Veronese, Andrea; Racca, Mattia; Pieters, Roel; Kyrki, Ville

Probabilistic Mapping of Human Visual Attention from Head Pose Estimation

Published in:
Frontiers in Robotics and AI

DOI:
10.3389/frobt.2017.00053
10.3389/frobt.2017.00053

Published: 30/10/2017

Document Version
Publisher's PDF, also known as Version of record

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 30 October 2017
doi: 10.3389/frobt.2017.00053

Probabilistic Mapping of Human Visual Attention from Head Pose Estimation

Andrea Veronese[1], Mattia Racca[1], Roel Stephan Pieters[1,2] and Ville Kyrki[1]*

[1] *Intelligent Robotics Group, Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland,*
[2] *Automation and Hydraulic Engineering, Tampere University of Technology, Tampere, Finland*

Effective interaction between a human and a robot requires the bidirectional perception and interpretation of actions and behavior. While actions can be identified as a directly observable activity, this might not be sufficient to deduce actions in a scene. For example, orienting our face toward a book might suggest the action toward "reading." For a human observer, this deduction requires the direction of gaze, the object identified as a book and the intersection between gaze and book. With this in mind, we aim to estimate and map human visual attention as directed to a scene, and assess how this relates to the detection of objects and their related actions. In particular, we consider human head pose as measurement to infer the attention of a human engaged in a task and study which prior knowledge should be included in such a detection system. In a user study, we show the successful detection of attention to objects in a typical office task scenario (i.e., reading, working with a computer, studying an object). Our system requires a single external RGB camera for head pose measurements and a pre-recorded 3D point cloud of the environment.

Keywords: object detection, attention detection, visual attention mapping, head pose, 3D point cloud, human–robot interaction

OPEN ACCESS

Edited by:
Serge Thill,
Plymouth University,
United Kingdom

Reviewed by:
Christian Balkenius,
Lund University, Sweden
Honghai Liu,
University of Portsmouth,
United Kingdom

****Correspondence:***
Roel Stephan Pieters
roel.pieters@tut.fi

Specialty section:
This article was submitted to Computational Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *28 April 2017*
Accepted: *06 October 2017*
Published: *30 October 2017*

Citation:
Veronese A, Racca M, Pieters RS and Kyrki V (2017) Probabilistic Mapping of Human Visual Attention from Head Pose Estimation. Front. Robot. AI 4:53. doi: 10.3389/frobt.2017.00053

1. INTRODUCTION

Modern-day robots are being developed to provide assistance and interaction with people. Such interaction can be physical (De Santis et al., 2008), social (Dautenhahn, 2007), or informative (Goodrich and Schultz, 2007), and involve robotic systems with varying degrees of complexity (Leite et al., 2013). Particular examples include support for the elderly in assisted living environments (Mast et al., 2015; Fischinger et al., 2016) and collaborative robots in manufacturing environments (Michalos et al., 2014). While the capabilities that a robot should have depend largely on its task and the context, the execution of these capabilities depends on the human. That is, when and how interaction or assistance occurs is coordinated (in)directly by human actions. The quality and effectiveness of human–robot interaction (HRI) therefore depends mainly on communication between the human and the robot. Similar to humans, effective interaction between humans and robots relies to great extent on the mutual understanding of actions and behavior (Gazzola et al., 2007; Fiore et al., 2013; Loth et al., 2015). The skills of communicating and recognizing actions apply for both the robot and the human. Expressive gestures will aid in correct recognition of actions and help proceed toward a common goal. Besides gestures, other actions can be used as a means of communication. Human attention, for example, is a mechanism that allows for selectively concentrating on individual components or tasks in the present world.

As such, tracking human attention allows to deduce (or at least estimate) the actions of a human.

Our approach is to enable the estimation of human actions by tracking human attention over time, and assume that paying attention for an extended period of time to a certain object implies engagement with it. Attention is modeled probabilistically in order to take into account the inaccuracy of using head pose as measure of attention. Head pose is chosen over gaze as gaze tracking glasses can hinder the natural behavior of the user, and the fact that tracking gaze by external cameras can be difficult. Additionally, head pose relates to human attention as humans align their head to their direction of gaze when something of interest is found and requires more attention. A weighted Normal distribution, centered around the head pose, is projected into a 3D point cloud of the environment in order to assign weight to individual points and build an attention map. This attention map is segmented by modeling as a Gaussian Mixture Model (GMM), where each component of the mixture corresponds to an object in the scene. Each object can then be related to a predefined user action. Tracking attention over time allows a Bayesian inference to objects that enable ignoring measurements that do not correspond to the current activity. For example, brief glances elsewhere in the scene or incorrect measurements do not lead to a misclassification of the current object of interest.

As scenario we adopt a table-top office environment where a person is assigned a task that involves reading, writing, and studying an object. The person is allowed to use a computer, and can move freely while staying seated. The task is explained on two pages that are fixed to the table and contains questions that the person should answer by writing them down on the paper. Information can be retrieved by accessing the computer and by studying the object. This arrangement assigns three distinct areas to which attention will be directed when performing the task: the documents on the table, the computer, and the object. Additionally, when thinking or distracted to the current task, the person can be directing attention elsewhere in the scene. This should be classified as outliers to the current task at hand. With this scenario, we aim to study whether objects of interest can be segmented from a person's attention to the scene, and whether human actions can be deduced from these segmented objects.

The main contributions of this work are: (1) the probabilistic modeling of human attention based on head pose, (2) the modeling (GMM) and tracking of attentional objects in the scene, and (3) a discussion on the relation between head pose, attention, and actions. We continue by first reviewing several related works that have studied attention and actions regarding HRI.

2. RELATED WORK

The recognition or detection of behavior has been studied in the past from both a psychological point of view, i.e., how humans show, share and process intention (Dominey and Warneken, 2011; Margoni and Surian, 2016), and from a robotics point of view, i.e., the role of intention in cognitive robotics (Dominey and Warneken, 2011; Anzalone et al., 2015; Vernon et al., 2016). As such, actions and perceptions depend heavily on context and setting. For example, social interaction unconsciously directs human perception to detect cues and patterns that are related to the social behavior at hand, and within its current setting. Considering a human–robot interaction scenario, it should therefore be identified that both the expression and recognition of actions are equally important. Regarding the expression of actions, robots and machines can do this either physically or via some method of projection. In the work by Gulzar and Kyrki (2015), it was shown that expressive gestures, such as whole-arm pointing, can benefit interaction. Deictic gestures can be used to anchor symbols to perceived objects and, by evaluating possible pointing locations, avoid geometric ambiguities. Experiments in a multi-agent scenario with the humanoid robot NAO and a Kuka YouBot shows that, when referring to a single object in a cluttered scene, communication can effectively be guided via gestures. Actions expressed by projecting a path or direction of motion has been the topic of research of Chadalavada et al. (2015). Future motion and shared floor space of autonomous vehicles is projected onto the floor for communication from robot to human. The approach increases the safety of bystanders and improves the efficiency of logistics, as vehicles are less halted due to personnel that is unaware of the robot's (future) behavior.

The (artificial) recognition of actions, on the other hand, is a more difficult task, as it depends on many different factors (e.g., visibility of the scene and the people in it, quality of training data and camera for recognition). One approach is to deduce actions from other, more easily measurable features, such as motions (Poppe, 2010; Herath et al., 2017), activities (Vishwakarma and Agrawal, 2013), or attention (Lemaignan et al., 2016). Visual attention is useful for estimating human attention in situations where people are visually interacting with objects. The human Field Of View (FOV) directs visual attention and can be extracted from camera images by detecting eye gaze (Palinko et al., 2016). The most robust techniques for gaze tracking require images of the human eyes and therefore use wearable devices, e.g., glasses as presented in Kassner et al. (2014). These kind of devices can be invasive and not suitable for certain scenarios, hindering the natural behavior of the user. Using external devices like cameras to track gaze is ineffective due to, for example, the relative dimension of the eyes in the image (the farther the camera, the smaller the eyes appear), illumination issues and occlusions (e.g., users with conventional glasses). When gaze tracking is not an option to infer attention, the typical alternative is head pose tracking (Palinko et al., 2016). As head pose estimation does not include information about eye movement most methods model the FOV as a pyramid or cone, growing from the user's face and following the direction of the head. The aperture of the pyramid or cone thus accounts for this missing gaze information (Sisbot et al., 2011; Palinko et al., 2016) and has to be set considering physiological, psychological, and application constraints. Several related works have adopted this approach of detecting human cues (such as visual attention) for the understanding of and assistance in human–robot interaction. As this is the approach in our work, these are explained in more detail in the following.

Human attention, as the cognitive process of concentrating on one particular task or activity, is a complex behavioral action and has been under study for many decades. Based on psychological studies (Rayner, 1998) it is identified that a person's direction of gaze correlates to what they pay attention to. Similarly, this is identified in Lemaignan et al. (2016) where the degree of interaction (or attention) has been studied for infants interacting with robots. The found level of interaction is of interest particularly for infants in a teaching scenario or for infants suffering from autism or any other related attention deficit disorder. According to the level of interaction, a robot can adjust or motivate an infant and provide better support. The chosen scenario consists of a Nao robot, two tablets and a human supervisor, and associates head pose to attentional targets in the scene.

Inferring engagement from non-verbal cues in human–robot interaction is studied with two cases in Anzalone et al. (2015). The first case experimentally analyses head and trunk poses both statically and dynamically. That is, single measurements and the temporal evolution of poses are processed by methods such as heat-maps, k-means, and temporal diagrams. The second case of study measures the engagement of children with an autism disorder. Behavior and joint attentional patterns are analyzed in a human–robot interactive task, considering both the robots Nao and iCub. While the scenario and tasks are relatively simple, and the attentional targets have to be known beforehand, pronounced differences between children with and children without autism can be detected.

Detecting visual attention from multiple people engaged in a meeting is studied in Murphy-Chutorian and Trivedi (2008). The aim of the study is to assess the person that receives the most visual attention among all meeting members. In particular, the joint attention of the participants is considered as a cue of saliency assigned to objects in the environment. The experimental scenario is a 5 min meeting between four people taking place in a room equipped with four RGB cameras. The presented system is fully automatic and builds the environment model (i.e., the positions of the members) and focus of attention model of each participant from head pose tracking.

Doshi and Trivedi (2010) introduce an approach for locating the attention of a human subject by observing both the subject and the environment. The images of the human are used to extract head pose and gaze direction, while the images of the scene are analyzed for modeling saliency maps. The task that the human is assigned to (i.e., a driving scenario) influences the attention and is included in the approach as well.

Summarizing, visual attention addressed by a person to the environment is of interest as this allows the recognition of objects and persons of interest in the scene (Sheikhi and Odobez, 2015). Gaze estimation is in most cases based on the direction provided by the head pose, as the analysis of eyes is difficult to perform in real-world scenarios, especially without obtrusive devices such as eye trackers. Our approach is different from existing works as our probabilistic attention model allows for a distribution of attention centered around the head pose of the human, and enables the modeling of the targets of attention as Gaussian Mixture Models (GMM). Moreover, our approach does not require the objects of interest to be known beforehand. The tracker gives as output the most likely viewed object in real-time, which can then be used to infer actions. The system requires a single external RGB camera for head pose measurements and a pre-recorded 3D point cloud of the environment.

This document proceeds in Section 3 with the modeling and tracking of attention. This includes the modeling of the targets of attention as a Gaussian Mixture Model (GMM) and the subsequent tracking of attention with respect to these targets. Section 4 describes the scenario and the evaluation of both the attention map and the attention tracking. Section 5 presents a discussion and conclusions are drawn in Section 6.

3. MATERIALS AND METHODS

We propose a probabilistic approach to model the attention distribution of the user on the environment. Our method requires the tracking of the user's head and a 3D point cloud representation of the environment. The general idea is to allocate a measure of attention on the environment, by means of analyzing the user's head pose. To supply for the missing gaze information, we model the attention distribution inside the FOV as a 2 dimensional Normal distribution with varying covariance. Combining the head pose tracking and the point cloud, we obtain an attention map of the environment. We can then model the targets of attention in the scene by segmenting the attention map. **Figure 1** summarizes the components of the proposed framework.

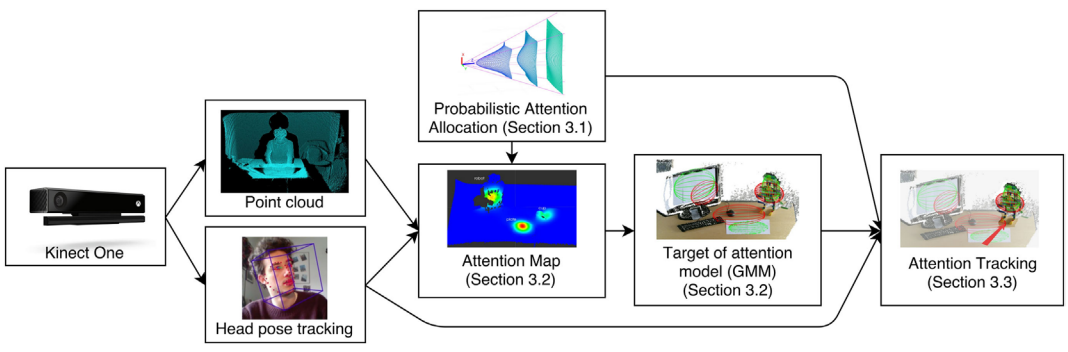

FIGURE 1 | Overall scheme of the proposed framework.

3.1. Probabilistic Attention Allocation

First, we model the FOV as a pyramid of aperture α_{FOV}, starting from the area between the eyes of the user. Any point lying outside the FOV is considered unseen and therefore not further processed. Our probabilistic model follows our assumption that most of the gazes are directed in the central area of the FOV, without however completely ignoring points in the peripheral area of the FOV. The attention measure $v_a(p)$ of point $p \in FOV$ is modeled with a 2D normal distribution as

$$v_a(p) = \frac{1}{2\pi\sqrt{|\Sigma_a(z, \frac{\alpha_{FOV}}{\tau})|}} \exp\left\{-\frac{1}{2}\begin{bmatrix} x & y \end{bmatrix}\Sigma_a\left(z, \frac{\alpha_{FOV}}{\tau}\right)^{-1}\begin{bmatrix} x \\ y \end{bmatrix}\right\}, \quad (1)$$

where $(x\ y\ z)'$ are the coordinate of point p in the FOV frame and $\Sigma_a(z, \frac{\alpha_{FOV}}{\tau})$ is the distribution covariance matrix. The parameter τ determines how strongly we assume that user's gazes are concentrated in the center of the FOV. High values of τ will produce more peaked distribution, increasing the attention measure of points directly in front of the user's face while decreasing it for points in the peripheral areas of the FOV. Small values of τ will produce flat attention distributions.

The covariance matrix Σ_a depends both on the distance of the point from the user's face z and on the fraction of the FOV's aperture $\frac{\alpha_{FOV}}{\tau}$. In particular,

$$\Sigma_a\left(z, \frac{\alpha_{FOV}}{\tau}\right) = z^2\begin{bmatrix} \left[\tan\left(\frac{\alpha_{FOV}}{\tau}\right)\right]^2 & 0 \\ 0 & \left[\tan\left(\frac{\alpha_{FOV}}{\tau}\right)\right]^2 \end{bmatrix}. \quad (2)$$

With this choice of Σ_a, the same amount of attention measure is allocated for points inside the central area of the FOV of aperture $\frac{\alpha_{FOV}}{\tau}$, at different distances from the user's face. As Σ_a increases with the distance z, the attention will be more spread, as can be seen in **Figures 2** and **3**. This reflects our observation that the farther the targets, the more difficult estimating the user's attention on them becomes. **Figures 2** and **3** show how the attention v_a is computed inside the FOV for different distances from the user's head, for two different values of τ. Following the assumption that the user's attention is concentrated around the center of the FOV, points located along the center will receive more attention measure. However, as this assumption becomes weaker for points far away from the user, also the attention distribution becomes flatter and flatter.

3.2. Targets of Attention Modeling

As previously mentioned, our approach represents the environment with a 3D point cloud representation, captured with an

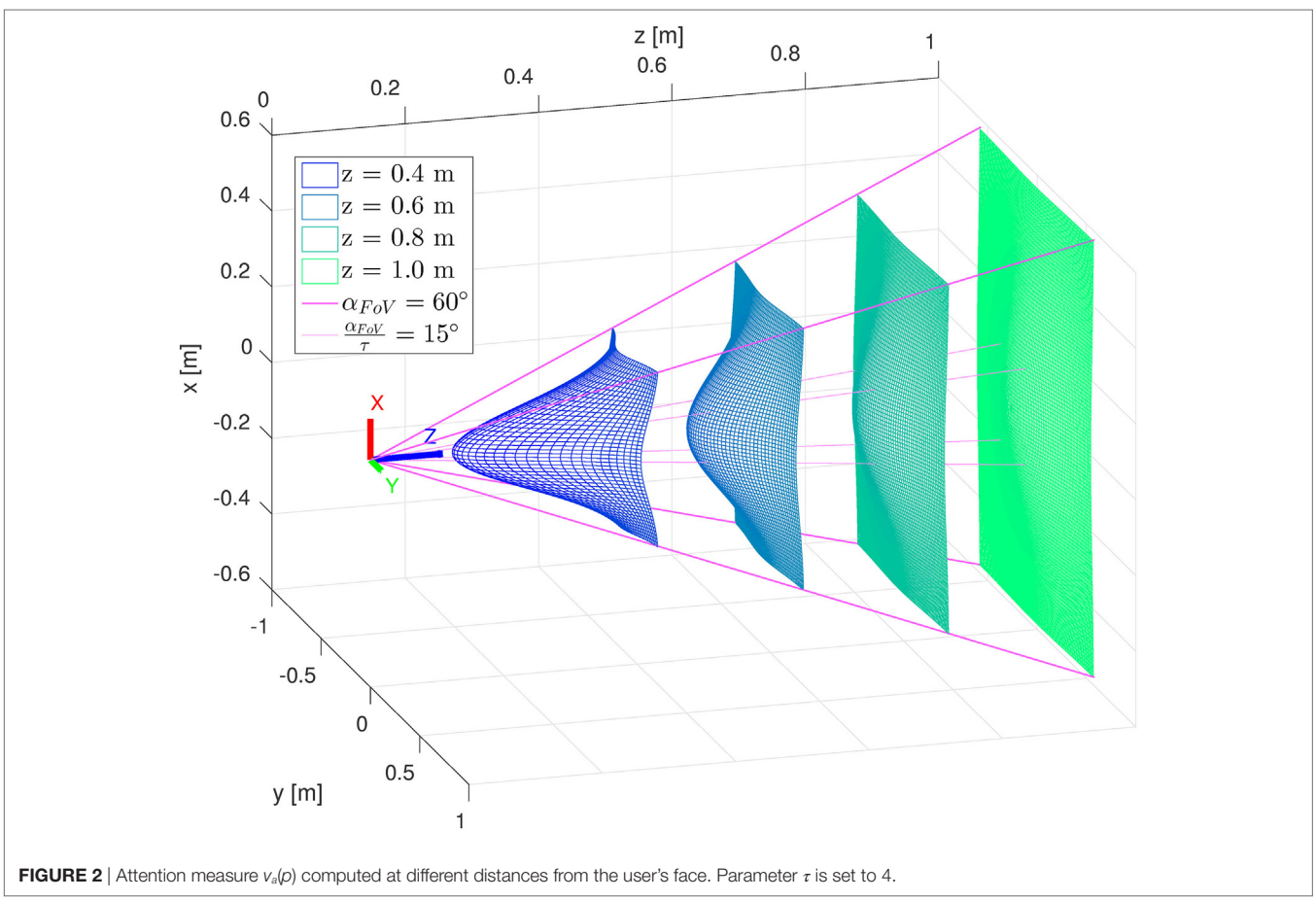

FIGURE 2 | Attention measure $v_a(p)$ computed at different distances from the user's face. Parameter τ is set to 4.

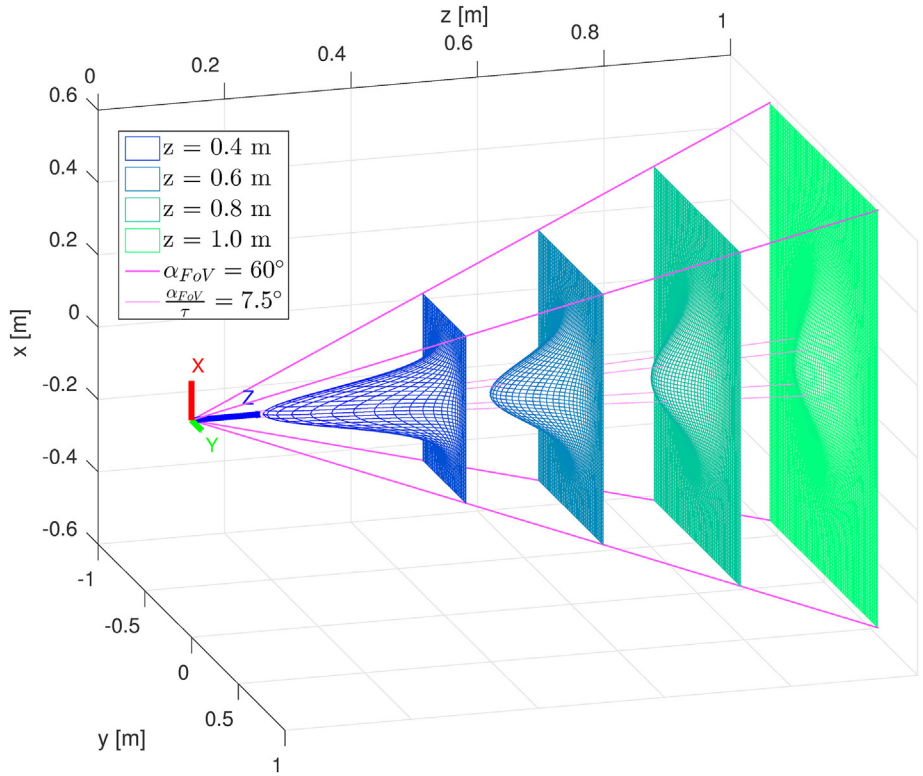

FIGURE 3 | Attention measure $v_a(p)$ computed at different distances from the user's face. Parameter τ is set to 8, decreasing the area where most of the attention is concentrated.

RGB-D camera. The projection of the FOV into the scene and the attention measure allows us to allocate attention over time. In practice, we augment the point cloud $\mathcal{D}$ by adding to each point x_i a cumulative measure of attention over time $V_a(x_i)$. We refer to the augmented point cloud as *attention map*.

Each point x_i starts with attention measure $V_a^1(x_i) = 0$. At each timestep t, we compute the measure of attention $v_a(x_i)$ and we sum this value to the attention measure from the previous timestep as

$$V_a^t(x_i) = V_a^{t-1}(x_i) + v_a(x_i) \qquad \text{if } x_i \in \text{FOV}. \tag{3}$$

As mentioned earlier, if a point is not lying inside the FOV, no attention is allocated to it. The combination of the point cloud $\mathcal{D}$ and the cumulative measure of attention at time t, V_a^t, creates the attention map $\mathcal{A}^t$. Once $\mathcal{A}^t$ is available, we can perform attention analysis over the environment and extract the attention targets automatically, increasing the flexibility and the robustness of our approach.

We model the set of attention targets as a Gaussian Mixture Model (GMM), i.e., a linear combination of Normal distributions. In particular, each component of the GMM will model one target of attention, based on the information stored in the attention map. The choice to use a GMM is supported by the following reasons: (1) the probabilistic nature of the approach makes it robust to noisy readings both from the point cloud and from the head tracking system, (2) the Normal distribution makes the least assumptions on the shape of the object, and (3) once trained, the information about the GMM can be stored in a memory efficient way, as each component has a 3 × 3 covariance matrix and a 3-vector of the mean position.

Our GMM consists of a set of K three-dimensional Normal distributions $\mathcal{N}$. The number of components K can be set *a priori* or chosen with model selection techniques like Bayesian Information Criterion (BIC). Each component $\mathcal{N}j$ of the mixture has its own parameters, i.e., mean μ_j and covariance Σ_j. Additionally, each component has a mixing coefficient π_j that describes the weight of the component in the mixture (with $\sum_{k=1}^{K} \pi_k = 1$). The GMM's parameters are usually learned from data by mean of the Expectation-Maximization algorithm (EM) (Dempster et al., 1977). However, since we want to take into account both the position of the points in the point cloud and their assigned attention, we use a weighted version of the EM algorithm.

EM is an iterative algorithm that performs Maximum Likelihood Estimation (MLE) or Maximum *a Priori* (MAP) estimation of the parameters θ of statistical models with latent variables. It alternates two steps: the expectation step (E) and the maximization step (M). The E step uses the current parameters' estimates and the data to compute the expected complete-data log-likelihood $Q(\theta, \theta^{t-1})$. By maximizing $Q(\theta, \theta^{t-1})$, the M step computes new values for the parameters. The EM algorithm yields to locally optimal parameters.

For GMMs (Bilmes, 1998), the EM algorithm works as follows. In the E step, the responsibility r_{ij} for each data point x_i is computed. r_{ij} is the probability that data point x_i is generated by the j-st component of the mixture. The responsibility r_{ij} is defined as

$$r_{ij} = p(j \mid x_i) = \frac{p(j)p(x_i \mid j)}{p(x_i)} = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}. \quad (4)$$

If $D = \{x_1, \ldots, x_N\}$ is a set of N points, their log-likelihood function $\ln P(D \mid \theta)$ is defined as

$$\ln P(D \mid \theta) = \prod_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k p(x_i \mid \mu_k, \Sigma_k). \quad (5)$$

In the M step, the parameters π_j, μ_j, and Σ_j of each component are updated in order to maximize the likelihood defined in equation (5) as follows

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \quad (6)$$

$$\mu_j = \frac{\sum_{i=1}^{N} r_{ij} x_i}{\sum_{i=1}^{N} r_{ij}} \quad (7)$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} r_{ij} x_i x_i^T}{\sum_{i=1}^{N} r_{ij}} - \mu_j \mu_j^T. \quad (8)$$

In the standard form, EM applied to our problem would only consider the points' locations, regardless of the attention measure they received. To obtain a weighted version of EM, we consider each point x_i with attention measure $V_a(x_i)$ conceptually as a set made of $M = V_a(x_i)$ points x_i with unitary attention measure. As an example, if a point x_i has $V_a(x_i) = 4$, we would instead consider 4 points located in x_i with unitary attention measure. With this trick, we can train the GMM while taking into account the allocated attention. First, we define the probability of a point x_i with cumulative attention measure $V_a(x)$ to be generated by the j-component of the mixture as

$$p(x_i, V_a(x_i) \mid j) = \prod_{w=1}^{V_a(x_i)} \mathcal{N}(x_i \mid \mu_j, \Sigma_j) = \mathcal{N}(x_i \mid \mu_j, \Sigma_j)^{V_a(x_i)}. \quad (9)$$

By plugging equation (9) into standard EM, we can redefine the E and M steps. In the E step, the responsibilities r_{ij} for each component $j = 1 \ldots K$ and data point x_i become

$$r_{ij} = p(j \mid x_i) = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)^{V_a(x_i)}}{\sum_{j=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_j, \Sigma_j)^{V_a(x_i)}}. \quad (10)$$

In the M step, we update the parameters for each component $j = 1 \ldots K$ as

$$\pi_j = \frac{\sum_{i=1}^{N} r_{ij} V_a(x_i)}{\sum_{i=1}^{N} V_a(x_i)} \quad (11)$$

$$\mu_j = \frac{\sum_{i=1}^{N} r_{ij} x_i V_a(x_i)}{\sum_{i=1}^{N} r_{ij} V_a(x_i)} \quad (12)$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} r_{ij} V_a(x_i) x_i x_i^T}{\sum_{i=1}^{N} r_{ij} V_a(x_i)} - \mu_j \mu_j^T. \quad (13)$$

Before running EM, the GMM's parameters must be initialized. We initialized the mean μ of each component with weighted k-means (Kerdprasop et al., 2005). As no prior information on the target's volumes is given, the covariance matrices Σ are initialized with the identity matrix I. The mixing coefficients π are set to $\frac{1}{K}$. In our implementation of EM, we perform all computations in log space in order to avoid the numerical errors.

3.3. Attention Tracking

After the GMM attention model is learned from the attention map, the user's attention can be tracked in real-time, starting from the head pose information and the environment's point cloud representation. First, the probabilities of each point x_i to belong to the different components of the GMM $p(x_i \mid j)$ are computed. If the point cloud is captured once and stay constant during the attention tracking phase, the aforementioned probabilities can be computed once to enhance the real-time performances of the algorithm. This results in a vector containing the K probabilities $\{p(x_i \mid 1), \ldots, p(x_i \mid K)\}$ that have been assigned to each point. Second, at each time step, the user's FOV is computed from the head pose measurements and projected into the point cloud. As for the attention allocation phase, each point lying inside the FOV pyramid is given the attention measure v_a as explained in Section 3.1. The only difference is that in this case points do not accumulate attention over time, on the contrary their level of attention is reset at the end of each iteration. Finally, we compute the probability for each component j of the mixture to explain the computed attention map, i.e., $p(j \mid \mathcal{A}^t)$.

First, we compute the likelihood of the attention map $\mathcal{A}^t$ to be generate by each of the GMM's j-st component as

$$p(\mathcal{A}^t \mid j) = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu_j, \Sigma_j)^{V_a^t(x_i)}. \quad (14)$$

By using Bayes' rule, we can obtain the desired $p(j \mid \mathcal{A})$ as

$$p(j \mid \mathcal{A}^t) = \frac{p(\mathcal{A}^t \mid j)p(j)}{p(\mathcal{A}^t)} = \frac{\pi_j \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu_j, \Sigma_j)^{V_a^t(x_i)}}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu_k, \Sigma_k)^{V_a^t(x_i)}} \quad \text{for each j} = 1 \ldots K. \quad (15)$$

The result of equation (15) is a vector containing K probabilities, where each element represents the likelihood of GMM's component j to be observed at time t by the user. With this result, we can track the user's attention over the targets in the environment.

4. RESULTS

4.1. Implementation

We implemented the entire framework with ROS as middleware. For the head pose tracking, we rely on OpenFace,[1] an opensource tool for facial behavior analysis (Baltrušaitis et al., 2016). The scene was observed by a Microsoft Kinect One. High-resolution images (1,280 × 1,024) were stored along with the computed head poses. Such images are used to determine the ground truth data regarding the person's real attention. The point cloud image of the environment is taken by the same sensor with a resolution of 640 × 480 before the experiment.

4.2. Experiment

The aim of the experimental study is to assess whether objects of interest can be automatically segmented from observing a person's visual attention. To evaluate this, a scenario is devised that directs a person's attention to different areas on the scene in such a way that it can be observed by an external camera. We consider a person placed in a working environment engaged in typical office tasks. In particular, these include reading a sheet of paper, writing on the sheet of paper, working with a computer and studying an object set on the desk. This scene therefore consists of three big objects (i.e., screen, paper, and building blocks object) and several smaller targets (i.e., mouse, keyboard, pen) as detailed in **Figure 4**. All objects, except the pen and the mouse are to remain stationary on the table. The subject is placed sitting in front of the desk and follows the instructions written in the sheets of paper. No external distractions are generated on purpose and the experiment is set up to last 5–10 min.

The sheets of paper contain 12 questions requiring written answers. Five questions have to be answered by searching for the answers on the internet, five questions require studying the building blocks object and two questions require the reading and writing of a paragraph of text (separately). The gaze shifts between the computer and the paper simulate the activity of studying both from a book and from a computer. The building blocks object is made of several components, varying in color and dimensions. The object's shape was chosen to not replicate the parallelepiped-shape of the screen and the sheets of paper. Questions regarding the building blocks object query the composition of the blocks (e.g., colors, number, and lengths).

All three targets, shown in **Figure 4**, are scattered on the desk: the screen on the left, the building block object on the right, and the question sheet in front of the subject. The presence of multiple and scattered objects allows the person to shift his/her head between the various targets. Such aspect is useful to validate in particular the attention tracker. Furthermore, objects are deliberately characterized by various shapes, to test the accuracy of the GMM in estimating the targets' position and volume.

We conducted the experiment with 4 participants between 20 and 30 years old. The participants were not instructed about the operating principle of the system (e.g., the fact that their head pose was tracked) and were not aware of our research questions.

[1] https://www.cl.cam.ac.uk/tb346/res/openface.html.

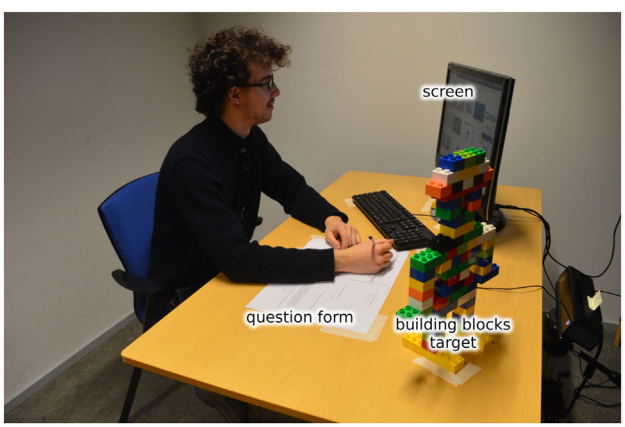

FIGURE 4 | Experimental setup.

4.3. Experimental Results

We separate the validation of our framework in two experiments. First, we evaluate the quality of the created attention map and the segmentation of the targets of attention. Second, we evaluate the online performance of the attention tracking. Following, we evaluate how well head pose is suited to detect and track attention.

4.3.1. Attention Map Evaluation

To evaluate our model for the targets of attention, we compare each component of the GMM with manually selected ground truth data. In particular, we manually segmented the point cloud in three set of points $S_{i,truth}$, one for each target i of the experiment. Based on these sets, we fitted ground truth normal distributions $N_{i,truth}$, one for each set $S_{i,truth}$. The FOV aperture α_{FOV} was set to 60°. The τ parameter was chosen to be 8, to narrow the area where we assume most of the eye gazes are concentrated. The attention model was built on the measurements from the whole experiment.

We evaluate the estimated position of the targets of attention. We compare the mean μ of the estimated components of our models with the mean of the ground truth normal distribution $N_{i,truth}$. We learned 4 attention models, one for each participant. **Figure 5** presents the boxplot for the displacement between the means, expressed in meters. The smaller error is obtained for the screen and the building block object, since the participants were concentrating their attention on the center of these objects. The questions sheet target achieved the worst results. The higher error is caused by the relative position of the sheets with respect to the user. We observed that participants preferred to move their eyes instead of their head in order to read the questionnaire. **Figure 6** shows the estimated components of the GMM for participant 2 and the ground truth Normal distribution $N_{i,truth}$, overlapped with the point cloud.

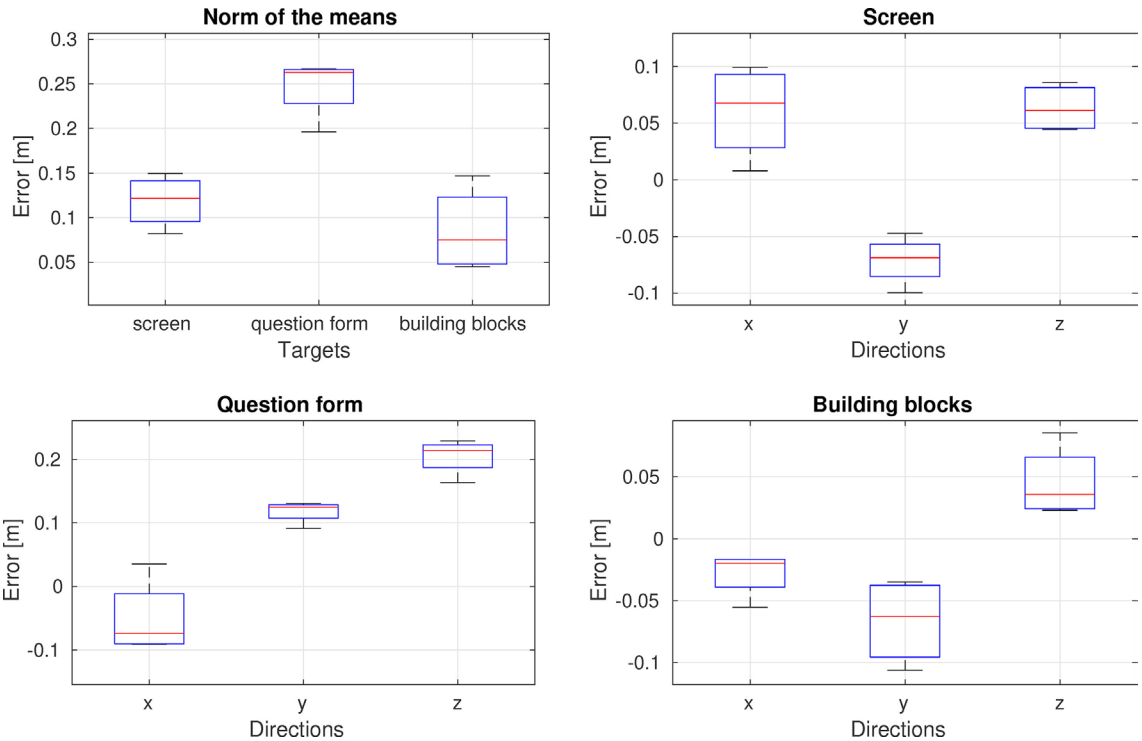

FIGURE 5 | Displacement between the estimated targets of attention and the ground truth normal distributions $N_{i,truth}$.

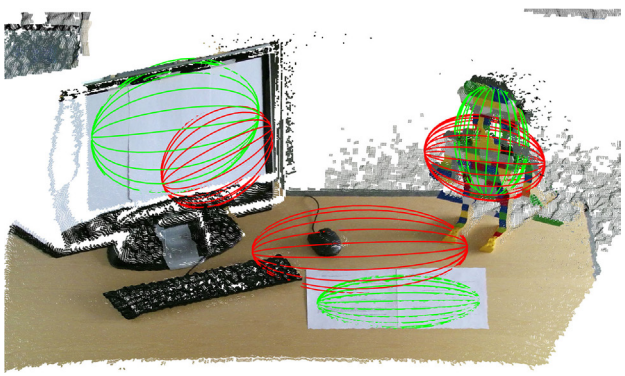

FIGURE 6 | Representation of the estimated components of the GMM (ellipsoids in red) and the ground truth normal distributions $N_{i,truth}$ (ellipsoids in green), overlapped with the point cloud.

The modeling of the targets of attention resulted to be accurate. In particular, the screen and the building block object were correctly estimated within the environment. Despite the fact that objects with no pseudo-ellipsoidal shape may not be estimated correctly with Normal distributions, this choice of distribution is in our view best when considering no prior information on the shape of the object.

4.3.2. Attention Tracking Evaluation

We evaluate the attention tracking capabilities of our model by comparing them against annotated data from video recordings of the experiments. The parameters α_{FOV} and τ were set as in the previous experiment. For each participant, the target of attention model (i.e., the GMM) was trained on the measurements captured during the first 60 s of the experiments. The duration of this initial phase was chosen in order to include significant glances to each target. During the rest of the experiment, the users' attention was tracked according to the built model. At each timestep t, we compute the probabilities of each component of the GMM to explain the current attention map, with equation (15).

We compared the estimated tracking of attention $p(j \mid \mathcal{A})$ to the annotated tracking data. In particular, we computed the tracked target as the $\text{argmax}_j\, p(j \mid \mathcal{A})$ and then compare it with the annotated target. **Table 1** presents several tracking measures, separated for each participant. First column presents the percentage of tracking errors (i.e., when the estimated target of attention does not match the annotated data). For this measure, we did not consider failures of the head pose tracking system

TABLE 1 | Attention tracking performance measurements.

	Attention tracker failures (%)	Head tracker failures (%)	Framework failures (%)
Participant 1	11.78	44.79	50.04
Participant 2	9.20	16.95	20.02
Participant 3	4.00	0.13	4.01
Participant 4	10.02	5.39	13.54
Mean	8.75	16.82	21.90

(i.e., column two) as failures of our method. The attention tracker was affected by error for maximum 12% of the experiment duration. The third column presents the percentage of framework failures (i.e., when our pipeline fails to detect the correct target of attention). **Figure 7** shows the comparison between estimated and ground truth targets of attention for participant two. As can be seen, errors occur mainly due to the gaze shift between two targets.

Additionally, we assess our attention estimation approach by considering Cohen's κ coefficient, which measures the inter-rater agreement, between our estimate and the ground truth. The Cohen's κ values can be seen in **Table 2**. The calculated agreement values exclude head pose tracking failures, and are moderate to good (between 0.60 and 0.96), depending on the participant.

4.3.3. Head Pose As Measure of Attention

Selecting head pose as directed attention is not a standard in human–robot interaction, as gaze is often a preferred measure (Rayner, 1998; Doshi and Trivedi, 2010). Besides the difficulty in tracking human gaze from real-world scenarios (e.g., resolution, mobility of participants), we see the following benefit for tracking the head pose. Human gaze results often from fast eye saccades, abruptly changing the point of fixation in the scene. While the continuous tracking of this would allow to deduce exactly what a person is looking at, it does not necessarily lead to acquiring human attention. Aligning our head to our direction of gaze, on the other hand, follows after something of interest is found that requires more attention (Doshi and Trivedi, 2012). For example, the decoupling between head pose and gaze occurred around 4% of the time (around 15 s) in the experimental scenario of participant 3. This misalignment happened when looking at the keyboard while keeping the head oriented toward the screen and when looking at the building blocks object, while keeping the head oriented toward the question form. Considering the head pose tracker, in the worst case (participant 1) tracking failed for 45% of the experiment. Despite this high failure rate, our system proved to be robust and could track attention. Our results are in accordance with (Stiefelhagen, 2002) and (Lemaignan et al., 2016) which report a high agreement between attention from head pose and ground truth data. Considering this, as well as the results in the previous section, we therefore conclude that head pose is suitable for inferring human attention.

5. DISCUSSION

As shown by the experimental results, our method can detect and track human visual attention to objects in a scene well. As main input, it relies on a current head pose estimate and a pre-recorded 3D point cloud of the environment. Estimating human attention from head pose instead of gaze turns out to be a suitable choice, as proven by our results and as pointed out in related work, e.g., Stiefelhagen (2002) and Lemaignan et al. (2016). The main limitation we see in using gaze is the limitations due to the sensor itself. Wearable glasses, as used in Kassner et al. (2014), hinder the natural behavior of the user, and cameras are limited with respect to range, due to the relative dimension of the eyes in the image. Regardless, future work will focus on a direct comparison between gaze and head pose as attention measure.

Regarding the attention allocation and the modeling of the targets of attention, our work distinguishes from others due to its probabilistic nature. Attention to the environment is distributed as a 2D Normal distribution, whereas works such as Lemaignan et al. (2016) and Anzalone et al. (2015) do not consider such distribution, when using head pose as attention estimate. Doshi and Trivedi (2010) do consider a probabilistic approach for attention allocation, however, use gaze as input and do not model the target. Similarly, modeling the attentional targets with a Gaussian Mixture Model (GMM) is not considered in other related work. The major benefit of

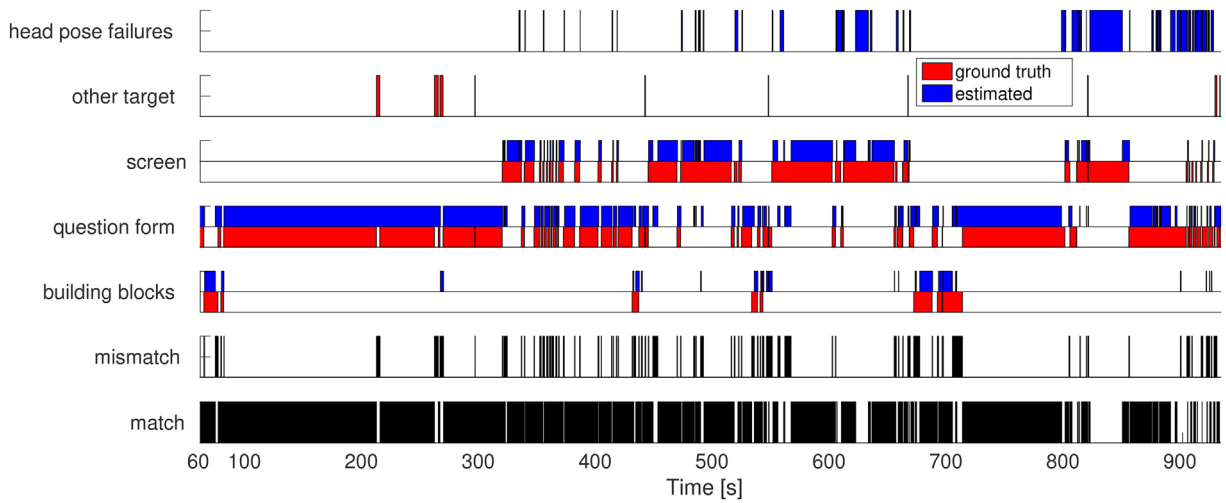

FIGURE 7 | Comparison of estimated and ground truth targets of attention for the Participant 2. The head pose tracker's failures are not considered.

TABLE 2 | Attention tracking agreement: Cohen's κ.

	Screen	Question form	Building blocks	Total without head pose failures	Total with head pose failures
Participant 1	0.74	0.60	0.73	0.70	0.42
Participant 2	0.86	0.80	0.68	0.85	0.77
Participant 3	0.96	0.92	0.86	0.93	0.93
Participant 4	0.86	0.75	0.79	0.85	0.81
Mean	0.86	0.76	0.77	0.84	0.73

the probabilistic nature of our approach is the robustness to outliers and head pose measurement failures. Moreover, as targets can be unknown, using a Normal distribution has the advantage of making the least assumptions on the shape of the object.

The choice of using a GMM to model the attentional targets comes with a limitation. A Normal distribution may not be the best representation of an object. Despite this issue, this choice of distribution is in our view best when considering no prior information on the shape of the object. In this work, our approach only considered static objects in the scene. This could be extended to dynamic objects as incremental ways of training GMMs are available (Calinon and Billard, 2007) and could be integrated in our method.

In our case, each individual object in the scene is related to an action. The engagement of the person with an object implies the action allocated to that object. This implies attention toward the computer means working with the computer, attention toward the building blocks means studying the building blocks object, and attention to the paper sheets means reading or writing. Attention directed elsewhere does not account to any action. In the evaluation of the scenario, it was shown that in most part (i.e., 88% of the time) the estimation is correct in assigning human attention to an object. A one-to-one relation between attention and action is, however, not always true. Indeed, a person can be staring at the sheet of paper and thinking inside their head and not be reading the words on the paper at all. In our view, these cases are very difficult, if not impossible to perceive, and we assume that, in general, most engagement with objects does account for the actions they are meant for. As one of the outcomes in this work, we recognize that observations of attentional targets alone are not sufficient to deduce actions (and intentions). Without knowing the object, the scenario or the context (i.e., prior knowledge) it is impossible to understand that directing attention to an object indicates a certain action or intention.

6. CONCLUSION

This work aimed to study human visual attention and its relationship to detecting objects and actions. Human visual attention is modeled in a probabilistic way in order to take into account the inaccuracy of using head pose as measure of attention. A weighted Normal distribution, centered around the head pose, assigns weight to a representation of the environment, creating an attention map. On the attention map, a model is built for the attentional targets, namely, a GMM. This allows to track attention on these objects in real-time. The approach only requires a single RGB camera for head pose measurements and a pre-recorded 3D point cloud of the environment. A user study shows the successful segmentation of objects and detection of attention in a typical office task scenario (i.e., reading, operating a computer, studying a toy). The tracking of attention on objects allows for directly allocating actions to objects. That is, directing attention to an object might imply the action that is associated to that object.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Aalto University ethics board with verbal and written consent from all adult subjects. Treatment of all subjects was in accordance with the Declaration of Helsinki. Written consent was received for publishing the images of the subject in the manuscript.

AUTHOR CONTRIBUTIONS

MR, RP, and VK designed the study. AV, MR, RP, and VK developed the methods. AV, MR, and RP performed experiments and analyzed the data. AV, MR, RP, and VK wrote the paper.

ACKNOWLEDGMENTS

We thank all members of the ROSE (Robots and the Future of Welfare Services) project for helpful discussion and ideas. We are grateful to all participants who took part in the experimental scenarios in order to develop and evaluate the work.

FUNDING

This work was supported by the Academy of Finland, Strategic Research Council (project: "Robots and the Future of Welfare Services," decision 292980).

REFERENCES

Anzalone, S. M., Boucenna, S., Ivaldi, S., and Chetouani, M. (2015). Evaluating the engagement with social robots. *Int. J. Soc. Robot.* 7, 465–478. doi:10.1007/s12369-015-0298-7

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision (WACV)* (Lake Placid, NY: IEEE), 1–10.

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Int. Comput. Sci. Inst.* 4, 126.

Calinon, S., and Billard, A. (2007). "Incremental learning of gestures by imitation in a humanoid robot," in *2nd ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, VA), 255–262.

Chadalavada, R. T., Andreasson, H., Krug, R., and Lilienthal, A. J. (2015). "That's on my mind! Robot to human intention communication through on-board

projection on shared floor space," in *European Conference on Mobile Robots (ECMR)* (Lincoln, UK: IEEE), 1–6.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philos. Trans. R. Soc. B Biol. Sci.* 362, 679–704. doi:10.1098/rstb.2006.2004

De Santis, A., Siciliano, B., De Luca, A., and Bicchi, A. (2008). An atlas of physical human-robot interaction. *Mech. Mach. Theory* 43, 253–270. doi:10.1016/j.mechmachtheory.2007.03.003

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B Methodol.* 39, 1–38.

Dominey, P. F., and Warneken, F. (2011). The basis of shared intentions in human and robot cognition. *New Ideas Psychol.* 29, 260–274. doi:10.1016/j.newideapsych.2009.07.006

Doshi, A., and Trivedi, M. M. (2010). "Attention estimation by simultaneous observation of viewer and view," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (San Francisco, CA), 21–27.

Doshi, A., and Trivedi, M. M. (2012). Head and eye gaze dynamics during visual attention shifts in complex environments. *J. Vis.* 12, 9–9. doi:10.1167/12.2.9

Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., and Axelrod, B. (2013). Toward understanding social cues and signals in human-robot interaction: effects of robot gaze and proxemic behavior. *Front. Psychol. Cogn. Sci.* 4:859. doi:10.3389/fpsyg.2013.00859

Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., et al. (2016). Hobbit, a care robot supporting independent living at home: first prototype and lessons learned. *Rob. Auton. Syst.* 75, 60–78. doi:10.1016/j.robot.2014.09.029

Gazzola, V., Rizzolatti, G., Wicker, B., and Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage* 35, 1674–1684. doi:10.1016/j.neuroimage.2007.02.003

Goodrich, M. A., and Schultz, A. C. (2007). Human-robot interaction: a survey. *Found. Trends Hum. Comput. Interact*, 1, 203–275. doi:10.1561/1100000005

Gulzar, K., and Kyrki, V. (2015). "See what I mean-probabilistic optimization of robot pointing gestures," in *IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)* (Seoul: IEEE), 953–958.

Herath, S., Harandi, M., and Porikli, F. (2017). Going deeper into action recognition: a survey. *Image Vis. Comput.* 60, 4–21. doi:10.1016/j.imavis.2017.01.010

Kassner, M., Patera, W., and Bulling, A. (2014). "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (Seattle, WA: ACM), 1151–1160.

Kerdprasop, K., Kerdprasop, N., and Sattayatham, P. (2005). "Weighted k-means for density-biased clustering," in *International Conference on Data Warehousing and Knowledge Discovery* (Copenhagen: Springer), 488–497.

Leite, I., Martinho, C., and Paiva, A. (2013). Social robots for long-term interaction: a survey. *Int. J. Soc. Robot.* 5, 291–308. doi:10.1007/s12369-013-0178-y

Lemaignan, S., Garcia, F., Jacq, A., and 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* (Christchurch), 157–164.

Loth, S., Jettka, K., Giuliani, M., and de Ruiter, J. P. (2015). Ghost-in-the-machine reveals human social signals for human-robot interaction. *Front. Psychol.* 6:1641. doi:10.3389/fpsyg.2015.01641

Margoni, F., and Surian, L. (2016). Explaining the u-shaped development of intent-based moral judgments. *Front. Psychol.* 7:219. doi:10.3389/fpsyg.2016.00219

Mast, M., Burmester, M., Graf, B., Weisshardt, F., Arbeiter, G., Španěl, M., et al. (2015). "Design of the human-robot interaction for a semi-autonomous service robot to assist elderly people," in *Ambient Assisted Living*, eds R. Wichert and H. Klausing (Berlin, Heidelberg: Springer), 15–29.

Michalos, G., Makris, S., Spiliotopoulos, J., Misios, I., Tsarouchi, P., and Chryssolouris, G. (2014). Robo-partner: seamless human-robot cooperation for intelligent, flexible and safe operations in the assembly factories of the future. *Proc. CIRP* 23, 71–76. doi:10.1016/j.procir.2014.10.079

Murphy-Chutorian, E., and Trivedi, M. M. (2008). "3D tracking and dynamic analysis of human head movements and attentional targets," in *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)* (Stanford), 1–8.

Palinko, O., Rea, F., Sandini, G., and Sciutti, A. (2016). "Robot reading human gaze: why eye tracking is better than head tracking for human-robot collaboration," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on* (Daejeon: IEEE), 5048–5054.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vis. Comput.* 28, 976–990. doi:10.1016/j.imavis.2009.11.014

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372. doi:10.1037/0033-2909.124.3.372

Sheikhi, S., and Odobez, J.-M. (2015). Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognit. Lett.* 66, 81–90. doi:10.1016/j.patrec.2014.10.002

Sisbot, E. A., Ros, R., and Alami, R. (2011). "Situation assessment for human-robot interactive object manipulation," in *RO-MAN, 2011 IEEE* (Atlanta, GA: IEEE), 15–20.

Stiefelhagen, R. (2002). "Tracking focus of attention in meetings," in *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces* (Pittsburgh, PA), 273–280.

Vernon, D., Thill, S., and Ziemke, T. (2016). "The role of intention in cognitive robotics," in *Toward Robotic Socially Believable Behaving Systems – Volume I. Intelligent Systems Reference Library,* eds A. Esposito and L. Jain, Vol. 105 (Cham: Springer).

Vishwakarma, S., and Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* 29, 983–1009. doi:10.1007/s00371-012-0752-6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.