Unsupervised Outlier Detection in Appearance-Based Gaze Estimation

Anonymous ICCV submission

Paper ID 38

Abstract

Appearance-based gaze estimation maps RGB images to estimates of gaze directions. One problem of 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, occlusion (by eye glasses), blur and failure of eye landmark detection. Training on these outliers degrades the performance of gaze estimator, 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 compared with baseline model trained without confidence score). The proposed algorithm is also able to detect outliers during testing, with a negative predictive value of 0.87 when the true negative rate is 0.5.

1. Introduction

Human gaze has been recognized as an important cue for inferring people's intent in many applications, such as human-computer interfaces [24, 28], human-robot interaction [17], virtual reality [25, 27], social behavioral analysis [15] and health care [11]. 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 [13]: model-based methods and appearance-based methods. Model-based methods mostly rely upon active illumination, e.g. infrared illumination used in pupil center corneal reflections (PCCR) [12]. 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 cheaper, easier to setup, calibration-free and more robust to head motion. Recently, the application of deep convolutional neural networks (CNNs) has reduced estimation error significantly [41]. There are a large number of high quality real and synthetic datasets [9, 10, 20, 31, 33, 35, 36, 37, 41]. Results based on these datasets 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, 5, 6, 20, 22, 29, 42].

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 occlusion, blur and failure of 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. [23, 8, 14, 21]. However, in most of this work, supervised learning scenarios are considered. Moreover, techniques for handling outliers in appearance-based gaze estimation have not been studied

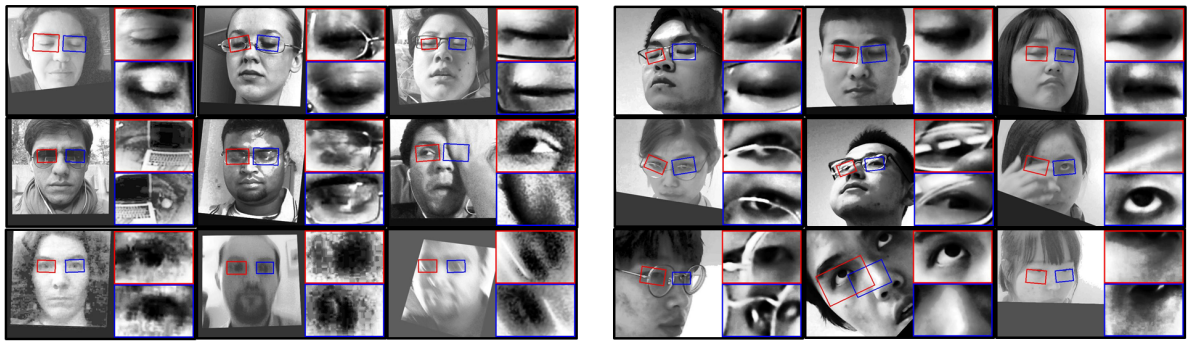


Figure 1. Outliers detected by our proposed algorithm in cross-subject experiments on the MPIIGaze dataset (left) and on an internal dataset (right). They are mainly caused by blinks, occlusion, blur and failure of the facial landmark detection.

previously to our knowledge.

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 successfully detect outliers 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 subtle intra-subject differences, large inter-subject differences, head poses and illuminations.

Past approaches to this problem have included k-Nearest Neighbors [35, 30], Support Vector Regression [30] and Random Forests [35]. More recently, 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 [41, 43]. They showed that deep CNNs improved accuracy significantly. To further improve the accuracy, Krafka *et al*. proposed a CNN with multi-region input (an image of the face, images of both eyes and a face grid) to estimate the gaze target on screens of mobile devices [20]. Zhang *et al*. proposed a network that takes the full face image as input and adopts a spatial weights method to emphasize features from particular face regions [42]. This work has shown that regions of the face other than the eyes also contain information about the gaze angle.

To further reduce estimation error, different directions have been explored. Some work has concentrated on the head-eye relationships [6, 29]. Estimation error was reduced by better utilizing the head pose information. Other work has focused on extracting better features from eye images, e.g., studying the "two eye asymmetry problem" [5], estimating the eye landmark locations and gaze directions jointly [39], learning an intermediate pictorial representation of the eyes [26], using dilated-convolutions to extract features at high resolution [3] and fusing information from images captured from multiple cameras [22].

2.2. Out-of-distribution detection

Out-of-distribution detection (OD detection or ODD) is an active research topic in the filed 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 [23]. Hendrycks *et al*. proposed to train anomaly detectors against an auxiliary dataset [14]. This two work 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 concentrated distribution for ID samples but uniform distribution for OD samples [21]. These methods use supervised learning for OD detection, where ID/OD labels are available for each image.

Our work is most similar to the work of DeVries and Taylor [8]. Their network learned confidence scores based on images, where an image that gives an incorrect prediction has low confidence score and is discounted in the loss function. We follow the same vein, where we use the mean squared error during training as the 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 [8] 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 the similar appearance as the normal samples.

In particular, we introduce a novel concept: the confidence pseudo-label. We use this pseudo label to design a proper loss function for appearance-based gaze estimation. We also use pseudo label 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 eye aspect ratio to indicate 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 uses long short-term memory (LSTM) to capture the temporal information [16]. However, these methods required labeled samples. Kassner *et al*. proposed to detect the pupil without supervision by a comprehensive algorithm depending on ellipse fitting [19].

Our proposed algorithm detects outliers in an unsupervised manner while training a gaze estimator, which reduces the burden of data labeling.

3. Methodology

3.1. Outliers

In this work, we refer to the samples in which the eyes are not fully visible as outliers, since these images have no or very limited information about the precise gaze directions. Some examples of outliers from the MPIIGaze dataset [41] and an internal dataset are illustrated in Fig. 1. Our analysis on these datasets indicate that outliers are mainly caused by blinks, occlusion, failure of the facial landmark detection and blur.

3.2. 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, since such outliers don't contain any information about gaze.

To alleviate the impact of outliers, we considered 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:

$$\sum_{i} \left[\hat{c}(x_i) e_i - \alpha \hat{c}(x_i) - \lambda \log\left(\hat{c}(x_i) \right) \right], \tag{3}$$

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

3.3. Architecture

The architecture of our proposed network is presented in Fig. 2. The general architecture is inspired by iTracker [20] 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 [4] to improve the performance of gaze estimation.

The input images x_i are first fed to three base CNNs. The architecture of the base CNN is shown in Fig. 2(d). It has five convolutional layers, one max-pooling layer and four dilated-convolutional layers [38] with different dilation rate, r. These dilated-convolutional layers can learn high-level features at high resolution to 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 this network by ϕ.

The outputs of the above three FC are concatenated and then fed to the gaze estimation branch and the confidence estimation branch. The gaze estimation branch has one FC and linear output layer (see Fig. 2(b)). The confidence estimation layer also has one hidden FC but uses a sigmoid function in the output layer (see Fig. 2(c)). We denote the parameters of the gaze estimation branch by θ and that of the confidence estimation branch by ψ.

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 weights of the first four convolutional layers are transfered from VGG-16 [32] pre-trained on the ImageNet

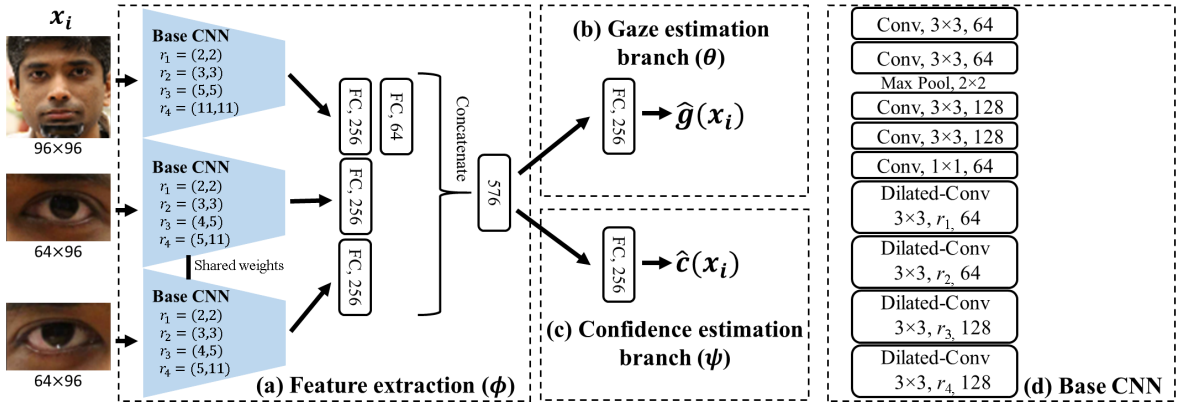


Figure 2. Architecture of the proposed network. (a) The main network that outputs $\hat{g}(x_i)$ based on the input image x_i. (b) The gaze estimation branch. (c) The confidence estimation branch. (d) The base CNN is the basic component of (a). (c) The regular CNN without dilated-convolution as baseline. FC denotes fully-connected layers, Conv denotes convolutional layers, Dilated-Conv denotes dilated-convolutional layers with r as the dilation rate.

dataset [7], while the others are trained from scratch. Batch renormalization layers [18] are applied to all layers trained from scratch. Dropout layers with dropout rates of 0.5 are applied to all FC layers.

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.

3.4. Training procedure

During training, we use different minibatches to train the gaze estimation branch and the confidence estimation branch. For the gaze estimation branch, we sample minibatches uniformly from the training set. For the confidence estimation branch, we try to balance the number of normal and outlier samples within a minibatch, since the number of outlier samples is far less than that of normal samples. Otherwise, the trained network would be strongly biased. Since we don't have outlier labels, we introduce a pseudo label $c_i^* \in [0, 1]$ for each sample x_i calculated based on the gaze estimation accuracy. The larger the $c_i^*$the higher the confidence we believe in the corresponding gaze estimate. The definition of c_i^* will be further explained in the next subsection.

We balance the samples according to their pseudo labels c_i^*. 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 a minibatch. We fix λ and dynamically adjust α in the loss function (3) during training. We fix λ to have a steady transition slope between normal and outlier samples. We adjust α to maintain the ratio of samples with $c_i^* < 0.5$ within a certain range.

In all our experiments, we set $\lambda = 5e^{-4}$ after grid search. We set the initial value of $\alpha = (\frac{\pi}{45})^2$ rad^2, update α during training to maintain the percentage of training samples with $c_i^* < 0.5$ to be between $TH_{\text{low}} = 5\%$ and $TH_{\text{high}} = 15\%$. The procedure is presented Algorithm 1.

Algorithm 1 Gaze Estimation with Outlier Detection

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_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 (5) based on (e_i, λ, α)
      if $c_i^* \geq 0.5$ **then**
        $\mathcal{S}_t^p$.append((x_i, g_i))
      else
        $\mathcal{S}_t^n$.append((x_i, g_i))
      end if
    end for
    Update (ϕ, θ) on s minimizing (3) by gradient descent
    if $t > 1$ **then**
      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) by gradient descent
    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

3.5. Confidence Pseudo label

Our proposed algorithm is designed based on two assumptions. First, we assume e_i is a good indicator of the quality of the input sample. During training, we assume normal samples to have low estimation error and the outliers to have estimation error that is on average large and that has large variance, since the network would generates unreliable outputs for them. An input sample with large e_i is more likely to be an outlier. Second, we assume that a deep network can learn to distinguish between normal samples and outlier samples.

We define pseudo label of sample x_i, c_i^* or $c^*(e_i)$, as the solution of the following optimization problem:

$$c_i^* = \underset{c}{\operatorname{argmin}}\, J(e_i, c) \quad \text{subject to} \quad 0 \leq c \leq 1 \tag{4}$$

where J is the differentiable loss function defined in (3). In this case, the pseudo label c_i^* has the following closed-form solution:

$$c_i^* = \begin{cases} 1 & if\ e_i \leq \alpha + \lambda \\ \frac{\lambda}{e_i - \alpha} & if\ e_i > \alpha + \lambda \end{cases} \tag{5}$$

As the gaze estimation error e_i increases, pseudo label c_i^* decreases. This is consistent with our first assumption described above. The hyperparameter α in linear penalty controls the rate that c_i^* decreases with e_i^*. 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 value of e_i and to have control of the rate of decrease for larger value of e_i.

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

3.6. Preprocessing.

We use the data normalization method introduced in [40]. 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.

4. Experiments

We evaluated our proposed algorithm on two datasets: MPIIGaze dataset and an internal dataset. We use a modified version of MPIIGaze dataset, where we create 10% corrupted images as the outliers. For the internal dataset, it contains outliers due to blinks, occlusion by eye glasses and failure of the landmark detection as it was recorded by video.

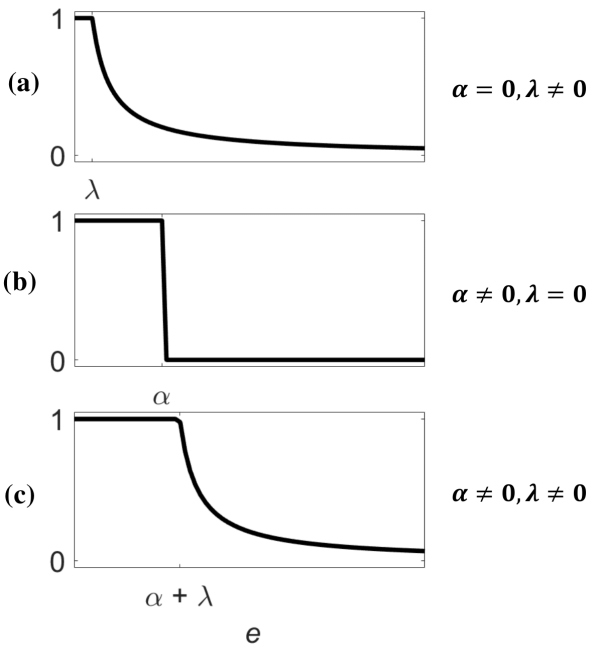


Figure 3. Pseudo label c^* vs. error e for different hyperparameter settings.

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

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

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 original samples, and this original dataset as **Clean MPIIGaze**.

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

In Task I, we trained on the Corrupted MPIIGaze and tested on the Clean MPIIGaze to evaluate the gaze estimation accuracy. In Task II, we trained on the Corrupted MPIIGaze and tested on the Corrupted MPIIGaze to evaluate the performance of outlier detection. Subjects used in training and testing were different.

Internal dataset. This dataset contains full face videos

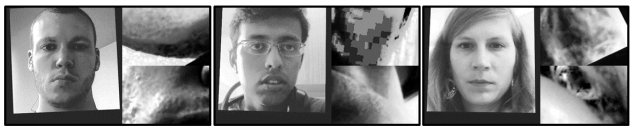


Figure 4. Examples of the generated corrupted samples of the Corrupted MPIIGaze.

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

of 21 subjects (10 female, 10 with glasses). It contains significant variances in head pose and face locations in the images. Some example images are presented in Fig. 5. OpenFace [1] is used for facial landmark detection. We sample images at 10 fps.

We used this dataset to create two new datasets. The first dataset, **Clean Internal**, contained all images in the dataset except those filtered out using some removal algorithms. We removed images whose confidence for the landmark detection given by OpenFace was lower than 0.02 or that with significantly abnormal landmarks. We also remove 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 images in which the faces are at the lower regions in the images. We use this set because it contains more samples that fail the landmark detection. It contains $185,357$ images in total (about $8,800$ images per subject).

Similar to MPIIGaze, in Task I we trained on the Corrupted Internal and tested on the Clean Internal. In Task II we trained on the Corrupted Internal and tested on the Corrupted Internal.

4.2. Dataset labeling

To evaluate the performance of outlier detection, we labeled a subset of each dataset. We label each sample whether it is normal or of low-quality. For the MPIIGaze dataset, we labeled the original samples whose estimated confidence scores given by our proposed method are 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 the Corrupted Internal by downsamping it from 10 fps to 1 fps. We labelled $18,535$ images in total, among which $1,326$ (7.15%) samples were labeled as outliers.

4.3. Results - MPIIGaze dataset

We conducted 15-fold leave-one-subject-out cross-validation. We trained our proposed method on the Corrupted MPIIGaze.

Performance of gaze estimation. We compared our proposed method with two baselines without confidence estimation: one was trained on the Clean MPIIGaze, and the other was trained on the Corrupted MPIIGaze. The mean angular errors over 15 subjects tested on Clean MPIIGaze are presented in Table 1. When trained on the Corrupted MPIIGaze, our proposed confidence estimation reduced the gaze estimation error from $5.1°$ to $4.7°$ (7.8%).

Without confidence estimation, the gaze estimation error is degraded from $4.5°$ to $5.1°$ when the corrupted images are added into training set. This indicates that including low-quality samples in the training set significantly degrades the performance. Our proposed method alleviates the degradation from $0.6°$ to $0.2°$. This small degradation is partly because that some normal samples were assigned low confidences as their training errors were high.

Outlier detection. We first tested on the Corrupted MPIIGaze. We compared the $\hat{c}$ of the original samples and corrupted samples. Overall, most values of $\hat{c}$ for the original samples were close to one, while most values of $\hat{c}$ for the corrupted samples were close to zero. For the original samples, over 95% samples have $\hat{c} \geq 0.9$. For the corrupted samples, except one sample with $\hat{c} = 0.69$, all of the others have $\hat{c} < 0.5$.

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 the Clean MPIIGaze.

Training set	Confidence estimation	Mean error
Corrupted Internal	*No*	$3.9°$
Corrupted Internal	*Yes*	$3.6°$

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

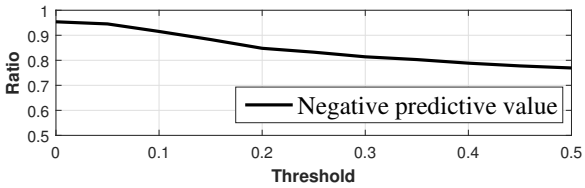


Figure 6. Negative predictive value as a function of the threshold of $\hat{c}$ tested on the original samples of the MPIIGaze.

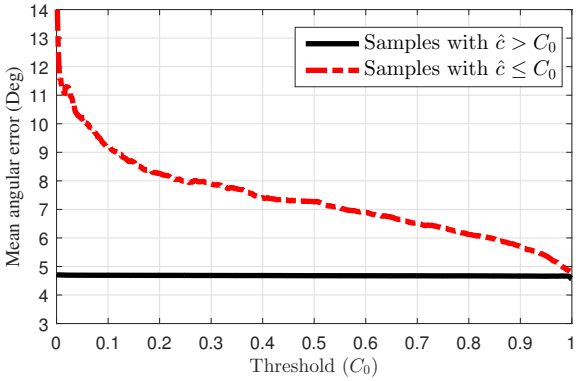


Figure 7. Mean angular error of predicted positive/negative samples as a function of decision threshold C_0 on the Clean MPIIGaze.

We then tested on the Clean MPIIGaze. Among the $45,000$ original 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 manually labeled the 659 detected outliers, and found that 507 out of 659 samples were true negatives, i.e., the negative predictive value (NPV) is 76.9%. We also present the NPV as a function of the threshold of $\hat{c}$ in Fig. 6. In general, the smaller the threshold, the higher the NPV. The NPV 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 estimation error and $\hat{c}$. Fig 7 presents the mean angular error of predicted positive/negative samples as a function of dicision 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 $\hat{c} > C_0$. These results indicate that the outliers detected by our proposed algorithm indeed have large estimation errors.

4.4. Results - Internal dataset

We conducted five-fold cross-subject cross-validation trained on the Corrupted Internal.

Performance of gaze estimation. We tested on the Clean Internal. We trained a network without confidence estimation as baseline. The mean angular errors over subjects are presented in Table 2. Our proposed method achieved an error of $3.6°$, which was $0.3°$ (7.7%) lower than the $3.9°$ achieved by the baseline method.

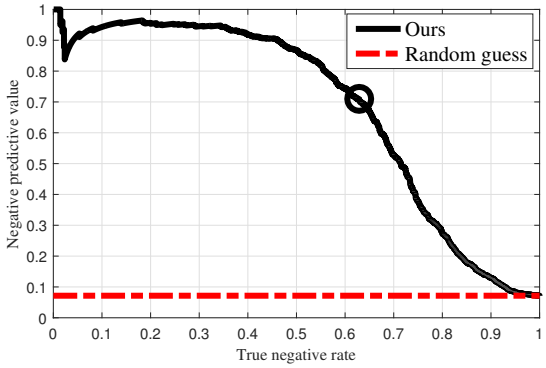


Figure 8. Negative predictive value as a function of true negative rate on the Corrupted Internal. The circle indicates their best harmonic average (0.67), which is obtained when the threshold is 0.12.

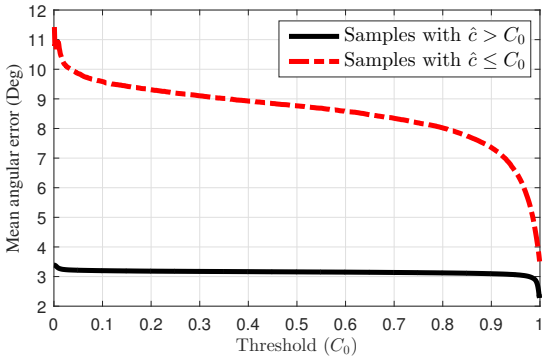
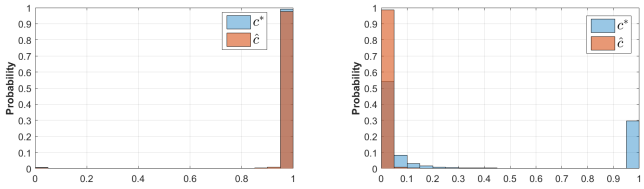


Figure 9. Mean angular error of predicted positive/negative samples as a function of decision threshold C_0 on the Corrupted Internal.

Outlier detection. We first tested on the subset of Corrupted Internal which was labelled. We use random guess as baseline, whose NPV was 7.15%. Fig. 8 presents NPV as a function of true negative rate (TNR), which is similar to a precision-recall curve. The area-under-curve (AUC) is 0.68. The circle on the black curve indicates the position of the best harmonic average of NPV and TNR (0.67). The results indicate that our proposed method had a good performance on outlier detection. For example, when the threshold was 0.5, the NPV was 0.87 and the TNR was 0.5. The best harmonic average was achieved when the threshold was 0.12.

We then tested on the whole Corrupted Internal. Fig. 9 presents the mean angular errors as the threshold on $\hat{c}$ var-



(a) Original samples (b) Corrupted samples

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

ied. Similar to the results obtained from the MPIIGaze, we observed a large gap between errors of samples with low $\hat{c}$ and that with 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 the Corrupted MPIIGaze. Fig. 10 presents the histograms of c^* and $\hat{c}$ for all folds.

For the original samples, the distributions of c^* and $\hat{c}$ are very similar. However, for the corrupted samples, their distributions are different. While over 30.2% of the samples have $c^* > 0.9$, the maximum value of $\hat{c} = 0.87$. Also, while only 61.7% samples have $c^* < 0.1$, over 99.4% samples have $\hat{c} < 0.1$. These results demonstrate that the estimates errors e_i, which determines c_i^* according to (5), is not a reliable indicator of outliers. Nonetheless, the neural network that estimates $\hat{c}$ learns to correctly assign low confidence values for outliers. This is because the parameters of the network are chosen to minimize (3), not the difference between $\hat{c}$ and c^*.

5. Conclusions

We focused on detecting outliers (low-quality samples) in appearance-based gaze estimation. These are caused by factors such as blinks, occlusions, blur and failures of the landmark detection. We proposed an effective algorithm that detects outliers during training of the appearance-based gaze estimator, where only labels of gaze directions are available. Outliers are assigned low confidence scores so that their impact on the trained network was reduced, leading to a $7.7\%-7.8\%$ reduction in error compared to a model trained without our proposed algorithm. During testing, the learned outlier detector was able to detect outliers reliable, with a negative predictive value 0.87 when the true negative rate was 0.5.

One limitation of this work is that we did not explicitly distinguish between outliers and samples that are difficult to obtain accurate gaze estimates for, but where eyes are clearly visible and well localised. These difficult samples should be very useful for the training, but their impact might be deducted by our current algorithm. We believe that a further improvement can be achieved by better modelling this problem.

We believe that 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 system.

References

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66. IEEE, 2018. 5, 6

[2] J. Cech and T. Soukupova. Real-time eye blink detection using facial landmarks. *21st Computer Vision Winter Workshop*, 2016. 6

[3] Z. Chen and B. E. Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018. 1, 2, 3

[4] Z. Chen and B. E. Shi. Appearance-based gaze estimation via gaze decomposition and single gaze point calibration. *arXiv preprint arXiv:1905.04451*, 2019. 3

[5] Y. Cheng, F. Lu, and X. Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision*, pages 100–115. Springer, 2018. 1, 2

[6] H. Deng and W. 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. 1, 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. 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. 4

[8] T. DeVries and G. W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 1, 2, 3

[9] T. Fischer, H. J. Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *European Conference on Computer Vision*, pages 334–352. Springer, 2018. 1

[10] K. A. Funes Mora, F. Monay, and J.-M. 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. 1

[11] A. Grillini, D. Ombelet, R. S. Soans, and F. 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. 1

[12] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006. 1

[13] 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*, 32(3):478–500, 2009. 1

[14] D. Hendrycks, M. Mazeika, and T. G. Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations*, 2019. 1, 2

[15] S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, 12:105, 2018. 1

[16] G. Hu, Y. Xiao, Z. Cao, L. Meng, Z. Fang, and J. T. Zhou. Towards real-time eyeblink detection in the wild: Dataset, theory and practices. *arXiv preprint arXiv:1902.07891*, 2019. 3

[17] C.-M. Huang and B. Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *ACM/IEEE International Conference on Human Robot Interaction*, pages 83–90. IEEE, 2016. 1

[18] S. Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems*, pages 1942–1950, 2017. 4

[19] M. Kassner, W. Patera, and A. 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. 3

[20] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016. 1, 2, 3

[21] K. Lee, H. Lee, K. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations*, 2018. 1, 2

[22] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2018. 1, 2

[23] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018. 1, 2

[24] R. Menges, C. Kumar, D. Müller, and K. 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. 1

[25] B. I. Outram, Y. S. Pai, T. Person, K. Minamizawa, and K. 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. 1

[26] S. Park, A. Spurr, and O. Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision*, pages 721–738. Springer, 2018. 2

[27] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6):179, 2016. 1

[28] J. Pi and B. E. Shi. Probabilistic adjustment of dwell time for eye typing. In *International Conference on Human System Interactions*, pages 251–257. IEEE, 2017. 1

[29] R. Ranjan, S. De Mello, and J. Kautz. Light-weight head pose invariant gaze tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2156–2164. IEEE, 2018. 1, 2

[30] T. Schneider, B. Schauerte, and R. Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *International Conference on Pattern Recognition*, pages 1167–1172. IEEE, 2014. 2

[31] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. 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. 1

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[33] B. A. Smith, Q. Yin, S. K. Feiner, and S. 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. 1

[34] T. Soukupova and J. Cech. Real-time eye blink detection using facial landmarks. In *Computer Vision Winter Workshop*, 2016. 3

[35] Y. Sugano, Y. Matsushita, and Y. 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. 1, 2

[36] K. Wang, R. Zhao, and Q. 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. 1

[37] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. 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. 1

[38] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[39] Y. Yu, G. Liu, and J.-M. Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *European Conference on Computer Vision*, pages 456–474. Springer, 2018. 2

[40] X. Zhang, Y. Sugano, and A. Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, page 12. ACM, 2018. 5

[41] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015. 1, 2, 3

[42] X. Zhang, Y. Sugano, M. Fritz, and A. 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. 1, 2

[43] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019. 2