# Misjudging the Machine: Gaze May Forecast Human-Machine Team Performance in Surgery

Sue Min Cho, Russell H. Taylor, and Mathias Unberath

Johns Hopkins University, Baltimore MD, USA
{scho72, rht, unberath}@jhu.edu

**Abstract.** In human-centered assurance, an emerging field in technology-assisted surgery, humans assess algorithmic outputs by interpreting the provided information. Focusing on image-based registration, we investigate whether gaze patterns can predict the efficacy of human-machine collaboration. Gaze data is collected during a user study to assess 2D/3D registration results with different visualization paradigms. We then comprehensively examine gaze metrics (fixation count, fixation duration, stationary gaze entropy, and gaze transition entropy) and their relationship with assessment error. We also test the effect of visualization paradigms on different gaze metrics. There is a significant negative correlation between assessment error and both fixation count and fixation duration; increased fixation counts or duration are associated with lower assessment errors. Neither stationary gaze entropy nor gaze transition entropy displays a significant relationship with assessment error. Notably, visualization paradigms demonstrate a significant impact on all four gaze metrics. Gaze metrics hold potential as predictors for human-machine performance. The importance and impact of various gaze metrics require further task-specific exploration. Our analyses emphasize that the presentation of visual information crucially influences user perception.

**Keywords:** Image-guided surgery · 2D/3D registration · Gaze

## 1 Introduction

Surgery is undergoing a digital transformation, transitioning from relying solely on human skill to integrating advanced tools such as imaging and robotics. Despite these technological advancements, the need for human judgment in this high-stakes, safety-critical domain remains unchanged [5]. The emergence of human-centered assurance in technology-assisted surgery underscores the importance of integrating human operators into the technological loop. Delving deeper into human-machine teaming is critical for providing system oversight for patient safety and the efficacy of surgical procedures [4]. Thus, understanding how operators perceive, interpret, and act upon the data and suggestions offered by the technology is imperative for the success of the overall safety assurance.

A crucial window into understanding this process is the human gaze, which serves as a bridge between external stimuli and internal cognition [1]. As a natural human behavior, the gaze reflects attention and cognitive processes [6, 8]. The

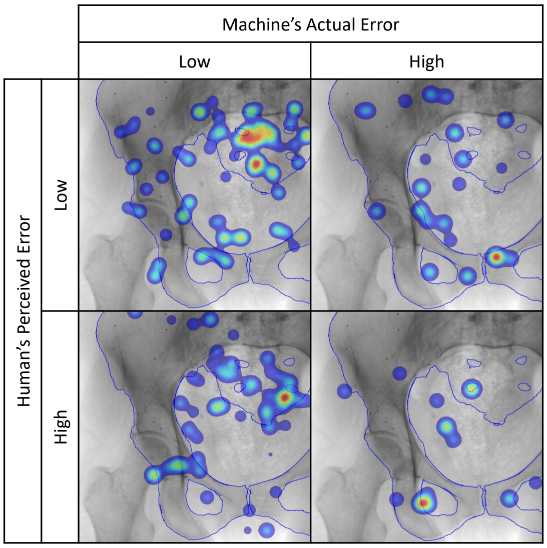


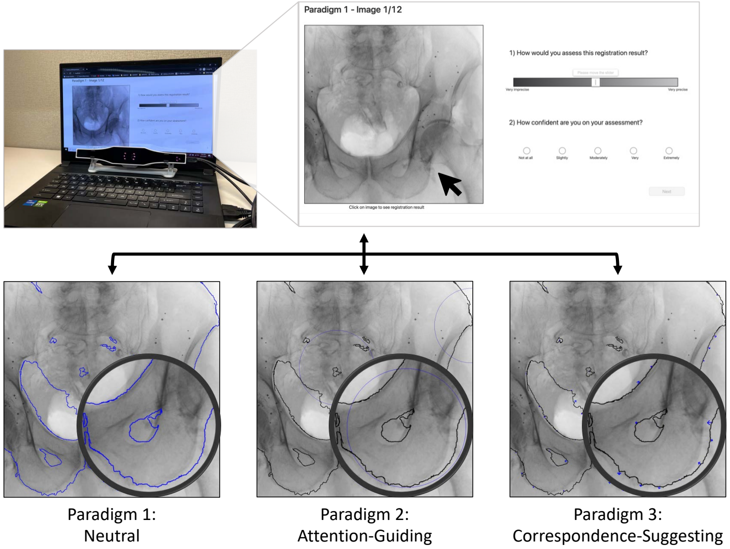
**Fig. 1.** Heatmaps of gaze patterns when human assesses machine's error.

patterns of gaze, including trajectories and fixation points, can offer profound insights into cognitive processes, understanding, and eventual decisions.

In the healthcare domain, technological progress has broadened the scope of gaze research. Gaze data from radiologists interpreting x-ray images can augment the training of deep learning models for computer-aided diagnosis systems [2, 11]. It is also emerging as a metric to distinguish expertise levels, with studies reporting different gaze patterns between experts and novices in radiography [10] and surgical simulations [9].

Expanding on these developments, we extend gaze analysis to another area of healthcare. Our focus centers on 2D/3D registration, a critical algorithm in image-guided surgery that enables optimal spatial alignment of 3D preoperative models with 2D intraoperative images, as depicted in the supplementary material. Since the true solution is not known during surgery, errors cannot be detected automatically. A feasible approach to ensuring safety is to involve human operators to verify the system's results [3]. In this study, we investigate users' gaze patterns during the visual assessment of the algorithm's results. This addresses a critical area that has yet to be explored in human-centered assurance in technology-assisted surgery: understanding gaze patterns during human-based spatial alignment assessment. Figure 1 provides visual examples of gaze patterns when users correctly versus incorrectly assess machine-generated registration errors, highlighting the intricate nature of the task and setting the stage for a detailed examination in our study.

We tailor gaze metrics well-established and standard in eye-tracking research to 2D/3D registration assessment by computing weighted metrics with image similarity metrics in the areas of interest. Analyzing these weighted metrics offers a valuable perspective on gaze patterns during human-based assessment of spatial alignment that has not been previously documented. By examining how users visually interact with algorithmic outputs, we can uncover the underlying factors that contribute to their assessment performance. This is not merely exploratory but holds practical significance; understanding gaze patterns can lead to the development of novel assurance mechanisms. Using gaze data, these mechanisms can potentially predict the accuracy of a user's assessments, thus providing a quantitative measure of the user's evaluation uncertainty, serving as an assurance metric for the human-machine team.


Paradigm 1 - Image 1/12

1) How would you assess this registration result?

2) How confident are you on your assessment?

Click on image to see registration result

**Fig. 2.** Experimental setup showcasing the gaze tracker and the user interface used in the study. Includes example cases for each visualization paradigm, with zoomed-in bubbles to emphasize key features: Paradigm 1 - contour only, Paradigm 2 - contour + attention-guiding circles, Paradigm 3 - contour + correspondence-suggesting arrows.

## 2 Methods

### 2.1 Gaze Data Collection

We recruited 22 participants, each with some background in medical imaging, image processing, or both, to evaluate simulated 2D/3D registration results using different visualization paradigms. The paradigms included: Paradigm 1: Neutral – outline of the projected digitally reconstructed radiograph (DRR, a computed 2D image from CT data simulating a conventional x-ray) only; Paradigm 2:

Attention-Guiding – outline and circular highlights to regions with possible mismatch; Paradigm 3: Correspondence-Suggesting – outline and arrows connecting the outline to possibly better matching visual features; examples can be seen in Figure 2. For specific details of the visualization paradigms and simulation of registration results, we kindly refer to [3].

The overall experimental setup is depicted in Figure 2. The study was conducted on a 15.6-inch display with a resolution of 3840 x 2160 to ensure clarity and detail in the images and overlays. To maintain a standardized viewing experience, participants could not pan or zoom into the registration correspondences, allowing focus on their gaze patterns in relation to the fixed images presented.

Participants encountered all visualization paradigms in a randomized order to negate any learning effects. The study began with an introductory briefing, followed by a preparatory step for each paradigm. In this step, participants were presented with examples demonstrating varied degrees of registration errors specific to that paradigm, thus familiarizing them with the range of potential misalignments for the subsequent evaluation phase. Participants then assessed the alignment of 12 x-ray images with their registration overlay, indicating their confidence level for each image.

During the study, participants' gaze data were collected using a Gazepoint GP3 eye tracker operating at 60Hz. We opted for a stationary screen-based eye tracker due to its anticipated accuracy for our tasks (i.e., inspecting an image and overlays on a screen). We excluded 12 participants due to failed gaze tracking. This exclusion was based on criteria such as frequent or prolonged failures in eye tracking recording, often caused by participants altering their initial position or wearing glasses that interfered with the tracker, impacting gaze recording. The integrity of our data was paramount, and this strict exclusion criterion was necessary to ensure the reliability of our findings.

### 2.2 Areas of Mismatch Quantification

We posit that areas that exhibit a large mismatch between edge information in the acquired x-ray image and DRR are most informative for alignment. To enable quantitative evaluation of how much participants spent visually assessing those regions, automated techniques to identify areas of mismatch are needed. We identify areas of mismatch through normalized cross-correlation (NCC).

Given two images, an x-ray, and its corresponding DRR of the simulated registration offset, the NCC between patches of these images serves as a metric to generate a set of weights for subsequent weighted aggregation of gaze metrics. The metric values are inverted so that regions with higher metric values are represented with smaller weights and lower metric values are with larger weights, thus putting more importance on the areas of higher mismatch.

### 2.3 Gaze Metrics

Given the $n$ square patches in the image, these are referred to as Areas of Interest (AOIs). Traditional eye tracking metrics, including fixation count and duration, are calculated in tandem with gaze entropies [7]. These combined metrics effectively quantify the attention distribution, scanning behavior, and overall exploration extent across the AOIs. The weights detailed in Section 2.2 attributed to each AOI play a pivotal role in these evaluations. Within the framework of analyzing 2D/3D registration results, these weights emphasize areas with pronounced mismatches, as such regions correlate directly with the registration's alignment errors. As a result, analyzing these weighted gaze patterns offers a valuable perspective on the human user's proficiency in focusing on mismatched areas for their assessment.

**Fixation Count** The weighted fixation count for the $i^{th}$ AOI is given by:

$$FC_{AOI_i} = \sum_{i=1}^{n} w_i \cdot f_i$$

where $w_i$ is the weight for $AOI_i$ and $f_i$ is the total fixation count in $AOI_i$.

**Fixation Duration** The weighted fixation duration for the $i^{th}$ AOI is represented as:

$$FD_{AOI_i} = \sum_{i=1}^{n} w_i \cdot d_i$$

where $w_i$ is the weight for $AOI_i$ and $f_i$ is the total fixation duration in $AOI_i$.

**Stationary Gaze Entropy** Stationary gaze entropy quantifies the overall spatial dispersion of gaze:

$$SGE = -\sum_{i=1}^{n} p_i \log(p_i + 0.00000001)$$

where $p_i$ is the probability of a gaze remaining stationary in the respective AOI.

**Gaze Transition Entropy** Gaze transition entropy quantifies the rate of fixation transitions between AOIs:

$$GTE = -\sum_{i=1}^{n} p_i \left( \sum_{j=1}^{n} t_{ij} \log(t_{ij} + 0.00000001) \right)$$

where $p_i$ is the stationary probability for the $i^{th}$ AOI and $t_{ij}$ is the transition probability from the $i^{th}$ AOI to the $j^{th}$ AOI.

### 2.4 Hypotheses

**H1.** Gaze metrics can be predictors for human-based 2D/3D registration assessment.

**H2.** Gaze metrics are influenced by different visualization paradigms.

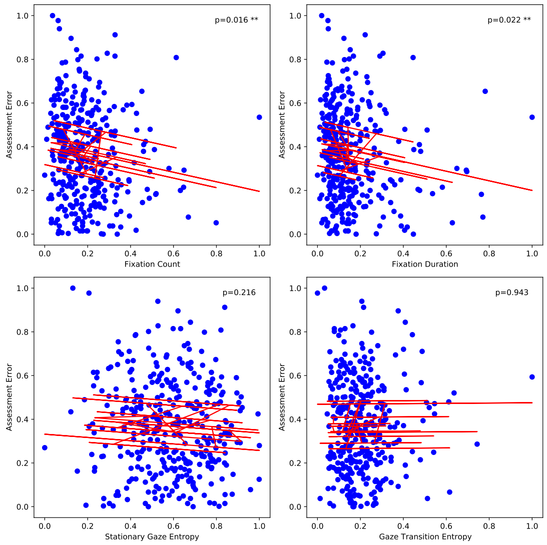


**Fig. 3.** Linear regression plots of gaze metrics against assessment error. Each line corresponds to an individual user, treated as a random effect in the model.

### 2.5 Data Analysis

To test our hypotheses, we utilized linear mixed models (LMM) for their ability to account for individual differences among participants, treating participants as a random effect to accommodate variability in gaze patterns. We ensured the dependent variable, assessment error, was continuous and approximately normally distributed, and we normalized the gaze metrics during preprocessing. To confirm normality in residual distributions, we conducted an extensive evaluation using Quantile-quantile (Q-Q) plots and D'Agostino's K-squared Test. The Q-Q plots were visually inspected to assess if the residuals followed a normal distribution. D'Agostino's K-squared Test was used to statistically verify the normality of residuals, with p-values above 0.05 indicating a normal distribution. Additionally, we evaluated homoscedasticity by examining the spread of residuals against fitted values. Independence of observations was checked by ensuring that the data collection methods did not introduce any dependencies among observations. Python was used for data analysis.

## 3 Results

### 3.1 Can gaze metrics be predictors for human-based assessment performance?

For Hypothesis 1, key assumptions were examined, including independence of observations and homoscedasticity, before proceeding with the analysis. Q-Q plot

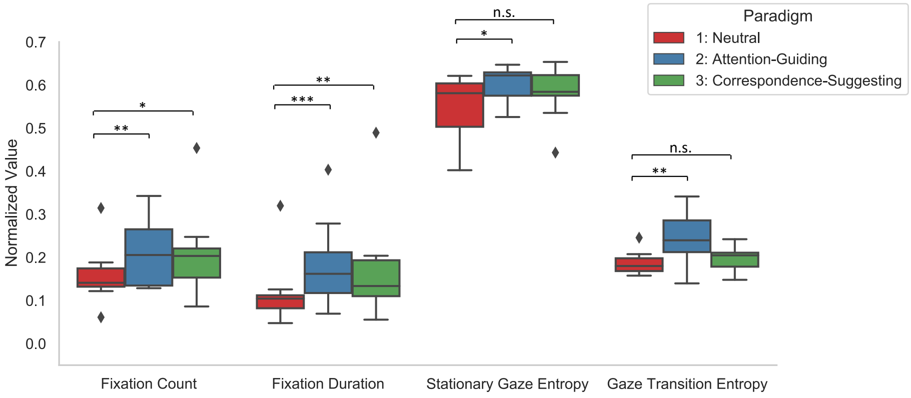


**Fig. 4.** Boxplots illustrating the variations in gaze metrics across different visualization paradigms.
$^{n.s}$not significant, $*p < 0.05$, $**p < 0.01$, $***\ p < 0.001$

observations revealed residuals were nearly normally distributed. This was corroborated by the D'Agostino's K-squared Test, which produced p-values above 0.05 across all four LMM models for various gaze metrics, implying residuals closely resembled a normal distribution.

The findings, summarized in the supplementary material and visually represented in Figure 3, indicate a statistically significant inverse relationship between two specific gaze metrics – fixation count and duration – and assessment error. The fixation count showed a coefficient of -0.219 (p=0.016), suggesting that an increase in fixation count correlates with decreased assessment errors. This relationship was significant at the p < 0.05 level. Similarly, fixation duration had a coefficient of -0.220 (p=0.022), indicating that longer fixation durations are associated with reduced errors, also significant at the p < 0.05 level. The descending trajectories of the regression lines for these metrics in the scatterplots underscore these trends. In contrast, stationary gaze entropy and gaze transition entropy did not demonstrate statistically significant relationships with assessment error. Their coefficients of -0.074 (p=0.216) and 0.007 (p=0.943), respectively, suggest a lack of correlation with assessment error, as further shown by the faint linear progression in their scatterplots.

These findings support our first hypothesis, demonstrating that certain gaze metrics, specifically fixation count and duration, can indeed serve as predictors for human-based assessment performance in 2D/3D registration tasks. The non-significant results for other metrics like stationary gaze entropy and gaze transition entropy highlight the specificity of the gaze metrics' predictive power.

### 3.2 Do visualization paradigms affect gaze metrics?

For Hypothesis 2, the categorical variable "paradigm" represents the three different visualization paradigms, with Paradigm 1 as the reference category. The independence of observations and homoscedasticity were confirmed. Visual inspection of Q-Q plots indicated that the residuals closely approximated a normal distribution. D'Agostino's K-squared Test further supported this, yielding p-values above 0.05 for LMM models for fixation count, fixation duration, and gaze transition entropy. However, stationary gaze entropy had a p-value below 0.05, so it should be interpreted with caution.

The coefficients for "Paradigm 2 vs. 1" and "Paradigm 3 vs. 1" in the model represent the changes in gaze metrics when shifting from Paradigm 1 to Paradigms 2 and 3, respectively. Statistically significant p-values in these coefficients would suggest meaningful differences in gaze metrics attributable to the visualization paradigms. Our findings, as graphically depicted in Figure 4 and detailed in the supplementary material, reveal distinct effects of these paradigms on gaze metrics. Specifically, both Paradigms 2 and 3 significantly influenced the fixation count and duration compared to Paradigm 1, with p-values of 0.007, 0.013, and less than 0.001 and 0.002, respectively. This indicates a notable impact of visualization paradigm on these gaze metrics. For stationary gaze entropy, a significant effect was observed only for Paradigm 2 (p=0.012), while Paradigm 3 did not show a significant difference. Similarly, gaze transition entropy was significantly influenced only by Paradigm 2 (p=0.001), with no significant effect observed for Paradigm 3. These results underscore the differential impact of visualization paradigms on various gaze metrics, affirming their role in shaping user interaction and interpretation in 2D/3D registration tasks.

## 4 Discussion and Conclusion

Two gaze metrics, fixation count and fixation duration, were significant indicators of human-based registration error assessment performance. These metrics underscore the importance of focused observation for accurate evaluation, suggesting that prolonged, deliberate gaze upon mismatch areas correlates with increased assessment precision. Conversely, stationary gaze entropy and gaze transition entropy showed no significant correlation. This may be because, despite the presence of misalignment cues, they were presented statically, leaving users to decide where and when to look. This insight informs the design of visualization paradigms that dynamically guide users' attention to mismatch areas, enhancing assessment accuracy. Future research could explore paradigms that sequentially reveal misalignment cues based on severity, potentially making entropy calculations more controlled across individuals.

Visualization paradigms also significantly impacted gaze metrics. Prior studies, such as [3], have indicated varying efficacies of different paradigms. This study resonates with their findings, demonstrating that the mode of visual presentation affects how users perceive the given information. Further research is needed to understand better how these findings can inform the design of more effective visualization and interaction methodologies.

Although our sample size is currently small, the observed significant effects and large effect size provide a compelling basis for our findings. Informed by this preliminary work, future studies can be designed with more robust power analysis and targeted participant recruitment.

Future work should refine data collection methods to minimize data exclusion. Despite achieving significant effects with corruption-free data post-exclusion, there is an opportunity for enhanced methodology. Participants altering their position and eyeglasses interference were identified as common sources of gaze recording failures. Future studies can explore advanced hardware for more stable, accurate gaze tracking. Considering alternative participant recruitment approaches, such as specific eyeglasses criteria, could improve data quality.

This study serves as a foundational step in understanding gaze metrics within the human-machine partnership in image-guided surgery. Our empirical evidence shows that gaze metrics have the potential to forecast error and uncertainty in human-based registration assessment and that the efficacy of assessments is affected by visual presentation. Yet, for a holistic comprehension, deeper and more meticulous exploration is necessary. Understanding these interrelationships is vital in formulating enhanced methodologies for human-in-the-loop technology-assisted interventions. As the frontier of human-machine collaboration in surgery continues to expand, these findings will be essential in establishing the reliability and adaptability of such systems in various pursuits.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Benedek, M., Stoiser, R., Walcher, S., Körner, C.: Eye behavior associated with internally versus externally directed cognition. Frontiers in psychology **8**, 1092 (2017)
2. Bigolin Lanfredi, R., Zhang, M., Auffermann, W.F., Chan, J., Duong, P.A.T., Srikumar, V., Drew, T., Schroeder, J.D., Tasdizen, T.: Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Scientific data **9**(1), 350 (2022)
3. Cho, S.M., Grupp, R.B., Gomez, C., Gupta, I., Armand, M., Osgood, G., Taylor, R.H., Unberath, M.: Visualization in 2d/3d registration matters for assuring technology-assisted image-guided surgery. International Journal of Computer Assisted Radiology and Surgery pp. 1–8 (2023)
4. Fiorini, P., Goldberg, K.Y., Liu, Y., Taylor, R.H.: Concepts and trends in autonomy for robot-assisted surgery. Proceedings of the IEEE **110**(7), 993–1011 (2022)
5. Grundgeiger, T., Hurtienne, J., Happel, O.: Why and how to approach user experience in safety-critical domains: the example of health care. Human factors **63**(5), 821–832 (2021)
6. Just, M.A., Carpenter, P.A.: A theory of reading: from eye fixations to comprehension. Psychological review **87**(4), 329 (1980)
7. Krejtz, K., Szmidt, T., Duchowski, A.T., Krejtz, I.: Entropy-based statistical analysis of eye movement transitions. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 159–166 (2014)

8. Lai, M.L., Tsai, M.J., Yang, F.Y., Hsu, C.Y., Liu, T.C., Lee, S.W.Y., Lee, M.H., Chiou, G.L., Liang, J.C., Tsai, C.C.: A review of using eye-tracking technology in exploring learning from 2000 to 2012. Educational research review **10**, 90–115 (2013)
9. Li, S., Duffy, M.C., Lajoie, S.P., Zheng, J., Lachapelle, K.: Using eye tracking to examine expert-novice differences during simulated surgical training: A case study. Computers in Human Behavior **144**, 107720 (2023)
10. McLaughlin, L., Bond, R., Hughes, C., McConnell, J., McFadden, S.: Computing eye gaze metrics for the automatic assessment of radiographer performance during x-ray image interpretation. International journal of medical informatics **105**, 11–21 (2017)
11. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)