Detecting Perceived Appropriateness of a Robot's Social Positioning Behavior from Non-Verbal Cues

'A robot study in scarlet'

Vroon, Jered; Englebienne, Gwenn; Evers, Vanessa

Publication date
2019

Document Version
Accepted author manuscript

Published in
The First IEEE International Conference on Cognitive Machine Intelligence

Detecting Perceived Appropriateness of a Robot's Social Positioning Behavior from Non-Verbal Cues.

'a robot study in scarlet'

Jered Vroon*
Knowledge and Intelligence Design
Delft University of Technology
Delft, the Netherlands
j.h.vroon@tudelft.nl

Gwenn Englebienne
Human Media Interaction
University of Twente
Enschede, the Netherlands
g.englebienne@utwente.nl

Vanessa Evers*
NTU Institute of
Science and Technology for Humanity
Nanyang Technical University
Singapore
vanessa.evers@ntu.edu.sg

***Abstract*—What if a robot could detect when you think it got too close to you during its approach? This would allow it to correct or compensate for its social 'mistake'. It would also allow for a responsive approach, where that robot would reactively find suitable approach behavior *through* and during the interaction. We investigated if it is possible to automatically detect such social feedback cues in the context of a robot approaching a person.**

We collected a dataset in which our robot would repeatedly approach people (n=30) to verbally deliver a message. Approach distance and environmental noise were manipulated, and our participants were tracked (position and orientation of upper body and head). We evaluated their perception of the robot's behavior through questionnaires and found no single or joint effects of the manipulations. This showed that, in this case, personal differences are more important than contextual cues – thus highlighting the importance of responding to behavioral feedback. This dataset is being made publicly available as part of this publication†.

On this dataset, we then trained a random forest classifier to infer people's perception of the robot's approach behavior from features generated from the response behaviors. This resulted in a set of relevant features that perform significantly better than chance for a participant-dependent classifier; which implies that the behaviors of our participants, even with our relatively limited tracking, contain interpretable information about their perception of the robot's behavior.

Our findings demonstrate, for this specific context, that the observable behavior of people does indeed contain usable information about their subjective perception of a robot's behavior. As such they, together with the dataset, provide a stepping stone for future research into the automatic detection of such social feedback cues, e.g. with other or more fine-grained observations of people's behavior (such as facial expressions), with more sophisticated machine learning techniques, and/or in different contexts.

***Index Terms*—Social robotics, Social positioning, Responsiveness, Social feedback cues, Social interaction dynamics.**

I. Introduction

In real-world social situations, it is impossibly difficult to fully predict which behavior will be appropriate. People make social 'mistakes' regularly, notwithstanding our extensive experience – from making someone slightly uncomfortable by getting closer to them than they prefer, to any kind of faux pas or bigger social blundering. So too, if not more so, do our current social machines – from notifications interrupting conversations, to mobile robots miscalibrating their social positioning behavior.

And yet, among others in the context of social positioning behavior for mobile robots, most current approaches still focus heavily on systems that try to predict the 'right' action through model-based reasoning and/or learning (e.g. [1]–[3]). While such work will be helpful in ensuring a reasonable starting point for social positioning behavior, it does not yet reliably account for those minor and major social mistakes that are prevalent in our social interactions. As we have argued [4], such purely model-based approaches may well be inherently unable to do so, given the size of the state space and the many unobservable variables that play a role in social interactions.

We propose that, instead or in addition, robots should detect when people think their behavior is not appropriate – infer when people feel it is getting uncomfortably close, too far away for a conversation, and so on. This would allow such robots to immediately and continuously use those detections to tune their behavior. We have previously referred to this dynamic 'dance' of adapting to feedback as **responsiveness** [4]. Such responsiveness is found in many forms in human-human interaction, e.g. during various dyadic interactions [5] and even during speed dates [6]. We have similarly found people to be responsive to robots. For example, when we left people relatively free to move during an interaction, they would actively and pro-actively adapt their positioning behavior to a robot, e.g. dynamically adjusting their distance to a robot, and urging a robot to pass through a group to ease its navigation [7]. Recent work suggests that people may even adapt their own distancing preferences to the effective sensor range of a robot [8].

In our theory of responsiveness, we have hypothesized that this dynamic of active adaptation is enabled by (subconscious) non-verbal cues that signal how people perceive the appropriateness of social behaviors, and that may even indicate the desired direction of improvement. We will here

This work was supported by the European FP7 project TERESA (Telepresent Reinforcement-learning Social Agent), grant number FP7-ICT-611153.

* The majority of this work has been conducted while Jered Vroon and Vanessa Evers were affiliated with the Human Media Interaction group at the University of Twente.

† http://doi.org/10.4121/uuid:b76c3a6f-f7d5-418e-874a-d6140853e1fa

refer to these signals as **social feedback cues**. For example, if someone thinks a robot (or other person) seems to be getting uncomfortably close, their non-verbal behaviors might indicate that they would like more space. Literature on human-human interaction has discussed various non-verbal behaviors that could potentially serve as social feedback cues, such as averting gaze and leaning behavior [5], [9], [10]. And previous work on social agents has used easy-to-detect cues, e.g. the use of estimated subjective task difficulty to try and adapt the difficulty of a learning task [11], and the use of specific non-verbal utterances to guide the adaptive behaviors of a conversational agent [12].

But can our systems automatically detect and correctly interpret subtle social feedback cues? Can a robot (learn to) detect from a person's non-verbal cues if that person thinks it got too close, did not get close enough, or is in a comfortable position? To our knowledge, there is no previous work in this direction, nor any datasets collected for the purpose of investigating these questions.

In this paper we investigate these questions, in the context of social positioning behavior for a mobile robot. More specifically, we investigate whether we can train a detector and find features from pose tracking that indicate that a robot's approach distance is subjectively perceived as too close, too far, or sufficiently comfortable. As our robot platform, we used the Giraff hardware (with modifications made in the context of the Teresa project [13]). To keep our scope focused, we have deliberately (1) limited ourselves to this specific context and robot, and (2) used a straightforward machine learning pipeline with little parameter tweaking. While this limits the generalizability of our current findings, and likely the performance of our detector as well, it *does* allow our findings to provide a first stepping stone and various leads for the automatic detection of social feedback cues.

To be able to train our detector, we first collected a dataset in which a robot would approach people, using a range of approach distances, to provide information through speech, and we collected both their response behaviors (through a tracking system) and their perception of the robot's behavior (through a questionnaire) (Section II). We then conducted several tests to confirm that the perception of our participants did not depend exclusively on the distancing behavior of the robot (Section III), as that would allow a detector to achieve a reasonable performance by simply using that distancing behavior (Figure 1). After that, we used our dataset to train a detector of social feedback cues – achieving a performance significantly better than chance and identifying various relevant features (Section IV). Together, these findings indicate that, at least in this context, there is scope for systems that detect and respond to social feedback cues (Section V).

II. A DATASET FOR DETECTING SOCIAL FEEDBACK CUES

Our first step, which we will describe in detail in this section, was to collect a dataset suitable for our purposes. To this end, we designed our data collection with three key requirements in mind. Firstly, to ensure a rich and sufficiently diverse dataset, we aimed to elicit a variety of internal states and leave participants relatively free to display external non-verbal behaviors as they please. Secondly, to ensure that the different data points are comparable, the interaction followed a somewhat controlled pattern. Thirdly, we wanted to make sure that detection of our participants' perception of the robot's behavior could not be derived from observable contextual cues. This to ensure that such detection would, in line with the aim of this paper, instead have to rely on the tracked behavioral responses (as illustrated in Figure 1).

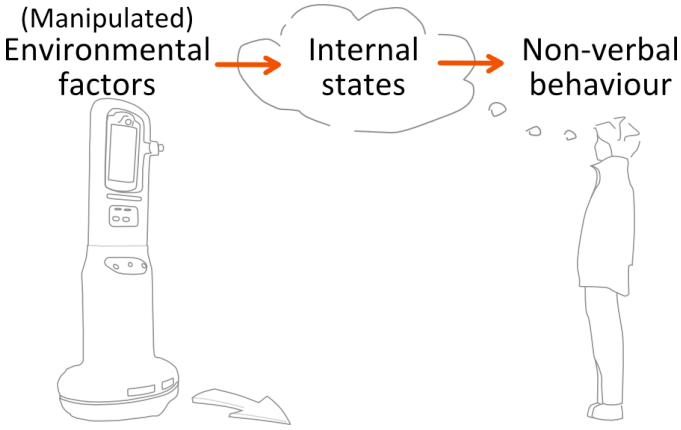


Fig. 1. Put abstractly, as in this model, (manipulated) environmental factors can influence internal states, which can in turn be reflected in (non-verbal) behavior. These two relations (represented by arrows) could both be used to detect internal states, provided that enough data is available. Since our focus is explicitly on the detection of internal states from non-verbal behavior (the right arrow), we should make the detection of internal states from environmental factors impossible, e.g. by introducing relevant environmental factors that are not represented in the dataset.

A. Task and context

To allow for the collection of multiple data points, we needed a task and context that would allow for the robot to approach the participants several times in a meaningful way. To this end, we reused a murder mystery task [7]. Within this task, the robot would, in 8 iterations: Approach the participant to give them a clue relevant to the murder mystery, briefly discuss the clue with the participant, indicate that it would go collect the next cue, and then retreat from the conversation.

The clues were designed to be comparable in length (20-30 seconds long) and information content; each clue started with some filler text on how it was collected, e.g. *"I have the fourth clue. The detective chief inspector had a hunch and also had someone ask around at several hardware stores"*, followed by information relevant to solving the murder mystery, e.g. *"Yesterday, around 6 p.m., the victim visited a local hardware store to purchase a crowbar."* To make the clues similar in length, some clues would end with more filler text, e.g. *"The shop assistant positively identified him."*

After sharing the clue, the robot would maintain a brief conversation about the clue for about 1 minute. To do so, we implemented a simple Wizard of Oz set-up in which an experimenter could select and play various pre-recorded audio

files. Beyond the clues, these fit two categories. Firstly, there were simple answers to questions the participants might ask, e.g. *"Yes"*, *"No"*, *"I did not catch that"*, *"I do not know"*. The experimenter was instructed to avoid giving opinions and to only give information that was also available in the clues shared thus far with the participants. Secondly, we included questions to engage participants in the murder mystery, e.g. *"What do you think happened?"*, *"Why?"*, *"Do you already have a suspect in mind?"*, *"Can you elaborate?"*.

After the brief conversation, the robot would wrap up the conversation by saying *"I will now go and collect the next clue,"* after which it would do so. Each participant would in this way be presented with a total of 8 clues, which together provided enough information to solve the murder mystery. After that, the robot would approach them a 9^{th} time, and ask them who they suspected. This 9^{th} approach was mainly included to allow participants to wrap up their interaction with the robot, and excluded from our analysis.

B. Conditions

We introduced two factors in our data collection. To ensure that the robot's behavior would elicit a rich range of reactions, we manipulated how close it would get during its approach (within-subject). In an attempt to introduce non-observable environmental factors and thus keep our detectors from using environmental factors as a short-cut to detect the internal states of our participants, we also manipulated environment noise (between-subject).

1) Interaction distance: Within-subject we manipulated approach distance of the robot, using the distances 30cm, 70cm, 110cm, and 150cm (measured from head-to-head, in the floorplane). These distances were chosen to be evenly distributed, while falling into four distinct informal social interaction distance classifications of Hall [14, p.126]; not close intimate, close personal, not close personal, and close social, respectively. These distances also align with literature in HRI, where, for human-sized mobile robots, distances around 30cm are often found as well invading personal space (e.g. [2], [15]–[17]). Furthermore, as the robot could only be controlled with limited precision, these distances were chosen such that even with those minor deviations they would still be distinguishable.

We used each approach distance twice, resulting in a total of 8 data points per participant. To counteract order and sequential effects, we used an 8×8 balanced latin square design to counterbalance.

For practical and safety reasons, the approach behavior of the robot was controlled by the experimenter using a Wizard-of-Oz approach.

2) Environment noise: For our other condition, we aimed to find a between-subject factor that would effectively influence participants' perception of different approach behaviors, without being directly observable from the tracking data recorded. Through pre-studies (stop-task) we tried a range of potential factors (smell, team membership, perception of the robot as safe/unsafe), but found no strong indication that these had an effect on the preferred approach distance.

In the end, we found environment noise to be a suitable factor. Previous work has suggested that perceptual challenges can be related to proxemic preferences in interactions with robots [8], [18], [19]. This also aligns with Hall's work, where he explicitly tied his informal social interaction distance classification to different perceptual qualities [14]. We conducted a plain stop task with high/low environment noise as a small pilot (n=12), which we explicitly framed in the context of having a conversation with the robot. While we found relatively big individual differences in stop distances (ranging from 5cm-95cm), the data suggested a clear effect of high/low environment noise on stop distance.

To implement our manipulation of environment noise, we hid 4 speakers above the drop ceiling of the experiment room and played white noise from them. In the low condition, we set it to a low volume such that it was audible but not invasive – sounding akin to the noises made by some air conditioning systems. In the high condition, we set it to the highest volume, such that to the experimenter it was challenging to follow the robot's speech if it was about 100cm away.

We started the white noise before the participants would enter the experiment room, apologizing for it if participants asked about it without suggesting it pertained to the experiment. The majority of participants in the high noise condition did ask about it, while none in the low noise condition did so. Participants were debriefed about this afterward.

C. Data collection

Throughout the experiment, we tracked the position of our participants and made video recordings of the interaction. In between each interaction with the robot, the experimenter would present participants with a between-session questionnaire, and after participants had gone through all interactions we presented them with a post-experiment questionnaire.

As discussed above, we needed to strike a balance between allowing our participants to move and react freely, while also keeping the collected data comparable. To this end we used three cover stories. Firstly, we told participants that all equipment used in the data collection was intended for autonomous robot behavior. This reinforced the idea that the robot was autonomous, while also serving to make the participants less aware of their actions and reactions being recorded. Secondly, we wanted to ensure that participants would be forced to let the robot approach them, and not the other way around. To achieve this, we used a wired skin conductance measuring device – the wire, connected to the participants' left hand, effectively limited their movement range to the wire's length (approximately 1 meter) around the device. Thirdly, when handing the participants the between-session questionnaires, the experimenter would always do so from the same position. This served as a means to roughly (and softly) 'reset' the position of the participants in between each approach. All participants were, of course, debriefed after the experiment about the deception involved in these cover stories.

1) Objective measures: To track the position of our participants, we equipped them with two uniquely identifiable markers. One marker was worn on the back of the chest, with two straps going around the shoulders. The other marker was worn on a cap. The robot was similarly equipped with markers. All markers were tracked by an OptiTrack (www.naturalpoint.com/optitrack) motion capture system using 12 infra-red cameras. This set-up allowed for sub-centimetre level precision tracking of both position and orientation of each marker.

In addition, we also recorded the whole interaction with a video camera present in the room. While we also equipped participants with the sensors to measure their skin conductance, the resulting data was not reliable and thus discarded.

2) In-between questionnaire: With our in-between questionnaires, we intended to get, for each of the interactions with the robot, a comparable indication of our participants' comfort level, perception of the robot, and how they would suggest the robot change its behavior. This had to be balanced with the need to keep the questionnaire short so as to not disturb the flow of the interaction too much.

Specifically, we asked participants how comfortable they were with the behavior of the robot (sliding scale, 1–100) and to rate the robot as being intelligent, sensitive, pleasant, and thorough (7-point Likert scales, Not at all–Very much). To keep them focused on the task, we also asked how relevant the latest clue was towards solving the case. Lastly, we asked them to indicate desired changes to positioning behavior (7-point scale, The robot should... get much closer–not change its position–stay much further away) and, similarly, to its volume settings (The robot should... increase its volume–not change its volume–decrease its volume). We concluded each in-between questionnaire with an open question in which participants could give other suggestions for improvement.

After the 9th interaction we used the same questionnaire, but swapped out the question on the relevance of the clue for an open question on who they suspected to have committed the murder.

3) Post-experiment questionnaire: After the experiment was over, we asked the participants for demographic information that could pertain to their social distancing preferences. Specifically, we asked for gender, age, education, country of origin, history of pet ownership [16], and prior experience with robots. In addition, given our manipulation of environment noise, we checked participants' hearing loss, and asked participants to indicate how they experienced the noise level in the lab (7-point Likert scale, no noise at all - a lot of noise).

D. Materials

For our data collection we used the hardware of a Giraff telepresence robot. However, rather than using it as a telepresence robot, we modified it to show two animated eyes on its screen which would occassionally blink. We prepared a Wizard-of-Oz set-up which allowed the experimenter to control the robot, and to quickly and efficiently select and play pre-recorded audio files on the robot. The experimenter controlled the robot from a laptop, located in a screened off corner of the experiment room (Figure 2).

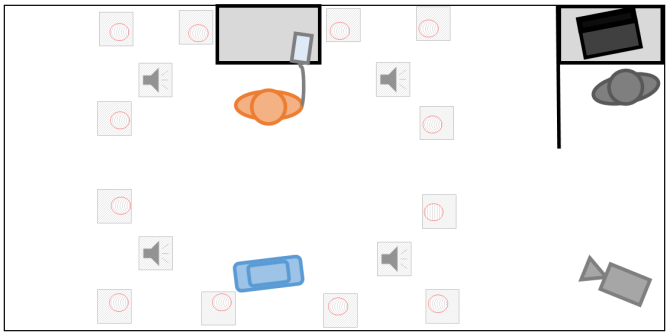


Fig. 2. Overview of the experiment room, showing the Wizard-of-Oz set-up with the experimenter (top-right), and the interaction between the participant and the robot (middle). Behind the participant was a table with a device for measuring skin conductance, to which they were connected through a wire. The overview also shows the location of the video camera (bottom-right). Located on and just below the drop ceiling were the infrared cameras (hatched squares with circle) and the speakers (hatched square with speaker icon).

E. Procedure

Participants came in, received a briefing, and read and signed an informed consent form. After that we equipped them with the markers, under the ruse that those markers would help the robot to navigate autonomously. We also hooked them up to the sensors for measuring skin conductance – which forced them to stay relatively close to the measurement tool.

They were then instructed on the task: Solve a murder mystery, the robot will go collect the cues. We told participants that the aim of the study was for them to collaborate with the robot and we encouraged them to talk with the robot, while warning them that its capacities for natural speech were limited. We further instructed them that each time the robot would go collect the next clue, the participants would be presented with a brief questionnaire (presented on a tablet). Just before the interaction started, we started the data collection and conducted a brief calibration.

The interaction then proceeded as described in Section II-A. Each time the robot went to collect the next clue, the experimenter would present the participant with an in-between questionnaire on a tablet. Otherwise, the experimenter hid in a screened off section of the experiment room and controlled the robot (Section II-D).

After the interaction with the robot was completed, participants were asked to fill in the post-experiment questionnaire. After the experiment was over, we fully debriefed our participants and offered them €6 for their efforts.

F. Participants

A total of 30 participants joined in our data collection. Of these, 21 (70%) identified as male, the rest as female. Most were students, with ages between 17 and 27 (mean age 21.73). The majority (73%) of our participants had the Netherlands as their country of origin. In our other demographic questions, we saw many participants who still/once owned or took care

of a pet (83%), and a fair distribution of prior experience with robots (2 with no prior experience, 14 who had seen robots before, 9 who had interacted with robots before, and 5 who had worked with or programmed robots before).

None of our participants ever wore a hearing aid, and a great majority did not feel they had a hearing loss (90%), nor did their friends or family think they had a hearing loss (90%). One participants rated their hearing as poor, while the rest rated it as fair (4 participants) or better.

III. Testing for effects of approach distance and environment noise on perception

The controlled data we collected can be seen as an experiment testing for the single and joint effects of approach distance and environment noise on perception. Within the context of this work, our main goal was to ensure that perception of the robot's behavior would not depend exclusively on the approach distance. Specifically, we expected a joint effect of approach distance and environment noise.

As mentioned above, in contrast to what we expected, our results were null-results. In this section we will briefly discuss the research question and hypotheses that guided our analysis (Section III-A), the results (Section III-B), and the implications of those results (Section III-C).

A. Research question

In our data collection set-up, we used two manipulations (approach distance and environment noise) and a subjective measure (the in-between questionnaires). The questions in the in-between questionnaires reflected several aspects of our participants' perception of the robot. As such, we defined the following research question;

What are the single and joint effects of approach distance and environment noise on the way a robot is perceived?

Our main interest in this question was to ensure that perception of the robot would not exclusively depend on its used approach distance (Figure 1). In addition, if we had found environment noise to have an effect, it would have been an additional piece of evidence within the relatively small body of literature on the effect of perceptual needs on proxemic preferences in human-robot interaction.

As discussed in more detail in Section II-B, both our conditions were chosen because, based on the literature and our small pilot, we would expect them to have an effect.

B. Results

A principal component analysis (PCA) was run on the 8 items of the in-between questionnaires. Inspection of the correlation matrix showed that the question about the relevance of the clue had no correlations with the other questions greater than 0.3, which is not surprising as it was primarily included to check that the relevance of the clue would not influence our measures; that question was excluded from further analysis. Overall Kaier-Meyer-Olkin was reasonable (0.779), though individual measures for the items on the robot changing its position (0.442) and its volume (.547) were low. When these items were recoded to a scale from 'strong change suggested' to 'no change suggested', these individual measures improved (to .541 and .559 respectively, with overall Kaier-Meyer-Olkin .774). Data was likely factorizable (Bartlett's test of sphericity, p=.000).

PCA revealed two components that had eigenvalues greater than one and which together explained 68.0% of variance (48.4% and 19.6%). The first component had strong loadings of the questions on perception of the robot (Comfortable, Intelligent, Sensitive, Pleasant, Thorough), while the second component had strong loadings of the two questions about the robot changing its position and its volume.

For the remainder of our analysis, we will use the component-based averaged scores for these components, labeled as 'Perception' and 'Suggested improvement'.

We conducted a two-way mixed ANOVA to investigate the single and joint effects of approach distance and environment noise on Perception and Suggested improvement.

For Perception, there was no significant interaction between our conditions ($F(3,78)=.306$, $p=.821$). Therefore, we looked into the main effects, but found no significant effects of either approach distance ($F(3,78)=1.357$, $p=.262$) or environment noise ($F(1,26)=.161$, $p=.691$).

For Suggested improvement, there was no significant interaction between our conditions either ($F(3,78)=1.100$, $p=.343$). Therefore, we looked into the main effects, but again found no significant effects of either approach distance ($F(3,78)=1.851$, $p=.165$) or environment noise ($F(1,26)=2.805$, $p=.106$).

C. Conclusions and discussion

These results show that, in this particular dataset, there is no strong effect of approach distance and/or environment noise on perception of our participants – i.e. we had null results.

These null results also pose an interesting question: why did we not find results, despite following an extensive body of work on social positioning behavior? It is important to note that absence of evidence is not evidence of absence; our findings do not disprove this body of work, especially not with our relatively limited sample size.

We want to here speculate about one possible explanation; a key difference between this study and much of the previous work on social positioning, is that we explicitly allowed our participants to move in response to the behaviors of the robot. In other words, it might be that being able to adapt your own position can alleviate the effects of a robot's distancing on the perception of people. This makes intuitive sense, as people frequently make small adaptations to find an interaction distance with which they are comfortable (as discussed in a.o. the equilibrium theory [20]). This also aligns with our findings in earlier studies, where we also allowed participants to move around, and also found no significant effect of approach distance on perception of the robot [7], [21].

IV. Detecting social feedback cues

Our aim was to get insights in the relation(s) between non-verbal cues from body posture and perception of a robot as

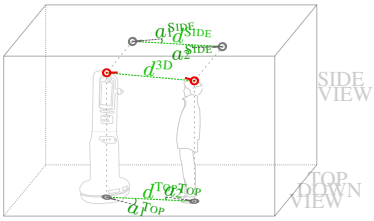
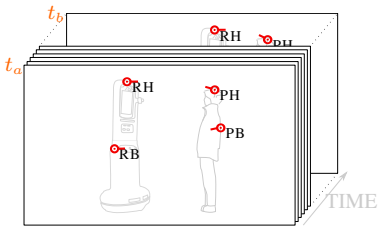
Relations between two markers $f(\mathbf{\sigma}, \mathbf{\sigma})$			**Markers** location pairs $(\mathbf{\sigma}_1^x, \mathbf{\sigma}_2^x)$	time windows (t_a, t_b)		**Configurations** to create features for each combination of a relation, a marker pair, and a time window	**Feature function templates**
						single moment	$f\left(\mathbf{\sigma}_1^{t_a}, \mathbf{\sigma}_2^{t_a}\right)$
						difference ($\Delta_{t=t_a,t_b}$)	$f\left(\mathbf{\sigma}_1^{t_a}, \mathbf{\sigma}_2^{t_a}\right) - f\left(\mathbf{\sigma}_1^{t_b}, \mathbf{\sigma}_2^{t_b}\right)$
						time-compare	$f\left(\mathbf{\sigma}_1^{t_a}, \mathbf{\sigma}_1^{t_b}\right)$
						Combine f-outcomes for each frame in the time window;	
Distance	in 3D space	d^{3D}	(robot head,	(-5,-3)	(-5,-1) (-5,0)	- average	$\text{AVG}(\{m \mid (t \in \mathbb{W}) f\left(\mathbf{\sigma}_1^t, \mathbf{\sigma}_2^t\right)\})$
	projected on side view	d^{SIDE}	participant head)	(-5,1)	(-5,3) (-5,5)	- standard-deviation	$\text{SD}(\{m \mid (t \in \mathbb{W}) f\left(\mathbf{\sigma}_1^t, \mathbf{\sigma}_2^t\right)\})$
	projected on top-down view	d^{TOP}	(robot body,	(-3,-1)	(-3,0) (-3,1)	- minimum	$\text{MIN}(\{m \mid (t \in \mathbb{W}) f\left(\mathbf{\sigma}_1^t, \mathbf{\sigma}_2^t\right)\})$
Angle	projected on side view	$a1^{\text{SIDE}}$	participant body)	(-3,3)	(-3,5) (-1,0)	- maximum	$\text{MAX}(\{m \mid (t \in \mathbb{W}) f\left(\mathbf{\sigma}_1^t, \mathbf{\sigma}_2^t\right)\})$
		$a2^{\text{SIDE}}$	(participant head,	(-1,1)	(-1,3) (-1,5)	- Δ average	$\text{AVG}(\{m \mid (t \in \mathbb{W}) \Delta_{u=t,t+1}(f\left(\mathbf{\sigma}_1^u, \mathbf{\sigma}_2^u\right))\})$
	projected on top-down view	$a1^{\text{TOP}}$	participant body)	(0,1)	(0,3) (0,5)	- Δ standard-deviation	$\text{SD}(\{m \mid (t \in \mathbb{W}) \Delta_{u=t,t+1}(f\left(\mathbf{\sigma}_1^u, \mathbf{\sigma}_2^u\right))\})$
		$a2^{\text{TOP}}$		(1,3)	(1,5) (3,5)	- Δ minimum	$\text{MIN}(\{m \mid (t \in \mathbb{W}) \Delta_{u=t,t+1}(f\left(\mathbf{\sigma}_1^u, \mathbf{\sigma}_2^u\right))\})$
						- Δ maximum	$\text{MAX}(\{m \mid (t \in \mathbb{W}) \Delta_{u=t,t+1}(f\left(\mathbf{\sigma}_1^u, \mathbf{\sigma}_2^u\right))\})$

TABLE I

OVERVIEW OF THE FEATURES WE GENERATED FROM OUR TRACKING DATA, USING DIFFERENT MEASURES OF THE DISTANCE AND ANGLE BETWEEN TWO MARKERS, ON A REPRESENTATIVE SET OF MARKER PAIRS AND TIME WINDOWS, IN A RANGE OF FEATURE FUNCTION TEMPLATES ('CONFIGURATIONS').

being appropriately positioned (or not) – such that it could be automatically detected. To this end, we tried to implement an effective classifier.

Within the scope of this paper, implementing an effective classifier is a means, not the end. In other words, we are and were not aiming for "perfect scores" and one should not expect them, as we are trying to read peoples' inner thoughts from a relatively small dataset. Instead, we have searched for relevant insights about what factors would play a role in developing and optimizing feature selection for such a classifier. For this reason, we focused on feature selection, and used a standard random forest classifier with 500 trees [22] as our classifier, without further tuning.

We will in this section discuss how from the raw data we derived a wide range of features and our labels (Section IV-A) and then tried to find which features were relevant for a classifier (Section IV-B).

A. Data preparation and feature extraction

From our data collection, we ended up with relatively clean data. The OptiTrack gave us temporal data on position (x,y,z) and 3D orientation (quaternion) for the four markers we used: participant head, participant body, robot head, and robot body. From the in-between questionnaires we further had a measure of participants' perception of the robot for each interaction (8 per participant). We used the questionnaire data to derive our labels (IV-A1) and the tracking data to derive our features (IV-A2).

1) Labels: As our labels, we used the participants' answers to the question on how the robot should change its positioning. We chose to do so, because those directly reflected their opinion, in contrast to the two constructs we *derived* from the questionnaire (Perception and Suggested improvement, Section III-B).

To limit the number of classes, we translated the 7-point scale into three bins: 1-3 'get closer', 4 'don't change position', and 5-7 'stay further away'. These bins resulted in a uneven distribution of the classes, with stay further away being chosen with roughly twice the frequency of get closer and don't change position.

2) Features: Even though we only tracked four markers, there are already many different aspects that could be relevant. These include different relations that exist between markers (we used several measures of distance and angle) and that have been hypothesized to serve as social cues in the context of social positioning (see, e.g. [5]), different time-windows that we could consider around the end of the approach, and different ways of combining these measures in a way that takes temporal aspects into account. A concise overview of the features we generated, and the way in which we generated them, can be found in Table I.

a) Marker pairs: We selected the two marker pairs that most richly encoded the relative positioning of the robot and our participants – (robot-head, participant-head) and (robot-body, participant-body). Since we did not change the robot's posture (just its position) and were mostly interested in the participant's behavior, other pairs that included markers on the robot did not really provide additional information and were thus excluded. We *did* include the pair of the two markers worn by the participant (participant-head, participant-body), as that would yield several features that could serve as social cues, such as gaze aversion and leaning behavior [5], [9], [10].

b) Time windows: Time windows were defined by their start and end time (t_a, t_b), with the window ($\mathbb{W}$) also encom-

passing all time frames in between ($\mathbb{W} = \{t|t \geq t_a, t \leq t_b\}$). As our temporal point of reference ($t = 0$), we used the end of the robot approach. This had the practical benefit that it was easily and reliably derivable from the robot reducing its speed to zero (we did this automatically, using an over-sensitive metric, and then manually removed the false positives and checked the outcomes).

c) Generating features: We deliberately generated an exhaustive set of features with all combinations of these aspects, rather than making assumptions on which features to pick. The downside hereof was that it resulted in 4410 unique features[1], which is excessive for this small a dataset and thus necessitated the use of feature selection. At the same time, this had the advantage that said feature selection could potentially provide us with information on which features were effective for our classifier.

B. Feature selection

After generating this many features, our next step was to try and conduct suitable feature selection. At first, we used a combination of feature pre-selection based on a variance threshold, selection of features having the highest chi-squared scores, and feature selection based on gini-importance in a random forest. Initial results on a participant-independent train set and test set seemed barely above chance level (Appendix A1), which improved to a very small but significant difference on a participant-dependent train set and test set (Appendix A2). Performance in both cases had a very high variance, which suggested that there were meaningful features to be found, but the feature selection had difficulty finding them. We confirmed this by using a set of features that was successful in one of the training folds and showing that, without further tuning, these features significantly improved performance on the participant-dependent test set.

Based on our findings with automatic feature selection we expected that suitable features existed and were occasionally but unreliably found by our automatic feature selection. To investigate this, we tried how effective classification would be if we used a set of features that had been found to be 'successful'.

To find this set of supposedly suitable features, we looked into the features that were found during automatic feature selection. Specifically, we selected features by using those from the (outlier) highest-performing classifier in one of the training folds (participant dependent, using all types of feature selection with t=3,k=10,n=50 (see Appendix A)), with a precision of .609. Cross-validation of these features on the other folds in the train set also seemed promising, with an average precision of .394. An overview of these four features can be found in Table II, along with our interpretations on what these features could entail. It is important to note that these are only hypotheses; further work will be necessary to investigate our interpretations, and to see if and how these features generalize.

Since we found these features based on their performance within one fold of the participant-dependent train set, we needed to test them on the participant-dependent test set as well.

Performance on the test set was reasonably good (see Table III), and better than what we had previously found with automatic feature selection. Performance on the individual classes aligned with their frequency in the train and test set, with performance on 'stay further away' being higher than performance on 'not change its position', which in turn was higher than performance on 'get closer'. The latter performs below chance level, which indicates that our classifier is better at detecting when the robot was perceived as uncomfortably close than when it is perceived as annoyingly distant. This might indicate that humans tend to give feedback when they feel their personal space is violated and less so, or would compensate themselves, if they felt their conversation partner is respecting their personal space too much.

We further investigated if performance was significantly better than what would be expected of a random classifier taking into account the relative frequencies of the different classes. For this, we used repeated holdout validation, splitting the full dataset into different random train and test sets, on which we then trained and tested our classifier with the chosen features. We repeated this a total of 20 times and then compared against the expected value of the random classifier. To compare the outcomes, we ran a one-tailed Wilcoxon signed-rank test, which showed a significant increase in average precision between our trained classifier (median of .483) and the random classifier (median of .366), Z=3.360, p=.001.

Overall, these findings show that in a participant-dependent case there are indeed social feedback cues that a robot might use to detect if people think it chose an appropriate interaction distance. As noted before, our approach in selecting these features here does not allow for generalizations to a participant-independent case.

V. Conclusions and discussion

In this paper we have investigated if it is possible to automatically detect and interpret social feedback cues, in the context of approach distance for our specific robot platform. In other words, we have investigated if people's nonverbal behaviors (position/orientation upper body and head, over time) contain detectable information about their perception of a robot's approach behavior.

We started by collecting an extensive dataset, manipulating approach distance and environment noise, measuring the perceived appropriateness, and tracking temporal positioning information. This dataset is being made publicly available as part of this publication.[2]. The two conditions in this dataset, approach distance and environment noise, were, unexpectedly, not found to have any significant single or joint effects on

[1] 7 relations $\times$ 3 marker pairs $\times$ 21 time windows $\times$ 11 configurations $=$ 4831, minus the duplicate5 features resulting from ignoring the second time-marker for single moment features ($7 \times 3 \times 14$) and ignoring the second marker for time-compare features ($7 \times 1 \times 21$).

[2] http://doi.org/10.4121/uuid:b76c3a6f-f7d5-418e-874a-d6140853e1fa

Feature	Interpretation of what the feature could capture	*Average gini-importance*
$\Delta_{t=\text{-}5,\text{-}3}\left(d^{\text{3D}}\left(\boldsymbol{\sigma}^{t}_{PH},\boldsymbol{\sigma}^{t}_{RH}\right)\right)$	**"anticipatory leaning"** anticipation based on earlier trials could cause the odd time window (-5,-3)?	*.263*
$\text{MAX}\left(\left\{m \middle\vert (t \in \mathbb{W}_{(\text{-}3,0)})\Delta_{u=t,t+1}\left(a^{\text{SIDE}}(\boldsymbol{\sigma}^{u}_{PB},\boldsymbol{\sigma}^{u}_{RB})\right)\right\}\right)$	**"abruptness of body up-down rotation"** caused by stepping away from/towards the robot?	*.232*
$\text{MAX}\left(\left\{m \middle\vert (t \in \mathbb{W}_{(\text{-}3,0)})\Delta_{u=t,t+1}\left(a^{\text{TOP}}(\boldsymbol{\sigma}^{u}_{PH},\boldsymbol{\sigma}^{u}_{RH})\right)\right\}\right)$	**"abruptness of head left-right rotation"** caused by participants aiming an ear to the robot (to hear it better)?	*.260*
$\Delta_{t=\text{-}5,1}\left(d^{\text{3D}}\left(\boldsymbol{\sigma}^{t}_{PH},\boldsymbol{\sigma}^{t}_{PB}\right)\right)$	**"increased tilting towards/away from the robot"** using the measure at t=-5 as a baseline?	*.244*

TABLE II

OVERVIEW OF THE FEATURES USED BY OUR EVALUATED CLASSIFIER, AND THE CLASSIFIERS USED IN THE CROSS-EVALUATION. AS AN INDICATOR OF THEIR (RELATIVE) IMPORTANCE WE HAVE GIVEN THEIR AVERAGE GINI-IMPORTANCE IN THE CLASSIFIERS USED IN THE CROSS-EVALUATION.

	The robot should... ...get closer	...not change its position	...stay further away	*Average*
Precision	.166	.500	.692	*.453*
Recall	.333	.364	.692	*.463*
F1-score	.222	.421	.692	*.445*

TABLE III

PERFORMANCE OF OUR CLASSIFIER ON THE TEST SET, TRAINED WITH ONLY THE SET OF FEATURES LISTED IN TABLE II. WE HAVE LISTED PERFORMANCE IN TERMS OF PRECISION, RECALL, AND F1-SCORE FOR EACH OF THE THREE CLASSES, AS WELL AS AVERAGE PERFORMANCE.

perceived appropriateness. We have hypothesized that this was caused by our participants having the freedom to compensate for the behavior of the robot by repositioning themselves, but further research will be necessary to test this hypothesis.

After initial struggles to get our classifier to detect and interpret social feedback cues in our dataset with the many features we generated, we managed a small but significant improvement over a random classifier. As we suspected that these struggles were caused by a difficulty in reliably identifying suitable features, we also looked at a specific set of features (Table II) that performed particularly well on one of the training folds. Use of these features resulted in a classifier with a somewhat more substantial improvement over random for a participant-dependent case.

Together these findings show that in this context, at least in the participant-dependent case, features can be found that provide information about subjectively perceived appropriateness. As the four found features all use time-windows that can be computed within 1 second after the end of the approach, this detection may well be quick enough to allow a robot to respond and try improving its behavior.

At the same time, there are several things that the current work deliberately does *not* do. Firstly, our findings are based on this specific context and robotic platform, so we cannot make strong claims on the generalizability of our findings. Secondly, though significantly better than chance, there is still a lot of space for improvement in the automatic detection of social feedback cues. And thirdly, we have only looked at the nonverbal cues encoded in position/orientation of our participants' upper body and head; there may be many other relevant cues, e.g. in facial expressions, in perspiration, in verbal utterances, or in various aspects of body language.

The different parts of the work reported here can, however, serve as a starting point for further work in these directions. The methods and considerations involved in our collection of the dataset could be a starting point for data collections with similar aims in a wide range of different contexts – with robots, or with other kinds of social agents. Furthermore, such data collections could focus on a rich variety of different nonverbal cues as well. In addition, the dataset we collected could be used with more sophisticated machine learning techniques to try and achieve a better performance.

Overall, we have taken the first steps, showing that a robot *could* detect it got too close during its approach. This provides an initial investigation of our main question of how to get robots to detect social feedback cues, and opens the door to the practical implementation of robots that behave responsively in social situations.

REFERENCES

[1] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013.

[2] C. Brandl, A. Mertens, and C. M. Schlick, "Human-robot interaction in assisted personal services: factors influencing distances that humans will accept between themselves and an approaching service robot," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 26, no. 6, pp. 713–727, 2016.

[3] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 137–153, 2015.

[4] J. Vroon, G. Englebienne, and V. Evers, "Responsive social agents: Feedback-sensitive behavior generation for social interactions," in *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings*, A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, Eds. Springer International Publishing, 2016, pp. 126–137.

[5] J. N. Cappella, "Mutual influence in expressive behavior: Adult–adult and infant–adult dyadic interaction." *Psychological bulletin*, vol. 89, no. 1, p. 101, 1981.

[6] A. Veenstra and H. Hung, "Do they like me? Using video cues to predict desires during speed-dates," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 838–845.

[7] J. Vroon, M. Joosse, M. Lohse, J. Kolkmeier, J. Kim, K. Truong, G. Englebienne, D. Heylen, and V. Evers, "Dynamics of social positioning patterns in group-robot interactions," in *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE, 2015, pp. 394–399.

[8] R. Mead and M. J. Mataric, "Robots have needs too: People adapt their proxemic preferences to improve autonomous robot recognition of human social signals," *New Frontiers in Human-Robot Interaction*, p. 100, 2015.
[9] A. Mehrabian, "Relationship of attitude to seated posture, orientation, and distance." *Journal of personality and social psychology*, vol. 10, no. 1, p. 26, 1968.
[10] M. L. Patterson, S. Mullens, and J. Romano, "Compensatory reactions to spatial intrusion," *Sociometry*, pp. 114–121, 1971.
[11] B. R. Schadenberg, M. A. Neerincx, F. Cnossen, and R. Looije, "Personalising game difficulty to keep children motivated to play with a social robot: A Bayesian approach," *Cognitive systems research*, vol. 43, pp. 222–231, 2017.
[12] H. Buschmeier and S. Kopp, "Towards conversational agents that attend to and adapt to communicative user feedback," in *International Workshop on Intelligent Virtual Agents*. Springer, 2011, pp. 169–182.
[13] K. Shiarlis, J. Messias, M. van Someren, S. Whiteson, J. Kim, J. Vroon, G. Englebienne, K. Truong, V. Evers, N. Pérez-Higueras, I. Pérez-Hurtado, R. Ramon-Vigo, F. Caballero, L. Merino, J. Shen, S. Petridis, M. Pantic, L. Hedman, M. Scherlund, R. Koster, and H. Michel, "Teresa: A socially intelligent semi-autonomous telepresence system," in *ICRA 2015: Proceedings of the IEEE International Conference on Robotics and Automation, Workshop on Machine Learning for Social Robotics*. IEEE, 2015.
[14] E. T. Hall, *The Hidden Dimension*. Anchor Books New York, 1966.
[15] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, D. S. Syrdal, and C. L. Nehaniv, "An empirical framework for human-robot proxemics," *Procs of New Frontiers in Human-Robot Interaction*, 2009.
[16] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 5495–5502.
[17] M. P. Joosse, R. W. Poppe, M. Lohse, and V. Evers, "Cultural differences in how an engagement-seeking robot should approach a group of people," in *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology*. ACM, 2014, pp. 121–130.
[18] R. Mead and M. J. Matarić, "Perceptual models of human-robot proxemics," in *Experimental Robotics*. Springer, 2016, pp. 261–276.
[19] J. Vroon, J. Kim, and R. Koster, "Robot response behaviors to accommodate hearing problems," in *Proceedings of New Friends 2015: the 1st International Conference on Social Robotics in Therapy and Education, Almere, the Netherlands*. Windesheim Flevoland University, 2015, pp. 48–49.
[20] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965.
[21] S.-D. Suvei, J. Vroon, V. Somoza, L. Bodenhagen, G. Englebienne, N. Kruger, and V. Evers, ""I would like to get close to you": Making robot personal space invasion less intrusive with a social gaze cue," in *(IN PRESS)*, 2018.
[22] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 278–282.

APPENDIX

A. Automatic feature selection

Our dataset consisted of 240 data points (30 participants $\times$ 8 interactions), with for each data point 4410 feature values and 1 label. This dataset was split in a train and a test set, further reducing the number of data points available for training. As such, our dataset is relatively small *and* any classifier is likely prone to overfitting. These challenges motivated our decision for which classification procedure to use.

Firstly, we chose to use two forms of feature selection, to reduce the number of used features and reduce the risk of overfitting. We used a chi-squared score based method to pre-select a subset of `k` features with the highest scores, after a pre-selection based on variance threshold (cutoff at .8). We then further selected from these features by training a random forest with `n` trees, and then selecting features that had a gini-importance higher than 1/`k`.[3]

For completeness, we tested four cases; no feature selection, only chi-squared score based feature selection, both forms of feature selection mentioned above, and manual feature selection. The latter, manual feature selection, was added as an alternative and had resulted in a selection of 8 features together representing stepping away, leaning away, and averting gaze in the (0,+1s) and (-1s,+3s) time frames. Using only random forest based feature selection (without chi-squared score based feature selection) was not included as a case as it is ineffective; with 4410 features, gini-importance of each individual feature would become so low that selection based on those gini-importances would be too sensitive to noise.

Secondly, as we were interested in the relevance of the different time-windows to the performance of the classifier, we also manipulated the time-windows that would be included in the dataset. Introducing `t`, as a variable taking one of the moments [-3s, -1s, 0s, +1s, +3s, +5s], we would only include features in the dataset that had values for time-windows up to and preceding that moment.

Together, this introduced several hyper-parameters that we wanted to investigate; `k`, `n`, `t`, and type(s) of feature selection to use. For this we used cross validation within the training set.

1) Participant-independent classification: In our first go, we tried to train a participant-independent classifier. That is, we split the dataset into a training set and a test set, such that all data points of a participant were in the same set. We similarly separated validation sets from the training set for our cross validation. This ensured that the classifier would always be tested with data points from a participant it had not been trained on – which would mean that, in the case of a good performance, our findings would likely be easily generalizable to new and previously unseen people. We wanted to avoid too large an influence of outlier participants, and thus split the dataset in 5 parts of 6 randomly chosen participants each; one formed the test set, the others the folds in the training set (4-fold cross-validation).

Already in our cross-validation, we saw that performance mostly was barely above chance-level[4] – despite feature selection, and across all hyper-parameters. To our surprise, performance on the training set was consistently near-perfect, while performance on the validate-set would drop to chance-levels. This seemed to be partly due to the curse of dimensionality – without feature selection, performance barely increased even when we included the correct label as a feature, demonstrating that the algorithm was unable to identify relevant features from

[3]Since the sum of gini-importances over all features for a random forest is equal to 1, and `k` is the number of features used, this method selects all features that had a higher gini-importance than could be expected based on chance.

[4]It is worth noting that performance for a `t` of -3s was especially low, being consistently below chance-level with an overall average precision of .262. This does make intuitive sense, as in that time-window the robot would have barely started its approach, thus providing participants with little to no reason to already judge the appropriateness of the robot's behavior.

a feature set of this size. However, even with small feature sets, we still had similar results.

We investigated several alternatives, trying to challenge our assumptions, but found no increase in performance. Firstly, we investigated our choice for the labels, by also testing with labels derived from binning participants' score on Perception and Suggested improvement (see Section III-B) – this did not seem to affect performance. Secondly, we investigated our choice for the random forest classifier. We tried several other classifiers (Naive Bayes, Support Vector Machines), but to no avail. We also tried further tuning of other parameters of the random forest classifier, aiming to make it less susceptible to overfitting to our feature set; trying several different numbers for the maximum number of features used by each tree, enforcing a maximum depth for the trees, and increasing the number of samples that were required to split an internal node. Again, this did not seem to affect performance in our cross-validation.

Since all these alternatives failed, the most likely explanation seemed to be that aspects of individual participants *did* matter and should be taken into account. This also posed an explanation for the recurring observation that performance in the training set was near perfect, dropping to close to random performance in the validation set only. We consequently decided to try participant-dependent classification instead.

2) Participant-dependent classification: As we suspected that aspects of individual participants played an important role, based on our first attempt discussed above, we tried participant-dependent classification. To do so, we split the dataset in a train (200 data points) and test set (40 data points) such that the data points of individual participants were spread across these two sets. We similarly separated validation sets from the training set for our cross validation, creating 5 folds with 40 data points each. This ensured that our classifier would usually have encountered a few data points from each participant in its training set before validation and testing. In our initial tests, we found that this already seemed to improve performance a bit, even without feature selection (average precision of .433 on the training set), and we thus investigated this more in-depth.

We then ran our full cross-validation, to find a hyperparameter setting where average performance was high and stable in terms of standard deviation and average performance with similar hyperparameter settings. Of the two peaks found, the peak around `t`=-1, `k`=10, and `n`=5000 (average precision of .462 with standard deviation .068) was discarded as standard deviations for those values were relatively high. We thus chose to go with feature selection based on both chi-squared score (`k`=45) and gini-importance in a random forest (`n`=100) for time frames up to `t`=0 (average precision of .452 with standard deviation of .038, and similarly low standard deviations for similar hyperparameters).

We trained our classifier on the full training set, with feature selection using the found hyperparameter settings, and then tested its performance on the test set (holdout validation). This resulted in an average precision of .38 on the test set.

As this is but a small improvement relative to what would be expected of a random classifier with three classes, we further wanted to investigate if performance was consistently better than random. To ensure a fair comparison, we used the expected precision of a random classifier that would take into account the relative frequencies of the different classes in the training set – which, given the distribution of labels in our dataset, resulted in a performance slightly better than pure random. For this comparison, we used repeated holdout validation. We took the full dataset and split it into different random train and test sets. On these splits, using the found hyperparameter settings, we then trained and tested our classifier. We repeated this a total of 20 times to get a reasonable sample. For each of these splits, we also computed the expected value of the random classifier. To compare the outcomes, we ran a one-tailed Wilcoxon signed-rank test, which showed a very small but significant increase in average precision between our trained classifier (median of .391) and the random classifier (median of .376), Z=1.792, p=.037.

While we thus found a small but significant improvement of our trained classifiers, it is worth noting that the variance of the trained classifiers was much higher than that in the random classifier, with standard deviations of .064 and .016, respectively. As performance improved upon random, we can conclude that features do contain (some) information on the labels. At the same time, the small difference and the high standard deviations strongly suggest that the used automatic feature selection could not (yet) reliably find these features.