Joint Visual Attention Modeling for Naturally Interacting Robotic Agents

Zeynep Yücel*, Albert Ali Salah†‡, Çetin Meriçli§, and Tekin Meriçli§
*Electrical and Electronics Engineering, Bilkent University - Ankara, Turkey
Email: zeynep@ee.bilkent.edu.tr
†ISLA University of Amsterdam - Amsterdam, The Netherlands
‡Centrum Wiskunde & Informatica - Amsterdam, The Netherlands
Email: a.a.salah@cwi.nl
§Department of Computer Engineering, Boğaziçi University - Istanbul, Turkey
Email: {cetin.mericli,tekin.mericli}@boun.edu.tr

***Abstract*—This paper elaborates on mechanisms for establishing visual joint attention for the design of robotic agents that learn through natural interfaces, following a developmental trajectory not unlike infants. We describe first the evolution of cognitive skills in infants and then the adaptation of cognitive development patterns in robotic design. A comprehensive outlook for cognitively inspired robotic design schemes pertaining to joint attention is presented for the last decade, with particular emphasis on practical implementation issues. A novel cognitively inspired joint attention fixation mechanism is defined for robotic agents.**

I. Introduction

The emerging field of Human-Robot Interaction is a rapidly growing research area which represents an interdisciplinary effort that addresses the need to integrate social informatics, human factors, cognitive science, and usability concepts into the design and development of a social robot. One of the main challenges that a social robot faces is the need to perceive the world as humans do and learn from the interactions with environment and humans. In order to do that, a social robot must be able to interpret human activity and behavior. The vision system of the social robot is responsible for accomplishing tasks like identifying faces, measuring head and hands poses, and recognizing gestures to emulate human social perception. In addition to that, the social robot should preferably have a similar look to humans with its animate limbs, hands, head, and face to communicate its mental state and intentions in a psychologically plausible manner.

Developmental robotics aims to design continuously learning robotic agents, which are capable of establishing natural interaction with humans in uncontrolled environments [1]. This type of agents are of special interest in real-life scenarios of human-robot interaction, as their skill progression follows a natural gradient of complexity, and promises to be robust and flexible in the face of unrestrained characteristics of natural settings.

A recent approach for building naturally interacting autonomous agents is to adopt a mutual standpoint of developmental psychology and robotics [2]. The collaboration of these fields was born out of the shift in the dominating paradigm of artificial intelligence towards situated cognition and the necessity of having embodiment to ground the physical experience of the agent, closing the sensor-actuator loop in a sense. A key insight is that the agent initially uses the environment as its own model, and gradually builds more complex representations that would enable new skills as the agent develops.

Developmental psychology examines the learning process of infants, as well as the evolution of auxiliary skills that contribute to learning. Among these skills are those that relate to the construction and restructuring of different types of memory, but also skills to actively explore the sensors and effectors of the agent, and consequently, the environment. In this paper, we focus on one such skill that enables a human infant to establish communication with a caregiver, thus playing a crucial role in supervised learning. We discuss the preconditions and mechanisms that would enable a robot to do the same.

In Section II the nuts and bolts of the joint attention mechanism in infants are discussed from an evolutionary perspective. Existing cognitive models that decompose the problem into different sets of biologically plausible functional and conceptual modules, are explained and the computational models of joint attention defined in literature are reviewed. Section III describes our approach and the robotic testbed for creating an autonomous agent capable of exhibiting joint attention with a human party.

II. Related Work

Cognitive development provides for a rich source of ideas for developing communicative skills for a robot, especially by suggesting ways for task decomposition and by identifying simpler cognitive skills. Here we briefly review joint attention models in this context.

A. Evolution of Joint Attention in Infants

Motives, which are described as the reasons behind initiation of voluntary behavior, are considered to play a crucial role in the learning process. A widely accepted classification scheme identifies two global motive classes as *drives* and *intrinsic motives* by distinguishing the role of end goals [3].

Drive theory claims that reward and punishment are the key elements which lead motivation, whereas intrinsic motivation theory focuses on ego motives. A multifaceted scheme of intrinsic motivation is described by Reiss based on 16 basic desires [4]. Although individuals prioritize these desires differently, several intrinsic motives such as social contact, status, and curiosity are prominent in terms of learning in infants. In order to establish social contact, to obtain social standing, and to fulfill the desire for knowledge, infants need to 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. It is observed that infants have a tendency to detect and track faces. Moreover, it is observed that infants are more sensitive to faces with open eyes [5]. After detecting face and establishing eye contact, the infant learns to relate this information with attention [6]. By 12 months, the infant tracks the caregiver's gaze and attends to the object of interest of the caregiver, which stands in his view of sight (geometric stage), whereas up to nine months, it attends to any salient object in his view of sight regardless of the caregiver's attention (ecological stage). After 18 months, the infant is able to turn around and track the gaze of the caregiver if the object of interest stands outside his first field of view (representational stage) [6].

B. Developmental Models for Joint Attention

"Theory of mind" expresses the collection of skills relating the attribution of beliefs, goals, and desires to other people [7]. In order to implement human-like complex social skills, one needs to develop a theory of mind from a robotic design perspective. This requires the decomposition of the process of communication into simpler cognitive skills, which can be implemented on a robot. In this respect, there are three decomposition approaches for joint attention we would like to mention here.

Module based decomposition of Baron-Cohen is one of the most prominent theories of shared attention [8]. He describes four modules as intentionality detector, eye direction detector, shared attention mechanism, and theory of mind mechanism. Although this model presents a useful decomposition for the key elements of shared attention, it provides little insight about how these mechanisms work and thus it is not convenient for robotic implementation. Kozima *et al.* describe the design of a robot that learns to communicate with human caregivers [9] based on three modules, namely intentionality, identification, and social communication, similar to the ones defined by [8].

Unlike module-based decomposition, task based decomposition proposed by Scassellati [3] presents practical advantages in terms of functional modularity. According to his hypothesis, the primary tasks are recognition, maintenance of eye contact, and gaze following, which enable getting engaged in joint attention. Subsequently, imperative and declarative pointing are considered to permit feedback between the infant and the caregiver.

Another developmental model for shared attention is given in [10] by introducing a basic set of key ingredients as motivational biases such as selective response to parents, a learning mechanism that benefits from predictable contingent interactions, and a structured environment, where the actions of the caregiver are not random, but predictable up to some extent. Based on these key ingredients, a suitable parameter setting is obtained which leads to a healthy natural development. Due to ease of adaptation, most cognitively inspired methods for joint attention employ functional modules which perform tasks similar to those defined in [3]. The next section provides a comprehensive outlook for joint attention models reported in the last decade.

C. Cognitively Inspired Joint Attention Models

One of the earlier studies in this field focuses on a biological model of human visual system and proposes a developmental learning model for joint attention from a biological point of view [11]. A neural network module composed of four layers is employed in modeling the visual system of the robot, where the layers represent the input, retina, visual cortex, and output. As learning proceeds, caregiver performs task evaluation by determining a reward in accordance with the output error of the robot and appropriate weights are obtained.

However, in order to implement a learning scheme, which truly mimics the cognitive development pattern of infants, one should rather go beyond the biological properties. Nagai *et al.* propose a developmental learning scheme [12], which improves learning by passing through the ecological, geometric, and representational stages of joint attention [6]. They further evaluate their system by imposing non-supervised learning conditions in an uncontrolled environment. The visual attention module evaluates the salient features based on color, edge, motion information, as well as the faces in the environment, so that the most salient object is attended at primary stages of learning, which corresponds to the pattern of six month old infants. As this process is repeated, the internal self-evaluation module provides feedback to the visual attention module and joint attention is improved gradually together with sensorimotor coordination corresponding to the pattern of 12 month old infants [6].

These initial methods treat the camera input as substantive images and omit the temporal connection. Humans, on the other hand, utilize motion information besides static information such as posture and face direction to infer about their desires and intentions. The information introduced by motion has also been shown to facilitate infants' learning [13]. For this reason, the robotic agent described in [14] alternates its gaze between a human caregiver and the object it attends by triggering motion, using the cues obtained from the motion of the caregiver's face. In [15], the temporal relationship between the frames is expressed in terms of optical flow vectors and thereby a coarse estimate for gaze shift providing initial motor output to follow the gaze is obtained. The proposed scheme also includes tracking of deictic gestures like pointing. After determining the edge information relating the hand of the caregiver, the robot obtains alternate directions for pointing gesture. The spatial dispersion of optical flow vectors deter-

mine the exact direction. In a similar approach, Haasch *et al.* implement an object attention scenario between a robot and a caregiver on the BIRON robotic platform [16]. The caregiver points to an object, and the robot tracks the hand gesture to look for an object in a small area constrained by the gesture. Verbal cues (such as "blue cup") are identified and used in conjunction with visual cues.

These approaches formulate the visual attention focus of the caregiver based on camera input, employing the 2D information available. However, morphological priors 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 head pose [17], one can model the head of the caregiver as a 3D object and resolve for the pose [18]. Hoffman *et al.* employ an ellipsoidal model for human head and the inferred head angles are used in the estimation of the gaze vector [19]. The assumption is that the robot can establish the relation between the pose of the caregiver and his focus of attention. The causal structure between the action variables, i.e. in most cases the gaze alteration, caregiver's face pose or object locations, are supposed to be given to the robot in advance. However, a low-level design must handle the problem in such a way that this relation is inferred by the robot itself, since these contingencies are reproduced during learning in a natural setting. In [20] a pair of contingent variables are derived using an information theoretic measure to obtain sensory-motor mapping. As human-robot interaction gets richer with the contribution of action modalities such as vocalization and pointing, the importance of the derivation of causal links between perception and action variables increase.

These methods mimic early stages of cognitive development of infants, i.e. mainly six to 12 months. As mentioned in [3], after 12 months, infants start providing feedback to the caregiver by utilizing imperative and declarative pointing, establishing reciprocal communication. Person identification, speech recognition and synthesis along with natural aligned gestures [21], mutually entrained body movements and complex eye movements are used as auxiliary modules in the realization of action-reaction pairs for interaction-oriented robots [22]. According to Kaplan & Hafner, joint attention requires skills for attention detection, attentional manipulation, social coordination and intentional understanding [23]. They are critical of the body of work which deals with elementary skills required for the task, arguing that deeper cognitive aspects are insufficiently addressed. While agreeing with their point, we duly note that the complete specification and implementation of a general joint attention system on a robotic platform is no less than a grand challenge of the field.

III. Proposed approach

Our proposed approach aims at developing the basic gaze following and object segmentation skills for the robot, and mimics the ecological strategy of resolution of focus of attention observed in infants. The proposed method is task-independent as long as the adequate training patterns are presented. Initially head pose of the caregiver is computed and gaze direction is estimated from the head pose. In addition to that, the depth of the object along the gaze direction is induced from the head orientation. Intersection of the gaze and depth gives us a coarse estimate for the object center [24]. Then by pooling a number of estimates and using the surrounding salient features such as color and intensity, we make a final estimate for the object center and perform segmentation around it. This section elaborates on the details of robot platform, head pose estimation, gaze direction resolution and saliency computation.

A. The Robot Platform

We have built a social interaction robot to be used for service and guiding purposes (Figure 1). The system is composed of three main components:

- The Aldebaran Nao humanoid robot [25] as the main interaction and animation unit
- The FESTO Robotino robot [26] as the navigation unit
- A laptop computer as the additional processing and monitoring unit

Aldebaran Nao is a $23''$ tall humanoid robot with 25-DOF in total, two vertically aligned color cameras with 640×480 resolution, and a 500Mhz Geode processor. A Linux based operating system is running on the robot and pre-installed text-to-speech packages allow the robot generate speech. In our design, we are utilizing the upper torso of the Nao robot as the primary visual input and human interaction unit and using the Robotino robot to make the whole body wander around. Robotino is a wheeled robot capable of moving omnidirectionally. It is surrounded by 9 IR sensors and a bump sensor, and it has a 300Mhz processor. Also a 5-meter range laser range finder is installed on the body of the Robotino robot to have more accurate range data from the robot's environment. Figure 2 illustrates an example for a video frame recorded by the robot, where the caregiver focuses his attention on one of the seven objects.

B. Joint Attention Modeling

Since head pose is an indicative of gaze direction, determination of head orientation provides a coarse estimate for center of attention fixation. Thus we employ a head tracking and pose estimation algorithm, which transforms the 2D information concerning the head into 3D pose vectors [24], [27]. This expands our understanding of orientation based on the general anthropomorphic measures. After resolving the gaze direction, a neural network regression is carried out to solve for the initial fixation point. By pooling three consecutive frames of the video recorded at $15fps$, a bin of video images is formed. Taking the fact that humans make three to five saccades per second into account, this bin is convenient to perform a single saccade. A prospective region for this bin is obtained by defining a distribution around estimated initial fixation points. Saliency computation is carried out on this prospective region and the eventual estimated object center is resolved.

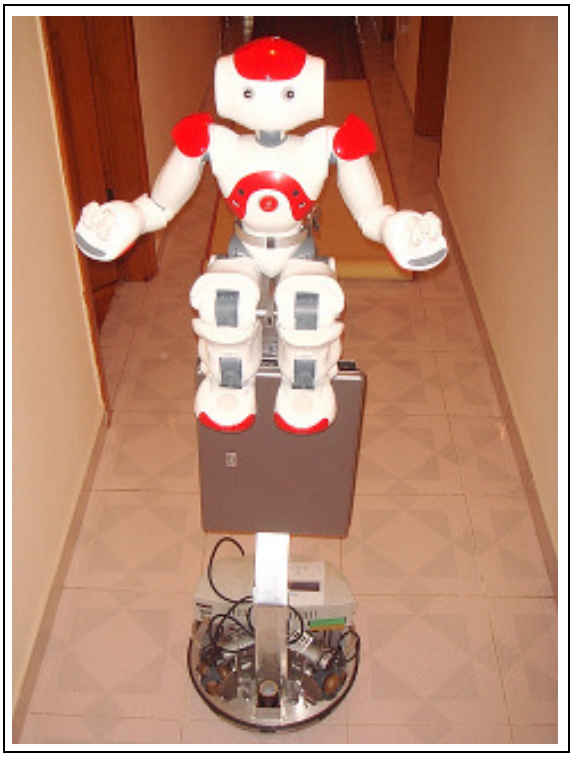

Fig. 1. The robot platform used in the experiments.

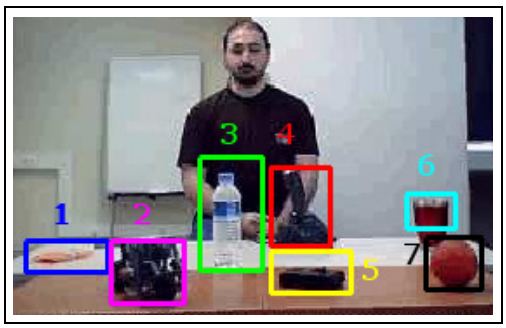

Fig. 2. An exemplary video frame grabbed by the robot showing a caregiver, the setup and object indices.

1) Head Pose Estimation: A session of joint attention is initialized once eye-contact is established between the robotic agent and the caregiver. An elliptic cylindrical head model with reasonable dimensions in agreement with anthropomorphic measures is fit to the corresponding head region [28]. Since this defines a frontal view, the pan, tilt and roll angles concerning this initial head pose are all set to zero. The translation parameters are initialized considering the location of face region on the video frame.

Lucas-Kanade optical flow algorithm is employed [29] in tracking the head. To decrease the computational load, a number of points are regularly sampled on the face region (see Figure 3). The relation of the 3D locations of these points on the cylinder to the 2D pixel coordinates is established by considering a simple pin hole camera model and performing perspective projection. Ray tracing is carried out to find the

(a) (b) (c)

Fig. 3. Cylindrical head model and pose vectors.

intersection of the rays arriving to the image plane and the cylinder.

In the resolution of pose update, initially the head is assumed to keep still for two video frames. By carrying out an iterative procedure, we gradually update the pose and minimize the error corresponding to the face region [30]. The pose vectors, which are computed in the above described manner, have a distribution as illustrated in Figure 4. Each pose value is demonstrated in corresponding colors same as Figure 2 depending on manual annotations obtained by users. As these distributions are modeled with Gaussians, the indicated regions in 3D come into view. The topological relationship between the localization of the objects on the table and corresponding head pose angles are preserved, which ascertains that head pose and gaze direction are closely related.

2) Gaze Direction Estimation: From Figure 4, we infer that there is a nonlinear relationship between the head pose and the gaze direction. We employ two different neural network modules to interpolate the gaze direction and depth of the object of interest from given 3D head pose vector estimates [30].

The video frames are manually annotated by a user, indicating the object of interest of the caregiver. We then define the gaze direction as the slope of the vector which connects the head center and the center of the object of interest. However, the gaze vector alone is not enough to determine the object of interest. Hence, we train another neural network module to estimate the depth of the object. Figure 5 presents examples for estimated gaze vectors. The vector starts from the head center, which is computed by the head tracking and pose estimation

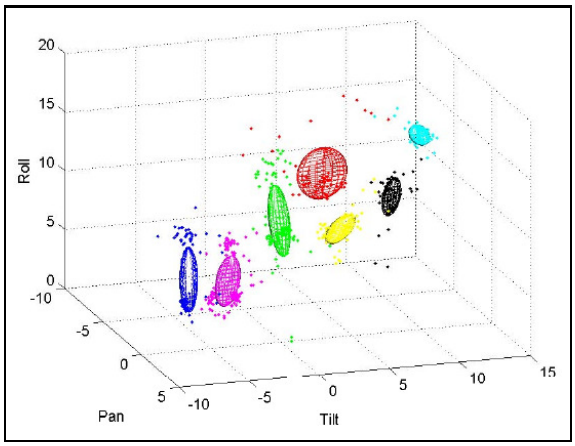

Fig. 4. The pose distributions.

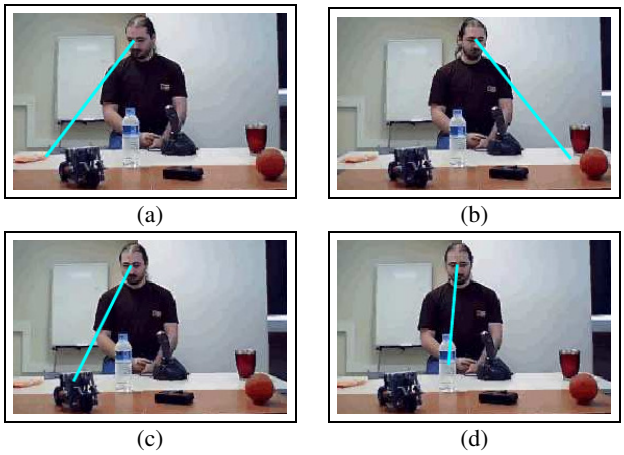

Fig. 5. Examples for estimation of gaze direction and object depth.

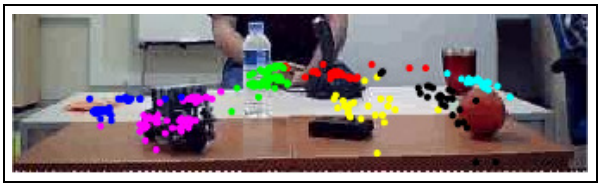

Fig. 6. Initial estimates for object centers.

module, goes along the gaze direction and ends as it reaches the estimated depth of the object of interest.

The point on the gaze direction vector with the estimated depth value is considered as an initial estimate for center of object of interest. Figure 6 shows these initial estimates for one video sequence.

3) Saliency Computation: Once we estimate the coarse object location, we rely on saliency to determine the exact location of the object of interest. To determine a prospective region to run the saliency computation, we pool three consecutive estimates and define a fixed-size search area around the three estimates. Then we employ the popular bottom-up saliency scheme proposed by [31]. 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 [31].

As a result of saliency computation and object segmentation, the regions shown in Figure 7 are obtained. Here the yellow curve indicates the estimated object boundaries and the center of this region is considered as the estimated object center.

C. Experimental Results

In the experiments we used 1200 frames of recorded video, where two different caregivers look at each of the seven objects on the table shown in Figure 2. We apply our algorithms to

Fig. 7. Saliency computation and segmentation in the prospective region.

TABLE I
CORRECT DETECTION RATE FOR THE OBJECTS

	Object Indices						
	1	2	3	4	5	6	7
M_1	0.47	0.32	0.93	0.67	0.44	0.40	0
M_2	0.67	0.58	1.00	0.67	0.67	0.90	0

determine the object of interest and quantify the performance with two different measures, M_1 and M_2.

M_1 indicates at which rate an estimated object center falls into the bounding box of the true object of interest, whereas M_2 shows the rate at which the estimated point is at shortest distance to the true center. Let p 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:

$$M_1(i) = |p \in B_i| \,/\, |p| \,,$$

where $|.|$ denotes the cardinality of a set. The explicit expression for M_2 is:

$$M_2(i) = |\{p | d(p, c_i) < d(p, c_j), \forall j = 1, \cdots, 7, j \neq i\}| \,/\, |p| \,,$$

as $d(a, b)$ denotes the Euclidean distance between points a and b, and c_i stands for the object center concerning object i.

In the first case an estimated point may not fall into any of the bounding boxes and thus it is not assigned to any of the objects, whereas in the second case the point is always assigned to the nearest object in the vicinity. For both measures, values vary between 1 and 0, where being closer to 1 indicates a higher success rate.

Table I summarizes the correct detection rates for each object. Objects lying in the central part of the table are detected correctly with a higher rate. One reason for this is that head pose is closer to the one of template image which is the from a frontal view and thus introduces a minor change in the face view. The objects lying on the sides, however, require more extreme head poses, which are hard to detect. Even though these poses are intuitively observed to be detected with a fairly good precision, the regression module is more likely to introduce some degradation on the extremes. The effect of this factor is prominent in the case of Object 7. The resolution of Object 7 is challenging, not only because it lies on the periphery, but also because it is quite close to Object 6. On the other hand, it is obvious that some objects are very close to each other (Figure 2), even partially occluding one another in some cases (Objects 6 and 7). Hence, in addition to calculating M_1 and M_2 for each object, we form clusters of objects such as left peripheral (L), right peripheral (R) and central (C),

TABLE II
CORRECT DETECTION RATE FOR THE CLUSTERS

	L	C	R
M_1	0.65	0.91	0.21
M_2	0.94	0.94	1.00

according to localization on the table, where L includes objects 1 and 2, R includes 6 and 7, and C covers 3, 4 and 5. The correct detection rates for these clusters are given in Table II.

IV. CONCLUSIONS

This paper provides a detailed insight into the design and training of naturally interacting robotic agents by first giving an overview of evolution of joint attention in infants from a developmental psychology point of view and then by describing the decomposition of progression of cognitive skills from a robotic implementation perspective. Several cognitively-inspired intelligent agent designs are elaborated and an algorithm is described to track the gaze of a caregiver in a joint-attention scenario. The proposed algorithm employs a 3D elliptic cylindrical head model to estimate the head pose, and uses regression analysis to interpolate the gaze direction. Bottom-up feature saliency is proposed to alleviate ambiguities and to segment objects of interest. Good initial results are obtained from a series of experiments performed on a robotic platform. Future work includes measuring the generalization performance of the proposed system accross subjects and experimental settings.

ACKNOWLEDGMENT

We would like to thank Roberto Valenti, Nicu Sebe and Theo Gevers of University of Amsterdam, and H. Levent Akın of Boğaziçi University for their helpful comments and inspiring discussions. We also would like to thank İsmet Meriçli for his helps on building the robot platform. This research is supported by the Dutch BRICKS/BSIK project, Turkish State Planning Organization TAM Project, Tübitak Mildar project with grant number 107A011, and Tubitak Project No 106E172.

REFERENCES

[1] M. Asada, K.F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37(2-3):185–193, 2001.

[2] J. Zlatev and C. Balkenius. Introduction: Why epigenetic robotics. In *Proceedings of the First International Workshop on Epigenetic Robotics: Modeling cognitive development in robotic systems*, volume 85, pages 1–4, 2001.

[3] B. Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Lecture Notes in Computer Science*, pages 176–195, 1999.

[4] S. Reiss. Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology*, 8(3):179–193, 2004.

[5] A. Batki, S. Baron-Cohen, S. Wheelwright, J. Connellan, and J. Ahluwalia. Is there an innate gaze module? Evidence from human neonates. *Infant Behavior and Development*, 23(2):223–229, 2000.

[6] G. Butterworth and N. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *Perspectives on the Child's Theory of Mind*, 9:55–72, 1991.

[7] B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[8] S. Baron-Cohen. *Mindblindness*. MIT Press Cambridge, Mass, 1995.

[9] H. Kozima and H. Yano. A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, pages 47–52, 2001.

[10] I. Fasel, G.O. Deák, J. Triesch, and J. Movellan. Combining embodied models and empirical research for understanding the development of shared attention. In *Proceedings of the 2nd International Conference on Development and Learning*, pages 21–27. IEEE Computer Society Press, 2002.

[11] Y. Nagai, M. Asada, and K. Hosoda. Developmental learning model for joint attention. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 932–937, 2002.

[12] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[13] A. Vinter. The role of movement in eliciting early imitations. *Child Development*, pages 66–71, 1986.

[14] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, 21(9):983–999, 2007.

[15] Y. Nagai. Joint attention development in infant-like robot based on head movement imitation. In *Proc. Third Int. Symposium on Imitation in Animals and Artifacts (AISB05)*, pages 87–96, 2005.

[16] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A multi-modal object attention system for a mobile robot. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005.(IROS 2005)*, pages 2712–2717, 2005.

[17] S.R.H. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception &# 38; Psychophysics*, 66(5):752–771, 2004.

[18] E. Murphy-Chutorian and M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 2008.

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

[20] H. Sumioka, Y. Yoshikawa, and M. Asada. Development of Joint Attention Related Actions Based on Reproducing Interaction Contingency. In *7th IEEE International Conference on Development and Learning*, pages 256–261, 2008.

[21] M. Imai, T. Ono, and H. Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics*, 50(4):636–643, 2003.

[22] T. Ono, M. Imai, and H. Ishiguro. A model of embodied communications with gestures between humans and robots. In *Proceedings of 23rd Annual Meeting of the Cognitive Science Society*, pages 732–737, 2001.

[23] F. Kaplan and V. Hafner. The challenges of joint attention. In *Proceedings of the 4th International Workshop on Epigenetic Robotics*, pages 67–74, 2004.

[24] Z Yücel and A. A. Salah. Resolution of focus of attention using gaze direction estimation and saliency computation. In *Proc. International Conference on Affective Computing and Intelligent Interfaces, to appear*, 2009.

[25] Aldebaran Nao humanoid robot. http://www.aldebaran-robotics.com/eng/Nao.php.

[26] FESTO Robotino robot platform. http://www.festo-didactic.com/int-en/learning-systems/education-and-research-robots-robotino/.

[27] R. Valenti, Z. Yücel, and T. Gevers. Robustifying Eye Center Localization by Head Pose Cues. 2009.

[28] C. C. Gordon, B. Bradtmiller, T. Churchill, C. E. Clauser, J. T. McConville, I. O. Tebbetts, and R. A. Walker. Anthropometric survey of us army personnel: Methods and summary statistics. Technical report, United States Army Natick Research, 1988.

[29] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, volume 3, 1981.

[30] 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.

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