Calibration-free gaze sensing using saliency maps

Yusuke Sugano
The University of Tokyo
Tokyo, Japan, 153-8505
sugano@iis.u-tokyo.ac.jp

Yasuyuki Matsushita
Microsoft Research Asia
Beijing, P. R. China, 100080
yasumat@microsoft.com

Yoichi Sato
The University of Tokyo
Tokyo, Japan, 153-8505
ysato@iis.u-tokyo.ac.jp

Abstract

We propose a calibration-free gaze sensing method using visual saliency maps. Our goal is to construct a gaze estimator only using eye images captured from a person watching a video clip. The key is treating saliency maps of the video frames as probability distributions of gaze points. To efficiently identify gaze points from saliency maps, we aggregate saliency maps based on the similarity of eye appearances. We establish mapping between eye images to gaze points by Gaussian process regression. The experimental result shows that the proposed method works well with different people and video clips and achieves 6 degrees of accuracy, which is useful for estimating a person's attention on monitors.

1. Introduction

Gaze estimation is important for predicting human attention, and therefore can be used for various interactive systems. There are a wide range of applications of gaze estimation including marketing analysis of online content and digital signage, gaze-driven interactive displays, and many other human-machine interfaces.

In general, gaze estimation is achieved by analyzing a person's eyes with an image sensor. Exact gaze points can be determined by directly analyzing gaze directions from observations of eyes. Many implementations of camera-based gaze estimator have been proposed including commercial products (see [3] for a recent survey). One of the limitations of camera-based gaze estimators is explicit calibration for learning person-dependent parameters. Although the number of reference points for calibration can be reduced using multiple light sources [18], or stereo cameras [12], it still requires a user to actively participate in the calibration task. In some practical scenarios, the active calibration is too restrictive because it interrupts natural interactions and makes the unnoticeable gaze estimation impossible.

To avoid active calibration, Yamazoe *et al*. used a simple eyeball model for gaze estimation and performed automatic calibration by fitting the model to appearance of a user's eye [19]. Sugano *et al*. proposed a method using input from a user's mouse as exemplar data for calibration [17]. However, these approaches are restricted to specific scenarios. Yamazoe *et al*.'s approach relies on a specific geometric model, and Sugano *et al*.'s approach can only be applied to interactive environments with user inputs.

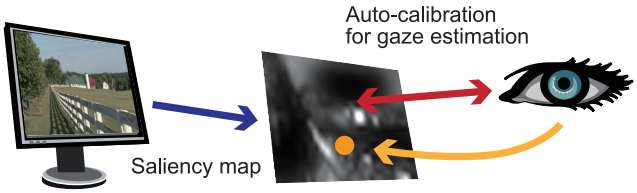

Figure 1. Illustration of our method. Our method uses saliency maps computed from video frames in bottom-up manner for automatically constructing a gaze estimator.

Apart from these gaze estimation studies, computational models of visual saliency have been studied to analyze visual attention on an image. While gaze estimation approaches aim at determining where people's eyes actually look at, the visual saliency give us information about which image region attracts more attention, as illustrated in Figure 1. Biologically, humans tend to gaze at an image region with high saliency, *i.e*., a region containing more unique and distinctive visual features compared with the surrounding regions. Hence, by knowing the visual saliency map of an image, the gaze point of a person looking at an image can be predicted. After Koch and Ullman proposed the original concept [11] of visual saliency, many bottom-up computational models of visual saliency maps have been proposed [7, 15, 5, 4, 1]. It is experimentally shown that there indeed exists a correlation between bottom-up visual saliency and fixation locations [14].

Gaze estimation and visual saliency models are closely related; nonetheless, not many previous studies relate these two. Kienzle *et al*. [10, 9] proposed a method for learning computational models of bottom-up visual saliency using gaze estimation data. Judd *et al*. [8] followed the approach

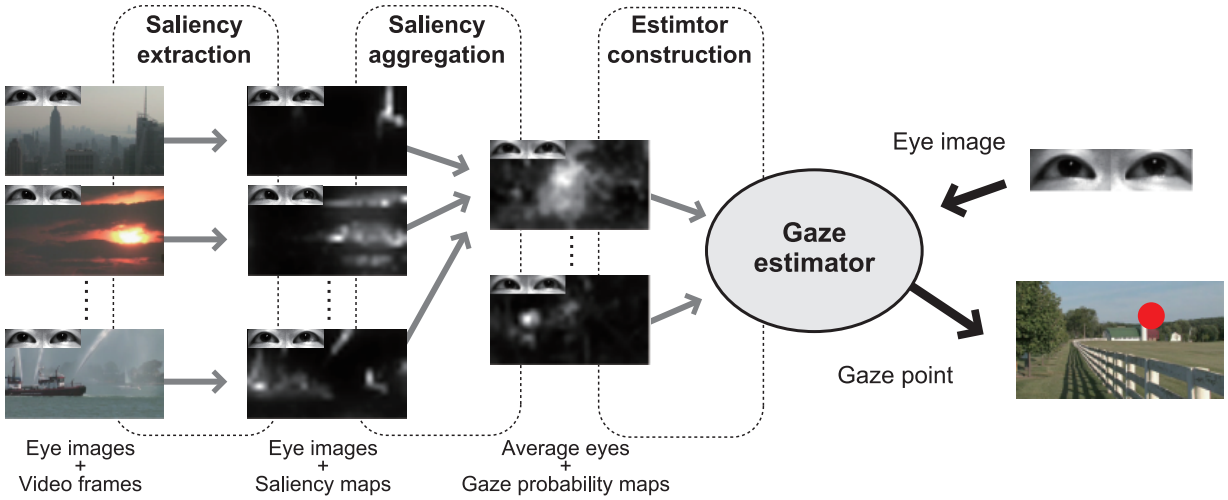

Figure 2. Proposed framework. Our method consists of three steps. Saliency extraction step computes saliency maps from the input video. Saliency aggregation step combines saliency maps to produce gaze probability maps. Using the gaze probability maps and associated eye images, the estimator construction step learns the mapping from an eye image to a gaze point.

with more features and a larger database. These approaches learn accurate saliency models using gaze points. In contrast to these methods, our goal is to construct a gaze estimator from saliency maps. To our knowledge, this is the first work to use visual saliency as prior information for gaze estimation.

We propose a novel calibration-free gaze sensing method using computational visual saliency. The key idea of our method is generating a probability distribution of gaze points using saliency maps. From eye images of a user watching a video clip, we acquire learning datasets that consists of saliency maps and eye images under a fixed head position. Gaze probability maps are generated by aggregating the saliency maps based on the similarity of eye appearances. Once the gaze probability maps are obtained, our method learns the relationship between the gaze probability maps and eye images. As a result, this leads to a completely ambient gaze estimator that exempts users from active calibration.

2. Gaze estimation from saliency maps

Our goal is to construct a gaze estimator without any explicit calibration stages. The inputs for our system are N video frames $\{\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N\}$ and associated eye images $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N\}$ of a person watching a video clip under a fixed head position. In our setting, eye images and video frames are synchronized; eye image $\boldsymbol{e}_i$ is captured at the same time when frame $\boldsymbol{I}_i$ is shown to the person. Using this dataset $\{(\boldsymbol{I}_1, \boldsymbol{e}_1), \ldots, (\boldsymbol{I}_N, \boldsymbol{e}_N)\}$, our goal is to construct a gaze estimator for estimating an unknown gaze point $\boldsymbol{g}$ from an eye image $\boldsymbol{e}$.

Our method consists of three steps; saliency extraction, saliency aggregation, and estimator construction as shown in Figure 2. Saliency extraction, is the step in which saliency maps from an input video are calculated. From the video clips, a visual saliency map that represents distinctive visual features is extracted from each frame. Saliency aggregation combines all saliency maps to obtain a gaze probability map that has a peak around the true gaze point. This step produces pairs of the average eye image and gaze probability map. The third step is the estimator construction. Using the gaze probability maps and associated eye images, the estimator construction step learns the mapping from an eye image to a gaze point. The resulting gaze estimator outputs gaze points for any eye image from the person. Details of each step are described in the following sections.

2.1. Saliency extraction

This step extracts visual saliency maps $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ from input video frames $\{\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N\}$. We use graph-based visual saliency [4] as a base saliency model. In our method, we use commonly used feature channels, *i.e.*, color, intensity, and orientations as static features, and flicker and motion are used as dynamic features. These features are detailed in, *e.g.*, [6]. All feature maps are combined with the same weight to form a low-level saliency map $\boldsymbol{s}^l$.

On top of the low-level saliency, we use a higher level saliency model. Humans tend to fixate on faces, especially the eyes, which are highly salient for humans. Cerf *et al*. [2] proposed a face channel-based saliency model using a face detector. We follow this approach to produce reliable saliency maps using a facial feature detector (OKAO Vision library developed by OMRON Corporation). The face channel saliency map $\boldsymbol{s}^f$ is modeled as 2-D Gaussian circles with a fixed variance at the detected positions of the center between two eyes. When the detector detects only a

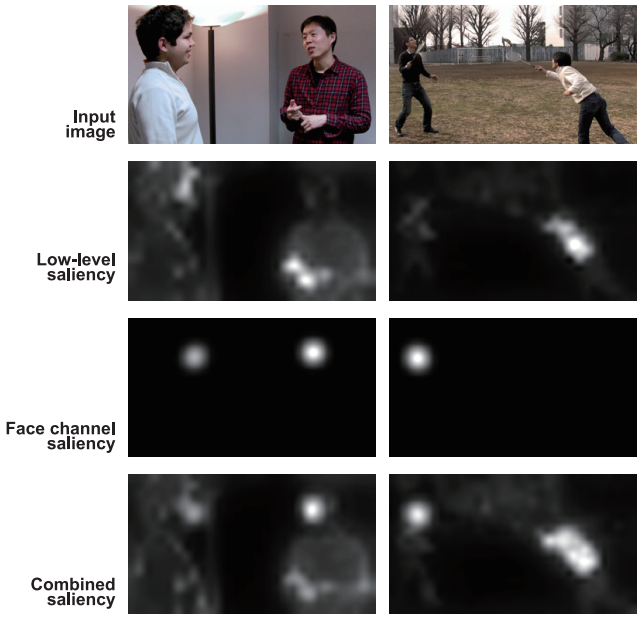

Figure 3. Examples of computed saliency maps. First row shows input images $\boldsymbol{I}$, second row shows corresponding low-level saliency $\boldsymbol{s}^l$, third row corresponds to face channel saliency $\boldsymbol{s}^f$, and the bottom row shows combined saliency maps $\boldsymbol{s}$.

face, *e.g.*, when the face region is small, the facial saliency is defined at the center of the facial region.

We combine the low-level saliency $\boldsymbol{s}^l$ and face channel saliency $\boldsymbol{s}^f$ after normalizing them to span in the same range. We also take the temporal average of the aggregated saliency maps to compute the final saliency map $\boldsymbol{s}$ as

$$\boldsymbol{s}_i = \frac{1}{2(n_s+1)} \sum_{j=i-n_s}^{i} (\boldsymbol{s}_j^l + \boldsymbol{s}_j^f), \tag{1}$$

where n_s is the number of frames used for temporal smoothing. Since humans cannot instantly follow rapid scene changes, only past frames are used for the smoothing to account for latency. As a result, synchronized pairs of saliency maps and eye images $\mathcal{D}_s = \{(\boldsymbol{s}_1, \boldsymbol{e}_1), \ldots, (\boldsymbol{s}_N, \boldsymbol{e}_N)\}$ are produced.

Figure 3 shows examples of the computed saliency maps. From top to bottom are input images $\boldsymbol{I}$, low-level saliency $\boldsymbol{s}^l$, face channel saliency $\boldsymbol{s}^f$, and combined saliency maps $\boldsymbol{s}$.

2.2. Saliency aggregation

Although saliency maps extracted in the previous step accurately predict gaze points, the accuracy is still not good enough to determine exact locations of gaze points.

The saliency maps $\boldsymbol{s}$ encode distinctive visual features in a video frame. While a saliency map does not provide exact gaze points, highly salient regions in a saliency map are likely to coincide with the true gaze point. Suppose we have a set of saliency maps that statistically have high saliency scores around the true gaze point, with random saliency scores at other regions. By aggregating the saliency maps, it is expected that the image region around the true gaze point has a vivid peak of saliency. The map can be used as the probability distribution of the gaze point. This step aims at producing such probability maps using the associated eye images.

Our basic idea is to use a similarity of eye images for the aggregation. The similarity measure w_s is defined as

$$w_s(\boldsymbol{e}_i, \boldsymbol{e}_j) = \exp(-\kappa_s ||\boldsymbol{e}_i - \boldsymbol{e}_j||^2). \tag{2}$$

When the gaze points of eye images $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are close, the appearances are similar and w_s becomes high.

In this step, we first eliminate unreliable eye images, *e.g.*, images during blinking, from the learning set. Eye images recorded during fixation are useful as a learning data. To identify such eye images, we use a fixation measure of an eye image $\boldsymbol{e}$ defined as

$$w_f(\boldsymbol{e}_i) = \exp(-\kappa_f \mathrm{Var}(\boldsymbol{e}_i)), \tag{3}$$

where $\mathrm{Var}(\boldsymbol{e}_i)$ denotes the variance of eye images $\{\boldsymbol{e}_{i-n_f}, \ldots, \boldsymbol{e}_{i+n_f}\}$ averaged over pixels. Since appearances of the eye images change rapidly during fast movement, w_f becomes lower when $\boldsymbol{e}_i$ is captured during eye movement or blinking. A subset $\mathcal{D}_{s'} = \{(\boldsymbol{s}_1, \boldsymbol{e}_1), \ldots, (\boldsymbol{s}_{N'}, \boldsymbol{e}_{N'})\}$ is created from $\mathcal{D}_s$ by removing eye images where w_f scores are lower than a predefined threshold τ_f.

Since variation in the gaze points is limited in $\mathcal{D}_{s'}$, and there can be many samples that share almost the same gaze point, eye images are clustered according to similarity w_s to reduce redundancy and computational cost. Each eye image $\boldsymbol{e}$ is sequentially added to the cluster whose average eye image $\bar{\boldsymbol{e}}$ is the most similar to $\boldsymbol{e}$. A new cluster is adaptively created if the highest similarity among all existing clusters is lower than the threshold τ_s. Finally, M clusters and their average eye images $\{\bar{\boldsymbol{e}}_1, \ldots, \bar{\boldsymbol{e}}_M\}$ are calculated.

After these steps, we compute the gaze probability map $\bar{\boldsymbol{p}}_i$ as

$$\bar{\boldsymbol{p}}_i = \frac{\sum_j^{N'} w_s(\bar{\boldsymbol{e}}_i, \boldsymbol{e}_j)(\boldsymbol{s}_j - \bar{\boldsymbol{s}}_{\mathrm{all}})}{\sum_j^{N'} w_s(\bar{\boldsymbol{e}}_i, \boldsymbol{e}_j)}, \tag{4}$$

where $\bar{\boldsymbol{s}}_{\mathrm{all}}$ is the average of saliency maps in $\mathcal{D}_{s'}$. Man-made pictures usually have higher saliency at the center of the image, Hence, without normalization, the gaze probability map $\bar{\boldsymbol{p}}_i$ tends to have higher value at the center regardless of $\bar{\boldsymbol{e}}_i$. The average saliency map $\bar{\boldsymbol{s}}_{\mathrm{all}}$ is used to eliminate this centering bias in the gaze probability map. Each gaze probability map $\bar{\boldsymbol{p}}_i$ is normalized to a fixed range. Finally, we obtain a dataset $\mathcal{D}_p = \{(\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{e}}_1), \ldots, (\bar{\boldsymbol{p}}_M, \bar{\boldsymbol{e}}_M)\}$.

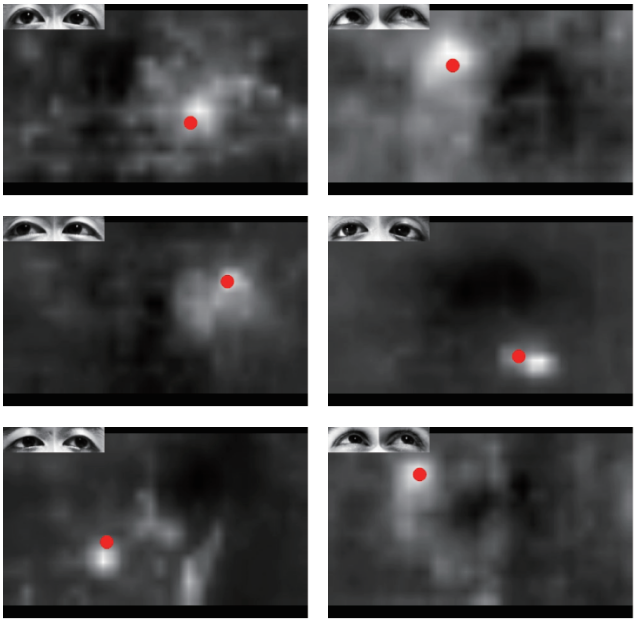

Figure 4. Examples of gaze probability maps $\bar{\boldsymbol{p}}$ and corresponding average eye images $\bar{\boldsymbol{e}}$. Overlaid circles depict true gaze points of $\bar{\boldsymbol{e}}$ to illustrate the correspondence between a gaze point and the peak in the gaze probability. The true gaze points are obtained using a calibration-based gaze estimator, and our method does not know the true gaze points.

Figure 4 shows examples of the obtained gaze probability maps. The eye images shown at the top-left indicate corresponding average eye images $\bar{\boldsymbol{e}}$. The six images are some examples taken from six different people. The overlaid circles indicate true gaze points of $\bar{\boldsymbol{e}}$. The true gaze points are unknown in our method, and these are obtained using a calibration-based gaze estimator and placed as a reference. Although the gaze probability maps $\bar{\boldsymbol{p}}_i$ are generated without knowing true gaze points, these highly correspond to the true gaze points.

Figure 5 shows the improvement of the correlation between the true gaze point and saliency maps by this aggregation step. The curves are drawn by changing saliency threshold values from minimum to maximum. The horizontal axis indicates a false positive rate, *i.e.*, rate of pixels in a map above a threshold. The vertical axis indicates a true positive rate, *i.e.*, rate of frames whose saliency value at the true gaze point is higher than the threshold. This plot is obtained using all data used in our experiment. The thin line shows the average receiver operating characteristic (ROC) curve (area under the curve (AUC) = 0.73) of the extracted saliency maps before aggregation. After aggregation, the accuracy is improved as shown by the bold line in Figure 5, which shows the average ROC curve (AUC = 0.90) of all the gaze probability maps.

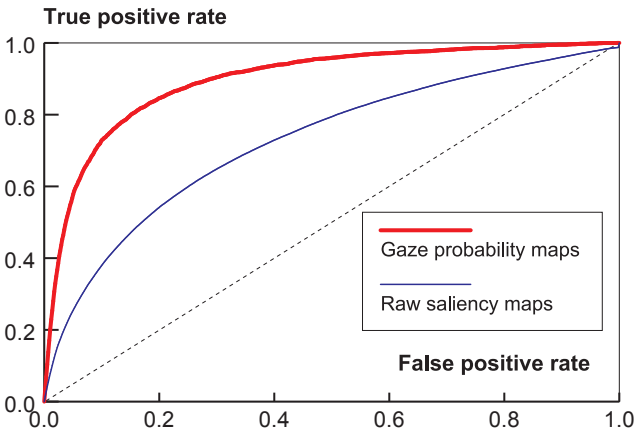

Figure 5. ROC curves of raw saliency maps and gaze probability maps. Horizontal axis indicates the false positive rate, *i.e.*, pixel rate above a threshold. Vertical axis indicates the true positive rate, *i.e.*, rate of frames which have a higher saliency value than a threshold at the true gaze point. Thin line (AUC = 0.73) indicates raw saliency maps extracted through process described in Section 2.1. Bold line (AUC = 0.90) corresponds to the gaze probability maps described in Section 2.2.

2.3. Estimator construction

In the previous step, the average eye images $\{\bar{\boldsymbol{e}}_1, \ldots, \bar{\boldsymbol{e}}_M\}$ and corresponding gaze probability maps $\{\bar{\boldsymbol{p}}_1, \ldots, \bar{\boldsymbol{p}}_M\}$ are produced.

In a standard Gaussian process regression, a model can be built to estimate the probability distribution $P(\boldsymbol{g}^*|\boldsymbol{e}^*, \mathcal{D}_g)$ of an unknown gaze point $\boldsymbol{g}^*$ of an eye image $\boldsymbol{e}^*$, given labeled data points $\mathcal{D}_g = \{(\boldsymbol{g}_1, \bar{\boldsymbol{e}}_1), \ldots, (\boldsymbol{g}_M, \bar{\boldsymbol{e}}_M)\}$. However in our case, we only know $\mathcal{D}_p = \{(\bar{\boldsymbol{p}}_1, \bar{\boldsymbol{e}}_1), \ldots, (\bar{\boldsymbol{p}}_M, \bar{\boldsymbol{e}}_M)\}$ where $\bar{\boldsymbol{p}}_i$ can be treated as a probability map of $\boldsymbol{g}_i$.

Therefore, we re-formulate Gaussian process regression using the probability maps as follows. By normalizing the gaze probability maps, we define probability distributions as

$$P(g|\bar{\boldsymbol{p}}) = \frac{\bar{\boldsymbol{p}}(g)}{\sum_x \sum_y \bar{\boldsymbol{p}}}, \tag{5}$$

where $\bar{\boldsymbol{p}}(g)$ indicates the value of $\bar{\boldsymbol{p}}$ at the gaze point g, and $\sum_x \sum_y \bar{\boldsymbol{p}}$ indicates overall summation of $\bar{\boldsymbol{p}}$. Here, we describe the case of estimating one-dimensional scalar g to simplify the notation; however, two regressors are independently built for each X- and Y-direction. Given Eq. (5), the target distribution $P(g^*|\boldsymbol{e}^*, \mathcal{D}_p)$ can be obtained by marginalizing over all possible gaze points $\{g_1, \ldots, g_M\}$ as

$$P(g^*|\boldsymbol{e}^*, \mathcal{D}_p) = \sum_{g_1} \cdots \sum_{g_M} P(g^*|\boldsymbol{e}^*, \mathcal{D}_g) P(\mathcal{D}_g|\mathcal{D}_p), \tag{6}$$

where

$$P(\mathcal{D}_g|\mathcal{D}_p) = \prod_i^M P(g_i|\bar{\boldsymbol{p}}_i). \quad (7)$$

In Eq. (6), g^* indicates the unknown gaze point of the eye image $\boldsymbol{e}^*$, and g_i is the gaze point corresponding to $\bar{\boldsymbol{e}}_i$.

Because the integral (summation) of Eq. (6) is computationally expensive, we solve Eq. (6) by Monte Carlo approximation. We randomly produce n_g sets of samples $\mathcal{D}_g^{(l)} = \{(g_1^{(l)}, \bar{\boldsymbol{e}}_1), \ldots, (g_M^{(l)}, \bar{\boldsymbol{e}}_M)\}_{l=1}^{n_g}$ according to the probability distribution defined by Eq. (5). Namely, $g_i^{(l)}$ in the l-th set is generated according to the distribution $P(g_i|\bar{\boldsymbol{p}}_i)$ defined by the i-th probability map. Because the gaze probability maps accurately predict true gaze points as shown in Figure 5, the low saliency values from the gaze probability maps are cut off to reduce the number of samples in the approximation. We use a threshold τ_s to set the probability to zero if $\bar{\boldsymbol{p}}(x, y)$ is lower than the threshold. Using these sets, Eq. (6) can be approximated as

$$P(g^*|\boldsymbol{e}^*, \mathcal{D}_p) = \frac{1}{n_g} \sum_{l=1}^{n_g} P(g^*|\boldsymbol{e}^*, \mathcal{D}_g^{(l)}). \quad (8)$$

Finally, each $P(g^*|\boldsymbol{e}^*, \mathcal{D}_g^{(l)})$ can be estimated based on a Gaussian process regression [16].

Gaussian process regression We assume a noisy observation model $g_i = f(\boldsymbol{e}_i) + \epsilon_i$, *i.e.*, a gaze point g_i is given as a function of $\boldsymbol{e}_i$with the data-dependent noise term $\epsilon_i = \mathcal{N}(0, \varsigma_i^2)$. In standard methods, the noise variance ς^2 is treated as an unknown parameter that takes a constant value across all data. In our case, because the sample distribution is known, the noise variance ς_i^2 can be set to an actual variance of generated samples $\{g_i^{(1)}, \ldots, g_i^{(n_g)}\}$. It explicitly assigns a higher noise variance for samples from ambiguous saliency maps with several peaks. $f(\boldsymbol{e}_i)$ is assumed to be a zero-mean Gaussian process with a covariance function k:

$$k(\boldsymbol{e}_i, \boldsymbol{e}_j) = \alpha \exp(-\beta ||\boldsymbol{e}_i - \boldsymbol{e}_j||^2), \quad (9)$$

with parameters α and β. With this assumption, $P(g^*|\boldsymbol{e}^*, \mathcal{D}_g^{(l)})$ is derived as a Gaussian distribution $\mathcal{N}(\mu_l, \sigma_l^2)$ with

$$\mu_l = \boldsymbol{K}^*(\boldsymbol{K} + \boldsymbol{S})^{-1}\boldsymbol{G}^{(l)}, \quad (10)$$

and

$$\sigma_l^2 = k(\boldsymbol{e}^*, \boldsymbol{e}^*) - \boldsymbol{K}^*(\boldsymbol{K} + \boldsymbol{S})^{-1t}\boldsymbol{K}^*, \quad (11)$$

where $\boldsymbol{K}_{ij} = k(\bar{\boldsymbol{e}}_i, \bar{\boldsymbol{e}}_j)$, $\boldsymbol{K}_i^* = k(\bar{\boldsymbol{e}}_i, \boldsymbol{e}^*)$, $\boldsymbol{S}_{ij} = \varsigma_i^2 \delta_{ij}$ and $\boldsymbol{G}_i^{(l)} = g_i^{(l)}$ [1]. As a result, the distribution $P(g^*|\boldsymbol{e}^*, \mathcal{D}_p)$ can be estimated as a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with

$$\mu = \frac{1}{n_g} \sum_{l=1}^{n_g} \mu_l, \;\; \sigma^2 = \frac{1}{n_g} \sum_{l=1}^{n_g} \sigma_l^2 = \sigma_1^2. \quad (12)$$

The variance σ^2 equals to σ_1^2, because σ_l^2 of Eq. (11) is independent of the index l. Therefore, σ^2 can be calculated by taking σ_1^2.

2.4. Gaze estimation

Once we have matrices $\boldsymbol{K}$, $\boldsymbol{S}$ and $\{\boldsymbol{G}^{(1)}, \ldots, \boldsymbol{G}^{(n_g)}\}$ in Eqs. (10) and (11), a gaze point can be estimated by taking any eye image $\boldsymbol{e}$ as input. The estimated distributions for each X- and Y-direction, $\mathcal{N}(\mu_x, \sigma_x^2)$ and $\mathcal{N}(\mu_y, \sigma_y^2)$, are converted to the display coordinates $\mathcal{N}(\hat{\mu}_x, \hat{\sigma}_x^2)$ and $\mathcal{N}(\hat{\mu}_y, \hat{\sigma}_y^2)$ as

$$\hat{\mu}_x = x_o + \frac{W_I}{W_s}\mu_x, \;\; \hat{\mu}_y = y_o + \frac{H_I}{H_s}\mu_y, \quad (13)$$

and

$$\hat{\sigma}_x^2 = \frac{W_I}{W_s}\sigma_x^2, \;\; \hat{\sigma}_y^2 = \frac{H_I}{H_s}\sigma_y^2, \quad (14)$$

where W_s, H_s indicates the width and height of the saliency maps, W_I, H_I indicates the actual width and height of the displayed images $\{\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N\}$, and (x_o, y_o) indicates the display origin of the images. The average $(\hat{\mu}_x, \hat{\mu}_y)$ corresponds to the estimated gaze point $\boldsymbol{g}$.

3. Experimental results

In this section, we show experimental results to evaluate our method. In the experiments, we used four video clips from four films: A) *2001: A Space Odyssey, Stanley Kubrick, 1968*, B) *Dreams, Akira Kurosawa, 1990*, C) *Nuovo Cinema Paradiso, Giuseppe Tornatore, 1988* and D) *Forrest Gump, Robert Zemeckis, 1994*. It is known that human gaze control is also strongly influenced by contexts and plots of films, however, such high-level attentions are not modeled by the bottom-up saliency model we employed. Hence, each film was shortened to a 10-minute video clip without audio signal by extracting 2-second sequences at regular intervals to remove these effects. The video clips were resized to a fixed dimension of 720×405, and the display resolution was set to $W_I = 1920$ and $H_I = 1080$. The video clips were shown at 15 fps; therefore, $N = 9000$ in the experiments. The saliency maps were calculated at a smaller resolution, $W_s = 32$ and $H_s = 18$.

Six novice test subjects $s_1 \ldots s_6$ were asked to watch two video clips. The combinations of video clips and test subjects are defined so that every clip was tested as learning data against three different subject persons as listed in Table 1. A chin rest is used to fix their head positions, and

[1] $\boldsymbol{K} \in \mathbb{R}^{M \times M}$, $\boldsymbol{K}^* \in \mathbb{R}^{1 \times M}$, $\boldsymbol{S} \in \mathbb{R}^{M \times M}$ and $\boldsymbol{G}^{(l)} \in \mathbb{R}^{1 \times M}$

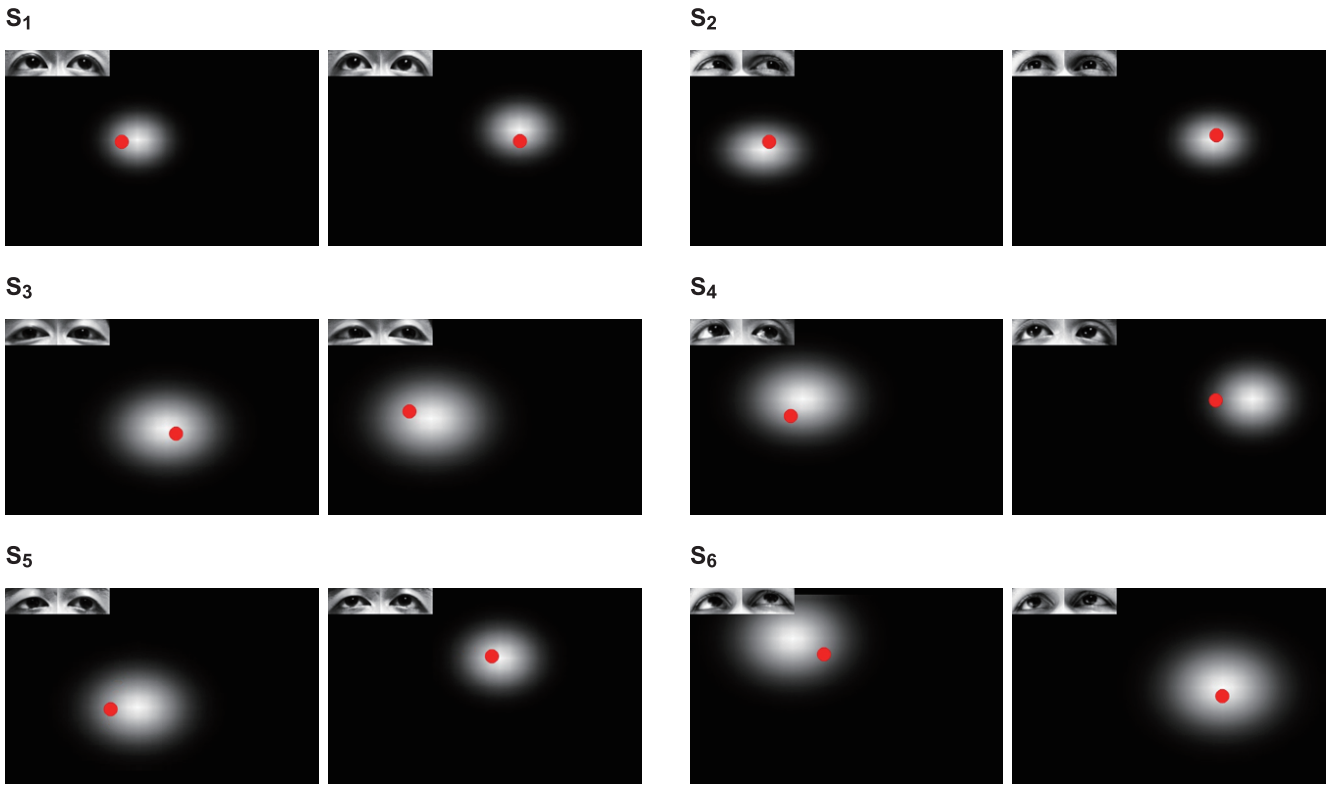

Figure 6. Estimation results. The estimation results are rendered as 2-D Gaussian circles. The corresponding input eye images are shown at the top-left corner. Overlaid circles are the ground truth gaze points obtained from a calibration-based gaze estimator.

Table 1. Combinations of video clips A to D and test subjects s_1 to s_6. For example, person s_1 watched clips A and B.

	Destination			
Source	A	B	C	D
A		s_1	s_2	s_3
B	s_1		s_4	s_5
C	s_2	s_4		s_6
D	s_3	s_5	s_6	

a 22.0-inches WUXGA (473.8×296.1 mm) display was placed 400 mm in front of the subject when video clips were shown. While the subjects were watching the clips, their eyes were automatically detected and captured using OMRON OKAO Vision library.

The ground truth calibration data were collected for each user by showing reference points in a separate stage. For this, 16×9 points were shown at 120×120-pixel intervals and eye images were captured in the same way. The ground truth was used to quantitatively assess our method in comparison with the gaze estimation method that involves an explicit calibration stage.

Throughout the experiment, the parameters were set as follows; $n_s = 5$, $\kappa_s = 7.8 \times 10^{-7}$, $\tau_s = 0.4$, $n_f = 5$, $\kappa_f = 0.02$, $n_g = 50$, $\alpha = 50$, $\beta = 5.0 \times 10^{-9}$, and τ_s was adaptively set to keep the top 20% of pixels and set remaining 80% to zero in each map. These parameter settings are empirically obtained from our experiment. In our current implementation, it took about 0.15 seconds per frame when $M \simeq 600$ using a Core 2 Quad 2.66GHz with simple code parallelization using OpenMP [13].

3.1. Gaze estimation results

Using the two clips $\times$ six subject people, we tested our method in two scenarios. In Scenario 1, we assessed our method using the learning dataset as a test dataset. Because the true gaze points are not known in the learning dataset, this experiment was designed to verify the performance of the algorithm. In Scenario 2, evaluations were performed using another dataset from the user as a test dataset to confirm the applicability of the trained gaze estimator to other datasets.

The ground truth gaze points of the datasets were obtained using a calibration-based gaze estimator. It was achieved by a standard Gaussian process regression method with a labeled dataset. Namely, pairs of the ground-truth gaze points and eye images were explicitly given to learn the relationship between gaze points and eye images. The same covariance function (Eq. (9)) was used, and α and β were set to be the same values as our estimator. The noise

variance ς^2 was empirically set to zero under the assumption that the ground-truth dataset is noise-free.

Figure 6 shows examples of the estimation results. Outputs of the estimators are rendered as 2-D Gaussian circles centered at $(\hat{\mu}_x, \hat{\mu}_y)$ with variance $(\hat{\sigma}_x^2, \hat{\sigma}_y^2)$ given by Eq. (12). The center coordinate $(\hat{\mu}_x, \hat{\mu}_y)$ corresponds to the estimated gaze point. The eye images shown at the top-left corner show input eye images for estimation, and the overlaid circles represent true gaze points obtained from the calibration-based estimator.

Table 2 summarizes the estimation results for each video clips. Each row corresponds to the average of three subjects' results where the corresponding video clip is used as the training dataset (see Table 1). First two columns indicate AUCs of average ROC curves of the raw saliency maps $\boldsymbol{s}$ and gaze probability maps $\bar{\boldsymbol{p}}$. The rest of the columns indicate estimation errors in distance and angle represented as *average* $\pm$ *standard deviation*. Distance errors are evaluated as the Euclidean distance between the estimated and ground-truth gaze points, and angular errors are calculated using the distance between eyes and the display.

From these results, it is observed that the gaze estimation accuracy depends on the accuracy of the gaze probability maps. When the AUC of the gaze probability maps $\bar{\boldsymbol{p}}$ is lower, the estimation error tends to become larger.

Table 3 lists the estimation error of each subject person. Each row corresponds to average of results of the corresponding test subject with two different training datasets. The columns show AUCs and estimation errors in the similar manner as in Table 2. In contrast to Table 2, subject dependency of our method is not clearly observed.

The accuracy of our method has dependency on the distribution of learning samples. Figure 7 shows the spatial distribution of average estimation errors. Each grid corresponds to a reference point that is used to capture the calibration data when producing the ground truth data. Using eye images obtained from the ground truth dataset as input to our method, we compute the errors of our method. Lower intensity corresponds to the lower estimation error. From this, the larger errors can be observed at edges of the display. Figure 8 shows the average saliency map and spatial histogram of gaze points. The left image shows the average of all raw saliency maps extracted from the four video clips used in our experiment. The right image shows the spatial histogram of ground-truth gaze points obtained from the experiment dataset. Higher intensity corresponds to larger amount of gaze points given at the grid. Usually salient objects are located at the center of video frames, and the gaze point also tends to concentrate at the center of the display. Because of these reasons, the number of learning samples at the display edges are limited, and these cause the bias of the estimation accuracy shown in Figure 7.

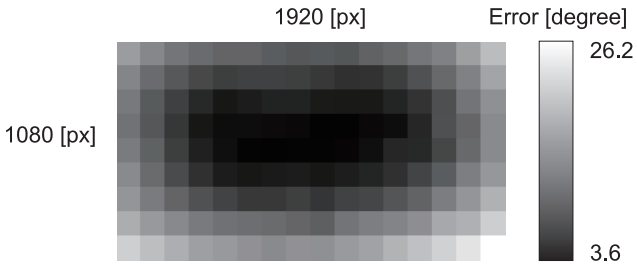

Figure 7. Spatial distribution of estimation errors in the display coordinate. Lower intensity corresponds to the lower estimation error as illustrated in the right bar.

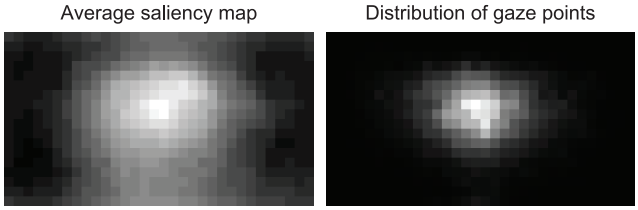

Figure 8. Average saliency map and spatial histogram of gaze points. Left image shows the average of all raw saliency maps extracted from four video clips used in the experiment. Right image shows the spatial histogram of the ground-truth gaze points of experimental dataset. Higher intensity corresponds to larger counts of gaze points.

4. Conclusions

We proposed a novel calibration-free gaze estimation framework using saliency maps. By only using a synchronized set of eye images and video frames, a gaze estimator can be constructed by treating saliency maps as probabilistic distributions of gaze points. To the best of our knowledge, this is the first work to use saliency maps as the key for gaze estimation. Our method naturally avoids an explicit and noticeable gaze calibration step that is often demanding for users. In our experimental setting with fixed head positions, our method achieves the accuracy of about 6-degree error.

The estimation accuracy of our method depends on the raw saliency maps extracted from input video clips. The mechanism of human gaze control has not been completely investigated, and there is a wide range of possibilities of more advanced saliency models for accurately predicting gaze. Our method can benefit from the further investigation of more accurate saliency models.

References

[1] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 2009.

[2] M. Cerf, J. Harel, W. Einhauser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *Proceedings of Advances in neural information*

Table 2. Average error for each video clip. Two AUC columns indicates AUCs of the average ROC curves of raw saliency maps $\boldsymbol{s}$ and gaze probability maps $\bar{\boldsymbol{p}}$. The rest of columns indicates distance and angular estimation errors (*average* $\pm$ *standard deviation*) in two estimation scenarios.

	$\boldsymbol{s}$	$\bar{\boldsymbol{p}}$	Scenario 1		Scenario 2	
Clip	AUC	AUC	error [mm]	error [deg.]	error [mm]	error [deg.]
A	0.75	0.90	38 ± 26	5.3 ± 3.6	45 ± 28	6.3 ± 3.8
B	0.71	0.87	60 ± 32	8.3 ± 4.4	56 ± 32	7.9 ± 4.3
C	0.74	0.92	31 ± 20	4.3 ± 2.7	36 ± 19	5.0 ± 2.5
D	0.70	0.89	36 ± 23	5.0 ± 3.1	42 ± 25	5.9 ± 3.4
Average	0.73	0.90	41 ± 25	5.7 ± 3.5	45 ± 26	6.3 ± 3.5

Table 3. Average error for each subject person. Columns indicate AUCs of the average ROC curves and estimation errors as in Table 2.

	$\boldsymbol{s}$	$\bar{\boldsymbol{p}}$	Scenario 1		Scenario 2	
Subject	AUC	AUC	error [mm]	error [deg.]	error [mm]	error [deg.]
s_1	0.74	0.90	48 ± 35	6.8 ± 4.8	48 ± 36	6.7 ± 5.0
s_2	0.75	0.93	30 ± 20	4.1 ± 2.7	30 ± 19	4.2 ± 2.6
s_3	0.72	0.87	42 ± 27	5.9 ± 3.6	58 ± 27	8.1 ± 3.7
s_4	0.71	0.87	43 ± 26	6.0 ± 3.5	48 ± 27	6.7 ± 3.6
s_5	0.71	0.89	51 ± 26	7.1 ± 3.6	52 ± 28	7.2 ± 3.8
s_6	0.74	0.92	33 ± 18	4.6 ± 2.5	34 ± 18	4.8 ± 2.4

processing systems (NIPS 2008), volume 20, pages 241–248, 2008.

[3] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on pattern analysis and machine intelligence*, 2009.

[4] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Proceedings of Advances in neural information processing systems (NIPS 2007)*, volume 19, pages 545–552, 2007.

[5] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems, Vol. 19 (NIPS 2005)*, pages 547–554, 2006.

[6] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, volume 5200, pages 64–78, 2003.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[8] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Proceedings of the Twelfth IEEE International Conference on Computer Vision (ICCV 2009)*, 2009.

[9] W. Kienzle, B. Scholkopf, F. A. Wichmann, and M. O. Franz. How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In *Proceedings of the 29th Annual Symposium of the German Association for Pattern Recognition (DAGM 2007)*, pages 405–414, 2007.

[10] W. Kienzle, F. A. Wichmann, B. Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *Proceedings of Advances in neural information processing systems (NIPS 2006)*, volume 19, pages 689–696, 2006.

[11] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

[12] T. Nagamatsu, J. Kamahara, T. Iko, and N. Tanaka. One-point calibration gaze tracking based on eyeball kinematics using stereo cameras. In *Proceedings of the 2008 symposium on Eye tracking research & applications (ETRA '08)*, pages 95–98, 2008.

[13] OpenMP. http://openmp.org/.

[14] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[15] C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence*, 22(9):970–982, 2000.

[16] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

[17] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *Proceedings of the 10th European Conference on Computer Vision (ECCV 2008)*, pages 656–667, 2008.

[18] A. Villanueva and R. Cabeza. A novel gaze estimation system with one calibration point. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(4):1123–1138, 2008.

[19] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the 2008 symposium on Eye tracking research & applications (ETRA '08)*, pages 245–250, 2008.