Comparing Models of Disengagement in Individual and Group Interactions

Iolanda Leite*, Marissa McCoy†, Daniel Ullman*,
Nicole Salomons*, Brian Scassellati*
*Dept. of Computer Science, †Dept. of Psychology
Yale University, New Haven, CT, USA
{iolanda.leite, marissa.mccoy, daniel.ullman,
nicole.salomons, brian.scassellati}@yale.edu

ABSTRACT

Changes in type of interaction (e.g., individual vs. group interactions) can potentially impact data-driven models developed for social robots. In this paper, we provide a first investigation in the effects of changing group size in data-driven models for HRI, by analyzing how a model trained on data collected from participants interacting individually performs in test data collected from group interactions, and *vice-versa*. Another model combining data from both individual and group interactions is also investigated. We perform these experiments in the context of predicting disengagement behaviors in children interacting with two social robots. Our results show that a model trained with group data generalizes better to individual participants than the other way around. The mixed model seems a good compromise, but it does not achieve the performance levels of the models trained for a specific type of interaction.

Keywords

Child-robot interaction; disengagement; multimodal classification; multiparty settings.

1. INTRODUCTION

Human behavior is largely dependent on social context. The way we behave alone is different than how we behave in a group [22]. For this reason, most data-driven perceptual systems developed for social robots rely on data collected in the same type of interaction where most future interactions are likely to occur. For example, a robot bartender is able to predict engagement using data collected in multiparty settings [7], while a chess-playing robot relies on data from a single user at a time [4]. Will the robot bartender be able to respond appropriately to the lonely costumer at the end of the night? How would the chess-playing robot behave when placed in a science fair?

HRI '15, March 02 - 05 2015, Portland, OR, USA

http://dx.doi.org/10.1145/2696454.2696466

Robots will inevitably have to interact with users in contexts with different group sizes. This might occur in a variety of domains, such as museums, hospitals or classrooms. So far, little is known about how perception models perform when they are tested in a group size different than the one they were trained on. However, this feature is critical for some perception problems. That is, the way the robot should interpret a user glancing to the side is different if that user is alone or if the user is in a group.

We provide a first investigation of this issue by addressing the following research questions: Using the same set of features, how does group size affect the performance of a data-driven perception model? Moreover, for the same classification problem, can a model trained on data from one group size generalize well when tested in data from another context? In this paper, we answer these questions in the particular case of predicting disengagement in child-robot interaction. According to Pohl and Murray-Smith [12], engagement (or disengagement) with technology is dependent on inhibitor factors from three main categories: physical (system is not accessible), social (people change their behavior built on what they envision people around them are thinking about them) and mental (primarily related to distraction). Because the two last inhibitor factors are highly dependent on type of interaction, the automatic prediction of engagement/disengagement is a very interesting case study to our research questions.

We start our analysis by training and testing two different disengagement models in the context of an interactive narrative scenario with social robots. The models were built on two different datasets, one with data collected from participants interacting alone with the robots, and another with data from participants interacting in groups of three. Even in the model collected from group data, the goal is to predict disengagement of one participant at a time. The models use the exact same set of audio, visual and contextual features, and the group model does not encode features from the other participants around the robots. This was a deliberate choice, not only to allow a more fair comparison between the models, but also because in future applications we may not have access to information of all participants in the group, such as due to occlusions or limitations in the robot sensors.

To address the second research question, we report the results of testing each model using data from participants of the opposite dataset. Finally, we train a mixed model with data from both datasets to see how this model performs in individual and group data. Our results show that while a mixed model (trained with data instances from the two group sizes) is a good compromise, higher performances are achieved when the models are trained using only data from a similar group size as the target interaction. We discuss these results in terms of potential implications for future research in this area.

2. BACKGROUND

Engagement, defined as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" [17], has been studied from two main perspectives in HRI. One perspective is dedicated to understanding which features or social cues robots should be endowed with to increase participant's engagement with the robot. For example, Sidner and colleagues [17] showed that people report higher levels of engagement when interacting with a robot capable of face-tracking and performing gestures. More recently, researchers investigated the impact of social cues including voice and facial expressions [10], while others studied the impact of a robot side-kick on perceptions of user engagement [20].

The second perspective, and the perspective more relevant to the work presented in this paper, has to do with the automatic recognition of engagement (or disengagement) in users interacting with robots or virtual agents. Seminal work in this area has focused on predicting engagement intention, i.e., the problem of whether users around a robot (or a virtual agent) express a desire to start interacting with the system. In this domain, Michalowski et al. [9] presented a spatial model of social engagement grounded in proxemics theory. Bohus and Horvitz [2] proposed the first data-driven approach to predict engagement intentions using spatiotemporal and attention cues for a conversational virtual agent. A similar model, using mainly visual features, was developed and tested in a social robot [21].

Other authors have investigated the automatic prediction of engagement as a continuous signal. The main goal here is typically to predict engagement or, more importantly, disengagement behaviors in real-time, so that the robot or agent can employ repair mechanisms to keep users engaged in the interaction. To address this problem, several models have been proposed using a variety of features, including visual and task-related information [4], eye gaze [11], speech and gestures [15], body postures [16] and EEG data [19]. Some authors compared the performance of predicting engagement using different machine learning and rule-based models [7].

The presented work shows that data-driven methods are clearly the most common and successful approaches for automatically predicting engagement in HRI, both in settings with one and multiple users. Despite the significant advances in this area, to our knowledge no multiparty engagement model has been tested in data from individual participants, or the other way around. In this work, we aim to advance the state-of-the-art in this area by performing the first experiment of this kind.

3. METHODOLOGY

3.1 Case Study

The case study used in this work utilizes two MyKeepon Robots (see Figure 1) that play out interactive stories around emotional words (e.g., frustration, inclusion, cooperation). At specific points, users can influence the story by choosing from among a set of optional scenes presented on a tablet. In other words, users can tell one of the robots what to do, and then see the impact of the selected action on the course of the story.

The robots use pre-recorded adult utterances with modified pitch signal to make them more childlike. The robots can display several animations during the interaction, such as speaking, idling (while they are waiting for children's choices or listening to the other robot) or bouncing (moving up and down and side to side – used in specific moments of the stories). The robots are autonomous, but at this point their only perception of the world is from the story choices selected on the tablet.

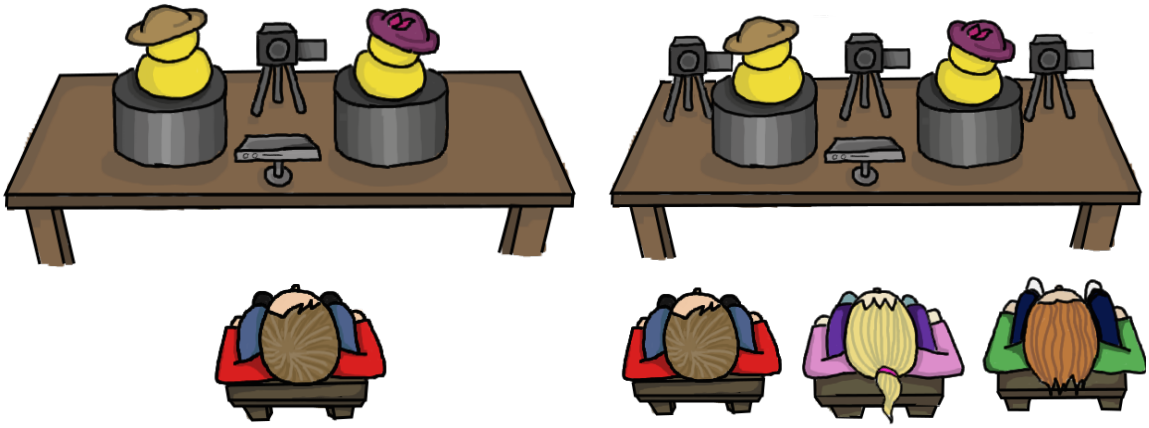
Our goal is to develop a classifier that allows the robots to accurately perceive when children are disengaged in the interaction, despite the number of children interacting at the same time. Upon detecting disengagement, the robots could employ repair strategies, such as displaying more active nonverbal behaviors to call attention to or change the story, in the attempt to re-engage users in the interaction.

For this analysis, the models will mostly rely on hand-annotated data. This was a deliberate choice because we wanted to distinguish the adequacy of the feature set and its effects on predicting disengagement in different contexts from the adequacy of particular feature detectors. Nevertheless, we plan to replace the hand-annotated features with autonomous modules, such that we can run a real-time implementation of the models in our robots.

3.2 Data Collection

Our data set consists of 40 children (22 female, 18 male), with ages between 6 and 8 years old ($M = 7.53$, $SD = 0.51$), interacting with the social robots in the interactive narrative scenario described in the previous section. Participants were recruited from an elementary school in the United States, where the data collection was conducted. The data collection took place in a small meeting room of the school. Participants were randomly assigned to a type of interaction condition: 19 were assigned to the single condition and 21 were assigned to the group condition (7 groups). The two conditions were balanced for gender.

One experimenter was present in the room for the entire session. The experimenter started by introducing the two robots and telling participants that the robots would play out a story, and then when the story stopped, they could decide what would happen next from the options that appeared on the tablet. In the group condition, participants were informed that they would have to choose the next story option together. No additional instructions were given. The story contained an introductory scene and three different options that participants could then freely explore. The interaction ended when participants explored all three of the story options. The average actual interaction time, from the



(a) single (b) groups

Figure 1: Sketch of the two interaction settings where data was collected.

moment when participants selected the first story option until the robots played all the possible scenes, was 4 minutes 36 seconds ($SD = 39$ seconds).

Three HD cameras were used to record the interaction in the group condition, while in the single condition one HD camera was used. Each camera recorded the upper body posture and face of one single participant. Log files containing the content of the robots' actions (speech and nonverbal behaviors), as well as the story choices made by the children, were generated automatically. The logs contained timestamps to allow future synchronization with the remaining data.

3.3 Feature Extraction

Our goal is to create a data-driven model that allows our robots to predict, in real-time, when each participant is disengaged in the interaction. To achieve this goal, we considered a set of features based on prior HRI research on automatic prediction of engagement and on research describing typical behaviors when people are engaged and disengaged [1, 14].

Using the collected videos and the ELAN annotation tool [3], one annotator (blind to our research questions) coded the start and end times of each participant's vocalizations, backchannel sounds, body posture (leaning forward/backward, arms on the table), gestures (smiles, mimicking robots, excitable bouncing and strong emotional reactions), concentration and boredom signs and off task behaviors. We ran an off-the-shelf face tracking algorithm[1] on the video recordings to automatically extract head orientation features – looking at the robots, looking up, looking down and rolling head. The contextual features – robots speaking, robots bouncing, and participant choosing an action – were extracted from the interaction logs. The final set of features considered to predict disengagement in this work are listed in Table 1.

Determining the optimal window size for analysis is still an open challenge in the processing of many social signals (like engagement), where it is hard to draw semantic or lexical boundaries [8]. Therefore, our approach was to use a small unit of analysis to simulate real-time decision-making in the disengagement predictions. From the hand-coded annotations, automatically generated face tracking analysis and interaction logs of each participant, we extracted a set of multimodal features into 500 msec time slices. The binary value of a feature for every time slice reflects what happened in the majority of the 500 msec interval. For example, voice activity will be set to true if a participant is speaking for more than 250 msec in that time slice.

[1]http://www.omron.com/r_d/coretech/vision/okao.html

3.4 Ground Truth Labelling

The ground truth labels for training and testing our data were based on human observations. Without having access to any of the extracted features, two independent coders (other than the one who coded children's behaviors) were given our working definition of engagement [17] and asked to rate participants' level of engagement during the interaction with the robots. Using the same video annotation tool, they coded the start and end times of disengagement, engagement and neutral episodes for all participants. The neutral category was used for moments where it was unclear whether participants were engaged or disengaged. In the group interactions, raters watched the interaction three times, once in the perspective of each participant. The videos of each participant in the group condition were trimmed to display the least possible information from the other children in the scene.

Both coders rated all the collected videos. As in the feature extraction process, the extracted ground truth labels for each rater were based on 500msec time slices reflecting the most predominant category in that segment. Inter-observer agreement between the two coders was 74% ($k = 0.41$, $p < .001$). The moderate agreement result was expected because perceived engagement can be a subjective observation. Nevertheless, to not undermine the comparison between models later on, we only consider the time slices where both raters agreed in the engagement category for future training and testing.

Table 1: Multimodal features considered for disengagement prediction.

Modality	Feature name	Description	Source
Audio	voice activity	Whether a participant is speaking	Hand annotated
	backchannel	Presence of backchannel sounds such as "uh-huh" and "hmm"	Hand annotated
Visual	look at robots	Looking at the robots or to the sides	Autom. extracted
	look up	Looking up (above the robots)	Autom. extracted
	look down	Looking down (not looking at the robots)	Autom. extracted
	rolling head	Rolling the head to the sides	Autom. extracted
	lean forward	Leaning forward	Hand annotated
	lean backward	Leaning backward	Hand annotated
	arms on table	Placing the arms on the table	Hand annotated
	smiling	Presence of smiles	Hand annotated
	head nods	Presence of head nods	Hand annotated
	mimicking robots	Micking the robots' movements	Hand annotated
	excitable boucing	Moving back and forth in the chair	Hand annotated
	emotional reaction	Strong emotional reactions (e.g. surprise)	Hand annotated
	boredom signs	Presence of boredom signs such as shrugs or fiddling	Hand annotated
	concentration signs	Presence of concentration signs such as fingers in the mouth	Hand annotated
	off task	Presence of off task behaviors (e.g., playing with sticker tag)	Hand annotated
Contextual	robots speaking	Whether any of the robots is speaking	Interaction logs
	robots bouncing	Whether any of the robots is displaying a bouncing animation	Interaction logs
	choosing	Two second window before a choice is made in the tablet	Interaction logs

(a) disengagement (b) engagement (c) disengagement (d) engagement

Figure 2: Snapshots of the collected datasets representing the disengagement different behaviors.

3.5 Two datasets: DS_I and DS_G

Two datasets were used in this analysis: DS_I, including the 500 msec segments from all participants who interacted alone with the robots, and DS_G, including segments for participants who interacted with the robots in small groups (see fig 2). Note that, in DS_G, the feature vectors only encode features related to the behavior of one participant, but all participants in each group are included in the dataset. Data from one participant from DS_I and from four participants from DS_G were excluded because they had no disengagement instances rated as such by the two coders. Table 2 provides a characterization of the two datasets. Each participant contributed an approximately similar number of instances to the data set.

Table 2: Characterization of the two datasets in terms of number of participants and number of 500msec instances for each label.

	Num. Participants	Disengaged	Engaged
DS_I	18	1283	5616
DS_G	17	853	4944

4. PREDICTING DISENGAGEMENT IN INDIVIDUAL AND SMALL GROUP INTERACTIONS

Our main goal is to investigate the effects of different group formations in the automatic prediction of disengagement in Human-Robot Interaction, rather than maximizing accuracy by trying different machine learning techniques or feature sets. As such, we focus our analysis in one classification technique and the same set of features in both cases. We decided to use Support Vector Machines (SVMs) as they have proven effective within similar classification problems in HRI [13, 18, 21].

4.1 Procedure

Using LibSVM Library [5], two SVM binary classifiers were trained, one using DS_I, which we will refer as M_I, and another using DS_G, referred from now on as M_G. We started by running the feature selection tool provided in this library (fselect), which performs feature ranking using F-scores [6]. The feature selection analysis was performed not only to find the optimal set of features, but also to rank and compare the features with the most discriminative power in each dataset.

The results of this latter analysis are reported in the next subsection. The fselect tool indicated that, in both datasets, the best classification accuracy was in the presence of 19 of the 20 extracted features (see Table 1). In both cases, head nods was the only feature that revealed no discriminative power for disengagement detection. This feature was therefore excluded from the analysis.

The two SVM models (type C-SVC) with a Radial Basis Function (RBF) kernel were trained with the 19 selected features, using different weights to account for the unbalanced number of disengagement and engagement instances in each data set. A utility tool also included in the LibSVM library (grid) was used to find the optimal parameters C and gamma ($C = 4$, $\gamma = 0.5$ for M_I and $C = 1$, $\gamma = 0.5$ for M_G).

The consistency of the two models was measured through leave-one-out cross-validation, using the data instances from one participant as the test set and training a model in the remaining participants of that data set. This process was repeated 18 times for M_I and 17 times for M_G, allowing data from each participant to serve once as test set in his/her respective data set.

4.2 Performance Evaluation

We averaged Accuracy, Area Under ROC Curve (AUC), True Positive Rate (TPR) and True Negative Rate (TPR) values from the cross-validation tests of each model. Note that, because our goal is to predict disengagement, TPR reflects the proportion of actual disengagement data points correctly classified as such, and TNR refers to the proportion of correctly predicted engagement instances. Because our datasets are unbalanced, AUC, TPR and TNR values are more informative than accuracy to understand how the models perform.

Cross-validation of M_I showed an average accuracy of 63%, with $AUC = 0.65$, $TPR = 0.68$ and $TNR = 0.62$. This performance was slightly better than M_G, which achieved 60% average accuracy, $AUC = 0.57$, $TPR = 0.53$ and $TNR = 0.61$. This result was not surprising, because when children are alone with the robots, interactions tend to be less chaotic, resulting in more accurate disengagement predictions.

Despite similar classification performances, the feature ranking analysis indicates that the two models are inherently different. Table 3 shows the top 10 most discriminative features in the two models. Although 7 of the top 10 features are the same, their rankings and weights are different in each model. For example, while in M_I the most discriminative feature for predicting disengagement is related to body posture (arms on the table), the most relevant feature for M_G is whether the participant is looking at the robots or not.

5. TESTING THE MODELS IN DIFFERENT DATA SETS

The models generated based on the datasets with different group size show slightly different performances and some differences on the features with most discriminative power. But are these models truly different? How do performance metrics change if we train a model with participants from the individual condition and test it with group data, or vice versa? Moreover, what happens if we combine data instances from both social settings (individual and group) in the same model?

Table 3: Top 10 most discriminative features in each model with normalized F-scores.

M_I		M_G	
Feature Name	**F-score**	**Feature Name**	**F-score**
arms on table	1.00	look at robots	1.00
look down	0.68	voice activity	0.65
look at robots	0.47	lean backward	0.27
lean forward	0.33	robots speaking	0.14
concent. signs	0.23	smiling	0.13
boredom signs	0.14	boredom signs	0.11
robots speaking	0.13	robots bouncing	0.08
choosing	0.10	backchannel	0.05
robots bouncing	0.10	off task	0.05
rolling head	0.06	choosing	0.03

In the previous section, we analyzed the ability of our models to predict disengagement in data from the same social setting (i.e., M_I was tested with data from other participants from the individual condition and M_G was tested with group data). We now investigate the behavior of the models when tested using data from the other data set, which offers a different social setting in the same scenario. To further investigate this issue, we report the performance metrics of a joint model trained with data from the two datasets.

5.1 Procedure

Using the SVM parameters obtained in the previous experiment for DS_I, we trained a model with the instances of the 18 participants from the individual condition (M_I) and tested this model in data from participants in the group condition, averaging the performances of every participant from DS_G. Similarly, a model using all 17 participants from DS_G was trained (MS_G) and tested in each participant from DS_I.

Finally, we investigated a model combining both datasets (M_A). This model was trained with the same subset of 19 features, using a C-SVS SVM with RBF Kernel parameters $C = 64$ and $\gamma = 0.125$. M_A was tested with participants from both data sets using a leave-one-participant-out cross-validation approach (the data instances from the participant used as test set were the only ones left out in the trained model in every cross-validation cycle).

5.2 Performance Evaluation

We used the same performance metrics as in the previous experiment: Accuracy, Area Under ROC Curve (AUC), True Positive Rate (TPR) and True Negative Rate (TPR), averaged across repeated validation tests.

When testing how M_I performs with data from DS_G, 17 test cycles (each cycle contains the data points from one participant in the group condition) show an average accuracy of 75%, $AUC = 0.59$, $TPR = 0.36$ and $TNR = 0.82$. On the other hand, the performance of M_G using data from D_I in

18 validation tests resulted in an average accuracy of 56%, $AUC = 0.58$, $TPR = 0.60$ and $TNR = 0.55$.

In the analysis of M_A, to better understand how performance is affected by the different datasets, we report separately the results obtained by testing M_A in participants from DS_I and DS_G. In 18-fold cross-validation with data points from participants in DS_I, M_A showed an average accuracy of 63%, $AUC = 0.61$, $TPR = 0.56$ and $TNR = 0.65$. When using data from DS_G as a test set, M_A accuracy increases to 73%, with $AUC = 0.62$, $TPR = 0.47$ and $TNR = 0.78$.

Table 4: Summary of classification results of all experiments.

Model	M_I		M_G		M_A	
Test dataset	DS_I	DS_G	DS_I	DS_G	DS_I	DS_G
Accuracy	63%	75%	56%	60%	63%	73%
TPR	0.68	0.36	0.60	0.53	0.56	0.47
TNR	0.62	0.82	0.55	0.61	0.65	0.78
AUC	0.65	0.59	0.58	0.57	0.61	0.62

6. DISCUSSION

Our results indicate that a disengagement model trained only with data from users interacting alone with the robot might not be appropriate for group interactions, but a model trained only with group data might perform reasonably well in HRI scenarios with a single user. Table 4 and the chart in Figure 3 summarize the results obtained in the classification experiments conducted in this paper. Overall, the results show that the selected multimodal features can be used to successfully predict disengagement in both types of interaction (i.e., a small group or a single participant).

In the cross-validation results using data collected in the same type of interaction, M_I seemed to be a slightly more coherent model than M_G. However, M_G showed greater flexibility in dealing with data from a different type of interaction. In the testing procedures with data from participants interacting alone with the robots (DS_I), the performance measures of M_G remained roughly the same. On the other hand, in tests with data from DS_G, although accuracy and TNR were fairly high, the TPR of M_I was extremely low.

The performance results of M_A, the model trained with instances from both datasets, lie in-between these two extreme comparisons. The average performance of M_A tested with DS_I instances shows better generalization than when DS_I instances are trained in M_G, but not as good as the performance of M_I. In the tests using group data (participants from DS_G), M_A performs better than M_G in some metrics but again, TPR is very poor. It is relevant to stress that M_A was trained with nearly twice as many data points as M_I and M_G, as it included both datasets (only excluding data from the participant used as test set at every cycle), but more data does not always mean better performance. In this case, M_A had to make generalizations from disengagement behaviors in different type of interactions (individual and group interactions), which could explain the decreased performance when compared to M_I and M_G individually.

A possible interpretation of these results is that in group settings there is more variety in the type of disengagement behaviors children exhibit. When disengaged, some children simply behave as if they were by themselves, while others, for example, start interacting socially with their peers, making the classification of disengagement more challenging because the model needs to make generalizations over a larger set of potential options.

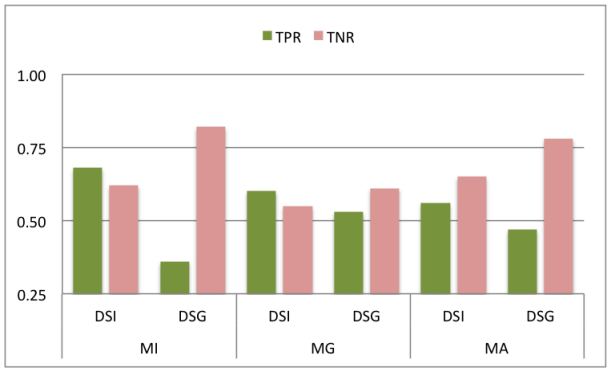


Figure 3: True Positive Rates (TPR) and True Negative Rates (TNR) obtained in the classification experiments.

7. LIMITATIONS

These experiments were conducted in an educational context using children's data. Although the generated models can potentially be re-used in other domains because most of the selected features are domain independent, we cannot be certain that similar performance results will be obtained in different application domains or, more importantly, in datasets with adults.

To further understand the lower performance of M_G compared to M_I, we plan to collect more group data and perform a more detailed analysis of the group interactions. In particular, we intend to study different behavioral profiles, and use that information to segment the group data in several sub-datasets – and eventually different models.

8. CONCLUSION

In this paper, we investigated the role of type of interaction (one participant versus groups of three participants) in the automatic prediction of disengagement in HRI. We reported a set of classification experiments comparing three distinct SVM-based disengagement models generated with the same set of features: a model trained with data from participants interacting alone with two social robots (M_I), a model trained with data from participants interacting with the robots in small groups (M_G), and a third model combining data from the two datasets (M_A).

Our results indicate that, while a model trained with data instances from both type of interactions is a good compromise, not surprisingly the prediction of disengagement episodes is better achieved when the model is trained using only data from a similar type of interaction as the target interaction. In practical terms, ideally the robot should have two different prediction models and, depending on the number of people around it, use the most appropriate model to predict disengagement. However, in cases when that is not possible, we found that a model trained in group data performs better

on single children than the other way around. Although we anticipate similar findings in the prediction of other social and motivational states, further research is needed in this direction.

The main contribution of this paper goes beyond providing a framework for the automatic prediction of disengagement in Human-Robot Interaction using domain independent features; to our knowledge, this is the first work exploring how automatic predictions of a social phenomenon are affected by manipulating the number of people around the robot. With this work, we expect to draw attention in the HRI community to the need for developing perception mechanisms tailored to the specific type of interactions where robots will interact with users.

9. ACKNOWLEDGMENTS

This work was supported by the NSF Expedition in Computing Grant #1139078 and SRA International (US Air Force) Grant #13-004807. We thank Rebecca Marvin for help in video coding, Rachel Protacio and Jennifer Allen for recording the voices for the robots, Emily Lennon for artwork creation, and the students and staff from the school where the study was conducted.

10. REFERENCES

[1] M. Argyle and M. Cook. Gaze and mutual gaze. 1976.

[2] D. Bohus and E. Horvitz. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proc. of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 244–252. Association for Computational Linguistics, 2009.

[3] H. Brugman, A. Russel, and X. Nijmegen. Annotating multi-media/multi-modal resources with elan. In *LREC*, 2004.

[4] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proc. of the 2009 International Conf. on Multimodal Interfaces*, ICMI-MLMI '09, pages 119–126, New York, NY, USA, 2009. ACM.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Trans. on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] Y.-W. Chen and C.-J. Lin. Combining svms with various feature selection strategies. In *Feature extraction*, pages 315–324. Springer, 2006.

[7] M. E. Foster, A. Gaschler, and M. Giuliani. How can i help you': Comparing engagement classification strategies for a robot bartender. In *Proc. of the 15th ACM on International Conf. on Multimodal Interaction*, ICMI '13, pages 255–262, New York, NY, USA, 2013. ACM.

[8] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120 – 136, 2013.

[9] M. Michalowski, S. Sabanovic, and R. Simmons. A spatial model of engagement for a social robot. In *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, pages 762–767, 2006.

[10] L. Moshkina, S. Trickett, and J. G. Trafton. Social engagement in public places: A tale of one robot. In *Proc. of the 2014 ACM/IEEE International Conf. on Human-robot Interaction*, HRI '14, pages 382–389, New York, NY, USA, 2014. ACM.

[11] Y. I. Nakano and R. Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proc. of the 15th International Conf. on Intelligent User Interfaces*, IUI '10, pages 139–148, New York, NY, USA, 2010. ACM.

[12] H. Pohl and R. Murray-Smith. Focused and casual interactions: Allowing users to vary their level of engagement. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '13, pages 2223–2232. ACM, 2013.

[13] P. Rani, C. Liu, N. Sarkar, and E. Vanman. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications*, 9(1):58–69, 2006.

[14] J. Read, S. MacFarlane, and C. Casey. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, volume 2, pages 1–23. Shaker Publishing Eindhoven, 2002.

[15] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. Recognizing engagement in human-robot interaction. In *Proc. of the 5th ACM/IEEE International Conf. on Human-Robot Interaction*, pages 375–382. IEEE, 2010.

[16] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proc. of the 6th International Conf. on Human-robot Interaction*, pages 305–312, New York, NY, USA, 2011. ACM.

[17] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1):140–164, 2005.

[18] C. L. Sidner, C. Lee, L.-P. Morency, and C. Forlines. The effect of head-nod recognition in human-robot conversation. In *Proc. of the 1st ACM SIGCHI/SIGART Conf. on Human-robot interaction*, pages 290–296. ACM, 2006.

[19] D. Szafir and B. Mutlu. Pay attention!: Designing adaptive agents that monitor and improve user engagement. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '12, pages 11–20, New York, NY, USA, 2012. ACM.

[20] M. Vázquez, A. Steinfeld, S. E. Hudson, and J. Forlizzi. Spatial and other social engagement cues in a child-robot interaction: Effects of a sidekick. In *Proc. of the 2014 ACM/IEEE International Conf. on Human-robot Interaction*, HRI '14, pages 391–398, New York, NY, USA, 2014. ACM.

[21] Q. Xu, L. Li, and G. Wang. Designing engagement-aware agents for multiparty conversations. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2233–2242, New York, NY, USA, 2013. ACM.

[22] R. B. Zajonc et al. *Social facilitation*. Research Center for Group Dynamics, Institute for Social Research, University of Michigan, 1965.