Joint Attention by Gaze Interpolation and Saliency

Zeynep Yücel* *Member, IEEE,* Albert Ali Salah† *Member, IEEE,* Çetin Meriçli‡ *Member, IEEE,* Tekin Meriçli† *Member, IEEE,* Roberto Valenti§ Theo Gevers§ *Member, IEEE,*
* ATR International
Intelligent Robotics and Communication Laboratories, Japan
† Department of Computer Engineering,
Boğaziçi University, İstanbul, Turkey
‡ Computer Science Department
Carnegie Mellon University, USA
§ Intelligent Systems Lab Amsterdam
University of Amsterdam, The Netherlands
Email: zeynep@atr.jp, {salah,tekin.mericli}@boun.edu.tr, cetin@cmu.edu, {rvalenti, th.gevers}@uva.nl

***Abstract*—Joint attention, which is the ability of coordination of a common point of reference with the communicating party, emerges as a key factor in various interaction scenarios. This paper presents an image-based method for establishing joint attention between an experimenter and a robot. The precise analysis of experimenter's eye region requires stability and high resolution image acquisition, which is not always available. We investigate regression-based interpolation of the gaze direction from the head pose of the experimenter, which is easier to track. Gaussian process regression and neural networks are contrasted to interpolate the gaze direction. Then we combine gaze interpolation with image-based saliency to improve the target point estimates and test three different saliency schemes. We demonstrate the proposed method on a human-robot interaction scenario. Cross-subject evaluations, as well as experiments under adverse conditions (such as dimmed or artificial illumination or motion blur) show that our method generalizes well and achieves rapid gaze estimation for establishing joint attention.**

***Index Terms*—Joint visual attention, head pose estimation, gaze following, saliency, developmental robotics, selective attention.**

I. Introduction and Motivation

Natural human-robot interaction requires the design and implementation of robot skills that enable adaptive behaviors in ways that mimic human-human interaction. In that respect, the perspective taken by developmental robotics adopts a mutual standpoint of developmental psychology and robotics, where the training of the robot follows a natural scheme [1]. Hence, automatic interpretation and imitation of human activity and behavior, as well as spontaneous generation of correct responses, emerge as key factors. These require *joint attention*, which needs to be established with the limited resources (i.e. sensors and computation) of a robotic agent.

Joint attention, which is the ability of coordination of a common point of reference with the communicating party [2], emerges in early stages of human social cognition, providing a basis for the development of further social skills and triggering imitation-based learning [3]. Subsequently, real-time tracking of the focus of attention of the communicating party (henceforth called *the experimenter*) and gaze following are crucial skills for the emergence of certain types of imitation [4]–[6].

In this paper, we propose a system to achieve joint attention when the visual resources of the robotic system are limited to a low resolution camera, which is a reasonable assumption considering the specifications of some state-of-the-art robotic platforms such as Nao or iCub. Joint attention is suggested to be achieved by estimating the point attended by the experimenter, namely the *target point*, and directing the focus of the robotic agent towards that point. To that end, the *gaze direction* needs to be determined, which we define as the direction that towards which the experimenters attention is steered. Gaze direction is proposed to be obtained by the composite motion of head and eyes, whereas the *head pose* is defined by the position and orientation of experimenter's head in object space. We use the term *eyeball orientation* to refer to the position and orientation of the eye within the head model. Due to the limitations in the hardware configuration of the robotic agent, eye region of the experimenter does not always provide sufficient information to estimate the orientation of the eyeball. Thereby, gaze direction is proposed to be interpolated from head pose only. Although head pose and gaze direction are suggested to be closely linked [7], our study underlines the fact that the head direction is not a substitute for gaze direction.

Most approaches to gaze direction estimation assume a calibration stage where the experimenter looks at certain predetermined points that serve as anchors for interpolation [8]. We assume here that there is no calibration stage. Instead, the structure of a visual scene provides additional cues to the embodied agent in guessing the focus of attention of the communicating party, namely the target point. We use *saliency* to detect objects of interest within the estimated gaze direction of the experimenter. The gaze vector itself traverses the whole scene in one direction, and whenever it goes through several locations of interest, additional information is necessary to resolve this conflict. We show that the head pose contains cues relating to the distance of the object of interest. These cues will be employed to further constrain the localization problem.

To summarize, the contributions of this work are the following:

- We propose to learn gaze direction from head pose esti-

mates via regression. We use Gaussian process regression and neural networks for the interpolation.
- We compare three different saliency schemes to improve estimation of target points.
- We systematically test the effects of resolution, changing illumination, and motion blur.

Our proposed system is tested on data collected from a Nao robot in realistic settings. In addition to establishing joint attention, estimating the gaze direction from camera input is relevant for different domains, such as human robot interaction [9], driver awareness [10], [11], attention tracking for meetings [12], and communication with virtual agents [13].

The outline of the paper is described as follows: Section II gives an overview of the related work in head pose and gaze estimation for embodied agents. Section III describes the proposed method. Section IV presents the head pose estimation module, based on tracking with a cylindrical head model (CHM), followed by Section V that describes gaze direction and object depth estimation via Gaussian process regression (GPR). In Section VI, the saliency-based fine tuning of the focus of attention is described. The proposed method is evaluated and discussed in Section VII, followed by our conclusions in Section VIII.

II. Related Work

The primary subtasks of the joint attention are recognition, maintenance of eye contact, and gaze following, which in tandem enable getting engaged in joint attention [14]. In natural settings, cues like imperative and declarative pointing are also considered to permit feedback between the interacting parties. While being first and foremost an image processing challenge (from a practical perspective), joint attention is a particularly important skill in early development, and as such, it has received interest from the developmental robotics community. Related findings of developmental psychology suggest that in order to establish social contact and fulfill the desire for knowledge, infants get engaged in communication and hence obtain joint attention with the caregivers. These social skills are observed to improve gradually at primary stages of infancy. Developmental psychologists have established that humans use head pose in estimation of focus of attention. It appears that young infants first follow the head movements of others, and only in time develop the ability to follow the gaze direction [4]. The perception of gaze direction depends to a large extent on head pose [15]. Subsequently, infants learn to relate this information with attention [16].

Begum *et al.* provide a detailed survey on computational models of visual attention for robot cognition [17]. In this work, we use operational constraints imposed by a robotic platform, and focus on related work with similar assumptions. For gaze estimation approaches that use detailed eye region analysis, refer to [8].

Earlier approaches to the problem focus on the developmental aspects of joint attention, and propose models with a biological motivation. For instance, in [18] a neural network is employed for modeling the visual system of a robot, where there are different layers representing the input, retina, visual cortex, and output. Since the problem of establishing joint attention is particularly difficult under these additional constraints, the contextual information must be employed. In [19], a visual attention module is used to detect salient objects in the robot's view, and attempts for joint attention are rectified by closing the sensorimotor loop. If the attended location is erroneous, the subsequent actions will be unsuccessful. The gaze direction of the experimenter is learned from the face image with a back-propagation neural network. The drawback of this approach is that the face image is a high-dimensional and complex representation, which necessitates large numbers of training samples for appropriate generalization. The approach resembles the method we propose in this paper, in that no special processing is used for the eye regions.

Baluja *et al.* proposed a neural network regression method to interpolate gaze direction from a high resolution face image [20]. In a similar approach, a simple local linear interpolation technique based on a Gaussian model assumption integrated with head pose estimation is proposed by Sugano *et al.* [21]. In [22], Sugano *et al.* assume that the head pose is fixed throughout the experiment and observe the high resolution eye region to establish the relationship of the gaze with the saliency of the scene viewed by the experimenter. In [23], Chen and Ji describe a method that, like the method we propose, requires no personal calibration. In contrast to our proposed approach, all these methods assume that the eye region can be cropped with a high resolution.

Most approaches proposed for gaze estimation make use of video or camera input, yet treat video frames as individual images and omit the temporal connection. Humans, on the other hand, utilize motion information along with static information, such as posture and face direction, to infer about desires and intentions. For this reason, the robotic agent described in [24] alternates its gaze between a human experimenter and the attended object by triggering motion actions, using the cues derived from the motion of the experimenter's face. In [25] the temporal relationship between subsequent frames is expressed in terms of optical flow vectors, and thus provide a coarse estimate for gaze shift. In our case, the effect of motion is negligible, as the objects are static at all times and motion of the camera affects all objects uniformly due to their close positioning.

Studies presented in [24] and [25] focus on available 2D information for formulating visual attention based on video frames. However, common morphological characteristics of faces can be employed in the derivation of 3D information from the 2D visual input. Since the perception of gaze direction depends to a large extent on the pose of the head [15], one can model the head of the experimenter as a 3D object and resolve for the pose [26]. This is the approach we take in this paper. In a similar vein, Hoffman *et al.* employ an ellipsoidal model for the human head and infer head angles for the estimation of the head pose vector [27]. This vector is also used as the estimated gaze vector. Saliency computation is performed around the estimated gaze, but some experimenter-specific priors are incorporated into the model, according to which each experimenter has different (but learnable) tendencies to look at certain objects [28].

The relation between head pose and gaze is further explored in [29], where a distinction is made between head movers and non-head movers. Given a target object, the head movers will orient their gaze by rotating the head towards the object, and the non-head movers will keep their head fixed and move their eyes only. These are two extremes of a continuous range of behaviors. A model based on this distinction is used in [30] to synthesize realistic head-gaze movements for an embodied conversational agent. Stahl provides empirical results regarding the range of customary ocular range and the onset of head movements for compensation of post-saccadic eye eccentricity to obtain a comfortable focus [7].

In [27] and [31] Bayesian principles are used to explore action spaces statistically, followed by gradual learning of action groups and communicative preferences. These approaches are based on Meltzoff and Moore's active intermodal mapping framework [32], which offers a theoretical basis for imitation based learning. In this paper we do not assume any communicative preference, and there are no object-specific prior probabilities that can help the system decide on the focus of attention. Our attention-based approach mimics early stages of infancy (mainly 6 to 12 months) used in robot learning and developmental robotics.

III. Outline of the Method

Our experimental setup follows a real robotic scenario proposed by Hoffman *et al.*, where an experimenter and a robot are facing each other over a table that contains several objects of interest [27]. This setup is typical in robotic joint learning scenarios (e.g. [9]).

We use a Nao robot, and observe that it is not sufficiently stable to extract an accurate estimate of the gaze direction directly by analyzing the eye and iris areas of the experimenter. The assumption that the camera input does not provide sufficient resolution for the analysis of the eye region is realistic for many application settings (including the recently popular interactive marketing scenarios), and it is imperative to extract the maximum amount of gaze information, even when this is the case.

The question of what resolution is appropriate for accurate gaze estimation is hard to answer. The most extensive survey about gaze estimation to date is [8], and while over a hundred methods are reviewed in this survey, the exact resolution ranges are not made explicit for any of them. There is a reason for this; external factors like camera movement, head movement, and tracking error can have strong influence on the minimum resolution. This issue obviously needs more research. Our analysis is based on the very practical observation that the iris area contains only a few pixels in our setup as described in Section VII-A, and no eye-region based method is applicable in such a low resolution setting.

Fig. 1 illustrates eye regions of different experimenters cropped from the camera input of the Nao robot used in our system (see Section VII). These regions have approximately 15×25 pixel resolution, and it is clear from Fig. 1 that one cannot make a reasonable estimate for gaze direction by just using the eye and iris area information of these patches.

(a) (b)

Fig. 1. Eye regions for different experimenters, acquired by the robot. The images are approximately 15×25 pixels.

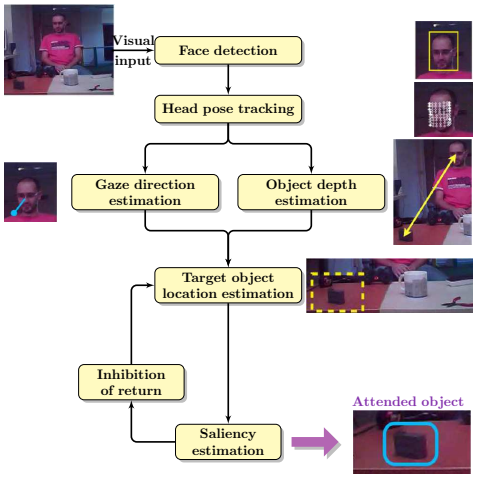


Fig. 2. Basic steps of the algorithm.

Additionally, non-frontal head pose affects the appearance of the eye region drastically.

Our proposed approach mimics the natural strategy of resolution of focus of attention observed in infants. The resolution of attention fixation points relies substantially on gaze direction, which depends to a large extent on head pose [15]. Subsequently, we seek to determine the gaze direction from head pose estimates, by noting that head pose is indicative of (but not equal to) the gaze direction. As a precaution, we note here that this is an ill-posed problem. It is well documented that the temporal alignment of eye and head movements is not strict in humans, as humans are capable of controlling them separately. However, in natural settings, there is a certain eye-head coupling [33], [34], and it is this coupling we seek to exploit in this paper.

The basic steps of the proposed algorithm are given in Fig. 2. The first step is detecting the face of the experimenter. The head pose of the experimenter is then resolved by adapting a 3D elliptic cylindrical model to the face region. By applying a pose update, derived with the Lucas-Kanade optical flow method, a continuous tracking is maintained [35].

Two Gaussian process regressors are employed in estimation of gaze direction and the distance of the target object along the gaze vector, from these pose values. These can be conceptualized as horizontal and vertical displacement estimators, respectively. The two estimates are then probabilistically combined to yield a coarse estimate for the center of the object of interest. By pooling a number of estimates regarding consecutive frames, a more robust decision on the target point is generated. The rough localization of the attended object is refined by a bottom-up saliency scheme. Additionally, if

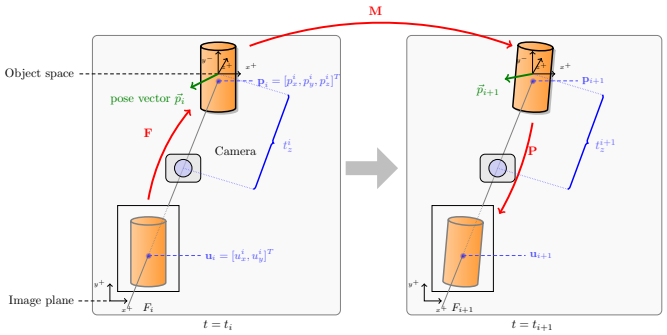


Fig. 3. Projection and transformations illustrated for two consecutive frames in which the head moves and tilts, together with the reference frames of object space and image plane.

the experimenter continues to maintain a certain head pose, alternative target locations are eventually explored as a result of an inhibition-of-return mechanism.

We stress the distinction of following the head pose and the gaze direction itself in particular. Most of the joint attention approaches in the literature do not explicitly correct for the discrepancy between the head pose and gaze direction, which is reported to be normally distributed with a mean of five degrees in natural settings [36], [37]. In computer graphics literature, there are some recent approaches that model this discrepancy explicitly, in order to synthesize virtual agents with more natural behavior [30], [38]. In the following sections we describe the steps of the proposed method in more detail.

IV. Head Pose Estimation

This section elaborates on the head tracking and pose estimation algorithm. The human head is modeled as an elliptic cylinder [39], with the actual width of the head and the radii in line with anthropomorphic measures [40]. The 3D head model is superposed on the detected face area found by the Viola-Jones algorithm which uses the Adaboost classifier with Haar wavelet features [41]. We require a joint attention session to be initialized with the experimenter facing the robot. All head pose angles are estimated with respect to the initial pose, and since the method uses actual pixel values, it is not affected by alignment of the cylinder to a non-frontal pose.

The pose for the cylindrical head model at frame F_i is represented by a vector $\vec{p}_i$. This vector is basically a collection of pitch, roll, and yaw angles (r_x^i, r_y^i, r_z^i) and translation parameters at the i^{th} frame :

$$\vec{p}_i = [r_x^i, r_y^i, r_z^i, t_x^i, t_y^i, t_z^i]^T. \tag{1}$$

The initial values for these parameters are determined by employing the initialization condition of a session of joint attention. In our scenario, we describe these conditions as the establishment of eye contact between the agent and the experimenter with a fully frontal view of the experimenter's face. This is a much more natural initialization assumption compared to explicit gaze calibration procedures. The rotation parameters are all set to 0 for the initial frame. The translations along $\mathbf{x}-$ and $\mathbf{y}-$axes with respect to the object space reference frame are determined by using the center of the face region as determined by the Haar classifier. The depth of the head, t_z^0, which describes the distance of the head from the camera, is set to an approximate fixed value. In practice, this value can be derived by relating the anthropomorphic measures concerning the height and width of the head to the face region found on image plane.

As soon as the initialization condition is satisfied, the agent sets the head pose parameters as described, and starts tracking the head. Since the initial pose, $\vec{p}_0$, is already determined, any pose value $\vec{p}_{i+1}$, $i \geq 0$, can be resolved by simply updating the previous value $\vec{p}_i$ [39],

$$\vec{p}_{i+1} = \vec{p}_i + \Delta\vec{\mu}_i, \tag{2}$$

where $\Delta\vec{\mu}_i = [\omega_x^i, \omega_x^i, \omega_x^i, \tau_x^i, \tau_x^i, \tau_x^i]^T$ stands for the rigid motion vector summarizing the pose update between time instants t_i and t_{i+1}. The efficacy of this dynamic approach is demonstrated in [42] in comparison to a variation of the cylindrical model based head pose estimation scheme from a fixed template image and it is concluded that the dynamic template yields better results. [42] also reports detailed head pose estimation accuracy results.

We note here that a 2D image (of the experimenter's head) is employed to derive pose values representing the direction and orientation in 3D space. In order to cope with the ambiguity ensuing from the dimensionality disparity, a suitable mapping needs to be defined. To that end, we use perspective projection and ray tracing through a pin hole camera for establishing the relation between the 3D locations of the points on the cylinder and their corresponding projections on the 2D image plane.

As seen in Fig. 3, the cylinder is observed at different locations and with different orientations at two consecutive frames F_i and F_{i+1}. Let $\mathbf{p}_i$ denote the 3D location of a point sampled on the cylinder on frame F_i. The new location of the point at F_{i+1} is found by applying a transformation model, $\mathbf{M}$, which is represented by a rotation matrix R corresponding to $\left(\omega_x^i, \omega_x^i, \omega_x^i\right)$ and a translation vector $T = [\tau_x^i, \tau_y^i, \tau_y^i]^T$,

$$\mathbf{M}(\mathbf{p}_i, \Delta\vec{\mu}_i) = R\mathbf{p}_i + T. \tag{3}$$

The location of the projected point $\mathbf{u}_{i+1}$ on F_{i+1} is found by using a 2D parametric function $\mathbf{F}$ and applying the rigid motion vector $\Delta\vec{\mu}_i$, followed by a projection operation denoted by $\mathbf{P}$. Thus the projection of the point at $t = t_{i+1}$ can be expressed in terms of the 3D location of the point at $t = t_i$ and the rigid motion vector as:

$$\mathbf{u}_{i+1} = \mathbf{P}(\mathbf{M}(\mathbf{F}(\mathbf{u}_i), \Delta\vec{\mu}_i)). \tag{4}$$

This equation describes the mapping in a comprehensive way, covering transformations from object space to image plane ($\mathbf{P}$), from image plane to object space ($\mathbf{F}$) and inter-frame motion ($\mathbf{M}$). If the illumination is assumed to be constant (i.e. the intensity of the pixel $I(\mathbf{u})$ does not change between the images), then the rigid motion vector can be obtained by minimizing the difference between the two image frames, i.e $I(\mathbf{u_{i+1}}) - I(\mathbf{u_i})$. The optimal value of $\Delta\vec{\mu}_i$ is found by using the identity given in Equation 4 and solving for this minimization problem with the Lucas-Kanade method [35].

V. Gaze Direction and Object Depth Estimation

The head pose is certainly indicative of the gaze direction, but does not completely specify it. This is due to the fact that gaze involves eye movements in addition to the head pose. In an experiment with children, it has been shown empirically that the head pose by itself is insufficient for determining children's focus of attention [43]. Some approaches deal with this problem by resolving gaze direction from head pose implicitly by incorporating additional assumptions. For instance, in [44], the focus of attention is assumed to rest on a person, and the estimated head pose is corrected to select the closest person as the target. However, this can only be done in specific application domains, for instance a meeting scenario, where the interaction between people is all that matters. As mentioned earlier, our human-robot interaction scenario involves an experimenter, who focuses on an object (called the object of interest), and the robot, facing the experimenter, is trying to determine the object of interest to initiate joint attention with the experimenter.

In this study, in order to quantify the gaze direction on image plane, we suggest employing the slope of the vector that connects the center of experimenter's head and the center of the bounding box of the object of interest. Moreover, the depth of an object is expressed in numerical terms as the y coordinate of the object center on the image reference frame (see Fig. 3). The intersection of gaze direction and object depth indicates the location of target point on image plane. Since experiments are set up to provide the ground truth for the object of interest at all times, gaze direction and object depth estimation become learning problems. Provided that a comfortable distance between the experimenter and the robot is kept so as to mimic a natural social interaction [45], [46], head pose clearly provides indications regarding gaze direction and object depth with the given definitions. Assuming that the distance between the communicating parties is close enough and the attention fixation points lie in this range, the pose of the head needs to be adjusted to obtain focal points with different depth and direction. We employ GPR to interpolate the gaze direction and depth of the object of interest from 3D head pose estimates [47]. Regression approaches based on Gaussian noise assumptions are previously applied to robotic settings successfully [48]. We now describe the GPR model we use, and justify our choice of model.

Let the variable $\vec{p}_i$ denote a head pose vector, where $\mathcal{P} = \{\vec{p}_i\}$ stands for the set of all observed poses. Assume that the gaze direction values (on image plane) are formulated as random variables $f(\vec{p})$, where $\vec{p} \in \mathcal{P}$, through a transformation f, which we assume to be a real Gaussian process. $f(\vec{p})$ is described by its mean function, $m(\vec{p})$, and covariance function, $k(\vec{p}, \vec{p}')$,

$$f(\vec{p}) \sim \mathcal{G}P(m(\vec{p}), k(\vec{p}, \vec{p}')), \tag{5}$$

where,

$$\begin{aligned} m(\vec{p}) &= E[f(\vec{p})], \\ k(\vec{p}, \vec{p}')) &= E[(f(\vec{p}) - m(\vec{p}))(f(\vec{p}') - m(\vec{p}'))]. \end{aligned} \tag{6}$$

For notational simplicity we let $m(\vec{p}) = 0$ at all times. This could be compensated by applying an offset to $\vec{p}$. An additional assumption, which makes the scenario close to the real life settings, is to consider the observations to be noisy,

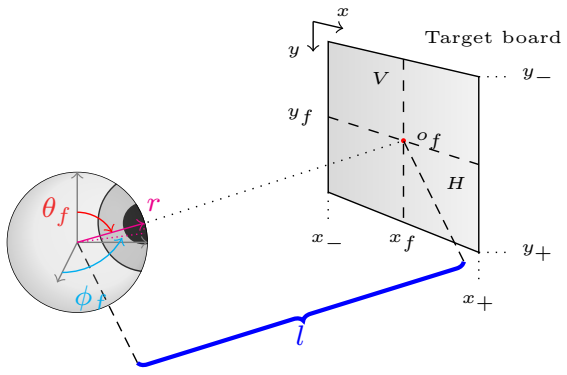


Fig. 4. Orientation of eyeball model and attended point.

$$y = f(\vec{p}) + \varepsilon, \tag{7}$$

where the additive noise ε is independent and identically distributed (iid) with variance σ_0^2. This may be regarded as accounting for the eye movements, by considering them as additive white noise. Before moving on to the details of the model, we discuss the validity of this assumption.

Let the eyeball be modeled as a rigid sphere with orientation (r, θ_f, φ_f) in spherical coordinates with respect to reference frame positioned at the eye center (see Fig. 4). Suppose an experimenter focuses on a point $o_f = (x_f, y_f)$ on a target board positioned in front of him. By changing the azimuth angle ϕ_f, the focus can be adjusted to any point lying on a line H passing through o_f. Line H is not exactly horizontal, but it has a small deviation from a horizontal line passing through o_f. In explicit terms, for a certain azimuth angle ϕ_f, the deviation of the focal point along $\mathbf{y}-$axis of the target board reference frame is $l \tan(\phi_f)$, where l stands for the shortest distance between the eye center and the target plane (see Fig. 4). While this deviation increases with θ_f, our initial assumption that the experimenter directs his field of vision towards the vicinity of the target point by adjusting the head pose accounts for it. The implication is that ϕ_f is expected to be small enough to result in a negligible deviation in the curvature of H. Furthermore, the comfortable distance between the experimenter and the robot l is small enough for the deviation term $l \tan(\phi_f)$ not to grow too much. By changing φ_f, the focus can be set to any point on a line V passing through o_f, which is approximately vertical due to similar reasons.

The angles ϕ_f and θ_f need to be independent and identically distributed for the noise term ε in Equation 7 to be modeled as iid. Let θ_f be known. Thereby, the focus is inferred to be lying on line H. However, the exact location of the focus can not be determined based on this information solely. Similarly, any information on φ_f does not give a clue on θ_f independent of the curvature of H or V. That is to say, as the probability distribution is denoted with $p(.)$,

$$p(\theta|\varphi) = p(\theta), \\ p(\varphi|\theta) = p(\varphi). \tag{8}$$

Thereby, the angles θ and φ are shown to be independent resulting in $\mathrm{cov}(\theta, \varphi) = 0$.

Moreover, since we do not adapt an inhomogeneous and experimenter-specific distribution for the focus points as in [27], the probability that the experimenter looks at any of the points on the target board is equal. In other words, as the borders of the target board are denoted with x_-, x_+ along the $\mathbf{x}-$axis and with y_-, y_+ along the $\mathbf{y}-$axis with respect to the target board axis (see Fig. 4), the distribution of x_f is described as follows:

$$p(x_f) = \begin{cases} \dfrac{1}{(x_+ - x_-)} & x_f \in [x_-, x_+], \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The distribution of y_f is similar to that of x_f in Equation 9. Therefore, from Equations 8 and 9, θ and φ are shown to be independent and identically distributed.

We now elaborate on the training of GPR. Let $n^\star$ input points, represented by $P^\star$, be drawn from this distribution. If there are n points to train the regressor, denoted by P, the covariance matrix $K(P, P^\star)$ has $n \times n^\star$ entries evaluated at every pair of training and test points. For the set of test points, a random Gaussian vector is generated as in [47]:

$$f^\star \sim \mathcal{N}(0, K(P^\star, P^\star)). \tag{10}$$

The prior for the joint distribution of training inputs f and the test outputs $f^\star$ is then:

$$\begin{bmatrix} y \\ f^\star \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(P,P) + \sigma_0^2 I & K(P, P^\star) \\ K(P^\star, P) & K(P^\star, P^\star) \end{bmatrix}\right). \tag{11}$$

The posterior distribution restricts the joint prior distributions to contain only those functions which agree with the observations. Hence, as the prior is conditioned on the observations, we get,

$$f^\star | P, y, P^\star \sim \mathcal{N}(\bar{f}^\star, \mathrm{cov}(f^\star)), \tag{12}$$

where,

$$\bar{f}^\star = K(P^\star, P)[K(P,P) + \sigma_0^2 I]^{-1} y, \\ \mathrm{cov}(f^\star) = K(P^\star, P^\star) - \\ K(P^\star, P)[K(P,P) + \sigma_n^2 I]^{-1} K(P, P^\star). \tag{13}$$

The Bayesian approach defined above is iterated until the marginal likelihood reaches an optimal value. The reader is referred to [47] for the details regarding the solution of free parameters of GPR.

The learning procedure determines the parameters of a function f_1 that interpolates the gaze direction from the head pose parameters, as well as a function f_2 that estimates the depth of the focused object. Both functions have the same form as presented in Equation 7; in principle it is straightforward to concatenate the output values and learn them jointly as a single multivariate output function. For the function f_1, the summation of an isotropic rational quadratic covariance function and a neural network covariance function is used, based on the inference presented above. For the second regression scheme f_2, we have used the summation of a linear covariance function and an independent covariance function. The linear covariance function is considered to account for the distance relation arising from the relative positions of the camera, the experimenter and the attended objects, whereas the independent covariance function represents the effect of the eye movements on object depth.

We now turn to the use of contextual information that complements these estimates.

VI. Target Object Location Estimation and the Saliency Model

Once the gaze direction and object depth are estimated, the intersection of those provides a prospective region for searching for the target point. However, since interpolation from head pose is a coarse indicative of the focus, there is an inherent uncertainty, which we seek to remedy via contextual cues. After narrowing down the search region using the coarse estimates, a saliency scheme is applied to single out the spot of attention. This combination is consistent with the view that an object can be salient because of its inherent properties like color, motion or proximity, but also because it is the focus of attention of the interacting party [49].

In [50], a robotic system is described where the bottom-up saliency of a visual scene is computed by color, edge, and motion cues. Top-down influences can be incorporated by modulating bottom-up channels, or by explicitly adding dedicated saliency components. For instance, faces are particularly important for natural interaction settings, consequently they can be separately detected and made salient [51]. Breazeal *et al.* point out to the importance of using such features of human pre-attentive visual perception for natural human-robot interaction [52].

We implement and contrast three saliency schemes to fine-tune the target location estimates. These are 1) a bottom-up method proposed by Itti *et al.* [53], 2) the more recent method of Judd *et al.* combining a set of low, mid- and high level features [54], and 3) the attention based information maximization approach of Bruce *et al.* [55]. From these, the first is by far the most popular saliency approach in the literature, but good results were reported under the other schemes as well. For segmenting out objects in the scene, region growing methods can be applied to the most salient location in an unsupervised manner [56]. More recent saliency approaches proposed for multimedia retrieval use supervised training, but require large amounts (tens of thousands) of annotated training samples [57].

1) **Itti *et al.*** present an approach in [53], which is based on the feature integration theory of Treisman and Gelade, and decomposes the saliency of a scene into separate feature channels [58]. A saliency map ordinarily uses the presence of cues like illumination intensity, colors, oriented features and motion to determine salient locations in a scene. The saccadic eye movements derived from the saliency map are simulated by directing a foveal window to the most salient location, determined by a dynamic and competitive Winner-Take-All (WTA) network [53]. Once a location is selected, it is suppressed

by an inhibition-of-return mechanism to allow the next most-salient location to receive attention.

2) **Judd *et al.*** extend [53] by incorporating a set of mid- and high level features. Similar to [53], they employ low level features accounting for the color, intensity, and orientation, whereas the mid-level features relate to the detection of horizon and high level features refer to contextual cues such as face and person detection, as well as a center prior. Since our interaction scenario precludes some of these channels (i.e. faces, people and the horizon), contextual detectors are turned off and the rest of the features (such as subbands of the steerable pyramid, the three channels of Itti and Koch's saliency model, and the color features) are employed in computation of saliency.
3) **Bruce and Tsotsos** argue that the saliency is equal to a measure of the information present locally within a scene in relation to its surroundings [55], [59]. In natural visual stimuli, there exist significant amounts of redundancy. For a given image, the authors employ a set of basis functions and extract independent features for each point. The projections to different feature spaces (one for each filter channel) spanned by the basis functions is followed by density estimation within each channel. Based on the independence assumption, the joint likelihood is assumed to be the product of the likelihoods concerning each filter type. The final saliency map is computed as the Shannon self-information derived from the joint likelihood concerning all filter types.

These three saliency resolution methods utilize static features and do not consider the effect of motion cues for the present scenario. These cues may be caused by movement of the target object, or of the robotic head. The effects of motion are negligible in our case, as the object of interest is always static (no motion due to target) and motion of robotic head affects all objects uniformly, since they are positioned closely on the table.

We use the saliency map (computed with one of the three methods explained above) for determining the most salient location in the prospective region. We consider those methods exploiting the low-level features, basis functions or entropy, more appropriate for our purposes than several others, which consider the issue from a practical perspective accounting for a biologically plausible non-uniform retinal sampling [60] or task engagement [61]. Namely, the previous experiences, intentions, the nature of performed task, or biological limitations do not play an important role in the detection of the salient points from a robot vision perspective. In addition, the collection of low-level features are suggested to perform better than, for instance, a simple edge detector due to the extended scope. If there is more information available as to the experimenter's intentions, or an instruction history that can provide background probabilities with regards to which objects are more likely to receive attention, these can be integrated into the saliency computation in a top-down manner, by for instance modulating the responses of individual feature channels appropriately. In [27], the probability that an experimenter selects a particular object is learned by fitting a Gaussian mixture model on the pixel distribution. We do not model the top-down influence at this stage, simply because in the absence of specific contextual models, this additional information would optimistically bias the results.

Since human eye makes three to five saccades per second, it is not realistic to compute saliency for each gaze direction and object depth estimate corresponding to a $10fps$ rate. Therefore, we form bins of consecutive frames and calculate the 2D location of focus of attention for each of these bins. We pooled 4 frames corresponding to 0.4 seconds. This procedure is referred to as *pooling* throughout the paper. Gaussian distributions are then positioned around the resulting estimates, with a 10 pixel standard deviation. The choice of the latter follows from considering a typical interaction scenario in conjunction with the resolution of the camera.

We postulate that the robot will typically consult the joint attention system when it is trying to figure out which object is in focus (i.e. there is a reference to an object which needs resolving) or when there is a large shift in the head pose of the experimenter, suggesting an imminent shift in the focus of attention. This implies that the estimated focus of attention can be assumed to rest on an object. For scenarios when this assumption cannot be justified, the saliency-based refinement described in this section should be disabled.

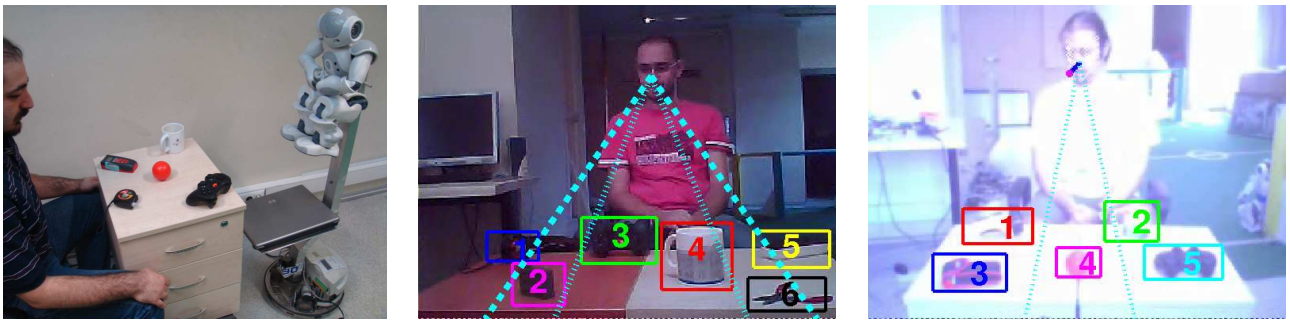
VII. Experimental Results and Discussion

A. Experimental Setup

In order to evaluate the proposed method, we performed two sets of experiments to model joint attention between a human experimenter and an embodied agent (see Fig. 5). In both cases, the experimenter is in front of a table, on which a number of objects are placed in a non-occluding fashion as in Fig. 5-(b). A session of joint attention is initialized when the experimenter establishes eye-contact with the robot, which leads to a fully-frontal face image being acquired by the robot. The experimenter fixates his/her attention to one of the six objects by looking at them in random order for a certain duration of time. In this setting, the experimenter moves his head in a range of $[-50, 50]$ degrees for pitch, $[-40, 40]$ degrees for yaw, and $[-20, 20]$ degrees for roll rotations. If the head pose is found to be outside of this range, tracking is assumed to be lost and it is restarted using the first image of the video sequence.

Considering all the design details given in Section I, we use a social interaction robot that is designed to be used for service and guiding purposes in our experiments. This system is composed of three main components: An Aldebaran Nao humanoid robot as the main interaction and animation unit, a Festo Robotino robot as the navigation unit, and a laptop computer as the additional processing and monitoring unit. The recordings are made with the color camera of the Nao.

To inspect the generalization capabilities of the proposed method, we provide experimental results including comparisons between recordings obtained from a single experimenter across multiple sessions, as well as results for cross-experimenter generalization both under normal and adverse



(a) (b) (c)

Fig. 5. The experimental setup consists of an experimenter and the robot at the opposite sides of a table, on which several objects are placed (a). To measure generalization capabilities, we vary the experimenter, the type, size, and number of objects, the background, and the illumination and blurring conditions between recording set-1 (b) and recording set-2 (c).

conditions. In each sequence, all objects are randomly attended for several seconds. The ground truth for the attended objects is obtained by manual annotation.

We collected two sets of videos, comprising 20 sequences, recorded at $10 fps$ frame rate with 320×240 resolution. In both sets, the distance between the robot and the experimenter is roughly 1.5 meters, so as to obtain a comfortable distance for a natural social interaction. The attention fixation points lie in the field of view of both parties. In this setting, the center-to-center object separation is approximately 20 to 30 cm. Recording set-1 is designed to evaluate the intraclass and interclass generalization capabilities under normal illumination conditions, providing an understanding of the *baseline* performance, where four different experimenters provide two sequences each in front of a stationary camera under sufficient natural illumination. Recording set-2 is composed of 12 video sequences collected from four experimenters under three different conditions. These include artificial illumination (see Fig. 5-(c)), dimmed illumination, and motion blur. The whole database constitutes 6691 images in total, for which the focus of attention ground truth is manually annotated.

We assess the performance for head-pose based interpolation and saliency-based fine-tuning separately. We compare the proposed GPR with a neural network (NN) regressor, which has a two-layer back-propagation architecture [62]. We use 10 hidden units, an initial learning rate of 0.1, which is exponentially decreased during training, and an online training scheme. Weights in both layers are initialized randomly from the interval $(-0.5, 0.5)$. During training, a validation set (drawn from the training sequence) is monitored for error decay to prevent over-fitting.

B. Quality Measures

We use two different measures, termed Q_1 and Q_2, for quantifying performance. Q_1 indicates the ratio that an estimated object center falls into the bounding box of the object of interest. This is a more fine-grained measure than the mean angular error, as the object depth is also taken into account. The second measure Q_2 shows the rate at which the estimated point has the shortest distance to the true object center among all targets. Thus, Q_2 shows the effect of imposing object selection as an additional constraint in the system.

Let q denote the pixel locations of the estimated object center for a set of frames which are labeled with object number i. Let B_i be the bounding box of this object. It follows:

$$Q_1(i) = |q \in B_i| \,/\, |q| \,, \tag{14}$$

where $|.|$ denotes the cardinality of a set.

Q_2 assigns an estimated point to the object, whose center has the shortest distance among the set of all objects. It follows that the explicit expression is:

$$Q_2(i) = |\{q | d(q, c_i) < d(q, c_j), \forall j = 1, \cdots, n, j \neq i\}| \,/\, |q| \,, \tag{15}$$

where $d(a, b)$ denotes the Euclidean distance between points a and b, c_i stands for center of object i, and n is the number of objects.

C. Qualitative Evaluation

In this section we qualitatively evaluate different parts of the proposed method.

We have mentioned that the camera input often does not provide sufficient resolution for the analysis of the eye region. Fig. 6 illustrates the performance of face detection on such low resolution videos. The original camera input from the robot is downscaled to as low as 10% of the original size, i.e. 32×24. We observe that the Haar-wavelet based face detector is able to find the face region correctly in about 90% of the frames when the images are 30% of their original size (320×240 for the entire original image, 1000-1500 pixels for the original face area). Therefore, we infer that the employed face detection algorithm is robust against low resolution inputs.

We next illustrate that the head pose distributions for different targets represent a learnable range of values, and graphically show the variation that relates to noise and eye movement. Fig. 7) illustrates estimated pose distributions for two exemplary video sequences obtained by the head pose estimation method explained in Section IV. We investigate the relation between the distribution of head pose values and the layout of the objects on the table. By fitting multivariate Gaussian distributions to the point clouds concerning each object, we obtain the indicated regions in 3D. As can be seen from Fig. 7, the topological relationship between the locations of the objects in Fig. 5-(b) and corresponding head

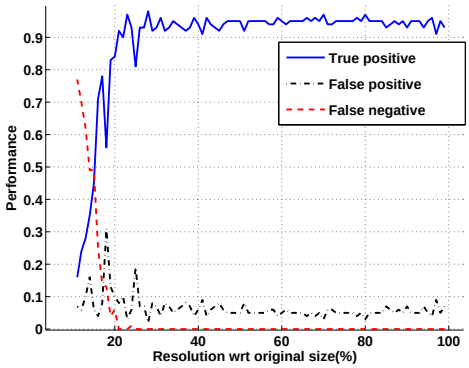
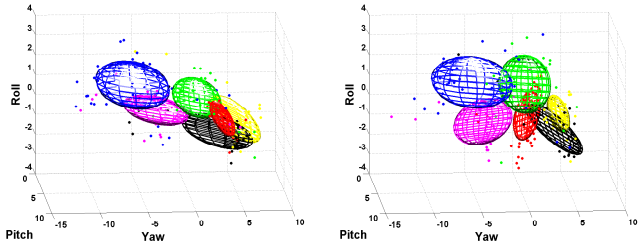


Fig. 6. Face detection performance vs resolution.



(a) (b)

Fig. 7. Distribution of estimated pose values for two video sequences. Each point is a pose value for one object location, ellipsoids are representations for clusters obtained by adapting multivariate Gaussian distributions for each object.

pose angles are preserved. Investigating the covariance of these poses (illustrated in Fig. 7 for each object) we conclude that the variation of pose values for different objects are not always significant compared to the accuracy of the pose estimator. Nonetheless, the accuracy of Xiao et al.'s method is regarded to perform satisfactorily good considering the difficulty of the problem. Furthermore, the grouping of the pose angles with respect to the target objects reveals not only a clear clustering, but also the nonlinear nature of the relation between head pose and gaze direction.

The location of the objects influences the accuracy of gaze estimation; the task is made more difficult by placing the objects closer to each other. As the incorrect decisions are

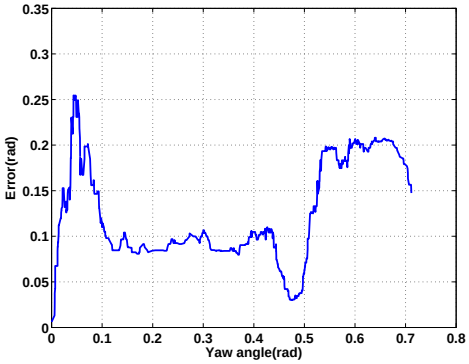
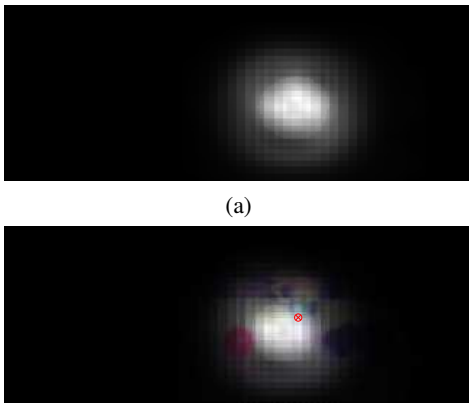


Fig. 8. Error in gaze direction with respect to estimated head pose values.



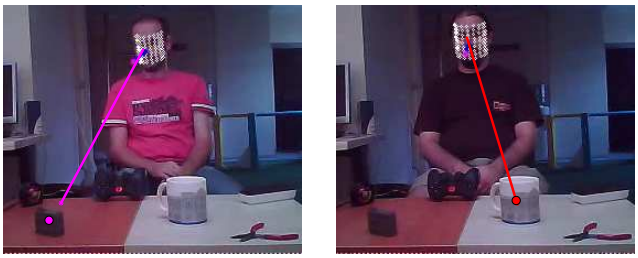
(a) (b)

Fig. 9. (a)The saliency map and (b) the target point resolved by the saliency method of Judd et al. [54] demonstrated on the masked image with the red marker.

investigated, it is observed that they mostly correspond to the neighboring objects. Fig. 8 illustrates the error versus the resolved values of yaw angle. From this figure, we observe that for a small angle range deviating from the frontal pose (up to approximately 0.1 radians) the head moves very little, and the target fixation is performed with eye movements mostly. This is due to the fact that the required eye movement for focusing on any point in this range is in the customary ocular motor range. Hence, focus can be attained only by eye movements [7]. This phenomenon causes increased error in estimation in the customary ocular motor range, as we rely on head pose. This phenomenon causes increased error in estimation in the customary ocular motor range, as we rely on head pose.Then for a wide range of yaw angles (0.1-0.5 radians) we have good prediction accuracy. Beyond that range, head pose estimation suffers from rotations, and this reflects on the accuracy.

We next discuss the integration of estimated gaze directions and object depths on image plane. By imposing object depth on gaze direction, a single point is obtained on the image plane, which is indicative of the best estimate for the target point. Fig. 10 presents two examples, where the gaze vector is depicted as starting from the head center, going along the gaze direction, and finally terminating at the estimated depth of the object of interest. For an entire video sequence, resulting integrated estimates are presented in Fig. 11. Subsequently, we pool 4 of the consecutive estimates and mask the image positioning a Gaussian distribution around each of them. Fig. 9 shows a masked image and corresponding saliency map obtain by Itti et al.'s method.

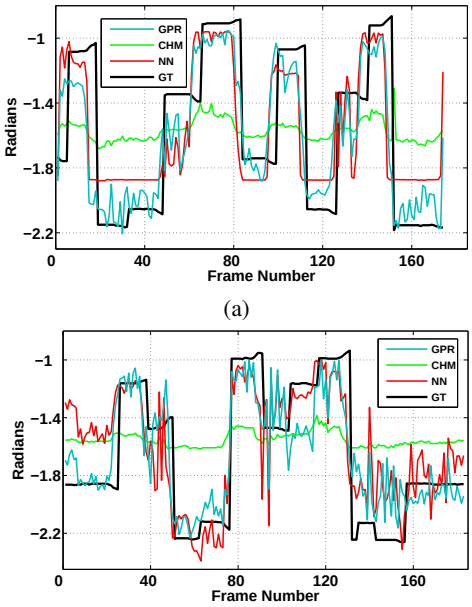
Our final remark addresses the computational aspects of the proposed method. The most computationally intensive part is the cylindrical-model based tracking, which works real-time on a standard PC ($10fps$ on a machine with 2GHz Intel Centrino CPU and 2Gb RAM). The saliency model needs to be evaluated on areas smaller than 5% of the images, and subsequently has negligible computation time. The two regression results are also obtained with very little computational effort.



(a) (b)

Fig. 10. Estimated gaze direction and object depth via GP regression (line) together with the center of the manually annotated object of interest (solid dot).

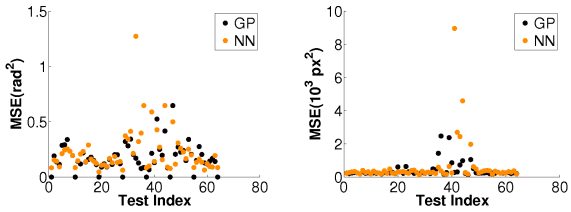
Fig. 11. Initial estimates for target points for a representative video sequence. Each estimated target point is depicted in the color presented in Fig. 5-(b) corresponding to the bounding box of the annotated object.



(a)

(b)

Fig. 12. Improvement in gaze direction shown for two exemplary training-test pairs. Ground truth (GT) is denoted in black.



(a) (b)

Fig. 13. Mean square errors for (a) gaze estimation (in terms of radian squared), (b) object depth estimation (in terms of 10^3 pixel squared).

D. Quantitative Evaluation

The two components of initial object location estimation, i.e. gaze direction and object depth, are examined separately first at an individual level and then in a collective manner.

For two exemplary video sequences, gaze estimation performance of CHM, GPR, and NN based schemes is presented Fig. 12. Among those, CHM based pose resolution clearly underestimates the actual gaze direction, while GPR and NN based methods follow the ground truth with better accuracy.

Subsequently, for the entire dataset, the mean square error (MSE) regarding gaze direction and object depth estimation of GPR and NN regressors are depicted in Fig. 13. From this figure, it follows that GPR generally results in slightly smaller error values in gaze estimation, where NN regressor converges prematurely to a local minimum for some of the training-test pairs in depth estimation (Fig. 13-(b)).

In addition to providing a performance comparison for GPR and NN based schemes, the MSE behavior is helpful in determining an adequate parameterization of the Gaussian mask to be applied around the pool of initial estimates. In accordance with the average of MSE values concerning gaze direction and object depth estimation illustrated in Fig. 13, the search envelope is restricted to 10 pixels. In the next section, we investigate initial (regression only) and final (regression + saliency) performance of the proposed method for the first group of recordings collected under normal conditions. The efficiency of the alternative NN based estimation is compared to this baseline performance and a discussion is provided.

E. GPR vs. NN based estimation

Performance quantifiers Q_1 and Q_2 are evaluated for self-referencing (intraclass, i.e. single experimenter) and cross-referencing (interclass, i.e. training and testing on different experimenters) cases. The mean value of true positive rates for individual objects are depicted in Tables I-(A) and I-(B) for GPR and NN regression based estimations, respectively.

From these two tables, it is evident that GPR based object location estimation yields higher accuracy compared to the NN based scheme at every evaluation condition. In addition to the overall rates presented in Table I, a closer look at the results at the individual object level indicates that the standard deviation of the NN based estimation method is generally higher than that of the GPR based scheme. We conclude that GPR is more suitable for the interpolation task, as it provides a more stable estimation. Nonetheless, the differences are not statistically significant, and we expect that any powerful regression method to be able to perform similarly under these conditions.

Comparing the results before and after the incorporation of saliency, we can see that modeling image saliency improves object location estimation, compared to relying only on regression. This is the case for all three saliency schemes we have considered. The improvements are close to each other, and we cannot single out any of the saliency schemes as being clearly superior to the others.

Comparing self-referencing and cross-referencing results, we see that the method generalizes well across experimenters.

TABLE I
PERFORMANCE QUANTIFICATION FOR (A) GPR AND (B) NN BASED REGRESSION.

(A)

	Self-referencing					Cross-referencing			
	Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce		Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce
Q_1	.24	.37	.45	.40	Q_1	.15	.29	.36	.32
Q_2	.39	.56	.52	.50	Q_2	.36	.51	.46	.43

(B)

	Self-referencing					Cross-referencing			
	Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce		Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce
Q_1	.18	.33	.44	.38	Q_1	.19	.29	.37	.35
Q_2	.36	.52	.53	.52	Q_2	.40	.45	.47	.46

It is natural that training and testing on the same experimenter produces higher predictive power, as the method is based on learning gazing behavior. Yet, the idiosyncratic variation is not so great as to prevent generalization.

F. Performance Under Adversarial Conditions

We assess the generalization capability of the proposed method under different illumination conditions and motion blur. For this purpose, a series of experiments are conducted, where the illumination conditions are changed for the test set by using artificial or dimmed lighting, as opposed to the previously used natural light or motion blur is introduced by moving the head of the robot. The performance rates concerning recording set-2 are presented in Table II. Table II-(A) is the cross-referenced baseline performance of recording set-2, i.e. trained and tested with the videos of different experimenters all under natural light. Tables II-(B,C,D) give the performance rates where training assumes natural light conditions and testing considers artificial illumination, dimmed illumination and blurred vision, respectively.

By comparing Table II-(A) to Table II-(B-D), it can be deduced that the effect of adversarial conditions does not lead to any decay in performance. The fluctuations are partly due to the different number of objects used in different experiments (five or six objects, respectively, see Fig. 5-(b) and (c)).

Next we look into the effect of pooling and changing quantifiers. From Table II-(A) it is clear that pooling the estimates and applying saliency computation leads to a significant improvement in resolution of attention fixation points in terms of both Q_1 (0.32 to 0.52) and Q_2 (0.51 to 0.73). Moreover, assigning the estimates to the object lying in closest proximity leads to higher detection rates, as pointed out in Section VII-D. Similar approaches in the literature also use pooling. In [9] a dual-camera system with high resolution is used in a similar robotic setup, and for three objects at 20 cm distance to the experimenter, $\%80$ correct detection rate is reported by pooling gaze estimations over more than 60 frames.

In order to investigate the level of degradation introduced by dimmed illumination, the regressors are trained using the pose values obtained from the videos that are collected under natural illumination and tested using the videos collected under dimmed illumination settings. Performance results are presented in Table II-(B). Comparing Tables I-(A) and II-(B), it is observed that there is no significant decline with respect to the baseline performance. Moreover, pooling and saliency estimation are inferred to improve performance in terms of both Q_1 (0.23 to 0.51) and Q_2 (0.47 to 0.60). Besides, similar to the artificial illumination case, following the assignment scheme of Q_2 leads to higher detection rates than that of Q_1.

We also test the performance under a motion blur condition with a set of recordings obtained when the robot was in motion. The results indicate that the proposed method is robust to motion artifacts. Similar to the previous cases, there is no significant decay in detection rates with respect to the baseline performance. Comparing initial and final performances in Table II-(C), it is observed that pooling and saliency estimation affects the results positively for both Q_1 (0.19 to 0.46) and Q_2 (0.34 to 0.53). Moreover, similar to the other test scenarios, performance rates calculated considering Q_2 are higher than those calculated considering Q_1.

TABLE II
PERFORMANCE UNDER (A) NATURAL LIGHT AND ADVERSE CONDITIONS OF (B) ARTIFICIAL ILLUMINATION, (C) DIMMED ILLUMINATION, (D) BLURRED VISION.

(A)

	Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce
Q_1	.18	.44	.35	.45
Q_2	.49	.63	.60	.58

(B)

	Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce
Q_1	.32	.52	.42	.48
Q_2	.51	.73	.70	.62

(C)

	Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce
Q_1	.23	.51	.44	.40
Q_2	.47	.60	.64	.58

(D)

	Regr.	Regr.+Itti	Regr.+Judd	Regr.+Bruce
Q_1	.19	.46	.36	.35
Q_2	.34	.53	.59	.45

There are several sources for the error in target point estimation. The basic source is the accuracy and effective range of the head pose estimation method [39]. The resolution values as reported in [42] do not permit us to distinguish very small angular variations. Nonetheless, considering the

difficulty of the problem, Xiao et al.'s method is regarded as performing well. A second source of error is suggested to be the experimenters' personal preferences regarding head pose onset or eye movement range. The pose distributions presented in Fig. 7 have a certain fluctuation even for the same experimenter across multiple instances of focusing on the same object. In addition to the accuracy of head pose estimation and personal variations, the proximity of objects poses an intrinsic challenge. Decreasing the number of objects for interaction helps improving target point estimation as expected, and this can be verified by comparing Tables I-(A) and II-(A).

VIII. Conclusions

This paper provides a method to estimate the focus of attention of an experimenter from a single, low-resolution camera. We employ a 3D elliptic cylindrical head model to estimate the head pose. Our model uses the estimate of the head pose to correct for gaze direction and object depth, and further refines the estimate by a saliency based selection for finding objects attended by an experimenter. We seek to remedy head pose and gaze direction discrepancy by employing two parallel Gaussian process regressors that correct for gaze direction by interpolation, as well as estimate object depth from the head pose. The proposed scheme is shown to work with good accuracy, although the problem we tackle is ill-posed under the realistic limited resource assumptions we make.

Using saliency to fixate on interesting objects serves a two-fold purpose. Firstly, it reduces the uncertainty in the estimation of the gaze direction. We may safely conjecture that since saliency computation in the early layers of the visual system precedes the estimation of gaze direction, the saliency-based grafting of the gaze to interesting objects should serve as a supervisory system for learning to estimate the gaze direction. In humans, a consequence of this learning is the developing ability of the infant to estimate the focus of the experimenter even when it lies beyond the visual field of the child. Secondly, saliency-based grafting compensates for the discrepancy between intended motor commands and executed physical actions, which is particularly relevant for robotic implementations. The movement of the simulated fovea effectively creates an object-centered coordinate system, which is a precondition of parsimonious mental object representations.

Our saliency scheme is feature-based, and does not assume the existence of any high-level information. In fact, if objects of interaction are known beforehand, it would be much easier to replace the saliency scheme with an object-specific search, for instance based on known sets of feature descriptors. Furthermore, assuming known objects would also permit us to learn object preferences per experimenter or object priors conditioned on the context. We argue that including these high-level cues does not make the approach more generally applicable, but less so, because in the present case the experimenter looks at all objects, without any preference or order, which is the least restricted of all interaction scenarios. Possible future directions include incorporating object-specific information, as well as the extension of experimenter's gaze range to allow for objects above the experimenter. Another extension would be to incorporate temporal information in the estimation; smaller head-shifts may imply that the shift of focus is performed with gaze only, giving clues about the reliability of estimation.

The combination of the proposed method with an accurate eye-region based gaze-estimation approach is possible under certain conditions. If the latter is doing a decent job, our intuition is that the direction estimator based on head pose becomes obsolete, as its seeks to solve an ill-posed problem. However, the depth estimator, as well as the saliency-based refinement can both be integrated, as the eye region does not provide these cues.

Estimation of body posture in addition to head pose might help to make the interaction more natural by the pointing or manipulation of the objects by the experimenter. In [63], the authors use pointing gestures (as opposed to gaze) in combination with saliency-based refinement to detect target objects. Another possibility is to add direct gaze estimation by using a higher-resolution camera to inspect experimenter's eyes as an additional physical cue. An accurate gaze following system for joint attention presents a suitable testbed for evaluating complex interaction models, for testing alternative teaching techniques for children and robots, for analyzing developmental disorders, and for running social simulations.

References

[1] M. Asada, K. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robot. Autom. Syst.*, vol. 37, no. 2-3, pp. 185–193, 2001.

[2] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Current Directions in Psychological Science*, vol. 16, no. 5, p. 269, 2007.

[3] R. Rao, A. Shon, and A. Meltzoff, "A Bayesian model of imitation in infants and robots," in *Imitation and Social Learning in Robots, Humans, and Animals: Behavioural, Social and Communicative Dimensions*, K. Dautenhahn and C. Nehaniv, Eds. Cambridge University Press, 2004, pp. 217–247.

[4] V. Corkum and C. Moore, "Development of joint visual attention in infants," in *Joint Visual Attention: Its Origins and Role in Development*, C. Moore and P. J. Dunham, Eds. Hillsdale, NJ:Erlbaum, 1995, pp. 61–84.

[5] A. Meltzoff, R. Brooks, A. Shon, and R. Rao, ""Social" robots are psychological agents for infants: A test of gaze following," *Neural Networks*, vol. 23, pp. 966–972, 2010.

[6] Z. Yücel and A. A. Salah, "Head pose and neural network based gaze direction estimation for joint attention modeling in embodied agents," in *Proc. Annual Meeting of Cognitive Science Society*, 2009.

[7] J. Stahl, "Amplitude of human head movements associated with horizontal saccades," *Experimental Brain Research*, vol. 126, no. 1, pp. 41–54, 1999.

[8] D. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 478–500, 2010.

[9] P. Ravindra De Silva, K. Tadano, S. Lambacher, S. Herath, and M. Higashi, "Unsupervised approach to acquire robot joint attention," in *Intl. Conf. Autonomous Robots and Agents*, 2009, pp. 601–606.

[10] L. Fletcher, G. Loy, N. Barnes, and A. Zelinsky, "Correlating driver gaze with the road scene for driver assistance systems," *Robot. Autom. Syst.*, vol. 52, no. 1, pp. 71–84, 2005.

[11] S. Lee, J. Jo, H. Jung, K. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, 2011.

[12] S. Ba and J. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, 2011.

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

[14] B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, C. Nehaniv, Ed. Springer Verlag, 1999, pp. 176–195.

[15] S. Langton, H. Honeyman, and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction." *Perception & Psychophysics*, vol. 66, no. 5, pp. 752–771, 2004.

[16] G. Butterworth and N. Jarrett, "What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy," *British J. Developmental Psychology*, vol. 9, no. 1, pp. 55–72, 1991.

[17] M. Begum and F. Karray, "Visual Attention for Robotic Cognition: A Survey," *IEEE Trans. Autonomous Mental Development*, vol. 3, no. 1, pp. 92–105, 2011.

[18] Y. Nagai, M. Asada, and K. Hosoda, "Developmental learning model for joint attention," in *Proc. IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, 2002, pp. 932–937.

[19] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003.

[20] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," *Advances in Neural Information Processing Systems*, pp. 753–753, 1994.

[21] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "An incremental learning method for unconstrained gaze estimation," *European Conf. Computer Vision*, pp. 656–667, 2008.

[22] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2667– 2674.

[23] J. Chen and J. Ji, "Probabilistic gaze estimation without active personal calibration," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 609–616.

[24] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada, "Acquisition of joint attention through natural interaction utilizing motion cues," *Advanced Robotics*, vol. 21, no. 9, pp. 983–1000, 2007.

[25] Y. Nagai, "Joint attention development in infant-like robot based on head movement imitation," in *Proc. Intl. Symp. Imitation in Animals and Artifacts*, 2005, pp. 87–96.

[26] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, 2009.

[27] M. Hoffman, D. Grimes, A. Shon, and R. Rao, "A probabilistic model of gaze imitation and shared attention," *Neural Networks*, vol. 19, no. 3, pp. 299–310, 2006.

[28] A. Shon, D. Grimes, C. Baker, M. Hoffman, S. Zhou, and R. Rao, "Probabilistic gaze imitation and saliency learning in a robotic head," in *Proc. Intl. Conf. Robot. Autom.*, 2005, pp. 2865–2870.

[29] C. Bard, M. Fleury, and J. Paillard, "Different patterns in aiming accuracy for head-movers and non-head movers," in *The Head-neck Sensory Motor System*, A. Berthoz, W. Graf, and P. Vidal, Eds. Oxford University Press, 1992, pp. 582–586.

[30] C. Peters and A. Qureshi, "A head movement propensity model for animating gaze shifts and blinks of virtual characters," *Computers & Graphics*, vol. 34, no. 6, pp. 677–687, 2010.

[31] A. Shon, J. Storz, A. Meltzoff, and R. Rao, "A cognitive model of imitative development in humans and machines," *Intl. J. Humanoid Robotics*, vol. 4, no. 2, pp. 387–406, 2007.

[32] A. Meltzoff and M. Moore, "Explaining facial imitation: A theoretical model," *Early development and parenting*, vol. 6, no. 3-4, pp. 179–192, 1997.

[33] H. Goossens and A. Opstal, "Human eye-head coordination in two dimensions under different sensorimotor conditions," *Experimental Brain Research*, vol. 114, no. 3, pp. 542–560, 1997.

[34] E. Freedman, "Coordination of the eyes and head during visual orienting," *Experimental Brain Research*, vol. 190, no. 4, pp. 369–387, 2008.

[35] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Intl. Joint Conf. Artificial Intelligence*, vol. 3, 1981, pp. 674–679.

[36] M. Hayhoe, M. Land, and A. Shrivastava, "Coordination of eye and hand movements in a normal environment," *Invest. Ophthalmol & Vis. Sci*, vol. 40, no. 4, p. S380, 1999.

[37] J. Triesch, H. Jasso, and G. Deák, "Emergence of mirror neurons in a model of gaze following," *Adaptive Behavior*, vol. 15, no. 2, pp. 149–165, 2007.

[38] L. Itti, N. Dhavale, and F. Pighin, "Photorealistic attention-based gaze animation," in *IEEE Intl. Conf. Multimedia and Expo*, 2006, pp. 521–524.

[39] J. Xiao, T. Moriyama, T. Kanade, and J. Cohn, "Robust full-motion recovery of head by dynamic templates and re-registration techniques," *Intl. J. Imaging Systems and Technology*, vol. 13, no. 1, pp. 85–94, 2003.

[40] C. Gordon, T. Churchill, C. Clauser, B. Bradtmiller, J. McConville, I. Tebbetts, and R. Walker, "Anthropometric Survey of US Army Personnel: Methods and Summary Statistics," *US Army Natick Research Dev. and Eng. Center Natick Massachusetts Tech. Report*, 1989.

[41] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[42] R. Valenti, Z. Yucel, and T. Gevers, "Robustifying eye center localization by head pose cues," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 612–618.

[43] L. Piccardi, B. Noris, G. Schiavone, F. Keller, C. Von Hofsten, and A. Billard, "Wearcam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children," in *Intl. Symp. Robot & Human Interactive Communication*, 2007, pp. 594–598.

[44] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing," in *Proc. Intl. Conf. Multimedia*, 1999, pp. 3–10.

[45] M. Walters, K. Dautenhahn, R. Te Boekhorst, K. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry, "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *IEEE Intl. Workshop on Robot and Human Interactive Communication*. IEEE, 2005, pp. 347–352.

[46] E. T. Hall, *The Hidden Dimension*. Doubleday, 1963.

[47] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[48] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *Proc. ACM/IEEE Intl. Conf. Human-robot Interaction*, 2007, pp. 262–270.

[49] C. Breazeal, "Social interactions in HRI: The robot view," *IEEE Trans. Syst., Man, Cybern. C*, vol. 34, no. 2, pp. 181–186, 2004.

[50] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "Emergence of joint attention based on visual attention and self learning," in *Intl. Symp. Adaptive Motion of Animals and Machines*, 2003.

[51] Y. Ma and H. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Intl. Conf. Multimedia*, 2003, pp. 374–381.

[52] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Trans. Syst., Man, Cybern. A*, vol. 31, no. 5, pp. 443–453, 2001.

[53] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[54] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE Intl. Conf. Computer Vision*, 2009.

[55] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vision*, vol. 9(3), no. 5, pp. 1–24, 2009.

[56] J. Han, K. Ngan, M. Li, and H. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, 2006.

[57] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011.

[58] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[59] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, pp. 155–162, 2006.

[60] B. Vincent, T. Troscianko, and I. Gilchrist, "Investigating a space-variant weighted salience account of visual selection," *Vision research*, vol. 47, no. 13, pp. 1809–1820, 2007.

[61] B. Tatler, M. Hayhoe, M. Land, and D. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of vision*, vol. 11, no. 5, 2011.

[62] C. Bishop, *Neural networks for pattern recognition*. Oxford University Press, 2005.

[63] B. Schauerte, J. Richarz, and G. Fink, "Saliency-based identification and recognition of pointed-at objects," in *Proc. IEEE/RJS Intl. Conf. Intelligent Robots and Systems*, 2010, pp. 4638–4643.