Gaze Estimation with Multi-Scale Channel and Spatial Attention

Song Liu
Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
+8615086671029
liusong@scieye.ac.cn

Danping Liu
School of Microelectronics and Communication Engineering, Chongqing University
+8613638392668
ldp@cqu.edu.cn

Haiyang Wu
Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
+8615111940626
wuhaiyang@cigit.ac.cn

ABSTRACT
Gaze estimation is well established as a significant research topic in computer vision given its importance for different applications. Recent studies demonstrate that other regions of the face beyond the two eyes contain valuable information for gaze estimation. Motivated by these works, we propose a novel and powerful deep convolutional network with multi-scale channel and spatial attention, which only takes the full-face image as input without additional modules to detect the eyes and estimate the head pose, to handle the gaze estimation task. It uses multi-scale channel and spatial information to adaptively select and increase important features and suppress some unnecessary facial regions which may not contribute to estimate gaze. By rigorously evaluating our module, we show that our method significantly outperforms the state-of-the-art for 3D gaze estimation on multiple public datasets.

CCS Concepts
• Human-centered computing→ Human computer interaction (HCI) → Interaction techniques → Pointing. • Computing methodologies→ Artificial intelligence→Computer vision.

Keywords
Eye tracking; attention mechanism; appearance-based gaze estimation; machine learning.

1. INTRODUCTION
Gaze estimation is well established as a significant research topic in computer vision given its importance for different applications, including human-computer interaction [1-4], medical diagnoses [6], psychological studies [5], VR/AR devices for controlling [7] and foveated rendering [8]. A variety of solutions proposed over the past decade can be divided into two categories: model-based and appearance-based. Early model-based methods are typically designed to exploit the mechanism of corneal reflection [23, 3, 25-28] or detect eye features and eye shapes [29-31, 61] such as pupil centers and iris edges to infer gaze direction. However, such methods require settings in which lighting conditions or head pose could be controlled, having high failure rate under real-world conditions. In line with the recent success of deep learning methods, many effective appearance-based techniques [15-17, 48, 62-64] using convolutional neural networks (CNN) were proposed to estimate gaze direction. It can work with a single ordinary camera to capture the eye appearance, then learn a mapping function to predict the gaze direction from the eye appearance directly in everyday real-world environments. In spite of these advances, most of these appearance-based methods that take eye images or both eye and face images as input to estimate gaze have two limitations: 1) an additional module is required to detect the eyes; 2) an additional module is required to estimate the head pose.

Compared with existing eye-only [15] and multi-region [17] methods, recent results by [48, 67] demonstrate that the methods which take a single full-face image as input to directly regress gaze direction can also achieve a high accuracy. They leverage additional information from other facial regions beyond eyes which can encode head pose or illumination-specific information across larger image areas. However, in general, duo to the fact that faces are portrayed in a myriad of poses, expressions, occlusions and more, the mechanism of the full-face approach for gaze estimation thus remains unclear.

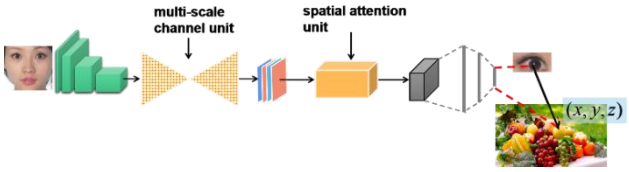

Figure 1. Overview of the proposed gaze estimation pipeline. Our method only takes the face image as input and performs 3D gaze estimation using a multi-scale channel unit and a spatial attention unit on the feature maps.

Motivated by recent reports [68-69] that attention mechanism focuses on selecting and increasing important features and suppressing unnecessary ones, the goal of our work is to learn a powerful and generic convolutional neural network (CNN) with mixed attention which can better obtain information from the full-face images for gaze estimation. Considering that the other regions of the face beyond the eyes contain valuable information for gaze estimation, we propose a multi-scale channel and spatial attention network, MCSA-Net, which focuses on selecting and increasing important features from full-face images and suppressing some unnecessary facial regions which may not contribute to estimate gaze. The overview of the MCSA-Net is shown in Fig. 1. It contains two crucial units that one is multi-scale channel unit and the other is spatial unit. The multi-scale channel unit is designed to effectively exploit multi-scale information and the inter-channel relationship of features by using multiple

ICCPR 2020, October 30-November 1, 2020, Xiamen, China

ACM ISBN 978-1-4503-8783-5/20/10…$15.00
https://doi.org/10.1145/3436369.3437438

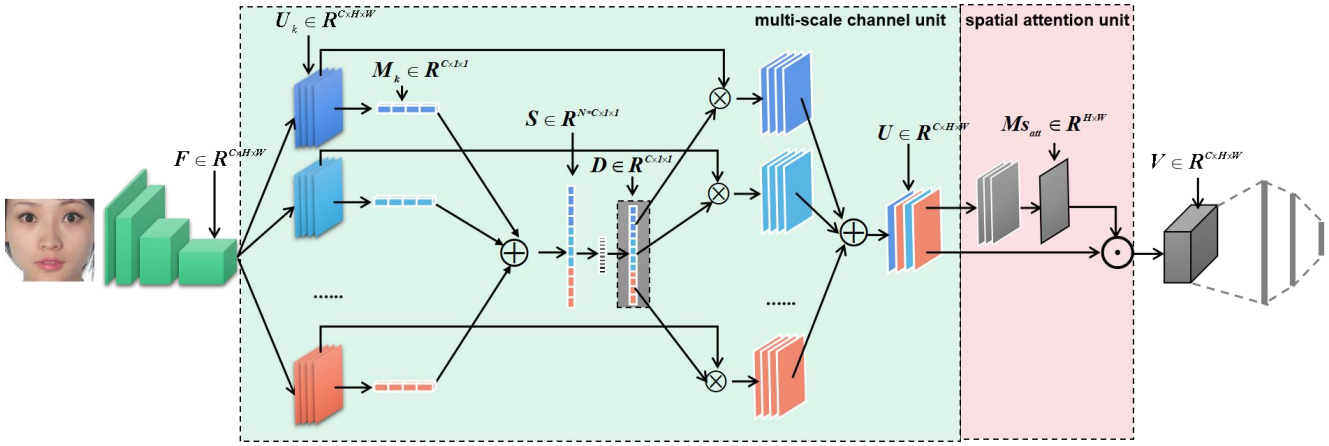

Figure 2. Overview of the proposed MCSA-Net. It consists of two major unit, namely, the multi-scale channel unit and the spatial attention unit.

convolutional kernels with additional channel attention layers. It can not only adaptively adjust the convolutional kernels sizes to select different scale information but also focus on what is meaningful features given an input image by a nonlinear approach. The spatial unit is to make the net enhance informative regions of full face, suppress some facial regions that do not contribute to estimate gaze and activate other subtle facial regions that can improve the power of gaze estimation. Naturally, we also illuminate the mechanism by providing a detailed analysis of our MCSA-Net for appearance-based gaze estimation.

Consequently, the specific contributions of this work are two-fold. First, in comparison to conventional methods in gaze estimation, we propose to learn a novel and powerful convolutional neural network (CNN) with mixed attention, MCSA-Net, to handle the gaze estimation task. It takes the full-face image as input to directly regress a map function from the input image to a 3D gaze vector without extra eye detection module and head pose estimation module. Second, we propose to use multi-scale channel unit and spatial unit to adaptively select and increase important features and suppress some unnecessary facial regions which may not contribute to estimate gaze.

2. RELATED WORK

There has been a plethora of researches proposed for the task of gaze estimation [9-14], which can be roughly divided into two major categories: model-based and appearance-based [22].

2.1 Model-Based Method

Model-based methods use a geometric eye model and can be further subdivided into corneal reflection and shape-based methods, depending on whether they rely on external light sources to detect eye features. Corneal-reflection-based methods focused on stationary settings [23-26] previously and were later extended to handle arbitrary head poses using multiple light sources or cameras [27-28]. On the other hand, shape-based methods [29-33] infer gaze direction from observed eye shapes, such as pupil centers and iris edges. However, these approaches tend to suffer with low image quality and variable lighting conditions and they are more suitable for being used in the controlled environments, e.g., in the laboratory, rather than in outdoor scenes or with large user-camera distances.

2.2 Appearance-Based Method

Appearance-based gaze estimation methods [34-39] that map images directly to gaze have recently surpassed classical model-based approaches for in-the-wild settings. They typically need a single camera to capture the user eye images [40]. These approaches cast the gaze estimation problem to learning a mapping function from eye images to gaze directions. Such a mapping function can be learned using various regression techniques, including neural networks [41-42], local interpolation [43-44], or Gaussian process regression [45-46]. The CNNs-based methods have already shown their ability to handle complex regression tasks, and thus they have outperformed traditional appearance-based methods. The availability of large-scale datasets such as MPIIGaze [15] and GazeCapture [17], and progress in CNNs have rapidly moved the field forward. Proposed advancements include the use of more complex CNNs [47]; more meaningful use of face [48, 17] and multi-modal input [17, 49-50]; explicit handling of differences in the two eyes [51]; greater robustness to head pose [52-53]; improvements in data normalization [54]; learning more informed intermediate representations [55]; using ensembles of networks [56]; and using synthetic data [57-60].

2.3 Attention Mechanism

Attention mechanisms [19-21] has become popular in many computer vision task [72-75]. It intends to focus on the most informative feature expressions and at the same time suppresses the less useful ones [68-69]. Some previous works implement the attention mechanism in the channel or spatial dimension. SENet [70] self-recalibrates the feature map according to the importance of channels by a lightweight gating mechanism. BAM [71] and CBAM [68] add an extra spatial attention unit beyond channel to encode spatial information. In this paper, we try to make further improvement by utilizing mixed attention mechanism.

3. METHOD

The gaze estimation task is formulated as a regression from the input image I to a 3D gaze vector. So, the goal of the task is to learn the regression function f. While some previous works have considered the use of the other facial regions beyond eyes for this task, they still cannot extract more valuable information from the full face. In this section, we aim to dissect the design of the CNN which can better obtain information from the full face for gaze estimation.

3.1 MCSA-Net Network Design

We are motivated by recent studies [48, 67] that other regions of the face beyond the eyes contain valuable information for gaze estimation. Our MCSA-Net network focuses on enhancing important features from full-face images and suppressing unnecessary ones due to the reason that some facial regions may not contribute to estimate gaze. The design of the MCSA-Net is illustrated in Fig. 2.

The input facial images, which are extracted by a face detector (e.g. [76]), are scaled to a standard size and then normalized by subtracting the training set mean image and dividing by its standard deviation.

Given an intermediate feature map $F \in \mathbb{R}^{C\times W\times H}$ as input from the four convolutional layers with intermittent max pooling layers(stride=2), where C is the number of feature channels and H and W are height and width of the output, we firstly utilize the 'multi-scale channel unit' to enable the network to adaptively choose appropriate local receptive field sizes and feature channels to focus on what is meaningful in the facial image for the gaze estimation task. The 'multi-scale channel unit' finally generates a feature attention map $U \in \mathbb{R}^{N\times C\times H\times W}$, which are fed into the next unit, 'spatial attention unit', to learn spatial weights for exploiting the spatial information from facial images efficiently. At the end, these are followed by a fully connected (dense) layer, which is then fully connected to an output with 3 values for the gaze direction. The main process can be summarized as:

$$F = convs(I) \tag{1}$$

$$V = \hat{S}\left(\tilde{S}(F)\right) \tag{2}$$

$$g = \delta(U) \tag{3}$$

Where $\tilde{S}$ denotes the map function from the feature map F to the feature attention map *U*, corresponding to the 'multi-scale channel unit', and $\hat{S}$ denotes the map function from the multi-scale channel feature map *U* to the spatial attention map *V*, corresponding to the 'spatial attention unit', and δ denotes the map function from attention map *U* to the output. The following describes the details of each unit.

3.2 Multi-Scale Channel Unit

As is known to all, when processing information at a certain step, the different local field sizes of neurons in the same area enables the neurons to obtain multi-scale spatial information [65-66]. Such as some InceptionNets [77-78] utilize several convolutional kernels, like 1x1, 3x3, 5x5, 7x7, to aggregate spatial information by a linear approach. However, just relying on the linear approach, they may not inject powerful adaptation ability to neurons. To more efficiently exploit the multi-scale information and the inter-channel relationship of feature maps, we propose to use multiple convolutional kernels to produce different feature maps and apply globe averaging pooling operations to generate multi-scale channel attention maps. For properly utilize multi-scale channel information, we integrate these multi-scale channel attention maps and fed them into next full connected layers to generate weight of each feature map which can straightly adaptively select channel feature and adjust the local field sizes by this nonlinear approach. We will explain the operation in detail below.

Given an intermediate feature map $F \in \mathbb{R}^{C\times W\times H}$ as input, we first conduct several transformations $\tilde{F}_k : F \to U_k^{C\times H\times W}$ with different scale kernel sizes $n_k, k = 1,2,...,N$ respectively where $\tilde{F}_k$ consists of grouped/depthwise convolutions, Batch Normalization and ReLU function in sequence. Due to the reason that each channel of a feature map can be considered as a feature detector [18] to focus on what is meaningful for a given input feature map on the specific kernel, we sequentially squeeze the spatial dimension of the feature map $U_k \in \mathbb{R}^{C\times H\times W}$ for each branch by using globe average-pooling operation, which can improve representation power, to generate a 1D specific kernel channel attention map $M_k \in \mathbb{R}^{C\times 1\times 1}$ as illustrated in Fig. 2. Specifically, the C-th element of M_k, i.e. $M_k^C \in \mathbb{R}^{1\times 1\times 1}$, is calculated by shrinking $U_k^C \in \mathbb{R}^{1\times H\times W}$ through spatial dimensions $H \times W$:

$$M_k^C = F_{gp}\left(U_k^C\right) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_k^C(i,j) \tag{4}$$

Further, to adaptively select channel features and adjust the kernel sizes $n_k, k = 1,2,...,N$ for properly exploiting multi-scales channel information, we integrate all channel attention maps from multiple branches to generate whole channel-wise statistics as $S \in \mathbb{R}^{N*C\times 1\times 1}$, which then forwarded to two simple fully connected (fc) layers to enable the guidance for the precise and adaptive feature map selections. Note that the finally fully connected layer is $D \in \mathbb{R}^{C\times 1\times 1}$. We then apply a SoftMax operator on the channel-wise digits $D \in \mathbb{R}^{C\times 1\times 1}$, it maps the output of multiple neurons to the [0, 1]. The final feature map $U \in \mathbb{R}^{C\times H\times W}$ is obtained through the attention weights on various channel maps:

$$U = \sum_{k=1}^{N} U_k \cdot D = \sum_{k=1}^{N}\sum_{c=1}^{C} U_k^c \cdot d_c \tag{5}$$

where U_k^c is c-th channel of *U*, is the c-th element of *D*.

3.3 Spatial Attention Unit

We propose a spatial weights mechanism that makes net focus on informative regions of full face and suppress some facial regions that do not contribute to estimate gaze and activate other subtle facial regions that can improve the power of gaze estimation. To generate the spatial attention map, we firstly apply a 1x1 convolutional layer to multi-scale channel map $U \in \mathbb{R}^{C\times H\times W}$ and then use an average-pool operation to produce a 2D attention map $Ms_{att} \in \mathbb{R}^{H\times W}$:

$$Ms_{att} = \sigma\left(convs(U)\right) \tag{6}$$

where σ denotes the sigmoid function.

The final spatial feature map $V \in \mathbb{R}^{C\times H\times W}$ is obtained from element-wise multiplication of the spatial attention map $Ms_{att} \in \mathbb{R}^{H\times W}$ with the original multi-scale channel map *U*:

$$V^C = \left[U^C \bullet Ms_{att}\right] \tag{7}$$

where U^C denotes the c-th channel of *U*, and V^C denotes the weighted activation map of the same channel, and $V = [V_1, V_2, ..., V_C]$. These maps are stacked to form the weighted

activation tensor V, and are fed into the next layer. Since the same weights are applied to all feature channels, the spatial attention unit encodes where to emphasize or suppress for facial regions.

4. EXPERIMENTS

4.1 Public Datasets

MPIIGaze [15] is a dataset which consists of 213659 images of 15 participants. The dataset contains a large variety of different illuminations, facial changes, eye appearances and head poses. It is already a common established benchmark dataset for in-the-wild gaze estimation. All the images and labels in the dataset have already been normalized as [54].

GazeCapture [17] is a mobile-based eye tracking dataset containing almost 1500 subjects from a wide variety of backgrounds, recorded under variable lighting conditions and unconstrained head motion. The dataset includes 1,500,000 frames with both face and two eyes. We also utilize a pre-processing method as same as what MPIIGaze has done for full-face frames to yield input images.

Gaze360 [16] has wide range of gaze and head poses, variety of indoor and outdoor capture environments and diversity of subjects. According to the 3D gaze ground truth, we used the face bounding box provided by the dataset to crop the face patch.

4.2 Baseline Methods

To validate the effectiveness of our method, we compared our approach with the following baseline methods on the public datasets. Results of these baseline methods are obtained from our implementation or the published paper.

-Full Face [48]: It also takes the full-face image as input and encodes the face image using a CNN with spatial weights applied on the feature maps. The performance has also been tested and reported on the same MPIIGaze dataset.

-Two Eyes [51]: It uses an asymmetric regression network to predict 3D gaze directions for both eyes with an asymmetric strategy, and at the same time uses an evaluation network adaptively adjusts the strategy by evaluating the two eyes in terms of their performance during optimization.

-Single Eye [6]: It is a classical appearance-based gaze estimation method which the input of this method is a single eye. According to the Caffe codes provided by [6], we can obtain all the results in our experiments.

-Both face and eyes [17]: The iTracker takes the full-face image, left and right eye, and a face grid as input. The performance of it has already been reported in [17] on the MPIIGaze dataset.

4.3 Within Dataset Evaluation

Since the MPIIGaze dataset contains abundant images of the eyes and full faces, we conduct experiments on this dataset to evaluate the performance. Note that the MPIIGaze dataset has been normalized as [54] and we perform a leave-one-person-out cross-validation strategy on all 15 participants like [48] to ensure that the experiments are done in a fully person-independent manner. Meanwhile, the gaze origin is defined at the face center for both the iTracker and Full-Face methods. Therefore, in order to make a fair comparison, we also convert our estimated two eye gaze vectors to have the same origin geometrically, and then take their average as the final output.

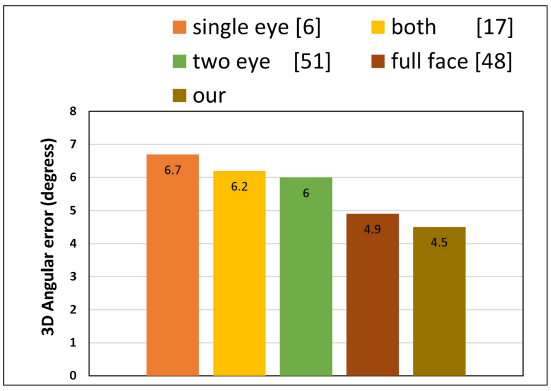

Figure 3. Experimental results of the within-dataset evaluation and comparison.

The comparison of different gaze estimation methods on MPIIGaze dataset is shown in Fig. 3. The 3D angular error was directly calculated from the estimated value and ground-truth 3D gaze vectors. In Fig. 3, we can see that all of these methods that take full-face as input significantly outperformed the single eye baseline and also have a competitive result than these elaborate two-eye and iTracker methods. As for the MCSA-Net, the average error is 4.5°, which achieves a significant performance improvement of 32.8% over the single eye method, and is more than 8.1% improved compared to [48]. Both the [48] and our method take the full-face image as input, but the [48] uses a CNN with spatial weights applied on the intermediate feature maps. Compared to the [48], our improvement is benefited from our new multi-scale channel attention mechanism.

4.4 Cross-Dataset Evaluation

We then perform a cross-dataset evaluation to rigorously evaluate our module. We choose the Gaze360 dataset as the training dataset since it has wide range of gaze and head poses, variety of indoor and outdoor capture environments and diversity of subjects and choose another MPIIGaze and Gazecapture datasets as test data. Moreover, we obtained face images from MPIIGaze corresponding to the same evaluation set and flipped the face images when they came from the right eye as [48]. Meanwhile, we set the middle point of face (the center of all six landmarks) as the origin of gaze direction. Note that the GazeCapture dataset has been normalized and face images were cropped and resized to 224 x 224 pixels.

Table 1. Average gaze errors of the cross-dataset evaluation

	MPIIGaze	Gazecapture
Singe Eye [6]	13.8	16.6
Two Eyes [51]	9.1	12.2
Full Face [48]	7.1	9.5
Our	6.3	8.1

As shown in Table 1, our method, MCSA-Net, outperforms the [6], [51] and [48] on both the MPIIGaze and the Gazecapture datasets. Compared to the [48], the performance improvement of our method is 11.3% on the MPIIGaze dataset, and 14.7% on the Gazecapture dataset. Note that the [48] only takes a full-face

image as input and uses a CNN with spatial weights module, but our MCSA-Net contains an extra multi-scale channel unit beyond the spatial attention unit. This demonstrates the effectiveness of the proposed multi-scale channel attention mechanism.

4.5 The Importance of Different Facial Regions

Our MCSA-Net network focuses on enhancing important features from full-face images and suppressing unnecessary ones. In our method, the multi-scale channel unit and the spatial attention unit play a key role in the network. From the [48], we know that the spatial attention mechanism can contribute to the gaze estimation task. Therefore, it is necessary to know if the multi-scale channel attention unit benefits the task. According to comparisons shown in Table 1, we have known that the performance of our method has a significant improvement compared with the full-face method, [48], which only contains a spatial attention unit. In addition, Fig. 4 illustrates that the error of different inputs in the MCSA-Net. Furthermore, Fig. 4 illustrates that if the MCSA-Net takes the full face as input, it can achieve a higher accuracy than these inputs of single eye or two eyes. When we take the full-face iamge as input of the MCSA-Net, the average angular error is 4.5°. It achieves a significant performance improvement of 13.5% over the two eyes input, and is more than 28.6% improved compared to single eye input.

Figure 4. MCSA-Net with different input region.

The multi-scale channel unit effectively exploit multi-scale information and the inter-channel relationship of features by using multiple convolutional kernels with additional channel attention layers. It can not only adaptively adjust the convolutional kernels sizes to select different scale information but also focus on what is meaningful features given an input face image.

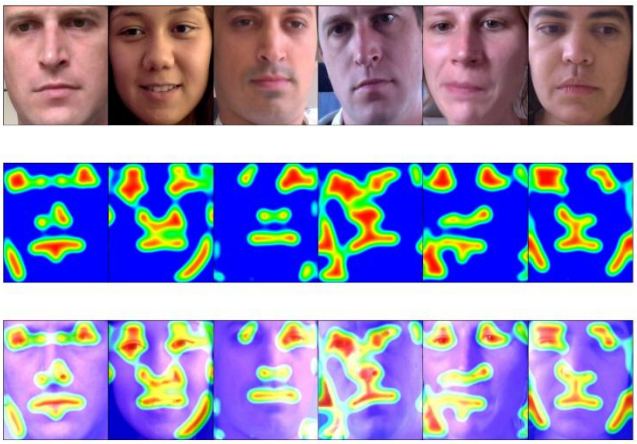

Figure 5. Important regions of estimating gaze.

From the Fig. 5, we can see that our MCSA-Net can extract more valuable information from the full-face image and effectively use the other facial regions beyond the two eyes for the 3D gaze estimation task. The region of eyes is of course an important area, but the location of the nose and mouth also make a contribution to the gaze estimation.

In these experiments, it can be clearly seen that the full-face input and the multi-scale channel attention mechanism and the spatial attention mechanism are particularly beneficial to improv-ing estimation performance and it also indicates that non-eye facial regions also have in general higher importance for 3D gaze estimation.

5. CONCLUSION

In this work we studied full-face appearance-based gaze estimation and proposed a novel and powerful deep convolutional network with multi-scale channel and spatial attention, MCSA-Net, which only takes the full-face image as input without additional modules to detect the eyes and estimate the head pose, to handle the gaze estimation task. By training the MCSA-Net, our method achieves good performances on these public datasets. We compared our approach with some baseline methods on the public datasets and it validated that our method achieved a significant improvement over the state of the art.

There are still future works to do along this line. First, we consider extending our MCSA-Net to more intermediate features maps exploit more multi-scale and inter-channel information for full-face based gaze estimation. Second, we can use more advanced network structures to enhance its performance. Third, we will extend our MCSA-Net with multi-scale channel and spatial attention to contribute to more generic computer vision tasks.

6. REFERENCES

[1] R. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. Mind, 2003.

[2] P. Majaranta and A. Bulling. Eye tracking and eye-based human-computer interaction. In Advances in Physiological Computing. Springer, 2014.

[3] C. H. Morimoto and M. R. Mimica. Eye gaze tracking techniques for interactive applications. CVIU, 2005.

[4] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In International Conference on Human-Robot Interaction (HRI), pages 61-68, 2009.

[5] K. Rayner. Eye movements in reading and information processing: 20 years of research. Psychological bulletin, 1998.

[6] P. S. Holzman, L. R. Proctor, D. L. Levy, N. J. Yasillo, H. Y. Meltzer,and S. W. Hurt. Eye-tracking dysfunctions in schizophrenic patients and their relatives. Archives of general psychiatry, 1974.

[7] S. Nilsson. Interaction without gesture or speech-a gaze controlled ar system. In Artificial Reality and Telexistence, 17th International Conference on, pages 280-281. IEEE, 2007.

[8] D. Pohl, X. Zhang, and A. Bulling. Combining eye tracking with optimizations for lens astigmatism in modern wideangle

hmds. In 2016 IEEE Virtual Reality, VR 2016, Greenville, SC, USA, March 19-23, 2016, pages 269-270, 2016.

[9] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell., 32(3):478-500, Mar. 2010.

[10] O. Ferhat and F. Vilarino. Low cost eye tracking: The current panorama. Computational intelligence and neuroscience, 2016.

[11] K. A. Funes-Mora and J.-M. Odobez. Gaze estimation in the 3d space using rgb-d sensors. International Journal of Computer Vision, pages 1-23, 2015.

[12] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. Learning gaze biases with head motion for head pose-free gaze estimation. Image and Vision Computing, 32(3):169-179, 2014.

[13] K. A. F. Mora and J.-M. Odobez. Person independent 3d gaze estimation from remote rgb-d cameras. In International Conference on Image Processing (ICIP), pages 2787-2791. IEEE, 2013.

[14] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In IEEE International Conference on Computer Vision (ICCV), 2015.

[15] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4511-4520, 2015.

[16] Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6912-6921 (2019)

[17] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2176-2184, 2016.

[18] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part I. LNCS, vol. 8689, pp. 818-833. Springer, Cham (2014).

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI, 20(11):1254-1259, 1998.

[20] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, and F.Nuflo, Modelling Visual Attention via Selective Tuning, Artificial Intelligence, vol. 78, no. 1-2, pp. 507-545, Oct. 1995.

[21] E. Niebur and C. Koch, Computational Architectures for Attention, R. Parasuraman, ed., The Attentive Brain, pp. 163-186. Cambridge, Mass.: MIT Press, 1998.

[22] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell., 32(3):478-500, 2010.

[23] S.-W. Shih and J. Liu. A novel approach to 3-d gaze tracking using stereo cameras. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 34(1):234-245, 2004.

[24] C. H. Morimoto, A. Amir, and M. Flickner. Detecting eye position and gaze from a single camera and 2 light sources. In Proc. ICPR, pages 314-317, 2002.

[25] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. Computer Vision and Image Understanding, 98(1):25-51, 2005.

[26] C. Hennessey, B. Noureddin, and P. Lawrence. A single camera eye-gaze tracking system with free head motion. In Proc. ETRA, pages 87-94, 2006.

[27] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In Proc. CVPR, pages 918-923, 2005.

[28] Z. Zhu, Q. Ji, and K. P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In Proc. ICPR, pages 1132-1135, 2006.

[29] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive driver gaze tracking with active appearance models. In Proc. 11th World Congress on Intelligent Transportation Systems, 2004.

[30] J. Chen and Q. Ji. 3d gaze estimation with a single camera without ir illumination. In ICPR, 2008.

[31] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. TIP, 2012.

[32] D. W. Hansen and A. E. Pece. Eye tracking in the wild. CVIU, 2005.

[33] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. IEEE Transactions on Image Processing, 21(2):802-815, 2012.

[34] Kar-Han Tan, David J. Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In WACV, 2002.

[35] W. Sewell and O. Komogortsev. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In SIGCHI, 2010.

[36] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. PAMI, 2014.

[37] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. Learning gaze biases with head motion for head pose-free gaze estimation. Image and Vision Computing, 2014.

[38] D. Torricelli, S. Conforto, M. Schmid, and T. D'Alessio. A neuralbased remote eye gaze tracker under natural head motion. Computer methods and programs in biomedicine, 2008.

[39] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical report, 1994.

[40] Tan, K., Kriegman, D., Ahuja, N.: Appearance-based eye gaze estimation. In:WACV. (2002) 191-195

[41] L.-Q. Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. In Proc. BMVC, 1998.

[42] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical Report CMU-CS-94-102, School of Computer Science, Carnegie Mellon University, 1994.

[43] K.-H. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In Proc. 6th IEEE Workshop on Applications of Computer Vision, pages 191-195, 2002.

[44] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In Proc. ICCV, pages 153-160, 2011

[45] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S3GP. In Proc. CVPR, volume 1, pages230-237, 2006.

[46] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. IEEE Trans. Pattern Anal. Mach. Intell., 35(2):329-341, 2013.

[47] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance based gaze estimation. TPAMI, 2019.

[48] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance based gaze estimation. In CVPR - Workshops, 2017.

[49] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In ECCV - Workshops, 2018

[50] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In ECCV, 2018.

[51] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance based gaze estimation via evaluation-guided asymmetric regression. In ECCV, 2018.

[52] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In ICCV, 2017.

[53] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light weight head pose invariant gaze tracking. In CVPR - Workshops, 2018.

[54] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In ETRA, 2018.

[55] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep Pictorial Gaze Estimation. In ECCV, 2018.

[56] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In CVPR, July 2017.

[57] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In CVPR, 2018.

[58] Kangwook Lee, Hoon Kim, and Changho Suh. Simulated+unsupervised learning with adaptive data generation and bidirectional mappings. In ICLR, 2018.

[59] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light weight head pose invariant gaze tracking. In CVPR - Workshops, 2018.

[60] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In ACM ETRA, 2018.

[61] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In Proc. ETRA, pages 245-250, 2008.

[62] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. Training person-specific gaze estimators from user interactions with multiple devices. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, page 624. ACM, 2018

[63] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 721-738, 2018.

[64] Yu Yu, Jean-Marc Odobez. Unsupervised Representation Learning for Gaze Estimation. In CVPR, 2020.

[65] M. Aubry and B. C. Russell, Understanding deep features withcomputer-generated imagery, in Proc. Int. Conf. Comput.Vision, 2015.

[66] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, How transferable are features in deep neural networks? in Neural Inform. Process. Syst., 2014, pp. 3320-3328.

[67] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang and Otmar Hilliges. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In ECCV, 2020.

[68] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. arXiv preprint arXiv:1807.06521, 2018.

[69] Xiang Li, Wenhai Wang, Xiaolin Hu, Jian Yang. Selective Kernel Networks. In CVPR, 2019.

[70] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, Squeeze-and-Excitation Networks, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[71] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. arXiv preprint arXiv: 1807.06514, 2018.

[72] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai. Person search via a mask-guided two-stream cnn model. arXiv preprint arXiv:1807.08107, 2018.

[73] K. Xu, D. Li, N. Cassimatis, and X. Wang. Lcanet: Endto-end lipreading with cascaded attention-ctc. In International Conference on Automatic Face & Gesture Recognition, 2018.

[74] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[75] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In CVPR, 2016.

[76] D. E. King, Dlib-ml: A machine learning toolkit. J. Mach. Learning Research, vol. 10, pp. 1755-1758, 2009.

[77] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, 2017.

[78] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016.