

CoGaze: Webcam-Based Gaze Sharing for Asymmetric Remote Collaboration in Practical Deployment Contexts

Takuma Masuda
Nojima Lab
The University of
Electro-Communications
Tokyo, Japan
Augmented Human Lab
National University of Singapore
Singapore, Singapore
takuma@ahlab.org

Prasanth Sasikumar
Augmented Human Lab
National University of Singapore
Singapore, Singapore, Singapore
prasanth@ahlab.org

Koichi Hirota
Faculty of Informatics and
Engineering
The University of
Electro-Communications
Chofu, Tokyo, Japan
hirota@vogue.is.uec.ac.jp

Sho Sakurai
Graduate School of Informatics and
Engineering
The University of
Electro-Communications
Chofu, Tokyo, Japan
sho@vogue.is.uec.ac.jp

Takuya Nojima
University of
Electro-Communications
Tokyo, Japan
tnojima@nojilab.org

Suranga Nanayakkara
Augmented Human Lab
School of Computing, National
University of Singapore
Singapore, Singapore
suranga@ahlab.org

Abstract

Shared gaze helps re-establish joint attention in asymmetric remote collaboration, yet existing MR systems source the expert's gaze from a dedicated infrared eye tracker. We present CoGaze, which shares webcam-estimated expert gaze into a local worker's MR view while an infrared reference records simultaneously, turning estimation noise into a measured variable. A four-participant pilot places that noise at 5.9° on average, with 95% of samples within 12.1° . Whether workers tolerate it remains untested: an instrumentation defect left the displayed cue dominated by the infrared stream, so the reduction in grounding cost we observed under gaze indicates the value of gaze presence, not tolerance of webcam noise. We contribute the system, the measured in-situ error bounds, a speech-based grounding measure sensitive where task accuracy was at ceiling, and a successor design calibrated to those bounds.

CCS Concepts

• Human-centered computing → Computer supported cooperative work.

Keywords

Remote Collaboration, Webcam, Eye tracking, Asymmetric Collaboration, MR

ACM Reference Format:

Takuma Masuda, Prasanth Sasikumar, Koichi Hirota, Sho Sakurai, Takuya Nojima, and Suranga Nanayakkara. 2026. CoGaze: Webcam-Based Gaze

Sharing for Asymmetric Remote Collaboration in Practical Deployment Contexts. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3830397.3841904>

1 Introduction

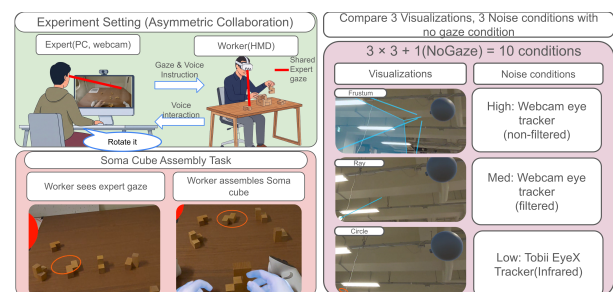


Figure 1: (a) Asymmetric setting: a Remote Expert at a webcam-equipped PC guides a Local Worker wearing an HMD in a physical workspace. (b) Assembly task: the Worker places Soma cube pieces under the Expert's gaze-directed guidance. (c) Study design: three visualizations (Ray, Circle, Frustum) crossed with three tracking sources (unfiltered webcam, filtered webcam, IR reference), plus a NoGaze control.

Gaze sharing helps re-establish joint attention in remote collaboration [4, 6, 12]. Yet existing gaze-sharing MR systems source the expert's gaze from dedicated infrared eye-tracking hardware [4, 13, 21]. Appearance-based webcam gaze estimation offers a hardware-free alternative, but no gaze-sharing system for physical remote assistance has adopted it, to our knowledge: its angular error of several degrees persists even in recent estimators [7, 15, 18].



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

UIST Adjunct '26, Detroit, MI, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2855-6/26/11

<https://doi.org/10.1145/3830397.3841904>

The noise arises on the expert’s side, but the worker must resolve the displaced or unstable cue into the correct physical target. The open question is therefore worker-side: can a worker establish joint attention from gaze cues carrying webcam-level noise, and does visualization design change the answer? We evaluate CoGaze from the worker’s side (Fig. 1), with an infrared reference recording simultaneously in every webcam trial, so we measure the noise rather than assume it.

In summary, we contribute the system, measured in-situ noise bounds from a four-participant pilot, and a study configuration revised from these findings; tolerance of that noise remains untested here, for reasons we state in Sec. 4.

2 Related Work

Shared gaze reduces verbal referencing overhead and supports spatial grounding in remote physical tasks [4, 6, 10, 12], and visualization form matters as much as gaze presence [3]. Existing MR gaze-sharing systems nevertheless rely on dedicated IR eye trackers [4, 11–13, 21]. Appearance-based webcam gaze estimation has been studied as a single-user screen-interaction modality [1, 7, 22]; even recent estimators report errors of several degrees in unconstrained settings [1, 7, 15, 18]. Prior noise-sensitivity studies tested only two visualization types, at noise levels that webcam estimation reaches or exceeds [6, 17]. Scaling visualization size to gaze uncertainty provides intuitive reliability feedback [2, 9, 19, 20], but no study has evaluated it in a collaborative task.

3 Study Design

CoGaze connects a Remote Expert at a webcam-equipped PC (Fig. 1(a)) to a Local Worker who wears a Meta Quest 3 in a physical workspace. Two tasks probe referential grounding and guided manipulation: identification and assembly.

Apparatus and pipeline: The HMD streams live video of the Worker’s workspace to the Expert’s PC. A commodity webcam (640×480, 30 fps) captures the Expert’s face. MediaPipe Face Mesh [16] extracts head-translation-invariant iris features, which a Ridge regression fitted on a 16-point calibration maps to screen coordinates. The pipeline also emits two per-frame confidence signals: c_{mesh} reflects face-landmark quality and c_{eye} reflects eyelid closure. A chin rest stabilizes the Expert’s head at approximately 50 cm; it is an experimental control for comparable calibration, not a requirement of the deployed system. To measure estimation error rather than assume it, an infrared tracker (Tobii EyeX) records simultaneously in all webcam conditions, giving every webcam sample a per-frame deviation from the IR reference. QR markers anchor the projection of the Expert’s screen-space gaze into the Worker’s HMD view and serve as targets in the identification task.

Visualization: Three primitives, namely Ray, Circle, and Frustum, differ in spatial extent and implied precision (Fig. 1(c)). Thresholding the confidence signals gates each primitive’s visibility, withdrawing the cue while estimation is unreliable and prompting verbal coordination. In the WebcamFiltered condition, a One-Euro filter [5] additionally smooths the coordinates.

Design: A 3×3 factorial design crosses tracking source with visualization type; a NoGaze control brings the total to ten conditions (Fig. 1(c)). A Williams square over tracking blocks counterbalances

presentation order. Each condition comprises gaze calibration in the webcam conditions, a 60 s identification trial in which the Worker grips the indicated QR marker, a 180 s assembly trial in which the Worker assembles Soma cube pieces at the positions the Expert indicates (Fig. 1(b)), and an in-HMD questionnaire. Four Local Workers (age $M = 22.5$ years, $SD = 0.58$), recruited within the authors’ laboratory without compensation, each completed all ten conditions in an instrumented pilot. A single trained researcher served as the Remote Expert throughout. We ran this pilot as an instrument test: the sample size was set to verify that the apparatus and the measures behaved as intended, not to detect differences between conditions.

Guidance and measures: The Expert guided assembly with gaze plus demonstratives such as “this one”; rotation words were always permitted because gaze cannot convey orientation. After a stall of roughly 15 s or accumulated misplacements, guidance escalated along a three-rung ladder: rung 1, deictic reference; rung 2, feature and relative descriptions such as “one cell to the right”; rung 3, absolute spatial language such as “top-right.” NoGaze guidance was speech-only. We measured identification accuracy and time-to-first-correct; verbal grounding demand, operationalized as the highest rung reached and the onset time of escalated language; NASA-TLX [8]; SSQ [14]; and post-session preference rankings. Grounding measures were coded from the Expert’s recorded speech (Sec. 4). With four Workers, we aggregate trials to participant-level means and compare conditions with paired t -tests, reporting Cohen’s d_z as the effect size and checking presentation-order effects.

4 Pilot Findings

Data validity: Running the pilot as an instrument test surfaced two defects, stated first to delimit the results below. (1) c_{mesh} returned a constant value, so confidence gating operated on c_{eye} alone. (2) Both gaze sources shared one display channel, so the displayed cue in webcam conditions was dominated by the higher-rate IR stream; contrasts between *displayed* tracking sources are attenuated. The results below rest on per-source gaze logs and this speech coding, which the defects do not affect; each defect has a localized fix.

Gaze estimation error: The time-matched webcam–IR recordings cover approximately 38 minutes across the four participants (67,701 frame pairs at 30 fps). The webcam pipeline deviated from the IR reference by 5.88° on average ($SD = 1.08^\circ$ across participants). At the 95th percentile, samples fell within 12.1° , which we adopt as a worst-case bound for design purposes. Per-condition calibration holdout errors were roughly 2° – 5° , so the calibration residual accounts for a large share of the total error. One-Euro filtering did not reduce the deviation (6.40° , $SD = 0.94^\circ$; $t(3) = -0.72$, $p = .52$, $d_z = -0.36$). One account is that the filter trades The filter may trade jitter for lag, which the per-frame metric penalizes, though tuning is not ruled out.

Verbal grounding: Post-hoc coding of 1,941 transcribed Expert utterances across all trials shows the expected asymmetry. In identification, gaze conditions stayed almost purely deictic (highest rung: $M = 1.25$, $SD = 0.29$), whereas every NoGaze trial required absolute spatial language ($M = 3.00$, $SD = 0.00$; $t(3) = -11.98$, $p = .001$, $d_z = -5.99$). In assembly, escalated language began $M = 6.9$ s into NoGaze trials ($SD = 5.4$) versus $M = 24.0$ s with gaze ($SD = 4.5$;

$t(3) = 6.12$, $p = .009$, $d_z = 3.06$). Identification accuracy was high in all conditions (0.84–0.97): at this scale, grounding cost, not task success, is the sensitive measure.

Workload, comfort, and preference: Raw NASA-TLX did not differ across tracking sources (participant-level means 3.2–3.4 on 0–6; Friedman $\chi^2(3) = 0.90$, $p = .83$), nor did identification accuracy ($p = .51$) or time-to-first-correct ($p = .55$). Preferences split across primitives: two of the four Workers ranked Ray first because it combines spatial context with a precise endpoint, one ranked Circle first for its usefulness in assembly, and one ranked Frustum first.

5 Discussion and Future Work

Presenting noisy gaze. The gaze-versus-NoGaze gap arose under an IR-dominated display (Sec. 4), so it shows the value of gaze presence, not tolerance of webcam noise. Within that boundary, the pilot yields design guidance. First-place rankings split two, one, and one, which at this sample size does not distinguish the primitives. A point-like endpoint implies precision a 6° signal lacks, while a Circle sized to the 12° bound occludes the workspace. Cue size and target spacing must therefore be co-designed around the measured bound [9, 19, 20]. With the confidence signal validated, uncertainty becomes a usable visual variable: extent scaled by momentary confidence through low-occlusion encodings (outline-only rings, transparency), rather than this pilot’s binary gating.

From pilot to revised design. Two structural limits end this design: identification sat at ceiling even speech-only, and ten conditions took two hours per participant. The successor places distractors inside the measured annulus, reduces conditions per sitting, and inherits the escalation ladder with onset time as its primary measure.

References

- [1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. 2023. L2CS-Net : Fine-Grained Gaze Estimation in Unconstrained Environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, 98–102. doi:10.1109/ICFSP59764.2023.10372944
- [2] Stavros Antifakos, Adrian Schwaninger, and Bernt Schiele. 2004. Evaluating the effects of displaying uncertainty in context-aware applications. In *International Conference on Ubiquitous Computing*. Springer, 54–69.
- [3] Jad A Atweh and Sara L Riggs. 2025. Gaze sharing, a double-edged sword: Examining the effect of real-time gaze sharing visualizations on team performance and situation awareness. *Human Factors* 67, 3 (2025), 196–224. doi:10.1177/00187208241272060
- [4] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [5] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1ε filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on human factors in computing systems*. 2527–2530.
- [6] Austin Erickson, Nahal Norouzi, Kangsoo Kim, Ryan Schubert, Jonathan Jules, Joseph J LaViola Jr, Gerd Bruder, and Gregory F Welch. 2020. Sharing gaze rays for visual target identification tasks in collaborative augmented reality. *Journal on Multimodal User Interfaces* 14, 4 (2020), 353–371.
- [7] Lucas Falch and Katrin Solveig Lohan. 2024. Webcam-based gaze estimation for computer screen interaction. *Frontiers in Robotics and AI* 11 (2024), 1369566.
- [8] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [9] Almoctar Hassoumi, Vsevolod Peysakhovich, and Christophe Hurter. 2018. Uncertainty visualization of gaze estimation to support operator-controlled calibration. *Journal of Eye Movement Research* 10, 5 (2018), 31.
- [10] Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2016. Can eye help you? Effects of visualizing eye fixations on remote collaboration scenarios for physical tasks. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5180–5190. doi:10.1145/2858036.2858438
- [11] Allison Jing, Gun Lee, and Mark Billinghurst. 2022. Using speech to visualise shared gaze cues in MR remote collaboration. In *2022 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 250–259.
- [12] Allison Jing, Kieran May, Brandon Matthews, Gun Lee, and Mark Billinghurst. 2022. The impact of sharing gaze behaviours in collaborative mixed reality. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.
- [13] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. eyemR-Vis: A Mixed Reality System to Visualise Bi-Directional Gaze Behavioural Cues Between Remote Collaborators. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–4. doi:10.1145/3411763.3451545
- [14] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.
- [15] Eduard Kuric, Peter Demcak, Jozef Majzel, and Giang Nguyen. 2025. Democratizing eye-tracking? Appearance-based gaze estimation with improved attention branch. *Engineering Applications of Artificial Intelligence* 149 (2025), 110494.
- [16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, Vol. 2019. Long Beach, CA, 2.
- [17] Nahal Norouzi, Austin Erickson, Kangsoo Kim, Ryan Schubert, Joseph LaViola, Gerd Bruder, and Greg Welch. 2019. Effects of shared gaze parameters on visual target identification task performance in augmented reality. In *Symposium on spatial user interaction*. 1–11.
- [18] Shreshth Saxena, Elke Lange, and Lauren Fink. 2022. Towards efficient calibration for webcam eye-tracking in online experiments. In *2022 symposium on eye tracking research and applications*. 1–7.
- [19] Sarah Schömb, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-assisted decision-making: Unveiling the role of uncertainty visualisation and embodiment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [20] Ching-Yi Tsai, Nicole Tacconi, Andrew D. Wilson, and Parastoo Abtahi. 2026. Uncertain Pointer: Situated Feedforward Visualizations for Ambiguity-Aware AR Target Selection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–31.
- [21] Yue Wang, Peng Wang, Zhiyong Luo, and Yuxiang Yan. 2022. A novel AR remote collaborative platform for sharing 2.5 D gestures and gaze. *The International Journal of Advanced Manufacturing Technology* 119, 9 (2022), 6413–6421.
- [22] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4511–4520.