Predicting gaze direction from head pose yaw and pitch

Predicting Gaze Direction from Head Pose Yaw and Pitch

David O. Johnson[1] **and Raymond H. Cuijpers**[2]
[1,2]Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, NL

Abstract - *Socially assistive robots (SARs) must be able to interpret non-verbal communication from a human. A person's gaze direction informs the observer where the visual attention is directed to. Therefore it is useful if a robot can interpret the gaze direction, so that it can assess whether a person is looking at it or some object in the environment. Gazing is a combination of head and eye movement, but detecting eye orientation from a distance is difficult in real life environments. Instead a robot can measure the head pose and infer the gaze direction. In this paper, we show that both the yaw and pitch of a human's gaze can be inferred from the measured yaw and pitch of the human's head pose with simple linear equations.*

Keywords: human robot interaction, gaze direction, head pose estimation

1 Introduction

Elderly people need more support due to their declining capabilities and age-related illnesses. Today that support is provided by younger caregivers. However, the ratio of younger caregivers to elderly people is decreasing in all developed countries. Worldwide, in 2010 there were fewer than nine persons of working-age per elderly person of 65 or older. By 2050 this ratio is expected to decrease to fewer than four working-age persons per elderly person [3]. This will lead to a problem of increasing demand for care and a shortage of caregivers [17].

Socially assistive robots (SARs) are one solution to this problem. SARs can provide reminders and instruction such as the nursing robot Pearl [14] and the Korean robotic language teacher EngKey [10]. They can also provide social support. Social support typically aims at reducing social isolation and enhancing well-being in the form of social interaction with users [5]. Humans use verbal and non-verbal communication. Thus, it is important for a SAR to be able to interpret non-verbal communication from a human. Gaze, or the direction in which the human is looking, is one form of non-verbal communication. Humans use gaze, with and without hand gestures, to point to an object. Humans also use gaze for turn-taking during conversations [7]. Gaze can also be used in navigation to position the robot so it is in the line of sight of the human when it is approaching. Gaze can also be used to determine if the human is paying attention to what the robot is saying.

Gazing is a combination of head and eye movement. The direction the head is looking, or head pose, can be measured by looking at the face [18][19][24]. The direction the eyes are looking is more difficult to determine (see Yamazoe et al. for a thorough discussion of appearance-based and model-based gaze estimation methods [23]). The goal of this research is to determine the gaze of the human (i.e., where the human is looking at) solely from the head pose (i.e., the direction of the head).

Many studies have modeled the relationship between head and eye movement in gazing [9][15][20][26]. But, these models are more suited to creating the behavior in the robot, than determining the gaze from head pose. Additionally as illustrated in Figure 1, these models only consider the yaw (i.e., left and right direction) of the gaze and not the pitch (i.e., up and down direction) or roll (i.e., tilt).

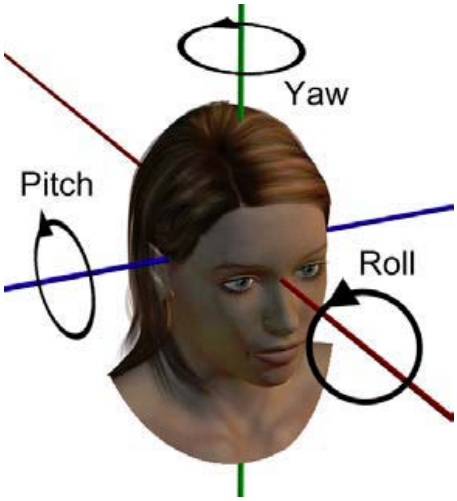

Figure 1. Head pose yaw, pitch, and roll.

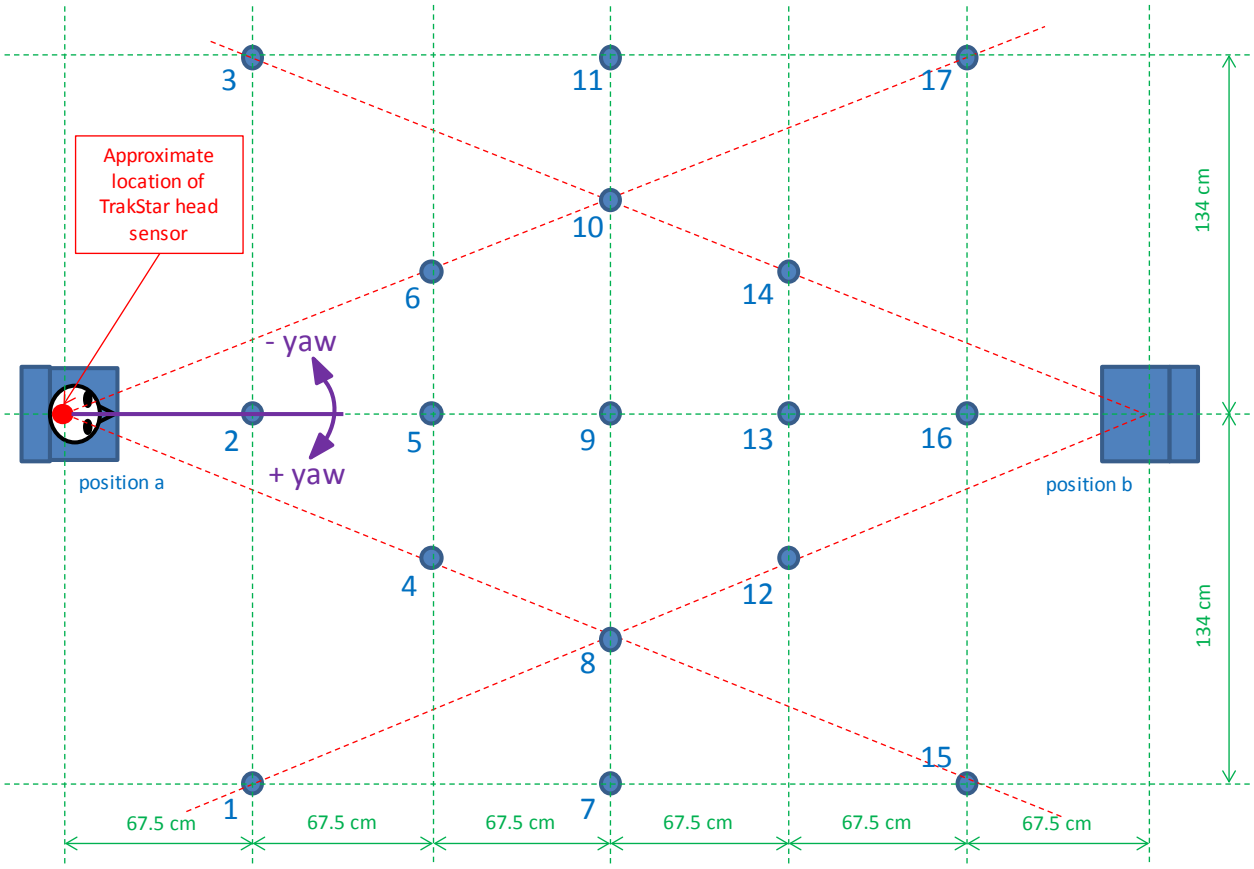

Figure 2. Overhead view of experimental set-up. The blue disks denote object locations on the ground floor. The locations consist of an equally spaced grid in terms of distance (locations 1, 2, 3, 7, 9, 11, 15, 16, 17) and angle (locations 4, 5, 6, 8, 9, 10, 15, 16, 17 from vantage point a, and locations 12, 13, 14, 8, 9, 10, 1, 2, 3 from vantage point b).

In this paper, we investigated the relationship between the yaw and pitch of a human's gaze and the yaw and pitch of the human's head pose. In an experiment we measured head orientations when participants looked at known object locations from two vantage points. The relative position of objects was chosen such that the viewer's gaze elevation, angle, and azimuth were systematically varied. With a linear model, which turns out to be sufficient, we relate the measured head pose to the ground truth of the gaze direction. From this relation the actual gaze direction can be inferred from the measured head pose.

2 Method

2.1 Design

The experiment was set up in a living room environment as illustrated in Figures 2 and 3.

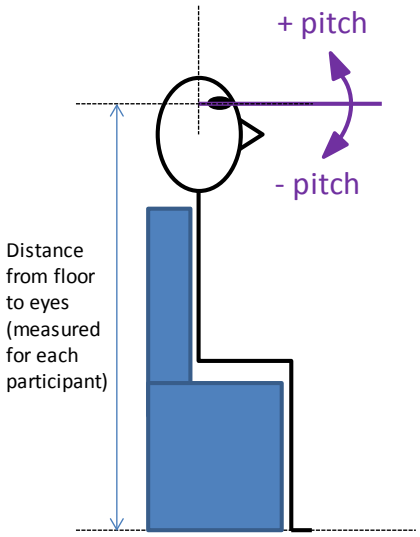

Figure 3. Side view of experimental set-up.

Two chairs were placed opposite from each other, at a distance of 405 cm. One chair was labeled 'position a' and the other chair was labeled 'position b'. In between the two

chairs, there were 17 numbers laid out symmetrically on the floor. The locations of the numbers consisted of an equally spaced grid in terms of distance (locations 1, 2, 3, 7, 9, 11, 15, 16, 17) and angle (locations 4, 5, 6, 8, 9, 10, 15, 16, 17 from vantage point a, and locations 12, 13, 14, 8, 9, 10, 1, 2, 3 from vantage point b). Because of the symmetry of the experimental layout, there are only nine ground-truth yaw values (see Table 1).

Table 1. Ground-Truth Yaw Values and Corresponding Number and Position

Ground-Truth Yaw (°)	Number (1-17) As Seen From Position (a or b)
-63.26	3a, 15b
-33.49	11a, 7b
-21.65	6a, 10a, 17a, 12b, 8b, 1b
-11.23	14a, 4b
0.00	2a, 5a, 9a, 13a, 16a, 16b, 13b, 9b, 5b, 2b
11.23	12a, 6b
21.65	4a, 8a, 15a, 14b, 10b, 3b
33.49	7a, 11b
63.26	1a, 17b

Also, there are only eleven ground-truth pitch values (see Table 2), if a constant distance from the eyes of the participant to the floor is assumed (see Figure 3). The ground-truth values in Table 2 were calculated using the mean distance from the eyes of the participant to the floor of 118.9 cm (standard deviation 2.9).

The participants wore a baseball cap. On this cap a device was attached which measured the yaw and pitch of the participants head.

This study used a within-subjects design. There were two independent variables: the yaw and pitch, with respect to the participant, of the location where we asked the participant to look. The dependent variables were the yaw and pitch of the participant's head when he or she was looking at the specified location. The measurements were counterbalanced with respect to the starting position, meaning that the starting position was varied across the participants. Participants with an odd participant number started at position a, and the participants with an even number started at position b.

Table 2. Ground-Truth Pitch Values and Corresponding Number and Position

Ground-Truth Pitch (°)	Number (1-7) As Seen From Position (a or b)
-18.13	1b, 3b, 15a, 17a
-19.41	2b, 16a
-23.37	4b, 6b, 12a, 14a
-23.77	5b, 13a
-26.09	71, 7b, 11a, 11b
-28.63	8a, 8b, 10a, 10b
-30.43	9a, 9b
-38.40	1a, 3a, 15b, 17b
-39.31	4a, 6a, 12b, 14b
-41.38	5a, 13b
-60.42	2a, 16b

2.2 Design

There were 7 participants (6 males, 1 female), whose ages varied from 20 to 23 (mean age 21.6, standard deviation 1.0). Six of the participants received course credit for participating in the experiment and one was paid 5 euros for participating. All the participants were told that the experiment was not about reaction speed. Other than that, no information was provided about the purpose of the experiment.

2.3 Apparatus

The yaw and pitch of the participant's head were measured using the trakSTAR manufactured by the Ascension Technology Corporation. The trakSTAR is a high-accuracy electromagnetic tracking device for short-range motion tracking applications [1]. The participants were told to look at one of the 17 numbers on the floor with a computer voice through a speaker. Before the experiment, the participants were told to press a hand-held button when they were looking at the number. When the button was

pushed, the yaw and pitch measured by the trakSTAR were recorded by the computer.

2.4 Procedure

The experiment was done in weeks 38 and 39 of 2012 during work hours (8:30-18:00). The participants were first given an eye test. After the eye test the participants were introduced to the experimental setup, which is shown in Figures 2 and 3. Half of the participants were seated in position a; the other half were seated in position b. The experimenter measured the distance from the ground to the eyes of the participant. The participant put on a baseball cap with the trakSTAR sensor on top of it. Then the experimenter explained the experiment.

The computer played pre-recorded messages, "Look at number x" with x being a number ranging from one to seventeen. The participant then had to find the number on the ground and look at it. When the participant was looking at the number he or she had to press a hand-held button which they were given before the experiment. Then, the next number was played until all fifty-one trials were completed. Each of the seventeen positions was measured three times in a random order.

After the first fifty-one trials the participant was asked to change seats and the experiment was repeated from the other vantage point. When the second session ended the participant was asked to take off the baseball cap. Participants were debriefed at the end and remarks were noted.

3 Results

Figure 4 shows the mean of the trakSTAR yaw measurements plotted as a function of the ground-truth yaw values. Regression analysis (R^2 = 0.98) shows a linear relation between the yaw of the head pose (trakSTAR yaw) and the participant's gaze (ground-truth yaw). We found a slope of 0.38 ± 0.02 (t [7] = 21.40, p < 0.001) and an intercept of 1.11 ± 0.64 (t [7] = 1.75, p = 0.124). The data are thus best described by the following equation:

$$y = 0.38x + 1.11 \quad (1)$$

where:

$$x = \text{yaw of measured head pose} \quad (2)$$

$$y = \text{yaw of gaze} \quad (3)$$

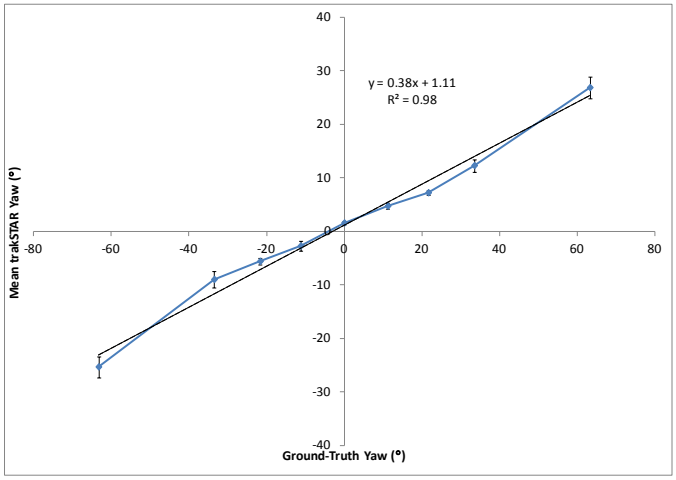

Figure 4. Mean trakStar Yaw (°) plotted as a function of Ground-Truth Yaw (°). Error bars denote standard errors of the sample mean.

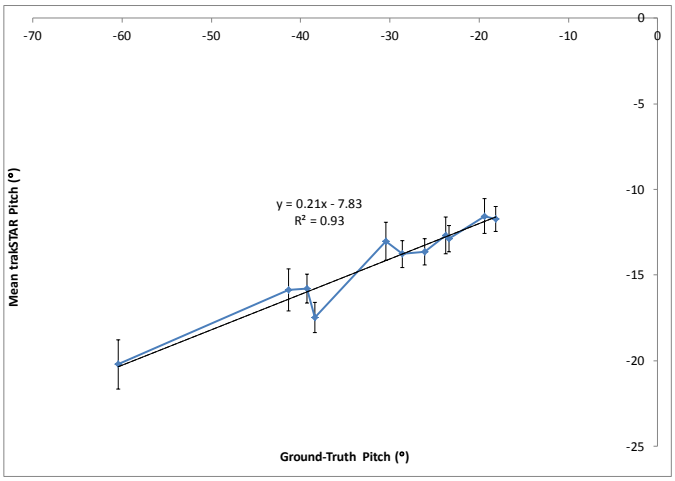

Figure 5. Mean trakStar Pitch (°) plotted as a function of Ground-Truth Pitch (°). Error bars denote standard errors of the sample mean.

Figure 5 shows the mean of the trakSTAR pitch measurements plotted as a function of the ground-truth pitch values. Regression analysis (R^2 = 0.93) shows the following linear relation between the pitch of the head pose (trakSTAR pitch) and the participant's gaze (ground-truth pitch). We found a slope of 0.21 ± 0.02 (t [9] = 11.23, p < 0.001) and an intercept of -7.83 ± 0.63 (t [9] = -12.51, p < 0.001). The data are thus best described by the following equation:

$$y = 0.21x - 7.83 \quad (4)$$

where:

$$x = \text{pitch of measured head pose} \quad (5)$$

$$y = \text{pitch of gaze} \quad (6)$$

4 Discussion and conclusions

We measured the head pose of human observers when looking at objects in the environment and compared it to the actual gaze direction. We found simple linear relations (see Equations 1 and 4) between the yaw and pitch of a human's head pose and the yaw and pitch of the human's gaze direction. As hypothesized, the gaze direction can be reliably estimated from the observed head pose using the fitted equations.

Others have estimated gaze direction from head pose, but not with simple linear relations. Yücel and Salah used a two-layer back-propagation neural network to estimate the gaze direction from a 3-dimensional head pose vector [25]. However, they did not check if there was a linear relation between the gaze direction and the head pose vector. Stiefelhagen et al., used an assumption to implicitly estimate the gaze direction from the head pose [16]. They assumed the focus of attention to be a person, and the estimated head pose was corrected to select the closest person as the target of the gaze.

The relationship between head pose and gaze direction turned out to be linear. For the yaw angle this is somewhat surprising considering that most people start looking against their own noses with an eye turn of say 40-60 degrees. Therefore one might expect a higher gain of head turn for large gaze angles than for small gaze angles. No such a change in gain was observed, however. Similar considerations apply to the pitch angle. Looking up is usually constrained by ones protruding eye brows whereas looking down is much less constrained. This could result in an asymmetry between looking up and looking down. However, from our data there is no reason to believe that the relationship between pitch of head pose and pitch of gaze direction is non-linear.

In real environments, objects may be large and cover a substantial part of the visual field like when admiring a new car of a friend, say. In such situations it is to be expected that many different locations within the interior of the object are being fixated. Our simple model does not say anything about how people perform gaze fixations within an object, nor how a scene of objects is scanned [6][13][21]. However, it seems reasonable to expect that the center of gravity of fixations within an object will adhere to the same simple relationships as we observed.

Previous research has established that humans estimate the gaze direction of another person from head pose and the features of the person's eyes [4][8][11][22]. Thus, it would seem logical that a robot should also use both head pose and eye features to estimate gaze direction. However, humans do not think in linear equations as easily as robots do, so it is also plausible that if the relationship between head pose and gaze is a simple linear relationship, as we have shown here, then a robot only needs to determine head pose to estimate gaze direction.

To summarize, we have shown that both the yaw and pitch of a human's *gaze* can be inferred from the measured yaw and pitch of the human's *head pose* with simple linear equations. Thus by measuring the human's head pose from its video stream, the robot can estimate where the human is gazing. Knowing where the human is gazing, will help the robot determine what object the human is pointing at, whether the human is paying attention to the robot, when it is the robot's turn in a conversation, and the direction to approach a human so it will be in the human's line of sight.

5 Acknowledgements

The research leading to these results is part of the KSERA project (http://www.ksera-project.eu) and has received funding from the European Commission under the 7th Framework Programme (FP7) for Research and Technological Development under grant agreement n2010-248085.

We would also like to thank Jan Roelof de Pijper, Ellen Hoefsloot, Milou de Louw, and Martijn van Vlijmen for their contributions to this work.

6 References

[1] Ascension Technology Cooperation (2011). 3D Guidance trakSTAR 2TM Installation and Operation Guide.

[2] Breazeal C, Kidd C, Thomaz A, Hoffman G, and Berlin M (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (Edmonton, Alberta, Canada, August 2-6, 2005), 708-713.

[3] Bremner, J., Frost, A., Haub, C., Mather, M., Ringheim, K., & Zuehlke, E. (2010). World population highlights: Key finding from prbs 2010 world population data sheet. Population Bulletin, 65(2).

[4] Cline, M. G. (1967). The Perception of Where a Person Is Looking. The American Journal of Psychology, 80(1), 41-50

[5] Feil-Seifer, D. & Mataric, M. (2005). Defining socially assistive robotics. In proceedings of the Rehabilitation robotics, 2005. icorr 2005. 9th international conference on (p. 465 - 468). http://dx.doi.org/ 10.1109/ICORR.2005.1501143

[6] Hardiess, G., Gillner, S., and Mallot, H. A. (2008). Head and eye movements and the role of memory limitations in a visual search paradigm. Journal of Vision, 8(1):7, 1–13, doi:10.1167/8.1.7

[7] Kendon, A. (1967). Some functions of gaze-direction in social interaction. Acta psychologica, 26(1), 22.

[8] Langton, S. R. H. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. The Quarterly Journal of Experimental Psychology, 53A (3), 825-845.

[9] Laurutis VP, Robinson DA (1986) The vestibulo-ocular reex during human saccadic eye movements. J Physiol 373: 209–233.

[10] Lee, S., Hyungjong, N., Jonghoon, L., Kyusong, L., Gary Geunbae, L., Seongdae, S., & Munsang, K. (2011). On the effectiveness of robot-assisted language learning. ReCALL, 23(01), 25-58. http://dx.doi.org/10.1017/S0958344010000273

[11] Lobmaier, J. S., Fischer, M. H., and Schwaninger, A. (2006). Objects Capture Perceived Gaze Direction. Experimental Psychology 2006, 53(2), 117–122. DOI 10.1027/1618-3169.53.2.117

[12] Lockerd A and Breazeal C (2004). Tutelage and Socially Guided Robot Learning. MIT Media Lab, Cambridge, MA, USA, 2004.

[13] Nguyen, A., Chandran, V., and Sridharan, S. (2006). Gaze tracking for region of interest coding in JPEG 2000. Signal Processing: Image Communication, 21(5), 359-377.

[14] Pineau, J., Montemerlo, M., Pollack, M., Roy, N., & Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. Robotics and Autonomous Systems, 42(3-4), 271–281. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0921889002003810

[15] Saeb S, Weber C, Triesch J (2011) Learning the Optimal Control of Coordinated Eye and Head Movements. PLoS Comput Biol 7(11): e1002253. doi:10.1371/journal.pcbi.1002253

[16] Stiefelhagen, R., Yang, J., & Waibel, A. (1999). Modeling focus of attention for meeting indexing. In Proc. Seventh acm int. conf. on multimedia, 1:3–10.

[17] Tapus, A., Mataric, M. J., & Scassellati, B. (2007). The grand challenges in socially assistive robotics. Robotics and Automation Magazine, 14(1), 1-7, http://dx.doi.org/10.1109/MRA.2007.339605.

[18] van der Pol, D., Cuijpers, R.H., & Juola, J.F. (2010). Head Pose Estimation for Real-Time Low Resolution Video. In proceedings of the European Conference on Cognitive Ergonomics, August 25-27, 2010, Delft, The Netherlands.

[19] van der Pol, D., Cuijpers. R.H., and Juola, J.F. (2011). Head pose estimation for a domestic robot. In proceedings of the 6th international conference on Human-robot interaction, March 6-9, 2011, Lausanne, Switzerland.

[20] Van Gisbergen JAM, Robinson DA, Gielen S (1981) A quantitative analysis of generation of saccadic eye movements by burst neurons. J Neurophysiol 45:417–442.

[21] Veneri G, Rosini F, Federighi P, Federico A, Rufa A. (2012). Evaluating gaze control on a multi-target sequencing task: the distribution of fixations is evidence of exploration optimisation. Comput Biol Med., 42(2), 235-44. doi: 10.1016/j.compbiomed.2011.11.013.

[22] Wilson, H. R., Wilkinson, F., Lin, L., and Castillo, M. (2000). Perception of head orientation. Vision Research, 40, 459–472.

[23] Yamazoe, H., Utsumi, A., Yonezawa, T., and Abe, S (2008). Remote Gaze Estimation with a Single Camera Based on Facial-Feature Tracking, In Proceedings of Eye Tracking Research & Applications Symposium (ETRA2008), pp.245--250.

[24] Yan, W., Torta, E., van der Pol, D., Meins, N., Weber, C., Cuijpers, R. H., and Wermter, S. (2013). Learning Robot Vision for Assisted Living. In J. Garcia-Rodriguez, & M. Cazorla Quevedo (Eds.), Robotic Vision: Technologies for Machine Learning and Vision Applications (pp. 257-280). Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-2672-0.ch015.

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

[26] Zee DS, Optican LM, Cook JD, Robinson DA, Engel WK (1976) Slow saccades in spinocerebellar degeneration. Arch Neurol 33: 243–251.