ManiGaze: a Dataset for Evaluating Remote Gaze Estimator in Object Manipulation Situations

Remy Siegfried
Idiap Reseach Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
remy.siegfried@idiap.ch

Bozorgmehr Aminian
Idiap Reseach Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
bozorgmehr.aminian@idiap.ch

Jean-Marc Odobez
Idiap Reseach Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
odobez@idiap.ch

ABSTRACT

Gaze estimation allows robots to better understand users and thus to more precisely meet their needs. In this paper, we are interested in gaze sensing for analyzing collaborative tasks and manipulation behaviors in human-robot interactions (HRI), which differs from screen gazing and other communicative HRI settings. Our goal is to study the accuracy that remote vision gaze estimators can provide, as they are a promising alternative to current accurate but intrusive wearable sensors. In this view, our contributions are: 1) we collected and make public a labeled dataset involving manipulation tasks and gazing behaviors in an HRI context; 2) we evaluate the performance of a state-of-the-art gaze estimation system on this dataset. Our results show a low default accuracy, which is improved by calibration, but that more research is needed if one wishes to distinguish gazing at one object amongst a dozen on a table.

CCS CONCEPTS

• **Computing methodologies** → Neural networks; *Activity recognition and understanding*; **Tracking**.

KEYWORDS

Human-robot interaction, gaze estimation, remote recording, dataset

ACM Reference Format:
Remy Siegfried, Bozorgmehr Aminian, and Jean-Marc Odobez. 2020. ManiGaze: a Dataset for Evaluating Remote Gaze Estimator in Object Manipulation Situations. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Short Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3379156.3391369

1 INTRODUCTION

As one of the main indicator of attention, gaze plays an important role in many human activities, and being able to estimate it can be useful in a large range of applications. In particular, as one very important non-verbal communication cue [Cook 1977; Muralidhar et al. 2018; Yarbus 1967], gaze has been much studied in Human-Human or Human-Robot social interaction contexts, for modeling turn-taking patterns [Ishii et al. 2016; Sheikhi and Odobez 2015], improving the dialog fluency [Andrist et al. 2013] or the robot anticipation during collaborative tasks [Huang and Mutlu 2016].

The role of gaze has also been well studied in object manipulation. Its importance has been demonstrated during handover, in which gaze pointing to objects eliminates reference ambiguities and allows partners to respond quicker [Moon et al. 2014]. Moreover, during object manipulations, the *proactive* use of the gaze informs about the intention and anticipation of the actor while its *reactive* use enlightens a particular attention [Admoni and Scassellati 2017; Bader et al. 2009; Hayhoe and Ballard 2005; Johansson et al. 2001]. Nevertheless, in spite of its importance, most studies in this domain relied on either hand-coded gaze events or the use of intrusive sensors like chin-rests [Johansson et al. 2001], head-mounted devices, or wearable glasses [Aronson et al. 2018; Newman et al. 2018], which can impact negatively the natural behavior of people. In fact, gaze estimation in such conditions is a particularly challenging task. Sensing conditions are usually quite different than in screen-gazing applications: higher pose variability, lower eye and face image resolution to accommodate potentially larger user mobility, larger illumination variations, potentially unknown user and absence of user cooperation. Recent methods like [Mora and Odobez 2016] aims at overcoming these challenges, but their performance remains to be investigated in the context of HRI.

In this paper, we investigate the use of non-intrusive remote sensors for gaze estimation in the framework of object manipulation tasks, which to the best of our knowledge had not been studied before. To that end, we collected the ManiGaze dataset, in which subjects performed different tasks related to this context and where the gaze ground truth can be obtained. Indeed, traditional gaze estimation datasets like Columbia Gaze [Smith et al. 2013], UT Multiview [Sugano et al. 2014], Eyediap [Mora and Odobez 2014], MPIIGaze [Zhang et al. 2017] and more recently RT-GENE [Fischer et al. 2018] or Gaze360 [Kellnhofer et al. 2019] are very useful to train and evaluate raw models, as they gather a lot of data with variability in head poses, illuminations or gaze directions. However, they may sometimes miss some higher natural variability and may not allow evaluating directly the usability of gaze estimation methods. Collecting data in an environment where gaze tracking is used in a system can address this issue, which is important as different gaze tracking characteristics (accuracy, consistency, robustness) might have different importance in different types of tasks.

To this end, we collected a medium-size gaze dataset involving natural interactions and behaviors which can help evaluate gaze estimation methods beyond classical screen gazing setups.

ETRA '20 Short Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7134-6/20/06…$15.00
https://doi.org/10.1145/3379156.3391369

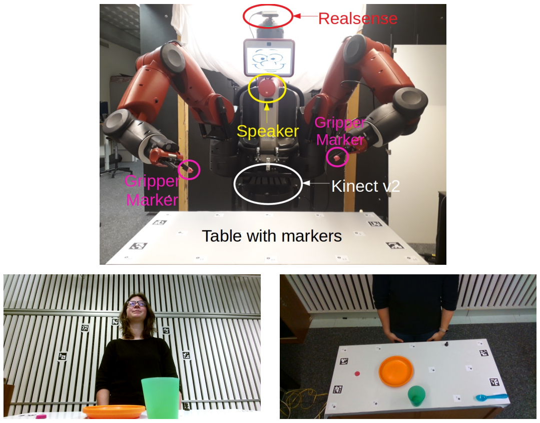

Figure 1: Experimental setup from the point of view of the user (top), the field of view of the Kinect v2 (bottom left) and of the Intel RealSense (bottom right).

The considered scenarios are characteristic of some human-robot collaboration where the goal is to perform tasks involving the environment (object manipulation, tool handover, etc.). In such conditions, the gaze range of the user is big, going from objects near his/her chest to targets at eye level, which is challenging for remote gaze estimation methods [Kellnhofer et al. 2019].

Using this dataset, we evaluated the performances of state-of-the-art 3D gaze estimations method. As an approach, we explored Appearance Based Method (ABM) methods since amongst existing systems [Chennamma and Yuan 2013], computer vision ones have a good ability to achieve non-invasive user-friendly remote gaze sensing. ABM algorithms (esp. thanks to deep learning) have also been shown to present the best trade-off between accuracy and operational range [Zhang et al. 2019]. More precisely, we adapted the approach proposed by [Mora and Odobez 2016], using the more recent GazeNet deep neural network (DNN) [Zhang et al. 2017] for learning the mapping from the eye image to the gaze direction. Furthermore, since due to eye variabilities (e.g. dominant eye, fovea point, eye shape, ...) across people, gaze trackers usually need person-specific calibration procedures to perform better, we study the effect of such calibration considering intra and cross-session situations as well as two bias models (constant and linear bias).

Contributions. In this context our contributions are:

Dataset. To evaluate gaze estimation and calibration during object manipulation tasks, we collected a multimodal dataset providing several scenarios with different characteristics (looking at objects on a table, picking objects), which we will make publicly available.

Remote sensor based gaze estimation evaluation. We evaluated gaze estimation within this context based on an unintrusive remote sensor. Experiments were designed to evaluate both the gaze estimation raw accuracy (error in degree) and the ability to distinguish between targets in the manipulation space (classification error).

2 DATASET

The multimodal dataset we collected involved 16 participants (24-36 years old, 4 women / 12 men, 6 glasses wearers). We described below its main characteristics. Both raw media files (e.g. videos) and annotations will be provided publicly (www.idiap.ch/dataset/manigaze). Some statistics on the dataset are provided in experiments.

2.1 Set-up and Calibration

The physical setup is illustrated in Fig. 1. It consists of a Baxter robot separated from the participant by a table on which the manipulation tasks take place. On this table, 14 markers (numbered black dots) were placed on three rows as a regular grid of points with a distance of around 20 cm between two rows and two markers within the same row. We used different sensors recording 3 modalities (color, depth, and audio): 2 RGB-D cameras and a microphone. Finding a good place to sense the gaze of the participants was difficult because of the place taken by Baxter, which is a rather big robot and the position of the potential targets (objects on the table, robot arms, and head) which were very spread in space. The best location we found was at the height of the table, and we used a Kinect v2 sensor due to its depth accuracy and large field of view. The low camera angle is unusual compared to classical gaze estimation datasets, but it can be cumbersome to get ideal sensing conditions (i.e. frontal view) in real-life setup, as sensors can not reasonably be placed in the workspace (here: the table). A second RGB-D Intel RealSense D435 camera was also placed on the robot head to record the table and participant hands from above. In addition to the sensor data, several features related to the interaction like mouse clicks, robot speech, or robot arms position were recorded.

Wizard-of-Oz approach. The recordings were made using a WoZ approach. The robot introduced himself at the beginning of the experiment and asked a few questions to make the participant used to the robot. Then it guided the participant through the whole set of experiments using voice to encourage interactions and natural behaviors. Randomness was introduced in the robot utterances ("look at X", "Can you look at X", "Now, look at X") and in the feedbacks ("ok", "congratulations", "well done") to make the interaction more natural. The robot also randomly asked sometimes to look at it to break the task monotony and gather gaze points to the robot. Also, at any time the participant could ask the robot to repeat the previous instruction. The result was qualitatively satisfactory: participants tend to ignore the experimenter, speaking naturally to the robot and turning the head toward it when the robot was speaking.

2.2 Recorded sessions

The experiment was organized in 4 sessions, going from the most artificial to the most natural, and the participant received some basic instructions before the start.

Markers on the Table Targets (MT). The participant stood approximately 1 meter in front of the robot and had the computer mouse in hand. The robot asked the participant to look at a numbered marker. The participant would then press the mouse button upon gazing at the target marker without blinking. After feedback, the procedure would repeat, with the robot designating another marker. The participants' gaze fixation locations and their occurrence time were deduced from the mouse events and the marker target. This session can be used to evaluate gaze estimation and calibration algorithms, as the markers build a dense and regular grid distributed on the whole manipulation space.

End-effector Targets (ET). This session is similar to the MT one except that the participants are asked to continuously look at one of the robot end-effectors and to press the mouse button when it stops moving (37 positions) while continuing to look at it without

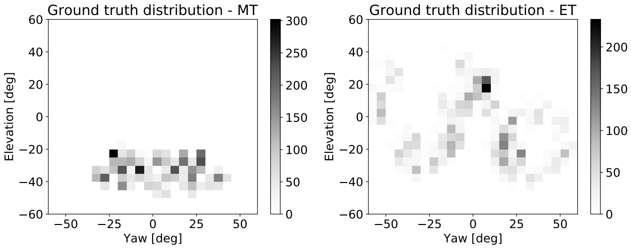

Figure 2: Gaze histogram for MT and ET sessions.

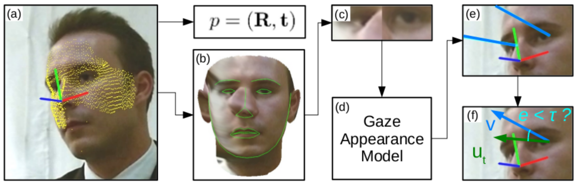

Figure 3: Framework. a) Head pose estimation [Yu et al. 2018]. b) Face image frontalization. c-d) Mapping of eye images to gaze angles g. e) Computation of gaze direction v from g and p. f) Extraction of the angle error e between the vector pointing to the target u_t and the gaze vector v.

blinking. Each robot's arm is used to cover half the space, to avoid participant's face occlusion. This session provides additional targets not limited to the manipulation area (the table) to evaluate gaze estimation and calibration. In this paper, we focus on the fixations indicated by the mouse events, but participants perform successions of smooth pursuits that could also be exploited in further studies.

Object Manipulation (OM). In this session, a backgammon pawn, a chess pawn, a spoon, a glass, and a plate are initially placed on markers on the table. The robot asks sometimes the participant to move objects to a designated numbered marker or on/in another object, and sometimes questions related to the actual position of the objects. Each participant repeats the recording for two out of three defined initial object positions. As pick and place actions of the participant are guided by the robot instructions and objects are always placed on markers, the positions of the objects are known at each robot's occurrence (i.e. before and after each pick and place action). This session simulates an object manipulation task to study gaze behavior during pick and place actions (pick and place moments were annotated). It is less controlled than previous ones and as a result, participants act even more naturally.

Set a Table (ST). In this final session, the participant is asked to set a table while explaining to the robot how to do it without referring to the markers on the table. The robot does not act or speak and the participant is free to choose how to proceed in order to encourage natural eye-hand coordination (e.g. usage of both hands at the same time, faster transitions between objects).

3 GAZE ESTIMATION APPROACH

The framework we evaluated comprised 3 steps: gaze estimation, bias calibration, and visual focus of attention estimation.

Gaze estimation. We followed the head pose independent gaze estimation framework of [Mora and Odobez 2016] with differences (see Fig. 3). The head pose $\mathbf{p} = (\mathbf{R}, \mathbf{t})$, where $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector, is tracked in RGB-D videos using the Headfusion method [Yu et al. 2018] which relies on the automatic fitting of both a 3D Morphable Model of the face and a 3D raw representation of the head. This method was shown to be highly accurate (≃ 2°errors) and much more robust to large head poses than [Mora and Odobez 2016] or color-based landmarks detection methods. Then, eyes were cropped using 3D frontalization and landmark alignment [Siegfried et al. 2017].

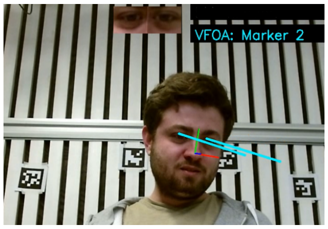

Figure 4: Gaze estimation example showing: head pose (red-green-blue coordinate system on the nose); gaze (cyan rays); extracted and exploited eye images after frontalization (top, together with the VFOA estimate).

To infer the gaze directions from the cropped eye images, we investigated two methods. The first one is the original approach presented in [Mora and Odobez 2016], which relies on a multi-level HoG SVR trained on the Eyediap dataset. The second one is the GazeNet model proposed by [Zhang et al. 2017]. It is based on a *vgg16* architecture and takes head pose angles as additional inputs. We validated our implementation on the Columbia Gaze and the UT Multiview datasets and obtained results consistent with the state-of-the-art (respectively 3.45°and 5.08°).

As a result, at each time we have for each eye its gaze $\mathbf{g} = (\phi, \theta)$ defined by its yaw (ϕ) and tilt elevation (θ) angles, which can be mapped into a corresponding gaze direction unitary vector $\mathbf{v}$ through a transform Φ, i.e. $\mathbf{v} = \Phi(\mathbf{g})$. See Fig. 4 for a result example. Finally, to have a unique estimation, we keep the gaze estimated from the closest eye to the camera, because it is usually the most visible and thus less prone to occlusions and deformations from the frontalization, resulting in more precise and stable estimations.

Bias computation. We relied on a set of calibration frames $\mathcal{F} = \{(\hat{\mathbf{g}}_i, \mathbf{g}_i)\}$ with $\hat{\mathbf{g}}$ the gaze estimate and $\mathbf{g}$ the ground truth gaze obtained by applying the reverse gaze transform Φ^{-1} to the vector defined by the 3D eye position x_{ei} and the known 3D target position t_i of the i^{th} calibration event, $\mathbf{g}_i = \Phi^{-1}(t_i - x_{ei})$. To perform calibration, given $\mathcal{F}$, we searched for the calibration function C so that $\mathbf{g}_i = C_{\mathcal{F}}(\hat{\mathbf{g}}_i)$. In this work, we investigated two models:

Constant: it considers a constant bias in both dimensions (yaw and elevation angles) that we want to compensate so that $\mathbf{g}_i = \hat{\mathbf{g}}_i - \mathbf{b}$. The bias is estimated by taking the mean of the gaze angle error

$$\widehat{\mathbf{b}} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} (\hat{\mathbf{g}}_i - \mathbf{g}_i). \tag{1}$$

Linear: it considers a linear relation between the estimation and the real gaze, so that $\mathbf{g}_i = \begin{pmatrix} s_\phi & 0 \\ 0 & s_\theta \end{pmatrix} \hat{\mathbf{g}}_i + \begin{pmatrix} t_\phi \\ t_\theta \end{pmatrix}$. The coefficients are estimated by solving two independent linear regression problems.

Visual focus of attention (VFOA) estimation. In collaborative tasks, the VFOA, i.e. which object or location the user is looking at, is often more useful than the gaze estimate itself. Deciding whether

Table 1: Mean angular errors (angErr) measured on MT or ET, as well as VFOA accuracy (vfoaAcc) measured on MT.

Method	angErr MT [deg]	angErr ET [deg]	vfoaAcc MT [%]
Baselines			
HSVR [Mora and Odobez 2016]	16.26	13.67	.26
GazeNet [Zhang et al. 2017]	22.35	17.90	.18
Supervised intra-session calibration (calibrated on session X, tested on session X)			
HSVR-cst	7.22	9.26	.44
HSVR-lin	5.63	**8.11**	.61
GazeNet-cst	6.00	9.97	.59
GazeNet-lin	**4.67**	8.96	**.70**
Supervised cross-session calibration (calibrated on session Y, tested on session X)			
HSVR-cst	9.04	**11.12**	.37
HSVR-lin	8.48	15.79	.42
GazeNet-cst	9.75	12.20	.43
GazeNet-lin	**8.30**	16.41	**.50**

a person looks at a target t (defined by the direction $\mathbf{u}_t$) can be done by comparing the gaze direction to $\mathbf{u}_t$ (see Fig. 3). It can be achieved by computing their angle difference: $e_t = \arccos(\Phi(C_{\mathcal{F}}(\mathbf{g})) \cdot \mathbf{u}_t)$. VFOA is estimated as the target t which has the lowest related e_t.

4 EXPERIMENTS

In this work, we let the less controlled OM and ST sessions aside, as they do not provide ground truth data (no VFOA or manipulation information). However, they open possibilities for further work, e.g. on hand-gaze coordination, at the cost of manual annotations.

Data statistics. Experiments were done using the 16 subjects in MT and ET sessions, for a total of respectively 75 and 49 minutes of video. There are an average of 807 ground truth labeled frames for the MT session (14 different targets) and 337 labeled frames for the ET session (37 different targets) by subject, for a total of respectively 5894 and 5396 annotated frames. The Fig. 2 presents the gaze ground truth distributions of the two sessions and shows that the ET session has a wider coverage of the gaze space compared to MT in which subjects only look at the table, resulting in challenging elevation angles range (-50°to -20°) as gaze estimation is usually less accurate when people look down [Kellnhofer et al. 2019].

Experimental protocol. As metrics, we use the mean angular error for MT and ET session and the VFOA accuracy for the MT session where 14 concurrent targets are present. Those metrics were computed for each subject, using all available gaze ground truth points and then averaged by session (see Tab. 1). The calibration function C was estimated using 20 annotated points taken at random in the calibration session and evaluated either on the same session (intra-session calibration) or on the other session (cross-session calibration). Experiments were repeated 10 times and results averaged to account for random effects.

Baseline results. Gaze estimation methods (see Sec. 3), namely HSVR and GazeNet, were trained in a cross-dataset fashion, using the Eyediap dataset (pose: M, scenario: FT, resolution: VGA). From Tab. 1, we see that without additional correction, both methods have high errors, above those reported in more traditional screen-based setups, and that the traditional method beats the neural network by a large margin. This can be explained by the experimental setup: the camera angle is unusual (i.e. far from a frontal view), and in the MT session, participants are looking at markers on the table which are close to them and below the camera. This is a difficult situation for gaze estimators, as eyelids cover a large portion of the eyeball. In ET, visual targets are on average higher in the user field of view and gaze estimation is more accurate, somehow confirming the impact of the visual target positions on performance.

Calibration. We tested both the constant (*-cst*) and linear (*-lin*) calibration models. After calibration, the difference between both gaze estimation methods fades away. As expected, absolute errors are much smaller since calibration provides person-specific correction and performs some kind of domain adaptation at the same time.

Cross-session calibration, which is closer to a real application case, performs worse than intra-session calibration. It still reaches decent performances on the MT session, but not on the ET session, especially when the linear correction is used. This is due to overfitting: the linear model trained on the MT session where the distribution of gazes is much smaller than in ET (esp. in elevation, see Fig.2) does not generalize well to the range of gazes of the ET session which were not contained in the (MT) calibration set. It also means that the proposed calibration models do not fully grasp the error sources. Indeed they make the hypothesis that the error only depends on the gaze, but it could also be related to the head pose or the eye location. Further experiments are needed to explore more complex models.

VFOA accuracy. The last column of Tab. 1 reports the VFOA estimation accuracy. Calibration improves significantly the results in all cases and the GazeNet model together with the linear correction outperforms all other methods. Interestingly, although the GazeNet performances are usually on par or worse than that of HSVR (on the more realistic cross-session case) its VFOA accuracy is better, suggesting that the gaze errors are not distributed similarly for both methods. Regarding the overall absolute performance, none of the presented methods provide a very satisfying estimation of attention on objects (which are separated by 20 cm). This means that for such task VFOA estimation remains difficult using the current system, even with supervised calibration.

5 CONCLUSION AND FUTURE WORK

We studied gaze estimation in an HRI scenario related to object manipulation. We presented a new public dataset that will allow the analysis of pick and place types of actions, involving static and dynamical visual targets in both the manipulation space (table) and the space between the user and the robot. We evaluated a state-of-the-art gaze estimation method and reported the difficulty to design a setup that can accurately estimate gaze with visual targets in a large space. The overall results indicate the need for further work to make gaze estimation fully exploitable in this kind of setup. Besides improving the base gaze estimator accuracy, research may include searching for a better bias model accounting for a potential head pose dependency and using task gaze priors (e.g. picking objects or hand activities) and other gaze priors (dialog between the user and the robot [Siegfried et al. 2017]) to sample more data points for online calibration and going beyond supervised calibration which has poor usability [Morimoto and Mimica 2005].

ACKNOWLEDGMENTS

This research has been supported by the European Union Horizon 2020 research and innovation program (grant agreement no. 688147, MuMMER project) and by the SNSF (grant agreement no. 30214, project ROSALIS).

REFERENCES

Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–53.

Sean Andrist, Bilge Mutlu, and Michael Gleicher. 2013. Conversational Gaze Aversion for Virtual Agents.. In *IVA*, Vol. 8108. 249–262.

Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. 2018. Eye-hand behavior in human-robot shared manipulation. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 4–13.

Thomas Bader, Matthias Vogelgesang, and Edmund Klaus. 2009. Multimodal Integration of Natural Gaze Behavior for Intention Recognition During Object Manipulation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI-MLMI '09).* 199–206.

H. R. Chennamma and Xiaohui Yuan. 2013. A Survey on Eye-Gaze Tracking Techniques. *CoRR* abs/1312.6410 (2013). http://arxiv.org/abs/1312.6410

Mark Cook. 1977. Gaze and Mutual Gaze in Social Encounters: How long—and when—we look others "in the eye" is one of the main signals in nonverbal communication. *American Scientist* 65, 3 (1977), 328–333. http://www.jstor.org/stable/27847843

Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV).* 334–352.

Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in Cognitive Sciences* 9, 4 (2005), 188–194.

Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory Robot Control for Efficient Human-Robot Collaboration. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16).* IEEE Press, 83–90.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Trans. Interact. Intell. Syst.* 6, 1, Article 4 (2016), 31 pages.

Roland S. Johansson, Göran Westling, Anders Bäckström, and J. Randall Flanagan. 2001. Eye–Hand Coordination in Object Manipulation. *Journal of Neuroscience* 21, 17 (2001), 6917–6932.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE International Conference on Computer Vision.* 6912–6921.

Ajung Moon, Minhua Zheng, Daniel M. Troniak, Benjamin A. Blumer, Brian Gleeson, Karon MacLean, Matthew K.X.J. Pan, and Elizabeth A. Croft. 2014. Meet me where I'm gazing: How shared attention gaze affects human-robot handover timing. In *HRI 2014 - Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction.* IEEE, Institute of Electrical and Electronics Engineers, 334–341.

Kenneth Alberto Funes Mora and Jean-Marc Odobez. 2014. Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras.. In *CVPR.* 1773–1780.

Kenneth Alberto Funes Mora and Jean-Marc Odobez. 2016. Gaze Estimation in the 3D Space Using RGB-D Sensors - Towards Head-Pose and User Invariance. *International Journal of Computer Vision* 118, 2 (2016), 194–216.

Carlos H. Morimoto and Marcio R. M. Mimica. 2005. Eye Gaze Tracking Techniques for Interactive Applications. *Comput. Vis. Image Underst.* 98, 1 (2005), 4–24.

Skanda Muralidhar, Rémy Siegfried, Jean-Marc Odobez, and Daniel Gatica-Perez. 2018. Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills?. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, MUM 2018, Cairo, Egypt, November 25-28, 2018.* 121–126. https://doi.org/10.1145/3282894.3282925

Benjamin A Newman, Reuben M Aronson, Siddartha S Srinivasa, Kris Kitani, and Henny Admoni. 2018. HARMONIC: A Multimodal Dataset of Assistive Human-Robot Collaboration. *arXiv preprint arXiv:1807.11154* (2018).

S. Sheikhi and J.M. Odobez. 2015. Combining dynamic head pose and gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters* 66 (Nov. 2015), 81–90.

Rémy Siegfried, Yu Yu, and Jean-Marc Odobez. 2017. Towards the use of social interaction conventions as prior for gaze model adaptation.. In *ICMI.* ACM, 154–162.

Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology.* ACM, 271–280.

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1821–1828.

Yarbus. 1967. Eye Movements and Vision. *Plenum* (1967).

Y. Yu, K. Funes, and J.-M. Odobez. 2018. HeadFusion: 360 degree Head Pose tracking combining 3D Morphable Model and 3D Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 40, 1 (Nov. 2018), 2653–2667.

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, 416.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 162–175.