Online engagement prediction in child-robot interaction

Master's Thesis in Artificial Intelligence

Pieter Wolfert

s4220366

Department of Artificial Intelligence

15 Augustus 2018

Internal supervisor:

Dr. W.F.G. Haselager

Donders Institute for Brain, Cognition and Behavior

Radboud University

External supervisor:

M. de Haas MSc.

Tilburg center for Cognition and Communication (TiCC)

Tilburg University

Second assessor:

Dr. J. Kwisthout

Donders Institute for Brain, Cognition and Behavior

Radboud University

Radboud University

Abstract

Robots are becoming increasingly popular in society. Within education, research is mainly focused on tutoring robots. Currently robots are compared with other methods of tutoring, but these robots are not yet adaptive, whereas adaptivity might make child-robot interaction more natural, and improve learning. Robots that can adapt to the child's knowledge and emotional state, could have a great effect on learning. One of these projects that aims to design a tutoring robot is L2TOR. Within L2TOR a Nao humanoid robot is used for tutoring children a second language. Some steps have been taken in making the robot more adaptive towards children based on the mistakes a child makes during a tutoring task. Currently the child's engagement with the robot is not taken into account. In this thesis a pipeline is suggested to account for the child's engagement with the robot. This pipeline is based on three features: gaze, smiling and posture. These features have been identified as important predictors for child-robot engagement, and can be used to adapt a robot's behavior. An experiment for validating this pipeline is described. Statistics for the individual features were collected, based on an L2TOR dataset annotated with task engagement, and it was found that smiling and a combination of the three features correlated significantly with engagement. Guidelines for future work are provided, since it is expected that online measuring of engagement can work given that enough (balanced) data is available.

Contents

Chapter 1

Introduction

Robots in education have a hard time tracking a child's engagement. Although a robot is able to track faces and make eye-contact, the robot itself does not have an understanding yet of whether a person is actively engaged with the robot. What is meant with engagement? Engagement is defined by Sidner, Lee, Kidd, Lesh, and Rich (2005) as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". It has been found that in an educational setting tutoring robots can have a beneficial effect on learning (Kanda, Hirano, Eaton, & Ishiguro, 2004). What would happen if a robot could improve its tutoring through measuring the child's engagement?

The communication of humans relies on making eye-contact and reading each other's body language. We can make each other feel uncomfortable through facial expressions, or show our engagement towards the other through smiling. Robots are able to smile, change posture and make eye-contact. This is however artificial, and at the moment a one-way street.

What if robots were able to read these communicative cues? And what if they were able to respond to these cues, in a way that is perceived natural by humans? For education this would mean that robots are able to keep up with the children, on a communicative level, which promotes better learning.

This thesis is embedded within the L2TOR project. In the L2TOR project the aim is to design a robot for second language tutoring with 5-year old children. Two questions are at the center of this thesis, the first one being:

"What are good indicators for child engagement in a child-robot task?"

This question has four separate subquestions:

1. Is gaze a good indicator for engagement in children?
2. Is posture a good indicator for engagement in children?
3. Is smiling a good indicator for engagement in children?
4. Is the combination of gaze, posture and smiling a good way to gauge engagement in children?

The second main question covers the feasability of this work:

"Is it possible to do online measuring of engagement in child-robot interaction?"

This question has two subquestions:

1. Would online engagement measuring provide a solid prediction of the engagement of a child with a robot in a learning environment?
2. Would these measurements enable the robot to display appropriate behaviors to re-engage the child?

Data from previous L2TOR studies -including ratings for the engagement of the child with the robot- is used in this thesis to create a model that, given the gaze, posture, and smiling can provide a prediction for the engagement. In this thesis it is investigated whether gaze, posture and smiling can be extracted through the use of machine-learning, such that in the future engagement can be predicted given these three features. An experiment to verify whether engagement predictions can be used for optimizing child-robot tutoring is described in chapter 5. The hypothesis for this experiment is that if a reliable prediction is available, then this would lead to improvements on learning in child-robot tutoring. This thesis aims to contribute to research within child-robot interaction and to provide insights on the automated measuring of engagement in child-robot tutoring, supporting further research in robot behavior adaption, which is outside the scope of this thesis. The hypothesis is that feature extraction can be performed, and provide a solid ground for engagement prediction.

The remainder of this thesis is organized as follows. The next chapter is divided into four sections and provides the foundations on which this thesis is based. First,

an introduction to human-robot interaction is given. Second, an analysis on robots in education is provided, with a focus on tutoring robots. In the third section work from the L2TOR project -in relation to the topic of this thesis- is presented. The fourth section introduces the topic of engagement in child-robot interaction, and provides a background for the selected features. Chapter three provides insight on the methods used in this thesis, and an explanation on the datasets used. Chapter four covers the results of the analysis of the data given the methods from chapter three. In chapter five these results are discussed. The last chapter provides a conclusion of the results and findings of this thesis, and some guidelines for future work.

Chapter 2

Background

This chapter discusses robots and educational robots. It gives an overview of different studies ran on child-robot interaction as well as an examination of the work done within the L2TOR project. As a final step, an analysis of the chosen features for measuring the engagement in child-robot interaction is provided.

2.1 Robots

The last decade has shown an increased demand for robots and automation. Many sectors are currently in the process of embracing robots, from health-care to education and beyond (Ford, 2015). Examples of robots currently employed are the Pepper and Nao by Softbank Robotics (Gouaillier et al., 2009).

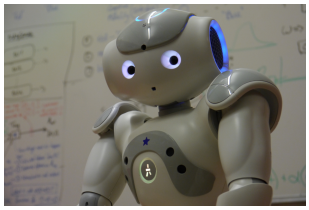


Figure 2.1: Nao robot as used in the ALIZ-E project

An example that can be found on robotics is the Nao robot used in the ALIZ-E project, as can be seen in Figure 2.1 (Belpaeme et al., 2013). One of the evaluation studies in this project was focused on using the robot in a clinical setting. Five children with diabetes played a quiz about diabetes against an adaptive or non-adaptive Nao robot. The difference between the adaptive and non-adaptive condition was that in the adaptive

condition the robot asked for their name, favourite color and sport. It was found that in both situations children were able to learn about diabetes, and that children learned the most in the adaptive robot condition. This shows that through personalizing the interaction, better learning is promoted.

2.2 Educational Robots

Robots can add additional value in education, as written in a review by Mubin, Stevens, Shahid, Al Mahmud, and Dong (2013). In another study by Bainbridge, Hart, Kim, and Scassellati (2011) it was found that robots are preferred over virtual agents in terms of cooperation and learning (Leyzberg, Spaulding, Toneva, & Scassellati, 2012). Robots for education can roughly be divided into three groups according to Chang, Lee, Po-Yao, Chin-Yeh, and Gwo-Dong (2010). They divide educational robots in either learning materials, learning companions/pets, and teaching assistants. An example of learning materials are the Lego Mindstorm robots, where the goal is to understand (robotic) programming (Weinberg & Yu, 2003). Regarding companion robots, the first example that pops up is the Sony AIBO robot, a small, dog-like robot. In a comparison study between a German Shepherd dog and the Sony AIBO it was found that even though both were touched and approached, the real dog was approached more often than the robot dog (Melson et al., 2005). But the fact that the AIBO also was approached in ways similar to the dog showed that some children saw no or little difference between the two. You, Shen, Chang, Liu, and Chen (2006) showed that humanoid robots have a potential to be used as a teaching assistant, when for example the robot tests the students. Most research in educational robotics however is focused on tutor robots, where the robot tutors the child on a topic (e.g. math or language skills).

An example of a robot placed in a school environment is that of Tanaka, Cicourel, and Movellan (2007). A social robot was placed in a classroom with toddlers, for 45 sessions over 5 months. The toddlers could decide to play with the robot, and there was no standard defined interaction of the robot with the toddlers. During the interactions a human operator assisted the robot in decision making. Annotators are showed videos with the interactions of children with the robot, and are asked to annotate the haptic behavior of the children towards the robot and other children. It was found that the

amount of haptic behavior from children towards the robot increased over time, which lead to the conclusion that the robot became a peer.

Figure 2.2: Robovie Robot (Kanda et al., 2004)

One of the first studies on child-robot second language tutoring was a study ran by Kanda et al. (2004). They set up a field trial at a Japanese elementary school. Two identical humanoid robots (Robovie 1 & Robovie 2) were used, and placed outside the classrooms in a corridor (see Figure 2.2). 119 first-grade (6 - 7 years old) and 109 sixth-grade students (11 - 12 years old) in total were exposed to the robots. The students received a RFID tag such that the robot could recognize the student, and personalize the interaction. The robots were placed in the school corridors for two weeks (9 school days). Except safety instructions, no more information was provided to the students. Any interaction that took place was initiated by the students themselves. The robots were able to recognize 50 different English words and reply from a limited set of basic utterances. It was found that students who kept on interacting during the two weeks improved their English, while students who only were able to stay interested for the first week, failed to do so. The interaction was personalized, since the robot knew the names of the children. The robots however did not provide different interactions based on how often they had already interacted with the students. In order to engage more children for a longer time, they ran a second study with interactive behaviors for long-term interaction (Kanda, Sato, Saiwaki, & Ishiguro, 2007). It was found that if the robot developed its behaviors over time, some children were starting to see the robot as a friend. These children were able to keep on interacting during the two months of the experiment. Unfortunately the robot only communicated in Japanese and no comparison with the previous study in relation to language learning could be made.

Another example of learning in child-robot interaction is the CoWriter project (Hood, Lemaignan, & Dillenbourg, 2015; Jacq, Lemaignan, Garcia, Dillenbourg, & Paiva, 2016), where children taught writing to a robot. This approach differs from tutoring robots,

but still the robot is used as a tool for learning. A child is tutoring the robot, such that it can learn through teaching. The investigators found that learning by teaching improved their confidence with handwriting. In a study by Lemaignan, Garcia, Jacq, and Dillenbourg (2016) located at a primary school, children were given individual sessions with the robot. Children had to teach the robot handwriting, following the 'learning by teaching' paradigm. With-me-ness was used as a metric and is defined as 'how much are they with me?' following the definition by Sharma, Jermann, and Dillenbourg (2014) In this study conceptual with-me-ness was measured using the gaze of the participant towards the learning materials on a slideshow. The average duration of these sessions was 20 minutes, and the average with-me-ness was 85.2%. It is important to mention this result as with-me-ness is a precursor of task engagement.

Lastly it is important to mention the study ran by Kennedy, Baxter, and Belpaeme (2015). They set up a study were a Nao robot had a tutoring role, teaching children the concept of prime numbers. There were four conditions:

1. Division only, this condition did not teach the concept of prime numbers.
2. Screen only, teaching the concept of prime numbers, minus the robot.
3. Asocial robot, taught the concept of prime numbers but lacked social skills and personalization.
4. Social personalized robot, taught the concept of prime numbers.

In every group there were on average 11 children. In the social robot condition, the robot could provide more types of feedback, and the interaction between the robot and the child was personalized, e.g. the robot used the name of the child when suggesting to make a move in the game and used gestures to support the communication. In the a-social condition the robot did not use personalized phrases, and made gestures at random times. Between the first and second condition a significant improvement on learning was found, which proves that children were able to learn the concept of prime numbers. There was no significant difference of improvement on the post-test between the social and the a-social robot. There was however a significant positive difference on learning gain between the presence of a robot and the screen only condition. The results from this study support child-robot tutoring, but do not provide evidence for making these robots social, as this does not have an effect on learning.

These examples show that robots can be used within an educational environment, be it as a peer robot interacting on a very basic level (Tanaka et al., 2007), as a robot helping with learning (Hood et al., 2015; Jacq et al., 2016) or as a robot tutoring a child (Kanda et al., 2004; Kennedy et al., 2015). An important feature of the discussed studies is that in all studies there was a form of bonding with the robot, and children were able to be engaged with the robot, resulting in a higher learning gain and longer interactions. However, the robot should not be too social, as the study by Kennedy et al. (2015) shows that such a robot actually entails a lower learning gain.

2.3 The L2TOR Project

As this thesis is embedded within the L2TOR project, an overview on the work resulting from the project, relevant for this thesis, is provided in this section. The L2TOR project is focused around using a humanoid robot for child-robot tutoring. Within the L2TOR the aim is to design a software platform, on a Nao humanoid, to support the teaching of a second language to preschool children (age 5). In the L2TOR project lessons for second language learning are designed around two different domains, namely: number domain and space domain. While a platform has been designed and implemented, the project did not focus on providing individual learning strategies towards children.

2.3.1 Adaptive Learning

Child-robot tutoring can be optimized through promoting a better interaction. One of the reasons that adaptive strategies are being explored is that adaptation can promote better learning (Craig, Graesser, Sullins, & Gholson, 2004; Ramachandran, Huang, & Scassellati, 2017). Schodde, Bergmann, and Kopp (2017) proposed an adaptation of Bayesian knowledge tracing for keeping track of the child's knowledge and to determine which task to offer next. Their model keeps track of the learned words and picks the next task based on the updated belief. They tested this model in a small pilot study were participants, with an average age of 24 years old (SD = 3.82), were randomly assigned to two groups. Participants had to learn German-'Vimmi' word pairs in a game of 'I spy with my little eye...'. Vimmi words are non-words that do not exist in real languages, so the participant has no clue of these words and starts learning from scratch. One group

learned words in a random order whereas the other group learned words where the order was based on the knowledge of the participant. Although they did not find significant differences in a post test, participants seemed to learn more words during the adaptive interaction.

2.3.2 Learning through Gestures

De Wit et al. (2018) performed a study where Dutch children were taught 6 animal names in English, and had to touch the correct animal on a tablet screen, when a name was given. This study had four conditions, in a 2x2 design: adaptive versus non-adaptive, and gestures versus non-gestures. The adaptive condition used was the same as in the previous section, based on work by Schodde et al. (2017). It was found that the use of iconic gestures increased learning and that the adaptive system promoted better engagement with the robot. The average interaction was 20 minutes. Engagement with the robot was seen to get lower overall during the interaction, which is given the average age of the target group (5 years old), an expected result, there was no difference of this lowering between the adaptive and the non-adaptive condition.

2.4 Engagement in Child-Robot Interaction

De Wit et al. (2018) found that during child-robot interaction the engagement seems to get lower at the end of an interaction. Although the engagement in the previous mentioned studies is rated by hand, it would be beneficial for learning if the robotic system can keep track of a child's engagement level. But what is meant by the word engagement? Sidner et al. (2005) defined engagement as: "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". This definition of engagement covers both social and task engagement, which in a child-robot interaction is between the child and the robot or between the child and the task.

Through measuring and accounting for engagement, more personalized ways of tutoring can be promoted. It has been found that personalized human tutoring resulted in better learning (Leyzberg, Spaulding, & Scassellati, 2014). In a study where children had to play chess against a robot, task and social features were recorded for predicting

the level of engagement. Predicting the level of engagement through combining these two features resulted in higher prediction values than when taking the features separately, by comparing this to human-annotated interactions (Castellano, Pereira, Leite, Paiva, & McOwan, 2009).

Schodde, Hoffmann, and Kopp (2017) conducted expert interviews where people with a pedagogical background were asked to rate videos containing child-robot interactions. Videos from the study performed by de Wit et al. (2018) were used for this. Three meta-level states of engagement were identified, namely: engagement, disengagement, and negative engagement. Cues for these different levels were also provided by the interviewees. Eye contact, smiling, and sitting still were among the features for engagement whereas gazing away from the robot and rubbing the eyes were signs of disengagement. Frowning, head tilting and lower mouth corners were identified as behavioral cues for negative engagement. The authors also asked for possible actions to repair disengagement. The interviewees stated that providing different feedback and a break could help re-engage the child. The features identified also re-appear in other research on child-robot interaction, and three features (smiling, gaze and posture) are selected to be used in a pipeline for online engagement prediction in child-robot interaction.

2.4.1 Features: Gaze

Schodde's expert interviews gave hints on which factors there are to determine the engagement level of a child in child-robot tutoring. Gaze is an indication for engagement. Ishii, Nakano, and Nishida (2013) ran a study to estimate engagement from gaze. In their experiment a participant was placed in a room with a screen on one wall, on which a virtual agent was projected. The task for the participant was to listen to an instruction on new model cell phones, and to guess which (out of six) cellphone was most popular among female high school students and businessmen. The participant was able to communicate with the agent and to ask questions. The average length of the conversations was 16 minutes, and data on speech, gaze (using an eye-tracker) and engagement was gathered. The disengagement of participants with the virtual agent was annotated by 10 annotators. Gaze transition, mutual gaze, gaze pattern duration , eye movement distance and pupil size were analysed. It was found that gaze duration and pupil size contributed the most to engagement estimation. Important to mention is the fact that gazing is

culture dependent (Akechi et al., 2013), and that the use of gaze as a metric is not the same among different cultures. Gaze has also been proposed by others as a metric for engagement in human robot interaction (Anzalone, Boucenna, Ivaldi, & Chetouani, 2015; Serholt & Barendregt, 2016). They found that the strongest indicator for a social event is gaze towards the robot's face. Gaze can be measured through eye-tracking devices or by using video-analysis tools. Eye-tracking devices need to be calibrated before they are used, and users need to sit or stand still so that their eye movements can be recorded. This makes measuring gaze through eye-tracking less useful in child-robot interaction, as children move a lot.

2.4.2 Features: Posture

Anzalone et al. (2015) evaluated the engagement with social robots through the use of people and head tracking. They used cheap off-the-shelve sensors, like Kinect and microphones. The Kinect is a device capable of stereo recording and extracting skeleton data of the person in front of it. The Kinect was used for posture extraction, whereas head detection was done after the body was detected. The head and gaze detection that was used relies on landmarks detected in the face, which differs from the approach that will be used in this thesis. In order to use the recordings for engagement evaluation, several metrics were defined: focus of attention, head stability, body posture stability, joint attention, synchrony, and imitation. These metrics were used in two case studies, one of which is particularly important, since posture was analyzed in this study. For this case study children affected by Autism Spectrum Disorder (ASD) and children with a Typical Development (TD) participated. 32 children participated, 16 in the ASD group and 16 in the TD group. The goal of the robot was to elicit joint attention, in the direction of a stimulus (cat vs. dog). It was found that posture variance (the amount of movement) was significantly lower in the TD group than in the ASD group. They conclude that since such differences can be found, it is a valid metric to be used for measuring the engagement.

2.4.3 Features: Smiling

In a study ran by Castellano et al. (2009) children had to play chess with the iCat robot. Eight children played two rounds of chess with the robot, one round was with low difficulty whereas the second had medium difficulty. As can be read in a previous section, task and social features were taken together for predicting the engagement. Important for this section are the social features, since in this study these entail user smiling and user looking at the iCat. Non-verbal behavior had a 93.75% recognition rate for engagement, with a small increase when it was combined with contextual information. This shows that user gaze and smiling can provide solid engagement prediction results.

Serholt and Barendregt (2016) set up a field trial with a Nao robot that took place in a Swedish primary school over the course of 3.5 months. 43 children took part in the experiment, of which 30 completed the full task (3 sessions of map reading with the robot). The robot was equipped with so called social probes, that can be a greeting, feedback or a question. The researchers had a particular interest in the response of the children on these social probes, and videos of the children interacting with the robot were coded according to an encoding scheme, with 'gaze', 'facial expressions', 'verbal', and 'gesture' as the categories. It was found that for facial expressions the children either looked serious or smiled (30% vs 60% of the exhibited behaviors). For gaze, children looked the most at the robot's face (30% of the exhibited behaviors). Gestures were almost unseen (80% as coded 'none'), and the same holds for verbal responses (70%). The intensity of the showed behaviors in the first session declined over the other sessions. The authors state that this is due to the novelty effect, which leads to a reduction in engagement of humans with robots over time as the novelty wears off.

2.5 Conclusions

This chapter gave an overview of the studies ran on child-robot tutoring and predicting user's engagement. Gazing at the robot, smiling, and variance in posture often come back as strong predictors for a user's engagement. The last study also suggests that due to the novelty effect, the measuring of engagement might get harder over time, which creates an extra challenge. Table 2.1 shows the selected features for this thesis as well as how these will be measured.

Table 2.1

Selected features and how they will be measured	
Feature	Way of measuring
Gaze	Number of gazes (per frame) directed at robot, tablet, and other.
Facial Expressions	Prediction of smiling per video clip.
Posture	Average movement of the head and body per video clip.

How can these three features be taken together in a data-driven approach, such that online engagement prediction can be performed? How this can be done, and whether this really works, will be discussed in the next few chapters.

Chapter 3

Methods

This chapter provides an overview of the methods and tools used in this thesis. First the datasets are described, followed by the features. Since the aim of this thesis is to provide a pipeline, it is necessary to make sure all the elements are written using the same programming language. The language of choice is Python, since this is both highly used and freely available to all. All the software used and described in this section is freely available for academic use.

3.1 Datasets

3.1.1 L2TOR Dataset

In order to learn how to detect engagement, a dataset containing video data from previous L2TOR studies is used. The dataset used in this study is the same dataset as used by Rintjema et al. (2018), although some adaptations are made.

Figure 3.1: Still from one of the fragments in the dataset. (Rintjema et al., 2018)

117 clips fragments are extracted from the video data of the study described in (Rintjema et al., 2018). Since a list with start times is not available, all videos have been watched and a list with start times has been created in order to retrieve the smaller fragments. The average duration of the fragments used originally is 5 seconds, but for this thesis a total of 10 seconds is taken, also because the original start times could not be retreived.

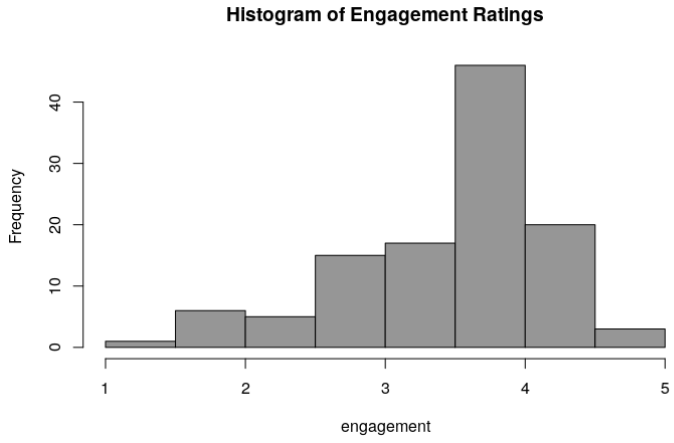


Figure 3.2: Histogram of engagement ratings in the dataset. (Rintjema et al., 2018)

By taking 10 second fragments it is very likely that this covers the original clips used by Rintjema et al. (2018) There are on average 8 fragments per child in the dataset, over 4 sessions (2 per session). Videos were rated by 11 participants (3 females, 8 males, M = 25 years, SD = 3 years). The particpants were asked to rate the engagement on a five-point scale. Before they participated, all raters were instructed on how to spot engagement. The intraclass correlation coefficient (ICC) was reported to be .886, which

is very high, and makes this dataset suitable to be used in this research. Analysis of the tools developed in this thesis will be run on the 117 fragments which will have an average length of 10 seconds.

3.1.2 EmoReact

Figure 3.3: Examples from the EmoReact dataset (Nojavanasghari et al., 2016).

A second dataset, EmoReact (Nojavanasghari et al., 2016) is used for emotion recognition in children. The dataset consists of 1254 clips on 63 children (32 females, 31 males). The clips are sourced from the React channel on YouTube, where children react to different subjects and stimuli (e.g. candy from Japan). These clips are annotated with 8 emotion labels (Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, and Frustration). As can be seen in Figure 3.4 there is a lot of co-occurence of labels in the dataset, meaning that fragments often cover multiple emotions.

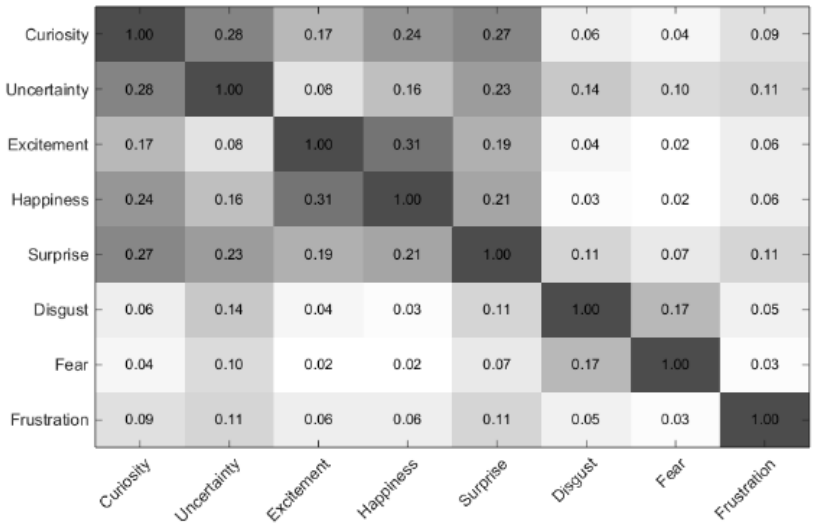


Figure 3.4: Co-occurence of labels in the dataset (Nojavanasghari et al., 2016).

3.2 Gaze Following

For gaze following, a pretrained model is used, developed by Recasens, Khosla, Vondrick, and Torralba (2015). This model has been trained on data annotated with head locations and gaze directions. The original model provided is implementend in Matlab with Caffe. Caffe is a deep learning framework, which is written in C++ but comes with front-ends for Python and Matlab. The code for pre- and postprocessing is completely rewritten into Python, and only the original network structure and weights are used. In order to

Figure 3.5: Example result from the Gaze Following model

calculate the gaze direction, an image with annotated head locations is needed as input. These head locations are calculated using the Face Recognition module for Python. The image is preprocessed and the mean of images in the imagenet dataset is used to normalize

the image (Deng et al., 2009). Three inputs are then calculated from the original image:

1. The normalized image, rescaled to 227 by 227 pixels.
2. Crop of the eyes, given the location of the head in the original image.
3. 13 by 13 grid with zeros, a one on the relative location of the head in the picture.

Once these inputs are calculated, they are used for a forward pass in the network. This results in five outputs. Every single output entails information on the heat map, but only for a part of the image. These outputs are then used for calculating a heat map, where the location with the highest activation is the predicted gaze location. Instead of single images, frames from videos are taken as input for the network.

The stimuli in the video clips from the L2TOR dataset have fixed locations for the tablet and the robot. By defining a region of interest (ROI), the number of gazes per stimulus in a movie clip can be count, using the gaze following model.

Different subsets of the L2TOR dataset will be used, as early tests showed that the face detection module preceding the gaze module had difficulty with children with an Asian background. It also appeared that videos of bad quality (mainly videos that were very yellow), could not be processed correctly, and these are also not part of the subset. In the original paper by Recasens et al. (2015), the gaze module was mainly used on pictures. To assess the quality of the gaze module predictions on video clips, a subset of the L2TOR dataset will be annotated with gaze, to make a direct comparison possible.

3.3 Smiling Detection

Several methods for smiling detection have been tried, which are described in this section. The last method is the method actually used for obtaining the results.

3.3.1 Emotion Detection

Emotion detection is done by using the EmoReact dataset (as described in chapter 3.1.2). First, the movieclips are preprocessed frame by frame, and per frame the face is identified using haarcascades in OpenCV. A crop of this frame is saved as a PNG image. For classification a convolutional neural network is trained and used. This network is

based on the network proposed by Arriaga, Valdenegro-Toro, and Plöger (2017), where the network was used for emotion and gender classification. This network is a fully convolutional neural network consisting of nine convolutional layers, a rectified linear unit, a batch normalization layer and a pooling layer. The network for emotion classification is implemented in Keras, which is a neural network interface for Python (Chollet et al., 2015).

The network was trained in combination with the EmoReact dataset and corresponding labels. It turned out that, due to class imbalance and many images with incorrect labels, the network did not learn.

3.3.2 OpenCV Smiling Detection

OpenCV is a framework that contains a variety of functions and methods for real-time computer vision (Bradski, 2000). In order to detect smiles, first a face needs to be detected. This is done by using the cascade classifier from OpenCV, that requires a cascade file, which comes together with the software. Cascade classification is done by using rectangle features, which are computed over the 'integral image' (Viola & Jones, 2001). A learning algorithm is then used to learn which features are important, since a higher number of features requires more computational effort. Once a face is detected, a second cascade classifier is used to detect smiling. This classifier is trained the same way as the classifier for detecting faces.

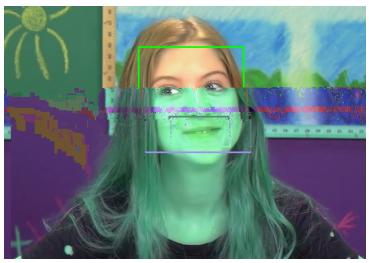


Figure 3.6: Example of face and smile detection

To prevent errors, smile detection is only valid when the center of the region of interest is below the center of the face. The reason for this is that cascade classifiers have a tendency to classify eyes and other parts of the face as a smile.

This approach turned out to classify not only a smiling mouth as smiling, but also other parts of the face, and therefore another method was implemented.

3.3.3 Neural Network for Smiling Detection

An improved method for smiling detection is implemented, as the other methods were not able to provide accurate results. This detection relies on a neural network, that is trained on images containing either a neutral or a smiling face. First face detection is performed using the Face Recognition package [1]. This package relies on HOG (Histogram of Oriented Gradients). Given an image, intensity gradients are identified. Objects and faces have distinctive features. HOGs contain these features, and a linear classifier is trained on these HOGs to perform basic face recognition.

A basic neural network containing two convolutional layers and two fully connected layers is trained on 9475 negative exaples and 3690 positive examples of images containing faces [2]. This network is trained for 20 epochs, where a final accuracy of 91% is achieved on the test set.

Following face recognition, the area of the face is cropped, resized, translated into a grayscale image and forwarded through the network for a prediction on whether there is a smiling face or not. The highest prediction for smiling of over all frames in a video clip is used as the prediction for smiling in that clip.

3.4 Posture

Figure 3.7: Example result from the OpenPose software

For posture, the software package OpenPose is used (Cao, Simon, Wei, & Sheikh, 2017). OpenPose is implemented in Caffe and OpenCV, and comes with many application hooks so that it can easily be incorporated in a pipeline. Since OpenPose offers a

[1] `https://face-recognition.readthedocs.io/en/latest/face_recognition.html`

[2] Source of dataset: `https://github.com/hromi/SMILEsmileD`

pretrained model, this is used for posture extraction from the L2TOR data, since a pre-trained model means that the individual poses do not need to be annotated. OpenPose normally provides 17 posture points per identifiable person. For the head movement, the point related to the nose is taken. For body movement, the position of the neck is used, as the hip is often not visible in the videos.

First all clips are processed with OpenPose, resulting in a 'json' file per frame per clip. This file contains the positions of the 17 posture points in a structured format. The variability in both head and body movement is used as a metric, as also done by Anzalone et al. (2015). The average amount of head and body movement is taken per clip, by using the values in the json files. For every single frame per clip the location of the body and head is determined, and the movement of the body and head between the frames is calculated by using the euclidean distance. This is summed per clip, and divided by the number of frames in the clip, such that this measurement corresponds to the average movement in the video clip.

3.5 Measurements

This section describes how the research questions will be answered.

3.5.1 RQ 1: What are good indicators for child-engagement in a child-robot task?

Is gaze a good indicator for engagement in children?

The number of frames where the gaze is directed at either the robot, tablet or other is counted. Per category, this number is correlated with the annotated engagement in the L2TOR dataset.

Is posture a good indicator for engagement in children?

To answer this question, the amount of movement per video clip is correlated with the rated engagement.

Is smiling a good indicator for engagement in children?

To answer this question, the prediction of smiling per video clip is correlated with the annotated engagement in the L2TOR dataset.

Is the combination of gaze, posture and smiling a good way to gauge engagement in children?

This question will be answered by looking at the correlation of the features (gaze, posture & smiling) with the annotated engagement in the L2TOR dataset. Especially at whether this correlation is strong and significant.

3.5.2 RQ 2: Is it possible to do online measuring of engagement in child-robot interaction?

Would online engagement measuring provide a solid prediction of the engagement of a child with a robot in a learning environment?

This question will be answered given the results from the evaluation study.

Would these measurements enable the robot to display the appropriate behaviors to re-engage the child?

This question will be answered given the results from the evaluation study.

Chapter 4

Results

4.1 Feature: Gaze

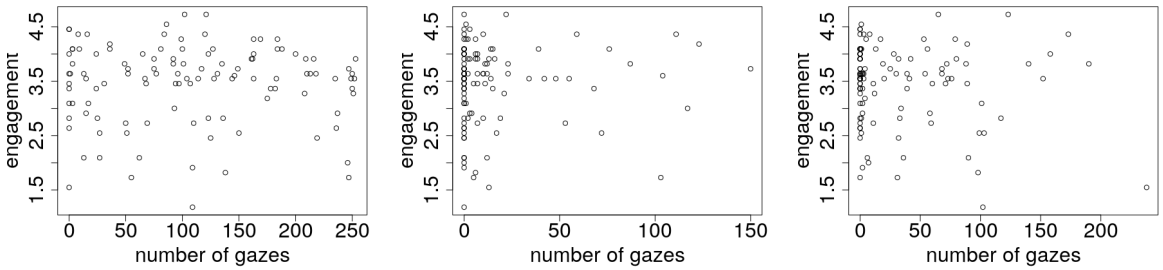
117 clips of 10 seconds were extracted from the L2TOR dataset by Rintjema et al. (2018). Every clip contained 250 frames (25 fps). The average rated engagement in the dataset was 3.45 (sd=0.73) on a scale from 1 to 5.

4.1.1 Results for the complete set

Table 4.1

Percentage of gazes spent on average (per video clip)	
Region	Percentage of gazes
Robot	42%
Tablet	6%
Other	13%
Missing	39%

Table 4.1 shows the average percentage of gazes spent looking at the robot, tablet and other, as recorded with the gaze module for the complete set. Figure 4.1(a) shows robot gaze versus the annotated engagement. To evaluate the relation between the number of gazes spent looking at the robot and the annotated engagement, we calculated the correlation.



(a) **(b)** **(c)**

Figure 4.1: (a) Gaze at robot (b) tablet (c) other

Number of gazes at the robot versus engagement are not correlated and not significant, r(113) = -0.01, p=0.9295.

To evaluate the relation between the number of gazes spent looking at the tablet and the annotated engagement, we calculated the correlation. Figure 4.1(b) shows the number of gazes at the tablet versus the annotated engagement. Number of gazes at the tablet versus engagement are not correlated and not significant, r(113) = 0.03, p=0.7127.

To evaluate the relation between the number of gazes not spent at both the tablet and the robot, we calculated the correlation of 'other' versus the annotated engagement. Figure 4.1(c) shows the number of gazes at other versus the annotated engagement. Number of gazes at other versus engagement are not correlated and not significant, r(113) = -0.12, p = 0.2041.

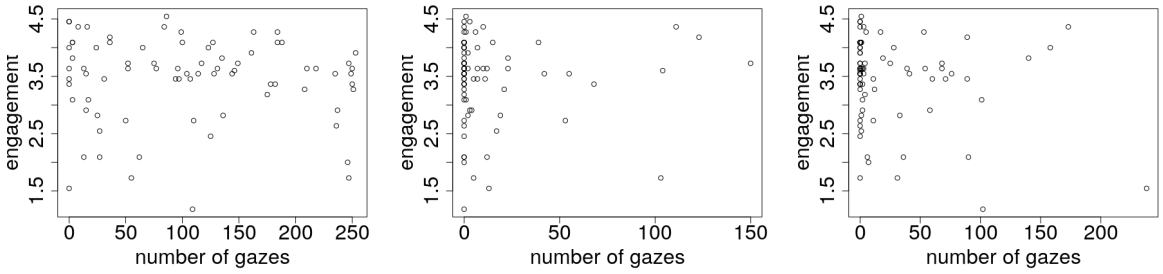
4.1.2 Results for the first subset

In the first subset 'yellow' videos and videos containing participants with an Asian background were excluded. The average annotated engagement for the subset was 3.45 (sd=0.73). Table 4.2 shows the average number of gazes spent at either the robot, tablet or other for this subset. Most gazes were spent looking at the robot.

Table 4.2

Percentage of gazes spent on average for the first subset	
Region	Percentage of gazes
Robot	42%
Tablet	6%
Other	10%
Missing	42%

To evaluate the relation between the number of gazes spent at the robot, in this subset, and the annotated engagement, we calculated the correlation. Figure 4.2(a) shows the number of gazes versus the annotated engagement.



(a) **(b)** **(c)**

Figure 4.2: (a) Gaze at robot (b) tablet (c) other

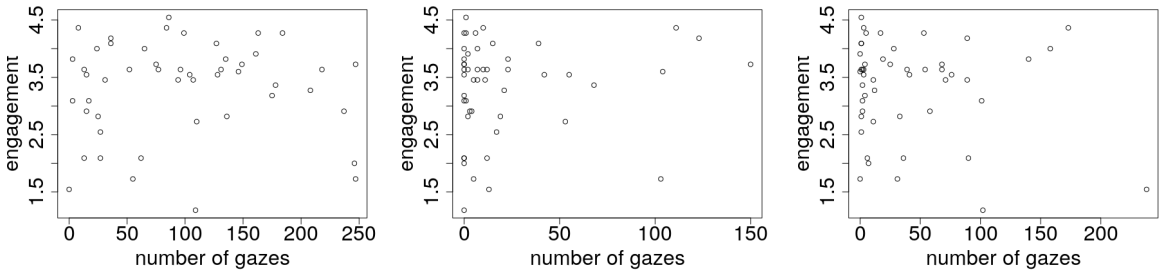
The number of gazes directed at the robot versus engagement are not correlated and not significant, r(78) = -0.06, p = 0.5862. To evaluate the relation between looking at the tablet and the annotated engagement, we calculated the correlation. Figure 4.2(b) shows the number of gazes spent looking at the tablet versus the annotated engagement. We found that the number of gazes at the tablet versus the engagement are not correlated and not significant, r(78) = 0.02, p = 0.8902. To evaluate the relation between not looking at both the tablet and the robot, and the annotated engagement, we calculated the correlation. Figure 4.2(c) shows the number of gazes at other versus the annotated engagement. We found that the number of gazes at other versus engagement are not correlated and not significant, r(78) = -0.19, p = 0.08781.

4.1.3 Results for the second subset

In the second subset, only clips with at least containing two non-zero scores, were taken into account. The average annotated engagement in this subset was 3.32 (sd=0.82). Table 4.3 shows the average number of gazes spent at the robot, tablet or other. Most gazes are directed at the robot.

Table 4.3

Percentage of gazes spent on average	
Region	Percentage of gazes
Robot	38%
Tablet	9%
Other	16%
Missing	37%



(a) **(b)** **(c)**

Figure 4.3: (a) Gaze at robot (b) tablet (c) other

Figure 4.3(a) shows the number of gazes directed at the robot per video clip versus the annotated engagement. To evaluate the relation between looking at the robot and the annotated engagement, we calculated the correlation. We found that the number of gazes at the robot versus engagement are not correlated and not significant, $r(49) = 0.01$, $p = 0.931$. Figure 4.3(b) shows the number of gazes directed at the tablet per video clip versus the annotated engagement. To evaluate the relation between looking at the tablet and the annotated engagement, we calculated the correlation. We found that the number

of gazes at the tablet versus engagement are not correlated and this correlation was not significant, r(49) = 0.11, p = 0.44. Figure 4.3(c) shows the number of gazes directed at other per video clip versus the annotated engagement. To evaluate the relation between other and the annotated engagement, we calculated the correlation. We found that the number of gazes at other versus engagement are not correlated and not significant, r(49) = -0.13, p = 0.4003.

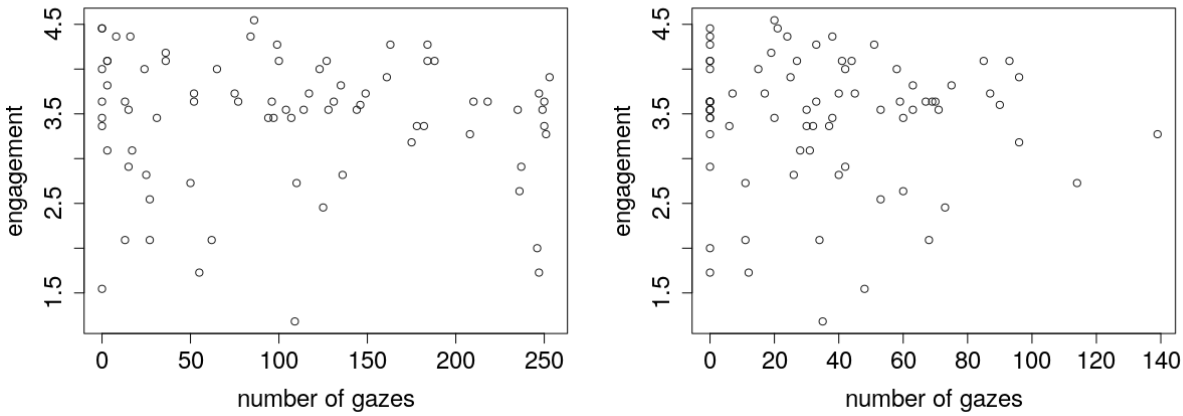
4.1.4 Results of comparison

This section covers the results of the comparison of the annotated clips with the statistics retrieved with the gaze module. The set used for this is the same as in section 4.1.2. 78 clips were annotated with gaze direction statistics, in the same way as how the gaze module gathered these statistics. The average engagement for this set was 3.45 (sd=0.73). Table 4.4 shows the average amount of gazes directed at the robot, tablet and other, both by the gaze module and by annotation of the gaze. It is clear that by manual annotation, most gazes are directed at the tablet, whereas the gaze module seems to miss this.

Table 4.4

Percentage of gazes spent on average		
Region	Gaze Module	Annotated gaze
Robot	42%	14%
Tablet	5%	80%
Other	10%	6%
Missing	43%	

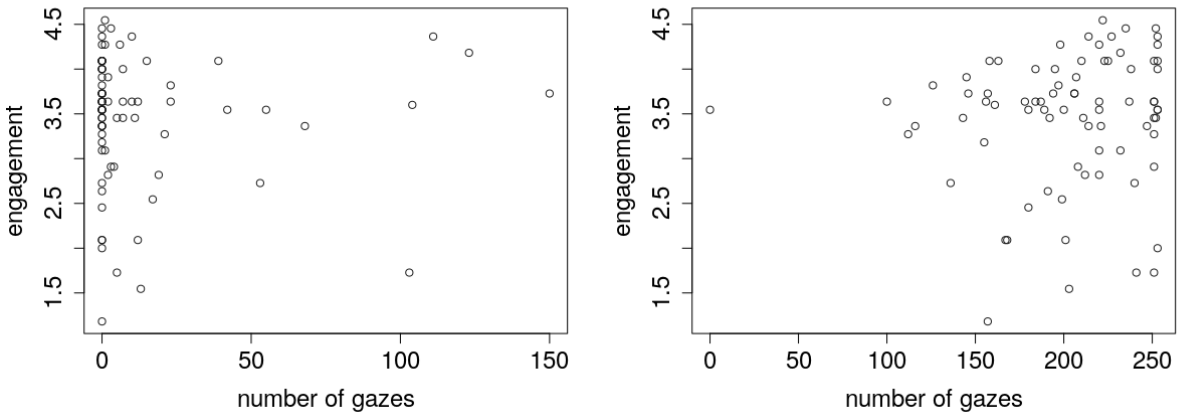
According to the manual annotation, most gazes were directed at the tablet.



(a) **(b)**

Figure 4.4: (a) Gaze at robot (not annotated) (b) gaze (annotated)

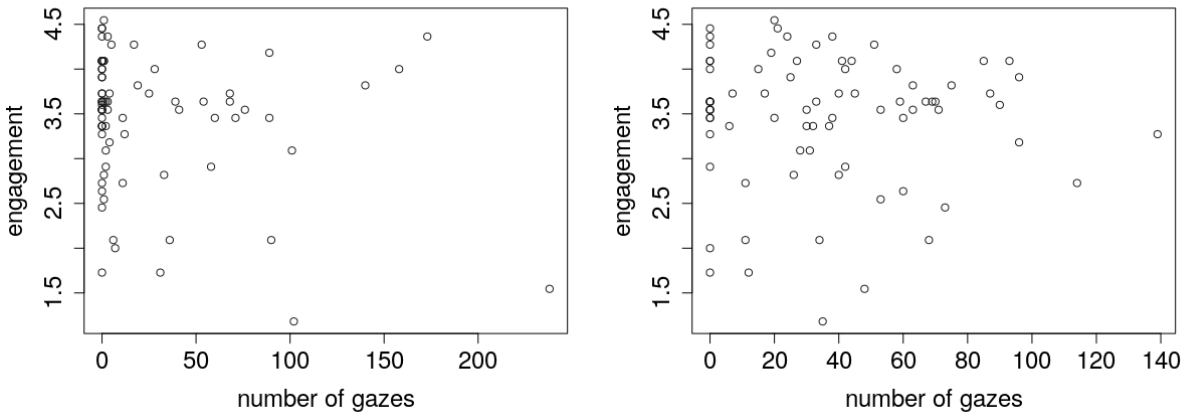
Figure 4.4(a) shows the number of gazes directed at the robot versus the annotated engagement, as captured by the gaze module. Figure 4.4(b) shows the number of annotated gazes directed at the robot versus the annotated engagement. For the second plot we evaluate the relation between the number of annotated gazes directed at the robot and the annotated engagement, by calculating the correlation. The number of gazes at the robot versus engagement is not correlated and this correlation is not significant, r(78) = -0.03, p = 0.7532. The number of gazes directed at the robot according to the gaze module is much higher than the number of annotated gazes.



(a) **(b)**

Figure 4.5: (a) Gaze at tablet (not annotated) (b) gaze (annotated)

Figure 4.5(a) shows the number of gazes directed at the tablet versus the annotated engagement. Figure 4.5(b) shows the number of annotated gazes directed at the tablet versus the annotated engagement. To evaluate the relation between the annotated number of gazes directed at the tablet and the annotated engagement, we calculated the correlation. We found that the number of gazes at the tablet versus engagement is not correlated and not significant, $r(78) = 0.08$, $p = 0.4792$. The number of gazes picked up by the gaze module is much lower in comparison to the annotated gaze. This is also clearly visible when looking at the figures.



(a) **(b)**

Figure 4.6: (a) Gaze at other (not annotated) (b) gaze (annotated)

Figure 4.6(a) shows the number of gazes directed at other versus the annotated engagement. Figure 4.6(b) shows the number of annotated gazes directed at other versus the annotated engagement. To evaluate the relation between the annotated number of gazes directed at other and the annotated engagement, we calculated the correlation. We found that the number of gazes at other versus engagement is not correlated and not significant, r(78) = -0.03, p = 0.7532. The gaze module often did not find gazes at 'other', given the high number of zero measurements. Gaze annotation showed that there were actually more non-zero cases than the results of the gaze module show.

To evaluate the relation between the annotated gaze for robot, tablet and other with the results from the gaze module, we calculated the correlations.

Table 4.5

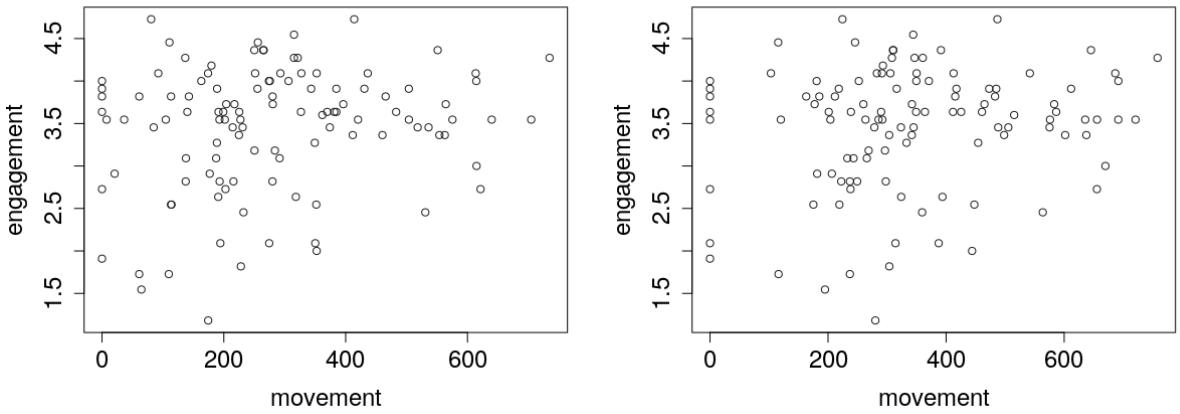
Percentage of gazes spent on average	
Gaze module vs. Annotated Gaze	Correlation
Robot	0.20
Tablet	0.00
Other	-0.03

*$p < 0.05$

4.2 Feature: Posture

We recorded the amount of movement per video clip. The average amount of head movement was 282 (sd=173) pixels, and the average amount of body movement was 343 (sd=179) pixels. Movement translates to the the displacement of the body or the head in relation to the frame.

Figure 4.7(a) shows the amount of body movement versus the annotated engagement.



(a) **(b)**

Figure 4.7: (a) body movement (b) head movement

To evaluate the relation between the amount of body movement and the annotated engagement, we calculated the correlation. We found that the amount of body movement versus engagement is not correlated and not significant, $r(113) = 0.17$, $p = 0.07$.

Figure 4.7(b) shows the amount of head movement versus the annotated engagement. To evaluate the relation between the amount of head movement and the annotated engagement, we calculated the correlation. We found that the amount of head movement versus engagement was not correlated and not significant, $r(113) = 0.16$, $p = 0.17$.

4.3 Feature: Smiling

We recorded the highest prediction for smiling per video clip. Figure 4.8 shows the smiling prediction versus the annotated engagement. Note that these statistics were gathered over the complete dataset.

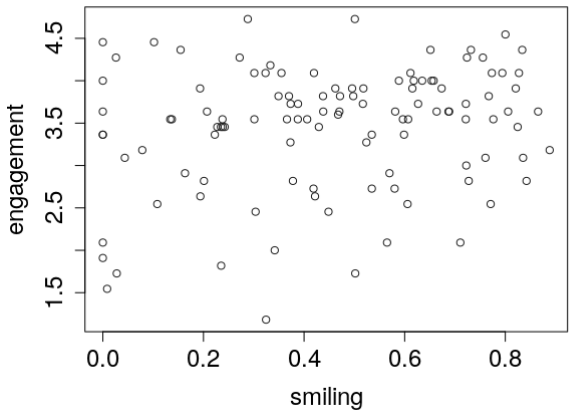


Figure 4.8: Smiling predictions per video clip versus engagement

To evaluate the relation between smiling and the annotated engagement, we calculated the correlation. Smiling versus engagement is correlated and this correlation is significant, $r(113) = 0.20$, $p = 0.03$.

4.4 All Features Together

Annotations were done for 78 video clips, which is the same set as in 4.1.2, and on this subset all variables are compared. Note that some correlations are different with the results earlier mentioned in this chapter, as the comparison of statistics is done on the smallest subset.

Table 4.6

Correlations of combinations of variables with engagement

Variables	Correlation
Robot Gaze	-0.06
Tablet Gaze	0.01
Other Gaze	-0.19
Smiling	0.18
Head Movement	0.25*
Body Movement	0.21
Smiling + Head Movement	0.25*
Smiling + Body Movement	0.21
Smiling + Head Movement + Body Movement	0.23*
Robot Gaze + Tablet Gaze + Other Gaze	-0.16
Robot Gaze + Tablet Gaze + Other Gaze + Smiling	-0.16
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Body Movement	0.10
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Head Movement	0.15
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Head & Body Movement	0.18

Note: Gaze is retrieved with the gaze module.

*$p < 0.05$

Table 4.7

Correlations of combinations of variables with engagement

Variables	Correlation
Robot Gaze	-0.03
Tablet Gaze	0.08
Other Gaze	-0.08
Smiling + Head Movement + Tablet Gaze	0.24*
Smiling + Body Movement + Tablet Gaze	0.28*
Robot Gaze + Tablet Gaze + Other Gaze	-0.02
Robot Gaze + Tablet Gaze + Other Gaze + Smiling	-0.02
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Body Movement	0.21
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Head Movement	0.25*
Robot Gaze + Tablet Gaze + Other Gaze + Smiling + Head & Body Movement	0.24*

Note: Gaze is manually annotated.

*$p < 0.05$

Chapter 5

Discussion

In this thesis the aim was to develop and implement a pipeline that would provide a prediction of the child's engagement with a robot in a robot tutoring task. Originally engagement with the robot was selected to be investigated, but this became impossible given the lack of a dataset with social engagement annotations, and instead task engagement was used.

5.1 Research Question 1

The first research question was "What are good indicators for child engagement in a child-robot task." This resulted in four sub questions, covering the three features (gaze, posture and smiling).

5.1.1 Is gaze a good indicator for engagement in children?

We hypothesized that gaze would be a good indicator for engagement in children. This followed the hints that came from the expert interviews conducted by Schodde, Bergmann, and Kopp (2017). Gaze was found by Ishii et al. (2013) to correlate with engagement, and reported by others as a predictor for child-robot engagement (Anzalone et al., 2015; Serholt & Barendregt, 2016). We did not find a significant correlation between gaze and engagement. We tested both on the results from the implemented gaze module and on the results of the manual annotations. Our hypothesis that there would be a correlation because gaze would be a good indicator, cannot be confirmed.

The gaze module as implemented is based on earlier work by Recasens et al. (2015). In

their work the focus was on extracting gaze given an image and a head location. As videos consist of many images (frames), this in theory should work with videos as well. However it appeared to not always work, and it did not provide good results in comparison to the manually annotated gazes. Camera position and blurred frames are among the reasons why this module performed badly, which can be seen in Figure 5.1. The camera position made it hard to capture the children's frontal face, making it very difficult to capture the full face and especially the eye region for gaze tracking. Blurred frames were a second reason why the module performed badly. Blurred frames are normally not visible in a video, but when individual frames are extracted, it becomes clear that on a lot of frames there is movement, this problem can be tackled by using a camera capable of recording 60 frames per second.

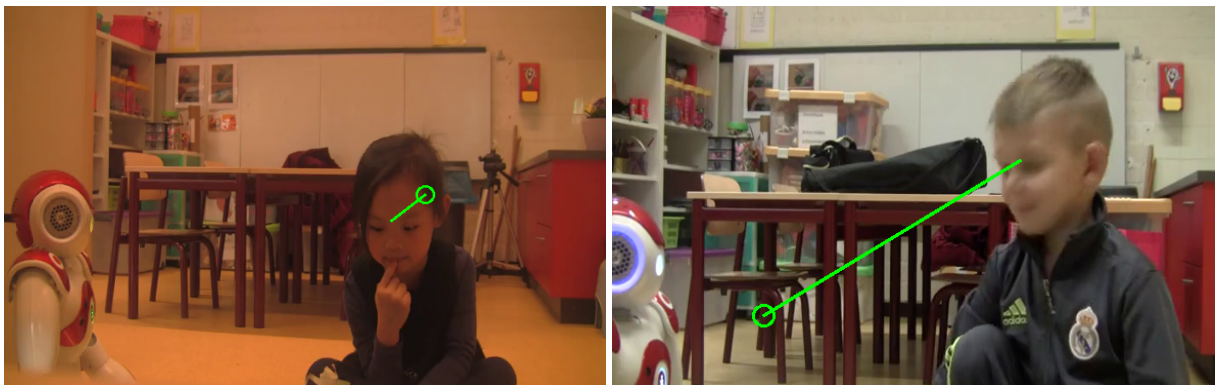


Figure 5.1: (a) yellow video frame (b) blurred video frame

One of the reasons that these separate frames are blurred has to do with lighting. When taking pictures with low lighting, these images often contain moving objects and subjects. For example Figure 5.1b shows a blurred frame, where it might not be directly determined whether the child is looking at the robot or at the tablet. In fig 5.1a it is clear that the child is looking down towards the tablet, but due to the lighting this is classified as gaze towards 'other'. Frames that contain moving faces cannot provide a correct gaze location prediction, given the gaze module. But these same frames can be annotated by a human as looking towards a stimuli, as humans are able to fill in missing details. The research question cannot be answered fully given the results from the gaze module, since these results are not good enough when comparing them to gaze annotations.

5.1.2 Is posture a good indicator for engagement in children?

We hypothesized that posture would be a good indicator for engagement in children. Anzalone et al. (2015) found that posture could be used to distinguish between neurotypical children and children with an autism spectrum disorder. Serholt and Barendregt (2016) found that head movement was positively correlated with engagement. We did not find a significant correlation between head and body movement, and engagement. For some frames the posture module was not able to extract a posture, but overall the performance of this module seemed to be on par. A lack of data might be the reason that we could not find a significant correlation between the amount of movement and engagement.

5.1.3 Is smiling a good indicator for engagement in children?

We hypothesized that smiling would be a good indicator for engagement in children. Serholt and Barendregt (2016) and Castellano et al. (2009) found that smiling can provide engagement prediction. There was, as we expected, a significant positive correlation of smiling with engagement. Given this correlation, our model would predict a high engagement for children who are smiling. It is important to mention that smiling does not occur that often, so when it occurs, it is a large predictor for engagement, if not then it is also not a predictor. Sometimes smiling or non-smiling could not be identified, when for example a hand is in front of the face or when the face is not fully visible (see figure 5.2).

Figure 5.2: Example of face not clearly visible.

Initially the aim was to identify the child's emotional expression (not only whether the child was smiling), but since the EmoReact dataset consisted of video clips only, many

frames did not contain an expression consistent with the label belonging to that video-clip. Therefore the neural network as described in chapter 3 section 5.1 failed to learn the right features to distinguish between the different emotional expressions available in the EmoReact dataset. In the end smiling prediction was performed by training a network on neutral and smiling faces, after several attempts with other techniques.

5.1.4 Is the combination of gaze, posture and smiling a good way to gauge engagement in children?

Correlation testing using all the variables (gaze, smiling and posture) showed a significant correlation of 0.28, but this is a weak correlation and is not strong enough to do engagement prediction. The correlation is an indication that this combination could be a predictor for engagement, but this correlation is probably higher with a bigger dataset. Figure 5.3 shows the pipeline which was meant to be used for this thesis to do engagement prediction.

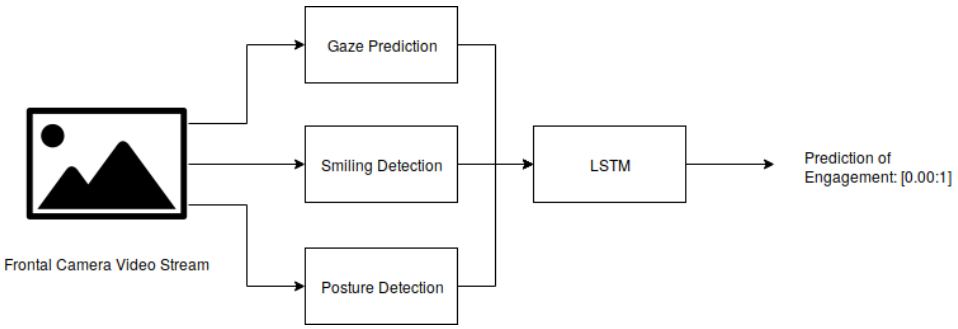


Figure 5.3: Schematic drawing of the pipeline for measuring engagement.

The pipeline comes down to feature extraction and a single Long Short Term Memory (LSTM) layer (Hochreiter & Schmidhuber, 1997). The different features are all implemented in Python and, where possible, weights from pre-trained networks are used. The outputs of the different features are the input for the last layer which is an LSTM. Since the input is a video stream, a recurrent layer is needed to learn temporal relations, such that a prediction is valid for a sequence of frames instead of for one frame. This pipeline would result in a prediction of engagement per frame, based on the frames seen earlier.

5.2 Research Question 2

5.2.1 Would online engagement measuring provide a solid prediction of the engagement of a child with a robot in a learning environment?

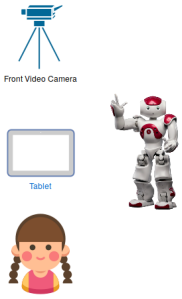
Currently it is not possible to provide a solid prediction of engagement. For this a dataset with enough samples is necessary. The dataset currently used (and the only one available) was too small (N=117) to gather enough information.

5.2.2 Would this measuring enable the robot to display appropriate behaviors to re-engage the child?

If a valid prediction would be available, appropriate adaptive behavior based on engagement predictions could be shown towards the child to re-engage the child. This question cannot be answered given the knowledge this thesis provides, but it is expected that if there is a valid prediction, the robot would be able to display behavior based on this prediction. To test whether accounting for the level of engagement has an effect on learning in child-robot tutoring, the next section describes an experiment which was originally meant to be carried out as part of this thesis.

Evaluation Study

This section describes an evaluation study in which the engagement predictions are used for optimizing and personalizing child-robot tutoring.



(a) Schematic drawing of evaluation study setup

(b) Tablet interface, taken with permission from the author.

Figure 5.4: Interface and setup for the evaluation study.

The aim of this experiment is to teach five-year old Dutch children English animal names with a robot and a tablet, and is based on the experiment described by de Wit et al. (2018). The robot in this experiment is a Nao robot, which is 58 cm tall and weighs 4.3kg (Gouaillier et al., 2009).

First, a pre-test is recorded on animal names, to benchmark the child's knowledge. A child takes place in front of the tablet with a video camera placed directly in front of the child (behind the tablet). On the right-hand side the robot is placed that tutors the child during the experiment. Second, the robot explains the experiment, after which the child is asked whether he or she understood the instruction by pressing a green or a red button. In case the red button is pressed, the experimenter will come in to answer questions. Once the experiment starts the robot partakes in the game "I spy with my little eye", for twenty rounds. Once an animal is spotted, it appears on the tablet together with two distractor images.

In the first condition engagement predictions are used for providing a break to the child, whereas in the second condition this is done halfway the session, and not related to the engagement. The break will have a fixed duration of two minutes, in which the robot asks whether the child would like to dance. A first round is played with Dutch animal names, after which a round with English animal names is played. A day after this experiment a post-test is scheduled to assess whether the child has learned something. These results will be compared between the two conditions. The hypothesis is that when a break is provided based on the engagement levels of the child with the robot, the child is able to learn more than when the same break is provided at a fixed time, halfway the experiment.

Chapter 6

Conclusion

The aim of this thesis was to contribute to the field of child-robot interaction, and specifically to improve second language tutoring. This was meant to be done by extracting features for engagement from videos of child-robot interaction, and use these features for engagement prediction during an interaction. We are able to extract the relevant features and to show the relation of these features towards engagement, but the results are not yet strong enough to be used in a pipeline. The main conclusion is that we are far away from having a pipeline to perform online engagement detection in child-robot tutoring, and provide adaptivity based on engagement. Only smiling and the combination of all features correlated with engagement, despite of the findings in the literature, which showed a significant correlation for each individual feature. We expected the pipeline and all the individual features to correlate with the data, and advance the field with online engagement prediction.

In this research the question was raised whether online engagement prediction would work. Unfortunately, due to a small dataset, we cannot answer this question, but we would like to make some recommendations for future work. First, we would advise to make use of cameras capable of recording 60 frames per second, which reduces blurring in individual video frames. Second, when the features are optimized, an experiment as described in the discussion should be executed to look for the effect of engagement measuring on learning gain.

This research raises new questions on the feasibility of online engagement prediction, and we would like to emphasize the importance of further research in this field, as a fully engaging social tutoring robot is not close yet.

References

Akechi, H., Senju, A., Uibo, H., Kikuchi, Y., Hasegawa, T., & Hietanen, J. K. (2013). Attention to eye contact in the west and east: Autonomic responses and evaluative ratings. *PLoS One*, *8*(3), e59312.

Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, *7*(4), 465–478.

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1), 41–52.

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., ... others (2013). Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, *1*(2), 33–53.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Cvpr* (Vol. 1, p. 7).

Castellano, G., Pereira, A., Leite, I., Paiva, A., & McOwan, P. W. (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* (pp. 119–126).

Chang, C.-W., Lee, J.-H., Po-Yao, C., Chin-Yeh, W., & Gwo-Dong, C. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, *13*(2), 13.

Chollet, F., et al. (2015). *Keras.*

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, *29*(3), 241–250.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., ... Vogt, P. (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* (pp. 50–58).

Ford, M. (2015). *Rise of the robots: Technology and the threat of a jobless future.* Basic Books.

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., ... Maisonnier, B. (2009). Mechatronic design of nao humanoid. In *Robotics and automation, 2009. icra'09. ieee international conference on* (pp. 769–774).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hood, D., Lemaignan, S., & Dillenbourg, P. (2015). When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 83–90).

Ishii, R., Nakano, Y. I., & Nishida, T. (2013). Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(2), 11.

Jacq, A. D., Lemaignan, S., Garcia, F., Dillenbourg, P., & Paiva, A. (2016). Building successful long child-robot interactions in a learning context. In *The eleventh acm/ieee international conference on human robot interaction* (pp. 239–246).

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, *19*(1), 61–84.

Kanda, T., Sato, R., Saiwaki, N., & Ishiguro, H. (2007). A two-month field trial in an

elementary school for long-term human–robot interaction. *IEEE Transactions on robotics*, *23*(5), 962–971.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 67–74).

Lemaignan, S., Garcia, F., Jacq, A., & Dillenbourg, P. (2016). From real-time attention assessment to with-me-ness in human-robot interaction. In *The eleventh acm/ieee international conference on human robot interaction* (pp. 157–164).

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*(March 2014), 423–430.

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Melson, G. F., Kahn Jr, P. H., Beck, A. M., Friedman, B., Roberts, T., & Garrett, E. (2005). Robots as dogs?: children's interactions with the robotic dog aibo and a live australian shepherd. In *Chi'05 extended abstracts on human factors in computing systems* (pp. 1649–1652).

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., & Dong, J.-J. (2013). A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, *1*(209-0015), 13.

Nojavanasghari, B., Baltrušaitis, T., Hughes, C. E., & Morency, L.-P. (2016). Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction* (pp. 137–144).

Ramachandran, A., Huang, C.-M., & Scassellati, B. (2017). Give Me a Break!: Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring. *ACM/IEEE International Conference on Human-Robot Interaction*, 146–155.

Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking? In *Advances in neural information processing systems* (pp. 199–207).

Rintjema, E., van den Berghe, R., Kessels, A., de Wit, J., & Vogt, P. (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* (pp. 219–220).

Schodde, T., Bergmann, K., & Kopp, S. (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*, 128–136.

Schodde, T., Hoffmann, L., & Kopp, S. (2017). How to manage affective state in child-robot tutoring interactions? In *Companion technology (icct), 2017 international conference on* (pp. 1–6).

Serholt, S., & Barendregt, W. (2016). Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. *ACM International Conference Proceeding Series*, *23-27-October*.

Sharma, K., Jermann, P., & Dillenbourg, P. (2014). "with-me-ness": A gaze-measure for students' attention in moocs. Boulder, CO: International Society of the Learning Sciences.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1-2), 140–164.

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, *104*(46), 17954–17958.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer vision and pattern recognition, 2001. cvpr 2001. proceedings of the 2001 ieee computer society conference on* (Vol. 1, pp. I–I).

Weinberg, J. B., & Yu, X. (2003). Low-cost platforms for teaching integrated systems. *Robotics & Automation Magazine*.

You, Z.-J., Shen, C.-Y., Chang, C.-W., Liu, B.-J., & Chen, G.-D. (2006). A robot as a teaching assistant in an english class. In *Advanced learning technologies, 2006. sixth international conference on* (pp. 87–91).