From Real-time Attention Assessment to "With-me-ness" in Human-Robot Interaction

Séverin Lemaignan, Fernando Garcia, Alexis Jacq, Pierre Dillenbourg
Computer-Human Interaction in Learning and Instruction Laboratory (CHILI)
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

***Abstract*—Measuring "how much the human is in the interaction" – the *level of engagement* – is instrumental in building effective interactive robots. Engagement, however, is a complex, multi-faceted cognitive mechanism that is only indirectly observable. This article formalizes *with-me-ness* as one of such indirect measures. *With-me-ness*, a concept borrowed from the field of *Computer-Supported Collaborative Learning*, measures in a well-defined way to what extent the human is *with* the robot over the course of an interactive task. As such, it is a meaningful precursor of engagement. We expose in this paper the full methodology, from real-time estimation of the human's focus of attention (relying on a novel, open-source, vision-based head pose estimator), to on-line computation of with-me-ness. We report as well on the experimental validation of this approach, using a naturalistic setup involving children during a complex robot-teaching task.**

***Index Terms*—Human-Robot Interaction; Visual Focus of Attention; With-me-ness; Real-time Head Pose Estimation.**

I. Introduction

Building capable social agents requires to endow them with a range of perceptual capabilities: however, while face and object tracking and recognition, path planning, speech recognition or task learning are some of the current major research directions, existing literature tends to evaluate each of these algorithms in their own metric space, without considering the interaction quality at a global level. Anzalone *et al.* [1] have recently argued in favor of such a global evaluation, and they propose to assess these algorithms in terms of their capability to *obtain the desired effect* in a human-robot interaction context. They correspondingly propose metrics built around the measurement of *engagement* as indicator of the quality of the experience.

"Engagement", the cognitive, affective and behavioral state of interaction with a computer application that "makes the user want to be there" [2], has actively been studied in a diverse set of domains. Specifically in robotics, several variables and social signals have been proposed in the literature to quantify it. A recent review of these is presented in [3].

For instance, [4] proposes to predict children's level of engagement by integrating non-verbal cues (gaze and smiles) with the current state of the interaction in a Bayesian model. While they report a high level of accuracy, their approach requires post-hoc video annotations, and is not applicable to on-line engagement assessment. Similarly, Baxter *et al.* [5] posit that the measure of the direction and timing of gaze in child-robot interactions is a proxy for engagement and attribution of social agency. However, they also conduct these measures as post-hoc analyses. [6] model the user's interest and engagement with a virtual agent by tracking eye gaze and head direction. Similarly, [7] estimates the user's engagement with a conversational agent based on the analysis of gaze patterns. In [8], a computational model based on the recognition of *connection* events such as directed gaze, mutual facial gaze is proposed. Not relying on gaze, [9] focus only on the back or trunk posture as a determining factor for the assessment. Finally, a recent study with social robots in face-to-face scenarios [1] explores a set of metrics based on non-verbal cues but they also underline possible limitations in long-term scenarios.

The variety of these approaches reflects the fact that *engagement* remains a broad concept, fairly ill-defined and thus difficult to operationalize. Therefore, instead of introducing "yet another metric of engagement", we introduce the more specific concept of *with-me-ness* [10]: to what extent the human is "with me", the robot, during the interaction.

We measure *with-me-ness* by comparing the attentional focus of the human (as estimated in real-time by the robot) with the expected, *a priori* targets of attention elicited by the task at hand.

The following sections expose the full methodology, starting with the estimation of attention: we present in the next section a novel method for on-line estimation of the focus of attention based on fast 6D head pose estimation. We validate this technique in section III with a real-world field study involving children. Section IV formally introduces the concept of *with-me-ness*. We present how to compute it over the course of the interaction to eventually build a new *in-the-moment* measurement of the quality of interaction. We finally validate and discuss this metric by comparing it to manual post-hoc annotations of the video-recordings of the interaction.

II. Visual Attention Assessment

A. Related Work

The relation between one's focus of attention and what he/she is looking at has long been established [11], [12], and more specifically, the existing relationship between gaze and attention during social interaction, and the related gaze

patterns, has been part of classic textbooks like [13] for decades. As such, there is little doubt that measuring the direction of gaze is a useful proxy to estimate the (visual) focus of attention of a social agent, and indeed this is one of the basic tools used in social psychology.

Estimating attention using gaze is not new to robotics either. A recent survey by Ruhland *et al.* [14] gives in a broad overview of eye gaze research in HCI and social robotics. It remains however an active field of research, as illustrated by several recent publications [1], [5], [15]. Performing such a measure on a robot, in real-time, and in ecologically valid environments (which rules out bulky or invasive apparatus like eye-trackers, or techniques requiring fine calibration and/or static interactions) remains a challenge in HRI.

Looking at techniques that both operate on-line and have been deployed in field experiments, one finds that most approaches rely on head pose estimation alone (no eye gaze tracking) and are generally based on depth sensors (RGB-D). Fanelli *et al.* provides an overview of these approaches in [16], and recent examples include [1], [5].

Approaches based on monocular 2D vision have been explored as well [6], with however limited robustness to occlusions or lightning conditions, and over-reliance on tracking to maintain real-time performances. Our work relies on recent advances in template-based face alignment [17] that allows fast (in the order of a few milliseconds) facial feature extraction on 2D images, combined with 3D model fitting, to obtain a fast, robust and stable 6D head pose estimate, that we successfully deployed in field experiments involving child-robot interactions.

We derive the field of attention from the head pose: this is supported by previous work, like [18] that shows that the head orientation's contribution in overall gaze direction is 68.9%, which further translates into a 88.7% accuracy in estimating the focus of attention from head pose only in a particular meeting scenario (using eye and head tracking).

While previous preliminary research in HRI seemed on the contrary to indicate that deriving attentional focus from head pose alone would not be accurate enough [15], we found in our case acceptable levels of agreement between the robot observations and manual post-hoc annotations, as detailled hereafter.

B. Head Pose Estimation

As explained, we derive the visual field of attention from the head pose. Our technique only involves a single monocular RGB camera used for facial feature extraction, and a static simplified 3D mesh of a human head. 68 facial features are extracted using a fast template-based face alignment algorithm by Kazemi and Sullivan [17], as implemented in the open-source `dlib` library [19]. Eight of these features (chosen to be far apart and relatively stable across age and gender) are then matched to their 3D counterparts (Figure 1) and we rely on an iterative PnP algorithm (OpenCV's implementation) to compute the translation and rotation of the head with respect to the camera frame. With this approach, knowing the intrinsic

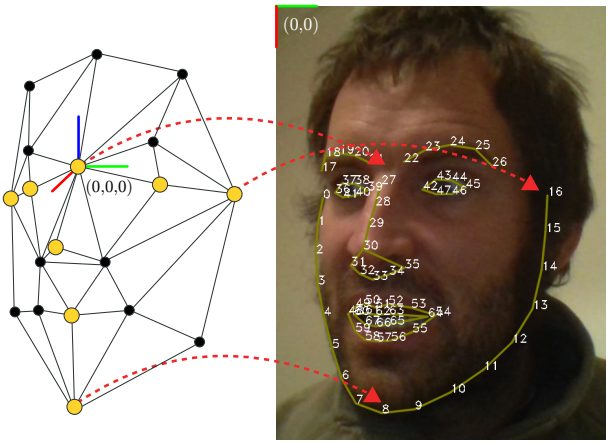

Fig. 1. The 6D head pose is estimated by fitting a 3D model of an adult head (left) onto the detected 2D features of the face (right). We rely on an iterative PnP algorithm, using 8 correspondence pairs (three are depicted: the sellion – the nasal depression –, the left tragion and the menton). The 3D origin of the head is set at the sellion.

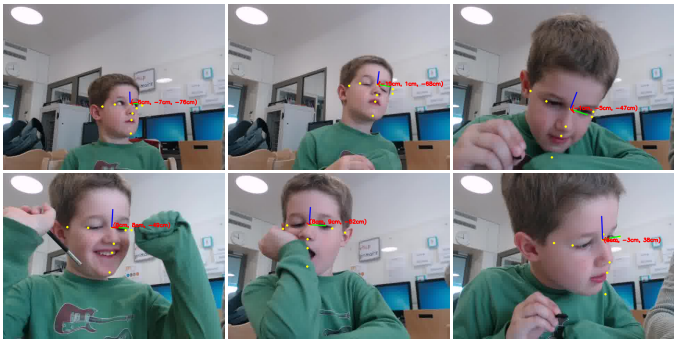

Fig. 2. Head pose results on images captured during a field experiment. Detection of face features (and therefore, estimation of the pose) is robust to significant occlusions and face rotations.

parameters of the camera (calibrated camera) is required for an accurate estimation of the absolute 3D localization of the head.

Besides being fast, the face alignment algorithm has been found to perform well in terms of robustness, including in a range of difficult situations encountered in field experiments, like partial occlusions or large head rotations (we have measured the default `dlib` model to be able to track a face with rotations up to $\pm 40°$ horizontally and $\pm 30°$ vertically). Figure 2 shows a selection of such difficult scenes with one child.

C. Field & Focus of Attention

We model the field of attention as the central region of the field of view. The field of view itself is approximated to a cone spanned from the nasal depression (sellion) of the human face. Different dimensions for the human field of view can be found in the literature: Holmqvist [20] models it with an horizontal aperture of $\pm 40°$ and a vertical aperture of $\pm 25°$, while Walker [21] for instance suggests $60°$ up, $75°$ down, $60°$

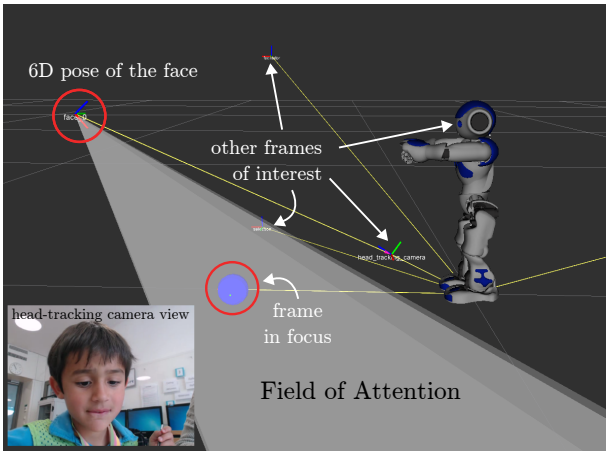

Fig. 3. Screenshot of the real-time attention estimation system. The visual field of attention is approximated to a 40° cone, spanning from the head's sellion. The objects whose 3D pose intersect with this cone are considered *in focus*.

inwards (towards the nose) and 95° outwards. Previous work on visual perspective taking for social robotics [22] model the field of attention as a cone of 30°. We retained in this work a slightly wider aperture of 40°. We then approximate the visual *focus* of attention (VFoA) of the human to the objects which lie inside this field of attention (Figure 3). At a given time, more than one object can therefore be *in focus*.

Our implementation has two limitations: objects are approximated to points (they are considered in focus if their *origin* lies in the field of attention), and we do not check actual visibility: one object could be hidden by another, it would still be considered as in focus. We did not address these limitations since our experimental setup (involving relatively small objects with no occlusions) did not necessitate it. Techniques for more accurate assessment of the visual perspective of the human peer can be found in [22] for instance.

Within these limitations, computing if object $A(x_A, y_A, z_A)$ is in the field of attention of the human requires first to transform the coordinates $A(X_A, Y_A, Z_A)$ into the frame of the face, and then to verify the simple inequality $\sqrt{Y_A^2 + Z_A^2} < tan\left(\frac{fov}{2}\right) \cdot X_A$ (with fov the aperture, and assuming that the main axis $\vec{x}$ of the field of attention points forward).

Our approach assumes that the pose of the objects of interest are available to the system: as described in section III-C, our implementation relies on the ROS TF framework to manage and make available to all software modules the list of poses of existing objects (represented as *frames*), and dedicated perception modules are in charge of publishing up-to-date informations regarding the location of the objects of interest (the so-called *situation assessment*). Due to the nature of the experiment, most of the points of interest considered for the experimental validation presented hereafter are static with respect to the robot, thus simplifying the scene perception.

III. Experimental Validation

As presented above, we use the 6D head pose as an approximation of the actual gaze direction, and we further approximate from here the participant's field of attention. The assumption that such an approximation of the field of attention allows to derive the actual focus of attention needs to be validated experimentally. Our proposed experiment involves child-robot interactions in the context of handwriting remediation. This section details the experimental procedure and presents our results.

A. Experimental Procedure

The experiment, part of the CoWriter project [23], involves a robot which tries to engage a child in handwriting tasks using a *learning by teaching* paradigm (*i.e.* the child is the teacher, and he/she attempts to improve the robot's handwriting). A tactile tablet is used as writing support. Figures 4 and 5 illustrate the experimental setup: a face-to-face child-robot interaction with an (autonomous) Aldebaran NAO robot, in the presence of a facilitator (one of the researchers).

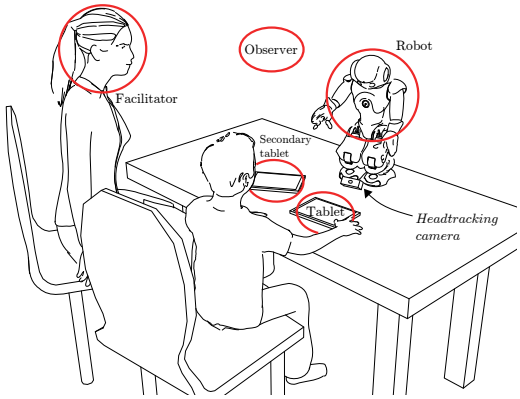

Fig. 4. **Experimental setup**: face-to-face interaction with a NAO robot. The robot writes on the tactile tablet, the child then corrects the robot by directly overwriting the words on the tablet with a stylus. The facilitator remains next to the child to guide the work. The secondary tablet allows the child to tell the robot what to write. The areas of interest – corresponding to potential target of attention – are circled in red with their name.

The subjects were typically located 50 cm away from the robot with the primary (writing) tablet in front and the secondary one 30 cm to the left of the first one. The facilitator was located about 60 cm to the left of the subject. Finally, two observers (visible by the child) were located further away from the interaction field. Figure 4 indicates accordingly the location of main areas of interest (the two tablets, the robot, the facilitator and the observers).

The dependent variable is the measurement of the participants' VFoA, assessed in terms of what the attentional targets of the child are over time. The face of the child is acquired through a fixed webcam (Logitech c920), placed on the table (see Figure 4), and the attentional targets are then computed as presented in section II.

Fig. 5. Picture of the interaction with one of the children.

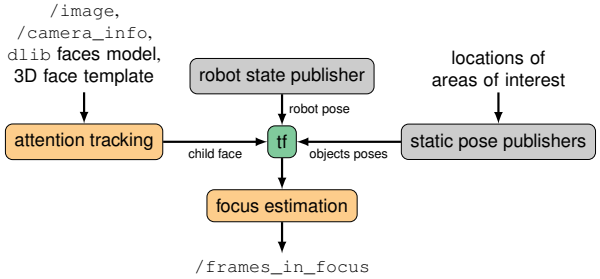

Fig. 6. ROS nodes involved in the VFoA estimation (orange nodes were specifically developed for this work).

B. Experimental Procedure

Six children (ages 5 to 6, 3 boys, 3 girls, none wearing glasses) were enrolled for this study. The study took place at school, in an isolated room (the computer lab). The participants were chosen by the teacher, and would come one after the other to interact with the robot (duration: $M = 19.6$ min, $SD = 1.58$).

The interaction is organized in rounds of writing: during a typical round, the child requests the robot to write something (a single letter, a number, or a full word), and presents a tactile tablet (equipped with a custom writing application) to the robot. The robot "writes" on the tablet by drawing in the air the letters that are displayed on the screen by the tablet application; the child then pulls back the tablet, corrects the robot's attempt by writing on top of or next to the robot's writing, and "sends" his/her demonstration to the robot by pressing a small button on the tablet. The robot learns from this demonstration and tries again. The child continues the turn-taking until they decide to train the robot on another word. In total, the children performed on average 12.16 ($SD = 2.61$) rounds of writing (complete details on the rationale and implementation of this experiment can be found in [23]).

Once per interaction, the robot interrupts the handwriting task to tell a story (taking about 2 min), and the turn-based hand-writing task continues afterwards. The intended purpose of the story-telling episode is to break the routine of the writing turns by creating a surprise, and thus, to elicit a different set of attention behaviors from the child.

C. System Implementation

The experiment was carried out with an Aldebaran NAO robot, using ROS as a middleware to build the attention estimation pipeline (Figure 6). Head pose estimation, presented in section II, builds on the `dlib` and OpenCV libraries; the pose transformations are handled by the ROS TF library. The same TF library is used to represent the possible point of interests as individual frames: an object is considered to be in focus when its frame lies within the field of attention of the participant (Figure 3). The implementation is open-source and available at `https://github.com/chili-epfl/attention-tracker`.

The implementation of the hand-writing activity itself has been presented by Hood *et al.* in [23].

D. Data Collection & Analysis

Successful detections of the head, and, when detected, the attentional targets of the children as estimated by the robot, were logged during the experiment (in total, $6 \times 19.6 = 117.6$ min of interaction). The only post-processing consisted in filtering out gaze shifts (short episodes – below 500ms – between two attentional targets).

We video-recorded the interactions, and performed a *post-hoc* manual coding of the focus of attention (24% double-coded, Cohen's $\kappa = 0.91$, high reliability). The manual coding forms our attentional *ground-truth*.

To assess the accuracy of the attention estimation by the robot, we computed over time the overlap between the ground-truth and the robot's estimate and the inter-rater agreement (Cohen's κ). The periods where the head was not detected were *excluded* from the agreement computation: at such times, the robot explicitly knows that it can not estimate the focus of attention, and as such, we do not consider that it *wrongly* estimates the focus.

E. Results

The main results are reported in Table I Figure 7 further gives a concrete picture of the ground-truth *vs.* computed attentional targets for subject 4 (the subject with the *least* successful tracking).

During the whole interaction, the head pose of the children was consistently tracked, 86% of the time in average, $SD = 3.0$. While this high score is expected for a face-to-face interaction with a static head-tracking camera (meaning that the child head would remain in the field of view of the camera most of the time), this is still comforting in terms of suitability of our approach for head pose estimation with children in field experiments of this kind. Expectedly, the primary causes of lost head pose were occlusions with the hands (similar to the middle-bottom picture in Figure 2), close proximity with the tablet while writing, and gaze directed to the facilitator (who was sitting directly on the left of the child, Figure 5).

TABLE I

ATTENTION TRACKING ACCURACY. *Head pose tracking* IS THE PERCENTAGE OF TOTAL TIME OF SUCCESSFUL DETECTION OF THE HEAD POSE; *Agreement* IS THE PERCENTAGE OF MATCHING TIME BETWEEN MANUALLY ANNOTATED FOCUS OF ATTENTION (GROUND-TRUTH) AND ROBOT'S COMPUTED FOCUS OF ATTENTION. TOTAL DURATION: 117.6 MIN.

Subject	1	2	3	4	5	6	**M**	*SD*
Head pose tracking (%)	88.2	83.5	90.5	83.1	87.9	85.0	**86.4**	*3.0*
Agreement (%)	58.9	67.1	79.2	48.3	65	77.1	**65.9**	*11.5*
Cohen's κ	0.48	0.56	0.68	0.26	0.47	0.68	**0.52**	*0.16*

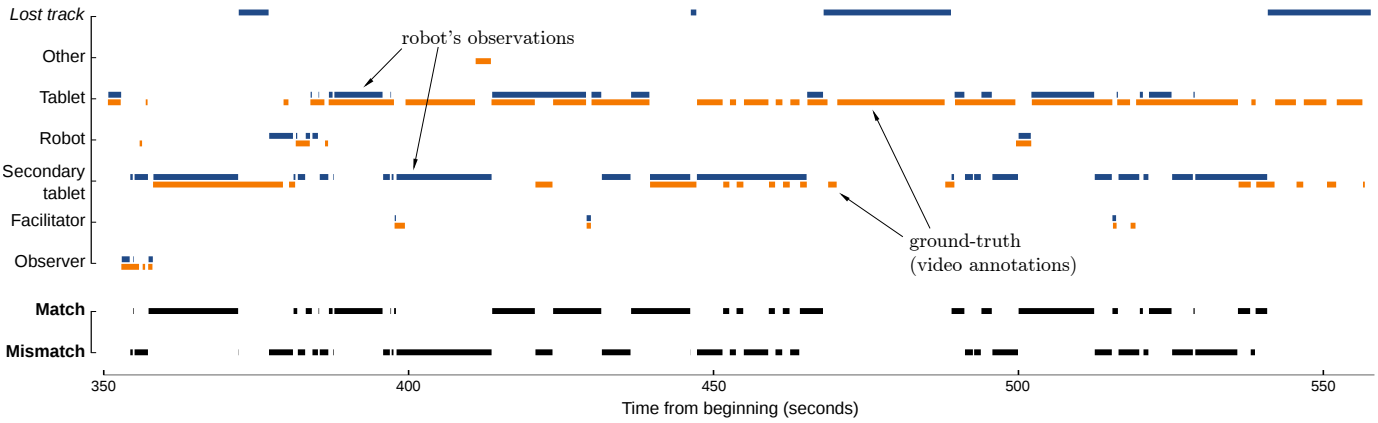

Fig. 7. **Comparison of computed focus of attention *vs.* ground truth** during a face-to-face child-robot interaction (subject 4 in table I, 3.5min-long excerpt). In blue (top lines), the focus of attention as computed by the robot; in orange (bottom lines), the focus of attention as manually annotated (ground-truth). The bottom section shows agreement between both (whenever the head is detected).

In terms of attention tracking, Cohen's κ values are between 0.47 and 0.68 with one subject resulting in significantly worst tracking, at 0.26. While the interpretation of Cohen's κ is subject to discussion (the number of the coded values – in our case 6 – and the distribution probability of values – in our case, values are *not* equiprobable – are factors impacting κ independently of the level of agreement), the levels of agreement are *moderate* to *substantial*, with one subject only showing *fair* agreement [24]. Further analysis of the videos shows that the child with the lowest level of agreement was particularly quiet and would indeed rely more on the eyes to direct his gaze than the other children, thus leading to a less accurate estimation of his focus of attention.

The next section builds upon this technique for real-time estimation of the focus of attention: by comparing the focus of attention with the set of attentional targets *a priori* expected by the robot, we can estimate to what extent the user is "with" the robot.

IV. WITH-ME-NESS

A. Concept & Calculation

The concept of *with-me-ness* has been introduced in the field of *Computer Supported Collaborative Learning* (CSCL) by Sharma *et al.* in [10]. Sharma *et al.* introduce this concept in an attempt to answer a recurrent teacher's question: *"how much are the students with me?"*. They distinguish what they call *perceptual with-me-ness* (the student follows what the teacher refers to with deictic gestures) from *conceptual with-me-ness* (the student follows what the teacher refers to verbally), and they show in an eye-tracking study that *conceptual with-me-ness* in particular correlates with better learning performance. This also relates to the concept of gaze cross-recurrence that has been shown to reflect the quality of the interaction [25] in collaborative learning tasks.

Sharma *et al.* simply define *conceptual with-me-ness* as the normalized percentage of time during which the student's gaze overlapped the areas of teaching slides currently referred to by the teacher. In order to apply it to human-robot interactions, we propose to extend this concept, and to define *conceptual with-me-ness* as the normalized ratio of time that the human interactant focuses its attention on the attentional target expected by the robot for the current task (or sub-task).

Algorithm 1 provides a formal way of computing the level of with-me-ness $\mathcal{W}$ between two time points $[t_{start}, t_{end}]$. A notable difference with the original definition by Sharma *et al.* is that, at a given time t, the task $task(t)$ performed by the robot may elicit more than one attentional target; thus, at a given time, more than one location can be regarded as possible *expected* focuses of attention for the human. For example, a robot which is writing, could typically elicit gazes to its hand as well as to its head. A human looking at either of these locations would be considered to be *with* the robot in terms of

Algorithm 1 Computation of *with-me-ness*. d_w stands for the duration the human is actually *with* the robot, while d_e stands for the total time where the human would be *expected to be with* the robot, $task(t)$ represents the task performed by the robot at time t (possibly none), $F(task)$ represents the (possibly empty) set of expected attentional targets associated to task $task$, $f(t)$ represents the actual focus of attention of the human measured at time t. $\mathcal{W}_{[start,end]}$ represents the level of *with-me-ness* from t_{start} to t_{end}.

procedure COMPUTE WITH-ME-NESS
  $d_w, d_e \leftarrow 0$
  $t \leftarrow t_{start}$
  repeat
    if $task(t) \neq$ nil **and** $F(task(t)) \neq \varnothing$ **and** $f(t) \neq$ nil **then**
      if $f(t) \in F(task(t))$ **then**
        $d_w \leftarrow d_w + \delta_t$
      end if
      $d_e \leftarrow d_e + \delta_t$
    end if
    $t \leftarrow t + \delta_t$
  until $t = t_{end}$
  $\mathcal{W}_{[start,end]} \leftarrow \frac{d_w}{d_e}$
  return $\mathcal{W}_{[start,end]}$
end procedure

TABLE II
MAPPING BETWEEN THE INTERACTION PHASES AND THE EXPECTED ATTENTIONAL TARGETS.

Phase	Expected targets
Presentation	robot
Waiting for word to write	secondary tablet
Writing word	tablet, robot
Waiting for feedback	tablet, secondary tablet
Story telling	robot
Bye	robot

interaction[1]. Also notable, we exclude from the computation of $\mathcal{W}$ all of the periods of time where the user's focus of attention can not be estimated (typically because the user's face is not visible at those times).

B. Experimental Measure & Interpretation

Over the course of the experiment presented in section III, the robot controller would associate a set of expected attentional targets to the phase of the interaction (Table II). For instance, while the robot was waiting for the child's handwriting demonstration ("*Waiting for feedback*"), the expected attentional target of the child was the tablet (since the child was supposed to write there) or the secondary tablet (that displayed a template of the word, used as a reference by the child). These expected targets (green lines on Figure 8) form the robot's attentional *a priori* knowledge and are used to compute the with-me-ness. With-me-ness can be calculated over the whole interaction or over shorter time windows. With-me-ness over the whole interaction for the six subjects is reported in Table III. The Pearson's correlation with the ground-truth is $r(4) = 0.46$ (significance not computed due to small sample size). Shorter time windows are interesting for two purposes: to analyse the level of with-me-ness in relation to specific interaction episodes; to allow a measurement of with-me-ness by the robot *over the course* of the interaction (*in-the-moment* measurement) – in the latter case, one may typically want to consider a sliding time window.

TABLE III
LEVELS OF WITH-ME-NESS. FOR EACH SUBJECT, THE WITH-ME-NESS LEVEL IS REPORTED OVER THE WHOLE INTERACTION, EITHER BASED ON THE ANNOTATED FOCUS OF ATTENTION (*i.e. ground-truth with-me-ness*), OR BASED ON THE FOCUS OF ATTENTION MEASURED BY THE ROBOT.

Subject	1	2	3	4	5	6	**M**	*SD*
$\mathcal{W}_{g.truth}$	79.4	81.6	90.5	87.9	90.7	80.9	**85.2**	*5.1*
$\mathcal{W}_{robot}$	52.6	55.3	74.3	52.9	59.5	63.9	**59.8**	*8.3*

The with-me-ness plotted at the bottom of Figure 8 is in fact computed on a sliding window of 30 seconds, and thus gives a picture of "how well the child is following the robot's expectations" at that time. As seen, the with-me-ness computed at run-time by the robot (blue line) is generally lower than the ground-truth (orange line, based on video-annotations), and sometimes quite off, such as during episode marked "A": during that phase, one can notice that the attention is mostly directed to undefined target Other, likely a consequence of inaccurate head detection. This kind of error (inaccurate head pose estimation) is the main source of discrepancy between the ground-truth and the attention distribution measured by the robot: ignoring all the episodes where the child's gaze is measured to be directed to Other, we indeed obtain levels of with-me-ness close to the ground-truth (over the six subjects, $M = 87.5, SD = 4.6$).

A chart like Figure 8 remains a useful tool to analyse the interaction, and several observations can be made from it: the green lines represent how the robot imagine, at a given time, the attention distribution of the child. They also provide an accurate picture of the overall turn-taking as viewed by the robot: for instance, the episode "B" on Figure 8 corresponds to one of the "*Robot writing*" episodes, surrounded by "*Waiting for feedback*" phases like "C"; episode "D" corresponds to the story telling; etc. In terms of interaction, the large variance of the duration of these phases reflects the fact that this child would sometimes take a lot of time to send feedback to the robot, and sometimes, on the contrary, be very quick.

Looking at the ground-truth focus of attention (orange lines), the first striking observation is that this child did generally *closely* follow what the robot was expecting: in that regard, it seems that the child was very much engaged in the interaction (we discuss in the next section the exact relationships between with-me-ness and engagement). The

[1]Considering a probabilistic model of attention expectations (an attention distribution) would be an interesting extension of this metric.

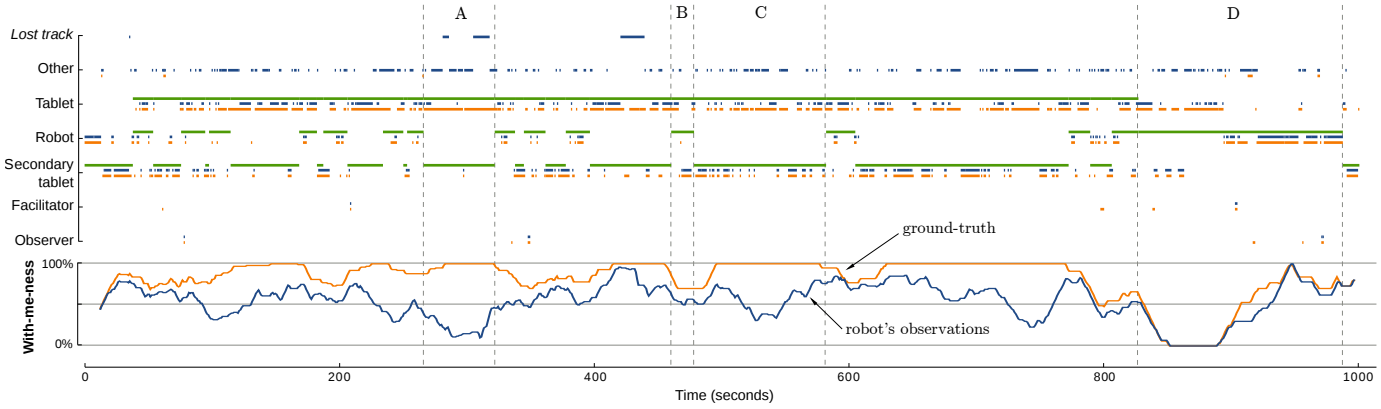

Fig. 8. **With-me-ness**. Evolution of the level of *with-me-ness* over the whole ≈17min long interaction of subject 2. The top chart is similar to Figure 7 with the *expected* attentional targets added in green. The bottom diagram represents the instantaneous level of with-me-ness over a sliding window of 30 seconds. The blue line is the with-me-ness as estimated by the robot, the orange line is the with-me-ness computed from manually annotated attentional targets. Pearson's correlation between both series for this subject: $r(973) = 0.58, p < .001$.

only major exception is the story-telling phase (episode "D"): the child was seemingly not captivated by the first half of the story, and their attention was not directed towards the robot (this actually matches the observed behavior of the children who mostly found the story boring).

Another interesting observation pertains to the facilitator: as one can see, this child only rarely turned to the facilitator, possibly indicating that the interaction and the task were meaningful and easy enough for him to follow alone.

More subtle patterns and events can also be observed: for instance, during the feedback phases like episode "C", we can notice numerous gaze shifts between the tablet (where the child writes) and the secondary tablet (that showed a template of the word). The episode "B" (robot writing) is also interesting: the child did not look at the robot, and instead remained focused on the secondary tablet. This situation is typically useful for the robot to detect as it may want to adapt its behavior to recover the child's attention.

V. Discussion

a) Head Pose to Assess Attention: is it Relevant?: We already stated the main limitations of our approach to estimating the focus of attention: eye gaze information is neglected and we do not perform visibility check of the in-focus objects (we simply approximate them to their origins, ignoring possible occlusions).

While the first issue is shared with most of the other vision (2D or 3D) or motion capture techniques for real-time gaze estimation found in robotics, our results are positive: we show that relying purely on head pose estimation to estimate gaze direction leads to real-world measures that are worth being considered and used. They may not match manual annotations, but they are definitely a valuable *in-the-moment* input for the robot. For certain children, we reach levels of accuracy traditionally considered as good.

Our approach relies on a simple, non-intrusive sensor (a RGB camera by the robot) and an open-source, fast pose estimation algorithm : we hope that this may contribute to the widespread adoption of such a technique on a range of robots, including the relatively common NAO platform.

b) With-me-ness: yet another metric of engagement?: Borrowing the neologism from the field of CSCL, we have also introduced in this article *with-me-ness* as a measure of "how much the user is *with* the robot during a task". This can be acquired over the course of the interaction, thus providing the robot with a real-time metric for a relatively high-level social construct, undoubtedly related to engagement.

One may reasonably wonder how different *with-me-ness* is from *joint attention* on one hand, and from *engagement* on the other hand. With-me-ness is related to both, with however noteworthy nuances: (Triadic) *joint attention* is understood as the cognitive realization of a shared attention to an object, itself building on a shared perception of that object (*i.e.* joint attention builds on a *perceptual* alignment of two agents). *Conceptual with-me-ness* as proposed by Sharma *et al.* in [10] is on the contrary *referential*: "you are *with* me if you focus on what I refer to, either explicitly or implicitly". We understand it here in a slightly broader sense that reflects the *interaction*: "you are *with* me if you focus on what is important for the interactive task at hand."

On the other hand, with-me-ness is only a precursor of engagement: it does not say much about the *cognitive* commitment of a user to an interaction. A user may closely adhere to the injunctions of the robot (or, actually, of the experimenters), with thus high levels of with-me-ness, without being *engaged* in the interaction. This is typically seen in child-robot interaction: children will attempt to closely follow what they are asked to do – which may *look like* they are engaged in the interaction – while they merely *obey orders*.

Compared to engagement, one of the strengths of with-me-ness is its specificity: it is well-defined, we can formalize it, and as such, it is valuable to assess and compare how users are willing or able to interact with a robot. We have hopefully

demonstrated in this article that with-me-ness is an operational *in-the-moment* metric that can also be used as a real-time feedback to the robot controller to build richer, more adaptive interactive behaviors for our robots.

Note however that, besides the actual focus of attention, the mapping *phase/expected attentional target* (*i.e.* our Table II) is a critical piece of information to interpret with-me-ness. The mapping is typically built by a domain expert, and is often subject to debate (for instance in our experiment, one could argue that during the "Waiting for feedback" phase, the child could have gazed toward the robot to make sure the robot was paying attention, and consequently, robot should be added to the expected target). For this reason, the chosen mapping should always be reported along with the computed with-me-ness levels, and with-me-ness should not be reported as an absolute metric, but rather as a mean of comparing different interactions within the same study.

VI. Conclusion

We have presented how a robot can effectively assess in real-time with a regular camera, the focus of attention of its interactants, and how we can combine it with the robot's *a priori* knowledge about the interaction to build a metric of *with-me-ness* over the course of the interaction.

The experimental validation has been conducted with six children in face-to-face interaction with an autonomous robot, over a total duration of about 2 hours. It shows that *1)* most of the time, we are able to estimate the head pose of these children; *2)* based on these head poses only, the instantaneous focus of attention as computed by the robot does reach a good level of accuracy, with however one inaccurate outlier out of the six participants; *3)* the robot is able to compute in real-time a level of *with-me-ness* that correlates strongly with the ground-truth.

The accuracy of the attention estimation could be improved, first by adding heuristics to detect and ignore erroneous/inaccurate face detections, second by implementing pupil tracking on top of head pose estimation. However, our results show that the accuracy levels that we reach already support the reliable computation of a metric measuring a high-level social construct, the *with-me-ness*, which we argued should be used as a well-defined, reliable precursor of engagement in building adaptive robot behaviors.

Acknowledgments

This research was partially supported by the Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by the Swiss National Science Foundation through the National Centre of Competence in Research Robotics.

References

[1] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the Engagement with Social Robots," *International Journal of Social Robotics*, pp. 1–14, 2015.

[2] H. L. O'Brien and E. G. Toms, "The Development and Evaluation of a Survey to Measure User Engagement," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 50–69, 2010.

[3] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in HRI," *International Journal of Social Robotics (submitted)*, 2016. [Online]. Available: http://arxiv.org/abs/1508.04603

[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 *Proceedings of the International Conference on Multimodal Interfaces*, 2009.

[5] P. Baxter, J. Kennedy, A.-L. Vollmer, J. de Greeff, and T. Belpaeme, "Tracking Gaze over Time in HRI As a Proxy for Engagement and Attribution of Social Agency," in *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, 2014, pp. 126–127.

[6] C. Peters, S. Asteriadis, and K. Karpouzis, "Investigating shared attention with a virtual agent using a gaze-based interface," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 119–130, 2010.

[7] R. Ishii, Y. Shinohara, T. Nakano, and T. Nishida, "Combining multiple types of eye-gaze information to predict user's conversational engagement," in *2nd workshop on eye gaze on intelligent human machine interaction*, 2011.

[8] C. Rich, B. Ponsleur, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, 2010.

[9] 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 *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, 2011.

[10] K. Sharma, P. Jermann, and P. Dillenbourg, "With-me-ness: A gaze measure for students' attention in MOOCs," in *International conference of the learning sciences*, 2014.

[11] A. L. Yarbus, *Eye movements during perception of complex objects*. Springer, 1967.

[12] P. Barber and D. Legge, "Perception and information, chapter 4: Information acquisition," *Methuen, London*, 1976.

[13] M. Argyle, *Social interaction*. Transaction Publishers, 1969, ch. The Elements of Social Behaviour, pp. 91–126.

[14] K. Ruhland, C. Peters, S. Andrist, J. Badler, N. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception," in *Computer Graphics Forum*. Wiley Online Library, 2015.

[15] J. Kennedy, P. Baxter, and T. Belpaeme, "Head pose estimation is an inadequate replacement for eye gaze in child-robot interaction," in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction, Extended Abstracts*, ser. HRI '15. ACM, 2015, pp. 35–36.

[16] G. Fanelli, J. Gall, and L. Van Gool, "Real time 3D head pose estimation: Recent achievements and future challenges," in *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*. IEEE, 2012, pp. 1–4.

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874.

[18] R. Stiefelhagen, "Tracking focus of attention in meetings," in *IEEE International Conference on Multimodal Interfaces*, 2002, p. 273.

[19] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[20] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[21] H. Walker, W. Hall, and J. Hurst, *Clinical Methods: The History, Physical, and Laboratory Examinations*, ser. Clinical Methods: The History, Physical, and Laboratory Examinations. Butterworth, 1980.

[22] E. Sisbot, R. Ros, and R. Alami, "Situation Assessment for Human-Robot Interaction," in *20th IEEE International Symposium in Robot and Human Interactive Communication*, 2011.

[23] D. Hood, S. Lemaignan, and P. Dillenbourg, "When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 83–90.

[24] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[25] P. Jermann and M.-A. Nüssli, "Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2012, pp. 1125–1134.