Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies

Pierre Chatelain, *Member, IEEE*, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble

***Abstract*—Gaze tracking is a promising technology for studying the visual perception of clinicians during image-based medical exams. It could be used in longitudinal studies to analyze their perceptive process, explore human–machine interactions, and develop innovative computer-aided imaging systems. However, using a remote eye tracker in an unconstrained environment and over time periods of weeks requires a certain guarantee of performance to ensure that collected gaze data are fit for purpose. We report the results of evaluating eye tracking calibration for longitudinal studies. First, we tested the performance of an eye tracker on a cohort of 13 users over a period of one month. For each participant, the eye tracker was calibrated during the first session. The participants were asked to sit in front of a monitor equipped with the eye tracker, but their position was not constrained. Second, we tested the performance of the eye tracker on sonographers positioned in front of a cart-based ultrasound scanner. Experimental results show a decrease of accuracy between calibration and later testing of 0.30° and a further degradation over time at a rate of 0.13°. month^{-1}. The overall median accuracy was 1.00° (50.9 pixels) and the overall median precision was 0.16° (8.3 pixels). The results from the ultrasonography setting show a decrease of accuracy of 0.16° between calibration and later testing. This slow degradation of gaze tracking accuracy could impact the data quality in long-term studies. Therefore, the results we present here can help in planning such long-term gaze tracking studies.**

***Index Terms*—Accuracy, biomedical imaging, calibration, data analysis, gaze tracking, testing, ultrasonography.**

I. Introduction

GAZE tracking is the process of determining one's point-of-gaze as a temporal sequence of coordinates. Spatially registering the points-of-gaze to a visual stimulus displayed on a monitor requires a calibration which depends on several factors related to the subject (physiological properties, presence of sight correction apparatus) or the environment (illumination, position of the monitor) [1]. While controlled cross-subject and cross-device comparative studies of gaze tracking accuracy are available in [2], these do not provide any information on whether the accuracy may degrade in time. In addition, vendor-provided performance information is an estimate assuming a specific use context [e.g., desktop-based human–computer interaction (HCI)].

Especially from the biomedical engineering perspective, there are several motivations for exploring human perception via gaze tracking. Namely, understanding the medical image perception process by assessment of visual search strategies of clinicians [3], studying human visual expertise during image acquisition and interpretation for potential skill assessment [4], improving and designing graphical visualizations and interfaces for HCI in medical settings [5], and developing computer-based image analysis methods, such as automatic detection of medical image contents and computational modeling of human visual attention [6]–[8].

First, the holistic understanding of expert perceptive and cognitive learning processes in terms of visual scan, search and recognition strategies, and decision-making has been of considerable interest to the scientific community. Studies in this direction include the evaluation of visual search tasks in radiology [9]–[12]; computed tomography [13]–[15]; mammography [16]–[19]; magnetic resonance imaging (MRI) [20], [21]; pathology [22]–[24]; and ultrasound imaging [25], [26].

Second, variations in visual expertise and behavior between experts, trainees, and novices have been analyzed in different clinical settings, such as endoscopic surgery [27]–[29]; radiology [30]–[33]; mammography [34]; pathology [35], [36]; computed tomography [37], [38]; pediatric neurology [39]; dermatology [40]; and ultrasound-guided anesthesia [41].

Third, gaze tracking has been widely explored for HCI and usability research [5], [42]. The information inferred from eye tracking devices can be utilized to determine areas of interest and visual search patterns in an interface, and to evaluate the visibility, usefulness, and position of its elements. This has been shown to be helpful in improving interface design for more efficient system interaction, e.g., reduced overall visual

Manuscript received March 3, 2018; revised June 10, 2018; accepted August 4, 2018. This work was supported by the European Research Council (project PULSE) under Grant ERC-ADG-2015 694581. The work of A. T. Papageorghiou was supported by the National Institute of Health Research Oxford Biomedical Research Center. The work of J. A. Noble was supported by EPSRC (project Seebibyte) under Grant EP/M013774/1. This paper was recommended by Associate Editor M. Shin. *(Corresponding author: Pierre Chatelain.)*

P. Chatelain, H. Sharma, and J. A. Noble are with the Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, U.K. (e-mail: pierre.chatelain@eng.ox.ac.uk; alison.noble@eng.ox.ac.uk).

L. Drukker is with the Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, U.K., and also with the Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford OX3 7DQ, U.K.

A. T. Papageorghiou is with the Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford OX3 7DQ, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2866274

clutter [43] and cognitive workload. An example of an HCI study for medical imaging is [44], where gaze tracking was used to analyze visual search paths of radiologists to provide useful insights for designing efficient radiology workstations.

Finally, recently there has been an emergence of interest in the image analysis literature in gaze tracking to understand its role in constraining image interpretation. For example, applications of gaze tracking-based computer vision include object recognition [45]–[47], action recognition [48]–[50]; and caption generation [51]. Likewise, expert human knowledge has been leveraged to design advanced medical image analysis algorithms for automatic detection and classification of image contents in mammography [52], retinal images [53], MRI [54], and ultrasound images [55], [56]. From the computer vision point of view, gaze tracking has informed saliency-based visual attention methods [57], leading to a number of computational visual saliency models. Correspondingly, prediction of visual attention in medical images has been performed using gaze data as ground truth in radiology and retinal images [58], endoscopy [59], and ultrasound images [60].

However, the construction of meaningful models of medical image perception requires a large amount of real-world data to account for the natural variability of the task as well as the natural variability in human perception. This involves the acquisition of gaze data across an extended period of time (potentially several months). In order to facilitate the analysis and interpretation of gaze data, it is important that the eye tracker is accurately calibrated for each observer, and that the calibration does not drift with time. Therefore, we wanted to understand how the accuracy of a commercial eye tracker varies in time, and whether a regular recalibration is necessary.

Longitudinal stability of gaze tracking accuracy has previously been reported at timescales of the order of minutes. Gòmez-Poveda and Gaudioso [61] evaluated the temporal stability of different eye tracking algorithms for webcams in a single continuous session, i.e., the stability of measurements across consecutive camera frames. Johansen *et al.* [62] compared the accuracy of two eye trackers with nine participants and at four different instances, separated by a pause of 2 min. The authors reported that the effect of time elapsed since calibration was not statistically significant at this timescale. Hence, to the best of our knowledge, there has not been a longitudinal study assessing performance at larger timescales. However, long-term temporal factors could have an impact on the accuracy of gaze tracking.

1) Changes in the external environment, such as head position [63], [64] or illumination [65].
2) Changes in the appearance of the user. This is relevant as most eye trackers use image-based tracking algorithms to take measurements from the user's face and eyes.

In this paper, we report the results of a longitudinal study of gaze tracking performance conducted over a period of one month. The aim of this paper was to evaluate whether the accuracy of gaze tracking is stable over this time interval after a single initial calibration for each user. This issue has been studied in related areas such as biometric recognition. For instance, it has been suggested that the accuracy of iris recognition may decrease in the very long term [66]. A recent study reported a diminution of face recognition accuracy with years [67]. However, accuracy over several years is not relevant to the range of applications of interest in biomedical image analysis. Indeed, the purpose of having a temporally robust gaze tracking system is to acquire data on human visual behavior while demanding minimum time and effort (i.e., a minimal number of calibrations) from the studied subjects (busy clinical professionals working in clinics).

Our investigation was divided into two studies. The first was a desktop study with 13 participants. It was intended as a reference for image viewing on a desktop monitor. The second is an *in situ* study, which specifically looked at accuracy in the context of cart-based ultrasonography (a cart-based scanner is the most commonly used device for ultrasound exams). The environment of an ultrasound exam is different from the desktop setting in which gaze tracking is usually performed. First, the amplitude of motion of a sonographer is typically larger than that of someone sitting at a desk in front of a computer, and the variability of head positioning is also larger due to the flexibility of the cart-based ultrasound scanner. Second, the ultrasound exam is performed in the dark, with the monitor as the main source of luminosity. Since the head position and room illumination are two important factors impacting eye tracking quality, it is of interest to estimate how an eye tracker performs in these conditions. Note that we purposefully did not constrain the movement of the participants, because the objective was to evaluate the temporal evolution of gaze tracking performance in a real use situation rather than the performance under optimal conditions.

TABLE I
MONITOR SPECIFICATIONS

	Monitor 1	Monitor 2
Resolution	1920×1200	1920×1080
Pixel pitch	2.7×10^{-4} m	2.4×10^{-4} m
Refresh rate	60 Hz	60 Hz

II. METHODS

A. Experimental Setup

For all experiments, we used a Tobii Eye Tracker 4C (Tobii, Sweden). This remote eye tracking device provides an estimate of the point-of-gaze and 3-D eye position for each eye at 90 Hz.

For the longitudinal desktop study, a Dell Ultrasharp U2413 monitor was used with the relevant specifications summarized in Table I under Monitor 1. For the *in situ* study, a Philips EPIQ 7G cart-based ultrasound scanner was used, consisting of an articulated monitor with the characteristics summarized under Monitor 2 in Table I. For each monitor, the eye tracker was rigidly attached under the display area with a magnetic mounting bracket, following the product's instructions. The experimental setup for the *in situ* study is represented in Fig. 1.

B. Calibration

A 9-point calibration was performed for each user, following the method in [1]. The calibration targets were defined on a

Fig. 1. Experimental setup for the *in situ* study. The eye tracker is fixed on the monitor of a cart-based ultrasound scanner. A visual stimulus is displayed to perform calibration and testing of the eye tracker.

regular grid at relative horizontal and vertical screen positions of 0.1, 0.5, and 0.9

$$\text{Targets} = \{0.1, 0.5, 0.9\} \times \{0.1, 0.5, 0.9\}. \tag{1}$$

The visual stimulus was a white disk on a black background. The calibration protocol was as follows:

procedure CALIBRATION(Targets)
  for $\mathbf{T} \in$ Targets **do**
    Expand stimulus to size s_0
    Move stimulus to $\mathbf{T}$
    Shrink stimulus to size s_1
    Acquire calibration data
  end for
  Compute calibration parameters
end procedure.

This protocol was implemented in C++ using Qt for the GUI and the Tobii Pro SDK (C language binding, version 1.1.4.5) for controlling the eye tracker. The acquisition of calibration data and the computation of calibration parameters were performed using functions of the Tobii Pro SDK. The radius of the stimulus was $s_0 = 10$ pixels between acquisitions, and $s_1 = 5$ pixels during acquisition. The duration of the stimulus shrinking and expanding animations was 500 ms and the duration of motion from one target to another was 2000 ms.

The calibration for each user was saved on the disk through the Tobii Pro SDK, so that it could be loaded later for testing.

C. Testing

The testing protocol was similar to the calibration protocol. Gaze data was acquired at the same nine targets as in Section II-B. The testing protocol was as follows:

procedure TESTING(Targets,numberOfSamples)
  Samples $\leftarrow$ EmptyList
  for $\mathbf{T} \in$ Targets **do**
    Expand stimulus to size s_0
    Move stimulus to $\mathbf{T}$
    Shrink stimulus to size s_1
    $n \leftarrow 0$
    while $n <$ numberOfSamples **do**
      $S \leftarrow$ acquireSample()
      if isValid(S) **then**
        Samples $\leftarrow$ append(Samples, S)
        $n \leftarrow n + 1$
      end if
    end while
  end for
  return Samples
end procedure.

D. Evaluation Metrics

Evaluation metrics were calculated as pixel and angular measures for the eye tracking device [1].

1) Pixel Measures: Let $\mathbf{G} = (g_x, g_y)$ denote the gaze measurement in screen coordinates, with x corresponding to the horizontal axis, and y to the vertical axis. Given a series $(\mathbf{G}_i)_{i=1}^N$ of N measurements corresponding to the same visual target, the fixation point $\mathbf{F}$ is defined as

$$\mathbf{F} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{G}_i \tag{2}$$

and the pixel accuracy A_{pixels} for a fixation $\mathbf{F}$ acquired over target $\mathbf{T} \in$ Targets from (1) is defined as

$$A_{\text{pixels}}(\mathbf{F}, \mathbf{T}) = \|\mathbf{F} - \mathbf{T}\|_2. \tag{3}$$

For the same series $(\mathbf{G}_i)_{i=1}^N$ of continuous measurements, ordered in chronological order, the precision P_{pixels} is defined as the root mean square of displacements [68]

$$P_{\text{pixels}}(\mathbf{G}_1, \ldots, \mathbf{G}_N) = \sqrt{\frac{1}{N}\sum_{i=1}^{N-1} \|\mathbf{G}_i - \mathbf{G}_{i+1}\|_2^2}. \tag{4}$$

2) Angular Measures: In order to convert the evaluation metrics from pixels to degrees, one needs to know the distance d between the eyes and the screen and the pixel pitch p (physical size of a pixel). The distance can be obtained for each eye using the eye position $\mathbf{E}$ and 3-D gaze point position $\mathbf{G}^{\text{3-D}}$ provided by the eye tracker

$$d_i = \|\mathbf{E}_i - \mathbf{G}_i^{\text{3-D}}\|_2 \tag{5}$$

where $i \in \{\text{left}, \text{right}\}$ indicates the left and right eyes. The average distance is $d = (1/2)(d_{\text{left}} + d_{\text{right}})$.

Then, the angular accuracy A_{ang} and angular precision P_{ang} can be computed as

$$A_{\text{ang}} = \text{atan}\left(\frac{pA_{\text{pixels}}}{d}\right) \tag{6}$$

$$P_{\text{ang}} = \text{atan}\left(\frac{pP_{\text{pixels}}}{d}\right). \tag{7}$$

E. Desktop Study

We evaluated the performance of the eye tracker on a cohort of 13 participants during one month. Each participant was calibrated at time T_0 and tested right after the calibration, and then at $T_0 + 1$ hour, $T_0 + 1$ day, then every week to (and including) the fourth week (seven test sessions per participants). The attendance rate was 95.6% (four sessions were missed).

1) Protocol:

a) First session: For the initial session, each participant was asked to sit in front of the monitor equipped with the eye tracker. They were free to adjust the monitor's height and inclination, and to move the chair. Then, they were asked to look at the visual stimulus displayed on the monitor, in order to perform the calibration. The lights were turned off, so as to have similar illumination conditions to the second study on sonographers. The calibration was performed as described in Section II-B. Note that the estimation of the calibration parameters performed by the Tobii Pro SDK can fail if the quality of the calibration data is insufficient. In such a case, the calibration procedure would be attempted one more time. Where the calibration was successful, the test procedure described in Section II-C was performed under the same conditions. The participant did not leave the experimental station between calibration and testing.

b) Test sessions: For each subsequent session, the experimental protocol was similar, and testing was performed as described in Section II-C. Therefore, the main differences from the first test were:

1) the participant had been away from the station between calibration and testing;
2) the configuration of the monitor and chair may have changed;
3) a certain amount of time had elapsed.

2) Models:

a) Linear model: We wanted to investigate whether the accuracy and precision (response variables) of an eye tracker are constant with respect to the time elapsed since the calibration. Explanatory variables are: participant; time since calibration; session type (initial/testing); target location; and distance to screen. Time since calibration is the specific focus of this paper. It is a continuous variable, therefore its effect can be modeled through regression. The participant variable is also of interest, because it can be used to study the interuser variability in accuracy and precision. Target location and distance to screen, on the other hand, cannot be controlled during eye tracking acquisitions without constraining the participant. Therefore, these can be considered as part of the residual variability in the response variables. In order to perform a regression while modeling interuser variability, we need to use a mixed-effect model. Thus, we fitted a generalized linear mixed-effect model [69] to the evaluation measure y (either A_{ang} or P_{ang}). The model is defined as

$$y_i(t)|\mu_i(t) \sim \text{Distr}\Big(\mu_i(t), \sigma^2\Big) \tag{8}$$

$$\mu_i(t) = \beta_0 + \beta_{0,i} + \big(\beta_1 + \beta_{1,i}\big)t \tag{9}$$

$$\begin{bmatrix} \beta_{0,i} \\ \beta_{1,i} \end{bmatrix} \sim \mathcal{N}\Big(\mathbf{0}, \sigma_r^2\mathbf{D}\Big) \tag{10}$$

where $i = 1 \ldots M$ denotes the participant, t is the time since calibration, $\text{Distr}(\mu, \sigma^2)$ is a statistical distribution with location parameter μ and dispersion parameter σ, $\mu_i(t)$ is the location parameter of the model with fixed-effects intercept and slope β_0, β_1, random-effects intercept and slope $\beta_{0,i}$, $\beta_{1,i}$ and σ_r is the mixing standard deviation, accounting for interuser variability. The matrix $\mathbf{D}$ parameterizes the correlations between random effects. This model is of interest to our problem because it performs a regression on the time variable t while explicitly modeling the interuser variability through σ_r and the residual variability through σ. The residual variability corresponds to the uncontrolled explanatory variables (target position) and testing conditions.

b) Two-stage linear model: One potential issue with the purely temporal model above is that it only accounts linearly for the time elapsed, without modeling the initial effect of leaving the station and coming back for testing (independently of the time delay). To account for this effect, we define a second mixed-effect model with an additional variable $I_i \in \{0, 1\}$ indicating whether the participant has been away between calibration and testing. The model can thus be defined as

$$y_i(t)|\mu_i(t) \sim \text{Distr}\Big(\mu_i(t), \sigma^2\Big) \tag{11}$$

$$\mu_i(t) = \beta_0 + \beta_{0,i} + \big(\beta_1 + \beta_{1,i}\big)t + \big(\beta_2 + \beta_{2,i}\big)I_i \tag{12}$$

$$\begin{bmatrix} \beta_{0,i} \\ \beta_{1,i} \\ \beta_{2,i} \end{bmatrix} \sim \mathcal{N}\Big(\mathbf{0}, \sigma_r^2\mathbf{D}\Big) \tag{13}$$

where β_2 is the parameter corresponding to the effect of leaving the station.

F. In Situ Study

We evaluated the performance of the eye tracker for a cohort of three sonographers. Each sonographer was calibrated at time T_0 and tested right after the calibration, and at $T_0 + t$, where $t > 1$ day is a variable parameter according to the availability of sonographers for the study.

1) Protocol: The protocol for the *in situ* study was similar to the desktop study with two main differences.

1) The monitor of the cart-based ultrasound scanner provided more flexibility than the desktop monitor, as the sonographers were free to adjust its inclination, height, rotation, and distance according to their preference.
2) The time intervals for subsequent test sessions were not fixed due to strict time constraints on the sonographers. However, the interest was in quantifying the initial shift in performance, due to the variability in the configuration of monitor of the ultrasound scanner. Indeed, the rate of change of performance due to time alone was expected to be similar to that observed in the desktop environment study. Thus, the measurements for individual test sessions after the first session were concatenated into a single group for each sonographer, instead of multiple test sessions at fixed time intervals for each participant in the desktop study.

2) Models: A linear mixed-effects model is not adequate to represent the variation in calibration metrics, as the data is recorded at fewer time instances for three sonographers. Hence, we performed a comparison of the statistical distribution of the performance metrics between the two time points.

TABLE II
FIXED EFFECTS COEFFICIENTS OF THE LINEAR MIXED-EFFECTS MODELS FOR ACCURACY AND PRECISION

	parameter	estimate	SE	correlation matrix[a]
$\log(A_{\text{ang}})$[b]	β_0	-4.156	0.040	$\begin{bmatrix} 1.00 & -0.70 & -0.07 & — \\ -0.70 & 1.00 & 0.05 & — \\ -0.07 & 0.05 & 1.00 & — \\ — & — & — & 1.00 \end{bmatrix}$
	β_1	0.008	0.002	
	σ_r	0.509	0.039	
	$\log(\sigma^2)$	-0.450	0.051	
$\log(P_{\text{ang}})$[b]	β_0	-5.774	0.133	$\begin{bmatrix} 1.00 & -0.43 & -0.59 & 0.28 \\ -0.43 & 1.00 & 0.09 & -0.09 \\ -0.59 & 0.09 & 1.00 & -0.22 \\ 0.28 & -0.09 & -0.22 & 1.00 \end{bmatrix}$
	β_1	0.002	0.003	
	σ_r	0.306	0.036	
	$\log(\sigma^2)$	-0.263	0.056	

[a] Correlation values smaller than 0.01 are replaced with —.
[b] Angle measurements in rad, rates in $\text{rad} \cdot \text{d}^{-1}$ (In logarithmic scale).

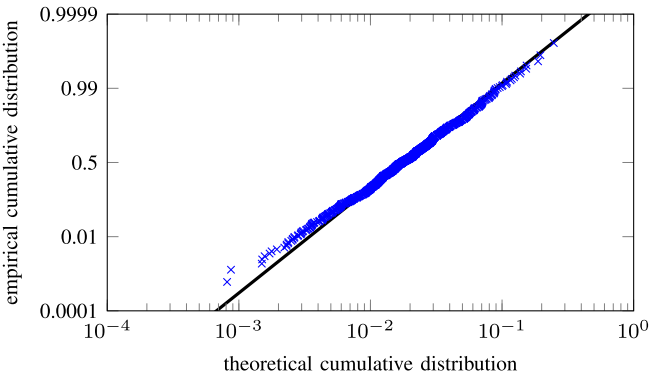

Fig. 2. Probability plot for accuracy against the lognormal distribution. The data is close to the theoretical distribution.

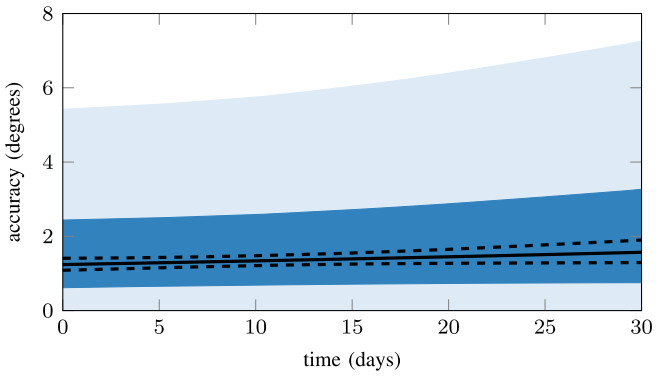

Fig. 3. Fitted linear model for accuracy. The cohort mean trend is shown as a solid line, with the 95% confidence band as dashed lines. The dark blue region represents the one-sided 95% confidence band for user variability. The light blue region represents the one-sided 95% confidence band for residual variability.

III. RESULTS

A. Calibration

The calibration succeeded at the first attempt for 12 out of 13 participants. In the single failure case, the calibration was successful at the second attempt.

B. Desktop Study

1) Exploratory Data Analysis: In order to identify a suitable statistical distribution to use in our model fitting we used probability plots to visually check the goodness of fit for a set of candidate distributions that could model the variability in accuracy: lognormal, Rayleigh, and Weibull. The visually best fit amongst those was obtained for the lognormal distribution (Fig. 2).

$Y \sim \text{Lognormal}(\mu, \sigma^2)$ is equivalent to $\log(Y) \sim \mathcal{N}(\mu, \sigma^2)$, so in the following we fit a normal distribution to the logarithm of accuracy and precision.

2) Linear Model: We estimated the parameters of the linear model (8)–(10) for accuracy and precision separately, using the generalized nonlinear mixed models function `gnlmm` of the R package `repeated` (v1.1.0). The scale and shape parameters where initialized using the `gnlr` function for generalized nonlinear regression from the R package `gnlm`(v1.1.0). The scale parameters (β_0, β_1), mixing standard deviation σ_r, and log-shape estimate $\log(\sigma^2)$ for the logarithm of accuracy $\log(A_{\text{ang}})$ and precision $\log(P_{\text{ang}})$ are reported in Table II, along with the corresponding standard error (SE) and correlation matrix. For each model we also present, in a figure, the fixed effects trend as a function of time, with the 95% confidence band for the fixed effects coefficients as dashed lines. The dark blue region represents the 95% confidence band for the interuser variability. The light blue region represents the one-sided 95% confidence band for the residual above the dark blue boundary.

a) Accuracy: The estimated parameters indicate a cohort mean accuracy at $t = 0$ of 1.24° and a slope of $0.30\,^\circ \cdot \text{month}^{-1}$ at $t = 0$ and $0.38\,^\circ \cdot \text{month}^{-1}$ at $t = 30\,\text{d}$. The correlation between random intercept and random slope is -0.70. The temporal model for accuracy and associated confidence bands are also represented as a function of time in Fig. 3.

b) Precision: The estimated parameters indicate a cohort mean precision at $t = 0$ of 0.26° and a slope of $0.01\,^\circ \cdot \text{month}^{-1}$ (not statistically significant). The temporal model for precision and associated confidence bands are presented in Fig. 4.

3) Two-Stage Linear Model: To fit the two-stage linear model (11)–(13), we defined an additional variable indicating whether the samples were acquired during the initial session or during a later one. Then, we used the same method as above to estimate the parameters of the model. The fixed effect scale coefficients $(\beta_0, \beta_1, \beta_2)$, mixing standard deviation σ_r, and log-shape estimate $\log(\sigma^2)$ for the logarithm of accuracy $\log(A_{\text{ang}})$ and precision $\log(P_{\text{ang}})$ are reported in Table III, along with the corresponding SE and correlation matrix.

TABLE III
FIXED EFFECTS COEFFICIENTS OF THE TWO-STAGE LINEAR MIXED-EFFECTS MODELS FOR ACCURACY AND PRECISION

	parameter	estimate	SE	correlation matrix[a]				
$\log(A_{\text{ang}})$[b]	β_0	-4.339	0.073	1.00	—	-0.84	-0.03	—
	β_1	0.004	0.003	—	1.00	-0.41	0.05	—
	β_2	0.259	0.087	-0.84	-0.41	1.00	-0.01	—
	σ_r	0.507	0.039	-0.03	0.05	-0.01	1.00	—
	$\log(\sigma^2)$	-0.461	0.051	—	—	—	—	1.00
$\log(P_{\text{ang}})$[b]	β_0	-5.574	0.082	1.00	0.01	-0.82	0.09	0.01
	β_1	-0.001	0.003	0.01	1.00	-0.41	0.02	-0.05
	β_2	0.132	0.096	-0.82	-0.41	1.00	-0.01	0.01
	σ_r	0.403	0.025	0.09	0.02	-0.01	1.00	0.12
	$\log(\sigma^2)$	-0.282	0.053	0.01	-0.05	0.01	0.12	1.00

[a] Correlation values smaller than 0.01 are replaced with —.
[b] Angle measurements in rad, rates in $\text{rad} \cdot \text{d}^{-1}$ (In logarithmic scale).

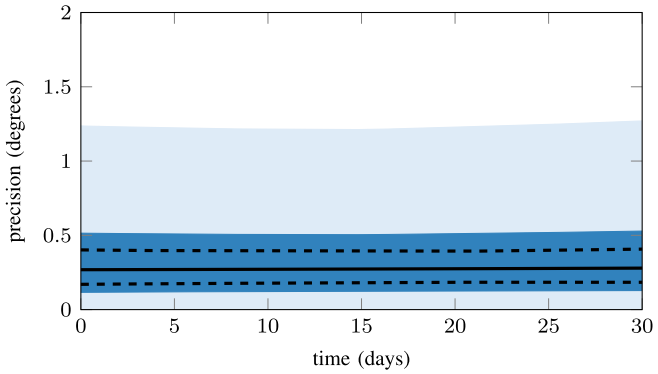

Fig. 4. Fitted linear model for precision. The cohort mean trend is shown as a solid line, with the 95% confidence band as dashed lines. The dark blue region represents the one-sided 95% confidence band for user variability. The light blue region represents the one-sided 95% confidence band for residual variability.

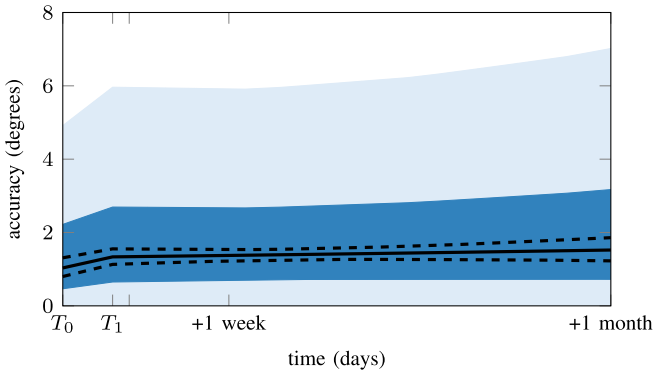

Fig. 5. Fitted linear model for accuracy. The cohort mean trend is shown as a solid line, with the 95% confidence band as dashed lines. The dark blue region represents the one-sided 95% confidence band for user variability. The light blue region represents the one-sided 95% confidence band for residual variability. The point T_0 corresponds to the initial session (following calibration) and the point T_1 corresponds to the first session thereafter, so that the evolution from T_0 to T_1 is discrete.

a) Accuracy: The estimated parameters indicate a cohort mean accuracy at $t = 0$ of 1.03°, an initial shift of 0.30°, and a slope of $0.13\,^\circ \cdot \text{month}^{-1}$ at $t = 0$ and $0.15\,^\circ \cdot \text{month}^{-1}$ at $t = 30\,\text{d}$. The correlation between intercept (β_0) and shift (β_2) is -0.84, the correlation between shift (β_2) and time (β_1) is -0.41 and the correlation between intercept (β_0) and time is less than 0.01. The model and associated confidence bands are presented in Fig. 5.

b) Precision: The estimated parameters indicate a cohort mean precision at $t = 0$ of 0.32°, an initial shift of 0.04° and a slope of $-0.01\,^\circ \cdot \text{month}^{-1}$ (not statistically significant). The model and associated confidence bands are represented in Fig. 6.

4) Pixel Performance: The results given above were in terms of angular errors. These allow the comparison between measurements taken at different distances from the monitor. Of practical interest are the metric or pixel measurements, which indicate how precisely the eye tracking system can locate the point-of-gaze on the display monitor. We represent the cohort-wide median accuracy and precision for each test target on a 24-inch monitor in Fig. 7. The overall median accuracy was 50.9 pixels (13.8 mm) and the overall median precision was 8.3 pixels (2.2 mm). The accuracy at the center of the display area (31.7 pixels, 8.6 mm) was significantly better than for the peripheral targets. Similarly, the precision was significantly better at the center (6.1 pixels, 1.6 mm).

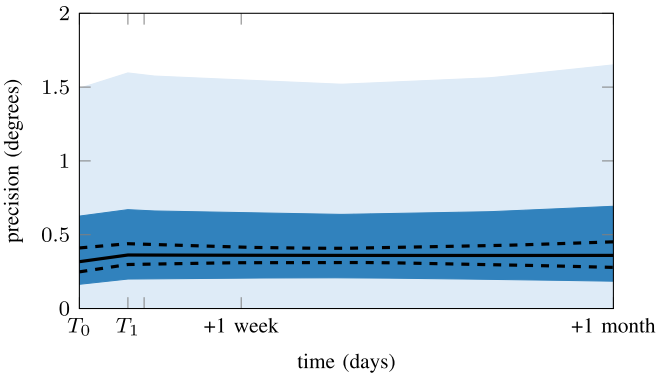

Fig. 6. Fitted linear model for precision. The cohort mean trend is shown as a solid line, with the 95% confidence band as dashed lines. The dark blue region represents the one-sided 95% confidence band for user variability. The light blue region represents the one-sided 95% confidence band for residual variability. The point T_0 corresponds to the initial session (following calibration) and the point T_1 corresponds to the first session thereafter, so that the evolution from T_0 to T_1 is discrete.

C. In Situ Study

Three sonographers participated in the study. For one of them, we performed three calibrations (on different days) in order to have more samples at T_0. The calibration of the eye

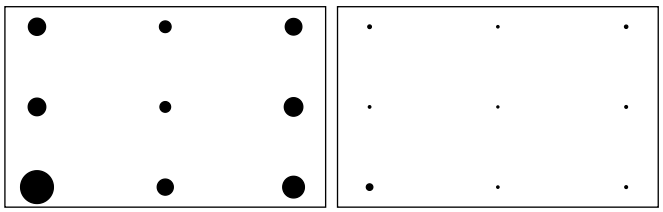

Fig. 7. Median accuracy (left) and precision (right) per test target on a 24-inch monitor. The rectangles represent the display area, and the circles' radius is scaled to the median performance.

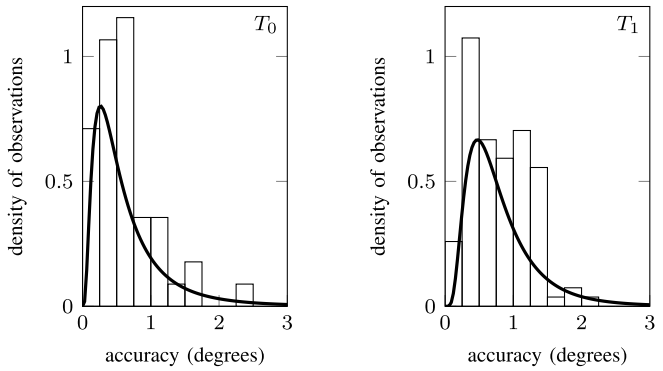

Fig. 8. Normalized histogram of gaze tracking accuracy on an ultrasound scanner and fitted lognormal probability density function at calibration (T_0, left) and at a later test session (T_1, right). Forty-five samples were used at T_0 and 108 at T_1. The histograms have bin widths of 0.25° and are normalized so that bin heights sum to one. The probability density functions (red curves) are multiplied by the histogram bin width.

tracker on the ultrasound scanner was successful for five out of five attempts. We performed ten testing sessions on different days with one sonographer and one testing session with the second sonographer. The third sonographer did not follow-up after the initial session.

Normalized histograms of the measurements for accuracy and precision are provided in Figs. 8 and 9, respectively.

a) Accuracy: The overall median accuracy was 0.65° (30.1 pixels, 7.4 mm). A two-sample Kolmogorov–Smirnov test indicated that the difference between the distributions of accuracy values at T_0 and T_1 is statistically significant ($p < 0.05$). The mean error at T_1 was 0.16° larger than at T_0.

b) Precision: The overall median precision was 0.09° (4.5 pixels, 1.1 mm). A two-sample Kolmogorov–Smirnov test indicated that the difference between calibration time (group T_0) and subsequent testing (group T_1) ($p = 0.3$) was not statistically significant.

D. Distance to Screen

The distance between the user's eyes and the screen (or the eye tracker) is an important factor of gaze tracking quality. In real-world conditions, when the user's position is not constrained, there is no guarantee that the distance to screen is within the optimal range for the eye tracker. The working range of commercial eye trackers is optimized for a desktop environment. However, the standard distance to screen when

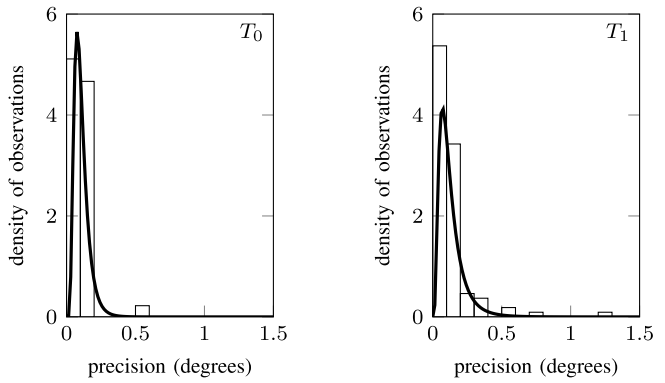

Fig. 9. Normalized histogram of gaze tracking precision on an ultrasound scanner and fitted lognormal probability density function at calibration (T_0, left) and at a later test session (T_1, right). Forty-five samples were used at T_0 and 108 at T_1. The histograms have bin widths of 0.10° and are normalized so that bin heights sum to one. The probability density functions (red curves) are multiplied by the histogram bin width.

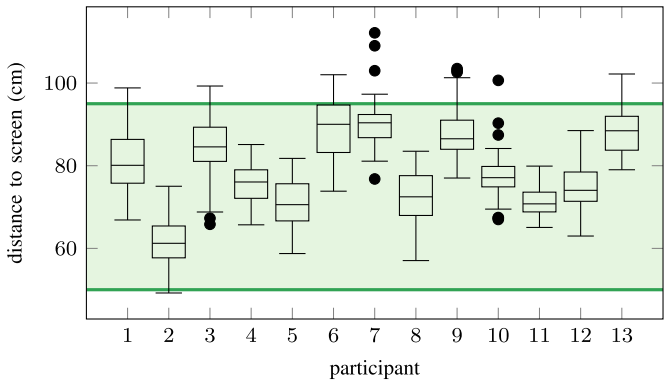

Fig. 10. Distance between the eye tracker and the eyes per user for the desktop study. The red lines indicate the recommended working range of the eye tracker.

using other types of work stations, such as a cart-based ultrasound scanner, is not known. In this section, we report some statistics on the distance to screen that was measured in this paper.

First, we report the distribution of the distance to screen per user for the desktop study as box-and-whiskers plots in Fig. 10. There is a significant interuser variability, but overall 94 % of observations are in the recommended working range of the eye tracker (50–95 cm). We also report in Fig. 11 the distribution of the distance to screen for different points in time. We observed no significant variation in time.

Finally, we report in Fig. 12 the distribution of distance to screen per user and per time point for the *in situ* study. All observations are within the recommended range for the eye tracker. As previously, we observed that the interuser variability is higher than the temporal variability.

IV. Discussion

The results of the desktop study support two main conclusions about the accuracy of the eye tracker. First, there is a statistically significant degradation of accuracy between the

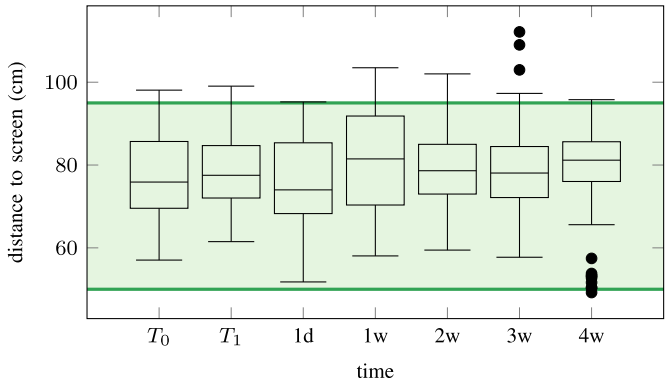

Fig. 11. Distance between the eye tracker and the eyes at different times for the desktop study. The red lines indicate the recommended working range of the eye tracker.

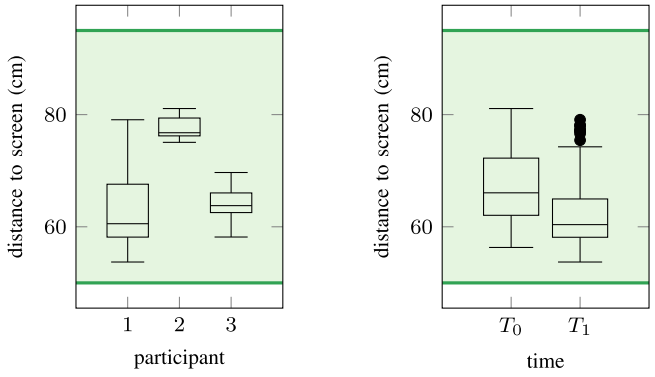

Fig. 12. Distance between the eye tracker and the eyes at different times for the *in situ* study. The red lines indicate the recommended working range of the eye tracker.

initial calibration time point and later testing time points. The mean shift was estimated at 0.30°, which corresponds to 30% of the initial accuracy. This suggests that, when more than one eye tracking session is planned after a single calibration, the accuracy during subsequent sessions may be reduced with respect to its initial level. A potential explanation for this effect is that the conditions of the acquisition, such as the illumination or the position of the eyes with respect to the monitor and eye tracker, have changed. These slight modifications in the conditions thus have an effect on the accuracy of the eye tracker. But note also that we did not see the same large shift between subsequent time points. Second, the results show a statistically significant degradation of accuracy with time, at an average rate of $0.13\,^\circ \cdot \text{month}^{-1}$. This suggests that when a gaze tracking study is planned over a certain period, the duration of the study should be taken into account when considering the gaze tracking accuracy requirements. Moreover, our results give an indication of the frequency at which the eye tracker should be recalibrated, in order to ensure a certain accuracy of measurements. For example, if it is critical to have a mean accuracy of 3° with a 95% confidence, then one should recalibrate the eye tracker after 24 days.

The correlation matrix of the linear mixed-effects model for accuracy shows a high negative correlation between intercept (accuracy at calibration) and temporal slope (−0.70), which suggests that the temporal decrease in accuracy is less important when the initial error is higher. However, the correlation matrix of the two-stage model shows that the most important correlations are between the intercept and the initial shift (−0.84), and between initial shift and slope (−0.41). This supports the conclusion that the two-stage model gives a more precise indication of the evolution of accuracy.

The results did not show any statistically significant change of precision with time. However, they also provide confidence intervals which may be useful when planning a gaze tracking study. This is particularly important because precision is arguably a more critical measure of quality for gaze tracking data. Indeed, while inaccuracy might be compensated *a posteriori*, low precision cannot be corrected.

The *in situ* results provide indications of the accuracy and precision of gaze tracking in a particularly challenging real-world environment, as the monitor of a cart-based ultrasound scanner is highly flexible, and there is more variability in the positioning of sonographers in front of the scanner than in the positioning of people in front of a desktop. The results give an indication of the performance that can be achieved in such an environment, and provide an estimation of the loss of accuracy between calibration and later use (0.16°). The experiments did not show any degradation of precision. Note that we found both the accuracy and precision to be better in the *in situ* study than in the desktop study. This is encouraging for the development of gaze tracking studies in real-world medical environments.

V. Conclusion

The main finding of this paper is that the accuracy of gaze tracking degrades slightly with time. We have identified two effects that may impact gaze tracking accuracy: 1) a natural change in conditions caused by leaving and returning to work station and 2) elapsed time, at a rate of about $8' \cdot \text{month}^{-1}$.

This was, to our knowledge, the first longitudinal study of gaze tracking performance at the timescale of weeks, and the first where the effect of natural changes in real-world experimental conditions was addressed. The practical significance of this paper is in providing guidance for the planning of real-world long-term gaze tracking studies, where repeated recalibration of the eye tracker is not desirable. While the rate of loss in accuracy is small, it was found to be practically significant at the scale of a month.

We expect these results to be of particular interest for the application of gaze tracking to the study of human understanding of biomedical images in real-world settings. Indeed, the acquisition of data from a clinical environment requires typically longer studies than in other domains, and it is often not feasible to perform daily calibrations. For such applications, the results we have presented will support the design of long-term gaze tracking studies while controlling the level of confidence in the measurements.

REFERENCES

[1] (Feb. 2011). *Accuracy and Precision Test Method for Remote Eye Trackers*. [Online]. Available: https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/use/what-affects-the-performance-of-an-eye-tracker/tobii-test-specifications-accuracy-and-precision-test-method.pdf/?v=2.1.1

[2] G. Funke *et al.*, "Which eye tracker is right for your research? Performance evaluation of several cost variant eye trackers," *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 60, no. 1, pp. 1240–1244, Sep. 2016.

[3] E. A. Krupinski, "Current perspectives in medical image perception," *Attention Percept. Psychophys.*, vol. 72, no. 5, pp. 1205–1217, Jul. 2010.

[4] S. E. Fox and B. E. Faulkner-Jones, "Eye-tracking in the study of visual expertise: Methodology and approaches in medicine," *Frontline Learn. Res.*, vol. 5, no. 3, pp. 43–54, Jul. 2017.

[5] A. Poole and L. J. Ball, "Eye tracking in HCI and usability research," in *Encyclopedia of Human Computer Interaction*, vol. 1, 1st ed., C. Ghaoui, Ed. Hershey, PA, USA: Idea Group Reference, 2006, pp. 211–219.

[6] P. L. Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proc. IEEE*, vol. 101, no. 9, pp. 2058–2067, Sep. 2013.

[7] Z. Liang, B. Xu, Z. Chi, and D. D. Feng, "Relative saliency model over multiple images with an application to yarn surface evaluation," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1249–1258, Aug. 2014.

[8] L. Zhang, X. Li, L. Nie, Y. Yang, and Y. Xia, "Weakly supervised human fixations prediction," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 258–269, Jan. 2016.

[9] E. L. Thomas and E. L. Lansdown, "Visual search patterns of radiologists in training," *Radiology*, vol. 81, pp. 288–292, Aug. 1963.

[10] H. L. Kundel and D. J. Wright, "The influence of prior knowledge on visual search strategies during the viewing of chest radiographs," *Radiology*, vol. 93, pp. 315–320, Aug. 1969.

[11] J. J. H. Leong, M. Nicolaou, R. J. Emery, A. W. Darzi, and G.-Z. Yang, "Visual search behaviour in skeletal radiographs: A cross-speciality study," *Clin. Radiol.*, vol. 62, no. 11, pp. 1069–1077, Nov. 2007.

[12] W. M. Reed, J. T. Ryan, M. F. McEntee, M. G. Evanoff, and P. C. Brennan, "The effect of abnormality-prevalence expectation on expert observer performance and visual search," *Radiology*, vol. 258, no. 3, pp. 938–943, Mar. 2011.

[13] H. Matsumoto *et al.*, "Where do neurologists look when viewing brain CT images? An eye-tracking study involving stroke cases," *PLoS ONE*, vol. 6, no. 12, Dec. 2011, Art. no. e28928.

[14] A. C. Venjakob, "Visual search, perception and cognition when reading stack mode cranial CT," Ph.D. dissertation, Fakultät V Verkehrs- und Maschinensysteme, Tech. Univ. Berlin, Berlin, Germany, 2015.

[15] G. Wen *et al.*, "Computational assessment of visual search strategies in volumetric medical images," *J. Med. Imag.*, vol. 3, no. 1, Jan. 2016, Art. no. 015501.

[16] E. A. Krupinski, "Visual search of mammographic images: Influence of lesion subtlety," *Acad. Radiol.*, vol. 12, no. 8, pp. 965–969, Aug. 2005.

[17] C. Mello-Thoms, "How does the perception of a lesion influence visual search strategy in mammogram reading?" *Acad. Radiol.*, vol. 13, no. 3, pp. 275–288, Mar. 2006.

[18] H. L. Kundel, C. F. Nodine, E. F. Conant, and S. P. Weinstein, "Holistic component of image perception in mammogram interpretation: Gaze-tracking study," *Radiology*, vol. 242, no. 2, pp. 396–402, Feb. 2007.

[19] K. K. Evans, R. L. Birdwell, and J. M. Wolfe, "If you don't find it often, you often don't find it: Why some cancers are missed in breast cancer screening," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e64366.

[20] L. Cooper *et al.*, "The assessment of stroke multidimensional CT and MR imaging using eye movement analysis: Does modality preference enhance observer performance?" in *Proc. SPIE Med. Imag. Image Percept. Observer Perform. Technol. Assess.*, vol. 7627. San Diego, CA, USA, 2010, Art. no. 76270B.

[21] C. Cavaro-Ménard, J.-Y. Tanguy, and P. L. Callet, "Eye-position recording during brain MRI examination to identify and characterize steps of glioma diagnosis," in *Proc. SPIE Med. Imag. Image Percept. Observer Perform. Technol. Assess.*, vol. 7627. San Diego, CA, USA, 2010, Art. no. 76270E.

[22] E. S. M. Tiersma, A. A. W. Peters, H. A. Mooij, and G. J. Fleuren, "Visualising scanning patterns of pathologists in the grading of cervical intraepithelial neoplasia," *J. Clin. Pathol.*, vol. 56, no. 9, pp. 677–680, Sep. 2003.

[23] D. Bombari, B. Mora, S. C. Schaefer, F. W. Mast, and H. A. Lehr, "What was I thinking? Eye-tracking experiments underscore the bias that architecture exerts on nuclear grading in prostate cancer," *PLoS ONE*, vol. 7, no. 5, May 2012, Art. no. e38023.

[24] E. A. Krupinski, A. R. Graham, and R. S. Weinstein, "Characterizing the development of visual search expertise in pathology residents viewing whole slide images," *Human Pathol.*, vol. 44, no. 3, pp. 357–364, Mar. 2013.

[25] A. Kosevoi-Tichie *et al.*, "THU0583 does eye gaze tracking have the ability to assess how rheumatologists evaluate musculoskeletal ultrasound images?" *Ann. Rheumatic Diseases*, vol. 74, no. S2, pp. 411–412, Jun. 2015.

[26] A. J. Carrigan, P. C. Brennan, M. Pietrzyk, J. Clarke, and E. Chekaluk, "A 'snapshot' of the visual search behaviours of medical sonographers," *Aust. J. Ultrasound Med.*, vol. 18, no. 2, pp. 70–77, May 2015.

[27] E. Kocak, J. Ober, N. Berme, and W. S. Melvin, "Eye motion parameters correlate with level of experience in video-assisted surgery: Objective testing of three tasks," *J. Laparoendosc. Adv. Surg. Techn. A*, vol. 15, no. 6, pp. 575–580, Dec. 2005.

[28] M. Wilson *et al.*, "Psychomotor control in a virtual laparoscopic surgery training environment: Gaze control parameters differentiate novices from experts," *Surg. Endoscopy*, vol. 24, no. 10, pp. 2458–2464, Oct. 2010.

[29] N. Ahmidi, M. Ishii, G. Fichtinger, G. L. Gallia, and G. D. Hager, "An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data," *Int. Forum Allergy Rhinol.*, vol. 2, no. 6, pp. 507–515, Nov. 2012.

[30] D. Manning, S. Ethell, T. Donovan, and T. Crawford, "How do radiologists do it? The influence of experience and training on searching for chest nodules," *Radiography*, vol. 12, no. 2, pp. 134–142, May 2006.

[31] T. Donovan and D. Litchfield, "Looking for cancer: Expertise related differences in searching and decision making," *Appl. Cogn. Psychol.*, vol. 27, no. 1, pp. 43–49, Jan. 2013.

[32] T. Drew, K. Evans, M. L.-H. Võ, F. L. Jacobson, and J. M. Wolfe, "Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images?" *Radiographics*, vol. 33, no. 1, pp. 263–274, Feb. 2013.

[33] G. Wood *et al.*, "Visual expertise in detecting and diagnosing skeletal fractures," *Skeletal Radiol.*, vol. 42, no. 2, pp. 165–172, Feb. 2013.

[34] C. F. Nodine, H. L. Kundel, S. C. Lauver, and L. C. Toto, "Nature of expertise in searching mammograms for breast masses," *Acad. Radiol.*, vol. 3, no. 12, pp. 1000–1006, Dec. 1996.

[35] E. A. Krupinski *et al.*, "Eye-movement study and human performance using telepathology virtual slides: Implications for medical education and differences with experience," *Human Pathol.*, vol. 37, no. 12, pp. 1543–1556, Dec. 2006.

[36] T. T. Brunyé *et al.*, "Eye movements as an index of pathologist visual expertise: A pilot study," *PLoS ONE*, vol. 9, no. 8, Aug. 2014, Art. no. e103447.

[37] R. Bertram, L. Helle, J. K. Kaakinen, and E. Svedström, "The effect of expertise on eye movement behaviour in medical image perception," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e66169.

[38] S. Mallett *et al.*, "Tracking eye gaze during interpretation of endoluminal three-dimensional CT colonography: Visual perception of experienced and inexperienced readers," *Radiology*, vol. 273, no. 3, pp. 783–792, Jul. 2014.

[39] T. Balslev *et al.*, "Visual expertise in paediatric neurology," *Eur. J. Paediatr. Neurol.*, vol. 16, no. 2, pp. 161–166, Mar. 2012.

[40] E. A. Krupinski, J. Chao, R. Hofmann-Wellenhof, L. Morrison, and C. Curiel-Lewandrowski, "Understanding visual search patterns of dermatologists assessing pigmented skin lesions before and after online training," *J. Digit. Imag.*, vol. 27, no. 6, pp. 779–785, Dec. 2014.

[41] L. K. Borg *et al.*, "Preliminary experience using eye-tracking technology to differentiate novice and expert image interpretation for ultrasound-guided regional anesthesia," *J. Ultrasound Med.*, vol. 37, no. 2, pp. 329–336, Feb. 2018.

[42] R. J. K. Jacob and K. S. Karn, "Commentary on section 4—Eye tracking in human–computer interaction and usability research: Ready to deliver the promises," in *The Mind's Eye*, J. Hyn, R. Radach, and H. Deubel, Eds. Amsterdam, The Netherlands: North-Holland, 2003, pp. 573–605.

[43] N. M. Moacdieh and N. Sarter, "The effects of data density, display organization, and stress on search performance: An eye tracking study of clutter," *IEEE Trans. Human–Mach. Syst.*, vol. 47, no. 6, pp. 886–895, Dec. 2017.

[44] M. S. Atkins, A. Moise, and R. Rohling, "An application of eyegaze tracking for designing radiologists' workstations: Insights for comparative visual search tasks," *ACM Trans. Appl. Percept.*, vol. 3, no. 2, pp. 136–151, Apr. 2006.
[45] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg, "Studying relationships between human gaze, description, and computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 739–746.
[46] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari, "Training object class detectors from eye tracking data," in *Computer Vision—ECCV 2014* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, Sep. 2014, pp. 361–376.
[47] X. Zhou *et al.*, "Eye tracking data guided feature selection for image classification," *Pattern Recognit.*, vol. 63, pp. 56–70, Mar. 2017.
[48] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.
[49] D. Stefic and I. Patras, "Action recognition using saliency learned from recorded human gaze," *Image Vis. Comput.*, vol. 52, pp. 195–205, Aug. 2016.
[50] F. Martinez, E. Pissaloux, and A. Carbone, "Towards activity recognition from eye-movements using contextual temporal learning," *Integr. Comput.-Aided Eng.*, vol. 24, no. 1, pp. 1–16, Jan. 2017.
[51] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, Jan. 2014.
[52] H. L. Kundel, C. F. Nodine, E. A. Krupinski, and C. Mello-Thoms, "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Acad. Radiol.*, vol. 15, no. 7, pp. 881–886, Jul. 2008.
[53] Ujjwal, K. S. Deepak, A. Chakravarty, and J. Sivaswamy, "Visual saliency based bright lesion detection and discrimination in retinal images," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, San Francisco, CA, USA, Apr. 2013, pp. 1436–1439.
[54] I. Mehmood, N. Ejaz, M. Sajjad, and S. W. Baik, "Prioritization of brain MRI volumes using medical image perception model and tumor region segmentation," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1471–1483, Oct. 2013.
[55] M. Ahmed and J. A. Noble, "Fetal ultrasound image classification using a bag-of-words model trained on sonographers' eye movements," *Procedia Comput. Sci.*, vol. 90, pp. 157–162, Jan. 2016.
[56] Y. Cai, H. Sharma, P. Chatelain, and J. A. Noble, "SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Washington, DC, USA, 2018, pp. 1475–1478.
[57] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Process.*, vol. 93, no. 6, pp. 1401–1407, Jun. 2013.
[58] V. Jampani, Ujjwal, J. Sivaswamy, and V. Vaidya, "Assessment of computational visual attention models on medical images," in *Proc. 8th Indian Conf. Comput. Vis. Graph. Image Process.*, Mumbai, India, 2012, pp. 1–8.
[59] A. J. Chung, F. Deligianni, X.-P. Hu, and G.-Z. Yang, "Extraction of visual features with eye tracking for saliency driven 2D/3D registration," *Image Vis. Comput.*, vol. 23, no. 11, pp. 999–1008, Oct. 2005.
[60] Y. Gao and J. A. Noble, "Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, Sep. 2017, pp. 305–313.
[61] J. Gómez-Poveda and E. Gaudioso, "Evaluation of temporal stability of eye tracking algorithms using webcams," *Expert Syst. Appl.*, vol. 64, pp. 69–83, Dec. 2016.
[62] S. A. Johansen, J. S. Agustin, H. Skovsgaard, J. P. Hansen, and M. Tall, "Low cost vs. high-end eye tracking for usability testing," in *Proc. CHI Extended Abstracts Human Factors Comput. Syst.*, Vancouver, BC, Canada, 2011, pp. 1177–1182.
[63] C. A. Hennessey and P. D. Lawrence, "Improving the accuracy and reliability of remote system-calibration-free eye-gaze tracking," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 7, pp. 1891–1900, Jul. 2009.
[64] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, 2017.
[65] R. Valenti and T. Gevers, "Accurate eye center location through invariant isocentric patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1785–1798, Sep. 2012.
[66] P. J. Grother, J. R. Matey, E. Tabassi, G. W. Quinn, and M. Chumakov, "IREX VI—Temporal stability of iris recognition accuracy," NIST Interagency/Internal, Gaithersburg, MD, USA, Rep. 7948, Jul. 2013.
[67] L. Best-Rowden and A. K. Jain, "Longitudinal study of automatic face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 148–162, Jan. 2018.
[68] K. Holmqvist *et al.*, *Eye-Tracking Data and Dependent Variables*, Lund Univ., Lund, Sweden, 2011.
[69] C. E. McCulloch, "Generalized linear mixed models," in *Proc. NSF CBMS Regional Conf. Series Probab. Stat.*, vol. 7, 2003, pp. 1–84.

Pierre Chatelain (S'15–M'16) received the B.S. degree in computer science and the B.S. degree in mathematics from the Université de Rennes 1, Rennes, France, in 2010, the M.S. degree in mathematics from the Ecole Normale Supérieure de Cachan, Cachan, France, in 2012, and the Ph.D. degree in signal processing and telecommunications from the Université de Rennes 1 and the Technical University of Munich, Munich, Germany, in 2016.

From 2012 to 2013, he was a Research Assistant with the Chair of Computer-Aided Medical Procedures, Technical University of Munich and with the Institute of Robotics and Mechatronics, German Aerospace Center, Cologne, Germany. He is currently a Postdoctoral Research Assistant with the University of Oxford, Oxford, U.K. His current research interests include medical robotics, medical image analysis, and machine learning.

Harshita Sharma received the B.Tech. degree in electronics and communication engineering from the Indira Gandhi Delhi Technical University of Women (formerly Indira Gandhi Institute of Technology, Guru Gobind Singh Indraprastha University), New Delhi, India, in 2010, the M.Tech. degree in electrical engineering from the Indian Institute of Technology Roorkee, Roorkee, India, in 2012, and the Ph.D. (Dr.-Ing.) degree in computer vision from Technical University Berlin, Berlin, Germany, in 2017.

From 2011 to 2012, she was an exchange Research Scholar with Technical University Berlin in collaboration with Charité University Hospital Berlin, Berlin. From 2012 to 2013, she was a Lecturer in electronics and communication engineering with the Jaypee Institute of Information Technology, Noida, India. She is currently a Postdoctoral Research Assistant with the University of Oxford, Oxford, U.K. Her current research interests include medical image analysis, computer vision, and machine learning.

Lior Drukker received the B.Med.Sc. and M.D. degrees from the Hebrew University of Jerusalem, Jerusalem, Israel, in 2005 and 2009, respectively.

From 2010 to 2017, he was a resident in Obstetrics and Gynecology, Shaare Zedek Medical Center, Jerusalem, and a Research Assistant with the Hebrew University of Jerusalem. He is currently a Clinical Research Fellow with the Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, U.K. His current research interests include obstetric ultrasound, automation of ultrasound, and computer vision image analysis.

Aris T. Papageorghiou received the M.B.Ch.B. degree from the University of Sheffield, Sheffield, U.K., in 1996 and the M.D. degree from the University of London, London, U.K., in 2009.

He has been a member of the Royal College of Obstetricians and Gynaecologists, London, U.K., since 1999. He divides his time between the University of Oxford (Clinical Research Director of the Oxford Maternal and Perinatal Health Institute), Oxford, U.K., and St. Georges Hospital, London (where he works clinically). His research is focused in the area of maternal, fetal and perinatal health using diverse methods, including basic science, clinical epidemiology, trials, knowledge transfer, and implementation science. A major interest is how artificial intelligence can improve pregnancy screening and diagnosis, particularly in low resource settings. For several years, he has researched collaboratively with biomedical engineers on these problems and co-supervised many joint graduate students.

Prof. Papageorghiou is a Trustee and the Honorary Secretary for the International Society of Ultrasound in Obstetrics and Gynaecology and an Executive Scientific Editor for the *British Journal of Obstetrics and Gynaecology*.

J. Alison Noble received the bachelor's degree in engineering science and the Ph.D. degree in computer vision from the University of Oxford, Oxford, U.K., in 1986 and 1989, respectively.

In 1995, she joined the faculty of the Department of Engineering Science, University of Oxford. She was a Research Scientist with the GE Corporate Research and Development Center, Schenectady, NY, USA, from 1989 to 1994. She was elected as the Technikos Professor of biomedical engineering in 2011. Her current research interests include ultrasound image analysis methodology with application to developed world healthcare settings and low and middle income countries.

Prof. Noble was a recipient of the OBE for services to science and engineering in 2013. She is a former President of the MICCAI Society from 2013 to 2016. She is a fellow of the Institution of Engineering Technology in 2001, the U.K. Royal Academy of Engineering in 2008, and the Royal Society in 2017.