iTracker(2016)

paper: Eye Tracking for Everyone

Network Structure

Similar to AlexNet

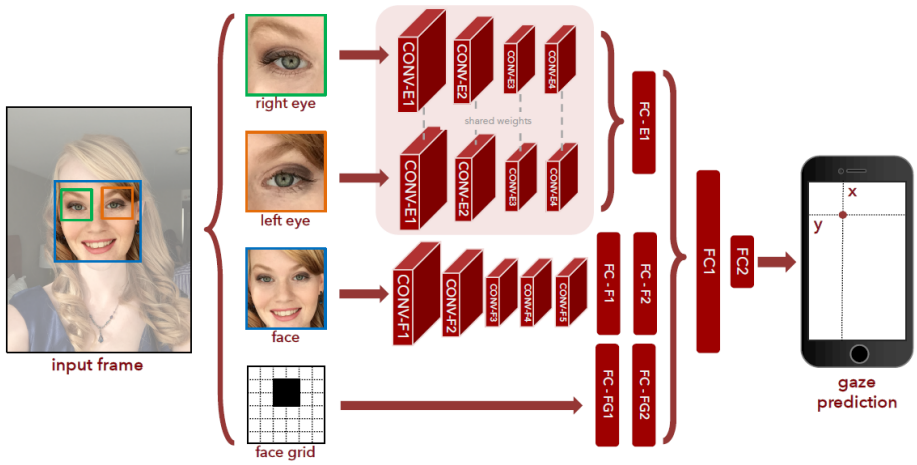


Figure 5: Overview of iTracker, our eye tracking CNN. Inputs include left eye, right eye, and face images detected and cropped from the original frame (all of size 224×224). The face grid input is a binary mask used to indicate the location and size of the head within the frame (of size 25×25). The output is the distance, in centimeters, from the camera. CONV represents convolutional layers (with filter size/number of kernels: CONV-E1, CONV-F1: $11 \times 11/96$, CONV-E2, CONV-F2: $5 \times 5/256$, CONV-E3, CONV-F3: $3 \times 3/384$, CONV-E4, CONV-F4: $1 \times 1/64$) while FC represents fully-connected layers (with sizes: FC-E1: 128, FC-F1: 128, FC-F2: 64, FC-FG1: 256, FC-FG2: 128, FC1: 128, FC2: 2). The exact model configuration is available on the project website.

Input & Output

- Input Data: Three images and one face grid:
- output: (x,y) axis

Loss function

using a Euclidean loss on the x and y gaze position.

Evaluation

Metric

report the error in terms of average Euclidean distance (in centimeters) from the location of the true fixation.

Results

Model	Aug.	Mobile phone		Tablet	
		error	dot err.	error	dot err.
Baseline	tr + te	2.99	2.40	5.13	4.54
iTracker	None	2.04	1.62	3.32	2.82
iTracker	te	1.84	1.58	3.21	2.90
iTracker	tr	1.86	1.57	2.81	2.47
iTracker	tr + te	1.77	**1.53**	2.83	2.53
iTracker*	tr + te	**1.71**	**1.53**	**2.53**	**2.38**
iTracker (no eyes)	None	2.11	1.72	3.40	2.93
iTracker (no face)	None	2.15	1.69	3.45	2.92
iTracker (no fg.)	None	2.23	1.81	3.90	3.36

Table 2: Unconstrained eye tracking results (top half) and ablation study (bottom half). The error and dot error values are reported in centimeters (see Sec. 5.1 for details); lower is better. *Aug.* refers to dataset augmentation, and *tr* and *te* refer to train and test respectively. *Baseline* refers to applying support vector regression (SVR) on features from a pre-trained ImageNet network, as done in Sec. 5.4. We found that this method outperformed all existing approaches. For the ablation study (Sec. 5.5), we removed each critical input to our model, namely eyes, face and face grid (*fg.*), one at a time and evaluated its performance.

- compare to other models

  Data set: TabletGaze

Method	Error	Description
Center	7.54	Simple baseline
TurkerGaze [41]	4.77	pixel features + SVR
TabletGaze	4.04	Our implementation of [13]
MPIIGaze [43]	3.63	CNN + head pose
TabletGaze[13]	3.17	Random forest + mHoG
AlexNet [20]	3.09	eyes (`conv3`) + face (`fc6`) + fg.
iTracker (ours)	**2.58**	`fc1` of iTracker + SVR

Table 4: Result of applying various state-of-the-art approaches to TabletGaze [13] dataset (error in cm). For the AlexNet + SVR approach, we train a SVR on the concatenation of features from various layers of AlexNet (`conv3` for eyes and `fc6` for face) and a binary face grid (*fg.*).

Tips

- we include the **eyes as individual inputs** into the network (even though the face already contains them) to provide the network with a higher resolution image of the eye to allow it to identify subtle changes.
- we found **finetuning the network to each device and orientation** helpful.
- Interestingly, we find that the **AlexNet + SVR** (applying support vector regression (SVR) to image features extracted using AlexNet pre-trained on ImageNet) approach outperforms all existing state-of-the-art approaches despite the features being trained for a completely different task.
- We find that the error decreases significantly as **the number of subjects is increased**, illustrating the importance of gathering a large-scale dataset.

code

MPIIGaze (GazeNet 2019)

MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation

Network Structure

Refer to VGG

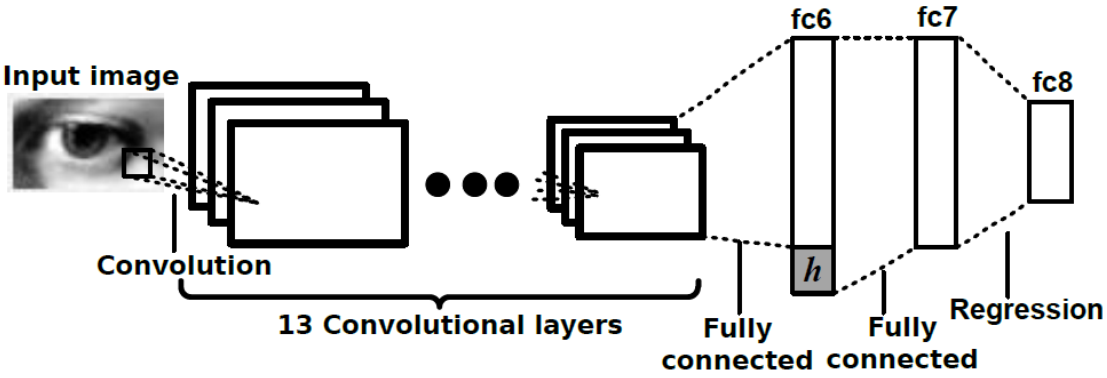


Fig. 10: Architecture of the proposed GazeNet. The head angle h is injected into the first fully connected layer. The 13 convolutional layers are inherited from a 16-layer VGG network [66].

based on the 16-layer VGGNet architecture that includes 13 convolutional layers, two fully connected layers, and one classification layer with five max pooling layers in between.

- We changed the stride of the first and second pooling layer from two to one to reflect the smaller input resolution.
- we injected head pose information h into the first fully connected layer (fc6)

We used the weights of the 16-layer VGGNet [66] pre-trained on ImageNet for all our evaluations, and fine-tuned the whole network in 15,000 iterations with a batch size of 256 on the training set. We used the Adam solver [71] with the two momentum values set to 1 = 0:9 and 2 = 0:95. An initial learning rate of 0.00001 was used and multiplied by 0.1 after every 5,000 iterations.

Input & Output

- input: 2D head angle h and eye image e
  - eye imag: a grey-scale single channel image with a resolution of 60*36 pixels.
  - head angle h
- output: a 2D gaze angle vector g, consisting of two gaze angles, gyaw and gpitch.

Loss function

the sum of the individual L2 losses measuring the distance between the predicted ^g and true gaze angle vector g.

Evaluation

Metric

Mean error? (not very clear)

Results

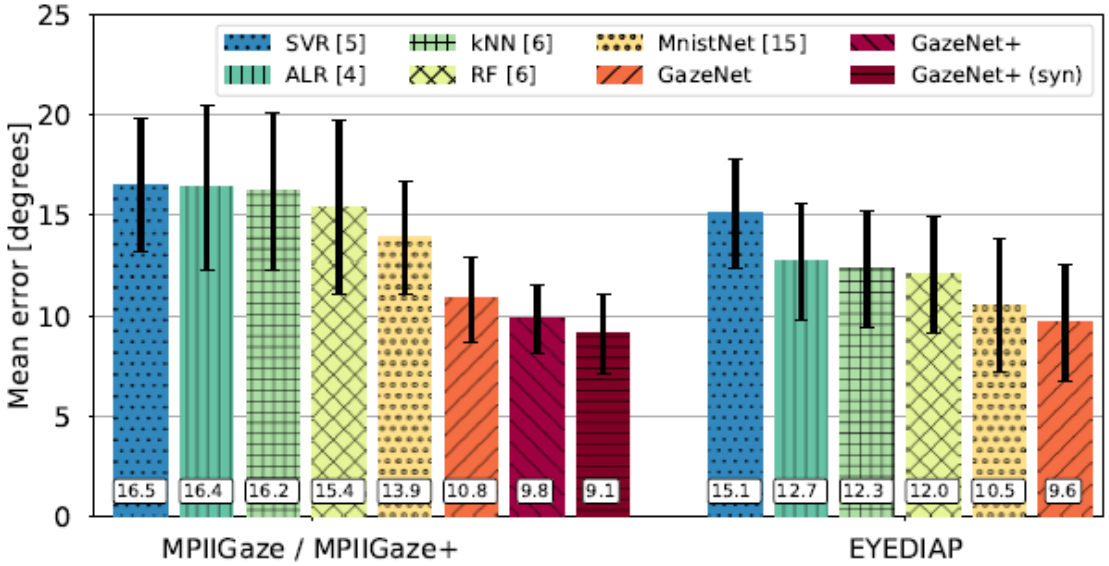


Fig. 11: Gaze estimation error for cross-dataset evaluation with training on 64,000 eye images in UT Multiview and testing on 45,000 eye images of MPIIGaze or MPIIGaze+ (left) and EYEDIAP (right). Bars show mean error across all participants; error bars indicate standard deviations.

Tips

- one of the most important challenges for unconstrained gaze estimation is **differences in gaze ranges** between the training and testing domains. Although handling the different gaze angles has been researched by combining geometric and appearance-based methods [75], it is still challenging for appearance-based gaze estimation methods.
- In GazeNet, we do not **use pupil centre** information as input. Although intuitively, eye shape features, such as pupil centres, can be a strong cue for gaze estimation. While there was an improvement between the models without and with the pupil centre feature, the improvement was relatively small (from 5.4 to 5.2 degrees). Performance improved more when using the manually annotated pupil centres, but still not significantly (5.0 degrees).
- use **both of eyes**.

ARE-Net(2018)

Network Structure

base-CNN structure

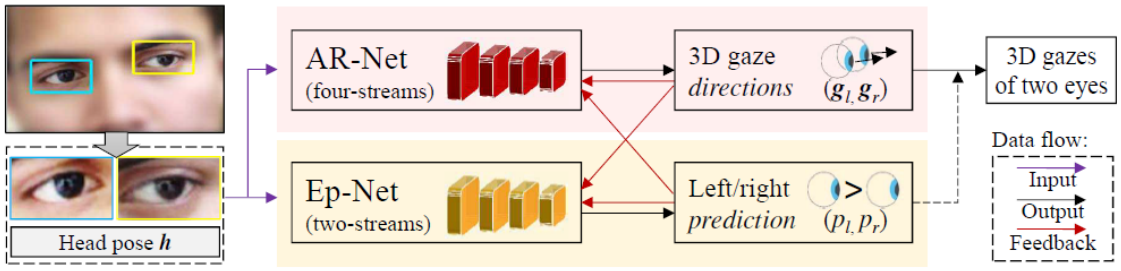


Fig. 1. Overview of the proposed Asymmetric Regression-Evaluation Network (ARE-Net). It consists of two major sub-networks, namely, the AR-Net and the E-Net. The AR-Net performs asymmetric regression for the two eyes, while the E-Net predicts and adjust the asymmetry to improve the gaze estimation accuracy.

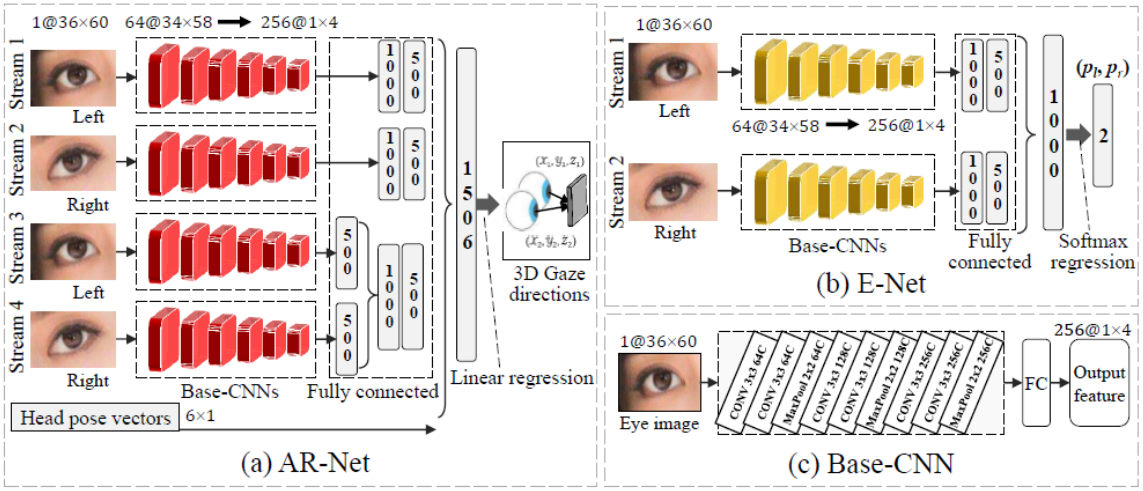


Fig. 2. Architecture of the proposed networks. (a) The AR-Net is a four-stream network to produce features from both the eye images. A linear regression is used to estimate the 3D gaze directions of the two eyes. (b) The E-Net is a two-stream network for two eye evaluation. The output is a two-dimensional probability vector. (c) The base-CNN is the basic component to build up the AR-Net and the E-Net. It uses an eye image as input. The output is a 1000D feature after six convolutional layers.

Input& Output

- input: two eye images and head pose vector
  - images: with a fixed resolution of 36 × 60.
- output: 3D gaze directions (x,y,z)

Loss function

- AR-Net:

左右两只眼的调和平均, angular error

$$l_{AR} = \frac{2e_l e_r}{e_l + e_r}$$

$$e_r = \arccos \frac{g_l f_l(x)}{\|g_l\| \|f_l(x)\|}$$

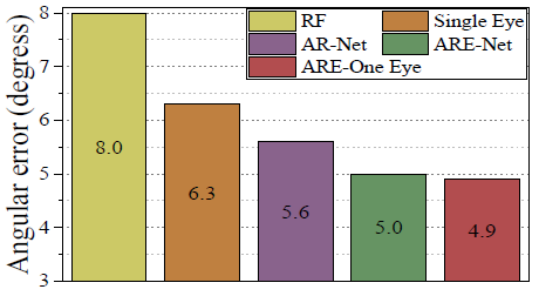
- E-Net: 某种意义上的cross-entropy

Evaluation

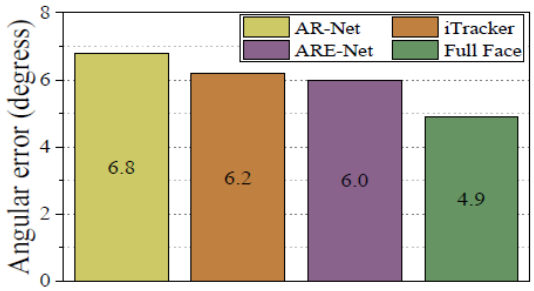
Metric

angular error

Results



(a) v.s. eye image-based methods.



(b) v.s. full face image-based methods.

shows better results than iTracker.

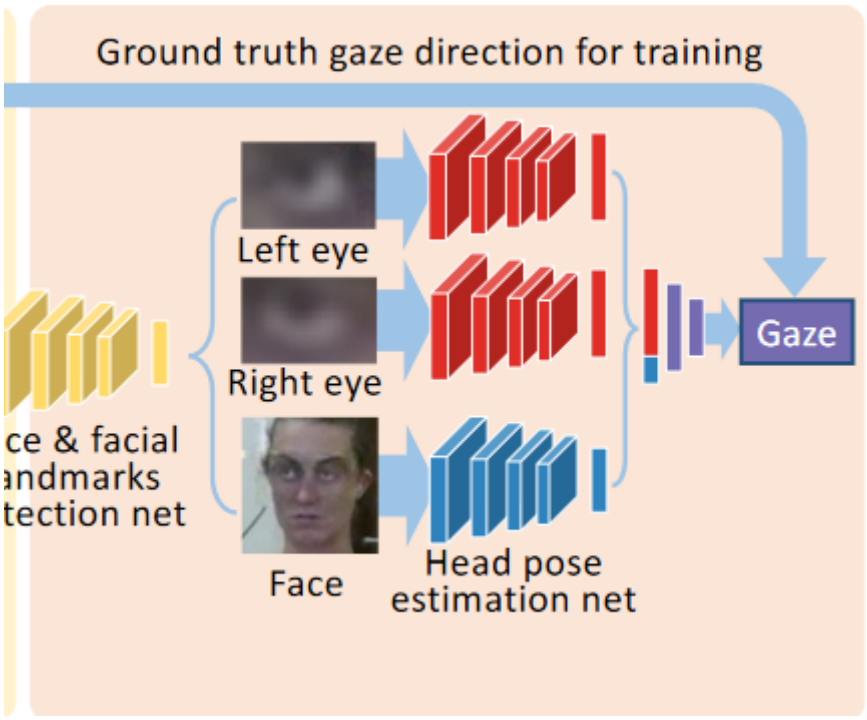
Tips

- Existing gaze regression methods handles the two eyes indifferently. However, in practice, we observe the **two eye asymmetry** regarding the regression accuracy.
    - Observation. At any moment, we cannot expect the same accuracy for two eyes, and either eye has a chance to be more accurate.

RT-GENE(2018)

Network Structure

VGG-16



The eye patches are fed separately

1. to VGG-16networks which perform feature extraction. Each VGG-16 network is followed by a fully connected (FC) layer of size 512 after the last max-pooling layer,followed by batch normalization and ReLU activation.
2. We then concatenate these layers, resulting in a FC layer of size 1024. This layer is followed by another FClayer of size 512.
3. We append the head pose vector to this FC layer, which is followed by two more FC layers of size 256 and 2 respectively.

The weights of the VGG-16 models are initialized using a pre-trained model on ImageNet.

The weights of the FC layers are initialized using the Xavier initialization.

We use the Adam optimizer with learningrate0.001,$\beta 1$= 0.9,$\beta 2$= 0.95and a batch size of256.

Input & Output

- input: two eye images, face image, head vector
- output: (yaw,pitch)

Loss function

l2 loss

Evaluation

Metrics

Results

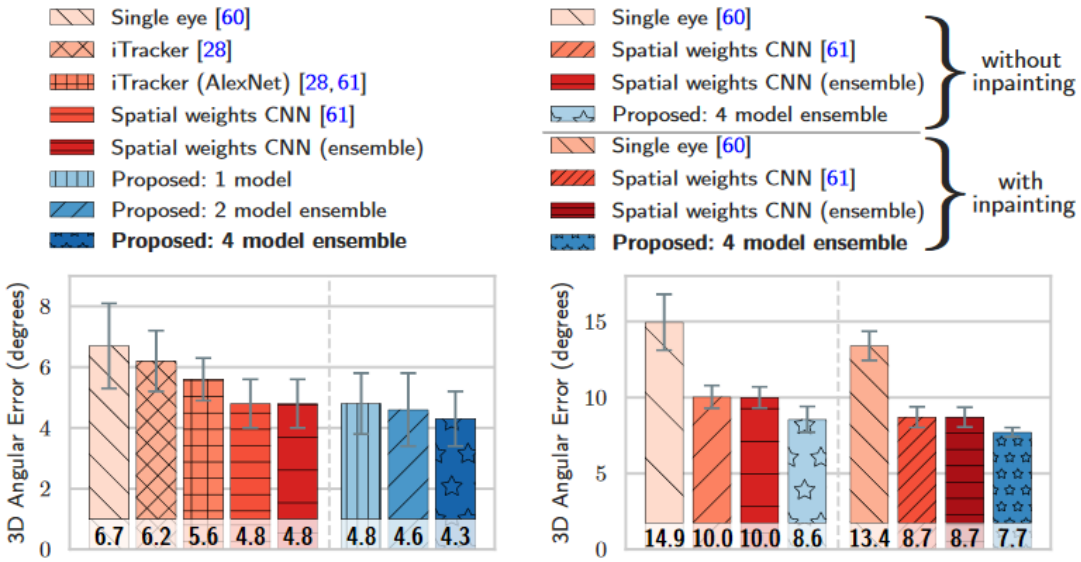


Fig. 7. Left: 3D gaze error on the MPII Gaze dataset. Right: 3D gaze error on our proposed gaze dataset. The inpainting improves the gaze estimation accuracy for all algorithms. Our proposed method performs best with an accuracy of 7.7 degrees.

Tips

- Image augmentation:
  - reduce the image resolution to1/2and1/4of its original resolution, followed by a bilinear interpolation to retrieve two augmented images of the original image size
  - to cover various lighting conditions, we employ histogram equalization. ???
  - convert color images to gray-scale images so that gray-scale images can be used as input as well
- For increased robustness, we use an ensemble scheme [29] where the mean of the predictions of the individual networks represents the overall prediction
- **ensemble schemes** were found to be particularly effective in our architecture. For a fair comparison, we also applied the ensemble scheme to the state-of-the-art method [61]. However, we did not observe any performance improvement over the single model (see Figure7). We assume that this is due to the spatial weights scheme that leads to similar weights in the intermediate layers of the different models. This results in similar gaze predictions of the individual models, and therefore an ensemble does not improve the accuracy for [61].

Pictorial(2018)

Network structure

the stacked hourglass architecture from Newell.

The hourglass architecture has been proven to be effective in tasks such as human pose estimation and facial landmarks detection

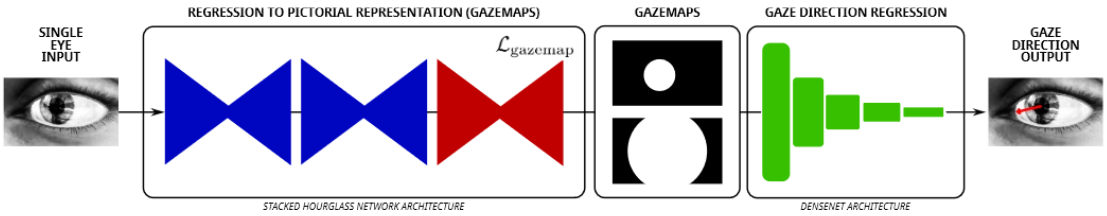


Fig. 1. Our sequential neural network architecture first estimates a novel pictorial representation of 3D gaze direction, then performs gaze estimation from the minimal image representation to yield improved performance on MPIIGaze, Columbia and EYEDIAP.

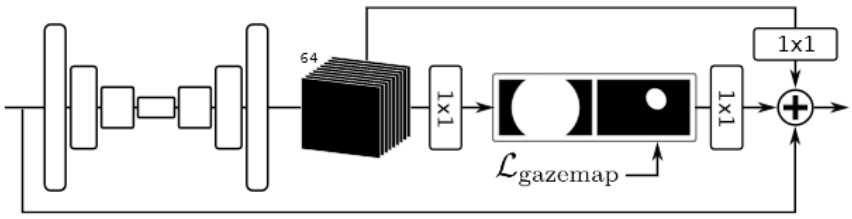


Fig. 3. Intermediate supervision is applied to the output of an hourglass module by performing 1×1 convolutions. The intermediate gazemaps and feature maps from the previous hourglass module are then concatenated back into the network to be passed onto the next hourglass module as is done in the original Hourglass paper [21].

- we use 3 hourglass modules with inter-mediate supervision applied on the gazemap outputs of the last module only.
- we select DenseNet which has recently been shown to perform well on image classification tasks for the final regression.
    - consists of 5 dense blocks (5 layers per block) with a growth-rate of 8, bottleneck layers, and a compression factor of 0.5.
    - This results in just 62 feature maps at the end of the DenseNet, and subsequently 62 features through global average pooling.
    - Finally, a single linear layer maps these features tog. The resulting network is light-weight and consists of just 66k trainable parameters.

input & output

- input: one-eye
- output: (yaw, pitch)

Loss function

- for the gazemap net (cross-entropy)

  Lgazemap=−α∑p∈Pm(p) logˆm(p),
- for regression: L2 loss

Training

leave-one-person-out evaluation

Evaluation

Metrics

Mean gaze estimation error in degrees

Results

(a) MPIIGaze (15-fold)

Model	kNN [47]	RF [47]	[45]	AlexNet	VGG-16	GazeNet [47]	**ours**
# params	0	-	1.8M	86M	158M	90M	0.7M
Inputs	e + h	e + h	e + h	e	e	e + h	e
Error	7.2	6.7	6.3	5.7	5.4	5.5	**4.5**

Tips

- we explicitly introduce a gaze-specific prior into the net-work architecture via **gazemaps**. This representation is formed of two boolean maps,which can be regressed by a fully convolutional neural network. We hypothesize that it is possible to learn **an intermediate image representation** of the eye. We reformulate the task of gaze estimation into two concrete tasks:
    1. reduction of input image to minimal normalized form (gazemaps),
    2. gaze estimation from gazemaps.