Unsupervised Outlier Detection in Appearance-Based Gaze Estimation

Zhaokang Chen* Didan Deng* Jimin Pi* Bertram E. Shi
The Hong Kong University of Science and Technology, Hong Kong SAR
{zchenbc, ddeng, jpi}@connect.ust.hk, eebert@ust.hk

Abstract

Appearance-based gaze estimation maps RGB images to estimates of gaze directions. One problem in gaze estimation is that there always exist low-quality samples (outliers) in which the eyes are barely visible. These low-quality samples are mainly caused by blinks, occlusions (e.g. by eye glasses), blur (e.g. due to motion) and failures of the eye landmark detection. Training on these outliers degrades the performance of gaze estimators, since they have no or limited information about gaze directions. It is also risky to give estimates based on these images in real-world applications, as these estimates may be unreliable. To solve this problem, we propose an algorithm that detects outliers without supervision. Based on the input images with only gaze labels, the proposed algorithm learns to predict a gaze estimates and an additional confidence score, which alleviates the impact of outliers during learning. We evaluated this algorithm on the MPIIGaze dataset and on an internal dataset. In cross-subject evaluation, our experimental results show that the proposed algorithm results in a better gaze estimator (8% improvement). The proposed algorithm is also able to reliably detect outliers during testing, with a precision of 0.71 when the recall is 0.63.

1. Introduction

Human gaze has been recognized as an important cue for inferring people's intent in many applications, such as human-computer interfaces [4, 25, 29], human-robot interaction [18], virtual reality [26, 28], social behavioral analysis [16], long-range tracking [36] and health care [12]. These successes have lead to more and more attention on generating good gaze estimates.

Gaze estimation methods can be generally classified into two main groups: model-based methods and appearance-based methods [14]. Model-based methods mostly rely upon active illumination, e.g. infrared illumination used in pupil center corneal reflections (PCCR) [13]. To obtain the parameter of the physiological eye model, calibration is needed before usage. While these methods provide high accuracy, they also place strong constraints on users' head movements. Accuracy rapidly degrades as the head pose changes and people need to do the calibration again to continue the usage. Meanwhile, eye trackers using model-based methods are relatively costly, as they rely upon custom hardware to provide the required illumination. On the other hand, appearance-based methods generate gaze estimates based on RGB images. They only require commonly available off-the-shelf cameras and provide relatively unconstrained gaze tracking. Although the accuracy is generally lower than the model-based methods, they are often cheaper, easier to setup, and more robust to head motion. Recently, the application of deep convolutional neural networks (CNNs) has reduced estimation error significantly [42]. Results on a large number of high quality real and synthetic datasets [10, 11, 21, 31, 33, 35, 37, 38, 42] show that deep CNNs can learn to compensate for the large variability caused by factors such as differences in individual appearance, head pose, and illumination [3, 6, 7, 21, 23, 30, 43].

Training an appearance-based gaze estimator requires a large number of training samples. Low-quality samples (outliers) are inevitable. Due to the fact that people blink occasionally, there will be closed-eye images in both training and testing scenarios. Besides blinks, outliers can also be caused by occlusions, blur and failures of the eye landmark detection (see Fig. 1 for examples). Fully trusting all the samples is risky for two reasons. First, they may degrade the learning, as they have no or limited information about gaze direction. Second, during deployment taking action based on unreliable gaze estimates is risky. For example, a gaze-based wheelchair should not follow commands generated by gaze estimates from images where there is a blink or when the eyes are occluded.

Alleviating the influence of low-quality samples is not a new topic. Detecting testing samples that are far away from the distributions of training samples has been well studied in classification problems, e.g. [9, 15, 22, 24]. However, most of these works considered supervised learning scenar-

*Indicates equal contribution

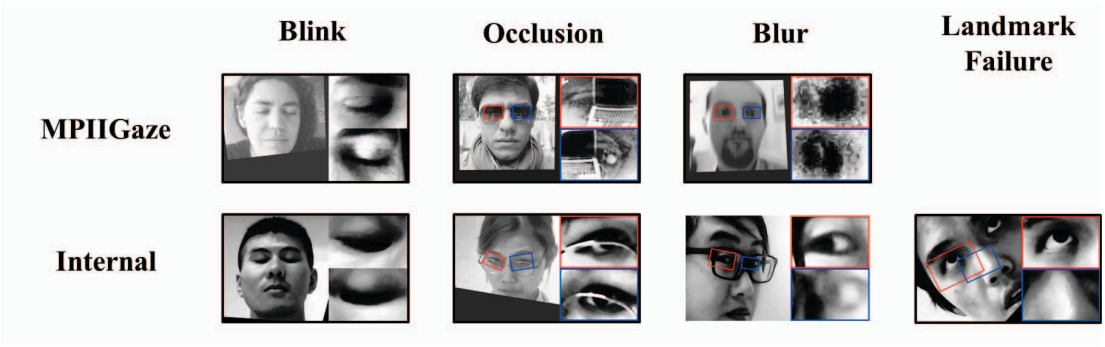


Figure 1. Outliers detected by our proposed algorithm in cross-subject experiments on the MPIIGaze dataset (top row) and on an internal dataset (bottom row). They are mainly caused by blinks, occlusions, blur and failures of the facial landmark detection. MPIIGaze does not contain images where landmark detection fails because the landmarks are refined manually.

ios. Moreover, to our knowledge, techniques for handling outliers in appearance-based gaze estimation have not been studied previously.

In this manuscript, we propose an algorithm which learns a subject-independent appearance-based gaze estimator and an outlier image detector simultaneously, without the need for outlier labels. The proposed algorithm learns to estimate a confidence score for the gaze estimate of each image, where the confidence score is low for outlier images. Our results in cross-subject experiments on the MPIIGaze dataset and an internal dataset show that this algorithm (1) improves the performance of subject-independent gaze estimation since the impact of outliers is alleviated during training, and (2) can detect outliers successfully during testing. Most importantly, our proposed algorithm does not require any extra labels about image quality since it learns to detect outliers without supervision.

2. Related work

2.1. Appearance-based gaze estimation

Methods for appearance-based gaze estimation directly regress from images to gaze estimates. To achieve relatively unconstrained gaze tracking, they need to address the large variability in real-world situation, such as differences between subjects, in head pose and in illumination.

The application of deep CNNs to this problem has received increasing attention. Zhang *et al*. proposed the first deep CNN for gaze estimation in the wild [42, 44]. They showed that deep CNNs improved accuracy significantly. To further reduce estimation error, different directions have been explored. Some work has focused on using information from the face region outside the eye regions [21, 43]. Some work has concentrated on the head-eye relationships [7, 30]. Estimation error was reduced by better utilizing head pose information. Other work has focused on extracting better features from eye images, e.g., studying the "two eye asymmetry problem" [6], estimating the eye landmark locations and gaze directions jointly [40], learning an intermediate pictorial representation of the eyes [27], using dilated-convolutions to extract features at high resolution [3] and fusing information from images captured from multiple cameras [23].

2.2. Out-of-distribution detection

Out-of-distribution detection (OD detection or ODD) is an active research topic in the field of classification. The goal of ODD is to identify testing samples that are far from the training samples, which are referred to as in-distribution (ID). For example, for a task of cat-dog classification, a cat/dog sample is ID, while a horse sample would be OD.

One approach to OD detection is to include OD samples during training for supervised learning. Liang *et al*. proposed ODIN, which increases the difference between the maximum softmax scores of ID and OD samples [24]. Hendrycks *et al*. proposed to train anomaly detectors with an auxiliary dataset [15]. These two works used external datasets as OD samples. Lee *et al*. proposed to use generative adversarial networks (GANs) to synthesize OD samples, which were used to train a classifier that produced a concentrated distribution for ID samples but a uniform distribution for OD samples [22]. These methods all use supervised learning for OD detection, assuming that ID/OD labels are available for each image.

Our work is most similar to the work of DeVries and Taylor [9]. Their network learned confidence scores based on images, where an image that gives an incorrect prediction has a low confidence score and is discounted in the loss function. We follow the same vein, where we use the mean squared error during training to define a measure of confidence. Our experimental results show that although this measure can not reliably distinguish ID and OD samples, it provides useful information that enables us to learn a reliable OD detector. We extend the approach in [9] to handle two problems. First, for our datasets, the number of outliers is far less than the number of normal samples. Second, the

outliers have similar appearance to the normal samples.

In particular, we introduce a novel concept: the confidence pseudo-label. We use this pseudo-label to dynamically adjust hyperparameters of the loss function for appearance-based gaze estimation and to balance the number of positive and negative samples during training.

2.3. Blink detection

Several approaches have been proposed to detect blinks based on RGB images. Soukupova and Cech defined an eye aspect ratio that measures the openness of eyes based on automatically detected facial landmarks [34]. They used a linear support vector machine (SVM) as the final classifier. Hu *et al*. proposed to use a long short-term memory (LSTM) to capture temporal information [17]. The methods were learning based, but required labeled samples. Kassner *et al*. proposed a hand-crafted algorithm to detect the pupil by ellipse fitting [20].

In contrast to these methods, our proposed algorithm learns how to detect and discount the effect of outliers without supervision during the training of gaze estimator. This reduces the burden of data labeling. It also enables the outlier detector and gaze estimator to share features, which reduces the computational load in comparison to testing the two problems separately.

3. Methodology

3.1. Outliers

In this work, we define outliers to be samples in which the eyes are not fully visible or corrupted, since these images have no or very limited information about the precise gaze directions. From available datasets, we have identified four main cases that significantly affect gaze estimation by human observers: blinks (more than half of the pupil is covered by the eyelid), occlusions (the center or more than half of the pupil is not visible), failures of the facial landmark detection (the bounding box of the eye given by the landmark detector does not cover the whole eye) and blur (the pupil is not clearly visible). Some examples of outliers from the MPIIGaze dataset [42] and an internal dataset are illustrated in Fig. 1.

In the following experiments, we refer to the normal samples as negative samples and the outlier samples as positive samples.

3.2. Image Preprocessing

We use the data normalization method introduced in [41]. This method rotates and scales an image so that the resulting image is taken by a virtual camera directed at a reference point on the face from a fixed distance and cancels out the roll angle of the head. Images are normalized by perspective warping, converted to gray scale and histogram-equalized. The ground truth gaze angles are also normalized correspondingly. We use OpenFace [1] for automatic facial landmark detection.

3.3. Weighted mean squared error

Our appearance-based gaze estimator estimates both yaw and pitch gaze angles. A common cost function used to train a gaze estimator is the mean squared error (MSE) between the estimated and ground truth gaze angles, i.e.,

$$\text{MSE} = \frac{1}{N}\sum_{i}^{N}\left\|g_i - \hat{g}(x_i)\right\|_2^2, \tag{1}$$

where i is sample index, g_i is the true gaze, $\hat{g}(x_i)$ is the estimated gaze, x_i is the image. In the rest of this manuscript, we define $e_i = \left\|g_i - \hat{g}(x_i)\right\|_2^2$. The MSE assumes that all samples in the training set should contribute equally. We expect a performance degradation if there exist a few outliers in the training set.

To alleviate the impact of outliers, we considered the weighted MSE, which can be written as follows:

$$\text{weighted MSE} = \frac{1}{N}\sum_{i}^{N}\left[\hat{c}(x_i)e_i\right], \tag{2}$$

where $\hat{c}(x_i)$ is a confidence score ranging from 0 to 1. We expect a high confidence score for a normal sample and a low confidence score for an outlier so that outliers contribute less to the cost function. During testing, $\hat{c}(x_i)$ can be used to detect outliers. To avoid $\hat{c}(x_i) = 0, \forall i$, we add penalties for $\hat{c}(x_i)$ being too small. The final loss function can be written as:

$$\frac{1}{N}\sum_{i}^{N} J(e_i, \hat{c}(x_i)), \tag{3}$$

where

$$J(e, c) = ce - \alpha c - \lambda \log c, \tag{4}$$

where α and λ are the hyperparamters of the penalties. We will explain the rationale for these penalties in detail later.

3.4. Architecture

The architecture of our proposed network is presented in Fig. 2. The general architecture is inspired by iTracker [21] and Dilated-Net [3]. It takes an image of the face and images of both eyes as input and outputs the gaze estimates. We also adopt the gaze decomposition method proposed in [5] to improve the performance of gaze estimation.

The input images x_i are first fed to three base CNNs that perform feature extraction on three image regions: the whole face and the two eyes. The architecture of the base

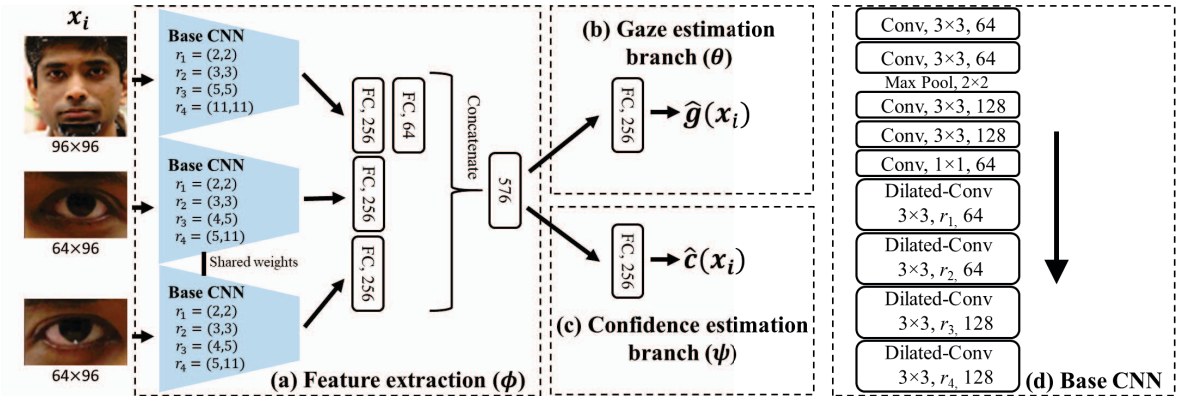


Figure 2. Architecture of the proposed network. (a) The main network that extracts features from the input image x_i. (b) The gaze estimation branch. (c) The confidence estimation branch. (d) The base CNN is the basic component of (a). FC denotes fully-connected layers, Conv denotes convolutional layers and Dilated-Conv denotes dilated-convolutional layers with r as the dilation rate.

CNNs is shown in Fig. 2(d). Each CNN has five convolutional layers, one max-pooling layer and four dilated-convolutional layers [39]. The CNNs differ in their dilation rates due to the differences in the sizes of the input image regions. The dilated-convolutional layers learn high-level features at high resolution and capture subtle appearance differences. The feature maps extracted by the base CNNs are then fed to fully-connected (FC) layers. The two base CNNs that take the eyes as input share the same weights. We denote the parameters of these networks by ϕ.

The outputs of the three feature extractors are concatenated and then fed to the gaze estimation branch and the confidence estimation branch. The gaze estimation branch has one FC followed by a linear output layer with two outputs corresponding to yaw and pitch. (see Fig. 2(b)). The confidence estimation layer also has one hidden FC layer but uses a sigmoid function in the output layer, which has only one output (see Fig. 2(c)). We denote the parameters of the gaze estimation branch and the confidence estimation branch by θ and ψ respectively.

Rectified Linear Units (ReLUs) are used as the activation functions. Zero-padding is applied to regular convolutional layers and no padding is applied to dilated-convolutional layers. The strides for all (dilated-) convolutional layers are one. The initial weights of the first four convolutional layers are transferred from VGG-16 [32] pre-trained on the ImageNet dataset [8]. The weights in all other layers are randomly initialized. Batch renormalization [19] is also applied to these layers. Dropout layers with dropout rates of 0.5 are applied to all FC layers. During training, all weights in all layers are updated.

We implement our network in TensorFlow. We use the Adam optimizer with its default parameters and a batch size of 64. An initial learning rate of 0.001 is used. It is divided by 10 after every ten epochs. The training proceeds for 25 epochs. We apply online data augmentation including random cropping, scaling, rotation and horizontal flipping.

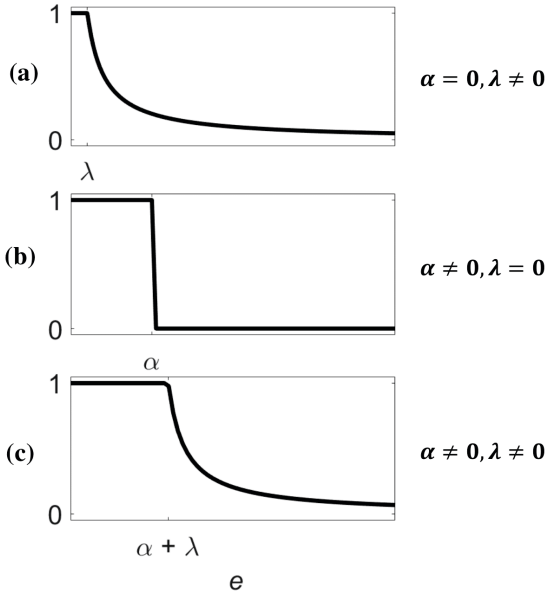


Figure 3. Pseudo-label c^* as a function of estimation error e for different hyperparameter settings.

3.5. Training procedure

We train the nework based on two assumptions. First, we assume e_i to be a good indicator of the quality of the input sample. Generally most normal samples will have low estimation error, whereas many outliers may have large estimation error. Second, we assume that a deep network can learn to distinguish between normal samples and outlier samples.

We define the confidence pseudo-label of sample x_i, c_i^* or $c^*(e_i)$, to be the solution of the following optimization

problem:

$$c_i^* = \underset{c}{\operatorname{argmin}} J\big(e_i, c\big) \quad \text{subject to} \quad 0 \leq c \leq 1 \tag{5}$$

where J is the differentiable loss function defined in (4). The pseudo-label c_i^* has the following closed-form solution:

$$c_i^* = \begin{cases} 1 & if\ e_i \leq \alpha + \lambda \\ \frac{\lambda}{e_i - \alpha} & if\ e_i > \alpha + \lambda \end{cases} \tag{6}$$

As the gaze estimation error e_i increases, the pseudo-label c_i^* decreases. Their relationship is shown in Fig. 3. This is consistent with the first assumption described above. We choose to use both a linear penalty and a log penalty in the loss (3) because we want the confidence pseudo-label to saturate at one for small values of e_i and to have control of the rate of decrease for larger values of e_i.

During training, we use different mini-batches to train the gaze estimation branch and the confidence estimation branch. To be specific, for the gaze estimation branch, we sample mini-batches uniformly from the training set. For the confidence estimation branch, we try to balance the number of normal and outlier samples within each mini-batch. Otherwise, the trained confidence estimation network would be strongly biased.

As the outlier labels are not availables, we use the confidence pseudo-label defined above to balance the samples. To be specific, we let the number of samples that have $c^* > 0.5$ equal to the number of samples that have $c_i^* < 0.5$ within each mini-batch.

We also use the pseudo-label to adjust the hyperparameter α of the loss function during training. We set the initial value of $\alpha = (\frac{\pi}{45})^2$ rad^2, which corresponds to $4°$, and update α during training to maintain the percentage of training samples with $c_i^* < 0.5$ to be between $TH_{\text{low}} = 5\%$ and $TH_{\text{high}} = 15\%$. To stabilize the training, we first train the network using the MSE (1) for $epoch\_warmup = 3$. The value of λ was fixed to $5e-4$ based on a grid search. The procedure is presented in Algorithm 1.

Our proposed algorithm may assign a high confidence pseudo-label to an outlier if it happens to have a small training error, or a low confidence pseudo-label to a normal sample if it happens to have a large estimation error. However, our experimental results show that although some confidence pseudo-labels c_i^* may be incorrect, they provide sufficient information for training an accurate gaze estimator and reliable outlier detector.

Algorithm 1 Training the Gaze Estimator/Outlier Detector

Initialization: $\mathcal{S} = \{x_i, g_i\}_{i=1}^N$, $m = 64$, $\lambda = 5e^{-4}$, $\alpha = (\frac{\pi}{45})^2$ rad^2, $TH_{\text{low}} = 5\%$, $TH_{\text{high}} = 15\%$, $epoch\_warmup = 3$, $epoch\_stop$, network parameters (ϕ, θ, ψ);
for $t = 1 : epoch\_stop$ **do**
  Initialization:$\mathcal{S}_t = \mathcal{S}, \mathcal{S}_t^p = \varnothing, \mathcal{S}_t^n = \varnothing$;
  while $\mathcal{S}_t \neq \varnothing$ **do**
    Sample s of size m from $\mathcal{S}_t$ without replacement
    for (x_i, g_i) in s **do**
      Calculate e_i on (x_i, g_i)
      Calculate c_i^* according to (6) based on (e_i, λ, α)
      if $c_i^* \leq 0.5$ **then**
        $\mathcal{S}_t^p$.append((x_i, g_i))
      else
        $\mathcal{S}_t^n$.append((x_i, g_i))
      end if
    end for
    if $t \leq epoch\_warmup$ **then**
      Update (ϕ, θ) on s minimizing (1)
    else
      Update (ϕ, θ) on s minimizing (3)
      # samples balancing
      Sample s^p of size $\frac{m}{2}$ from $\mathcal{S}_{t-1}^p$ with replacement
      Sample s^n of size $\frac{m}{2}$ from $\mathcal{S}_{t-1}^n$ with replacement
      Update (ϕ, ψ) on (s^p, s^n) minimizing (3)
    end if
  end while
  if $|\mathcal{S}_t^n|/|\mathcal{S}| < TH_{\text{low}}$ **then**
    $\sqrt{\alpha} \leftarrow \sqrt{\alpha} - \frac{\pi}{180}$
  else if $|\mathcal{S}_t^n|/|\mathcal{S}| > TH_{\text{high}}$ **then**
    $\sqrt{\alpha} \leftarrow \sqrt{\alpha} + \frac{\pi}{180}$
  end if
end for

4. Experiments

We evaluated our proposed algorithm on two datasets: MPIIGaze dataset and an internal dataset. We created a modified version of MPIIGaze dataset, where 10% of the images were corrupted. The internal dataset already contains many outliers due to blinks, occlusions and failures of the landmark detection as it was recorded by video and was not filtered.

We conducted two tests for each dataset: Test I evaluated the gaze estimation accuracy, and Test II evaluated the performance of outlier detection.

4.1. Datasets

MPIIGaze. This dataset contains full face images of 15 subjects (six female, five with glasses). We used the "Evaluation Subset", which contains 3,000 randomly selected images for each subject. We refer to these samples as uncorrupted samples, and this dataset as **Clean MPIIGaze**.

We created a modified version of MPIIGaze (**Corrupted MPIIGaze**) by adding 300 outlier images per subjects, which were generated by significantly disturbing the facial landmarks (see examples in Fig. 4).

Internal dataset. This dataset contains full face videos of 21 subjects (10 female, 10 with glasses). It contains

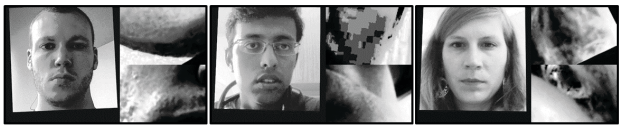


Figure 4. Examples of the generated corrupted samples of Corrupted MPIIGaze.

Figure 5. Example images of the internal dataset. This dataset contains 21 subjects with large variability of head pose and face location.

significant variations in head pose and face location. Some example images are presented in Fig. 5. OpenFace [1] was used for facial landmark detection. The sample rate was 10 fps.

We used this dataset to create two new datasets. The first dataset, **Clean Internal**, contained most of the collected images in the dataset. We removed images whose confidence for the landmark detection given by OpenFace was lower than 0.02 or images with significantly abnormal landmarks. We also removed images during blinks, which were detected by the algorithm proposed in [2] with an empirical threshold. Clean Internal contains $496,695$ images (about $24,000$ images per subject).

The second dataset, **Corrupted Internal**, contains the images in which the faces are in the lower regions in the images. This set contains more samples that fail landmark detection. It contains $185,357$ images in total (about $8,800$ images per subject).

4.2. Dataset labeling

To evaluate the performance of outlier detection, we labeled a subset of each dataset. We manually labeled each sample as being either normal or an outlier. For the MPIIGaze dataset, we labeled the set of uncorrupted samples whose estimated confidence scores were less than or equal to 0.5, i.e., $\{x_i : \hat{c}(x_i) \leq 0.5\}$. The size of this subset is 659.

For the internal dataset, we labeled 10% of Corrupted Internal by downsamping it from 10 fps to 1 fps. This subset contained $18,535$ images in total, among which $1,326$ (7.15%) samples were labeled as outliers.

4.3. Results - MPIIGaze dataset

In Test I, we trained on Corrupted MPIIGaze and tested on Clean MPIIGaze to evaluate the gaze estimation accuracy. In Test II, we trained on Corrupted MPIIGaze and tested on Corrupted MPIIGaze to evaluate the performance of outlier detection. Subjects used in training and testing were different. We conducted 15-fold leave-one-subject-out cross-validation.

Test I: Performance of gaze estimation. We compared our proposed method with two baselines without confidence estimation: one was trained on Clean MPIIGaze, and the other was trained on Corrupted MPIIGaze. The mean angular errors over 15 subjects from Clean MPIIGaze are presented in Table 1.

Without confidence estimation, the gaze estimation error degraded from $4.5°$ to $5.1°$ when the corrupted images were added into training set. This indicates that including low-quality samples in the training set significantly degrades the performance. Our proposed confidence estimation reduced the gaze estimation error from $5.1°$ to $4.7°$ (7.8%), i.e., the degradation decreased by 66.7% (from $0.6°$ to $0.2°$). This small degradation remaining is partly because that some normal samples were assigned low confidences as their training errors were high.

Test II: Outlier detection. We first tested on Corrupted MPIIGaze. We compared the confidence score from the confidence estimator, $\hat{c}$, of the uncorrupted and corrupted samples. For the uncorrupted samples, 95% samples had $\hat{c} > 0.9$. For the corrupted samples, 99% samples had $\hat{c} < 0.1$. All of the corrupted samples had $\hat{c} < 0.5$, except for one sample for which $\hat{c} = 0.69$.

We then tested on Clean MPIIGaze. Among the $45,000$ original uncorrupted samples, 659 samples (1.5%) were assigned $\hat{c} \leq 0.5$. These samples have low $\hat{c}$ mainly due to blinks or blur. Some examples are presented in Fig. 1. We

Training set	Confidence estimation	Mean error
Clean MPIIGaze	*No*	$4.5°$
Corrupted MPIIGaze	*No*	$5.1°$
Corrupted MPIIGaze	*Yes*	$4.7°$

Table 1. Cross-Subject Gaze Estimation on Clean MPIIGaze.

Training set	Confidence estimation	Mean error
Corrupted Internal	*No*	$4.1°$
Corrupted Internal	*Yes*	$3.8°$

Table 2. Cross-Subject Gaze Estimation on Clean Internal.

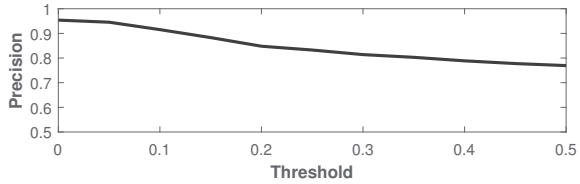


Figure 6. Precision as a function of the threshold of $\hat{c}$ tested on the uncorrupted samples of Clean MPIIGaze.

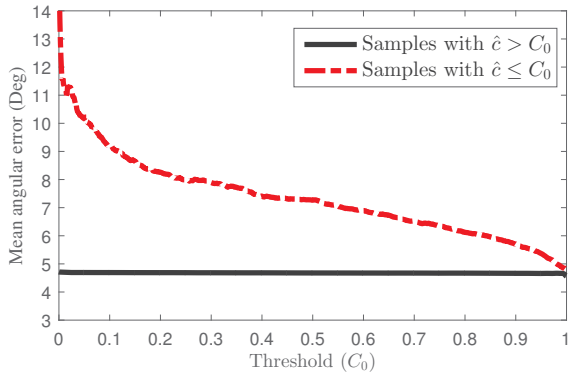


Figure 7. Mean angular error of predicted positive/negative samples as a function of decision threshold C_0 on Clean MPIIGaze.

manually labeled the 659 detected outliers, and found that 507 out of 659 samples were outliers, i.e., the precision was 76.9%. We plot the precision as a function of the threshold of $\hat{c}$ in Fig. 6. The smaller the threshold, the higher the precision. The precision was above 90% when the threshold was 0.1. These results are significant given that the number of normal samples is far larger than that of outliers.

We also evaluated the relationship between the estimation error and $\hat{c}$. For a decision threshold $C_0 \in [0, 1]$ on $\hat{c}$, we calculated the mean angular of predicted positive samples ($\hat{c} \leq C_0$) and that of predicted negative samples ($\hat{c} > C_0$). Fig 7 presents the mean angular errors as a function of decision threshold C_0. The results show that when the threshold was small, e.g. $C_0 = 0.5$, the mean angular error of samples with $\hat{c} \leq C_0$ was significantly greater than the mean angular error of samples with $\hat{c} > C_0$. This indicates that the outliers detected by our proposed algorithm indeed have large estimation errors.

4.4. Results - Internal dataset

Similar to MPIIGaze, in Test I we trained on Corrupted Internal and tested on Clean Internal. In Test II we trained on Corrupted Internal and tested on Corrupted Internal. We conducted five-fold cross-subject cross-validation.

Test I: Performance of gaze estimation. We trained a network without confidence estimation as a baseline. We tested on Clean Internal. The mean angular errors over

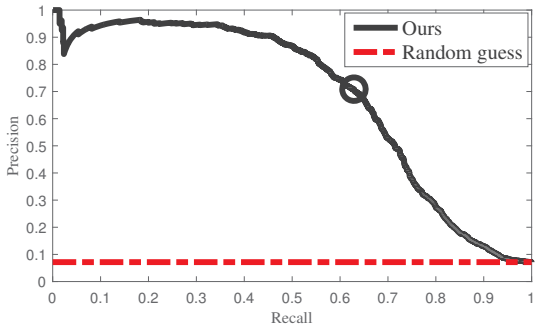


Figure 8. The precision-recall curves of different algorithms on Corrupted Internal.The circle indicates the best F1-score.

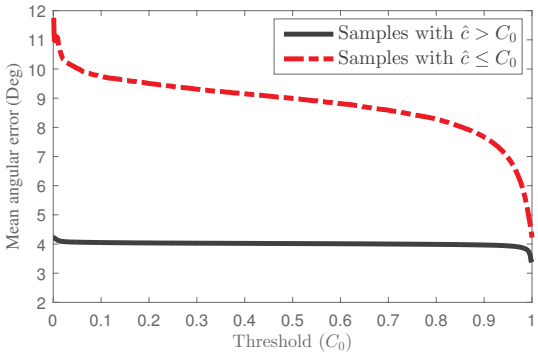


Figure 9. Mean angular error of predicted positive/negative samples as a function of decision threshold C_0 on Corrupted Internal.

subjects are presented in Table 2. Our proposed method achieved an error of 3.8°, which was 0.3° (7.3%) lower than the 4.1° achieved by the baseline.

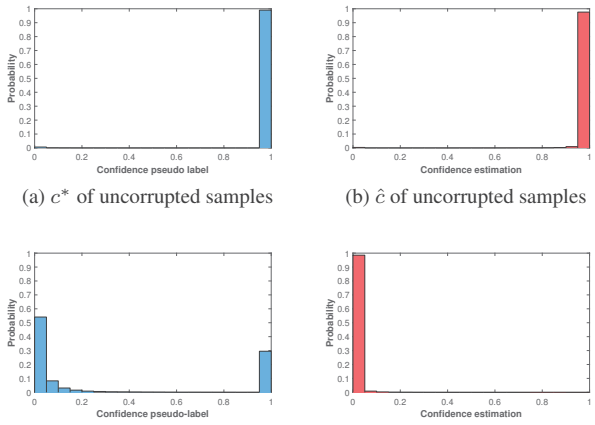
Test II: Outlier detection. We first tested on the subset of Corrupted Internal that was labelled. Fig. 8 presents the precision-recall curves. The precision of a random guess was 7.15%. The area-under-curve (AUC) of our proposed algorithm was 0.68. The circle on the black curve indicates the position of the best F1-score (0.67), where the precision was 0.71 and the recall was 0.63.

We then tested on the entire Corrupted Internal dataset. Fig. 9 presents the mean angular errors as the threshold on $\hat{c}$ varied. Similar to the results obtained from the MPIIGaze, we observed a large gap between errors of samples with low $\hat{c}$ and high $\hat{c}$.

4.5. Analysis

We evaluated the relationship between the pseudo-label c^* and the estimated label $\hat{c}$ at the end of training by 15-fold cross-validation trained on Corrupted MPIIGaze. Fig. 10 presents the histograms of c^* and $\hat{c}$ for all folds.

For the uncorrupted samples, the distributions of c^* and $\hat{c}$ were very similar. The distributions were both concentrated



(a) c^* of uncorrupted samples (b) $\hat{c}$ of uncorrupted samples

(c) c^* of corrupted samples (d) $\hat{c}$ of corrupted samples

Figure 10. The histograms of c^* and $\hat{c}$ for the uncorrupted and corrupted samples at the end of training on Corrupted MPIIGaze. The distribution of c^* and that of $\hat{c}$ is quite different for the corrupted samples.

around 1. This is not surprising, since the vast majority of samples are not outliers and have low estimation errors.

For the corrupted samples, their distributions were different. While 30.2% of the corrupted samples had $c^* > 0.9$, none of them had $\hat{c} > 0.9$ (the maximum value of $\hat{c} = 0.87$). Also, while only 61.7% of the samples had $c^* < 0.1$, 99.4% of the samples had $\hat{c} < 0.1$. The confidence pseudo-label c^* is not a reliable measure of whether or not a sample is an outlier. However, our network can still use this information to train a confidence score estimate $\hat{c}$ that can be used to detect outliers reliably. This is because the parameters of the network are chosen to minimize (3), not the difference between $\hat{c}$ and c^*.

5. Conclusions

Outliers (low-quality samples) in appearance-based gaze estimation are caused by factors such as blinks, occlusions, blur and failures of the facial landmark detection. We proposed an effective algorithm that learns to detect outliers during training of an appearance-based gaze estimator. Only gaze direction labels are required. Outlier labels are not required. This reduces manual work required in labelling the dataset. Outliers are assigned low confidence scores so that their impact on the trained network is reduced. In our experiments, this lead to a $7.3\% - 7.8\%$ reduction in error compared to a model trained without our proposed algorithm. The learned outlier detector was able to detect outliers reliably, with a precision 0.71 when the recall was 0.63.

One limitation of this work is that the confidence pseudo-label does not distinguish between outliers and samples where eyes are clearly visible and well localised but the estimation errors are high. These difficult samples should be very useful for the training, but their impact might be reduced by our current algorithm. Further improvement may be possible by better modelling this problem.

Appearance-based gaze estimation can play an important role in many real-world scenarios, e.g. human-robot interaction, and driver monitoring. The proposed algorithm can reduce the risk caused by low-quality samples, increasing the reliability of gaze-based control systems.

Acknowledgements

This work was supported in part by the Hong Kong Innovation and Technology Fund under Grant ITS/406/16FP.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66. IEEE, 2018.

[2] J Cech and T Soukupova. Real-time eye blink detection using facial landmarks. *21st Computer Vision Winter Workshop*, 2016.

[3] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[4] Zhaokang Chen and Bertram E Shi. Using variable dwell time to accelerate gaze-based web browsing with two-step selection. *International Journal of Human–Computer Interaction*, pages 1–16, 2018.

[5] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation via gaze decomposition and single gaze point calibration. *arXiv preprint arXiv:1905.04451*, 2019.

[6] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision*, pages 100–115. Springer, 2018.

[7] Haoping Deng and Wangjiang Zhu. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3162–3171. IEEE, 2017.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[10] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *European Conference on Computer Vision*, pages 334–352. Springer, 2018.

[11] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d

cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258. ACM, 2014.

[12] Alessandro Grillini, Daniel Ombelet, Rijul S Soans, and Frans W Cornelissen. Towards using the spatio-temporal properties of eye movements to classify visual field defects. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 38. ACM, 2018.

[13] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006.

[14] Dan Witzner Hansen and Qiang Ji. 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, 2009.

[15] Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations*, 2019.

[16] Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, 12:105, 2018.

[17] Guilei Hu, Yang Xiao, Zhiguo Cao, Lubin Meng, Zhiwen Fang, and Joey Tianyi Zhou. Towards real-time eyeblink detection in the wild: Dataset, theory and practices. *arXiv preprint arXiv:1902.07891*, 2019.

[18] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *ACM/IEEE International Conference on Human Robot Interaction*, pages 83–90. IEEE, 2016.

[19] Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems*, pages 1942–1950, 2017.

[20] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An ppen source platform for pervasive eye tracking and mobile gaze-based interaction. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1151–1160. ACM, 2014.

[21] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016.

[22] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations*, 2018.

[23] Dongze Lian, Lina Hu, Weixin Luo, Yanyu Xu, Lixin Duan, Jingyi Yu, and Shenghua Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2018.

[24] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018.

[25] Raphael Menges, Chandan Kumar, Daniel Müller, and Korok Sengupta. Gazetheweb: A gaze-controlled web browser. In *Proceedings of the Web for All Conference on The Future of Accessible Work*, page 25. ACM, 2017.

[26] Benjamin I Outram, Yun Suen Pai, Tanner Person, Kouta Minamizawa, and Kai Kunze. Anyorbit: Orbital navigation in virtual environments with eye-tracking. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 45. ACM, 2018.

[27] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision*, pages 721–738. Springer, 2018.

[28] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6):179, 2016.

[29] Jimin Pi and Bertram E Shi. Probabilistic adjustment of dwell time for eye typing. In *International Conference on Human System Interactions*, pages 251–257. IEEE, 2017.

[30] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Lightweight head pose invariant gaze tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2156–2164. IEEE, 2018.

[31] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial traininfg. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 271–280. ACM, 2013.

[34] Tereza Soukupova and Jan Cech. Real-time eye blink detection using facial landmarks. In *Computer Vision Winter Workshop*, 2016.

[35] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828. IEEE, 2014.

[36] Haofei Wang, Jimin Pi, Tong Qin, Shaojie Shen, and Bertram E Shi. SLAM-based localization of 3D gaze using a mobile eye tracker. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 65. ACM, 2018.

[37] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[38] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an

appearance-based gaze estimator from one million synthesised images. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, pages 131–138. ACM, 2016.

[39] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[40] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *European Conference on Computer Vision*, pages 456–474. Springer, 2018.

[41] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 12. ACM, 2018.

[42] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015.

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Its written all over your face: Full-face appearance-based gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2299–2308. IEEE, 2017.

[44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019.