Exploration of automatic child-robot engagement measuring

Pieter Wolfert
Ghent University - imec
Ghent
pieter.wolfert@ugent.be

Mirjam de Haas
Tilburg University
Tilburg
m.dehaas@uvt.nl

Paul Vogt
Tilburg University
Tilburg
p.a.vogt@uvt.nl

Pim Haselager
Radboud University
Nijmegen
w.haselager@donders.ru.nl

ABSTRACT

In this paper an exploratory study is presented in which gaze, the frequency of smiling, and posture are used as features for detecting the engagement of children with a robot in a second language tutoring task. We found that a linear combination of these features correlated the strongest with the annotated engagement. While the correlations are currently not strong enough to be used in decision making for a robot's behavior, we hope that this exploratory study can provide new insights on child-robot engagement.

ACM Reference Format:
Pieter Wolfert, Mirjam de Haas, Paul Vogt, and Pim Haselager. 2018. Exploration of automatic child-robot engagement measuring. In *Proceedings of Symposium on Robots for Language Learning*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There is an increasing body of research on the role and effects of robots in education. It has been found that in an educational setting, social robots can have a positive effect on second language learning [7]. Social robots should be capable of interpreting how engaged someone is with that robot, so that in the case of teaching a second language, a child can be offered the best possible learning environment. But what is meant with engagement? Engagement is defined by Sidner et al. [12] as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". We expect that learning improves when the interaction is adapted for the child's (emotional) state and knowledge levels. Schodde et al. [10] used a Bayesian knowledge tracing model to track the word knowledge of participants during an experiment where they were taught 'VIMMI' words (words that do not exist in a real language) by a robot. It was found that participants were able to learn more words when the order of presentation was based on the knowledge of the participant (the adaptive condition) in comparison to a condition where the words were presented in a random order. In another study on adaptivity and the use of gestures by de Wit et al. [4] Dutch children, age 5, were taught 6 animal names in English by a robot. The child's task was to click on the correct animal on a tablet, after a word was presented. This study had four conditions, in a 2x2 design: adaptive versus non-adaptive, and gestures versus non-gestures. The adaptive condition was based on the Bayesian knowledge tracing model, used earlier by Schodde et al. [10]. They found that the engagement got lower towards the end of the task, which is expected given the target group, but there was a positive difference on the level of engagement between the gesture and the non-gesture condition. Anzalone et al. [1] found that a child's engagement is correlated with the movement of that child. In a study by Castellano et al. [3] children had to play chess against a robot. They found that the combination of social features (gazing and smiling towards the robot) with task features lead to a better prediction of the child's engagement. Even though the last study was not in the domain of second language tutoring, the results give rise to the idea that a combination of task and social features could improve second language tutoring. But on which features should we focus when looking at a child's engagement with a robot? Schodde et al. [11] conducted expert interviews to find out where people were looking at when deciding whether a child is engaged. Videos from the study performed by de Wit et al. [4] were used for this. From these interviews three meta-level states of engagement could be identified: engagement, disengagement and negative engagement. Eye-contact, smiling and sitting still were among the features for engagement, whereas gazing away from the robot and rubbing the eyes were seen as features for disengagement. Frowning, head tilting and lowering of the corners of the mouth were identified as features for negative engagement. In this paper an exploratory study is presented in which a bottom-up approach is taken, where gaze, smiling, and posture are selected as features for a pipeline to do online engagement detection in child-robot interaction.

2 METHODS

2.1 L2TOR dataset

For this study the L2TOR dataset by Rintjema et al. [9] is used. This dataset contains 117 video fragments of children interacting with a social robot and a tablet, which have been rated on engagement by 11 annotators(3 females, 8 males, M = 25 years, SD = 3 years) on a five-point likert scale. The intraclass correlation coefficient is .886, and therefore justifies the use of this dataset in this study. A subset of 78 clips with a length of 10 seconds is taken, as not all video fragments are of a high quality.

2.2 Features

Three features are explored in this study; gaze, smiling and posture. The gaze implementation is based on work by Recasens et al. [8].

Symposium on Robots for Language Learning, December 2018, Istanbul, Turkey
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

They trained a convolutional neural network (CNN) on a dataset containing images annotated with head and gaze locations. Individual frames are first processed by a face detection tool [5] after which the frames are used for a forward-pass in the CNN. This results in a list of gaze locations per frame. Since the original model by Recasens et al. is trained on high quality images, individual frames from the L2TOR dataset are manually annotated with gaze locations, to check whether the model can actually be applied to low quality images. For posture, poses are extracted from individual frames with the use of OpenPose [2]. Per frame the head and body movement is calculated by taking the difference in position between the current and the previous frame. Neck movement is used as an indicator for body movement, since the children have to sit down, occluding the lower body for posture processing. Smiling is detected with the use of a neural network that is trained on a dataset with 9475 negative samples and 3690 positive samples[6]. This CNN consists of two convolutional layers and two fully connected layers. It is trained for 20 epochs that results in a final test accuracy of 91%. Per clip the highest smiling prediction is recorded.

3 RESULTS

To look at the quality of the gaze module, manually annotated gaze locations were compared with gaze locations that resulted from inferencing on the model by Recasens et al. [8]. Table 1 shows the results from this comparison, it can be seen that there is a large difference between gaze results from the gaze module and the annotated gaze.

Table 1: Percentage of gazes spent per location per video on average

Region	Gaze Module	Annotated gaze
Robot	42%	14%
Tablet	5%	80%
Other	10%	6%
Missing	43%	

Table 2 shows the results of multiple Pearson correlation tests with engagement (N=78). The annotated gaze is used given the poor results of the gaze module. The only significant correlations can be found when a linear combination of features is tested against engagement.

Table 2: Features versus engagement

Variables	Correlation
Smiling + Head Movement + Tablet Gaze	0.24*
Smiling + Body Movement + Tablet Gaze	0.28*
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Head Movement	0.25*
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Head & Body Movement	0.24*

4 DISCUSSION

Only a combination of variables leads to significant correlations. Individual features do not entail significant correlations. The first reason is that the used dataset is too small. Second, the gaze model by Recasens et al. [8] does not perform well on individual frames extracted from videos. This is mainly due to poor video quality (yellow videos, low lighting conditions, blur caused by a slow shutter speed). Improvements can be made on the gaze model, which could be retrained with extra data containing images extracted from videos annotated with face and gaze locations. Other improvements can be made by adding more features (feature stacking) or using a data driven approach where an end-to-end model is trained on only the videos annotated with engagement, which eliminates the need for feature engineering.

5 CONCLUSION

In this paper an exploratory study is presented in which three features, that are identified in the literature as important, are used to construct a pipeline to predict the engagement of a child in a child-robot tutoring task. Significant correlations are found for linear combinations, but these correlations are not strong enough to be used for decision making. However, the findings indicate that the features can be identified through the use of machine learning tools, and that a combination of these features can provide a prediction of the engagement level.

ACKNOWLEDGMENTS

This work has been supported by the EU H2020 L2TOR project (grant 688014).

REFERENCES

[1] Salvatore M Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. *International Journal of Social Robotics* 7, 4 (2015), 465–478.
[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, Vol. 1. 7.
[3] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 119–126.
[4] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 50–58.
[5] Adam Geitgey. 2018. face_recognition. https://github.com/ageitgey/face_recognition.
[6] Daniel D. Hromada. 2010. SMILEsmileD. https://github.com/hromi/SMILEsmileD.
[7] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction* 19, 1 (2004), 61–84.
[8] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*. 199–207.
[9] Emmy Rintjema, Rianne van den Berghe, Anne Kessels, Jan de Wit, and Paul Vogt. 2018. A Robot Teaching Young Children a Second Language: The Effect of Multiple Interactions on Engagement and Performance. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 219–220.
[10] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (2017), 128–136.
[11] Thorsten Schodde, Laura Hoffmann, and Stefan Kopp. 2017. How to manage affective state in child-robot tutoring interactions?. In *Companion Technology (ICCT), 2017 International Conference on*. IEEE, 1–6.
[12] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.