

# Auto-Calibrated Gaze Estimation Using Human Gaze Patterns

Fares Alnajar<sup>1</sup> · Theo Gevers<sup>1</sup> · Roberto Valenti<sup>1</sup> · Sennay Ghebream<sup>2</sup>

Received: 5 November 2015 / Accepted: 24 April 2017 / Published online: 17 May 2017  
© The Author(s) 2017. This article is an open access publication

**Abstract** We present a novel method to auto-calibrate gaze estimators based on gaze patterns obtained from other viewers. Our method is based on the observation that the gaze patterns of humans are indicative of where a new viewer will look at. When a new viewer is looking at a stimulus, we first estimate a topology of gaze points (initial gaze points). Next, these points are transformed so that they match the gaze patterns of other humans to find the correct gaze points. In a flexible uncalibrated setup with a web camera and no chin rest, the proposed method is tested on ten subjects and ten images. The method estimates the gaze points after looking at a stimulus for a few seconds with an average error below  $4.5^\circ$ . Although the reported performance is lower than what could be achieved with dedicated hardware or calibrated setup, the proposed method still provides sufficient accuracy to trace the viewer attention. This is promising considering the fact that auto-calibration is done in a flexible setup, without the use of a chin rest, and based only on a few seconds of gaze initialization data. To the best of our knowledge, this is the first work to use human gaze patterns in order to auto-calibrate gaze estimators.

**Keywords** Eye gaze estimation · Calibration free · Auto-calibration

## 1 Introduction

Gaze estimation is the process of determining where a person is looking at in a predefined plane. It is an important task in computer vision and has numerous applications: i.e. human–computer interaction, assisting disabled users (e.g. eye typing) (Majaranta and Rih 2002), and human behavior analysis (Smith et al. 2008).

In general, gaze estimation methods fall into two categories: (1) appearance-based methods (Hansen et al. 2002; Lu et al. 2011; Valenti and Gevers 2012; Sugano et al. 2014) and (2) 3D-eye model-based methods (Villanueva et al. 2006; Guestrin and Eizenman 2006, 2008; Chen and Ji 2011; Dracos et al. 2015; Xiong et al. 2015). The former class extracts features from images of the eyes and map them to points on the gaze plane (i.e. gaze points). The latter aims to construct a 3D model of the eye to estimate the visual axis. Then, the intersection of the axis and the gaze plane determines the gaze point. Regardless of which gaze estimation method is used, a calibration procedure is always needed. The calibration can be camera-based (estimating the camera parameters), geometric calibration (estimating the relationships between the scene components like the camera, the gaze plane, and the user), personal calibration (determining the angle between visual and optical axes), or gaze mapping correlation (Hansen and Ji 2010). An overview of the different approaches of gaze estimation and calibration can be found in Hansen and Ji (2010).

3D-eye models require special equipment like cameras with multiple light sources and infrared. The costs and the usage requirements (infrared, for example, is not reliable when used outdoors) limit their range of applicability. On the other hand, appearance-based approaches are less accurate than 3D-eye-models and less invariant to head pose changes. Yet, low-cost cameras are common and sufficient

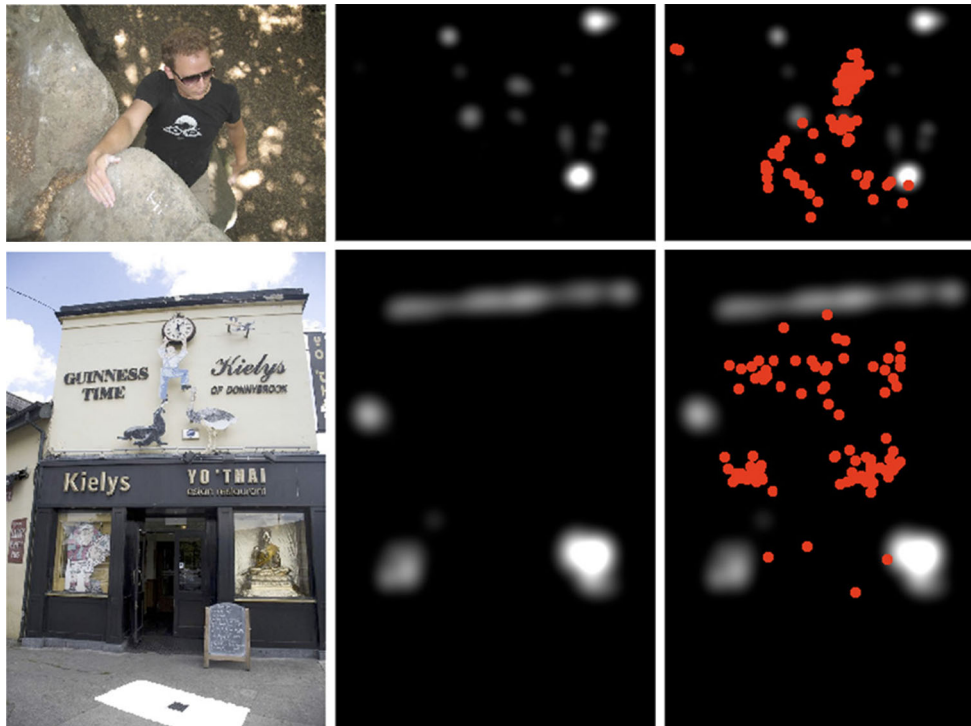
---

Communicated by Antonio Torralba.

✉ Fares Alnajar  
F.Alnajar@uva.nl

<sup>1</sup> Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands

<sup>2</sup> Amsterdam University College, Amsterdam, The Netherlands



**Fig. 1** (Taken from Judd et al. 2009). Examples where saliency models do not match the human fixations. *Bright spots* indicate the saliency model predictions and the *red dots* refer to the human gaze points (Color figure online)

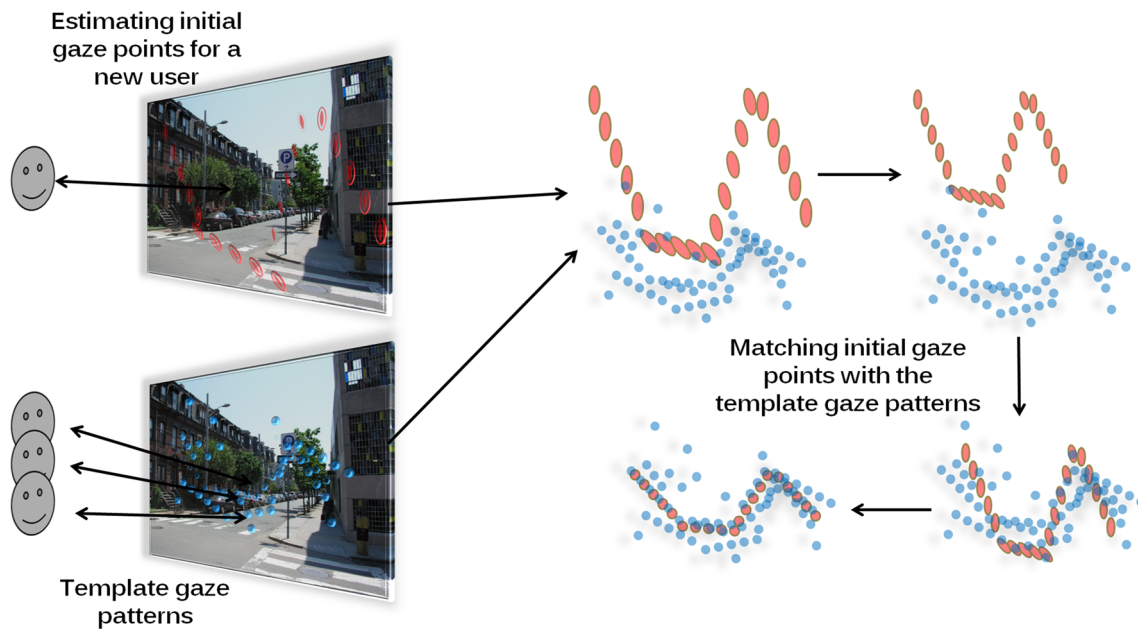
for appearance-based approaches which makes them suitable for applications where high accuracy is not essential. Consider for example an application of people looking at advertisements for market research. Asking each participant to buy dedicated cameras or to do the experiment in the lab is time consuming and costly. Because low-cost cameras are integrated in almost every laptop or tablet nowadays, appearance-based methods are more suitable in such an application.

Besides the choice of the recording equipment, the adopted approach allows for a certain level of flexibility in the setup and calibration. During calibration, users are usually asked to fixate their gaze on certain points while images of their eyes are captured. This procedure is cumbersome and sometimes impractical. In case of, for example, tracing costumers' attention in shopping malls, estimating the gaze points or regions should be done passively. Hence, some approaches propose methods to reduce the number of calibration points. However, in the case of passive gaze estimation, the calibration should be done completely automatically without an active calibration procedure imposed on the user.

Some recent studies focus on visual saliency information in images and videos to avoid applying active human calibration. Sugano et al. (2010, 2013) treat saliency maps extracted from videos as probability distributions for gaze points. Gaussian process regression is used to learn the map-

ping between the images of the eyes and the gaze points. Chen and Ji (2011) use 3D models of the eye and incrementally estimate the angle between the visual and optical axes by combining the image saliency with the 3D model. The argument for using saliency is that people look at salient regions with higher probability than other regions. However, as shown in Judd et al. (2009), the computational saliency models do not frequently match the actual human saccades (Fig. 1). We propose that the gaze patterns of several viewers provide important cues for the auto-calibration of new viewers. This is based on the assumption that humans produce similar gaze patterns when they look at a stimulus. The assumption is supported by Judd et al. (2009), where the authors show that fixation locations of several humans are strongly indicative, in general, of where a new viewer will look at. To the best of our knowledge, our work is the first to use human gaze patterns in order to auto-calibrate gaze estimators.

In this paper, which is an extension of our previous work (Alnajjar et al. 2013), we present a novel approach to auto-calibrate gaze estimators based on the similarity of human gaze patterns. In addition, we make use of the topology of the gaze points. Consider, in a fully uncalibrated setting, a person following a stimulus from left to right. It would be difficult to indicate where the gaze points are on the gaze plane. However, their relative locations can still be inferred and used for auto-calibration. In a fully uncalibrated setting, when a new



**Fig. 2** Graphical illustration of the proposed method. Template gaze patterns refer to the gaze points of other individuals for the same gaze plane (display). When a new user looks at the stimulus, his or her ini-

tial gaze points are first estimated which preserves the relative locations between the gaze points. These points are transformed so that they match the template gaze patterns

subject looks at a stimulus, initial gaze points are inferred. Then, a transformation is computed to map the initial gaze points to match the gaze patterns of other users. In this way, we use all the initial gaze points to match the human gaze patterns instead of using each gaze point at the time. Consequently, the transformed points represent the auto-calibrated estimated gaze points.

The rest of the paper is organized as follows. The proposed method is explained in Sect. 2. Next, we describe the experimental setup and evaluation in Sect. 3. The results are discussed in Sect. 4. Finally, the conclusions are given in Sect. 5.

## 2 Auto-Calibrated Gaze Estimation Using Human Gaze Patterns

We start from the observation that gaze patterns of individuals are similar for a certain stimulus (Judd et al. 2009). Although, there is no guarantee that people always look at the exact same regions, human gaze patterns will provide important cues about the locations of the gaze points of a new observer. The pipeline of the proposed method is as follows: when a new user is looking at a stimulus, the initial gaze points are computed first. Then, a transformation is inferred which maps the initial gaze points to gaze patterns of other individuals. In this paper, we consider transformations which combine translation and scaling (per dimension). Including

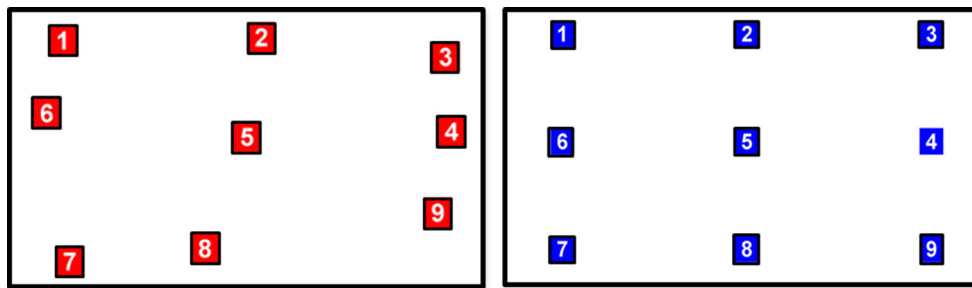
other transformations like rotation or shearing may yield better mapping. However, they are not taken into account, since (1) translation and scaling are more common for gaze estimation, and (2) to reduce the search space. Figure 2 illustrates the pipeline.

### 2.1 Gaze Points Initialization

The final gaze points should eventually match the human gaze patterns. However, we need to start from an *initial* estimation of the gaze points. Hereafter, we present two methods to achieve this: estimation of initial gaze points from eye templates and estimation based on 2D-manifold.

#### 2.1.1 Eye Templates

In this approach, the eye images of a (template) subject are captured while fixating the eyes on points on a gaze plane. The images of the eyes of a new user are captured and compared with the template eye images. The idea is to reconstruct eye images based on the eye image templates. Note that here the eye templates are captured once for a single subject. When a new subject uses the gaze estimator, his or her eye images are compared with the already-collected eye templates. This is different from the traditional calibration-based gaze estimator where the eye templates are captured and stored for each subject and/or each different setting. The process can be performed at the raw intensity level or at the feature level.



**Fig. 3** The projection of features of 9 eye images on a 2-D manifold (red, left) and the positions of the corresponding gaze points on the gaze plane (blue, right). The 2D manifold is computed using 800 eye images corresponding to various locations on the gaze plane (Color figure online)

We will refer to both eye image representations as feature vectors. Consider  $\{\mathbf{t}_i\}$  to be the template feature vectors, and  $\{p_i\}$  denotes the corresponding gaze points. Furthermore,  $\{w_i\}$  corresponds to the computed weights to reconstruct the feature vector of a new eye image  $\hat{\mathbf{t}}$ :

$$\hat{\mathbf{t}} = \sum_i w_i \mathbf{t}_i \quad \text{s.t.} \quad \sum_i w_i = 1. \quad (1)$$

Then the corresponding gaze point  $\hat{p}$  for  $\hat{\mathbf{t}}$  is calculated as follows:

$$\hat{p} = \sum_i w_i p_i. \quad (2)$$

To find the weights  $\{w_i\}$ , Tan et al. (2002) suggest to first select a subset of  $\{\mathbf{t}_i\}$  where the first and the second neighbors of the sample (in feature space) are used for training. The weight values are then computed as in Roweis and Saul (2000). Lu et al. (2011) select only the direct neighbors as a training subset. Here, we select only the direct neighbors as in Lu et al. (2011).

For a new user, potentially in a different unknown scene setup, the initial gaze points will be incorrect (without calibration). However, the relative locations between the gaze points are preserved.

### 2.1.2 2D Manifold

In their work, Lu et al. (2011) find that the (template) eye features correspond to a 2D manifold while retaining most of the important information about the relative eye movements. The reason is that eyes move, in the appearance-based representation, in two degrees of freedom. Figure 3 shows the projection of features of nine eye images on a 2D manifold and their corresponding nine gaze points on the gaze plane. It can be derived that the feature projections preserve the relative locations of the corresponding gaze points.

The 2D manifold can be obtained by projecting the template features on the first two principal components.

However, the locations on the 2D manifold may be interchanged, transposed, or rotated when compared with the corresponding gaze points. For example, when the eyes move mainly vertically, the first principal component represents the pupil changes on the Y dimension and the second principal component represents the X dimension. Hence, the projected locations need to be transposed. As this step is performed once offline, the projected locations are checked once and transformed to match the corresponding gaze points locations. As in the eye templates method, this procedure is followed once with a single (template) subject. When a new user looks at a stimulus, the eye features are projected on the offline-learned 2D manifold and the projected values are treated as initial gaze points.

The previous two methods (eye templates and 2D manifold) provide a way to find the initial gaze points. In the next section, we explain how to map these points to match the template (human) gaze patterns.

## 2.2 Gaze Points Mapping

Judd et al. (2009) show that the fixation points of several humans correspond strongly with the gaze points of a new user. We aim to exploit this observation to perform calibration without the need for active user participation. To this end, we transform the initial (uncalibrated) gaze points so they match the template gaze patterns for a stimulus. By applying the aforementioned transformation, we aim to transfer the gaze points to their correct positions without explicit calibration. We present two different methods to find the transformation: K-closest points and mixture model fitting. Let the set  $\mathbb{P} = \{\mathbf{p}^1, \mathbf{p}^2, \dots, \mathbf{p}^M\}$  denotes the gaze patterns of  $M$  users (hereafter, we call them *template gaze patterns*) where  $\mathbf{p}^u = \{p_1^u, p_2^u, \dots, p_{S_u}^u\}$  consists of the  $S_u$  gaze points of user  $u$ . Let  $\mathbf{p} = \{p_1, p_2, \dots, p_S\}$  be the initial gaze point set for a new user. The following two methods aim to transform and hence match  $\mathbf{p}$  with the template gaze patterns  $\mathbb{P}$ .

### 2.2.1 K-Closest Points

This method aims to find a mapping which minimizes the sum of distances for each point  $p_j \in \mathbf{p}$  to its  $K$  closest neighbors of  $\mathbb{P}$ . Assume  $\Phi$  is the set of all mappings. The method tries to find a mapping  $\bar{\phi} \in \Phi$  which satisfies:

$$\bar{\phi} = \arg \min_{\phi} \mathcal{L}(\mathbf{p}, \mathbb{P}, \phi), \tag{3}$$

where

$$\mathcal{L}(\mathbf{p}, \mathbb{P}, \phi) = \sum_{j=1}^S \sum_{k=1}^K \|\phi(p_j) - N(\phi(p_j), \mathbb{P}, k)\|. \tag{4}$$

$N(p_j, \mathbb{P}, k)$  is the  $k$  closest point from  $\mathbb{P}$  to  $p_j$ .  $\bar{\phi}$  is the computed mapping and  $\bar{\mathbf{p}} = \bar{\phi}(\mathbf{p})$  represents the mapped auto-calibrated gaze points. Note that we match the initial gaze points  $\mathbf{p}$  with all the gaze patterns in  $\mathbb{P}$  simultaneously. To find  $\bar{\mathbf{p}}$  and  $\bar{\phi}$ , we adopt a gradient-descent approach. To search for a local minimum (or maximum) using gradient descent methods, first, the gradient of the objective function is computed w.r.t to the corresponding parameters. Second, the parameters step toward the negative (positive) direction of the gradient in case of cost (reward) function. These two steps are repeated multiple times (epochs). We restrict the transformation to translation and scaling. The transformation of a point  $p = \begin{bmatrix} x \\ y \end{bmatrix}$  by  $\phi = [s_1, s_1, h_1, h_1]$  is

$$\phi(p) = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \cdot p + \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} s_1 \cdot x + h_1 \\ s_2 \cdot y + h_2 \end{bmatrix}.$$

Here, we assume the origin to be the mean of  $\mathbf{p}$ . The parameter set  $\phi$  is updated based on the derivative of the cost function  $\mathcal{L}$  w.r.t  $\phi$ :

$$\phi \leftarrow \phi - \gamma \nabla_{\phi} \mathcal{L}, \tag{5}$$

where  $\gamma$  is the learning rate and:

$$\nabla_{\phi} \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial h_1} \\ \frac{\partial \mathcal{L}}{\partial h_2} \\ \frac{\partial \mathcal{L}}{\partial s_1} \\ \frac{\partial \mathcal{L}}{\partial s_2} \end{bmatrix}.$$

The derivative w.r.t  $h_1$  is computed as follows:

$$\frac{\partial \mathcal{L}}{\partial h_1} = \sum_{j=1}^S \sum_{k=1}^K \frac{\partial \|\phi(p_j) - N(\phi(p_j), \mathbb{P}, K)\|}{\partial h_1}. \tag{7}$$

Let  $N(\phi(p_j), \mathbb{P}, K) = \{g_1, g_2, \dots, g_K\}$ , then:

$$\frac{\partial \mathcal{L}}{\partial h_1} = \sum_{j=1}^S \sum_{k=1}^K \frac{\partial \sqrt{(\phi(p_j)_x - g_{x,k})^2 + (\phi(p_j)_y - g_{y,k})^2}}{\partial h_1}, \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial h_1} = \sum_{j=1}^S \sum_{k=1}^K \frac{s_1 \cdot p_{x,j} + h_1 - g_{x,k}}{\sqrt{(\phi(p_j)_x - g_{x,k})^2 + (\phi(p_j)_y - g_{y,k})^2}}. \tag{9}$$

and

$$\frac{\partial \mathcal{L}}{\partial s_1} = \sum_{j=1}^S \sum_{k=1}^K \frac{s_1 \cdot p_{x,j}^2 + h_1 \cdot p_{x,j} - g_{x,k} \cdot p_{x,j}}{\|\phi(p_j) - g_k\|}. \tag{10}$$

$\frac{\partial \mathcal{L}}{\partial h_2}$  and  $\frac{\partial \mathcal{L}}{\partial s_2}$  can be derived in a similar manner.

### 2.2.2 Mixture Model

For the K-closest points method, the matching is measured by the distance between each point of the initial gaze set and its closest neighbors in the template gaze patterns. Here, the initial gaze points are mapped to match a mixture model which is fit to the template gaze patterns. More specifically, we first model the template gaze patterns by a Gaussian mixture model. Next, the initial gaze points are transformed so that the probability density function of the transformed points is maximized. Formally, the method searches for a mapping  $\bar{\phi} \in \Phi$  so that:

$$\bar{\phi} = \arg \max_{\phi} \sum_{j=1}^S pdf(\phi(p_j)), \tag{11}$$

where

$$pdf(p) = \sum_{k=1}^K \omega_k \mathcal{N}(p | \mu_k, \Sigma_k), \tag{12}$$

and

$$\mathcal{N}(p | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left(-\frac{1}{2}(p - \mu)^T \Sigma^{-1}(p - \mu)\right). \tag{13}$$

$K$  is the number of model components,  $\omega_k$  is the mixing coefficient of the  $k_{th}$  Gaussian component  $\mathcal{N}(\mu_k, \Sigma_k)$  with  $\mu_k$  mean and  $\Sigma_k$  covariance matrix.  $\bar{\phi}$  is computed again by a gradient descent approach. The parameter set  $\phi$  is updated as follows:

$$\phi \leftarrow \phi + \gamma \nabla_{\phi} \mathcal{F}, \tag{14}$$

where  $\mathcal{F}$  is the reward function we aim to maximize:

$$\mathcal{F} = \sum_{j=1}^S \sum_{k=1}^K \omega_k \mathcal{N}(\phi(p_j) | \mu_k, \Sigma_k). \quad (15)$$

The derivative w.r.t  $h_1$  is computed as follows:

$$\frac{\partial \mathcal{F}}{\partial h_1} = \sum_{j=1}^S \sum_{k=1}^K \omega_k \frac{\frac{1}{\sqrt{(2\pi)^2 |\Sigma_k|}} \exp\left(-\frac{1}{2}(\phi(p_j) - \mu_k)^T \Sigma_k^{-1} (\phi(p_j) - \mu_k)\right)}{\partial \phi(p_j)} \frac{\partial \phi(p_j)}{\partial h_1}. \quad (16)$$

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial h_1} &= \sum_{j=1}^S \sum_{k=1}^K \omega_k \frac{1}{\sqrt{(2\pi)^2 |\Sigma_k|}} \exp\left(-\frac{1}{2}(\phi(p_j) - \mu_k)^T \Sigma_k^{-1} (\phi(p_j) - \mu_k)\right) \\ &+ \sum_{j=1}^S \sum_{k=1}^K \frac{\partial\left(-\frac{1}{2}(\phi(p_j) - \mu_k)^T \Sigma_k^{-1} (\phi(p_j) - \mu_k)\right)}{\partial \phi(p_j)} \frac{\partial \phi(p_j)}{\partial h_1} \end{aligned} \quad (17)$$

$$\frac{\partial \mathcal{F}}{\partial h_1} = \sum_{j=1}^S \sum_{k=1}^K \omega_k \mathcal{N}(\phi(p_j) | \mu_k, \Sigma_k) + \sum_{j=1}^S \sum_{k=1}^K (-\phi(p_j) - \mu_k)^T \Sigma_k^{-1} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (18)$$

and

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial s_1} &= \sum_{j=1}^S \sum_{k=1}^K \omega_k \mathcal{N}(\phi(p_j) | \mu_k, \Sigma_k) \\ &+ \sum_{j=1}^S \sum_{k=1}^K (-\phi(p_j) - \mu_k)^T \Sigma_k^{-1} \cdot \begin{bmatrix} p_{j,x} \\ 0 \end{bmatrix}. \end{aligned} \quad (19)$$

$\frac{\partial \mathcal{F}}{\partial h_2}$  and  $\frac{\partial \mathcal{F}}{\partial s_2}$  can be derived in a similar manner.

### 3 Experimental Results

In this section, we describe the experimental setup and the data used to evaluate the performance of our method. The first ten images of the eye tracking dataset of Judd et al. (2009) are used as stimuli (Fig. 4). The dataset has the advantage of containing the eye tracking data of 15 subjects for 1003 images collected from Flickr and LabelMe (Russell et al. 2005). Hence, this data is used as template gaze patterns. The dataset contains landscape and portrait images with a  $1024 \times 768$  resolution. The images contain multiple objects and they do not necessarily contain faces or objects centered in the middle of the image, representing a realistic stimuli set.

To obtain the ground truth for a new user, the Tobii T60XL gaze estimator (<http://www.tobii.com/>) is used. It uses four infrared diodes mounted at the bottom of a 24 inch display with a resolution of  $1920 \times 1200$  pixels. The reported error of the gaze estimator is within  $1^\circ$ .

The design of the scene setup is to allow the subjects to look at the stimuli without hard constrains e.g. using a chin rest or sitting at a fixed distance from the stimuli. To collect the eye images, a web camera is mounted above the screen to record the face of the subject. The eye image resolution is around  $60 \times 30$ . The coordinates and direction of the camera is unknown with regard to the gaze plane and can change for

each new subject. Ten subjects were asked to sit where they wanted but within the allowed range of the Tobii system. The subject's distance from the display ranged from 55 to 75 cm. No chin rest is used in the experiments. Heads of the subjects are allowed to move during the experiment.

The subjects were asked to look at each image for 3 s followed by 1 s of showing a gray image. No specific task was asked and the subjects freely viewed the stimuli. The recording of each subject is stored and later analyzed to estimate the gaze points. We follow Lu et al. (2011) to extract the images of the eyes. For each of the ten stimuli, the first corresponding web camera frame is taken as an input by the landmarker (Zhu and Ramanan 2012) to detect the eye corners. In Sugano et al. (2013), the eye corners are detected using the OMRON OKAO vision library. To detect the eye corners for the subsequent frames, we apply template matching using the eye corners of the first frame (for each stimulus) as templates. The eye images are then cropped from the corner and resized to  $70 \times 35$ . Histogram equalization is applied to alleviate illumination changes. Regarding the gradient descent search in the matching methods, the number of epochs is set to 50. To prevent over downscaling the initial gaze points, we set a lower bound of scaling equal to 90% of the scale of the template gaze patterns.

#### 3.1 Results on Artificially Distorted Data

Our assumption is that a collection of gaze patterns of individuals can be used to automatically infer the gaze calibration of a new user. In this section, we validate the assumption on

**Fig. 4** The 10 images used as stimuli in our experiments. The images show landscapes and street views where multiple objects are present in the scene



artificially distorted data. More specifically, we use the eye tracking dataset in Judd et al. (2009) and apply a distortion to the subject fixations. The distorted fixations are considered as a simulation of the initial (uncalibrated) gaze points. For each stimulus, we apply a random translation and scaling to the fixation set of each subject. Then, the proposed methods, i.e. K-closest points and mixture model methods are used to transform the distorted gaze points to their correct locations. The first 30 images in the dataset are used in this experiment. For each image, we tested the subjects with 10 or more fixations. We discarded the images where the number of subjects (10 or more fixations) was below 6 to ensure sufficient gaze patterns. Using the K-closest points, the mean error across all images is  $3.3^\circ$ , while the error is  $3.5^\circ$  using the mixture model fitting (the scene setup details can be found in Judd et al. (2009)). The same procedure is applied on the ground truth gaze points obtained from our collected data. For this dataset, the K-closest points and mixture model fitting obtained accuracies of  $2.6^\circ$  and  $2.4^\circ$  respectively. The results show the validity of the proposed methods to bring the distorted (uncalibrated) gaze points closer to their correct locations for different sets of template gaze patterns. Regarding the parameter setting, we set  $K$  in the K-closest points method to 5 and the number of Gaussian components to 7. We examined different values of  $K$  and components number and the performance difference was not significant.

### 3.2 Results on Real Data

The previous section shows how artificially distorted gaze points can be transformed to their correct locations with sufficient accuracy using the K-closest points. In this section, we use the aforementioned collected data to automatically calibrate the gaze estimator and find the gaze points from the videos acquired from a web camera. We apply the K-closest points and mixture model methods (Sects. 2.1.1 and 2.1.2) to find the initial gaze points.

For the eye templates method, 25 eye templates were captured while a subject was fixating their eyes at 25 points on a 21.5 inch display. This process is followed once for a single (template) subject. Therefore, reconstructing an eye image of a new subject from the eye templates will not be optimal due to the changes in eye appearance between the template subject and the other subjects. However, we assume that it still gives a good representation of the topology of the gaze points. As in Lu et al. (2011) we divide the eye image into a  $5 \times 3$  grid and sum up the intensity of the pixel inside each grid cell. The resulting 15 values constitute the feature vector of the eye image.

Regarding the 2D manifold method, a template subject was asked to look at random points on the screen while his face was video recorded. The eye images are cropped and their feature vectors are computed as previously explained.

**Table 1** Accuracies over different methods and template gaze pattern sets

	Template gaze patterns from Judd et al. (2009)		Template gaze patterns from our data	
	KCP	GMM	KCP	GMM
Eye templates	$4.6^\circ$	$4.6^\circ$	$4.7^\circ$	$4.7^\circ$
2D manifold	$4.2^\circ$	$4.3^\circ$	$4.4^\circ$	$4.5^\circ$

KCP denotes K-closest points method, GMM refers to Gaussian mixture model fitting. The best accuracy is yielded using 2D manifold and K-closest points

Then, the feature vectors are projected on the first two principal components to constitute a 2D-manifold. The eye images of a new subject (while looking at a stimulus) are cropped. Then the feature vectors are extracted and projected on the same manifold to determine their relative locations. The distances between the initial gaze points are significantly larger than the actual corresponding gaze points. Yet, this will not affect the results as the initial gaze points will be scaled down, while finding the mapping, to match the initial gaze points with the template gaze patterns.

We select the gaze template patterns in two ways: First, we use the fixation points provided by the eye tracking dataset (Judd et al. 2009). Second, the ground truth of our collected data (via the Tobii gaze estimator) is used. In the second case, for each subject, we consider the gaze points of the other subjects as template gaze patterns. The K-closest points and the mixture model fitting methods are applied to the initial gaze points. Table 1 shows the results.

The results show that the K-closest points method achieves lower error than using the mixture model method while 2D manifold outperforms eye templates for both template gaze pattern sets. The best accuracy ( $4.2^\circ$ ) is obtained using the K-closest points and 2D manifold. Table 2 shows the results per subject/stimulus. Figure 5 shows the results for the first four images with subjects 3 and 7.

Regarding the template gaze patterns, the accuracies are similar for both sets (the eye tracking dataset of Judd et al. (2009) and our collected dataset) with a slight improvement when using the gaze patterns from Judd et al. (2009) dataset. The template gaze pattern sets were collected in two different experiments on two different groups of subjects. This is interesting as it shows the general similarity of gaze patterns and hence suggests the validity of using them in auto-calibration regardless of the viewers. The gaze estimation accuracies vary for different subjects. The relatively lower accuracies for some subjects might be either due to errors in estimating the initial gaze points, i.e. because of eye appearance variations with the template subject eye templates which leads to incorrect initialization, or because of the gaze behavior of the subjects and its variation with the template gaze patterns.



**Table 2** Accuracies of the gaze estimation auto-calibrated using K-closest points and 2D manifold

	Stim. 1	Stim. 2	Stim. 3	Stim. 4	Stim. 5	Stim. 6	Stim. 7	Stim. 8	Stim. 9	Stim. 10	Average (Std-Dev)
Subject 1	4.8°	3.1°	2.1°	2.7°	6.3°	5.3°	4.9°	6.7°	6.4°	4.5°	4.7° (± 1.6)
Subject 2	4.7°	2.1°	3.6°	2.1°	4.1°	3.8°	3.7°	5.9°	5.5°	4.8°	4.0° (± 1.3)
Subject 3	4.4°	2.9°	1.8°	2.2°	3.6°	3.8°	3.4°	5.0°	5.3°	6.6°	3.9° (± 1.5)
Subject 4	3.7°	2.3°	2.0°	2.8°	2.1°	2.4°	3.6°	6.2°	5.2°	6.7°	3.7° (± 1.7)
Subject 5	5.5°	2.9°	2.8°	2.6°	3.4°	3.2°	3.6°	6.1°	4.6°	5.7°	4.0° (± 1.3)
Subject 6	3.9°	3.0°	1.6°	3.9°	2.9°	3.5°	4.6°	5.1°	6.5°	5.3°	4.0° (± 1.3)
Subject 7	4.2°	3.7°	3.1°	3.2°	3.5°	4.7°	5.2°	6.3°	7.7°	6.1°	4.8° (± 1.5)
Subject 8	3.5°	3.1°	3.6°	5.0°	5.0°	5.3°	4.9°	5.4°	5.0°	4.0°	4.5° (± 0.8)
Subject 9	3.8°	2.6°	2.7°	4.0°	4.4°	3.6°	5.5°	5.7°	5.5°	4.1°	4.2° (± 1.1)
Subject 10	4.4°	3.3°	3.8°	4.2°	3.3°	4.7°	4.6°	6.0°	6.6°	4.8°	4.6° (± 1.1)
Average (Std-Dev)	4.3° (± 0.6)	2.9° (± 0.5)	2.7° (± 0.8)	3.3° (± 1.0)	3.9° (± 1.2)	4.0° (± 0.9)	4.4° (± 0.7)	5.8° (± 0.5)	5.8° (± 0.9)	5.3° (± 1.0)	4.2° (± 1.3)

The accuracies are shown per subject/stimulus

The stimuli set contains landscape and street view images, which makes the auto-calibration more challenging than images with clearly salient objects that humans usually focus on. Yet, the reported accuracy (4.2°) and the results in Fig. 5 show the validity of our approach.

Finally, we conduct an experiment to investigate the influence of the number of subjects on the calibration error. More specifically, we vary the number of template gaze patterns starting from 1 and increasing by 2 to reach to 15 patterns (corresponding to the 15 subjects in the dataset (Judd et al. 2009)). The results are shown in Fig. 6. As expected, the more template gaze patterns (and users) the lesser the error. Interestingly, with only one template gaze pattern, the resulting error is 4.8° which can be still valuable for attention estimation.

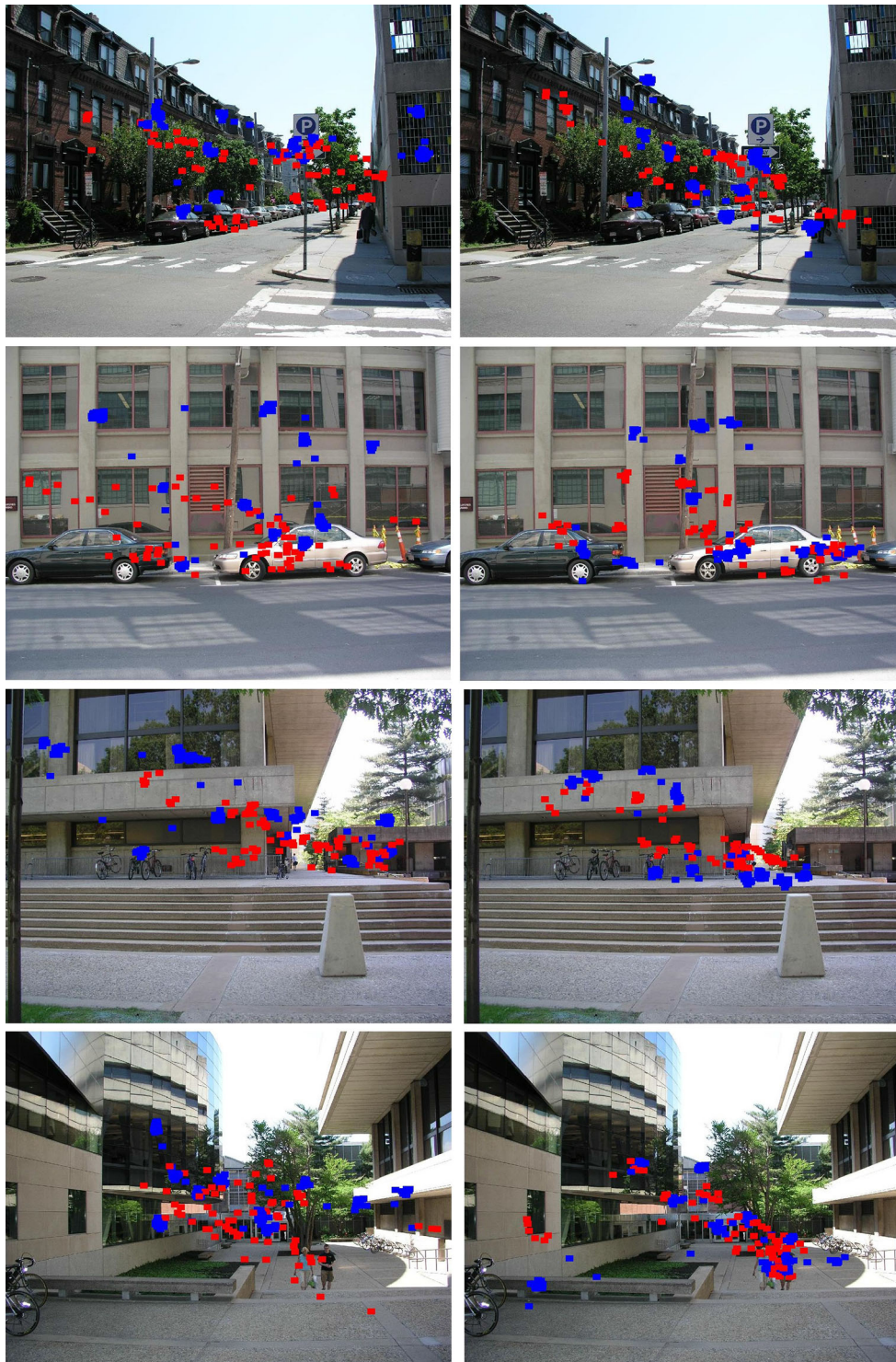
### 3.3 Gaze Estimation Error versus Image Content

The primary observation behind our method is the similarities between the gaze patterns of different viewers when looking at the same stimulus (Judd et al. 2009). These patterns may be influenced by the complexity or the scattering of the stimulus. This may, consequently, affect the gaze estimation error. In this section, we look further into the relationship between the image complexity and the auto-calibration performance (measured by the gaze estimation error).

A number of approaches are proposed to model image statistics (i.e. complexity) (Geusebroek and Smeulders 2005; Scholte et al. 2009; Torralba and Oliva 2003). Here, first, we follow the approach of Geusebroek and Smeulders (2005) and Scholte et al. (2009) by fitting a Weibull distribution to the contrast values of the stimulus image. The Weibull distribution is defined as:

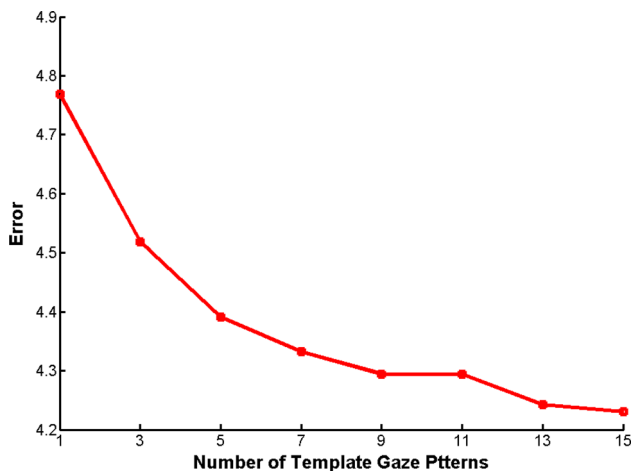
$$f(x) = C \exp\left(-\left|\frac{x - \mu}{\beta}\right|^\gamma\right). \tag{20}$$

where  $C$  is a normalization constant and  $\mu$ ,  $\beta$ , and  $\gamma$  are the parameters of the Weibull distribution corresponding to the location, the scale, and the shape respectively.  $\beta$  and  $\gamma$  indicate some perceptual characteristics of the image such as regularity, coarseness, roughness, and contract (Geusebroek and Smeulders 2005; Tamura et al. 1978). Geusebroek and Smeulders (2005) and Scholte et al. (2009) found that images which correspond to low values of beta and gamma represent isolated objects in a plain background while their content changes gradually to contain multiple objects with higher values of beta and gamma. This suggests that image complexity can be characterized by the Weibull parameters. More importantly, the Weibull parameters, gamma and beta, highly correlate with neural responses in the early visual system (Scholte et al. 2009).



**Fig. 5** Gaze estimation results for the first four images with subjects 3 (*right column*) and 7 (*left column*). The *red traces* represent the estimated gaze points while the *blue traces* represent the ground truth

obtained from the Tobii gaze estimator. The results are achieved using 2D-manifold and K-closest points (Color figure online)



**Fig. 6** Gaze estimation error versus number of template gaze patterns

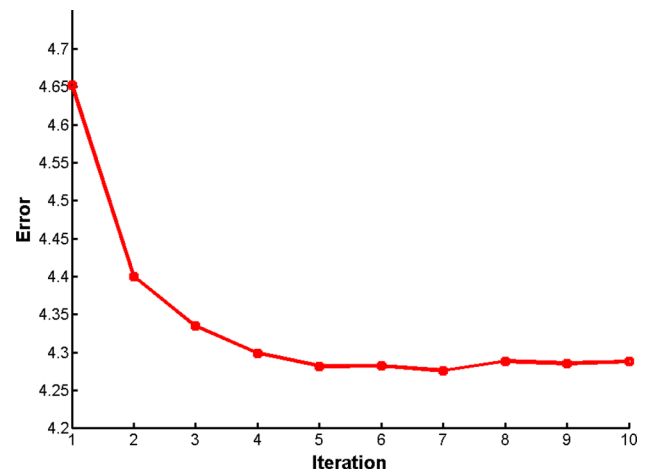
Another approach is measuring the “density” of objects in the scene. Our rationale here is that the more objects the scene contains, the more complex the image content is. Since we are not interested in particular objects, we use general-purpose object proposal methods (e.g. Selective Search Uijlings et al. 2013).

To evaluate the relationship between the image complexity and the gaze estimation error, we compute the correlation between the predicted gaze error and the corresponding beta and gamma parameters of a stimulus. Since we do not have the estimation error for all 1003 stimuli, we artificially distort the ground truth by random transformations (translation and scaling) and applying the mapping algorithm to transform the gaze points back to their correct positions (similar to the experiment in Sect. 3.1). The gaze estimation error, here, is associated with only the gaze mapping error (i.e. feature extraction and initialization error is not applied) which makes the correlation more indicative. We select the 50 samples with highest errors and the 50 samples with the lowest errors to show how the error correlate with the image complexity.

Using the object proposal as a complexity measure, the results do not indicate a significant correlation ( $r < 0.1$ ) between the calibration error and the complexity of the image. When using Weibull distribution fitting, with  $\alpha = 0.01$ , the results show a correlation of ( $r = 0.3$ ) which indicates an effect of image complexity (characterized by Weibull parameters as in Scholte et al. 2009; Geusebroek and Smeulders 2005) and the auto-calibration performance and, hence, the gaze estimation error.

### 3.4 Gaze Patterns versus Saliency Maps

Our approach consists primarily of two parts: first, finding a topology of the gaze points and, second, mapping this point topology to the gaze patterns of other users. In this section, we compare the gaze patterns with the computational saliency as



**Fig. 7** Gaze estimation error after incrementally updating the template gaze patterns with estimated gaze points. Saliency information (Harel et al. 2006) is used as a starter gaze pattern

a cue source for mapping the point topology. To this end, we simply substitute the gaze patterns with saliency information (Harel et al. 2006). The results show an error of  $4.7^\circ$  when using the saliency information compared to  $4.2^\circ$  when using the gaze patterns of other users.

The previous results show the advantage of using the gaze patterns over the saliency maps. An intuitive step is to combine the saliency information with the gaze patterns of other users. However, here, we relax some of the prerequisites and assume that the accurate gaze patterns of other users are unavailable and, instead, the estimated ones are utilized and modified gradually. The assumption is that even if the gaze points are estimated, they still provide some cues for other uncalibrated gaze points. More specifically, the method starts with saliency information (Harel et al. 2006) as the first template gaze pattern. For each new user, the gaze points are estimated and added to the template gaze patterns. After a certain number of users, since the template gaze patterns are modified, the gaze points of all users are re-estimated. Hence, the accuracy is improving gradually. In the experiments, the gaze points are re-estimated after 10 users. This process is repeated a number of times. The gaze estimation error over the iterations is plotted in Fig. 7. Table 3 shows the errors after 10 iterations per user/stimulus. The results show that the error decreases gradually when adding or updating estimated gaze points. After 10 iterations, the error decreases from  $4.7^\circ$  to  $4.3^\circ$ . This suggests the advantage of adding “estimated” gaze points to the saliency information.

### 3.5 Initialization Error versus Mapping Error

The previous experiments show how our method achieves, using visual features, an error of  $4.2^\circ$  without any kind of active calibration. Early steps of cropping the eye regions

**Table 3** Accuracies of the gaze estimation using *uncalibrated* gaze patterns after 10 iterations

	Stim. 1	Stim. 2	Stim. 3	Stim. 4	Stim. 5	Stim. 6	Stim. 7	Stim. 8	Stim. 9	Stim. 10	Average (Std-Dev)
Subject 1	4.5°	2.6°	2.8°	5.2°	5.2°	5.0°	3.5°	6.5°	5.0°	4.0°	4.4° (± 1.2)
Subject 2	4.4°	2.9°	3.9°	4.2°	4.3°	3.8°	4.0°	5.3°	4.9°	4.5°	4.2° (± 0.7)
Subject 3	3.7°	3.4°	2.0°	4.5°	3.6°	3.6°	3.4°	5.1°	4.8°	5.7°	4.0° (± 1.1)
Subject 4	4.2°	2.9°	2.8°	3.3°	3.1°	2.1°	3.3°	5.3°	3.7°	6.5°	3.7° (± 1.3)
Subject 5	5.0°	4.2°	3.5°	3.1°	4.5°	3.0°	3.9°	5.5°	4.5°	5.3°	4.3° (± 0.9)
Subject 6	2.8°	3.1°	2.0°	6.0°	3.7°	3.3°	3.6°	4.9°	5.4°	5.3°	4.0° (± 1.3)
Subject 7	4.5°	3.4°	3.5°	3.2°	4.1°	4.7°	5.2°	6.6°	6.6°	5.5°	4.7° (± 1.3)
Subject 8	4.0°	3.3°	3.6°	5.4°	4.9°	5.3°	4.6°	4.2°	4.0°	3.2°	4.2° (± 0.8)
Subject 9	4.4°	3.1°	3.3°	5.5°	5.4°	3.6°	4.0°	7.4°	5.4°	4.1°	4.6° (± 1.3)
Subject 10	4.5°	3.6°	3.7°	4.8°	3.5°	4.8°	4.3°	5.9°	6.0°	4.4°	4.5° (± 0.9)
Average (Std-Dev)	4.2° (± 0.6)	3.5° (± 0.5)	3.1° (± 0.7)	4.5° (± 1.0)	4.2° (± 0.8)	3.9° (± 1.0)	4.0° (± 0.6)	5.7° (± 0.9)	5.0° (± 0.9)	4.9° (± 1.0)	4.3° (± 1.1)

The accuracies are shown per subject/stimulus

and extracting the visual features are likely to introduce some noise which propagates to later steps and, consequently, contributes to the final error. In this experiment, we aim to give analysis of the contributions of initialization and optimization steps to the overall gaze estimation error. To this end, we use the relative movements of eye centers provided by Tobii's infrared diodes as initial gaze points for the new users. Note that these are not the ground truth gaze points provided by the actively-calibrated Tobii system but just the changes of eye center positions measured by the infrared diodes. Since the infrared diodes produce more accurate and stable measurements than an RGB webcam (especially when the head moves slightly during recording), we assume that such measurements alleviate the influence of initial gaze points estimation error. Please note that this is different from the experiment in Sect. 3.1, where the ground truth is distorted and realigned by the auto-calibration method. In this experiment, the measurements are more robust than the ones obtained by the RGB webcam, however, they are still prone to some form of noise. Using K-closest points, the error drops to 3.1°. The results show that part of the gaze estimation error is attributed to noise in the feature extraction step (and hence the initial gaze point estimation).

### 3.6 Temporal Gaze Patterns

Besides the spacial cues the gaze patterns provide us to auto-calibrate gaze estimation systems, here, we argue that temporal information can further improve the calibration and hence the performance. Therefore, as a proof of concept, we employ this information to correct the calibrated (estimated) gaze patterns. To this end, for each subject-stimulus test case, we learn the temporal patterns from the template gaze patterns and use them to update the predicted gaze points. More specifically, our training data is the sampled subsequences of  $T$  points from all the template gaze patterns, denoted as  $SS$ . For the newly estimated gaze sequence, for each subsequence,  $ss$ , we find the closest subsequence  $ss^{closest}$  from  $SS$  and update the next point with a factor of  $\gamma$ :

$$ss_T^{new} = \gamma ss_T^{old} + (1 - \gamma) ss_T^{closest} \quad (21)$$

We set  $T$  to 5 and  $\gamma$  to 0.5.  $ss_T$  is the last point of  $ss$ . The closest subsequence is selected using straightforward KNN. The overall error drops to 4.1° which suggests the impact of temporal information.

### 3.7 Comparison to the State-of-the-Art

We compare our method with existing state-of-the-art auto-calibration approaches. The recent work of [Chen and Ji](#)

(2011) uses a single camera with multiple infrared lights to reconstruct the 3D eye model. They use saliency information to estimate the angle between the visual and optical axes. The authors reported less than  $3^\circ$  error using five images and five subjects. Clearly, the comparison with this method is not feasible as the authors use different equipment to reconstruct an accurate 3D eye model.

Sugano et al. (2013) adopt an appearance-based gaze estimator and use visual saliency for auto-calibration. The authors reported an error of  $3.5^\circ$ . However, their experimental setup differs from ours in the following aspects: First, a chin rest is used in Sugano et al. (2013) to fixate the head during the experiment while the subjects in our experiment do not use any tool to fixate their heads. Second, the authors in Sugano et al. (2013) ask the subjects to look at a number of 30 s videos for training (5–20 videos), while in our method the subject needs to look at a single image for 3 s. Images contain less cues than videos in which moving objects attract the viewers attention. However, experimenting on still images is more natural and requiring motion in the scene limits the applicability of the gaze estimator. Finally, Sugano et al. analyze the performance variations with respect to different number of training videos. When training on 5 videos (each lasts 30 s), the average error is about  $5.2^\circ$  (the exact accuracy is not reported as the results are plotted on a graph). While our method achieves an average error of  $4.2^\circ$  by looking at a single image for 3 s.

## 4 Discussion

Our method provides sufficient accuracy to predict the areas of attention with a flexible setup and a webcam. This is especially important for tasks where gaze estimation is required with no active participation from the user and using off-the-shelf hardware. In this work, we propose a flexible setup and use low-cost publicly available web cameras. There is a trend nowadays to use eye gaze estimation for electronic consumer relationship marketing which aims to employ information technology to understand and fulfill the needs of the consumers (Wedel and Pieters 2008). These applications usually collect the data passively without active user participation. Our method is suitable for such applications. Tracing consumers attention when shopping in malls or when exploring advertisements on their laptops are examples of use.

When compared with saliency information, the gaze patterns of other users produce lower error. We further relax the prerequisite of having accurate gaze patterns and substitute them with estimated gaze points of the subsequent subjects. By gradually updating the template gaze patterns (using the estimated gaze points), the gaze estimator is gradually auto-calibrated and the accuracy improves.

## 5 Conclusion

We presented a novel method to auto-calibrate gaze estimators in an uncalibrated setup. Based on the observation that humans produce similar gaze patterns when looking at a stimulus, we use the gaze patterns of individuals to estimate the gaze points for new viewers without active calibration.

The proposed method was tested in a flexible setup using a web camera without a chin rest. To estimate the gaze points, the viewer needs to look at an image for only 3 s without any explicit participation in the calibration. Evaluated on 10 subjects and 10 images showing landscapes and street views, the proposed method achieves an error of  $4.2^\circ$ . A number of experiments were conducted to give further insight into the method and its contribution in different cases. When relaxing some prerequisites and using estimated gaze points (compared to accurate gaze patterns), the gaze estimator was auto-calibrated gradually and so the accuracy improved. The estimation error was comparable to the one with accurate gaze patterns. Experiments show that the heterogeneity between the gaze patterns of the viewers has an impact on the auto-calibration error. To the best of our knowledge, this is the first work to use human gaze patterns in order to auto-calibrate gaze estimators.

**Acknowledgements** This research is supported by the Dutch national program COMMIT.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Alnajar, F., Gevers, T., Valenti, R., & Ghebreab, S. (2013). Calibration-free gaze estimation using human gaze patterns. In *IEEE international conference on computer vision (ICCV)* (pp. 137–144).
- Chen, J., & Ji, Q. (2011). Probabilistic gaze estimation without active personal calibration. In *IEEE computer vision and pattern recognition (CVPR)* (pp. 609–616).
- Draeos, M., Qiu, Q., Bronstein, A., & Sapiro, G. (2015). Intel realsense=real low cost gaze. In *International conference on image processing*.
- Geusebroek, J.-M., & Smeulders, A. W. M. (2005). A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62, 7–16.
- Guestrin, E. D., & Eizenman, M. (2006). General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6), 1124–1133.
- Guestrin, E. D., & Eizenman, M. (2008). Remote point-of-gaze estimation requiring a single-point calibration for applications with infants. In *Symposium on eye tracking research and applications (ETRA)* (pp. 267–274).

- Hansen, D. W., Hansen, J. P., Nielsen, M., Johansen, A. S., & Stegmann, M. B. (2002). Eye typing using Markov and active appearance models. In *Sixth IEEE workshop on applications of computer vision* (pp. 132–136).
- Hansen, D. W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 478–500.
- Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems (NIPS)* (pp. 545–552).
- Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE international conference on computer vision (ICCV)* (pp. 2106–2113).
- Lu, F., Sugano, Y., Okabe, T., & Sato, Y. (2011). Inferring human gaze from appearance via adaptive linear regression. In *IEEE international conference on computer vision (ICCV)* (pp. 153–160).
- Majaranta, P., & Rih, K.-J. (2002). Twenty years of eye typing: Systems and design issues. In *Symposium on eye tracking research and applications (ETRA)* (pp. 15–22).
- Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 229, 2323–2326.
- Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2005). Labelme: A database and web-based tool for image annotation. In *MIT AI Lab Memo AIM-2005-025, MIT CSAIL*.
- Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., & Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, 9, 1–9.
- Smith, K., Ba, S. O., Odobez, J., & Gatica-Perez, D. (2008). Tracking the visual focus of attention for a varying number of wandering people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7), 1212–1229.
- Sugano, Y., Matsushita, Y., & Sato, Y. (2010). Calibration-free gaze sensing using saliency maps. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2667–2674).
- Sugano, Y., Matsushita, Y., & Sato, Y. (2013). Appearance-based gaze estimation using visual saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2), 329–341.
- Sugano, Y., Matsushita, Y., & Sato, Y. (2014). Learning-by-synthesis for appearance-based 3D gaze estimation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1821–1828).
- Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Systems, Man and Cybernetics*, 8, 460–473.
- Tan, K., Kriegman, D., & Ahuja, N. (2002). Appearance-based eye gaze estimation. In *Applications of computer vision* (pp. 191–195).
- Tobii Technology: <http://www.tobii.com/>.
- Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, 14, 391–412.
- Uijlings, J. R. R., Van De Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International journal of computer vision*, 104, 154–171.
- Valenti, R., & Gevers, T. (2012). Accurate eye center location through invariant isocentric patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9), 1785–1798.
- Villanueva, A., Cabeza, R., & Porta, S. (2006). Eye tracking: Pupil orientation geometrical modeling. *Image and Vision Computing*, 24(7), 663–679.
- Wedel, M., & Pieters, R. (2008). *A review of eye-tracking research in marketing. Review of marketing research*. New York: Emerald Group Publishing Limited.
- Xiong, C., Huang, L., & Liu, C. (2015). Remote gaze estimation based on 3D face structure and iris centers under natural light. *Multimedia Tools and Applications*, 75, 1–15.
- Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2879–2886).