Resolution of Focus of Attention Using Gaze Direction Estimation and Saliency Computation

Zeynep Yücel
CWI, Science Park 123,
1098 XG, Amsterdam, The Netherlands
zeynep@ee.bilkent.edu.tr

Albert Ali Salah
ISLA, Science Park 107
1098 XG, Amsterdam, The Netherlands
a.a.salah@cwi.nl

Abstract

Modeling the user's attention is useful for responsive and interactive systems. This paper proposes a method for establishing joint visual attention between an experimenter and an intelligent agent. A rapid procedure is described to track the 3D head pose of the experimenter, which is used to approximate the gaze direction. The head is modeled with a sparse grid of points sampled from the surface of a cylinder. We then propose to employ a bottom-up saliency model to single out interesting objects in the neighborhood of the estimated focus of attention. We report results on a series of experiments, where a human experimenter looks at objects placed at different locations of the visual field, and the proposed algorithm is used to locate target objects automatically. Our results indicate that the proposed approach achieves high localization accuracy and thus constitutes a useful tool for the construction of natural human-computer interfaces.

Keywords: Head pose estimation, gaze estimation, joint attention modeling, saliency, intelligent interaction.

1. Introduction

Creating a more natural interface between humans and intelligent agents is a challenging problem, which has drawn considerable interest in the recent years. One popular approach in the construction of naturally interacting agents relies on the exploration of interaction between humans. From communication point of view, establishing and maintaining social interaction requires sustaining a continuous joint attention. As such, building a module that enables an embodied agent to determine the focus of joint attention of a human is clearly a major step in the construction of naturally interacting embodied agents.

Visual cues are extensively used for implementing working models on embodied agents, and the visual distinctions that can be perceived by the embodied agent serve as affordances [7]. Intuitively, joint attention modeling relies on the estimation of gaze direction to a large extent. On the other hand, head pose provides a coarse estimate for gaze direction. Recent approaches in this direction employ Bayesian principles to explore action spaces statistically, followed by gradual learning of action groups and communicative preferences [4, 8]. In this paper we show that the head pose is not directly usable as the gaze vector in most cases, and seek to improve the estimation for the gaze vector by learning an interpolation function. As in [4], we assume that the eye region does not provide sufficient resolution for gaze direction estimation, and we do not make use of the eyes.

The outline of the paper is as follows. In Section 2, a short overview of the algorithm is presented. Sections 3 and 4 elaborate on head pose and gaze direction estimation algorithms, respectively. The resolution of focus of attention and the segmentation of the object of interest are given in Section 5. We define three quality measures in Section 6, under which we report experimental results in Section 7.

2. General overview of the system

The basic steps of the method are given as a pseudocode in Algorithm 1. The algorithm is initialized with a frontal face image, which initiates a sequence of interaction with a human, henceforth called the experimenter. The face detection step can be performed until a frontal image is obtained. Subsequently, the initialization assumption is not an overly restrictive assumption. The first step of the proposed algorithm is to adapt an elliptic cylindrical model based on the face region for estimating the head pose. The head pose is tracked with a Lukas-Kanade method, producing head pose angles for each processed frame.

Once the pose angles are determined, two different neural network regressors are employed in the estimation of gaze direction and depth of the object of interest along the path of the gaze vector. The intersection of these two yields a coarse estimate for the center of object of interest. Assuming that the experimenters attention is deployed for more

978-1-4244-4799-2/09/$25.00

than 0.2 seconds for any object of significance, estimates from five consecutive frames are pooled to give a more robust decision. Each estimate corresponds to a probability distribution of pre-determined size around a mean location. The combination gives a broad indication of the feasible region for target object. Finally, feature saliency is computed within this region to segment out the object of interest.

Algorithm 1 Object Detection Through Joint Attention.

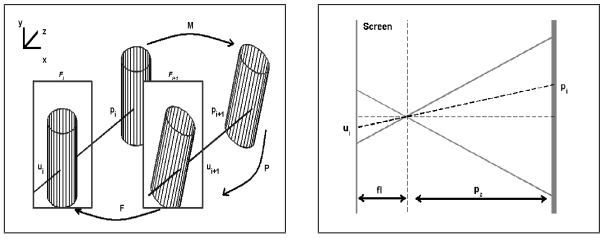
```
Initialize algorithm
while receiving visual input do
    Use Viola-Jones algorithm to detect a frontal face
end while
Adapt cylindrical head model and initialize pose
while receiving visual input do
    for i= 1 to 5 do
        Get the next frame
        Update head pose via Lukas-Kanade algorithm
        Estimate gaze direction by NN regression
        Estimate depth of the object by NN regression
        Pool estimates
    end for
    Determine estimated target region
    Compute saliency
    Perform object segmentation
end while
```

3. Head pose estimation

This section details the real-time head tracking and 3D head pose estimation algorithm, which is based on Lucas-Kanade optical flow method [6]. A single pin hole camera model is assumed for perspective projection. The human head is modeled as an elliptic cylinder and the 3D head model is superposed on the detected face area, which is found by Viola-Jones algorithm [10].

The pose of the head is represented as a pose vector $\mathbf{p_0} = \left[r_x^0, r_y^0, r_z^0, t_x^0, t_y^0, t_z^0\right]$, which is a collection of rotation and translation parameters. This vector is initialized by assuming that the initial frame F_0 contains a fully frontal face, where the eye-contact is established between the agent and the experimenter, and a session of joint attention is initialized. From this point on, the agent tracks the gaze of the experimenter. For frame F_0, the pitch and yaw angles (r_x^0 and r_y^0) are both set to zero. The roll angle (r_0^z) is initialized to the arctangent of the angle between the horizontal axis and the line connecting the two eye centers. The translations along $\mathbf{x}$- and $\mathbf{y}$-axes are initialized in relation with the radius of the elliptic cylinder and the face region obtained from the Haar classifier after a normalization with respect to the center point of the image. The depth of the head, t_z^0, which describes the distance of the head from the camera, is set to a pre-determined value.

The 3D coordinates of each point with respect to the reference frame is determined by regularly sampling points on the cylinder, which are then projected on to the 2D image plane. Initially the cylindrical head model is assumed to be centered and aligned along $\mathbf{y}$-axis of the reference frame as seen in Figure 1-(a). Any point $p = (p_x, p_y, p_z)^T$ on the cylinder satisfies the following explicit equation:



(a) (b)

Figure 1. (a) Orientation of the cylinder and its visualization on image plane, (b) Perspective projection of point p onto image plane by a pin hole camera assumption.

$$\left(\frac{p_x}{\mathbf{r_x}}\right)^2 + \left(\frac{p_z}{\mathbf{r_z}}\right)^2 = 1, \tag{1}$$

where $\mathbf{r_x}$ and $\mathbf{r_z}$ stand for the radii of the ellipse along $\mathbf{x}$- and $\mathbf{z}$-axes, respectively. The visible part of the cylinder is sampled from a $N_s \times N_s$ grid on $x-y$ plane and corresponding depth values are obtained by using Equation 1. These points are used in the Lucas-Kanade optical flow algorithm.

We use a perspective projection to obtain the 2D pixel coordinates in the image plane. Let $p = (p_x, p_y, p_z)^T$ in Figure 1-(a) be a point sampled from the surface of the cylinder and $u = (u_x, u_y)^T$ be its projection on the image plane. Figure 1-(b) illustrates this setting and the simplifying pin hole camera assumption. Using similarity of triangles in Figure 1-(b), the following equations apply for the relation between p and u:

$$\begin{aligned} p_x &= \tfrac{p_z u_x}{fl}, \\ p_y &= \tfrac{p_z u_y}{fl}, \end{aligned} \tag{2}$$

where fl stands for the focal length of the camera. The perspective projection function $\mathbf{P}(p) = u$ maps the 3D points onto the 2D image plane. As seen in Figure 1-(a), the cylinder is observed at different locations and with different orientations at two consecutive frames F_i and F_{i+1}. This is expressed as an update in pose vector $\mathbf{p_i}$ by the rigid motion vector $\Delta\mu_i = [\omega_x^i, \omega_y^i, \omega_z^i, \tau_x^i, \tau_y^i, \tau_z^i]$. In order to compute this motion vector, we need to establish the relation between p_i and u_i of F_i and their corresponding locations on F_{i+1}. In formulation of this relation, three transformation functions are employed as illustrated in Figure 1-(a). The 3D transformation $\mathbf{M}$ maps p_i to p_{i+1}, whereas the 2D transformation $\mathbf{F}$ maps u_i to u_{i+1} and the perspective projection function $\mathbf{P}$ maps p_i to u_i.

Let p_i denote the 3D location of a point sampled on the cylinder at frame F_i. The new location of the point at F_{i+1} is found by applying the transformation model, $\mathbf{M}$,

which is represented by a rotation matrix $\mathbf{R}$ corresponding to $(\omega_x^i, \omega_y^i, \omega_z^i)$ and a translation vector $\mathbf{T} = (\tau_x^i, \tau_y^i, \tau_z^i)^T$,

$$p_{i+1} = \mathbf{M}(p_i; \Delta\mu_i) = \mathbf{R}p_i + \mathbf{T}.$$

The location of the projected point on F_{i+1} is found by using the 2D parametric function $\mathbf{F}$ and applying the rigid motion vector $\Delta\mu_i$, which summarizes the motion between t_i and t_{i+1}:

$$u_{i+1} = \mathbf{F}(u_i; \Delta\mu_i).$$

If illumination is assumed to be constant across consecutive frames, the rigid motion vector can be obtained by minimizing the difference between the two image frames:

$$I(\mathbf{F}(u_i; \Delta\mu_i)) = I(u_{i+1}),$$

$$min(E(\Delta\mu_i)) = \sum_{u_{i+1} \in \Omega} \{I(\mathbf{F}(u_i; \Delta\mu_i)) - I(u_i)\}^2,$$

where Ω stands for the set of points sampled on F_i and which are still visible on F_{i+1}. The minimization problem is solved by the Lucas-Kanade method [6]:

$$\Delta\mu_i = -(\sum_{u \in \Omega} (I_u F_{\Delta\mu_i})^t (I_u F_{\Delta\mu_i}))^{-1} \sum_{u \in \Omega} (I_t (I_u F_{\Delta\mu_i})^t),$$

where I_u an I_t are the spatial and temporal image gradients, and $F_{\Delta\mu}$ is the interframe distance. The projection of the point at $t = t_{i+1}$ can be expressed in terms of the 3D location of the point at t_i and the rigid motion vector as:

$$u_{i+1} = \mathbf{P}(\mathbf{M}(p_i, \Delta\mu_i)).$$

Between two consecutive frames, the rotation can be assumed to be very small, thus the rotation matrix $\mathbf{R}$ can be approximated as [2]:

$$\mathbf{R} = \begin{bmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{bmatrix}.$$

Hence, the explicit representation of the perspective projection function in terms of the rigid motion vector parameters and the previous coordinates of the point becomes:

$$\mathbf{P}(\mathbf{M}(p_i, \Delta\mu_i)) = \begin{bmatrix} p_i^x - p_i^y\omega_z + p_i^z\omega_y + \tau_x \\ p_i^x\omega_z + p_i^y - p_i^z\omega_x + \tau_y \end{bmatrix} \times \frac{fl}{-p_i^x\omega_y + p_i^y\omega_x + p_i^z + \tau_z}.$$

4. Gaze direction estimation

Head pose estimation is primarily used to determine the focus of attention of a person. Wu and Toyama previously developed a method that is based on fitting an ellipsoidal head model to the 2D video image to estimate the pose angle, not unlike our approach detailed in the previous section [11]. This method was employed in Hoffman et al. to follow the gaze of the instructor in a shared-attention scenario [4].

Once the face is localized, an analysis of the eye region can reveal useful information about the gaze direction. While humans are very successful at estimating the gaze direction of an interacting party, several factors make this task very challenging for computers: Eyelids create occlusion problems, face morphology effects eye shape and characteristics, light conditions and reflectance change the appearance of the iris [3]. In this application we assume that the preconditions for such detailed processing are not met, and we rely on the estimated head pose to reveal the gaze specifics.

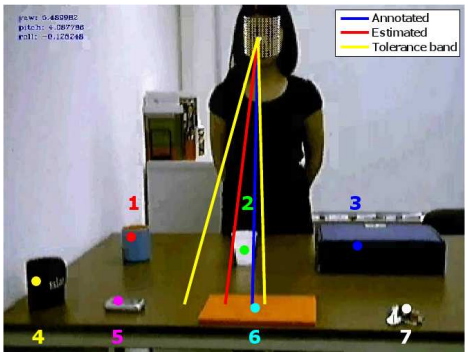


Figure 2. The experimental setup. Object indices and centers manually annotated. The lines show the gaze ground truth, estimated gaze direction, and a tolerance band around it.

The head pose is certainly indicative of the gaze direction. However, it does not completely specify the gaze direction, since gaze involves the eye movements in addition to the head pose. In [12], we have shown that the deviation of gaze direction from the head pose direction is a predictable, but nonlinear function. In the absence of restricting assumptions with regards to the context of the application (as in [9], for instance), as well as low-resolutiosn face information, we opt for a two-layer backpropagation neural network to interpolate the gaze direction from given 3D head pose vector estimates [1]. The training samples required for the supervised training of the neural network are obtained by manual annotation of the target object locations and the ground truth is composed of the slope of the gaze direction vector. The nonlinear nature of the mapping suggests a multi-layer architecture as opposed to having a single layer. As in [12], a second neural network is used to estimate the depth of object of interest along the estimated gaze direction.

5. Saliency model

Once gaze direction is estimated, one needs to determine a feasible region for searching focus of attention. Consid-

ering that the head pose provides a coarse estimate for gaze direction, a tolerance interval is positioned around the gaze vector and the resulting conic region is regarded as an initial feasible region as in Figure 2. To determine the focus of attention of the experimenter, we employ the popular bottom-up saliency scheme proposed by [5]. This approach decomposes the saliency of a scene into separate feature channels. The presence of illumination intensity, colors, oriented features and motion are indicative of salient locations in the scene. Each feature channel is separately used to determine a feature-specific saliency map, which are then combined to a saliency master map. In the original model, the saccadic eye movements are simulated by directing a foveal window to the most salient location, determined by a dynamic and competitive Winner-Take-All (WTA) network [5]. Once a location is selected, it is suppressed by an inhibition-of-return mechanism to allow the next most-salient location to receive attention.

We use this model for determining the most salient object in the gaze direction of the experimenter. If there is more information available as to the experimenters 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. For instance in [4], the probability that an experimenter selects a particular object is learned by fitting a Gaussian mixture model on the pixel distribution.

Using saliency to fixate on the interesting objects reduces the uncertainty in the estimation of the gaze direction. In our model, the bottom-up saliency model receives a modified image from the gaze estimation module, where a particular region around the estimated gaze retains image information and the rest of the visual field is suppressed. This forces the WTA to attend only to salient parts within the gaze cone.

6. Quality measures

We employ three measures for quantifying the performance in detection of the focus of attention. We first check the estimated gaze direction and compare it to the ground truth. We then manually annotate the bounding box for each object and consider the ratio of the estimates falling inside the bounding box to the number of all estimates for a particular object to denote the performance rate in detection of focus of attention.

Quality measure Q_1 indicates the error as deviation of gaze direction estimated from the ground truth, where both are measured on the image plane. It is the absolute difference between estimated gaze direction and the annotated object center, denoted as $\Delta\gamma$ and measured in radians. This value should be as close to zero as possible.

Quality measure Q_2 indicates the number of times the target object falls within the estimated gaze area. Since the segmentation step can recover from gaze estimation errors, it is important to distinguish between cases of complete miss and cases where the gaze cone touches the object, and with high probability the saccadic search will visit the correct object in time. Thus, Q_2 is the ratio of times the estimated gaze intersects the bounding box of the target object to all estimates. This value is ideally close to unity.

Finally, Q_3 measures how much the algorithm deviates from the correct object center on image plane, in relation to the rest of the objects. Let the true object centers be denoted by $\mathbf{a_i}$, with $i = 1, \ldots, N$, and N being the number of objects in the scene. If a target estimate for object j is denoted by $\mathbf{e_j}$,the distance to the true object is:

$$D_t = \|\mathbf{e_j} - \mathbf{a_j}\| .$$

The distance to the closest non-target object to the estimate can be denoted by

$$D_c = \arg\min_{i, i \neq j} \|\mathbf{e_j} - \mathbf{a_i}\| .$$

The Q_3 measure pools the ratio of deviations of these distances:

$$Q_3 = \sum \frac{D_t}{D_c}.$$

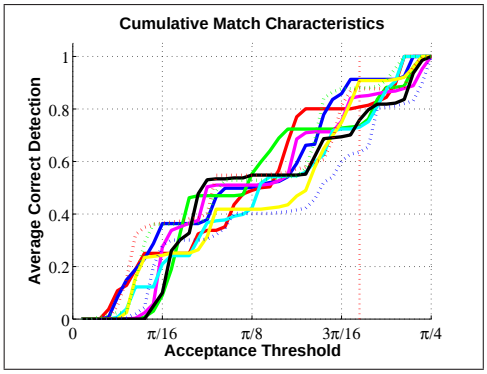
While using this measure, we expect to have values smaller than 1, meaning that the distance to the target object is smaller than the distance to the closest non-target object.

7. Experimental results

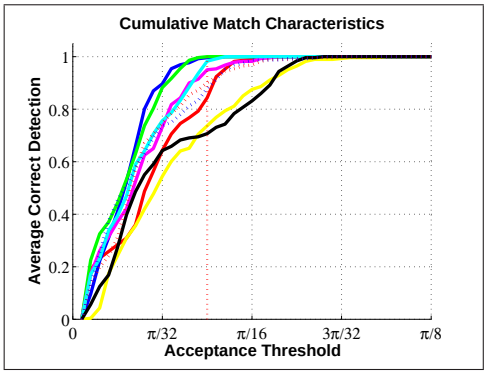
The experiments use ten video sequences recorded at $25fps$ for a total of 4211 frames. All videos are recorded in the same envorinment, where the experimenter and the objects do not change. For all experiments, we employ a ten-fold cross validation schemeand the mean and standard deviations are reported for ten folds. The ground truth is manually annotated.

In the first experimental setting (**only head**), the head pose is assumed to be exactly the same as gaze direction, and the tolerance band is positioned directly around the pose vector. In the second setting (**head + gaze**), the neural network regressor for the gaze estimation is taken into account. Finally, for the third setting (**head + gaze + depth**), the neural network regressor for the depth estimation is used as well to determine the focus of attention.

We position a tolerance interval ($\pm\tau$, expressed as angular deviation) around the estimated gaze vector to search for the object of interest. The impact of this tolerance parameter is shown with a cumulative match characteristics curve (CMC). The CMC curve plots the accuracy of the system for a range of τ values. This curve is shown separately for



a



(b)

Figure 3. Cumulative Match Characteristics (CMC) curves for estimation of tolerance intervals around (a) head pose and (b) gaze direction vectors.

head pose estimates and gaze estimates in Figure 3. Figure 3-(a) shows that the head pose, by itself, cannot appropriately constrain the search area for the target object. The curve in Figure 3-(b), on the other hand, shows that a tolerance interval of $\pi/64$ leads to a reasonable detection rate.

Since accuracies depend on the placement of the objects, we partition the objects into groups that indicate distance from the experimenter (i.e. **near** and **far**, about 50cm. apart), as well as into groups that indicate angular distance from the frontal gaze direction (i.e. **central** and **peripheral**, about 45° apart). The average deviation from target in radians is 0.04 in the near and peripheral conditions, and 0.06 in the central and far conditions, respectively. The gaze direction is correctly estimated in the majority of cases, and there are no significant differences between object groups. Furthermore, it is observed that the difference presents an acceptable deviation, close to the tolerance value derived from Figure 3-(b).

For the **head + gaze + depth** setting, we have an additional depth estimation module, which enables us to get an intersection of gaze vector and object depth, resulting in a coarse object center estimation. Figure 4 shows these intersection points, indicating that the coarse estimates for object locations are reliable. The object location is represented by a Gaussian distribution centered around the initial estimate with a standard deviation of 50 pixels, which provides a window width sufficiently close to the one derived from the CMC curve.

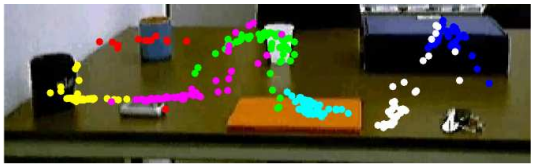
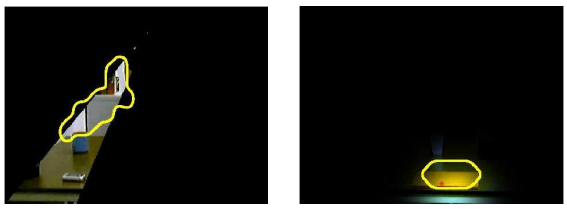


Figure 4. Estimates for object center location.

Since human eye makes three to five saccades per second, it is not realistic to form a feasible region for each video frame and then compute saliency for a $25fps$ rate. Therefore we form bins of five consecutive frames and calculate a feasible region for each of them. Since we do not expect the focus of attention to change drastically in this short time interval, we perform a smoothing operation on the head pose vectors, gaze directions and estimated object centers.

Figure 5 presents two feasible regions (for one bin) for the (**head + gaze**) and (**head + gaze + depth**) settings, respectively. For the (**only head**) setting, taking $\pi/5$ radians around the head pose estimate results in a large feasible region, where part of the background and several objects are visible. For the setting (**head + gaze**), comparing Figures 3-(a) and (b) it is clear that the tolerance band is narrower, giving a smaller region to look for the object of attention. However, the background is still present, and that may lead to a confusion, as in the particular case shown in Figure 5-(a). The feasible region is reduced in the third setting, and only the object of interest is present. The segmentation is shown in Figure 5-(b), with the object center indicated in red.

The estimated salient region centers for one video sequence are depicted in Figure 6. In the (**only head**) setting, the presence of the white cup in most of the feasible regions lead to misleading decisions. In Figure 6-(a), the drawback introduced by the background can be seen clearly. The objects on the sides are mostly mapped to the back-

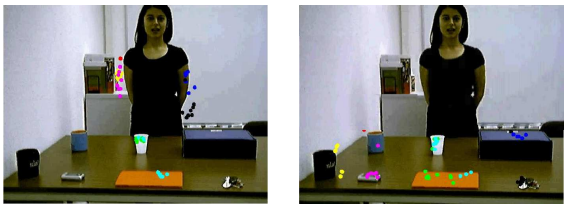


(a) (b)

Figure 5. Sample feasible regions and saliency-based segmentation for (a) **head + gaze** and (b) **head + gaze + depth**.

Table 1. Performance quantification for (**only head**), (**head + gaze**) and (**head + gaze + depth**).

	(**only head**)		(**head + gaze**)		(**head + gaze + depth**)	
	Q_2	Q_3	Q_2	Q_3	Q_2	Q_3
	μ	$\mu \pm \sigma$	μ	$\mu \pm \sigma$	μ	$\mu \pm \sigma$
$near$	0.33	11.38 ± 2.85	0.33	0.67 ± 0.02	0.87	0.45 ± 0.63
far	0	26.03 ± 17.41	0.16	2.17 ± 0.15	0.72	1.03 ± 1.69
$central$	0	14.60 ± 12.69	0.55	1.67 ± 1.47	0.80	0.81 ± 1.49
$peripheral$	0.25	23.88 ± 10.03	0	1.47 ± 0.06	0.76	0.76 ± 1.04



(a) (b)

Figure 6. Estimated object center locations for Cases (a) (**head + gaze**) and (b) (**head + gaze + depth**).

ground where the ones in the middle are determined with a higher accuracy, as the background is not present in the corresponding feasible regions. For the (**head + gaze + depth**) setting, we see that the estimated salient region centers are more accurate when compared to the (**head + gaze**) setting. Moreover, compared to Figure 5, which shows the intersection of gaze direction vector and estimated depth, these points are closer to annotated object centers and fall mostly within the object boundaries. These intuitive results are confirmed by quality measures Q_2 and Q_3. Table 1 summarizes the performance values for the settings (**only head**), (**head + gaze**), and (**head + gaze + depth**), respectively. It is clear that the extension of the method with gaze direction estimation and depth estimation leads to a significant improvement in the performance.

There are several factors leading to degradation in performance. First of all, it is harder to find the small objects, since they are defined by a smaller bounding box. This has a bearing on quality measure Q_2. On the other hand, location of the object on the table affects the performance rate as well. As the yaw and pitch angles increase, the head pose is harder to determine, since the view is less similar to the template obtained from a frontal view. In that case, it is more probable that the gaze direction and consequently the estimated object location deviate from the correct localization.

8. Conclusions

We have proposed a method for determining an attended object location by using head pose estimates, which is useful for establishing joint attention between a human and an embodied agent. Our model uses estimation of head pose, correction for gaze direction, and attention-based selection for finding objects attended by the experimenter. We point out to a shortcoming in the literature, in which the head pose is taken for specifying the focus of attention. We seek to remedy this by employing a neural network regressor that interpolates the gaze direction from the head pose. We also use a second regressor to further reduce the target search area by estimating the depth of the object of focus in the gaze field. By this means, we provide a first approximation to an otherwise complex cognitive phenomenon.

9. Acknowledgments

This research is supported by the Dutch BRICKS/BSIK and TUBITAK BTT-105E065 projects.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, USA, 1995.

[2] C. Bregler and J. Malik. Learning Appearance Based Models: Mixtures of Second Moment Experts. *Advances in Neural Information Processing Systems*, pages 845–851, 1997.

[3] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on PAMI*, 2010.

[4] M. Hoffman, D. Grimes, A. Shon, and R. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19(3):299–310, 2006.

[5] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on PAMI*, pages 1254–1259, 1998.

[6] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. IJCAI*, volume 3, 1981.

[7] R. Moratz and T. Tenbrink. Affordance-Based Human-Robot Interaction. *LNCS*, 4760:63–76, 2008.

[8] A. Shon, J. Storz, A. Meltzoff, and R. Rao. A Cognitive Model of Imitative Development in Humans and Machines. *International Journal of Humanoid Robotics*, 4(2):387, 2007.

[9] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *Proc. Seventh ACM Int. Conf. on Multimedia (Part 1)*, pages 3–10, 1999.

[10] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *IEEE CVPR*, volume 1, 2001.

[11] Y. Wu and K. Toyama. Wide-range, person-and illumination-insensitive head orientation estimation. In *Proc. IEEE Conf. on Automatic Face and Gesture Recognition*, pages 183–188, 2000.

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