Int J of Soc Robotics (2017) 9:659–674
DOI 10.1007/s12369-017-0414-y

Automatically Classifying User Engagement for Dynamic Multi-party Human–Robot Interaction

Mary Ellen Foster[1] · **Andre Gaschler**[2] · **Manuel Giuliani**[3]

Accepted: 23 May 2017 / Published online: 20 July 2017

Abstract A robot agent designed to engage in real-world human–robot joint action must be able to understand the social states of the human users it interacts with in order to behave appropriately. In particular, in a dynamic public space, a crucial task for the robot is to determine the needs and intentions of all of the people in the scene, so that it only interacts with people who intend to interact with it. We address the task of estimating the engagement state of customers for a robot bartender based on the data from audio-visual sensors. We begin with an offline experiment using hidden Markov models, confirming that the sensor data contains the information necessary to estimate user state. We then present two strategies for online state estimation: a rule-based classifier based on observed human behaviour in real bars, and a set of supervised classifiers trained on a labelled corpus. These strategies are compared in offline cross-validation, in an online user study, and through validation against a separate test corpus. These studies show that while the trained classifiers are best in a cross-validation setting, the rule-based classifier performs best with novel data; however, all classifiers also change their estimate too frequently for practical use. To address this issue, we present a final classifier based on Conditional Random Fields: this model has comparable performance on the test data, with increased stability. In summary, though, the rule-based classifier shows competitive performance with the trained classifiers, suggesting that for this task, such a simple model could actually be a preferred option, providing useful online performance while avoiding the implementation and data-scarcity issues involved in using machine learning for this task.

Keywords Human–robot interaction · User engagement classification · Joint action · Socially appropriate behaviour · Multi-party interaction

1 Introduction

Robots will become more and more integrated into daily life over the next decades, with the expectation that the market for service robots will increase greatly over the next 20 years [31]. Everyday interactions, especially in public spaces, differ in several ways from the companion-style interactions that have been traditionally considered in social robotics (e.g., [9,14,34]). First, interactions in public spaces are often short-term, dynamic, multimodal, and multi-party. Second, in a public setting, it is not enough for a robot simply to achieve its task-based goals; instead, it must also be able to satisfy the social goals and obligations that arise through interactions with people in real-world settings. Therefore, we argue that task-based, social interaction in a public space can be seen as an instance of multimodal joint action [32,58].

In this work, we consider the socially aware robot bartender shown in Fig. 1, which has been developed as part

This article integrates and extends the work described in the following conference papers: [17,20,23,24].

✉ Mary Ellen Foster
MaryEllen.Foster@glasgow.ac.uk

Andre Gaschler
gaschlera@gmail.com

Manuel Giuliani
manuel.giuliani@brl.ac.uk

[1] School of Computing Science, University of Glasgow, Glasgow, UK

[2] Fortiss GmbH, Munich, Germany

[3] Bristol Robotics Laboratory, University of the West of England, Bristol, UK

A customer attracts the bartender's attention	
ROBOT:	[Looks at Customer 1] How can I help you?
CUSTOMER 1:	A pint of cider, please.
Another customer attracts the bartender's attention	
ROBOT:	[Looks at Customer 2] One moment, please.
ROBOT:	[Serves Customer 1]
ROBOT:	[Looks at Customer 2] Thanks for waiting. How can I help you?
CUSTOMER 2:	I'd like a pint of beer.
ROBOT:	[Serves Customer 2]

Fig. 1 The JAMES socially aware robot bartender

of the JAMES project.[1] The JAMES robot bartender supports interactions like the one shown in the figure, in which two customers enter the bar area and each attempt to order a drink from the bartender. Note that when the second customer appears while the bartender is engaged with the first customer, the bartender reacts by telling the second customer to wait, finishing the transaction with the first customer, and then serving the second customer. In the bartending scenario, the first step in ensuring successful joint action between robot and customer is to correctly classify the engagement of all potential customers in the scene, both at the start of the interaction and as it progresses: that is, in order to carry out its interactive task, the bartender must be able to understand the social scene in front of it to ensure that it only interacts with potential customers who are actually seeking to engage with it. In this paper, we present the collected findings from the engagement classification work in the context of the JAMES project.

We make use of rule-based and data-driven methods for estimating the desired engagement of customers of the robot bartender. We begin with an off-line experiment for social signal recognition using hidden Markov models. We then compare two classification strategies in the context of the full robot bartender system: a simple, hand-coded, rule-based classifier based on the observation of human behaviour in real bars, and a range of supervised-learning classifiers trained on an annotated corpus based on the sensor data gathered from an initial human–robot experiment. We first compare the two classification strategies through offline cross-validation; then we integrate the rule-based classifier and the top-performing trained classifier into the full robot bartender system and compare them experimentally through interactions with real human users. Because the ground-truth engagement-seeking behaviour of the users in that experimental study is not available, making the practical implications difficult to interpret, we therefore also test the performance of all of the classifiers (rule-based and trained) on a newly-recorded, fully annotated, more balanced test corpus. Finally, we examine the impact of incorporating temporal features into the classifier state by using an alternative classification strategy—Conditional Random Fields—which is particularly suited to this sequence classification task.

2 Related Work

Gaze contact is crucial for establishing social rapport [3], and a number of researchers have addressed the task of estimating engagement based on gaze and other signals. Bohus and Horvitz [6,7] pioneered the use of data-driven methods for this task: they trained models designed to predict user engagement based on information from face tracking, pose estimation, person tracking, group inference, along with recognised speech and touch-screen events. After training, their model was able to predict intended engagement 3–4 s in advance, with a false-positive rate of under 3%. A number of more recent systems have also used machine learning to address this task. For example, Li et al. [44] estimated the attentional state of users of a robot in a public space, combining person tracking, facial expression recognition, and speaking recognition; the classifier performed well in informal real-world experiments. Castellano et al. [10] trained a range of classifiers on labelled data extracted from the logs of children interacting with a chess-playing robot, where the label indicated either high engagement or low engagement. They found that a combination of game context-based and turn-based features could be used to predict user level engagement with an overall accuracy of approximately 80%. McColl and Nejat [48] automatically classified the social accessibility of people interacting with their robot based on their body pose, with four possible levels of accessibility: the levels estimated by their classifier agreed 86% of the time with those of an expert coder. MacHardy et al. [47] classified the engagement states of audience members for an online lecture based on information from facial feature detectors; the overall performance was around 72% on this binary classification task. Hernandez et al. [29] used wearable electrodermal activity sensors to detect the engagement of children during social interactions with adults. Their goal was to automatically predict which children are difficult to engage with in social interactions. Leite et al. [43] compared models for detecting disengagement. They found that mod-

[1] http://www.james-project.eu.

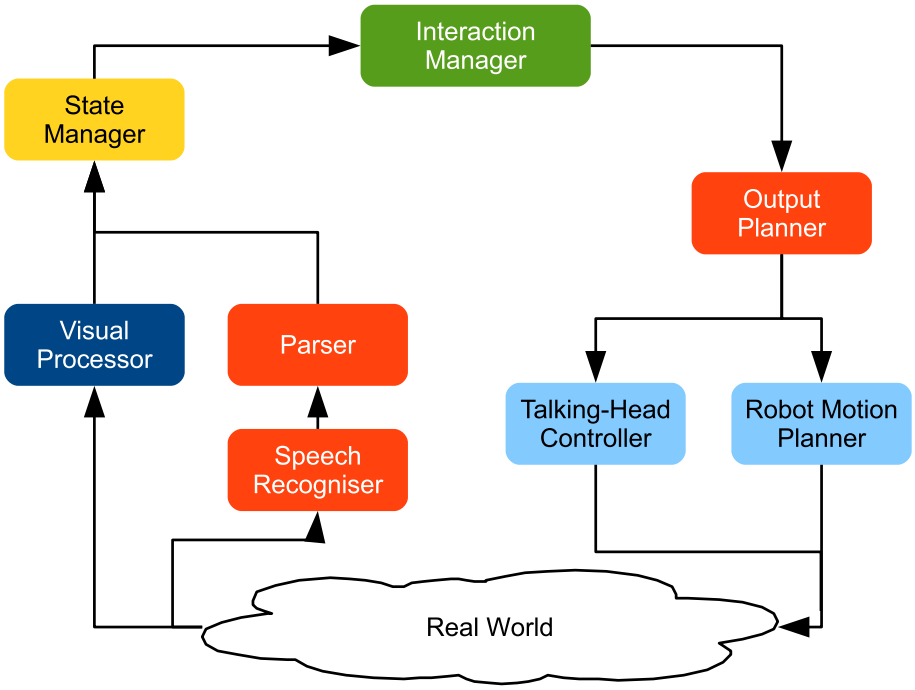

Fig. 2 Software architecture of the JAMES robot bartender

els trained on data from group interactions between several humans and two social robots scale better when applied to single user interactions than the other way around. Further related problems are the prediction of responses in dialogues with embodied agents [15] and turn-taking in general [60].

Our work is also closely related to automatic human activity recognition. Ke et al. [35] survey the use of human activity recognition in single person activity recognition, multiple people interaction and crowd behaviour, and abnormal activity recognition. Aggarwal and Xia [1] give details on human activity recognition with 3D data, similar to the data we are using in our approach. Lara and Labrador [40] give an overview of human activity recognition using wearable sensors. Brand et al. [8] used coupled hidden Markov models for robust visual recognition of human actions. Torta et al. [61] addressed the dual problem of how a robot should attract a human's attention, and found that speech and body language were the most successful, while gaze behaviour was useful only in cases where the human was already attending to the robot. Figueroa-Angulo et al [16] trained a compound hidden Markov model to recognize human activity with RGB-D skeleton data of humans for a service robot. For the related problem of face-to-face conversation, conversation estimation has been demonstrated using visual tracking alone [51–53] or combined RGB-D sensing to analysing and generating multimodal behaviour [49]. Mihoub et al.'s approach for social behaviour modelling and generation is based on incremental discrete hidden Markov models. It can be used to recognise the most likely sequence of cognitive states of a speaker, given his or her multimodal activity, and to predict the most likely sequence of the following activities. Finally, Chen et al. [12] conducted experiments in a "drinking at a bar" scenario. In contrast to our user engagement classification, their intention recognition system with two-layer fuzzy support vector regression identifies most likely orders based on age, gender, nationality, and detected emotions.

Most previous work in engagement classification has approached it as a machine learning problem. The related work shows that machine learning works well for engagement classification for several different interaction settings and using various machine learning algorithms. In comparison to this previous work, we have a more holistic approach for studying error classification. In particular, one of our goals is to compare a simpler classification system with handwritten rules to a machine-learned approach. We have also tested a large variety of different machine learning algorithms on the same data set and interaction scenario. Finally, we tested our rule-based and machine-learned approaches not only in an offline evaluation, as most previous work did, but also in an online human–robot interaction user study with naïve participants.

3 Social Signal Processing in the JAMES Robot Bartender

The JAMES robot bartender incorporates a large number of hardware and software components; Fig. 2 illustrates the software architecture. In summary, the robot senses events in its surroundings through the **speech recogniser** and **visual processor** modules, while the **parser** component processes the output from speech recognition. The **state manager** takes the

output from visual processing and parsing and transforms it into symbolic representations for the **interaction manager** module. The interaction manager then selects high-level actions for the robot, which are processed by the **output planner** for execution as concrete actions by the **talking-head controller** and the **robot motion planner**. Full technical details of the system can be found in [19,25].

The work presented in this paper takes place largely in the context of the state manager (SM), whose primary role is to turn the continuous stream of sensor messages produced by the low-level input-processing components into a discrete representation of the world, the robot, and all entities in the scene, integrating social, interaction-based, and task-based properties. Petrick and Foster [55] give a formal description of the inputs and outputs of the SM. In summary, the input consists of a set of timestamped sensor readings, while the output is a set of first-order predicates denoting properties of all agents in the scene, their locations, torso orientations, engagement states, and drink requests if they have made one. In addition to storing and discretising all the low-level sensor information, the SM also infers additional relations that are not directly reported by the sensors. For example, it fuses information from vision and speech to determine which user should be assigned a recognised spoken contribution, and estimates which customers are in a group. Most importantly in the current scenario—where one of the main tasks is to manage the engagement of multiple simultaneous customers, as in Fig. 1—the SM also informs the rest of the system every time a customer is seeking to engage with the bartender.

The low-level sensor data that is relevant for classifying intended user engagement is available on two input channels. The visual processor [5,54] tracks the location, facial expressions, gaze behaviour, and body language of all people in the scene in real time, using a set of visual sensors including two calibrated stereo cameras and a Microsoft Kinect depth sensor. The data from the vision system is published as frame-by-frame updates approximately every 200 ms. The other primary input modality in the system is linguistic [56], combining a speech recogniser with a natural-language parser to create symbolic representations of the speech from all users. For speech recognition, we use the Microsoft Speech API together with a Kinect directional microphone array; incremental hypotheses are published constantly, and recognised speech with a confidence above a defined threshold is parsed using a grammar implemented in OpenCCG [65] to extract the syntactic and semantic information.

Concretely, for these experiments in user state classification, we make use of the following data from the input sensors:

- The (x, y, z) coordinates of each customer's head, left hand, and right hand as reported by the vision system (Fig. 3);
- The angle of each customer's torso in degrees, where 0° indicates that the customer is facing directly towards the counter;
- The customer's three-dimensional head pose (roll, pitch, yaw); and
- An estimate of whether each customer is currently speaking, derived from the estimated source angle of each speech hypothesis along with the location information from vision.

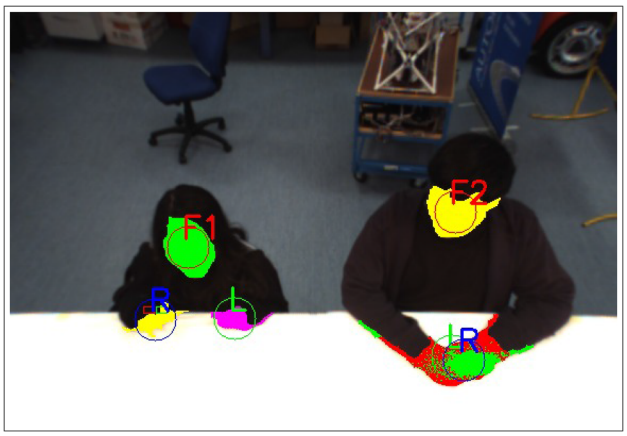

Fig. 3 Output of face and hand tracking (image from [19])

4 Experiment 1: Offline State Classification with Hidden Markov Models

As an initial experiment to test the utility of the sensor data for this task, we trained a supervised hidden Markov model (HMM) to recognize a set of communicative states based on the vision data: that is, using the customers' body and hand coordinates, body angles, and head poses with position (xyz coordinates) and orientation (roll, pitch, yaw). To create the training data, a total of 200 interactions were enacted by five human customers, with up to three customers in one interaction. In this experiment, there was no feedback from the robot system. For later training and off-line evaluation, we recorded the robot's RGB and depth camera views. Customers' interactions contained a total of eight different states: Entering or leaving the scene, an idle state, attention to the robot bartender, attention to a written menu, interaction with another customer, and visible "cheers" gestures and drinking actions. We labelled all data by hand, resulting in a total of 1720 interaction states, which we divided into a training set of 1010 states, a cross-validation set for model optimization of 319 states, and a testing set of 391 states. All participants appeared in both training and test data sets.

As components of the feature vector, we selected body position, body orientation, head orientation (represented both as a normal vector and as pitch and yaw angles), hand posi-

Table 1 Results of the HMM social state recognition experiment

Recognized interaction states	Labelled interaction states									
	b	o	g	c	d	e	i	l	D	%Corr.
Bartender	52	1	0	1	1	2	0	0	6	91.2
Object	0	24	0	0	0	0	1	0	2	96.0
Guest	0	0	36	0	0	0	1	1	7	94.7
Cheers	2	0	0	12	1	1	1	1	6	66.7
Drink	0	0	1	0	35	1	1	0	6	92.1
Enter	0	0	0	0	0	30	0	0	1	100.0
Idle	0	1	1	1	0	1	105	0	17	96.3
Leave	0	0	0	0	0	0	0	30	1	100.0
I	8	10	5	7	4	5	18	8		
Correctness									82.9%	H/N
Accuracy									66.2%	(H–I)/N
Correctly recognized states									324	H
Deletions									46	D
Substitutions									21	S
Insertions									65	I
Number of interaction states									391	N

tions, and two horizontal and frontal distance features to other customers. The fuzzy distance features were computed from the set of body positions of all customers and responded to whether another customer was located in front of or next to an actor. We derived this collection of features through systematic, manual evaluation of the available visual input data and its different representations while observing the correctness and accuracy on the cross-validation set. When experimenting with different features, we observed that hand positions were not significant to detect the interaction states "bartender" or "guest" [23] (i.e., the states relevant to user engagement), but were necessary for detecting the "cheers" and "drink" gestures.

We then modelled the behaviour and state of interaction of each costumer by a separate, continuous-valued multi-dimensional hidden Markov model. As an emission model, full covariance matrices showed slightly more accurate than standard, diagonal variance, which we compared on the cross-validation data set. To prepare training of the model, we measured transition frequencies between states in the training data to bootstrap the hidden state graph and its transition matrices. For the hidden state graph, we defined a linear graph of three hidden states (or, inner states) within each interaction state (or, outer state), with transitions between interaction states if labelled transition frequencies were higher than 5%.

The results of this off-line evaluation are listed in Table 1. The confusion matrix indicates substitutions, false insertions, and deletions of states, and therefore shows the editing distance between the recognized and labelled sequence of states; it does not depend on time frames. In general, we could correctly recognize 82.9% of all states, and the accuracy of the recognition was 66.2%.

We can draw two main conclusions from this stand-alone study using HMMs for state recognition. First, we have confirmed that the attributes available from the visual processor do, in principle, support the recognition of user social states. Secondly, this study agrees with the results of related human–human studies [30,45,46] which suggest that the combined features of head pose and torso orientation are adequate for classifying user engagement in this bartender context. In our HMM experiment, hand positions were not significant for detecting the "bartender", "idle", or "guest" interaction states, which are the states relevant to the overall property of user engagement.

5 Experiment 2: Strategies for Engagement Detection

The preceding section described a stand-alone experiment which tested the performance of HMMs at estimating user states based on a small amount of high-quality data provided by trained actors; the results of that study confirm that the available sensor data is useful for determining user states. We now turn our attention to the task of online detection of user engagement, which—as mentioned previously—is fundamental to interactions with the full robot bartender (e.g., Fig. 1). For this task, we will explore two classification strategies: a **rule-based** classifier that uses a simple, hand-crafted rule derived from the observation of natural interactions in a

real bar, and a set of **trained** classifiers based on an annotated corpus of actual human–robot interactions.

The rule-based engagement classifier relies on the signals observed in real bar customers who signalled that they wanted to engage with the bartender [30]: (1) standing close to the bar, and (2) turning to look at the bartender. These signals were extremely common in the natural data; and, although they seem very simple, in a follow-up classification experiment based on still images and videos drawn from the natural data, the two signals also proved both necessary and sufficient for detecting intended customer engagement [45]. Also, when a different group of human participants were asked to play the role of the human bartender based on a "Ghost-in-the-Machine" paradigm (where the participants had access only to the data detected by the robot sensors), they also paid attention primarily to the signals of position and pose when determining whether to initiate an interaction [46].

Based on the details of the bartender environment, we therefore formalised these two signals into a rule-based classifier that defined a user to be seeking engagement exactly when (1) their head was less than 30cm from the bar, and (2) they were facing approximately forwards (absolute torso angle under 10°)—note that since the bartender robot (Fig. 1) is very large compared to the bar, facing forwards is used as a proxy for looking towards the bartender. In Experiment 1, we also included hand positions in the feature vector; however, that study found that signal to be relevant only for classifying gestures that did not relate to user engagement; the latter can be reliably detected without hand poses [23].

The trained classifiers, on the other hand, make use of a multimodal corpus derived from the system logs and annotated video recordings from the first user study of the robot bartender [19]. In particular, the engagement state of each customer visible in the scene was annotated by an expert with one of three (mutually exclusive) levels: *NotSeeking-Engagement*, *SeekingEngagement*, and *Engaged*. For the current classification task—where we aim to detect users who have not yet engaged with the system but are seeking to do so—the *Engaged* state is not relevant, so the corpus was based on the time spans annotated with one of the other labels. In total, the corpus consisted of 5090 instances: each instance corresponded to a single frame from the vision system, and contained the low-level sensor information for a single customer along with the annotated engagement label. 3972 instances were in the class *NotSeekingEngagement*, while 1118 were labelled as *SeekingEngagement*.

For this initial experiment in trained classification, we used the Weka data mining toolkit [27] to train a range of supervised-learning classifiers on this corpus, using a set of classifiers designed to provide good coverage of different classification styles. To ensure that we selected a wide range of classifiers, we chose the classifier types based on those listed in the Weka primer [63]; The full list of classifiers is given in Table 2. All classifiers were treated as "black boxes", in all cases using the default configuration as provided by Weka version 3.6.8. For training and testing, we treated the corpus as a set of 5090 separate instances; that is, each instance (i.e., frame) was separately classified.

Table 2 Classifiers considered

CVR	Classifies using regression: the target class is binarised, and one regression model is built for each class value [22]
IB1	A nearest-neighbour classifier that uses normalised Euclidean distance to find the closest training instance [2]
J48	Classifies instances using a pruned C4.5 decision tree [57]
JRip	Implements the RIPPER propositional rule learner [13]
LibSVM	Generates a Support Vector Machine using LIBSVM [11]
Logistic	Multinomial logistic regression with a ridge estimator [42]
NaiveBayes	A Naïve Bayes classifier using estimator classes [33]
ZeroR	Baseline classifier; always predicts the most frequent value

Before integrating any engagement classifier into the system for an end-to-end evaluation, we first tested the classifiers in a set of offline experiments to compare the performance of the trained classifiers with each other and with that of the rule-based classifier. This study provides an initial indication of which classification strategies are and which are not suitable for the type of data included in the training corpus, and also gives an indication of the performance of the rule-based classifier on the same data.

5.1 Cross-Validation

We first compared the performance of all of the classifiers through 5-fold cross-validation on the 5090-item training corpus. For each classifier, we computed the following measures: the overall classification accuracy, the area under the ROC curve (AUC), along with the weighted precision, recall, and F measure. Note that the baseline accuracy score for this binary classification task is the size of the larger class (*NotSeekingEngagement*): $3972/5090 = 0.78$. The results of this evaluation are presented in Table 3, sorted by accuracy; the overall performance of the hand-coded rule on the full training corpus is also included. The groupings in Table 3 reflect differences among the accuracy scores that were significant at the $p < 0.01$ level on a paired T test based on 10 independent cross-validation runs. In other words, the IB1 classifier (nearest-neighbour) had the highest performance on this measure; J48 (decision trees), CVR (regression) and JRip (propositional rule learner) were statis-

Table 3 Cross-validation results, grouped by accuracy

Classifier	Accuracy	AUC	Precision	Recall	F
IB1	0.954	0.926	0.954	0.954	0.954
J48	0.928	0.921	0.928	0.928	0.928
CVR	0.912	0.955	0.910	0.912	0.911
JRip	0.910	0.877	0.908	0.910	0.909
LibSVM	0.790	0.521	0.830	0.790	0.706
Logistic	0.781	0.738	0.730	0.781	0.711
ZeroR	0.780	0.500	0.609	0.780	0.684
NaiveBayes	0.665	0.654	0.728	0.665	0.687
Hand-coded rule	0.655	na	0.635	0.654	0.644

Table 4 Output of attribute selection

	Head			HandL			HandR			Ori	Spk	Acc
	x	y	z	x	y	z	x	y	z			
IB1	•	•	•	•			•	•	•		•	0.963
J48	•	•	•	•		•	•	•	•		•	0.932
CVR	•	•	•		•	•	•	•	•		•	0.926
JRip	•	•	•	•		•	•	•			•	0.921
LibSVM	•						•				•	0.830
Logistic												0.780
ZeroR												0.780
NaiveBayes			•	•	•						•	0.786
Hand-coded rule			•							•		0.655
CBF	•	•	•	•	•		•					

tically indistinguishable from each other; LibSVM (support vector machines), Logistic (logistic regression), and ZeroR (baseline—chooses most frequent class) were again indistinguishable (these classifiers generally labelled all instances as *NotSeekingEngagement*); while NaiveBayes (naïve Bayes) and the hand-coded rule (distance + orientation) had the lowest overall accuracy by a significant margin. Figure 4 shows the ROC curves for all classifiers based on the *SeekingEngagement* class: as expected, the curves for all of the high-performing classifiers are close to optimal, while those for the other classifiers are closer to the chance performance of the baseline ZeroR classifier.

5.2 Attribute Selection

The above cross-validation results made use of the full set of sensor attributes included in the corpus; however, it is likely that not all of the sensor data is equally informative for the classification task. To get a better assessment of which sensor data was most relevant, we carried out two forms of attribute selection. We first determined the sensor attributes that were the most informative for each of the individual classifiers, using a wrapper method [37] to explore the relationship between the algorithm and the training data. We then analysed the corpus as a whole using Correlation-Based Feature Selection (CBF) [28], a general-purpose selection method known to have good overall performance [26].

The results of this attribute selection process are shown in Table 4. The main body of the table indicates with a bullet (•) the attributes that were determined to be most informative for each of the classifiers; for reference, the last row shows the two features that were used by the rule-based classifier (z head position and body orientation). The final *Acc* column shows the cross-validation accuracy of a classifier making use only of the selected attributes. As can be seen, most of the high-performing classifiers made use of the full 3D location of the customer's head, along with the 3D location of the hands and the "speaking" flag. The accuracy of most classifiers was very slightly better with the classifier-specific attribute subset when compared to the results from Table 3, but in no cases was this improvement statistically

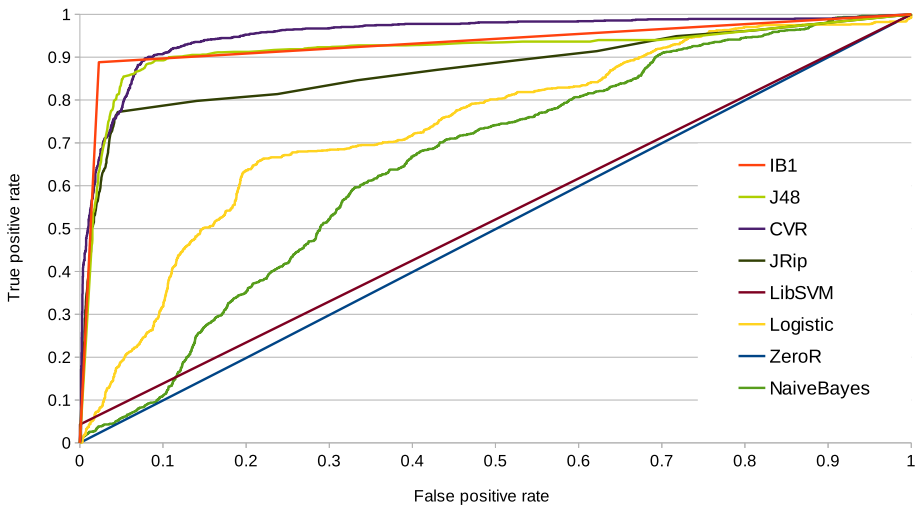

Fig. 4 ROC curves for *SeekingEngagement* class

significant. The bottom row of the table shows the attributes that were found to be most informative by the CBF selector, which were similar to those used by the high-performing classifiers: namely, the full 3D position of the customer's head, along with some of the hand coordinates. The selected attributes correspond very well with the results of the HMM-based study from the previous section.

It is notable that body orientation—which was one of the two main engagement-seeking signals found in the human–human data, and which was found to be necessary for making offline engagement judgements based on that same data—was not determined to be informative by any of the attribute selectors. This is most likely due to the performance of the initial vision system that was used to create the corpus data, which turned out to have difficulty in detecting body orientation reliably, making this attribute unreliable for engagement classification. The unreliability of this signal in the corpus data likely also affected the cross-validation performance of the hand-coded rule (which used both factors found to be relevant to engagement based on the real-world study), which had lower accuracy even than the baseline ZeroR classifier. Also, the right hand was generally found to be more informative than the left: this is probably because, assuming that most customers were right-handed, they would have used this hand more often, thus providing more useful vision data.

6 Experiment 3: Online Comparison of Rule-Based and Trained Classifiers

The offline results presented in the preceding section are promising: in cross-validation against real sensor data, the top-performing trained classifier (IB1) correctly labelled over 95% of the video-frame instances. However, this study was based on frame-by-frame accuracy; and as Bohus and Horvitz [7] point out, for this sort of classifier, a better runtime evaluation is one that measures the errors per person, not per frame.

As a step towards such an evaluation, we therefore integrated the top-performing trained classifier into the robot bartender's state manager (SM) and tested its performance against that of the rule-based classifier through an online evaluation, with human participants playing the role of customers for the robot bartender. This study used the drink-ordering scenario illustrated in Fig. 1: two customers approached the bar together and attempted to engage with the bartender, and—if successful—each ordered a drink. The bartender was static until approached by a customer, and did not engage in any interaction other than that required for the target scenario. As soon as the robot detected a customer intending to engage with it, it would acknowledge their presence by turning its head towards them and speaking: either a greeting (if they were the first customer) or a request to wait (if they were the second)—Fig. 1 contains an example of both behaviours.

For this experiment, in half of the trials, the SM used the rule-based engagement classifier, while for the rest, it instead made use of the IB1 classifier trained on the complete 5090-instance corpus used in the experiment in the preceding section.

6.1 Participants

41 participants (29 male), drawn from university departments outside the German robotics group involved in developing the bartender, took part in this experiment. The mean age of the participants was 27.8 (range 16–50), and their mean self-rating of experience with human–robot interaction systems was 2.51 on a scale of 1–5. Participants were given the choice of carrying out the experiment in German or English; 27 chose to use German, while 14 chose English.

6.2 Scenario

The study took place in a lab, with lighting and background noise controlled as far as possible. In each trial, the participant approached the bartender together with a confederate, with both customers seeking to engage with the bartender and order a drink (as in Fig. 1). Each participant was given a list of the possible drinks that could be ordered (Coke or lemonade), but was not given any further instructions. The robot was static until approached by a customer, and the confederate did not attempt to speak at the same time as the participant. Each participant carried out two interactions, with the order and selection of classifiers counter-balanced across participants.

6.3 Dependent Measures

To evaluate the performance of the classifiers, we used the system logs to compute a number of objective measures which specifically address the interactive performance of the two engagement classifiers. Note that the ground-truth data about the participants' actual behaviour is not available; however, based on the scenario (Fig. 1), it is reasonably safe to assume that the majority of customers were seeking to engage with the bartender as soon as they appeared in the scene, and that the participants behaved similarly in the two classifier conditions. We collected the following objective measures:

- *Detection Rate* How many of the customers detected in the scene were classified as seeking to engage. Under the above assumptions, this measure assesses the accuracy of the two classifiers.
- *Initial Detection Time* The average delay between a customer's initial appearance in the visual scene (i.e., the point at which the vision system first noticed them)

and the time that they were considered to be seeking engagement. Again, under the assumption that all participants behaved similarly, this measure assesses the relative responsiveness of the two engagement classifiers.
- *System Response Time* The average delay between a customer's initial appearance in the visual scene and the time that the system generated a response to that customer. Since the system would only respond to customers that were detected as seeking engagement, this is a secondary measure of classifier responsiveness, but one that is more likely to have been noticed by the participants.
- *Drink Serving Time* The average delay between a customer's initial appearance in the visual scene and the time that the system successfully served them a drink. Since serving a drink ultimately depends on successful engagement between the customer and the bartender, this is an even more indirect measure of responsiveness.
- *Number of Engagement Changes* The average number of times that the classifier changed its estimate of a user's engagement-seeking state over the course of an entire experiment run. In the experimental scenario, only the initial detection affected the system behaviour: as soon as a customer was determined to be seeking engagement, the system would engage with them and the interaction would continue. However, the engagement classifier remained active throughout a trial, so this measure tracks the performance over time. Although the actual behaviour of the experimental participants is not known, we assume that it was similar across the two groups, so any difference on this measure indicates a difference between the classifiers.

The participants also completed a subjective usability questionnaire following the experiment, including questions about perceived success, ease and naturalness of the interaction, and overall satisfaction. In general, the participants gave the system reasonably high scores on perceived success, interaction ease, and overall quality, with somewhat lower scores for naturalness. However, the choice of engagement classifier made no significant difference to any of the responses to this questionnaire, so we do not discuss those results further here—see Foster et al. [20] for more details.

6.4 Results

A total of 81 interactions were recorded in this study. However, due to technical issues with the system, only 58 interactions could be analysed, involving data from 37 of the 41 subjects: 26 interactions using the rule-based classifier, and 32 using the trained IB1 classifier. All results below are based on those 58 interactions.

Table 5 summarises the objective results, divided by the classifier type. Overall, the detection rate was very high, with 98% of all customers determined to be seeking engagement, generally within 4–5 s (and, in many cases, in under 1 s). The robot acknowledged a customer on average about 6–7 s after they first became visible, and a customer received a drink about a minute after their initial appearance—note that this last number includes the full time for the spoken interaction, as well as the 20 s normally taken by the robot arm to physically grasp and hand over the drink. Over the course of an entire interaction, a customer's estimated engagement changed an average of 15 times.

Table 5 Objective results (significant difference highlighted)

Measure	Rule (SD)	Trained (SD)
Detection rate	0.98 (0.10)	0.98 (0.09)
Time to first detection (s)	5.4 (7.9)	4.0 (9.7)
Time to system response (s)	7.0 (7.9)	6.4 (10.4)
Time to drink served (s)	62.2 (22.2)	53.7 (14.0)
Num. engagement changes	12.0 (10.2)	17.6 (7.6)

Each study participant took part in two interactions; however, as mentioned above, due to technical issues we could not analyse the full paired data. Instead, we analysed the data using a linear mixed model [4,64], treating the participant identifier as a random factor, with the classification strategy and all demographic features included as fixed factors. This analysis found that the effect of the classification strategy on the number of changes in estimated engagement was significant at the $p < 0.05$ level; however, while the numbers in Table 5 suggest that the trained classifier was somewhat more responsive, none of those differences were found to be statistically significant.

Several demographic factors also affected the objective results: the participants who carried out the experiment in German took significantly longer to receive their drinks than did those who interacted in English (48.1 vs. 62.0 s; $p < 0.05$), while the classifiers changed their estimate of the female participants' engagement state significantly more often over the course of an interaction (21.1 vs. 13.3 times; also $p < 0.05$).

6.5 Discussion

The objective results of this study indicate that the system was generally successful both at detecting customers who wanted to engage with it and at serving their drinks: despite the minimal instructions given to the participants, the objective success rate was very high. The choice between the two classification strategies had one main objective effect: the trained classifier changed its estimate of a customer's engagement state more frequently than did the rule-based classifier; in other words, the rule-based classifier was more stable over

the course of an interaction than the trained classifier. While the data in Table 5 suggests that the trained classifier may have been more responsive than the rule-based classifier (i.e., with a faster response time), no significant difference was found in these results.

The demographics had several effects on the results. First, the participants who used German took significantly longer to receive their drink, and also gave lower overall ratings to the system. We suspect that this was likely due to the decreased performance of the Kinect German language model, which was added to the Kinect Speech API much later than the English recognition. The system only responds to speech utterances with a confidence above a threshold—and on average, nearly twice as many attempted user turns were discarded due to low confidence for the German participants (4.1 per interaction) as for the English participants (2.2). Also, both classifiers' estimate of customer engagement changed more often over the course of a trial for the female participants than for the male participants: we hypothesise that this may be due to the vision system having been trained primarily on images of male customers.

Note that all of the dependent measures in this study are based only on the data from the log files, along with some underlying assumptions about user behaviour based on the scenario given to the participants (Fig. 1): namely, we assume that all customers were seeking to engage with the bartender from the moment that they appeared, and that the behaviour of the participants in the two conditions did not differ over the course of an interaction. The difference in classifier stability between male and female participants suggests that this assumption may not hold in practice; however, to assess the true performance of the classifiers, we require ground-truth data as to the actual engagement-seeking behaviour of the customers in the scene. Such ground-truth information would also allow us to analyse the impact of the demographic factors more directly.

7 Experiment 4: Evaluation with Novel Test Data

In the user evaluation summarised above, the ground truth about the customers' actual engagement-seeking behaviour was not available. This makes the results of the user study difficult to interpret, as it is impossible to know which of the classifiers actually estimated customer engagement more accurately in practice. Note that, due to the study design (in which all subjects were instructed to engage, and which was carried out in parallel with the end-to-end system evaluation described by Keizer et al. [36]), even if the recordings were annotated, there would be very few true negative examples in any case.

Instead, we therefore carried out a new evaluation of the engagement classifiers, making use of a specially-recorded test corpus addressing the weaknesses of the previous study: namely, the engagement-seeking behaviour of all customers is fully annotated, and the data includes a much more balanced set of positive and negative instances.

(a)

(b)

Fig. 5 Sample images from the test data. **a** Customer not seeking engagement. **b** Customer seeking engagement

The test data is based on six videos, each showing a single customer in front of the bar, as in the sample images in Fig. 5. Two different customers were recorded: one who was involved in the human–robot interactions making up the original training corpus, and one who was not. The customers were instructed to move around in front of the bartender; for half of the videos, they were instructed to engage with the bartender, while for the others, they were told to move around but not to engage; the details of how to behave were left up to the subjects.

After the recordings were made, the ELAN annotation tool [66] was used to annotate the videos, using the same labels as the original training data: the customer's engagement state was labelled as either *NotSeekingEngagement* (Fig. 5a) or *SeekingEngagement* (Fig. 5b). The video annotations were synchronised with the frame-by-frame information produced by the JAMES vision system, and a corpus instance was then created from the relevant data in each vision frame, using the annotation for the relevant time stamp as the gold-standard label. In total, the test corpus consisted of 361 instances: 233 labelled as *NotSeekingEngagement*, and 128 labelled as *SeekingEngagement*.

Table 6 Classifier performance on the test set, sorted by F score

Classifier	Accuracy	AUC	Precision	Recall	F
Rule	0.681	na	0.694	0.681	0.687
J48	0.648	0.583	0.661	0.648	0.653
CVR	0.598	0.576	0.612	0.598	0.604
NaiveBayes	0.571	0.528	0.638	0.571	0.578
LibSVM	0.645	0.500	0.417	0.645	0.506
ZeroR	0.645	0.500	0.417	0.645	0.506
JRip	0.421	0.350	0.557	0.421	0.432
Logistic	0.438	0.329	0.390	0.438	0.411
IB1	0.349	0.341	0.388	0.349	0.363

We then trained each classifier from Table 2 on the full training corpus from the previous study, and used each trained classifier to predict labels for every instance in the test data. The results of this test are shown in Table 6, sorted by weighted average F score. As shown by the groupings in the table, the results fell into three broad categories: at the top, the hand-coded rule and the J28, CVR, and NaiveBayes classifiers all had F scores well above the baseline ZeroR classifier, which always chooses the highest-frequency label (*NotSeekingEngagement*); the LibSVM classifier exactly reproduced the baseline ZeroR behaviour; while the JRip, Logistic, and IB1 classifiers all did worse than this baseline.

These results contrast strongly with the cross-validation results from Table 3. Firstly, the overall numbers are much lower: while the top performing classifiers from the previous study had scores well above 0.9 on all measures, the top results in this study were in the range of 0.6–0.7. Also, the relative ordering of the classifiers is very different: while the IB1 (instance-based) and JRip (rule learner) classifiers did well on cross-validation, they were both among the lowest-performing classifiers on the test data; this suggests that these classification strategies may have ended up over-fitting to the training data and did not generalise well. On the other hand, the NaiveBayes classifier and the hand-coded rule—which were both near the bottom on the cross-validation study—both scored at or near the top on the test data. Other classifiers such as J48 (decision trees) and CVR (classification via regression) did well in both studies; for this binary classification task, it is not surprising that these classifiers—which are particularly suited to binary classifications—showed generally good performance.

To better understand the performance of the classifiers, we inspected the classifier output on each of the test-data videos. Figure 6 shows the gold-standard (reference) annotation for three of the test videos, along with the labels produced by each classifier on those same videos. The light yellow regions correspond to the frames labelled with the *NotSeekingEngagement* class, while the dark blue regions correspond to the *SeekingEngagement* class. The figure clearly suggests differences among the classifiers: for example, the hand-coded rule selected *SeekingEngagement* very rarely; on the other hand, the lowest-performing classifiers (JRip, Logistic, IB1) selected this state frequently, even in cases where the customer never actually sought to engage (e.g., Video 3).

Note also that even the best-performing classifiers changed their engagement estimate much more frequently than the gold standard. Table 7 shows the mean number of engagement switches per test video produced by each classifier; with the exception of the two classifiers which always select *NotSeekingEngagement*, all of the numbers are well above the reference value of 2.0. Recall that in the online user study in Experiment 3, stability was also an issue: the hand-coded rule changed its estimate an average of 12.0 times per interaction, while the value for the IB1 classifier was 17.6.

8 Experiment 5: Adding Temporal Context with Conditional Random Fields

Although we used an HMM in the stand-alone study in Experiment 1—which implicitly incorporates temporal context in its processing—for all of the subsequent engagement studies, the input to the classifier consisted only of the sensor data at a given instant, without taking into account any of the temporal context provided by the interaction. However, real customers switch their engagement-seeking state relatively infrequently, so—as noted at the end of the preceding section—classifying each input frame independently tends to overestimate the number of engagement changes.

If an engagement classifier—even one with high overall accuracy—changes its estimate too frequently, the job of the system's interaction manager is made more difficult, in that responding to every change in estimated state is likely to produce undesirable behaviour. In an alternative, unsupervised, POMDP-based approach to interaction management, this issue is addressed by making the POMDP "sticky"; that is, biasing it towards self-transitions [62]. As an initial effort to address this issue in the current context, we experimented with various methods of incorporating information from previous frames into the state used to train the supervised classifiers; however these modifications were not found to improve either the stability or the performance of the classifiers (see [17] for details of these experiments).

Instead, we address this issue by turning to a completely different classification model: Conditional Random Fields (CRFs) [39,59], which are probabilistic graphical models particularly suitable for segmenting and labelling sequence data such as the user-engagement data considered in this paper. In particular, for these experiments, we have used the freely-available CRF implementation CRFSuite [50]. Just as we did in the previous studies with Weka, we used CRFSuite

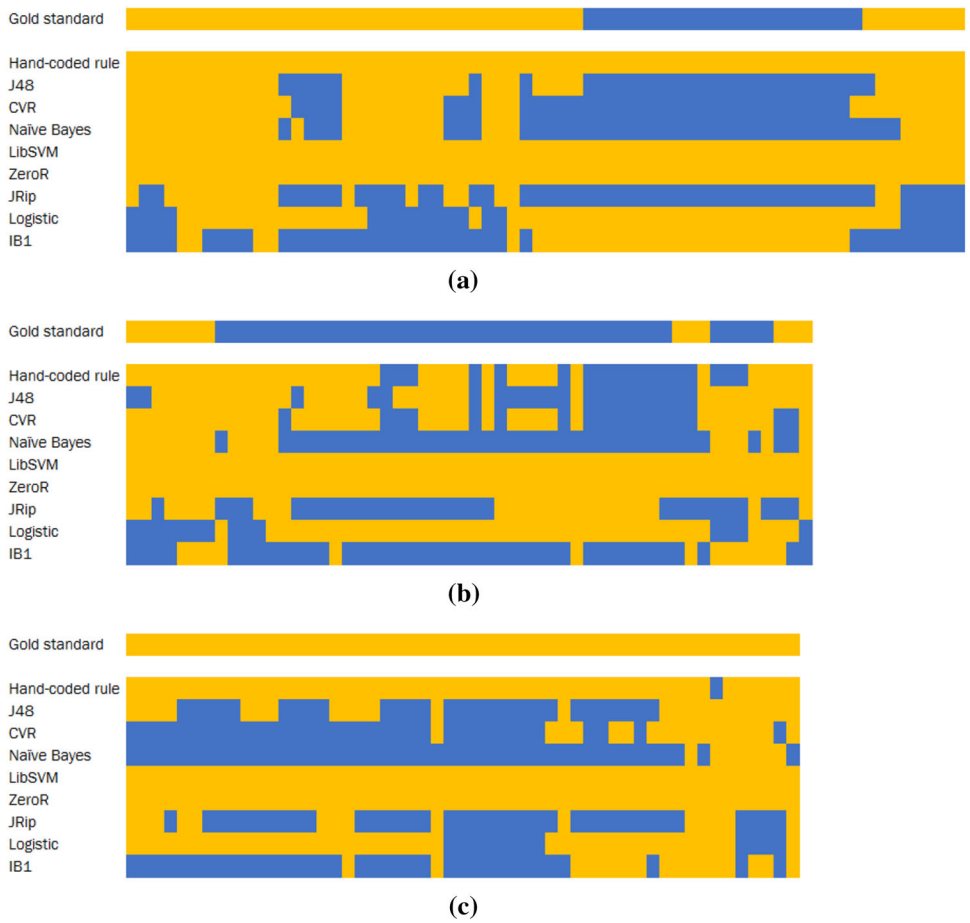

Fig. 6 Reference annotations and classifier predictions for three sample videos (*yellow* indicates NotSeekingEngagement, *blue* indicates SeekingEngagement). **a**)Video 1. **b** Video 2. **c** Video 3. (Color figure online)

Table 7 Mean engagement changes per classifier

Rule	J48	CVR	NaiveBayes	LibSVM	ZeroR	JRip	Logistic	IB1	Gold
4.7	10.5	8.8	5.8	0.0	0.0	11.3	5.3	9.3	2.0

in its default configuration: a first-order Markov CRF, trained through gradient descent using the L-BFGS method [67]. To make the data suitable for use by CRFSuite (which does not deal with continuous attributes), we first rounded all locations in the training and test data to the nearest 50mm, and all body-orientation values to the nearest degree. Rounding parameters were chosen to provide good discrete approximations of the continuous data.

To test the performance of the CRF model for the current engagement classification task, we carried out the same studies as on the supervised Weka classifiers in Experiments 2 and 4: 5-fold cross-validation against the training corpus, and evaluation against the separately recorded test data. The results of the cross-validation study are presented in Table 8; the results from the IB1, J48 and ZeroR classifiers and the hand-coded rule are repeated from Table 3 for context. Note that a paired t test found a significant difference at the $p < 0.01$ level between the accuracy scores of all classifiers in this table. Clearly, the cross-validation performance of the CRF is much lower than that of the previous classifiers, including the hand-coded rule; but as noted earlier, this measure itself does not necessarily reflect the practical utility of a classifier for the current task.

Table 8 Cross-validation results for CRF

Classifier	Accuracy	Precision	Recall	F
IB1	0.954	0.954	0.954	0.954
J48	0.928	0.925	0.928	0.928
ZeroR	0.780	0.609	0.780	0.684
Hand-coded rule	0.655	0.635	0.654	0.644
CRF	0.589	0.606	0.627	0.503

We then developed a CRF model based on the full training corpus and tested its performance on the test data; the results of this study are presented in Table 9, again with the results for the IB1, ZeroR and J48 classifiers and the hand-coded rule repeated for context. Here, the advantages of using a

Table 9 CRF performance on the test set

Classifier	Accuracy	Precision	Recall	F	Changes
Hand-coded rule	0.681	0.694	0.681	0.687	4.7
J48	0.648	0.661	0.648	0.653	10.5
CRF	0.615	0.614	0.624	0.606	1.0
ZeroR	0.421	0.557	0.421	0.432	0.0
IB1	0.349	0.388	0.349	0.363	9.3

CRF rather than a frame-by-frame classifier are becoming clearer: the CRF accuracy, precision, recall, and F score on this test set are all comparable to those of the hand-coded rule and the best-performing trained classifiers such as J48.

Finally, we revisit the main motivation for exploring a temporal classifier such as CRF in the first place: does using this sort of sequence model improve the overall stability of the classifier? Based on the performance on the test data (included in the final column of Table 9), the answer is clearly yes: in contrast to the previous classifier, the CRF classifier changed its estimate of the user's engagement state an average of 1.0 times per video across the test set—recall that the number from the gold-standard data was 2.0. The CRF output on the same three gold-standard (reference) videos is shown in Fig. 7. While the predictions are obviously not perfect—especially on Video 3—the overall pattern is closer to realistic, and is much more stable than that of any of the previous classifiers.

9 Summary, Conclusions, and Future Work

In the context of real-world human–robot joint action, a crucial task is to understand the social states of every person in the dynamic, changing scene. We have discussed the role of user engagement detection in the context of the JAMES robot bartender, and have shown how understanding the intended engagement of the customers is vital to supporting socially appropriate joint action in this bartender context.

We have then summarised our efforts in engagement detection in the context of this particular social HRI scenario. We began with a proof-of-concept study using HMMs to estimate user state based on a small corpus of specially-recorded training data (Experiment 1). In the light of the subsequent experiments, we can draw two main conclusions from this study. First, the visually recognized attributes available in this human–robot interaction scenario allow, in principle, classification into a larger set of user states. Of course, larger numbers of states would require more high-quality training data with perfectly recognized head poses, which is difficult to collect with uninformed customers that are not familiar with the limitations of the depth camera. Second, the HMM experiment confirms that the attribute selection of head pose and body posture features is necessary to classify user engagement in the bartender scenario, independently from related human–human studies [30,45,46] and the newer classifiers defined in Sect. 5.

Next, we have described two approaches to the task of online estimation of customers' intended engagement: the first version used a hand-coded rule based on findings from annotated human behaviour in real bars, while for the second version, we trained a range of supervised-learning classifiers using a multimodal corpus derived from user interactions with the initial system. In a cross-validation study using real sensor data (Experiment 2), nearly all of the trained classifiers significantly outperformed the hand-coded rule. The best-performing classifier in terms of accuracy was the instance-based IB1 classifier, which had an overall accuracy of 0.954 in frame-based cross-validation. When we carried out feature selection, it was found that the most informative features were the 3D position of the customer's head, along with some of the coordinates of their hands; body orientation—which was one of the two features used by the rule-based classifier—was actually not informative based on the corpus data, which we hypothesise was mainly due to the noisiness of this signal in the vision data used for training.

In an online user study (Experiment 3) comparing the rule-based classifier with the top-scoring IB1 classifier in the context of the full robot bartender system, we found one main difference between the two classifiers: namely, the trained classifier changed its estimate of the customers' engagement state significantly more often over the course of an interaction than did the rule-based classifier, suggesting that the former

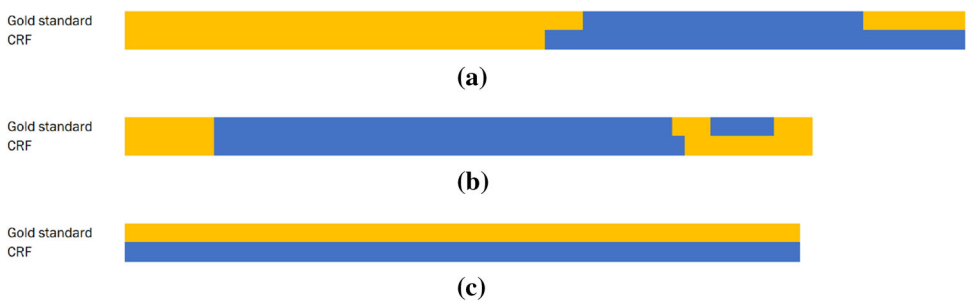

Fig. 7 Reference annotations and CRF predictions for three sample videos (*yellow* indicates NotSeekingEngagement, *blue* indicates SeekingEngagement). **a** Video 1. **b** Video 2. **c** Video 3. (Color figure online)

is less stable in practice. However, due to the details of the user experiment, these results have some limitations: in particular, the gold-standard engagement data was not available, and in any case the scenario would have led to very few true negative testing instances.

To address these limitations, we then carried out a targeted evaluation (Experiment 4) of the classifiers using a corpus of separately recorded, fully annotated, more balanced test data, and found that the relative performance was different. In the cross-validation study, the instance-based IB1 classifier had the highest performance and the hand-coded rule the lowest. On this study, we found instead that the J48 decision-tree classifier gave the best estimate of the users' engagement state, while the hand-coded rule actually had the overall best performance. We suspect that this result may also have been influenced by the noisy body orientations in the training data, particlarly when contrasted with higher-quality body orientation detection in the test data.

In all cases, and across all of Experiments 2–4, even the top-performing classifiers changed their estimate of the customers' engagement state much more frequently than the gold standard, likely because they all operate by classifying individual sensor data frames. To address this issue, we used the same data to train a classification model based on Conditional Random Fields, which are explicitly designed for sequence labelling problems of this type. The cross-validation results for the CRF were not as high as those for the previous frame-level classifiers; however, the overall stability of the classifier was much better, indicating that this sort of sequence model is a fruitful future direction for this classification task.

In summary, the main conclusion that we can draw from these studies is that, while data-driven methods can be useful for this engagement classification task, care must be taken in several areas. First of all, we have confirmed the message from Bohus and Horvitz [7] that online, run-time evaluation is crucial for evaluating any classifier for this task: the results from offline, frame-by-frame evaluation may not be indicative of online performance. Also, we have found that using a CRF, which explicitly incorporates the temporal sequence information, shows comparable frame-level performance to the frame-level classifiers but also greatly improves the overall stability of the classification. Even though the performance of all classifiers was likely affected by the noisy body orientation information from the training data, the stability difference with the CRF was so dramatic that it still seems that this is a better classification strategy.

Perhaps most importantly, we must also point out that the performance of the hand-coded, rule-based classifier—which used an extremely simple rule derived from the observation of human performance—was competitive with that of the highest-scoring trained classifiers in all of the experiments. While this may not be the case for every audiovisual processing task, this result does remind researchers to consider such simpler, easier to implement models, particularly if training data may be missing or of potentially uncertain quality.

Regarding future work, we first note that in all of the classification studies, we have made a deliberate choice to treat all of the classifiers as "black boxes", in all cases using the default parameter settings provided by the tools (Weka and CRFSuite, respectively). This is a similar approach to that taken, for example, by Koller and Petrick [38], who compared the off-the-shelf performance of a number of AI planners when applied to tasks derived from natural language generation. However, it is certain that the relative and absolute performance could be significantly affected by appropriate parameter tuning [41], and in future we will explore the space of parameters more fully.

Another direction for future work is to explore methods for making improved use of the classifier output in the context of end-to-end interactions with the robot bartender. In particular, where the classifier provides not only a class, but also an estimated confidence in that class, that additional information can be incorporated into the state and used in the interaction. Indeed, the state representation used by the final JAMES bartender system retains and exploits the uncertainty coming from the underlying input sensors to improve interactive performance [18,21]. The use of classifiers such as J48 and CRFs—which provide such confidence estimates—could also prove useful in this context.

The anonymised, annotated training and test corpora from Experiments 2–5 are available for download from http://downloads.maryellenfoster.uk/, and we encourage other researchers to test their models on this data.

Acknowledgements The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2014) under Grant Agreement No. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems (james-project.eu). Thanks to Ingmar Kessler and Sören Jentzsch for helping run the experiments, and to all of our JAMES colleagues for fruitful collaboration throughout the project.

References

1. Aggarwal JK, Xia L (2014) Human activity recognition from 3d data: a review. Pattern Recognit Lett 48:70–80
2. Aha D, Kibler D (1991) Instance-based learning algorithms. Mach Learn 6:37–66
3. Andrist S, Pejsa T, Mutlu B, Gleicher M (2014) Designing effective gaze mechanisms for virtual agents. In: Proceedings of

ACM/SigCHI conference on human factors in computing (CHI). Canada, Toronto, pp 705–714
4. Baayen R, Davidson D, Bates D (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59(4):390–412. doi:10.1016/j.jml.2007.12.005
5. Baltzakis H, Pateraki M, Trahanias P (2012) Visual tracking of hands, faces and facial features of multiple persons. Mach Vis Appl 23(6):1141–1157. doi:10.1007/s00138-012-0409-5
6. Bohus D, Horvitz E (2009a) Dialog in the open world: platform and applications. In: Proceedings of ICMI-MLMI 2009, Cambridge, MA, pp 31–38, doi:10.1145/1647314.1647323
7. Bohus D, Horvitz E (2009) Learning to predict engagement with a spoken dialog system in open-world settings. Proc SIGDIAL 2009:244–252
8. Brand M, Oliver N, Pentland A (1997) Coupled hidden Markov models for complex action recognition. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, 1997. IEEE, pp 994–999
9. Breazeal C (2005) Socially intelligent robots. Interactions 12(2):19–22. doi:10.1145/1052438.1052455
10. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan P (2012) Detecting engagement in HRI: an exploration of social and task-based context. In: Proceedings of SocialCom'12, pp 421–428. doi:10.1109/SocialCom-PASSAT.2012.51
11. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27:1–27:27. doi:10.1145/1961189.1961199
12. Chen LF, Liu ZT, Wu M, Ding M, Dong FY, Hirota K (2015) Emotion–age–gender–nationality based intention understanding in human–robot interaction using two-layer fuzzy support vector regression. Int J Soc Robot 7(5):709–729
13. Cohen WW (1995) Fast effective rule induction. In: Twelfth international conference on machine learning, Morgan Kaufmann, pp 115–123
14. Dautenhahn K (2007) Socially intelligent robots: dimensions of human–robot interaction. Philos Trans R Soc B Biol Sci 362(1480):679–704. doi:10.1098/rstb.2006.2004
15. de Kok IA (2013) Listening heads. PhD thesis, Enschede. http://doc.utwente.nl/87077/
16. Figueroa-Angulo JI, Savage J, Bribiesca E, Escalante B, Sucar LE (2015) Compound hidden Markov model for activity labelling. Int J Intell Sci 5(05):177
17. Foster ME (2014) Validating attention classifiers for multi-party human–robot interaction. In: Proceedings of the HRI 2014 workshop on attention models in robotics, Bielefeld, Germany
18. Foster ME, Petrick RPA (2014) Planning for social interaction with sensor uncertainty. In: Proceedings of the ICAPS 2014 scheduling and planning applications workshop (SPARK). Portsmouth, NH, pp 19–20
19. Foster ME, Gaschler A, Giuliani M, Isard A, Pateraki M, Petrick RPA (2012) Two people walk into a bar: dynamic multi-party social interaction with a robot agent. In: Proceedings of ICMI 2012
20. Foster ME, Gaschler A, Giuliani M (2013) How can I help you? Comparing engagement classification strategies for a robot bartender. In: Proceedings of the 15th international conference on multimodal interaction (ICMI 2013), Sydney, Australia. doi:10.1145/2522848.2522879
21. Foster ME, Keizer S, Lemon O (2014) Action selection under uncertainty for a socially aware robot bartender. In: Proceedings of HRI. doi:10.1145/2559636.2559805
22. Frank E, Wang Y, Inglis S, Holmes G, Witten I (1998) Using model trees for classification. Mach Learn 32(1):63–76
23. Gaschler A, Huth K, Giuliani M, Kessler I, de Ruiter J, Knoll A (2012a) Modelling state of interaction from head poses for social human–robot interaction. In: Proceedings of the gaze in human–robot interaction workshop held at the 7th ACM/IEEE international conference on human–robot interaction (HRI 2012), Boston, MA
24. Gaschler A, Jentzsch S, Giuliani M, Huth K, de Ruiter J, Knoll A (2012b) social behavior recognition using body posture and head pose for human–robot interaction. In: IEEE/RSJ international conference on intelligent robots and systems (IROS). doi:10.1109/IROS.2012.6385460
25. Giuliani M, Petrick RPA, Foster ME, Gaschler A, Isard A, Pateraki M, Sigalas M (2013) Comparing task-based and socially intelligent behaviour in a robot bartender. In: Proceedings of the 15th international conference on multimodal interfaces (ICMI 2013), Sydney, Australia
26. Hall M, Holmes G (2003) Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 15(6):1437–1447. doi:10.1109/TKDE.2003.1245283
27. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor Newsl 11(1):10–18. doi:10.1145/1656274.1656278
28. Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the seventeenth international conference on machine learning (ICML 2000), pp 359–366
29. Hernandez J, Riobo I, Rozga A, Abowd GD, Picard RW (2014) Using electrodermal activity to recognize ease of engagement in children during social interactions. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, ACM, pp 307–317
30. Huth K, Loth S, De Ruiter J (2012) Insights from the bar: a model of interaction. In: Proceedings of formal and computational approaches to multimodal communication
31. International Federation of Robotics (2015) Service robot statistics. http://www.ifr.org/service-robots/statistics/
32. Iqbal T, Gonzales MJ, Riek LD (2015) Joint action perception to enable fluent human–robot teamwork. In: 24th IEEE international symposium on robot and human interactive communication (RO-MAN), 2015. IEEE, pp 400–406
33. John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Eleventh conference on uncertainty in artificial intelligence, San Mateo, pp 338–345
34. Johnson DO, Cuijpers RH, Juola JF, Torta E, Simonov M, Frisiello A, Bazzani M, Yan W, Weber C, Wermter S et al (2014) Socially assistive robots: a comprehensive approach to extending independent living. Int J Soc Robot 6(2):195–211
35. Ke SR, Thuc HLU, Lee YJ, Hwang JN, Yoo JH, Choi KH (2013) A review on video-based human activity recognition. Computers 2(2):88–131
36. Keizer S, Foster ME, Lemon O, Gaschler A, Giuliani M (2013) Training and evaluation of an MDP model for social multi-user human–robot interaction. In: Proceedings of the 14th annual SIGdial meeting on discourse and dialogue
37. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1):273–324
38. Koller A, Petrick RPA (2011) Experiences with planning for natural language generation. Comput Intell 27(1):23–40. doi:10.1111/j.1467-8640.2010.00370.x
39. Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '01, pp 282–289. http://dl.acm.org/citation.cfm?id=645530.655813
40. Lara OD, Labrador MA (2013) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor 15(3):1192–1209
41. Lavesson N, Davidsson P (2006) Quantifying the impact of learning algorithm parameter tuning. In: Proceedings of AAAI

42. le Cessie S, van Houwelingen J (1992) Ridge estimators in logistic regression. Appl Stat 41(1):191–201
43. Leite I, McCoy M, Ullman D, Salomons N, Scassellati B (2015) Comparing models of disengagement in individual and group interactions. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction, ACM, pp 99–105
44. Li L, Xu Q, Tan YK (2012) Attention-based addressee selection for service and social robots to interact with multiple persons. In: Proceedings of the workshop at SIGGRAPH Asia, WASA '12, pp 131–136. doi:10.1145/2425296.2425319
45. Loth S, Huth K, De Ruiter JP (2013) Automatic detection of service initiation signals used in bars. Front Psychol. doi:10.3389/fpsyg.2013.00557
46. Loth S, Jettka K, Giuliani M, De Ruiter JP (2015) Ghost-in-the-machine reveals human social signals for human–robot interaction. Front Psychol. doi:10.3389/fpsyg.2015.01641
47. MacHardy Z, Syharath K, Dewan P (2012) Engagement analysis through computer vision. In: Proceedings of CollaborateCom, pp 535–539
48. McColl D, Nejat G (2012) Affect detection from body language during social HRI. In: Proceedings of 2012 IEEE RO-MAN, pp 1013–1018. doi:10.1109/ROMAN.2012.6343882
49. Mihoub A, Bailly G, Wolf C (2013) Social behavior modeling based on incremental discrete hidden markov models. In: Salah AA, Hung H, Aran O, Gunes H (eds) Human behavior understanding. HBU 2013. Lecture Notes in Computer Science, vol 8212. Springer, Cham
50. Okazaki N (2007) Crfsuite: a fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/
51. Otsuka K (2011) Conversation scene analysis. IEEE Signal Process Mag 28(4):127–131. doi:10.1109/MSP.2011.941100
52. Otsuka K, Takemae Y, Yamato J (2005) A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In: Proceedings of the 7th international conference on multimodal interfaces, ACM, pp 191–198
53. Otsuka K, Yamato J, Takemae Y, Murase H (2006) Conversation scene analysis with dynamic Bayesian network basedon visual head tracking. In: IEEE international conference on multimedia and expo, 2006, pp 949–952
54. Pateraki M, Sigalas M, Chliveros G, Trahanias P (2013) Visual human–robot communication in social settings. In: Proceedings of ICRA workshop on semantics, identification and control of robot–human–environment interaction
55. Petrick RPA, Foster ME (2013) Planning for social interaction in a robot bartender domain. In: Proceedings of the ICAPS 2013 special track on novel applications, Rome, Italy
56. Petrick RPA, Foster ME, Isard A (2012) Social state recognition and knowledge-level planning for human–robot interaction in a bartender domain. In: AAAI 2012 workshop on grounding language for physical systems, Toronto, ON, Canada
57. Quinlan R (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo
58. Sebanz N, Bekkering H, Knoblich G (2006) Joint action: bodies and minds moving together. Trends Cogn Sci 10(2):70–76
59. Sutton C, McCallum A (2006) An introduction to conditional random fields for relational learning. In: Geloor L, Tasker B (eds) Introduction to statistical relational learning. MIT Press, pp 93–128
60. Thórisson KR (2002) Natural turn-taking needs no manual: computational theory and model, from perception to action. In: Granström B, House D, Karlsson I (eds) Multimodality in language and speech systems. Springer, Netherlands, pp 173–207. doi:10.1007/978-94-017-2367-1_8
61. Torta E, Heumen J, Cuijpers R, Juola J (2012) How can a robot attract the attention of its human partner? A comparative study over different modalities for attracting attention. In: Ge S, Khatib O, Cabibihan JJ, Simmons R, Williams MA (eds) Social robotics, Lecture notes in computer science, vol 7621, Springer, Berlin, pp 288–297. doi:10.1007/978-3-642-34103-8_29
62. Wang Z, Lemon O (2012) A nonparametric Bayesian approach to learning multimodal interaction management. In: Proceedings of SLT. doi:10.1109/SLT.2012.6424162
63. Weka (n.d.) Weka primer. http://weka.wikispaces.com/Primer
64. West B, Welch KB, Galecki AT (2006) Linear mixed models: a practical guide using statistical software. CRC Press, Boca Raton
65. White M (2006) Efficient realization of coordinate structures in Combinatory Categorial Grammar. Res Lang Comput 4(1):39–75. doi:10.1007/s11168-006-9010-2
66. Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H (2006) ELAN: a professional framework for multimodality research. In: Proceedings of LREC 2006
67. Zhu C, Byrd RH, Lu P, Nocedal J (1997) Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans Math Softw (TOMS) 23(4):550–560

Mary Ellen Foster is a Lecturer in the School of Computing Science at the University of Glasgow. Her primary research interests are human–robot interaction, social robotics, and embodied conversational agents. She is a member of the Glasgow Social Robotics group and the Glasgow Interactive Systems Group (GIST), and an Associate Academic of the Institute of Neuroscience and Psychology. She is the coordinator of the MuMMER project, a European Horizon 2020 project in the area of socially aware human–robot interaction.

Andre Gaschler is a scientist with the fortiss An-Institut at Technische Universität München, Germany, working on the EU-funded project SMErobotics. He received a doctoral degree in computer science from Technische Universität München in 2016. His research interests include human–robot interaction, robot task and motion planning, and manipulation planning.

Manuel Giuliani is a Professor at the Bristol Robotics Laboratory, University of the West of England, Bristol. Before coming to Bristol, he led the Human–Robot Interaction group at the Center for Human–Computer Interaction, Department of Computer Sciences, University of Salzburg. He received a Master of Arts in computational linguistics from Ludwig-Maximilian-University Munich, a Master of Science in computer science from Technische Universität München, and a PhD in computer science from Technische Universität München. He worked on the European projects JAST (Joint Action Science and Technology), JAMES (Joint Action for Multimodal Embodied Social Systems), ReMeDi (Remote Medical Diagnostician) and the Austrian Christian-Doppler-Laboratory "Contextual Interfaces". His research interests include human–robot interaction, social robotics, natural language processing, multimodal fusion, multimodal output generation, and robot architectures.