Head pose and neural network based gaze direction estimation for joint attention modeling in embodied agents

Zeynep Yücel (Zeynep@Ee.Bilkent.Edu.Tr)
Centre for Mathematics and Computer Science (CWI),
Science Park 123,
1098 XG, Amsterdam, The Netherlands

Albert Ali Salah (A.A.Salah@Cwi.Nl)
Centre for Mathematics and Computer Science (CWI),
Science Park 123,
1098 XG, Amsterdam, The Netherlands

Abstract

Imitation is a powerful capability of infants, relevant for bootstrapping many cognitive capabilities like communication, language and learning under supervision. In infants, this skill relies on establishing a joint attentional link with the teaching party. In this work we propose a method for establishing the joint attention between an experimenter and an embodied agent. The agent first estimates the head pose of the experimenter, based on tracking with a cylindrical head model. Then two separate neural network regressors are used to interpolate the gaze direction and the target object depth from the computed head pose estimates. A bottom-up feature-based saliency model is used to select and attend to objects in a restricted visual field indicated by the gaze direction. We demonstrate our system on a number of recordings where the experimenter selects and attends to an object among several alternatives. Our results suggest that rapid gaze estimation can be achieved for establishing joint attention in interaction-driven robot training, which is a very promising testbed for hypotheses of cognitive development and genesis of visual communication.

Keywords: Head pose estimation; gaze following; selective attention; saliency; communication; joint attention; neural networks; robotics; autonomous mental development; imitation-based learning.

Introduction and Motivation

Embodied agents are vital tools for testing developmental hypotheses in controlled simulation environments. In the last few years, there is a marked effort to create robots that learn like babies do. These systems allow the experimenter to test ranges of experimental conditions under similar assumptions and obtain quantitative results about the preconditions and developmental stages of various skills. They are particularly relevant for studying how joint attention develops, as real infants must be properly motivated and coerced into following the experimenter's attention, which imposes certain limitations on the experimental setup (Flom, Deák, Phill, & Pick, 2004). In this paper we propose an algorithm that allows an embodied agent to establish a joint-attentional link with the experimenter. This skill is also important for developing language and communication, as well as for *imitation-based learning*, which allows the experimenter to demonstrate a behaviour rather than explicitly design algorithms to produce the behaviour in the agent.

Recent models of imitation-based learning rely on Meltzoff and Moore's active intermodal mapping (AIM) framework for action imitation learning (1997). Important work in this area includes (Shon, Storz, Meltzoff, & Rao, 2007) and (Hoffman, Grimes, Shon, & Rao, 2006), which use Bayesian principles to explore action spaces statistically, followed by gradual learning of action groups and communicative preferences. In (Hongeng, Nevatia, & Bremond, 2004) a goal-based action model is used to classify intentional actions in a controlled environment. The embodied agent extracts a large number of visual features from the scene and by tracking the trajectory of the experimenter's hand, determines which of the predefined actions is being performed. As a common factor of most research in this area, 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 (Moratz & Tenbrink, 2008).

In experiments concerning human robot interaction, the learned structure of a visual scene provides additional cues to the embodied agent in guessing the focus of attention of the communicating party. Subsequently, most approaches incorporate *saliency* as a part of the joint-attention system, and select appropriate saliency measures that will indicate what is inherently interesting in the scene depending on the application domain. The saliency can be a function of natural image statistics. For instance in (Nagai, Hosoda, Morita, & Asada, 2003), 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 also be incorporated by modulating bottom-up channels, or by explicitly adding dedicated saliency components. Faces are important for the natural interaction settings, consequently they are separately detected and made salient. In what follows, the party that interacts with the embodied agent is referred to as the *experimenter*.

When the saliency of a scene is determined, a visual feedback controller provides motor control commands to direct the gaze of the robot to a salient location, both for attending to the face of the experimenter and to other objects in the environment. Estimating the head pose of the other party is a visual skill necessary for joint attention modeling. In (Nagai et al., 2003) a separate module learns to associate facial ap-

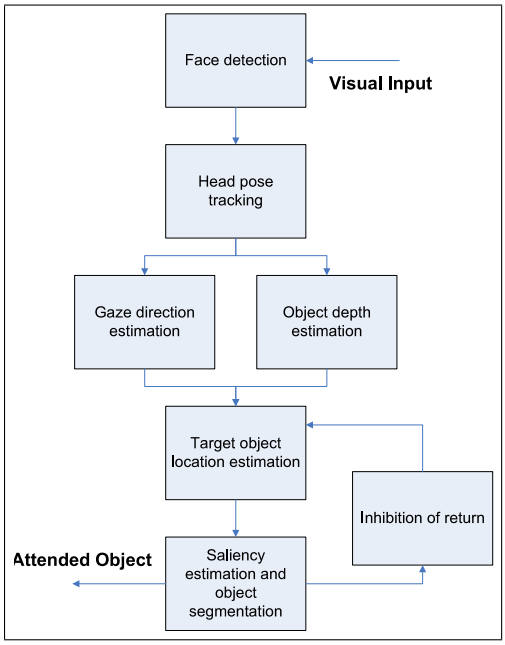


Figure 1: Basic steps of the algorithm.

pearance of the experimenter with angles that specify its pose. For each facial appearance in the training set, the head orientation is manually annotated.

The distinction of following the head pose and the gaze direction itself is an important one that we explicitly stress in this paper. It appears that young infants first follow the head movements of others, and only in time develop the ability to follow the gaze direction (Corkum & Moore, 1995). 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 (Hayhoe, Land, & Shrivastava, 1999 ; Triesch, Jasso, & Deák, 2007). Estimating the gaze direction has received a lot of attention for obvious reasons, see (Hansen & Ji, to appear) for a recent overview of eye and gaze models.

In the next section, we describe a fast model for joint attention modeling, which is based on estimating the head pose of the experimenter. The individual components of the system are described with dedicated sections, followed by our experimental results.

Overall Description of the Model

The basic steps of the proposed algorithm are summarized in Figure 1. The first step of the proposed metod is detecting the face of the experimenter with the Viola-Jones algorithm (Viola & Jones, 2001). The details of this step are omitted, as the method is fairly mainstream, and a widely used implementation exists in the OpenCV library. The head pose of the experimenter is tracked by adapting a 3D elliptic cylindrical model to the face region. The pose vector consists of the roll, pitch, and yaw angle parameters of the cylinder. Once the pose angles are determined, a neural network regressor estimates the gaze direction. This step is necessary, since small head pose changes towards peripheries of the visual field are usually indicative of larger deviations of the gaze direction. We assume that the embodied agent is not sufficiently stable to extract an accurate estimate of the gaze direction directly by analysing the eye and iris area of the experimenter.

A second neural network regressor is used to estimate the distance of the target object along the gaze direction. These two estimates are probabilistically combined to yield a coarse estimate for the center of the target object. By pooling a estimates from a number of consecutive frames, a more robust decision on the target is generated.

The rough localization of the attended object is refined by a bottom-up saliency scheme, which also segments out the target object. If 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 now describe each of these steps in more detail.

Head Pose Tracking

The real-time head tracking and 3D pose estimation algorithm is initialized using the popular Viola-Jones face detection method, which employs an Adaboost classifier with Haar wavelet features (Viola & Jones, 2001). The 3D pose estimation is implemented via continuous tracking with the Lucas-Kanade optical flow method (Lucas & Kanade, 1981).

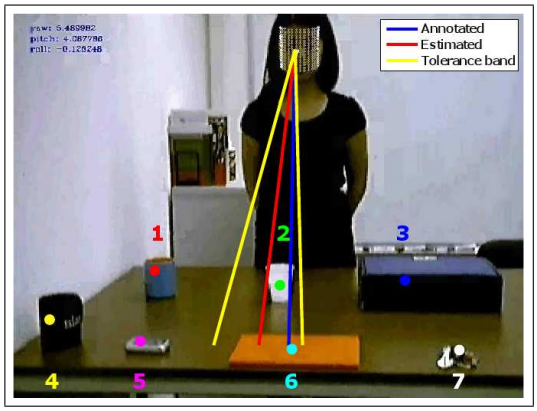
The pose of the head relating frame F_t at time t is represented with a pose vector $\mathbf{p_t}$, which is initialized by assuming that F_0 contains a fully frontal face, where the eye-contact is established between the agent and the experimenter. Thus the rotation parameters are all set to 0 and the translation parameters are initialized considering the detected face location of the experimenter.

For simplicity and fast computation, the 3D motion is summarized by a set of points that are obtained by regular sampling on the cylinder surface (See Figure 2 (a)). The relation between these points and their corresponding projections on the 2D image plane is established by a perspective projection based on a simple pin hole camera model. Let p_i be a point sampled from the surface of the cylinder at F_i and u_i be its projection on the image plane. If 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 $\Delta\mu_i$,

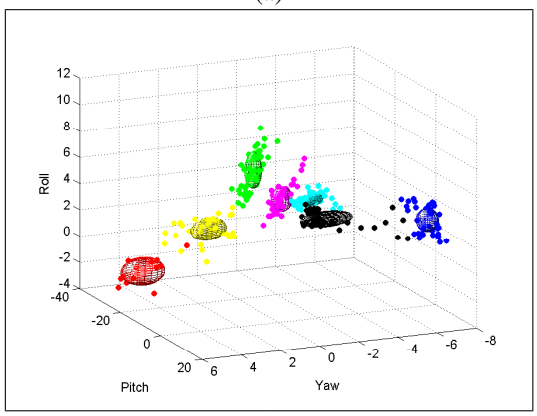
$$\mathbf{p_{i+1}} = \Delta\mu_i \mathbf{p_i}.$$

In order to compute this motion vector, we need to establish the relation between p_i and u_i for F_i and their corresponding locations on F_{i+1}. The new location of the point at F_{i+1} is found by projecting u_i onto the cylindrical model, applying the pose update and mapping back to the image coordinates. If the intensity of the pixel $I(u)$ is assumed to be constant between the images, the pose update satisfies:

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



(a)



(b)

Figure 2: (a) The experimental setup. Object indices and centers are manually annotated by the user. The annotated gaze direction points towards the object centre, the estimated gaze directions are shown with a tolerance band around it. (b) Distribution of pose angles.

where I_u and I_t are the spatial and temporal image gradients, respectively (Lucas & Kanade, 1981). Solving for $\Delta\mu$ for each frame F_i, we obtain a continuously updated pose vector for all frames in the sequence. For further details, the reader is referred to (Valenti, Yücel, & Gevers, 2009).

Gaze Direction and Target Depth 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 (Wu & Toyama, 2000). This method was also employed to follow the gaze of the instructor in a shared-attention scenario (Hoffman et al., 2006).

The head pose is certainly indicative of the gaze direction. However, it does not completely specify the gaze direction, since gaze involves eye movements, in addition to the head pose. Our experimental setup involves an experimenter looking at several objects placed on a flat surface, shown in Figure 2 (a). Figure 2 (b) illustrates the distribution of head pose angles obtained as the experimenter looks at each object for a few seconds. The head pose angles are grouped (and coloured) according to the target object, which reveals a clear clustering, as well as the nonlinear nature of the relation between head pose and gaze direction.

Some approaches resolve gaze direction from head pose implicitly by incorporating additional assumptions. For instance in (Stiefelhagen, Yang, & Waibel, 1999), 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 of the gaze. In this paper we assume that precise eye-center positioning and 3D interpolation of the gaze vector in real time is not realistic for the embodied agent. We use a two-layer backpropagation neural network to interpolate the gaze direction from a given 3D head pose vector estimate (Bishop, 1995).

The input layer of the feedforward artificial neural network receives the three-dimensional estimated pose vector and maps this input to a gaze direction, represented by a single angle on the image plane. We have used 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 $(-0.5, 0.5)$ interval. A validation set is monitored for error decrease to prevent overfitting. The training samples required for the supervised training of the neural network are obtained by manual annotation of the target object location for each frame of the video.

Our experiments indicate that the angle with which the head is turned towards the focused object underestimates the actual gaze direction, both horizontally and vertically. Figure 3 illustrates the estimated gaze direction through head pose computation and the gaze direction estimated through the neural network regressor. The neural network interpolation (or extrapolation, in most cases) achieves both 3D and 2D coordinate mapping, and provides more accurate estimates of the gaze direction. For the particular case presented in Figure 3, the improvement per frame is 0.14 radians as measured on the image plane.

The actual depth is not specified by the gaze direction vector on the image plane, yet this information is present to some extent in the 3D head pose vector estimated in the first step. Therefore another neural network module is trained to obtain an estimate for the depth of object of interest. The parametrization is similar to the first regressor, with three input values and a single depth value measured from the head centre as the output value.

Target Object Location Estimation

Once the gaze direction is estimated, we determine a feasible region for directing the focus of attention. The estimate for the gaze direction is allowed a tolerance interval, shown in Figure 2 (a) with a yellow gaze cone, and the target for the joint attention is assumed to fall within this gaze cone. We have experimentally set the maximum deviation from the es-

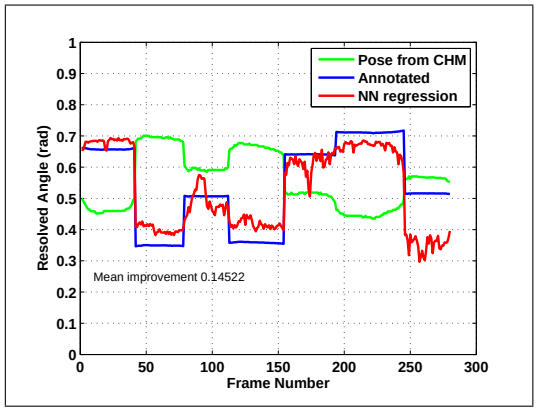


Figure 3: Improvement of gaze direction estimation over head pose estimation introduced by neural network regression.

timated gaze angle to $3\pi/64$. The rough depth estimate helps to further narrow down the search.

The intersection of the gaze vector and the line of depth gives a single point in the image plane, indicative of the best estimate for the target location. Examples of the resulting estimates for focus of attention on the 2D image plane are presented in Figure 4.

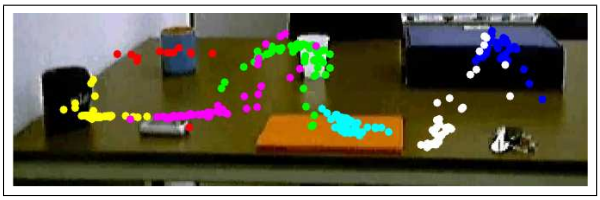


Figure 4: Estimates for focus attention. This figure is best viewed in colour.

Saliency Model

Once the gaze direction is estimated, the agent attempts to determine the focus of attention of the experimenter. For this purpose, we employ the popular bottom-up saliency scheme proposed in (Itti, Koch, & Niebur, 1998). This approach is based on the feature integration theory of Treisman and Gelade, and 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 (Itti et al., 1998). 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 immediate neighbourhood of the estimated target location. The tracking and interpretation of the head pose itself is noisy, and by itself not sufficient to single out the target. 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, by for instance modulating the responses of individual feature channels appropriately. In Hoffman et al., 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 presented to the system would optimistically bias the results.

Using saliency to fixate on the interesting objects serves a twofold 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. A consequence of this learning is the developing ability of the infant to estimate the attention focus of the experimenter even when it lies beyond the visual field of the child.

Secondly, saliency-based grafting compensates the discrepancy between intended motor commands and executed physical actions, an issue 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.

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.

Since human eye makes three to five saccades per second, it is not realistic to compute saliency for a $25 fps$ rate. Therefore we form bins of consecutive frames by considering five consecutive frames to belong to the same bin and calculate the 2D location of focus of attention for each of them. Since we do not expect the focus of attention change drastically in this short time interval, we perform a smoothing operation on the estimated point by using a low pass filter. Five Gaussian distributions are then positioned around the resultant estimates and an eventual feasible region is obtained. Saliency computation followed by object segmentation is performed in the eventual feasible region and thereby the object of interest is resolved. It is observed that in most of the cases the coarse object location estimates fall on the object.

Experimental Results

We have collected ten video sequences at $25 fps$ for a total of 4211 frames, where the ground truth for experimenter's attention is manually annotated. The results are reported by ten-fold validation, where one session is used for training, and the remaining nine are used to evaluate the accuracy of the system for each fold. The mean values are reported for ten such batches. For each sequence, the experimenter focuses on each of the seven objects for several seconds in random order. 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**), as well as into groups that indicate angular distance from the frontal gaze direction (i.e. **central** and **peripheral**).

We assume that if the computed focus of attention is sufficiently close to the target object, the detection is successful. The threshold for accepting success, however, can be determined arbitrarily. In order to determine a reasonable value for the tolerance interval around the estimated gaze direction, we inspect the cumulative match characteristics curve (CMC), given in Figure 5. The CMC curve plots the accuracy of the system for a whole range of thresholds, where a particular value τ of the threshold means that angular deviations from the target less than τ are acceptable at this stage. The final threshold to be used in the actual deployment also depends on the attention module; a larger threshold means that a larger area needs to be searched by the attention module, and increases the probability of off-object fixations. We determine from the curve that a tolerance interval of $3\pi/64$ leads to a reasonable detection rate.

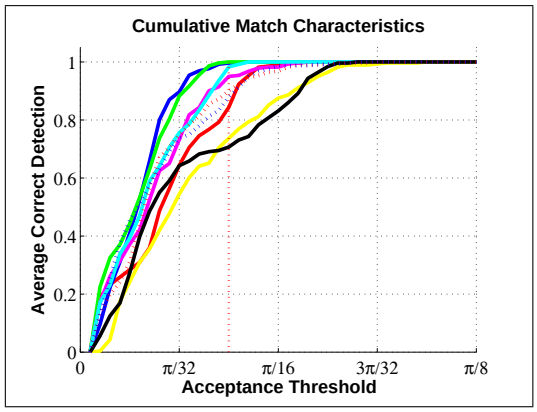


Figure 5: Cumulative match characteristics of the head-pose estimation module on test recordings.

The first row in Table 1 shows the average deviation from target in radians for the whole system, denoted by Q_1. It can be seen that 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 the CMC curve.

In order to provide comparative results indicating the contribution of each part of our proposed method, we present results in three different experimental settings. In the first 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 to determine the focus of attention. The last three rows of Table 1 show the ratio of times the estimated gaze intersects the bounding box of the target object to all estimates for each of these settings, denoted by Q_2. This value is ideally close to unity. 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.

Table 1: Performance evaluation, see text for details.

Quality Measure	*near*	*far*	*central*	*peripheral*
Q_1	0.04	0.06	0.06	0.04
$Q_2(h)$	0.33	0.00	0.00	0.25
$Q_2(hg)$	0.33	0.16	0.55	0.25
$Q_2(hgd)$	0.87	0.72	0.80	0.76

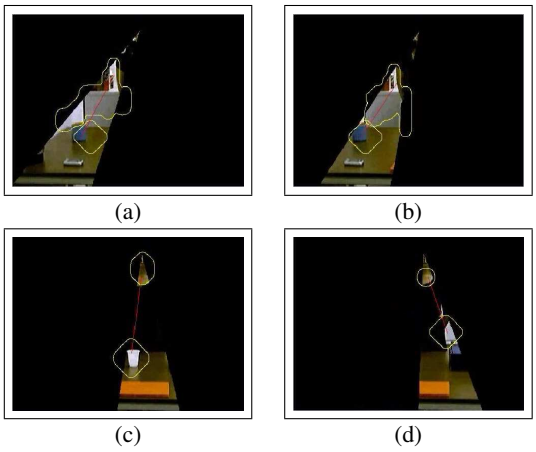
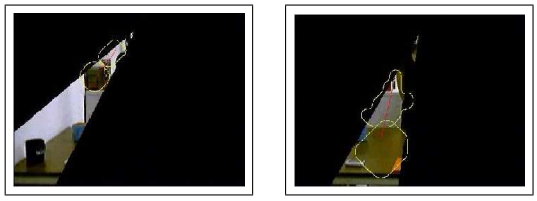


Figure 6: Example frames where the object of interest is detected.

Figure 6 illustrates several examples for which the proposed approach detects the target object. The visible image indicates the tolerance band around estimated gaze direction.

Figure 7 illustrates several example frames where the target was not detected. There are various reasons for misdetection. It may be the case that the pose vector is not estimated with high accuracy, so that the cone does not include the object of interest. The other possibility is that the pose vector is estimated correctly, but the objects falls into an image segment with more salient objects, which draw the focus of attention.



(a) (b)

Figure 7: Example frames where the object of interest is not detected.

Conclusions

We have proposed a method 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 an 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.

The proposed method is meant to provide a first approximation to an otherwise complex cognitive phenomenon. Possible future directions include direct gaze estimation by using a higher-resolution camera to inspect the eyes of the experimenter, as additional physical cues to determine the focus of attention. Yet one should not forget the contribution of context in the interaction. As Kaplan and Hafner (2006) rightly point out, the existence of top-down influences and the considerations imposed by higher-level cognitive functions make joint attention a very difficult egg to crack.

Acknowledgements

This research is supported by the Dutch BRICKS/BSIK program and TUBITAK project with grant number BTT-105E065.

Références

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, USA.

Corkum, V., & Moore, C. (1995). Development of joint visual attention in infants. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 61–83). Erlbaum.

Flom, R., Deák, G., Phill, C., & Pick, A. (2004). Nine-month-olds shared visual attention as a function of gesture and object location. *Infant Behavior and Development*, *27*(2), 181–194.

Hansen, D., & Ji, Q. (to appear). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hayhoe, M., Land, M., & Shrivastava, A. (1999). Coordination of eye and hand movements in a normal environment. *Investigative Ophthalmology & Vision Science*, *40*.

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

Hongeng, S., Nevatia, R., & Bremond, F. (2004). Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, *96*(2), 129–162.

Itti, L., Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1254–1259.

Kaplan, F., & Hafner, V. (2006). The challenges of joint attention. *Interaction Studies*, *7*(2), 135–169.

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Int. Joint Conf. on Artificial Intelligence* (Vol. 3).

Meltzoff, A., & Moore, M. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting*, *6*, 179–192.

Moratz, R., & Tenbrink, T. (2008). Affordance-based human-robot interaction. *Lecture Notes in Computer Science*, *4760*, 63–76.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). Emergence of joint attention based on visual attention and self learning. In *Proc. 2nd Int. Symposium on Adaptive Motion of Animals and Machines* (Vol. SaA-II-3).

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

Stiefelhagen, R., Yang, J., & Waibel, A. (1999). Modeling focus of attention for meeting indexing. In *Proc. seventh acm int. conf. on multimedia* (Vol. 1, pp. 3–10).

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Triesch, J., Jasso, H., & Deák, G. (2007). Emergence of Mirror Neurons in a Model of Gaze Following. *Adaptive Behavior*, *15*(2), 149.

Valenti, R., Yücel, Z., & Gevers, T. (2009). Robustifying eye center localization by head pose cues. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Viola, P., & Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* (Vol. 1).

Wu, Y., & Toyama, K. (2000). Wide-range, person-and illumination-insensitive head orientation estimation. In *Proc. Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition* (pp. 183–188).