hell. *Perspectives on Psychological Science* 7, 6 (2012), 643–644.

B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America* 115, 11 (2018), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Daniel J O'Keefe. 2007. Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Communication methods and measures* 1, 4 (2007), 291–299.

Open Science Collaboration and others. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

SAS Institute. 2004. *Getting Started with the SAS Power and Sample Size Application.*

Thomas Schäfer and Marcus Schwarz. 2019. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology* 10 (2019), 813.

J. Schimel. 2012. *Writing science: how to write papers that get cited and proposals that get funded.* OUP USA.

J. D. Schoenfeld and J. P. A. Ioannidis. 2012. Is everything we eat associated with cancer? A systematic cookbook review. *The American Journal of Clinical Nutrition* 97, 1 (11 2012), 127–134. https://doi.org/10.3945/ajcn.112.047142 arXiv:http://oup.prod.sis.lan/ajcn/article-pdf/97/1/127/23818321/127.pdf

J. P. Simmons, L. D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

T. D. Stanley, E. C. Carter, and H. Doucouliagos. 2018. What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin* (2018).

W. Świątkowski and B. Dompnier. 2017. Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology* 30, 1 (2017), 111–124.

Len Thomas. 1997. Retrospective power analysis. *Conservation Biology* 11, 1 (1997), 276–280.