Journal of Mathematical Imaging and Vision (2020) 62:585–605
https://doi.org/10.1007/s10851-020-00947-8

Screen-Light Decomposition Framework for Point-of-Gaze Estimation Using a Single Uncalibrated Camera and Multiple Light Sources

Carlos H. Morimoto[1] · Flávio L. Coutinho[1] · Dan W. Hansen[2]

Received: 31 August 2018 / Accepted: 28 January 2020 / Published online: 15 February 2020

Abstract

The use of a single uncalibrated camera is desirable for eye tracking to reduce the overall complexity and cost of the system. Quite often, at least one external light source is used to enhance image quality and generate a corneal reflection used as a reference point to estimate the point-of-gaze (PoG). Though the use of more than one light source has shown to enhance accuracy and robustness to head motion, it is unlikely that all corneal reflections appear in the eye images during natural eye movements. In this paper, we introduce the Screen-Light Decomposition (SLD) framework as a generalized model for PoG estimation using a single uncalibrated camera and a variable number of light sources. SLD synthesizes existing uncalibrated video-based eye trackers and can be used as a modeling tool to compare and design eye trackers. We have used the framework to design a novel eye-tracking technique, called SAGE, for single normalized space adaptive gaze estimation, that can gracefully degrade the gaze tracker performance when one or more corneal reflections are not detected, even during the calibration procedure. Results from an user experiment are presented to demonstrate its improved performance over other designs.

Keywords Adaptive · Estimation · Normalized · Space · eye · Tracking · Calibration · Framework

1 Introduction

Eye-tracking applications are rapidly growing beyond traditional neuro- and psychophysical laboratories and are now reaching toward more general purpose areas such as human factors research [12,39], marketing analysis [6,35], gaze-based interaction [29,31,37,42,46,53], and several other scenarios [26]. Unfortunately the applicability of eye trackers is still limited mostly due to calibration issues [12].

The basic problem of point-of-gaze (PoG) estimation is to compute the point being gazed by the user's eye(s) over a surface of interest or, alternatively, to compute the 3D line of sight. According to Hansen and Ji [17], the methods for PoG estimation can be classified into feature-based and appearance-based methods. They describe two types of feature-based gaze estimation approaches: the *interpolation-based* (regression-based) [1,13,23,34,38] and the *model-based* (geometric) [15,30,40,45,47]. In real-world application [2,14,27], i.e., in less constrained environments where near-infrared light cannot be used, appearance-based gaze estimation methods [28,48,49,51,52] have been used more frequently, though with lower accuracy than feature-based methods.

Despite the differences in device setups for distinct application scenarios, gaze estimation is performed mostly on 2D surfaces, based on interpolation, cross-ratio, or appearance-based methods. These methods are preferred because they can be implemented in uncalibrated setups requiring just a single uncalibrated camera, lowering setup complexity and cost when compared to the calibrated setups required by the model-based methods. Although these types of gaze estimation methods share this characteristic of not requiring calibrated setups, the working principle of appearance-based gaze estimation is considerably different from the other two. Appearance-based methods use the image of the eye region directly to estimate 2D PoG. In general, appearance-based methods do not necessarily use IR light sources, working in the visible spectrum, making them more suitable to be used with standard RGB cameras. The challenge of dealing with

✉ Carlos H. Morimoto
hitoshi@ime.usp.br

Flávio L. Coutinho
flcoutinho@usp.br

Dan W. Hansen
witzner@itu.dk

[1] University of São Paulo, São Paulo, Brazil

[2] IT University, Copenhagen, Denmark

the large variations in light conditions can be addressed during training of the model using a training dataset that covers a wide range of illumination conditions. Appearance-based methods, on the other hand, do not yet achieve the same accuracy level of current interpolation and model-based methods.

Model-based PoG estimation methods compute the 3D line of sight in space. They require knowledge of user specific parameters (e.g., the angle between the optical and visual axis of the eye, known as the κ-angle) and the geometry of hardware components (e.g., relative positions of lights and cameras) [17]. Much of the theory behind geometric models using fully calibrated setups has been formalized by Guestrin and Eizenman [15], covering setups with a variable number of light sources and cameras. Eye trackers with fully calibrated cameras rely on eye models to allow free head motion, but requires system recalibration when a component is changed. Therefore, they require rigid frames to prevent drifts between components.

Interpolation-based methods compute the PoG over a 2D surface of interest, typically the computer screen. Eye trackers with uncalibrated setups do not require hardware calibration or rigid frames, but they typically require user calibration. The use of multiple corneal reflections (CRs) improves robustness to the user's natural head motions, but might reduce the working volume of the system. For example, when the multiple light sources are placed on a large surface, such as a wide screen TV or computer monitor, not all CRs might be formed on the cornea for some eye orientations (e.g., looking at a corner). Reducing the number of required CRs simplifies the placement of the light sources such that all reflections are distinctly formed on the cornea surface and might provide a larger working volume, but still the light sources must be restricted to a relatively small planar surface (such as around a computer screen).

The main contributions of this paper are threefold:

- The introduction of the Screen-Light Decomposition (SLD) gaze estimation framework. The main objective of the SLD framework is to unify the theory of point-of-gaze estimation using single uncalibrated cameras with one or more light sources.
- Wide review and discussion of the literature about gaze estimation methods using one single uncalibrated camera. Inspired by the stratification theory introduced by Faugeras [11], we show that the framework can be used to generalize the methods described in the literature.
- The development and evaluation of the SAGE (single normalized space adaptive gaze estimation) technique that uses a single uncalibrated camera and four near-infrared (nIR) lights for PoG estimation. Though the use of multiple light sources has been suggested before, SAGE is adaptable, i.e., its performance gracefully degrades when CRs are not detected even during the user calibration procedure.

The next section introduces the SLD framework along with a review of the literature about gaze estimation in uncalibrated setups. We show in Sect. 3 how the framework can be used to design and evaluate adaptive gaze estimation methods, allowing us to compare and predict the performance of the methods based on the functions that are chosen to decompose the light and screen components. Section 4 describes the experiment designed to evaluate the performance of SAGE, comparing it to state-of-the-art methods such as homography and polynomial estimation. The results of the experiment are presented and discussed in Sect. 5, and Sect. 6 concludes the paper.

2 The Screen-Light Decomposition Framework

In this section, we systematically review existing techniques used for PoG estimation using a single uncalibrated camera and a variable number of external light sources and, at the same time, build the Screen-Light Decomposition (SLD) framework that synthesizes the basic structure of these techniques.

Though several parameters might influence the performance of the PoG estimation, such as illumination conditions, computer vision algorithms used for feature detection, and camera parameters (e.g., optical lens distortions, sensor resolution, frame rate, etc.), the focus of the SLD framework is on geometric issues that relates the optical system (camera and lights), the eye, and the gaze estimation surface (computer screen). It is a step toward the generalization of the geometry of PoG estimation methods available in the literature that use a single uncalibrated camera with one or more CRs.

Point-of-gaze estimation using a single uncalibrated camera typically seeks a function S that maps eye features (such as the center of the pupil or iris) extracted from images of the eye to corresponding coordinates on the computer screen [17]:

$$\mathbf{g} = S \cdot \mathbf{p} \tag{1}$$

where $\cdot$ is a composition. We adopt this notation to reduce clutter created by a large number of parenthesis.

Figure 1 illustrates the specific case where $\mathbf{p}$ is the position of the center of the pupil in the image and $\mathbf{g}$ is the position of the point-of-gaze on the computer screen. (Note that $\mathbf{p}$ could be other eye features that reflects the eye orientation, such as the center of the iris or limbus.)

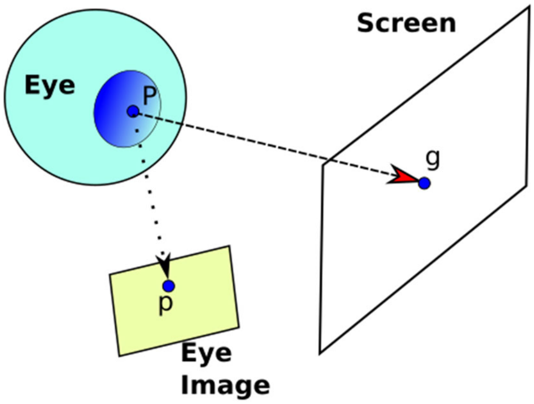


Fig. 1 Point-of-gaze estimation using a single uncalibrated camera

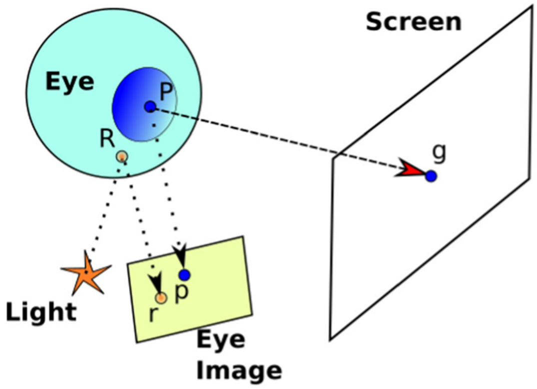


Fig. 2 The PCR technique uses a single CR to estimate the point-of-gaze **g**

The mapping function S depends on both geometric (e.g., location of the eye, camera, and screen) and personal parameters (e.g., the κ-angle). The coefficients of S are computed through user calibration by minimizing the errors of the coefficients with respect to known gaze positions [43].

Simple mapping functions, e.g., based only on the center of the pupil in the eye image, work well when the head is fixed using a chin rest or bite bar or when used in a head mounted eye tracker, but are not reliable when the head is not restrained. Fortunately, S can be made more robust to head motion if the pupil position is considered relative to a feature that moves with the head, such as some facial feature or CR created by an external light source.

Figure 2 illustrates the commonly used pupil–corneal reflection (PCR) technique often employed for PoG estimation [3]. It uses a single CR to ensure robustness to small head pose changes. The use of an external light source also improves illumination conditions for capturing the eye images. PCR requires one user calibration per session and is prone to drifts mostly due to head motion from the calibration position. Let $\mathbf{r}$ be the center of the CR in the eye image, so the PoG $\mathbf{g}$ can now be computed as:

$$\mathbf{g} = S \cdot (\mathbf{p} - \mathbf{r}) \tag{2a}$$

齐次坐标系

$$\mathbf{g} = S \cdot L \cdot \mathbf{p} \tag{2b}$$

距离更远 w更大
(x,y,w)

Observe that the pupil translation by $\mathbf{r}$ can also be represented by a composition with L. For example, by presenting p as a 2D point in homogeneous coordinates, L represents a translation (matrix) by $-r$.

Polynomials have been frequently used to interpolate S [4,38]. Morimoto and Mimica [38], for example, used a full second-order polynomial and a single CR. Considering the pupil–corneal reflection vector $\mathbf{v} = \mathbf{p} - \mathbf{r} = (x, y)$, the mapping function S in (2) can be expanded as:

$$\begin{aligned} g_x &= a_0 x^2 + a_1 y^2 + a_2 xy + a_3 x + a_4 y + a_5 \\ g_y &= a_6 x^2 + a_7 y^2 + a_8 xy + a_9 x + a_{10} y + a_{11} \end{aligned} \tag{3}$$

and the coefficients (a_0 to a_{11}) of (3) are estimated through a calibration procedure. Cerrolaza et al. [4] have studied the behavior of other polynomial mapping functions in gaze estimation.

Only 6 calibration points are required to estimate the coefficients in (3), but more points can be used to improve accuracy by solving an over determined system of equations. For example, using 9 calibration points PCR can achieve an average accuracy of about 1^o of visual angle over the area of a 17" monitor when the head is kept at the calibration position [38]. However, as the head moves away from the calibration position, the accuracy is affected more by depth changes of the head relative to the computer screen than to translations parallel to the screen, suggesting that the PCR technique tolerates small side-to-side head motions and it is more sensitive for back-and-forth motions. 可以左右动

The PCR performance can be improved by using more than one external light source [3,21,32]. For example, using the CRs from two light sources, Cerrolaza et al. [3] have shown that the distance between two CRs can be used to normalize the PCR vector (illustrated in Fig. 3), so the accuracy of the technique becomes more robust to depth changes of the head position. Let $\mathbf{r}_1$ and $\mathbf{r}_2$ be the centers of the two CRs in the eye image, and s the norm of the vector ($\mathbf{r}_2$-$\mathbf{r}_1$). Then the point-of-gaze $\mathbf{g}$ can be computed as:

$$\mathbf{v} = \mathbf{p} - (\mathbf{r}_1 + \mathbf{r}_2)/2 \tag{4a}$$

$$\mathbf{g} = S \cdot (\mathbf{v}/s) \tag{4b}$$

$$\mathbf{g} = S \cdot L \cdot \mathbf{p} \tag{4c}$$

where $\mathbf{v}$ in (4a) is the new PCR vector that is dynamically scaled by s in (4b). (4c) shows that we can still represent translation and scaling by a composition L.

平移

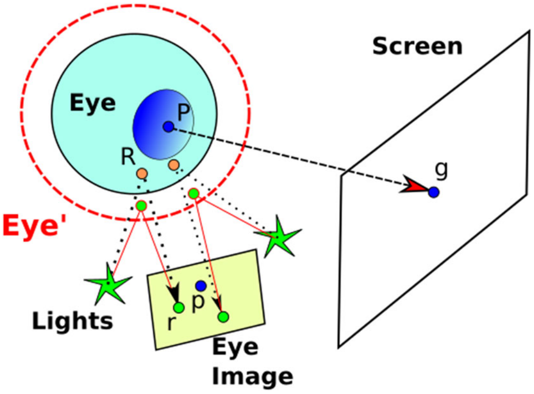


Fig. 3 PCR using two CRs, the image distance between reflections can be used to compensate small depth changes. When the eye gets closer to the eye camera (shown as the dotted circle Eye'), the distances between all image features increase

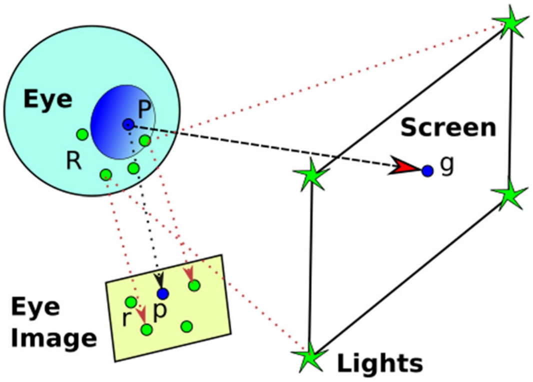


Fig. 4 The cross-ratio method uses four light sources placed around the computer screen to estimate the point-of-gaze

Loosely speaking, one could say that by tracking one CR it is possible to compensate for head translation, and by using 2 CRs, we can also compensate for scale changes due to depth movements. From these two observations, we begin to introduce the Screen-Light Decomposition framework. It is clear from (2) and (4) that the CRs (the light component) can be used to transform the eye features such as the pupil **p** using a transformation L to compensate for head motion, before the mapping S to the screen is applied.

2.1 Projective Transformation Methods

Consider an eye-tracking system as shown in Fig. 4, where four light sources are placed on a plane (e.g., at the corners of a computer screen). Assuming the cornea is a flat surface, the reflections of the lights on the cornea, when seen by the camera, are formed by a planar projective transformation (homography). Considering that the center of the pupil defines the gaze direction and that it lies on the flat cornea surface, the image of the pupil can be back projected to the surface containing the light sources using the cross-ratio, which is invariant under projective transformations [20].

This method was first introduced by Yoo and Chung [50], and we consider this as a constructive (step-by-step) solution. Unfortunately, in practice, the cross-ratio method performs poorly, with accuracy about 3^o to 5^o. Several extensions have been proposed to improve the accuracy of the method [1,5,7–10,16,25,50], and they all resort to some calibration, at least a one-time calibration per user to measure the κ-angle (the offset between the optical and visual axis of the eye). The main sources of error, as described by Kang et al. [24], are that the CRs are not coplanar, that the pupil is not coplanar with the cornea surface, and that the pupil center defines the optical and not the visual axis of the eye. For most extensions, such as those presented in [7,8,16,50], because the estimated parameters depend on the calibration position, the new methods loose their invariance to head motion (despite being considerably more robust than PCR, for example).

The analytical transformations of coplanar points, for which the cross-ratio is invariant, are homographies [20]. Hansen et al. [16] introduced the homography normalization method that models the back projection of the pupil center into the PoG by means of two mapping functions: a homography and second mapping function that can be a homography or any other appropriate function such as polynomial, Gaussian process, or a geometric model. The first homography removes the perspective distortion from the image of the 4 CRs (caused by head movements) and maps the pupil center to the so-called normalized space (a square of unitary size). The second mapping transforms the pupil in the normalized space to the screen space implicitly accounting for personal specific parameters (such as the angle between the optical and visual axes aka angle-κ [17]). Because such hidden parameters depend on a calibration procedure and the position it takes place, the accuracy of the method varies with head movements, with depth changes accounting for a significant decay of gaze estimation accuracy.

Using the SLD framework, the point-of-gaze **g** computed by the homography normalization method can be described as:

$$\mathbf{g} = S \cdot L \cdot \mathbf{p} \tag{5}$$

where L is the homography that transforms the image of the pupil to its position on the "flat" cornea surface and S is the mapping that transforms the pupil to screen coordinates.

To understand the homography normalization method using the SLD framework, recall from (2) that the CRs can be used to compensate for head translations in the PCR technique, while from (4) that head depth changes can also be compensated when the transformation L is a scale factor computed from the size variation of two CRs, using the size at the calibration position as reference. Homography normalization uses the transformation L to compensate for perspective transformations through the normalized space. Assuming that the light sources are placed at the corners of a computer monitor, forming a rectangle, the relationship between the normalized space and the screen space is simply given by scaling the x and y coordinates appropriately. However, because of the reasons pointed by Kang et al. [24], a simple scale transformation cannot be directly used to map the normalized pupil $L \cdot \mathbf{p}$ to the PoG $\mathbf{g}$ in screen space, requiring a user calibration to determine S.

单应性

The cross-ratio-based method introduced by Yoo and Chung [50] uses a fifth light source placed near the camera center. The CR generated by this fifth source is used as a virtual plane where the remaining four CRs are projected. This virtual plane works in a similar way to the homography normalization L. Recently, Cheng et al. [5] have proposed a variation of this method by creating the virtual plane on the center of the pupil instead, to reduce the error caused by the non-coplanarity of the pupil center with the four CRs.

Coutinho and Morimoto [9,10] introduced the planarization method that maintains the cross-ratio invariance property to head motions. The method explicitly models the angle-κ and the non-coplanarity of the pupil with the cornea surface, correcting the major sources of error of the original method by creating a virtual corneal plane and projecting (planarizing) the relevant eye features to this plane before applying the cross-ratio. The planarization method requires a one-time calibration per user to measure the personal eye parameters used in the model. Once the cross-ratio main error sources are corrected, the planarization method remains accurate within a large working volume.

Considering the SLD framework, we can say that the planarization method uses a calibration procedure to compute personal parameters to adjust an eye model where the cross-ratio method can be applied. Though a 3rd transformation could be integrated to the framework to model personal parameters such as the κ-angle, we will consider that such parameters are embedded in S (since it includes user calibration). Therefore, this simple framework is well suited to describe all of the methods that estimate gaze using a single uncalibrated camera.

We have shown in this section that the SLD framework can be used to interpret how different eye-tracking methods work. These methods use a single uncalibrated camera and a variable number (one, two, or four) of lights. An interesting first consequence of building the SLD framework is the lack of PoG estimation systems in the literature using affine models with three external lights. In the next section, we show how such a system could be build and propose the SAGE technique.

3 Adaptive Normalization

The Screen-Light Decomposition framework reveals that gaze estimation methods using one, two, and four light sources have been suggested in the literature. Each combination of light sources enables different head motion compensation models for L, corresponding to translation, similarity, and homography transformations. Therefore, affine transformations have been mostly ignored.

In our search for affine methods for gaze estimation, we only found the work of Ma et al. [33]. It is an extension of the homography normalization method that uses four light sources and tolerates up to two missing CRs. This method can also be interpreted using the SLD framework.

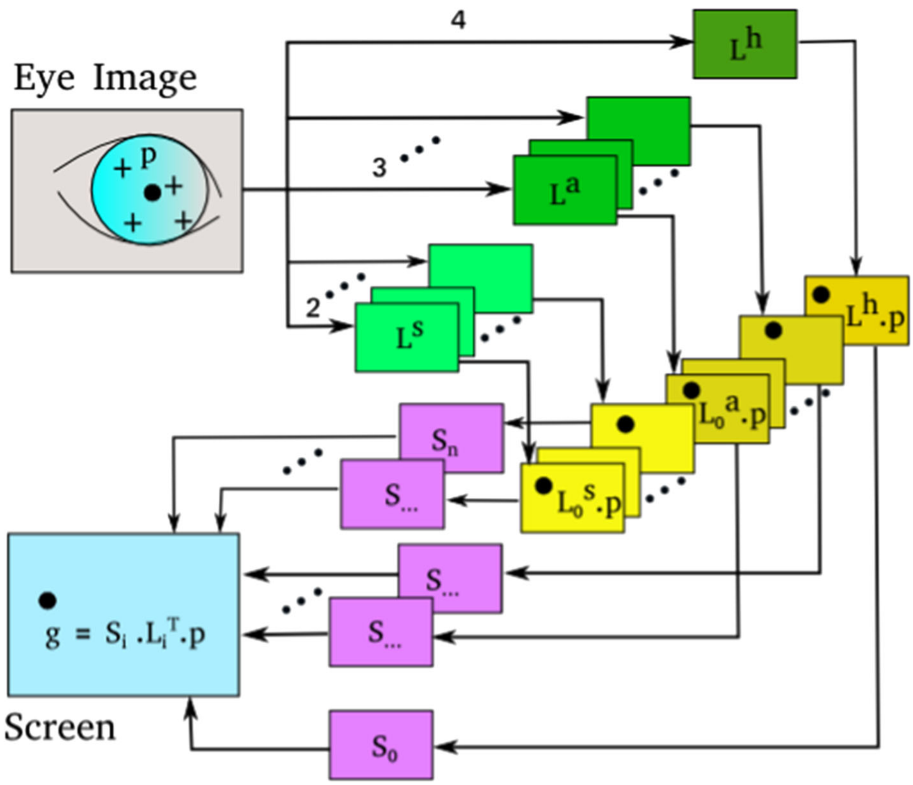
During the calibration procedure, all 4 CRs must be properly formed and detected to compute the following L transformations: one homography using 4 CRs, 4 affine using triples of CRs, and 4 similarity using consecutive pairs of CRs. In the remaining of this paper, we will use the notation L^h, L^a, and L^s when appropriate to explicitly denote the homography (h), affine (a), and similarity (s) transformations. For each one of these transformations, a corresponding transformation S is computed and stored, as shown in Fig. 5, where L_i^T denotes the type of transformation T used for normalization (i.e., similarity, affine, or homography), and the index i denotes a particular combination of CRs (i.e., a specific pair of reflections for similarity or a particular triplet for affine).

During gaze estimation (after calibration), the set of detected CRs defines which L_i^T and its associated S_i are used to compute the PoG g. Therefore, $L_i^T.p$ can be considered as a set of normalized pupils. Ma et al. associate a particular second-order polynomial S_i to each $L_i^T.p$.

It is natural to think about other possible adaptive variations of the homography normalization method. Because the framework decouples the transformation L to the normalized space and S to the screen, it is possible to choose any appropriate transformation L based on the number of available CRs and a (possibly different types of) transformation S based on the desired number of user calibration targets, independently. In fact, we show next that only a single transformation S could be used instead of several S_i.

First, to compute L, assume the correspondences of the visible CRs to the corners of the normalized space are known. Methods for computing such correspondences are presented in [18,22,33]. With one missing CR, L can be estimated using the remaining three CRs as an affine mapping L^a to the normalized space. In rare situations when two CRs are missing, the remaining two CRs can be used to map pupil positions to the normalized space using a similarity transformation L^s. A similarity transformation (4 degrees of freedom) is able to model the translation, rotation, and uniform scale components of the perspective distortion observed in the images of

Fig. 5 Adaptive normalization using multiple normalized spaces $L_i^T.p$, where T defines the type of transformation (either homography, affine, or similarity) and i defines the particular transformation of that type (for 4 light sources, for example, there are 4 possible affine and 4 similarity transformations using consecutive pairs of CRs)



the CRs due to head movements. In extreme situations when all CRs but one are missing, the normalization transformation becomes a simple translation L^t. It is, therefore, related to the PCR method since a translation by a CR **r** can be represented by L^t.

Using similar arguments, the calibration function S can be adjusted by varying the number of calibration targets. When 4 or more calibration targets are used, S can be defined by a homography S^h. If 3 calibration targets are used, then an affine mapping S^a can be employed. In situations where the number of calibration targets needs to be further reduced, a similarity mapping S^s can be defined by just 2 calibration targets. Although a reduction from 4 to 2 calibration targets might not seem very significant, the capacity to deal with a small number of calibration targets may be of special interest in some cases, particularly when the user is noncooperative and/or very hard to track (such as small children and some people with disabilities).

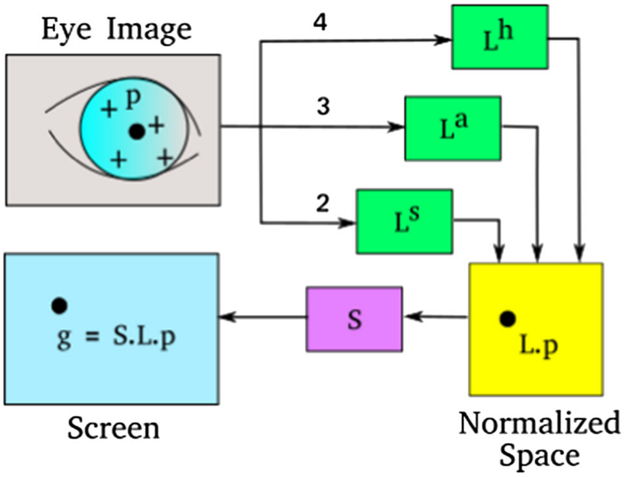


Fig. 6 The single normalized space adaptive gaze estimation technique uses a single transformation to the screen S and applies the best L^T transformation available based on the number of detected CRs (4, 3, or 2)

Figure 6 shows a diagram of the single normalized space adaptive gaze estimation technique (SAGE). For comparison purposes, we will denote the system described by Ma et al. [33] as MAGE for multiple normalized space adaptive gaze estimation. The single normalized space to screen space transformation S is estimated based on the normalized pupil centers associated with each calibration target. Observe that each normalized pupil center may have been obtained using a different L_i^T (defined by the set of available CRs), ensuring robustness to missing CRs also during the calibration procedure. The working principle of SAGE is summarized in Algorithms 1 and 2. Algorithm 1 describes how SAGE works during the calibration procedure, and Algorithm 2 describes how PoG estimation is performed by SAGE after calibration.

An important difference from SAGE to MAGE, and to homography normalization [16], is that the normalized space for SAGE must have an aspect ratio equal to the aspect ratio

ALGORITHM 1: SAGE calibration

```
Input: the set of calibration targets $\mathbf{T}$; the $min$ minimum
        acceptable number of calibration targets to be used to
        compute $S$.
Output: the $S$ transformation.

// This set will be used to store tuples,
   each one composed of a
// normalized pupil center and the
   associated calibration target.
$\mathbf{C} \leftarrow \emptyset$;

for $\mathbf{t_k} \in \mathbf{T}$ do
    // $\mathbf{t_k}$ is displayed on screen and then
       user is supposed to gaze at it.
    $frame \leftarrow acquire\_from\_camera()$;
    $\mathbf{p_k} \leftarrow detect\_pupil\_center(frame)$;
    $\mathbf{CR_k} \leftarrow detect\_CRs(frame)$; // the labelled set
       of detected CRs.

    if $\mathbf{p_k}$ is valid and $|\mathbf{CR_k}| \geq 2$ then

        if $|\mathbf{CR_k}| = 4$ then
            $L_k \leftarrow calc\_homography(\mathbf{CR_k})$;
        else if $|\mathbf{CR_k}| = 3$ then
            $L_k \leftarrow calc\_affine(\mathbf{CR_k})$;
        else // $|\mathbf{CR_k}| = 2$
            $L_k \leftarrow calc\_similarity(\mathbf{CR_k})$;

        $\mathbf{p'_k} \leftarrow L_k \cdot \mathbf{p_k}$;
        $\mathbf{C} \leftarrow \mathbf{C} \cup \{(\mathbf{p'_k}, \mathbf{t_k})\}$;
    end
end

if $|\mathbf{C}| \geq min$ then
    compute $S$ from tuples in $\mathbf{C}$; // the type of $S$ is
       choosen based on $|\mathbf{C}|$
    return $S$;
else
    return NULL; // calibration failed
end
```

ALGORITHM 2: SAGE PoG estimation

```
Input: the $S$ transformation computed during the calibration
        procedure.

while running do
    $frame \leftarrow acquire\_from\_camera()$;
    $\mathbf{p} \leftarrow detect\_pupil\_center(frame)$;
    $\mathbf{CR} \leftarrow detect\_CRs(frame)$;

    if $\mathbf{p}$ is valid and $|\mathbf{CR}| \geq 2$ then

        if $|\mathbf{CR}| = 4$ then
            $L \leftarrow calc\_homography(\mathbf{CR})$;
        else if $|\mathbf{CR}| = 3$ then
            $L \leftarrow calc\_affine(\mathbf{CR})$;
        else // $|\mathbf{CR}| = 2$
            $L \leftarrow calc\_similarity(\mathbf{CR})$;

        $\mathbf{g} \leftarrow S \cdot L \cdot \mathbf{p}$;
        broadcast $\mathbf{g}$ to registered clients;
    else
        // PoG estimation failed for current
           frame.
    end
end
```

of the quadrilateral formed by the light sources. Let w and h be the width and height of such quadrilateral. We define our normalized space as a rectangle with width equal to w/h and height equal to 1.0. We also assume that the camera sensor preserves the aspect ratio of the scene it captures. This requirements are necessary because similarity transformations just model translation, rotation, and uniform scaling. Therefore, pupil mappings from the image space to the normalized space using a similarity for L, or from the normalized space to screen space using a similarity for S, will be more accurate if all spaces share the same aspect ratio. For affine or homography transformations, the correct aspect ratio is not an issue since they can model non-uniform scaling.

Remember that the main purpose of introducing a normalized space is to remove the effects of head motion. If the pupil center can be normalized by different L_i^T transformations, an associated S_i transformation can be optimized to minimize the gaze estimation error, similar to [33] as shown in Fig. 5. A limitation of MAGE is that the set of CRs used to compute L_i^T must be visible for all calibration targets required to compute the associated S_i. To deal with this limitation, Ma et al. [33] required all 4 CRs to be visible during the whole calibration procedure.

Because SAGE uses a single transformation S, we might not expect to achieve the same level of accuracy that is observed when multiple S_i are used. Nonetheless, we achieve greater flexibility and robustness to missing CRs during calibration since the normalized pupils required to compute S can be obtained by distinct L_i^T. Also, the number of calibration targets can be dynamically adjusted based on the number of calibration samples successfully detected. (A successfully detected sample must include the pupil center and a minimum number of CRs, e.g., 2 CRs.) For instance, if SAGE is being calibrated using 4 targets, and eye features for one of the targets are not properly detected, it may still be possible to calibrate the system using 3 calibration samples, which may avoid repetition of the whole calibration procedure.

Facilitating the calibration procedure is highly desirable because it allows the use of larger displays and wider tracking areas. It might also increase the number of possible participants, such as people with glasses, and less cooperative users such as children and people with certain disabilities. For example, Fig. 7 shows a few cases where, even with careful placing of the camera, illuminators, and computer screen,

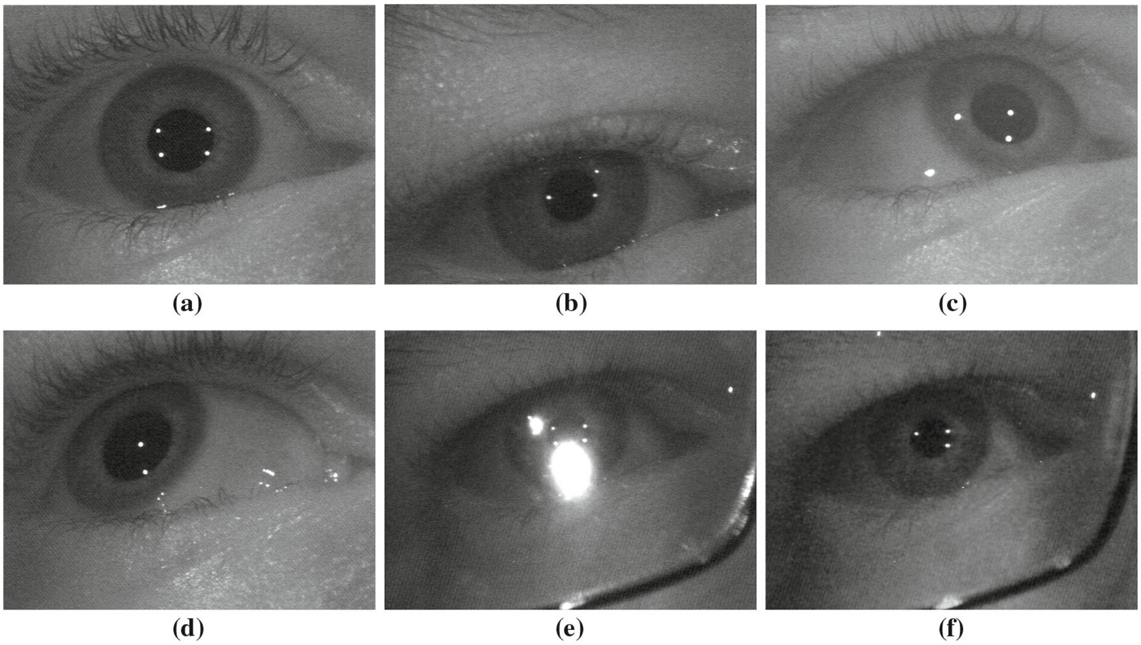


Fig. 7 Eye image samples illustrating scenarios that show the advantage of SAGE's flexible calibration (i.e., robustness to missing CRs during the calibration procedure)

operators would find MAGE very hard to use, while SAGE would be easily applicable.

The images shown in Fig. 7 were captured with the same camera used in the experiments described in Sect. 4, placed at the bottom of a 24" monitor with IR LEDs placed at its corners. The user's head was about 55 cm from the center of the screen. Figure 7a shows that the four CRs can be seen when the user's eyes are wide opened. However, for a more natural and comfortable opening position shown in Fig. 7b, some of the CRs (particularly the top ones) might be missing due to occlusion by eyelids/eyelashes when the user looks at the bottom region of the screen. Figure 7c, d shows that CRs might be missing when the user looks at positions close to the corners of the screen. In Fig. 7c, one CR is missing (it is actually being formed on the sclera), while in Fig. 7d two CRs are missing. Figure 7e shows a user wearing glasses where reflections on the surface of the glasses occlude the pupil and cornea. SAGE allows the system operator to turn off the interfering light source (or cover it), resulting in the image in Fig. 7f, and estimate the gaze (or continue the calibration procedure) using three CRs.

4 Experimental Design

We have conducted an experiment to assess the robustness of SAGE to missing CRs and fewer number of calibration targets and, also, compare its performance with MAGE. We also investigate how these methods are affected by head movements. The same dataset used by Coutinho and Morimoto in [10] was used for this experiment. The dataset is available at http://latin.ime.usp.br/datasets.

The dataset contains eye images from 7 subjects (i.e., 14 eyes) with normal or corrected to normal vision, captured with the use of a chin rest placed at the 9 distinct positions as depicted in Fig. 8, marked P_0 to P_8. At each position, users were asked to fixate their gaze at 49 screen targets evenly distributed on a $7{\times}7$ grid. Targets were shown one at a time on the computer screen, and 20 image samples were acquired for the left and right eyes simultaneously. This procedure was repeated for the 9 head positions. Not all of the 20 image samples were used due to eye feature detection failure though. However, special care was taken to ensure that, for each screen target, at least one image sample contained valid features (pupil and all 4 CRs properly detected). A more detailed description about how data were collected can be found in [10].

Note that the monitor used in the creation of the dataset used had a 17" screen (different from the setup used to capture image samples presented in Fig. 7 that used a 24" screen). The use of a smaller screen assured that all 4 CRs were properly detected for at least one image sample for every screen target. Observe that it would not have been possible to compare SAGE against MAGE if a larger screen had been used due to the CR formation issues illustrated in Fig. 7.

A 30 Hz differential lighting eye tracker [36], originally developed as a single light source PCR eye tracker, was used to record images of resolution 640 × 480 pixels. The eye

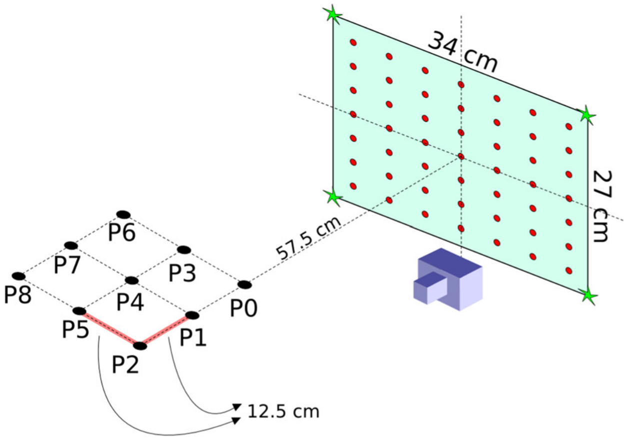


Fig. 8 Experimental setup (top view)

tracker is composed of an interlaced CCD camera with long lens appropriate for a desktop remote setup. A 17¨ monitor with one near-infrared LED (HSDL 4220—875 nm) at each corner was used to display the visual targets. The pupils were segmented using the difference between the bright and dark pupil images. The CRs were detected by thresholding bright pixels from the dark pupil image and selecting the best candidates according to their size, shape, and relative position to the pupil center. Our current implementation of SAGE was developed in C++ and OpenCV, and it is capable of processing frames at more than 100 fps on a single-core modern notebook. This performance is relatively easy to achieve since, once eye features are detected, the PoG is computed from simple matrix multiplications. Therefore, the real-time capability of our implementation relies more on the image processing algorithms used for feature detection.

Let $\epsilon_{ij}^{L^T}$ be the gaze estimation error (in degrees of visual angle) measured for screen target i, sample j and using a transformation L^T. The mean gaze estimation error for L^T, computed at a specific head position, is defined using equation (6), where N corresponds to the total number of available samples.

$$E^{L^T} = \frac{1}{N} \sum_i \sum_j \epsilon_{ij}^{L^T} \tag{6}$$

When all 4 CRs are available, just a single homography transformation L^h is defined, but there are several subsets of CRs that can determine an affine or similarity transformation. More specifically, there are 4 distinct subsets of CRs that can define affine transformations, and 6 subsets that define similarities. Therefore, for a given transformation type, more than one error measure E^{L^T} can be computed (homography being the exception).

In our evaluation, instead of considering each possible affine and similarity transformation that can be obtained from distinct subsets of CRs, and computing E^{L^T} for each of them, we condense the error measurements by the number of CRs used. To do this, a heuristic rule was applied to both SAGE and MAGE. CRs farther from the pupil center in the eye image are, due to the corneal curvature, more likely to disappear because they are formed closer to the limbus (the cornea boundary—see Fig. 7c, d for examples). Therefore, only the two or three CRs closest to the pupil center were used to estimate a similarity transformation L^s or an affine transformation L^a, respectively. Observe that, using this heuristic, the specific transformation of a certain type used in the evaluation was, in fact, determined by the screen target being gazed.

5 Results and Discussion

To understand how these methods were calibrated, consider R_7 to be the outer most rectangle of the 7 × 7 grid of Fig. 8, consisted of 24 points, and let C_1 be the point at the center of the grid (center of the screen). For 9-point calibration, the 4 corners and the 4 midpoints of the edges of R_7 and C_1 were used to estimate the coefficients of a second-order polynomial (see equation 3) using least squares. For 4-point (or less) calibration, a corresponding number of corners of R_7 were used. The homography and affine transformations are computed using linear regression methods available in OpenCV such as the direct linear transform (DLT) algorithm from Hartley and Zissermann [19].

The methods were calibrated at position P_0, and the grand mean gaze estimation error and standard deviation were computed at each of the 9 positions using the 49 targets and 14 eyes.

5.1 MAGE × SAGE

MAGE and SAGE are compared by how missing CRs affect accuracy as the head gets farther from the calibration position.

Figure 9 presents the evaluation results for both SAGE and MAGE at each of the 9 head positions. Observe that P_0, the calibration position, is shown at the top right corner to mimic the actual physical setup used in the experiment shown in Fig. 8. Therefore, the graphs at P_1 and P_2 show how performance decays with larger distances from the monitor, while P_3 and P_6 show results for different lateral translations from the central position. Considering that the error is somewhat symmetrical when the head moves left and right, for the experimental setup shown in Fig. 8, we can investigate the systems' performance for a ±25 cm lateral translation. The use of P_0 as the calibration position also follows the procedure described in [10] that uses the same dataset, and

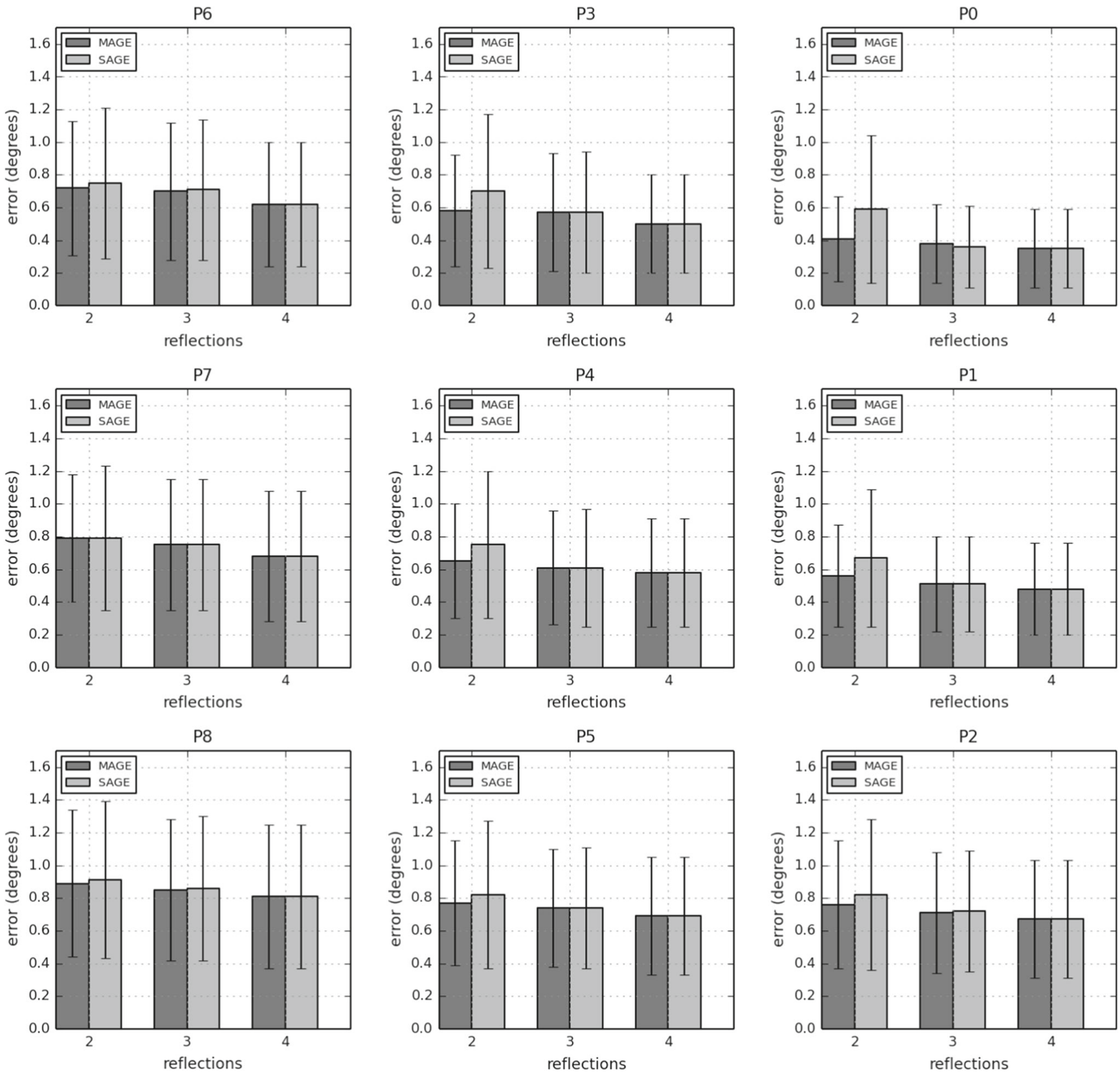


Fig. 9 Performance evaluation of MAGE (based on [33]) and SAGE. The graph positions P_0 to P_8 match the physical head positions shown in Fig. 8. Both systems were calibrated at P_0 (top right) using 9 calibration targets to fit a second-order polynomial. The vertical bars show the grand mean and standard deviation of the gaze estimation error computed for the 49 grid points and 14 eyes at each head location. Each graph shows how the error varies with the number of available CRs, from 2 to 4 CRs

facilitates the comparison of results presented later in this paper.

Each graph in Fig. 9 shows how the grand mean gaze estimation error is affected by the number of available CRs. In this experiment, user calibration was performed with 9 points, second-order polynomial functions were used for S, for both MAGE and SAGE. (Remember that for MAGE several S_i^T were obtained as the result of the calibration procedure, while for SAGE just a single S was estimated.) For MAGE, it was assumed that all 4 CRs were visible during calibration, even when fewer CRs were considered for PoG estimation. For SAGE, the number of CRs considered during calibration was the same that was considered for PoG estimation. For instance, to evaluate the PoG estimation error of SAGE in a scenario where 3 CRs were available, it was also assumed that 3 CRs were available during calibration. Note that the set of 3 CRs considered to define L^a varied depending on the screen target, according to the heuristic previously described.

The best results were obtained when all 4 CRs were available, while the worst results (larger error) were obtained when 2 CRs were used. Observe that the performances of SAGE and MAGE are the same when 4 CRs are available. When one CR is missing, the performance of SAGE is very similar to MAGE, which may be a bit surprising since MAGE uses multiple normalized spaces where each mapping S_i^T is specifically optimized for each possible L_i^T. When two CRs are missing, the performance of SAGE is slightly below MAGE due to SAGE's more flexible calibration conditions.

Considering each head position separately, the mean error difference between the two methods never exceeds 0.2^o when two CRs are missing, and it is much lower for only one missing CR. MAGE presents slightly better robustness to missing CRs. For any position, its accuracy from the 4 CRs condition never decays more than 0.1^o when two CRs are missing, while the accuracy of SAGE decays up to 0.24^o. It is interesting to notice that the maximum decay occurs at the calibration position, implying that head movements must have a larger impact on accuracy than the missing CRs.

5.2 Varying the Number of Calibration Targets Using SAGE

Schnipke and Todd [44] describe how hard it can be to use an eye tracker. To set up their usability experiment of a popular software application, an operator with 1-year experience with the eye tracker was not able to make the system track 10 out of 16 subjects. Though eye trackers have improved, as pointed out by Nyström et al. [41], few studies have focused on practical issues about collecting eye movement data. One main motivation behind SAGE's design is to create robust and accurate eye trackers that are easier to use by end users and operators. Besides being more flexible than MAGE (and other single camera eye trackers) for the practical situations shown in Fig. 7, SAGE relies on fewer calibration points, allowing the system to be used by noncooperative users such as children and people with certain disabilities.

Unlike MAGE, SAGE can dynamically adapt to the number of detected CRs even during the calibration procedure. It can also dynamically adapt to the number calibration points if eye feature detection fails (pupil not detected or fewer than 2 CRs) for some calibration targets. In this section, we investigate how the performance of SAGE is influenced by the number of calibration points and the number of available CRs. Figures 10, 11, and 12 show how the accuracy of SAGE varies when 2, 3, 4, and 9 calibration targets are used. Similar to Fig. 9, the system is calibrated at position P_0 (the top right corner of the figure) to mimic the same physical setup shown in Fig. 8. The results were computed using the corners of R_7 as calibration points when 4 or less targets are used for calibration, as described in Sect. 3. For 9 calibration targets, the figures show the results of two different S functions: a second-order polynomial (shown as the 9* column) and a homography transformation (shown as the 9 column). For fewer calibration targets (4, 3, and 2), only linear transformations were used (homography, affine, and similarity, respectively).

Figure 10 shows the performance of the different S calibration functions when all 4 CRs are assumed to be visible at all times. Thus, L^T is a homography for all graphs shown in this figure. Observe that the polynomial function with 9 calibration points (column 9*) consistently outperforms the other methods probably because it is able to compensate for some nonlinearities. Nonetheless, its performance is very similar to the linear methods using 3, 4, and 9 calibration targets (about $0.4^o \pm 0.2^o$ at the calibration position P_0). This suggests that a simple affine model with only 3 calibration points might in practice be an appropriate solution to achieve reasonable eye data quality with simple and fast calibration. The average error using 2 calibration points with 4 CRs, though much higher than the other methods at the calibration position P_0, is below 1^o for most positions and approximately $0.75^o \pm 0.5^o$ at P_0, which is still better than the performance of single CR methods reported in the literature [38] of about 1^o. Also note that the accuracy decay due to head movements is observed independently of the number of calibration targets.

When all 4 CRs are available, the results in columns 4 and 9 are equivalent to the homography normalization method using 4 and 9 calibration points [16], while column 9* is equivalent to SAGE and MAGE results shown in Fig. 9.

Figure 11 shows the performance of the different S calibration functions when only 3 CRs are assumed to be visible (i.e., L^T is an affine transformation). Observe that the error distribution is similar to Fig. 10, which suggests that affine transformations in practice produce as good results as homographies.

Figure 12 shows the performance of the different calibration functions of S when only 2 CRs are used (i.e., L^T is a similarity transformation). The average error at the calibration position is higher at P_0 (about $0.6^o \pm 0.4^o$). Therefore, when only two CRs are available, the overall performance of SAGE is compromised, but still better than methods that use a single CR [38].

Compared to other algorithms available in the literature, the column 9* using the polynomial calibration function is similar to the 2 CRs methods described by Cerrolaza et al. in [3,4].

5.3 Error Distributions

To better understand the behavior of the gaze estimation error under these conditions, we present in Fig. 13 heatmaps that show how the error is distributed across the screen at the calibration position P_0. Darker points correspond to higher errors. Each gaze estimation error shown in the heatmap corresponds to the grand mean for each test target, considering

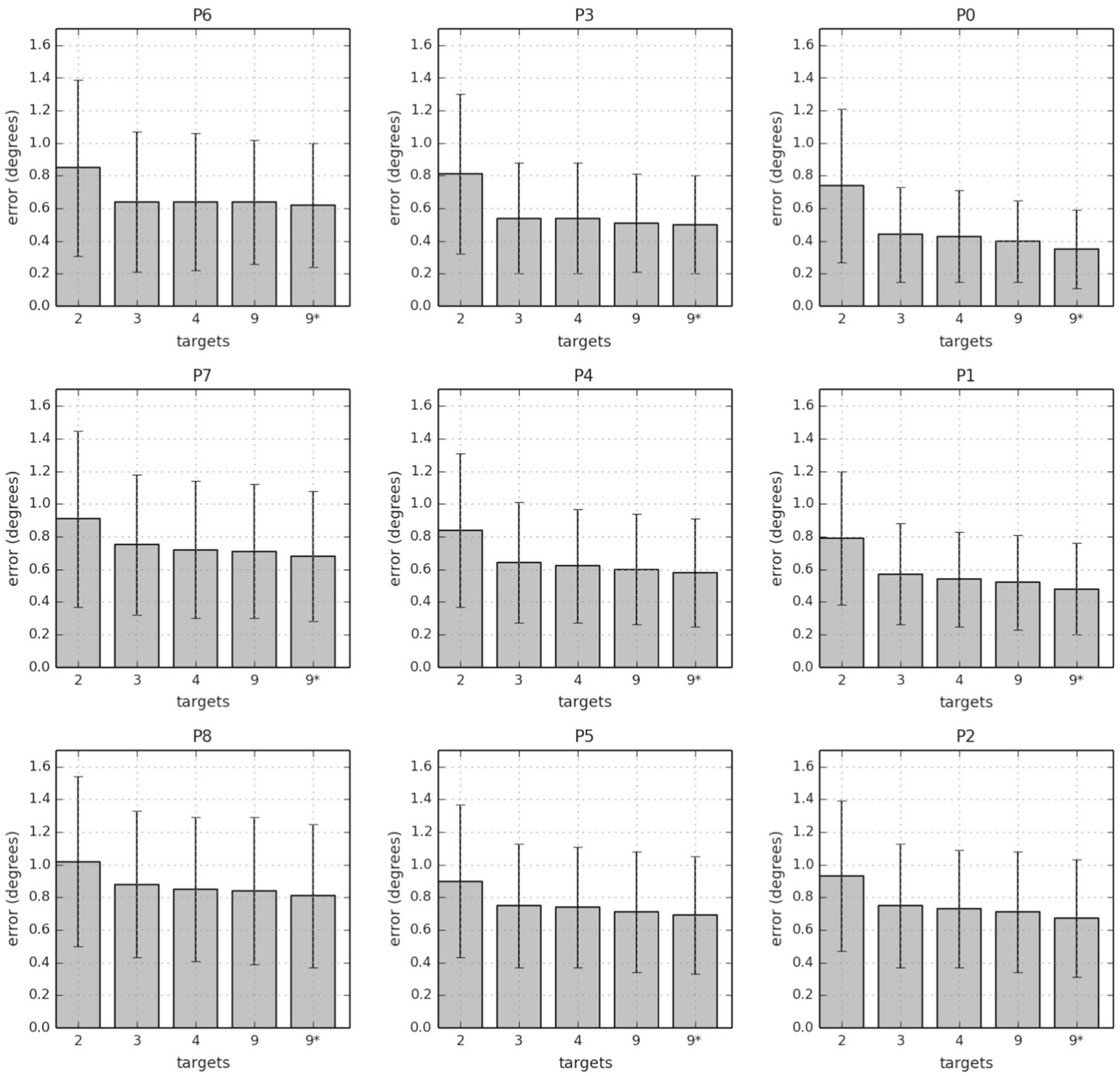


Fig. 10 Performance of SAGE when all 4 CRs are visible, using a variable number of calibration targets. The graph positions P_0 to P_8 match the physical head positions shown in Fig. 8. The vertical bars show the grand mean and standard deviation of the gaze estimation error. The horizontal axis shows the number of calibration targets used. Columns 2 to 9 use linear calibration functions and 9* a second-order polynomial

all 49 targets and 14 eyes (left and right eyes of all participants).

The top line in Fig. 13 presents the results for MAGE when 2, 3, and 4 CRs are available. MAGE was calibrated using 9 calibration targets to fit a second-order polynomial, and it was assumed that all 4 CRs were visible during calibration. The uniformity of light gray areas shows that MAGE presents good accuracy and precision over the whole computer screen at P_0, regardless of the number of CRs available during estimation (i.e., regardless of which L^T was used).

The remaining lines in Fig. 13 present the results for SAGE when 2, 3, and 4 CRs are available during calibration and gaze estimation. In the second line, SAGE was configured to use a second-order polynomial estimated from 9 calibration points. In this case, the distribution of errors using 4 CRs is similar to MAGE's. The third and fourth lines show the results for SAGE using homography transformations calibrated using 9

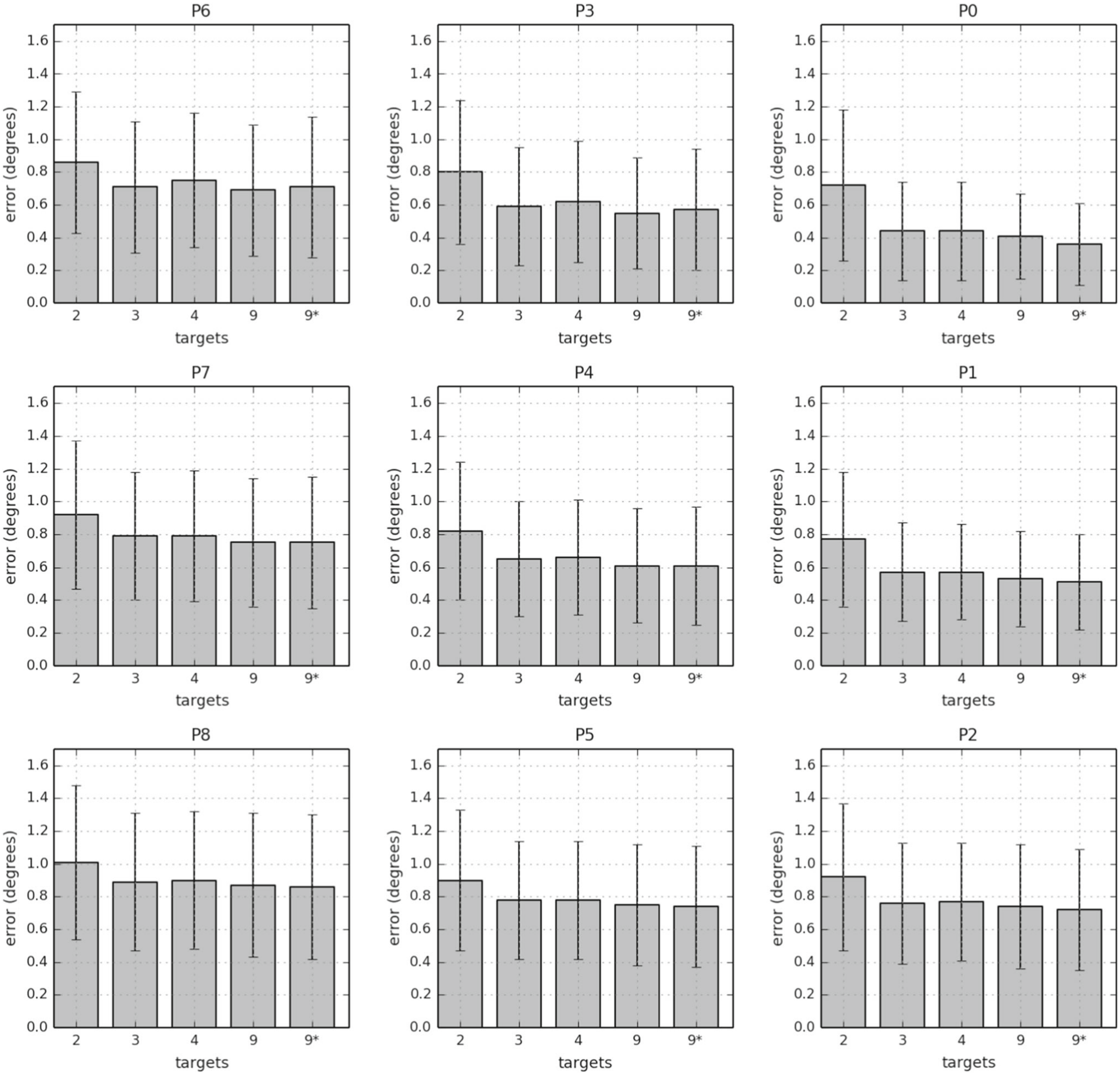


Fig. 11 Performance of SAGE when 3 CRs are visible, using a variable number of calibration targets. The graph positions P_0 to P_8 match the physical head positions shown in Fig. 8. The vertical bars show the grand mean and standard deviation of the gaze estimation error. The horizontal axis shows the number of calibration targets used. Columns 2 to 9 use linear calibration functions and 9* a second-order polynomial

and 4 targets, respectively. The last 2 lines show the error distributions for the affine and similarity conditions.

These heatmaps show that the error distribution over the screen is not as uniform as observed for MAGE, especially when calibration is done using just 2 targets, and also (in a lesser extent) when just 2 CRs are used. Notice that MAGE could not be used at all when only 2 or 3 CRs are available during calibration. The less homogeneous error distribution can be explained by the use of a single S function which cannot be customized to each L_i^T. Nevertheless, the mean gaze estimation error below 1.0^o was achieved even with two missing CRs during calibration, with an upper bound of about 1.5^o. If at least 3 CRs and 3 calibration targets are used, the upper bound is reduced to about 0.9^o. Note that this upper bound value corresponds to the gaze estimation error around particular corners of the screen, but for most screen regions the estimation error is below 0.6^o.

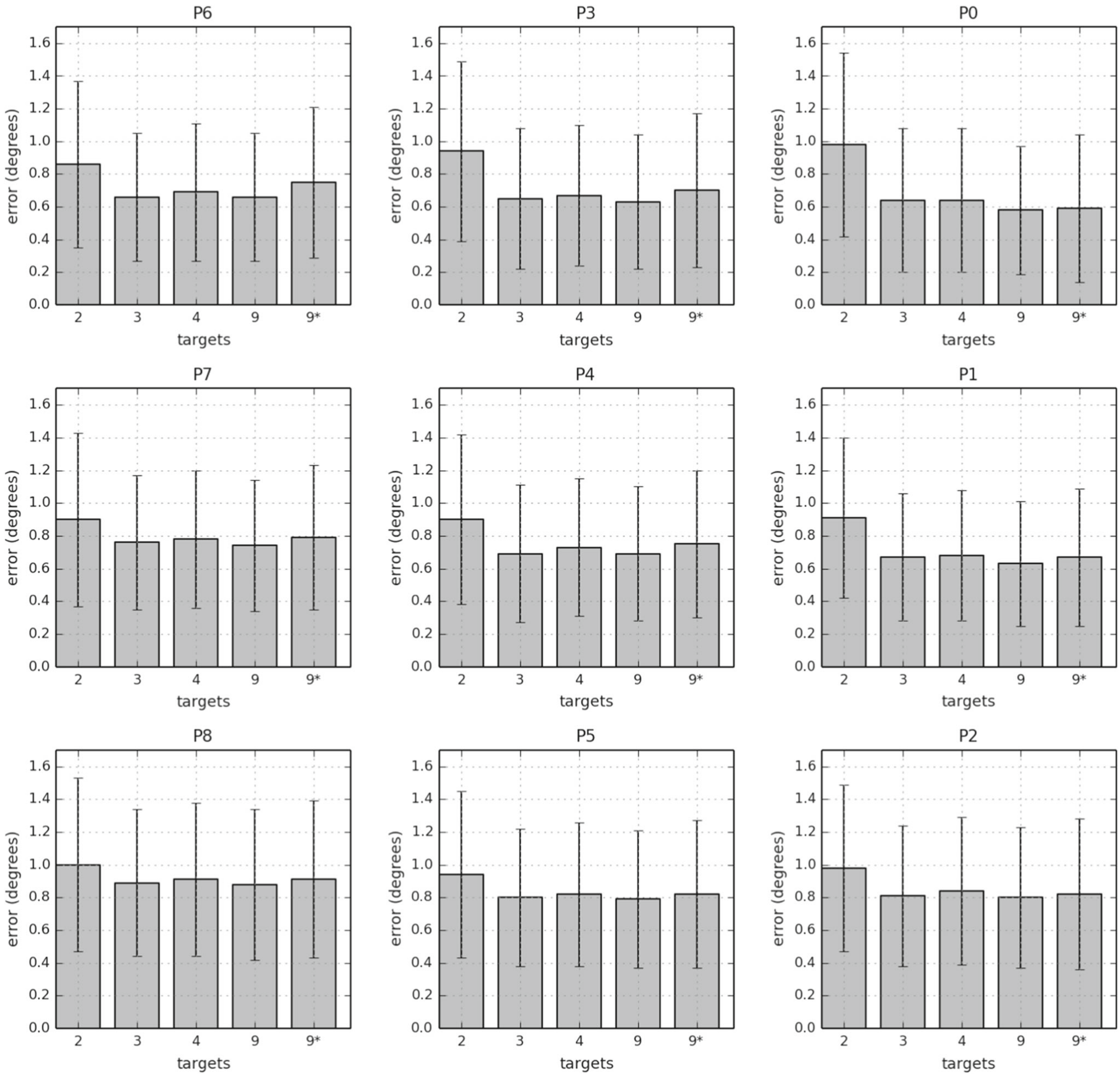


Fig. 12 Performance of SAGE when 2 CRs are visible, using a variable number of calibration targets. The graph positions P_0 to P_8 match the physical head positions shown in Fig. 8. The vertical bars show the grand mean and standard deviation of the gaze estimation error. The horizontal axis shows the number of calibration targets used. Columns 2 to 9 use linear calibration functions and 9* a second-order polynomial

Figure 14 shows the error distribution at P_2, when the head moves 25 cm away from the screen. The darker shades indicate higher average errors for all configurations compared to P_0, though the error distribution remains relatively homogenous. Results using 3 CRs are still very similar to 4 CRs and the overall average error remains under 1^o. Only when using 2 CRs and 2 calibration points, errors get larger, but mostly under 1.5^o.

Figure 15 shows the error distribution at P_6, when the head moves 25 cm to the left of the calibration position P_0. Note that the average error is not as homogenous across the screen when compared to the heatmaps in Figs. 13 and 14. This can be explained by the larger perspective distortions observed in the pattern of CRs caused by lateral head movements. Under these circumstances, the use of a transformation L with fewer degrees of freedom (affine, similarity) will not be as good an approximations for an homography. Nevertheless, the overall

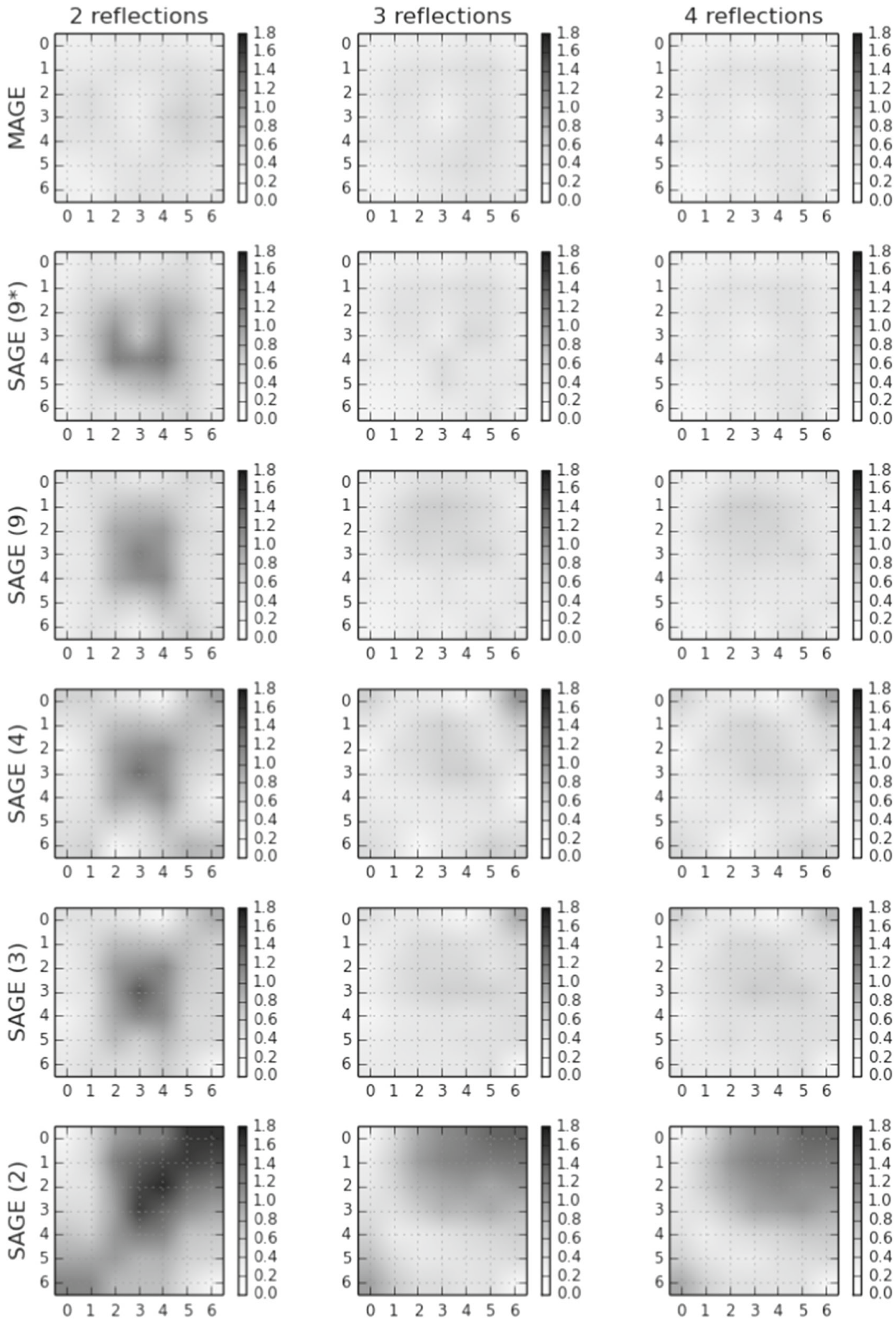


Fig. 13 This heatmap illustrates the grand mean error distribution across the screen of MAGE (based on [33]) and SAGE at the calibration position P_0. MAGE uses a second-order polynomial estimated from 9 calibration targets. For SAGE, results are presented for 9, 4, 3, and 2 calibration targets using linear calibration functions, and 9* using a second-order polynomial

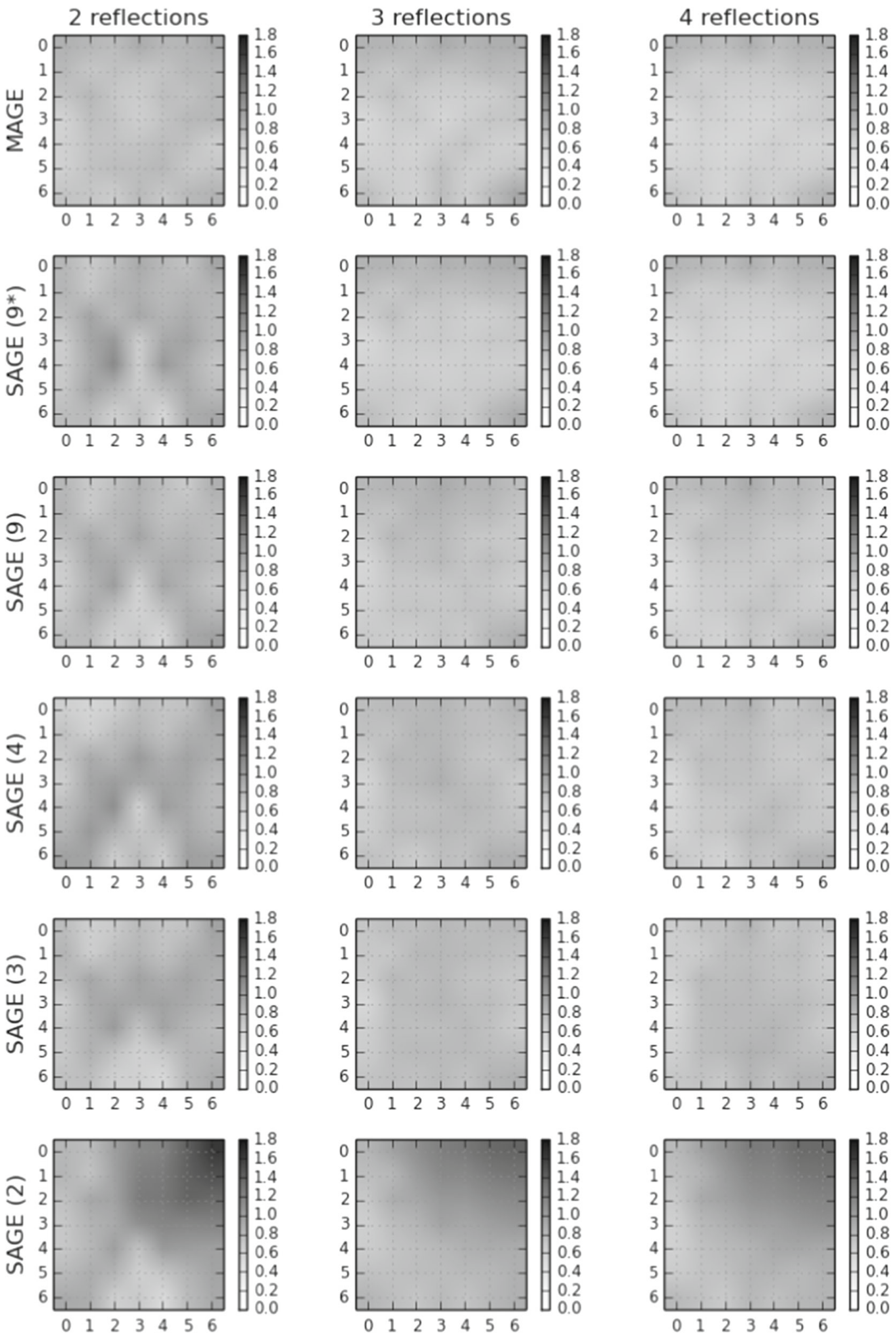


Fig. 14 This heatmap illustrates the grand mean error distribution across the screen of MAGE (based on [33]) and SAGE at the calibration position P_2 which is 25 cm away from the calibration point P_0. MAGE uses a second-order polynomial estimated from 9 calibration targets. For SAGE, results are presented for 9, 4, 3, and 2 calibration targets using linear calibration functions, and 9* using a second-order polynomial

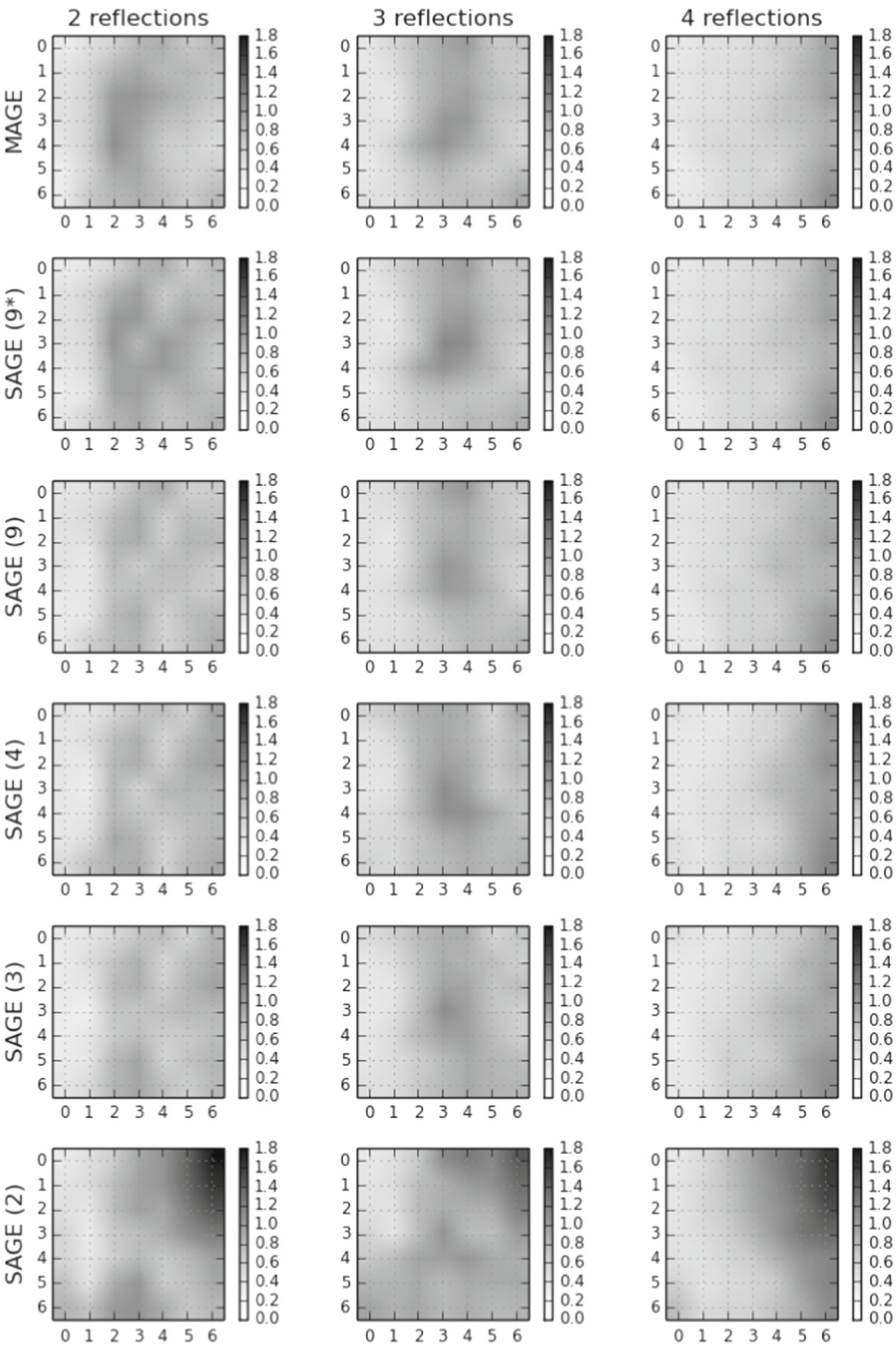


Fig. 15 This heatmap illustrates the grand mean error distribution across the screen of MAGE (based on [33]) and SAGE at the calibration position P_6 which is 25 cm to the left of the calibration point P_0. MAGE uses a second-order polynomial estimated from 9 calibration targets. For SAGE, results are presented for 9, 4, 3, and 2 calibration targets using linear calibration functions, and 9* using a second-order polynomial

average error remains under 1.2^o for any number of CRs as long as at least a 3-point calibration is performed.

5.4 Comparison with Other Methods

Several gaze estimation methods that used a single uncalibrated camera were discussed in Sect. 2, but since the focus was on the methods that can adapt to missing CRs, just SAGE and MAGE were evaluated and discussed in detail so far. Now we compare our results with those previous works. In increasing order on the number of light sources required to work, the following PoG estimation methods have been presented in Sect. 2: PCR using a single light source [38]; PCR with two light sources [3,4]; and cross-ratio/homography-based methods such as the homography normalization [16] and the planarization [9,10] methods.

We have already mentioned that some of these methods are, in fact, similar to particular conditions of SAGE. When just 2 CRs are used to define L and 9 calibration points are used to estimate a second-order polynomial for S, SAGE's behavior is similar to the PCR method with two light sources [3,4]. When it is assumed that all 4 CRs are used to define L and 4 or 9 calibration points are used to estimate S, then SAGE is equivalent to the homography normalization method [16].

Although the PCR using a single light [38] source also fits the SLD framework presented (with L being a translation transformation, and S a second-order polynomial), results for this method were not presented due to its high sensitivity to head motion. Its mean gaze estimation error at P_0 is comparable to the error achieved by SAGE (and MAGE) when all 4 CRs and 9 calibration points are used, but for any head position other than P_0 large errors are observed.

The only method that consistently outperforms SAGE and MAGE is the planarization method [9,10]. Though its performance is comparable at the calibration position P_0, the error variation is smaller when the head moves to other positions. Nonetheless, it is not adaptive, i.e., it requires all 4 CRs to be available at all times.

Table 1 shows the performances of the PCR using 1 single light source and 9 calibration points (PCR-1-9) and the planarization methods. To facilitate the comparison with the adaptive methods, results for MAGE in its best scenario, where 4 CRs are always used, calibrated using 9 points (MAGE-4-9), and SAGE in its worst scenario, where 2 CRs and 2 calibration targets are used (SAGE-2-2) are also shown. The results for PCR-1-9 and planarization were computed using the same dataset described in Sect. 4. Table 1 shows the grand mean gaze estimation error (for all 14 eyes) at each of the 9 head position (P_0 to P_8). Calibration of the methods was also performed at P_0.

Table 1 Grand mean and standard deviation of gaze estimation error (in degrees of visual angle) for the PCR method using just 1 single light source and 9 calibration points (PCR-1-9), SAGE in its worst scenario, where just 2 CRs and 2 calibration targets are used (SAGE-2-2), MAGE in its most favorable scenario, where 4 CRs are used using 9 calibration points (MAGE-4-9), and for the planarization method. Results are shown at each head position (P_0 to P_8), and calibration was performed at P_0

	PCR-1-9	SAGE-2-2	MAGE-4-9	Planarization
P_0	0.35 ± 0.23	0.98 ± 0.56	0.35 ± 0.24	0.38 ± 0.21
P_1	6.07 ± 2.08	0.91 ± 0.49	0.48 ± 0.28	0.42 ± 0.18
P_2	7.93 ± 2.72	0.98 ± 0.51	0.67 ± 0.36	0.51 ± 0.20
P_3	2.52 ± 1.06	0.94 ± 0.55	0.50 ± 0.30	0.48 ± 0.22
P_4	6.65 ± 2.27	0.90 ± 0.52	0.58 ± 0.33	0.49 ± 0.21
P_5	8.18 ± 2.79	0.94 ± 0.51	0.69 ± 0.36	0.53 ± 0.21
P_6	4.80 ± 1.61	0.86 ± 0.51	0.62 ± 0.38	0.51 ± 0.22
P_7	7.43 ± 2.50	0.90 ± 0.53	0.68 ± 0.40	0.53 ± 0.20
P_8	8.31 ± 2.83	1.00 ± 0.53	0.81 ± 0.44	0.59 ± 0.21

Observe that the accuracy of PCR-1-9 is comparable to MAGE-4-9 at the calibration position P_0, which is expected because they share the same second-order polynomial function computed using the 9 calibration points. Observe that, at P_0, PCR-1-9 outperforms SAGE-2-2, which uses a simpler linear function and only 2 calibration points. On the other hand, PCR-1-9 is very sensitive to head movements, while the accuracy of SAGE-2-2 remains about 1^o.

Remember that, as described in Sect. 2.1, the planarization method corrects the main sources of error of projective transformation methods by "planarizing" eye features. It requires the 4 CRs to be available at all times to compute the PoG, and a single calibration point to estimate the κ-angle. Despite not being adaptive to missing CRs, the results in Table 1 show that planarization better compensates head motion than the second-order polynomial function used by MAGE-4-9, and might be the preferred method for more constrained scenarios.

6 Conclusions

The first contribution of this paper was the introduction of the Screen-Light Decomposition (SLD) framework for point-of-gaze estimation. The framework synthesizes all available eye-tracking methods that relies on a single uncalibrated camera and a variable number of light sources. By decomposing the gaze estimation function into two components, a light component L that uses the CRs to reduce the effects of head movements from the eye features, mapping the features to a normalized space, and a screen component S that maps the normalized eye features to screen coordinates, SLD provides a deeper understanding of the various methods and allows us to predict the performance of an eye tracker based on the choice of functions used for L and S.

SLD was used to design SAGE, a single normalized space adaptive gaze estimation technique. This new technique exploits the decoupling of the L normalization and the S calibration mappings to gracefully adapt the gaze model to the number of available CRs and intended number of calibration targets. SAGE advances the state of the art of adaptive eye-tracking techniques by allowing missing CRs during calibration and requiring fewer calibration points. This feature is very important in realistic scenarios because, if CRs are missing for particular head position or orientation, it is likely that the CRs will be missing during the calibration procedure as well.

The performance degradation of SAGE when fewer CRs are available has been investigated using a publicly available dataset. The results show that normalization mappings with higher degrees of freedom are preferred, but the difference between homographic and affine models was relatively small. These models outperform the similarity model. An important characteristic of SAGE is its robustness to missing CRs even during calibration.

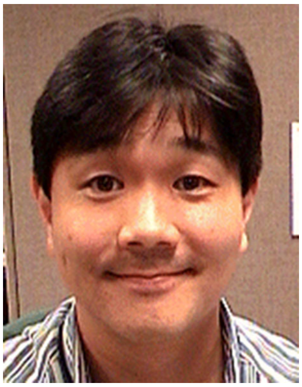
The results also show that the use of linear transformations (such as affine and homography) reduces the number of required calibration targets with low impact on the overall performance of the gaze estimation, when compared to second-order polynomials.

Acknowledgements This research was supported by Grant 2011/00267-1 and 2016/10148-2 from the São Paulo Research Foundation (FAPESP).

References

1. Arar, N.M., Gao, H., Thiran, J.: A regression-based user calibration framework for real-time gaze estimation. IEEE Trans. Circuits Syst. Video Technol. **27**(12), 2623–2638 (2017). https://doi.org/10.1109/TCSVT.2016.2595322
2. Baltrušaitis, T., Robinson, P., Morency, L.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10 (2016). https://doi.org/10.1109/WACV.2016.7477553
3. Cerrolaza, J.J., Villanueva, A., Cabeza, R.: Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems. In: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, ETRA '08, pp. 259–266. ACM, New York, NY, USA (2008). https://doi.org/10.1145/1344471.1344530
4. Cerrolaza, J.J., Villanueva, A., Cabeza, R.: Study of polynomial mapping functions in video-oculography eye trackers. ACM Trans. Comput. Hum. Interact. **19**(2), 10:1–10:25 (2012). https://doi.org/10.1145/2240156.2240158
5. Cheng, H., Liu, Y., Fu, W., Ji, Y., Yang, L., Zhao, Y., Yang, J.: Gazing point dependent eye gaze estimation. Pattern Recognit. **71**(1), 36–44 (2017). https://doi.org/10.1016/j.patcog.2017.04.026
6. Cortiñas, M., Chocarro, R., Villanueva, A.: Image, brand and price info: Do they always matter the same? In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA '19, pp. 1–8. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3317960.3321616
7. Coutinho, F.L., Morimoto, C.H.: Free head motion eye gaze tracking using a single camera and multiple light sources. In: Oliveira Neto, M.M.D., Carceroni, R.L. (eds.) Proceedings, pp. 171–178. IEEE Computer Society (2006)
8. Coutinho, F.L., Morimoto, C.H.: A depth compensation method for cross-ratio based eye tracking. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA '10, pp. 137–140. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1743666.1743670
9. Coutinho, F.L., Morimoto, C.H.: Augmenting the robustness of cross-ratio gaze tracking methods to head movement. In: Proceedings of the 2012 Symposium on Eye-Tracking Research & Applications, ETRA '12, pp. 1–8 (2012)
10. Coutinho, F.L., Morimoto, C.H.: Improving head movement tolerance of cross-ratio based eye trackers. Int. J. Comput. Vis. **101**(3), 459–481 (2013)
11. Faugeras, O.: Stratification of three-dimensional vision: projective, affine, and metric representations. J. Opt. Soc. Am. A **12**(3), 465–484 (1995)
12. Feit, A.M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., Morris, M.R.: Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pp. 1118–1130. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025599
13. García-Dopico, A., Pérez, A., Pedraza, J.L., Córdoba, M.L.: Precise non-intrusive real-time gaze tracking system for embedded setups. Comput. Inform. **36**(2), 257–282 (2017)
14. George, A., Routray, A.: Real-time eye gaze direction classification using convolutional neural network. In: 2016 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5 (2016). https://doi.org/10.1109/SPCOM.2016.7746701
15. Guestrin, E.D., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Eng. **53**(6), 1124–1133 (2006)
16. Hansen, D.W., Agustin, J.S., Villanueva, A.: Homography normalization for robust gaze estimation in uncalibrated setups. In: Proceedings of the 2010 Symposium on Eye-Tracking Research &; Applications, ETRA '10, pp. 13–20 (2010)
17. Hansen, D.W., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010). https://doi.org/10.1109/TPAMI.2009.30
18. Hansen, D.W., Roholm, L., Ferreiros, I.G.: Robust glint detection through homography normalization. In: Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '14, pp. 91–94. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2578153.2578165
19. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, New York (2003)
20. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). ISBN: 0521540518
21. Hennessey, C.A., Lawrence, P.D.: Improving the accuracy and reliability of remote system-calibration-free eye-gaze tracking. IEEE Trans. Biomed. Eng. **56**(7), 1891–900 (2009)
22. Hennessey, C.A., Lawrence, P.D.: Improving the accuracy and reliability of remote system-calibration-free eye-gaze tracking. IEEE Trans. Biomed. Eng. **56**(7), 1891–1900 (2009). https://doi.org/10.1109/TBME.2009.2015955
23. Ji, Q., Yang, X.: Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. Real-Time Imaging **8**(5), 357–377 (2002)
24. Kang, J.J., Guestrin, E.D., Eizenman, E.: Investigation of the cross-ratio method for point-of-gaze estimation. Trans. Biomed. Eng. **55**(9), 2293–302 (2008)

25. Kang, J.J., Guestrin, E.D., Maclean, W.J., Eizenman, M.: Simplifying the cross-ratios method of point-of-gaze estimation. In: 30th Canadian Medical and Biological Engineering Conference (CMBEC30) (2007)
26. Kar, A., Corcoran, P.: A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. IEEE Access **5**, 16495–16519 (2017). https://doi.org/10.1109/ACCESS.2017.2735633
27. Koutras, P., Maragos, P.: Estimation of eye gaze direction angles based on active appearance models. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2424–2428 (2015). https://doi.org/10.1109/ICIP.2015.7351237
28. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
29. Kurauchi, A.T.N., Feng, W., Joshi, A., Morimoto, C.H., Betke, M.: Eyeswipe: Dwell-free text entry using gaze paths. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 1952–1956. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858335
30. Lai, C., Shih, S., Hung, Y.: Hybrid method for 3-d gaze tracking using glint and contour features. IEEE Trans. Circuits Syst. Video Technol. **25**(1), 24–37 (2015). https://doi.org/10.1109/TCSVT.2014.2329362
31. Lander, C., Gehring, S., Krüger, A., Boring, S., Bulling, A.: Gazeprojector: Accurate gaze estimation and seamless gaze interaction across multiple displays. In: Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, UIST 2015, Charlotte, NC, USA, November 8–11, 2015, pp. 395–404 (2015). https://doi.org/10.1145/2807442.2807479
32. Li, F., Munn, S., Pelz, J.: A model-based approach to videobased eye tracking. J. Mod. Opt. **55**(4–5), 503–531 (2008). https://doi.org/10.1080/09500340701467827
33. Ma, C., Choi, K.A., Choi, B.D., Ko, S.J.: Robust remote gaze estimation method based on multiple geometric transforms. Opt. Eng. **54**(8), 083103 (2015). https://doi.org/10.1117/1.OE.54.8.083103
34. Ma, Z., Liu, Z., Ho, M., Yen, J., Chen, Y.: Long range gaze estimation with multiple near-infrared emitters. In: 2017 International Automatic Control Conference (CACS), pp. 1–5 (2017). https://doi.org/10.1109/CACS.2017.8284270
35. Meyerding, S.G., Merz, N.: Consumer preferences for organic labels in Germany using the example of apples–combining choice-based conjoint analysis and eye-tracking measurements. J. Clean. Prod. **181**, 772–783 (2018). https://doi.org/10.1016/j.jclepro.2018.01.235
36. Morimoto, C.H., Koons, D., Amir, A., Flickner, M.: Pupil detection and tracking using multiple light sources. Image Vis. Comput. **18**(4), 331–335 (2000)
37. Morimoto, C.H., Leyva, J.A.T., Diaz-Tula, A.: Context switching eye typing using dynamic expanding targets. In: Proceedings of the Workshop on Communication by Gaze Interaction, COGAIN '18, pp. 1–9. ACM, New York, NY, USA (2018). https://doi.org/10.1145/3206343.3206347
38. Morimoto, C.H., Mimica, M.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**(1), 4–24 (2005)
39. Nguyen, C., Liu, F.: Gaze-based notetaking for learning from lecture videos. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 2093–2097. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858137
40. Noureddin, B., Lawrence, P., Man, C.: A non-contact device for tracking gaze in a human computer interface. Comput. Vis. Image Underst. **98**(1), 52–82 (2005)
41. Nyström, M., Andersson, R., Holmqvist, K., van de Weijer, J.: The influence of calibration method and eye physiology on eyetracking data quality. Behav. Res. Methods **45**(1), 272–288 (2013). https://doi.org/10.3758/s13428-012-0247-4
42. Ramirez Gomez, A., Gellersen, H.: Looking outside the box: reflecting on gaze interaction in gameplay. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play, CHI PLAY '19, pp. 625–637. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3311350.3347150
43. Santini, T., Fuhl, W., Kasneci, E.: Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pp. 2594–2605. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025950
44. Schnipke, S., Todd, M.: Trials and tribulations of using an eye-tracking system. In: Proceedings of the 2000 CHI Conference on Human Factors in Computing Systems, CHI'2000. ACM (2000)
45. Shih, S.W., Wu, Y.T., Liu, J.: A calibration-free gaze tracking technique. In: Proceedings of the 15th International Conference on Pattern Recognition, pp. 201–204 (2000)
46. Tula, A.D., Morimoto, C.H.: Augkey: Increasing foveal throughput in eye typing with augmented keys. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 3533–3544. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858517
47. Wang, K., Ji, Q.: Real time eye gaze tracking with 3d deformable eye-face model. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1003–1011 (2017). https://doi.org/10.1109/ICCV.2017.114
48. Wang, Y., Shen, T., Yuan, G., Bian, J., Fu, X.: Appearance-based gaze estimation using deep features and random forest regression. Knowl. Based Syst. **110**(1), 293–301 (2016). https://doi.org/10.1016/j.knosys.2016.07.038
49. Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3756–3764 (2015). https://doi.org/10.1109/ICCV.2015.428
50. Yoo, D.H., Chung, M.J.: A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. Comput. Vis. Image Underst. **98**(1), 25–51 (2005)
51. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
52. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: MPIIGaze: real-world dataset and deep appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 162–175 (2019). https://doi.org/10.1109/TPAMI.2017.2778103
53. Zhang, Y., Pfeuffer, K., Chong, M.K., Alexander, J., Bulling, A., Gellersen, H.: Look together: using gaze for assisting co-located collaborative search. Pers. Ubiquitous Comput. **21**(1), 173–186 (2017). https://doi.org/10.1007/s00779-016-0969-x



Carlos H. Morimoto received his B.Sc. and M.Sc. in Electronic Engineering from the University of São Paulo (USP). PhD in Computer Science from the University of Maryland at College Park. He is currently an associate professor at the Department of Computer Science at Institute of Mathematics and Statistics (IME) at USP. His main areas of interest include Computer Vision, Image Processing, and Human Computer Interaction. His projects involve the understanding of human activities using computer vision to enhance human interaction, with focus on person detection, tracking of body, face, and eye movements.

Flávio L. Coutinho received his Bachelor degree in Computer Science from Institute of Mathematics and Statistics of University of São Paulo, Brasil, in 2001. He also received his M.Sc. and Ph.D. in Computer Science from the same institution in 2006 and 2011. He is currently an assistant professor at School of Arts, Sciences, and Humanities of University of São Paulo, Brasil. His current research interests include computer vision, image processing, human-computer interaction, eye gaze tracking and gaze based interfaces.

Dan W. Hansen currently works at the Software and Systems, IT University of Copenhagen. Dan does research in Artificial Intelligence, Human-computer Interaction and Computing in Mathematics, Natural Science, Engineering and Medicine. He is particularly interested in EYE INFORMATION in these areas.