L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments

Ahmed A. Abdelrahman*
department of Neuro-Information Technology
Otto-von-Guericke-University
Magdeburg, Germany
ahmed.abdelrahman@ovgu.de
*Corresponding author

Thorsten Hempel
department of Neuro-Information Technology
Otto-von-Guericke-University
Magdeburg, Germany
thorsten.hempel@ovgu.de

Aly Khalifa
department of Neuro-Information Technology
Otto-von-Guericke-University
Magdeburg, Germany
aly.khalifa@ovgu.de

Ayoub Al-Hamadi
department of Neuro-Information Technology
Otto-von-Guericke-University
Magdeburg, Germany
Ayoub.Al-Hamadi@ovgu.de

Laslo Dinges
department of Neuro-Information Technology
Otto-von-Guericke-University
Magdeburg, Germany
laslo.dinges@ovgu.de

Abstract—Human gaze is a crucial cue used in various applications such as human-robot interaction and virtual reality. Recently, convolutional neural network (CNN) approaches have made notable progress in predicting gaze direction. The performance of existing appearance-based gaze methods remains unsatisfactory due to the uniqueness of eye appearance, lighting conditions, and the diversity of head pose and gaze directions. In this paper, we propose a novel multi-loss two-branch CNN architecture (L2CS-Net) to explicitly learn the discriminative features for each gaze angle by predicting each gaze angle using a separate fully connected layer and loss function. In addition, we introduce a new multi-loss gaze function that consists of combined classification and regression losses to further enhance the model performance. We perform gaze classification utilizing a softmax layer along with cross-entropy loss. To obtain fine-grained predictions, we calculate the expectation of the gaze-class probabilities followed by a Mean Squared Error (MSE) loss. We evaluated our model with two popular datasets collected with unconstrained settings. Our proposed model achieves state-of-the-art performance on the MPIIGaze and Gaze360 datasets, respectively. We make our code open source at https://github.com/Ahmednull/L2CS-Net.

Index Terms—Appearance-based gaze estimation, Gaze Analysis, Gaze Tracking, Convolutional Neural Network.

I. INTRODUCTION

Eye gaze is one of the essential cues used in a large variety of applications. It indicates the user’s level of engagement in human-robot interaction [1], [8], [16], and open dialogue systems [11]. Furthermore, it is used in augmented reality [15] to predict the user’s attention. Gaze estimation methods are divided into two categories: model-based and appearance-based approaches. Model-based methods generally require dedicated hardware that makes them difficult to use in an unconstrained environment. On the other hand, appearance-based methods regress the human gaze directly from the images captured by inexpensive off-the-shelf cameras, making them easy to generate in different locations with unconstrained settings.

Recently, CNN-based appearance-based methods have become increasingly prevalent for gaze estimation due to their ability to extract essential gaze features, resulting in improved gaze performance [2], [3], [5], [12]. Most of the early work [2], [3], [5], [9], [22] focused on the development of innovative CNN-based networks that primarily utilized popular backbones (e.g., VGG [7], ResNet-18 [9], ResNet-50 [20]) to extract gaze features and ultimately predict gaze direction.

The most commonly used loss function for the gaze estimation task is the mean square loss or $\ell_2$ loss. However, Petk et al. [9] proposed a novel pinball loss that estimates the gaze direction and error bounds together, which improves the accuracy of gaze estimation, especially in unconstrained settings. Several appearance-based methods [2], [3], [5], [9], [22] have attempted to estimate gaze by directly regressing two gaze angles pitch and yaw using one fully connected layer and single loss function. This way, the unique features of each angle are not emphasized or adequately utilized, preventing the model from learning discriminative features for each angle.

In this paper, we introduce a novel method for estimating 3D gaze angles from RGB images using a two-branch CNN architecture. Our approach involves independently regressing each gaze angle (yaw and pitch) by utilizing a dedicated fully connected layer and a loss function for each angle. This allows us to explicitly learn the distinct gaze features associated with each angle. To further enhance the performance of our model, we incorporate a multi-loss function that includes...
gaze binary classification and regression losses. For the gaze binary classification, we employ a softmax layer in conjunction with cross-entropy loss. On the other hand, we calculate the expectation of the gaze class probabilities and apply MSE loss for regression. By combining these losses, we create two signals that backpropagate through the network, resulting in precise fine-tuning of the network weights and stabilizing the gradients.

Based on our proposed loss function and the inclusion of a softmax layer (consisting of $\ell_2$ loss, cross-entropy loss, and softmax layer), we introduce a new network called L2CS-Net, which predicts 3D gaze vectors in unconstrained scenarios. To assess the performance of our network, we evaluate its robustness on two widely used datasets: MPIIFaceGaze and Gaze360. The experimental results demonstrate that our proposed L2CS-Net achieves state-of-the-art performance on both MPIIFaceGaze and Gaze360 datasets.

In summary, this paper makes the following contributions:

- A novel L2CS-Net for predicting gaze angles using two-branch CNN architecture to learn the discriminative features associated with each gaze angle.
- A new gaze multi loss function of classification and regression losses. Using classification loss yields robust and stable network learning while still providing fine-grained predictions through the regression loss.
- Our network achieves state-of-the-art gaze performance on MPIIFaceGaze and Gaze360 datasets.

II. RELATED WORK

According to the literature, appearance-based gaze estimation can be divided into conventional and CNN-based methods. Conventional gaze estimation methods use a regression function to create a person-specific mapping function to the human gaze, e.g., adaptive linear regression [13] and gaussian process regression [18]. These methods show reasonable accuracy in constrained setup (e.g., subject-specific, fixed head pose and illumination), however they significantly decrease when tested on unconstrained settings.

Recently, researchers have gained more interest in CNN-based gaze estimation methods, as they can model a highly non-linear mapping function between images and gaze. For example, Zhang et al. [22] proposed a VGG CNN-based architecture to predict gaze using a single eye image. Additionally, they designed a CNN spatial weights in [21] to give more weight to those regions of the face related to the gaze in appearance. Krafft et al. [10] proposed a multichannel network that takes eye images, full-face images, and face grid information as inputs.

Combining statistical models with deep learning is a viable solution for gaze estimation. Chen et al. [19] introduced a mixed effect model that integrates statistics information within the CNN architecture based on eye images. Chen et al. [3] adopted dilated convolutions to make use of high-level features extracted from images without decreasing spatial resolution. In addition, they expanded their work by proposing GEDDNet [4], which uses gaze decomposition with dilated convolutions. Fischer et al. [7] add the head pose vector along with the features extracted using a VGG CNN with eye crops to predict gaze angles. Additionally, they used an ensemble scheme to improve gaze accuracy.

Motivated by the two-eye asymmetry property, Cheng et al. [6] proposed FAR-Net that estimates 3D gaze angles for both eyes with an asymmetric approach. They give asymmetric weights to each loss of the two eyes and finally sum these losses. The proposed model showed excellent performance on multiple public datasets. Cheng et al. [5] proposed a coarse-to-fine adaptive network (CA-Net) that first uses face image to predict primary gaze angles and adapt it with the residual estimated from eye crops. Then, they proposed a bigram model to bridge the primary gaze with the eye residual. Kellnhofer et al. [9] used a temporal model (LSTM) with a sequence of 7 frames to predict gaze angles. In addition, they adapt pinball loss to jointly regress the gaze direction and error bounds together to improve gaze accuracy.

Recently, transformers with a self-attention module bring high performance to various computer vision tasks. In [14], they use a transformer with a self-attention module to extract crucial gaze features from images with high variance. Further, they effectively filter irrelevant gaze information using convolution projection and maintain detailed image features using deconvolution layer. Similarly, AGE-Net [2] proposed two parallel networks for each image of the eye, one generating a feature vector using CNN and the other generating a weight feature vector using an attention-based network.

III. METHOD

A. L2CS-Net Architecture

We propose a simple network architecture (L2CS-Net) based upon the proposed classification and regression losses. Our proposed approach employs a CNN backbone ResNet50, followed by two fully connected layers, to predict two separate gaze angles: pitch and yaw. Unlike prior methods that regressed both angles in one fully-connected layer, our network predicts each angle individually with separate loss using two fully-connected layers that share the previous convolutional layers. To improve the network’s stability and robustness, we employ two loss functions for each angle, each of which is a linear combination of classification and regression losses. Using this approach will help to improve the learning of the network, as it has two signals that backpropagate through the network.

The network’s output dimension for each fully connected layer is adapted to predict a specific number of output angle bins based on the range of angles in the dataset being evaluated. To achieve this, we used a non-linear softmax layer for two reasons. Firstly, it allows us to obtain a categorical cross-entropy loss by formulating the gaze estimation problem as a multi-class classification task. Second, it enables us to obtain a continuous gaze value by calculating the expectation of the output probabilities.

For each output from the fully connected layer, we first use a softmax layer to convert the network output logits into a probability distribution. Then, we apply a cross-entropy
loss to calculate the bin classification loss between output probabilities and target bin labels. Next, we calculate the expectation of the probability distribution to get fine-grained gaze predictions. Finally, we calculate the mean square error for this prediction and add it to the classification loss. The detailed architecture of L2CS-Net is shown in Fig. 1.

B. Proposed loss function

Most CNN-based gaze estimation models predict 3D gaze as the angles of gaze direction (yaw, pitch) in spherical coordinates. Furthermore, they adopt the mean squared error ($\ell_2$ loss) to penalize their networks. We propose to use two identical losses for each gaze angle. Each loss contains a combined cross-entropy loss and MSE. Instead of directly predicting continuous gaze angels, we used a softmax layer with cross-entropy to predict binned gaze classification. Then, we estimate the expectation of the gaze binned output followed by MSE to fine-grain the predictions.

For the classification branch, we apply softmax layer to predict gaze bin probabilities as follows:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{b} e^{x_j}}$$  \hspace{1cm} (1)

Then, we adapt the cross entropy loss to compute the classification loss by:

$$H(y, p) = -\sum_i y_i \log p_i$$  \hspace{1cm} (2)

For the regression branch, we estimate the expectation of the gaze bins using:

$$E(X) = \sum_{i=1}^{b} p_i * i$$  \hspace{1cm} (3)

Finally, we add MSE to obtain the regression loss as follows:

$$MSE(y, p) = \frac{1}{N} \sum_{0}^{N} (y - p)^2$$  \hspace{1cm} (4)

Our proposed loss for each gaze angle is a linear combination of the mean-squared error and cross-entropy losses, which is defined as:

$$CLS(y, p) = H(y, p) + \beta \cdot MSE(y, p)$$  \hspace{1cm} (5)

Where $CLS$ is the combined loss, $p$ is the predicted values, $y$ are the ground truth values and $\beta$ is the regression coefficient. We change the weight of the mean-squared loss during the experiments in Section IV to obtain the best gaze performance.

To the best of our knowledge, all related works which estimated gaze using CNN-based methods do not consider the combined classification and regression loss in their techniques.

C. Datasets

With the development of appearance-based gaze estimation methods, large-scale datasets have been proposed to improve gaze performance. These datasets were collected with different procedures, ranging from laboratory settings to unconfined indoor and outdoor environments. In order to get a valuable evaluation of our network, we train and evaluate our model using two popular datasets collected with unconstrained settings: Gaze360 and MPIIGaze.

**Gaze360** [9] provides the widest range of 3D gaze annotations with a range of 360 degrees. It contains 238 subjects of different ages, genders, and ethnicities. Its images are captured using a Ladybug multi-camera system in different indoor and outdoor environmental settings like lighting conditions and backgrounds.

**MPIIGaze** [21] provides 213,659 images from 15 subjects captured during their daily routine over several months. Consequently, it contains images with diverse backgrounds, time, and lighting that make it suitable for unconstrained gaze estimation. It was collected using software that asks the participants to look at randomly moving dots on their laptops.
TABLE I
COMPARISON OF MEAN ANGULAR ERROR BETWEEN OUR PROPOSED MODEL AND SOTA METHODS ON MPIIFaceGaze and Gaze360 DATASETS.

| Methods                  | MPIIFaceGaze° | Gaze360° |
|--------------------------|---------------|----------|
| tTracker (AlexNet) [10]  | 5.6°          | -        |
| MeNets [19]              | 4.9°          | -        |
| FullFace (Spatial weights CNN) [21] | 4.8°  | 14.99°  |
| Dilated-Net [3]          | 4.4°          | 13.73°   |
| GEEDNet [4]              | 4.5°          | -        |
| RT-Gene [7]              | 4.3°          | 12.26°   |
| Bayesian Approach [17]   | 4.3°          | -        |
| FAR-Net [6]              | 4.3°          | -        |
| CA-Net [5]               | 4.27°         | 11.20°   |
| AGE-Net [2]              | 4.09°         | -        |
| Gaze360 [9]              | 4.06°         | 12.20°   |
| Selfatt [14]             | 4.04°         | 10.70°   |
| L2CS-Net (β = 1)         | 3.96°         | 10.54°   |
| L2CS-Net (β = 2)         | 3.92°         | 10.41°   |

TABLE II
COMPARISON OF MEAN ANGULAR ERROR BETWEEN OUR PROPOSED MODEL WITH ResNet50 and ResNet18 BACKBONES.

| Methods         | MPIIFaceGaze° | Gaze360° |
|-----------------|---------------|----------|
| ResNet-18       | 4.00°         | 10.7°    |
| ResNet-50       | 3.92°         | 10.41°   |

TABLE III
COMPARISON OF MEAN ANGULAR ERROR BETWEEN OUR PROPOSED MODEL WITH SINGLE FULLY CONNECTED LAYER AND TWO FULLY CONNECTED LAYERS.

| Methods | MPIIFaceGaze° | Gaze360° |
|---------|---------------|----------|
| SFC     | 4.02°         | 10.53°   |
| TFC     | 3.92°         | 10.41°   |

IV. EXPERIMENTS

A. Data preprocessing

We follow the same procedures in [21] to normalize images in the two datasets. In summary, this process applies rotation and translation to the virtual camera to remove the head’s roll angle and keep the same distance between the virtual camera and a reference point (the center of the face). Furthermore, we split the continuous gaze target in each dataset (pitch and yaw angles) into bins with binary labels for classification based on the range of the gaze annotations. As a result, both datasets have two different target annotations: continuous and binned labels make them suitable for our combined regression and classification losses. Moreover, we change the regression coefficient in the combined loss function during the experiments to obtain the best gaze performance. For the MPIIFaceGaze dataset, we utilize leave-one-subject-out cross-validation as previously used methods [2], [5], [21]. Kellnhofer et al. [9] divided the Gaze360 dataset into train-val-test sets and presented three evaluation scopes based on the range of gaze angles: 360°, front 180°, and front facing (within 20°). We follow the same evaluation criteria in [9], but only with the front 180°and front-facing for a fair comparison with all appearance-based gaze methods that are trained and evaluated on datasets within 180°range.

B. Implementation details

We use an ImageNet-pretrained ResNet-50 as the backbone network. Our proposed network (L2CS-Net) was trained in the PyTorch framework using Adam optimizer with a learning rate of 0.00001. We train our proposed network for 50 epochs using a batch size of 16. We evaluate our proposed network on MPIIFaceGaze and Gaze360 datasets. We change the regression coefficient during the experiments and compare the output performance with the state-of-the-art gaze estimation methods.

C. Performance Measurement

We utilize gaze angular error (°) as the evaluation metric following most gaze estimation methods. Assuming that the ground truth gaze direction is \( \hat{g} \in \mathbb{R}^3 \) and the predicted gaze vector is \( g \in \mathbb{R}^3 \), the gaze angular error (°) can be computed as:

\[
\mathcal{L}_{\text{angular}} = \frac{\| \hat{g} \cdot g \|}{\| \hat{g} \| \| g \|} \quad (6)
\]

D. Comparison with the state of the art

We conduct experiments on MPIIFaceGaze and Gaze360 to compare the performance of our L2CS-Net with the SOTA gaze estimation methods. Table I shows the angular error results of our L2CS-Net and SOTA methods include tTracker [10], MeNets [19], Full-Face [21], Dilated-Net [3], RT-Gene [7], GEEDNet [4], Bayesian Approach [17], FareNet [6], CA-Net [5], AGE-Net [2] Eth-Gaze [20], Gaze360 [9], and SelfAtt [14].

For MPIIFaceGaze, our proposed L2CS-Net (β = 1) achieved state-of-the-art gaze performance with 3.92°mean angular error of 3.92 °with an improvement of 0.12°compared with Selfatt [2]. Note that Selfatt [2] uses a self-attention mechanism to extract crucial gaze features with 74.8 million parameters. Although our L2CS-Net uses less resources (Resnet-50 as a backbone with 22 million parameters) and output higher performance. For Gaze360, our method provides improvement over Selfatt [2] with approximately 0.3°. Overall, our findings suggest that the proposed L2CS-Net with β = 1 is a robust and effective model for gaze estimation on datasets within the 180°range. Its superior performance compared to other state-of-the-art methods demonstrates its potential to be utilized in various gaze-related applications.

E. Ablation study

In order to investigate the effectiveness of L2CS-Net, we conduct ablation study on MPIIFaceGaze and Gaze360 datasets. First, we study the effect of swapping out ResNet-50 in our proposed method for a smaller ResNet-18 backbone. Our L2CS-Net with ResNet-50 has already shown its superiority through table I. Nevertheless, we want to evaluate our method using a more compact backbone to better understand the effect of model size on the performance of our method. As shown in Table II, our proposed approach retained its SOTA performance, even with the smaller ResNet-18 backbone,
with a slight decrease in performance compared to ResNet-50. This demonstrates the robustness of our L2CS-Net. This is important when considering real-world applications, as it signifies the capability of our model to operate effectively in resource-constrained environments without compromising accuracy.

Secondly, to prove the advantage of separating the gaze angles using two fully connected layers, we experimented a network (SFC) with one fully connected layer for predicting two gaze angles’ pitch and yaw. We use the same backbone and loss function as our L2CS-Net which predict gaze using two separate fully connected layers (TFC). As shown in Table III, our proposed network improves gaze performance by approximately 0.1% compared to the SFC network. This proves that predicting gaze angles separately makes the network obtain more discriminative gaze features which enhance the overall gaze performance. Finally, we present some qualitative examples of L2CS-Net in Fig 2. From the figure, our L2CS-Net is able to predict accurate gaze direction from images with different settings.

V. CONCLUSION

In this paper, we present a novel multi loss two-branch CNN-based model (L2CS-Net) for predicting 3D gaze angles in unconstrained environments. We propose to predict the gaze angles separately using two fully connected layers and two loss functions. To further improve network learning, we adapt a combined loss function of regression and classification losses. We use a reliable softmax layer along with cross-entropy loss to get a robust gaze classification. On the other hand, we achieve the gaze regression by calculating the expectation of the gaze class probabilities followed by a gaze regression loss. To show the robustness of our model, we validate our network using two of the most unconstrained gaze datasets: MPIIGaze and Gaze360, and we followed the same evaluation criteria used in each dataset. Our model achieved state-of-the-art gaze accuracy with the lowest angular error in both datasets.

REFERENCES

[1] Abdelrahman, A.A., Strazdas, D., Khalifa, A., Hintz, J., Hempel, T., Al-Hamadi, A.: Multimodal engagement prediction in multiperson human–robot interaction. IEEE Access 10, 61980–61991 (2022)

[2] Biswas, P., et al.: Appearance-based gaze estimation using attention and difference mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3143–3152 (2021)

[3] Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: Asian Conference on Computer Vision. pp. 309–324. Springer (2018)

[4] Chen, Z., Shi, B.E.: Geddnet: A network for gaze estimation with dilation and decomposition. arXiv preprint arXiv:2001.09284 (2020)

[5] Cheng, T., Huang, F., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10635–10640 (2020)

[6] Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. IEEE Transactions on Image Processing 29, 5259–5272 (2020)

[7] Fischer, T., Chang, H.J., Demiris, Y.: Ri-gen: Real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 334–352 (2018)

[8] Hempel, T., Al-Hamadi, A.: Slam-based multistate tracking system for mobile human-robot interaction. In: International Conference on Image Analysis and Recognition. pp. 368–376. Springer (2020)

[9] Hilthofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6912–6921 (2019)

[10] Krafka, K., Khosla, A., Kellhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2176–2184 (2016)

[11] Li, L., Yu, X., Li, J., Wang, G., Shi, J.Y., Tan, Y.K., Li, H.: Vision-based attention estimation and selection for social robot to perform natural interaction in the open world. In: 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 183–184. IEEE (2012)

[12] Liu, S., Liu, D., Wu, H.: Gaze estimation with multi-scale channel and spatial attention. In: Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition. pp. 303–309 (2020)

[13] Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE transactions on pattern analysis and machine intelligence 36(10), 2033–2046 (2014)

[14] O Oh, J., Chang, H.J., Choi, S.I.: Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4992–5000 (2022)

[15] Patney, A., Salvi, M., Kim, J., Kaplanayan, A., Wyman, C., Benty, N., Luebke, D., Lefohn, A.: Towards foveated rendering for gaze-tracked virtual reality. ACM Transactions on Graphics (TOG) 35(6), 1–12 (2016)

[16] Strazdas, D., Hintz, J., Khalifa, A., Abdelrahman, A.A., Hempel, T., Al-Hamadi, A.: Robot system assistant (rosa): Towards intuitive multimodal and multi-device human-robot interaction. Sensors 22(3), 923 (2022)

[17] Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11907–11916 (2019)

[18] Williams, O., Blake, A., Cipolla, R.: Sparse and semi-supervised visual mapping with the s’ 3gp. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). vol. 1, pp. 230–237. IEEE (2006)

[19] Xiong, Y., Kim, H.J., Singh, V.: Mixed effects neural networks (menets) with applications to gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7743–7752 (2019)

[20] Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Ethxgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: European Conference on Computer Vision. pp. 365–381. Springer (2020)

[21] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It’s written all over your face: Full-face appearance-based gaze estimation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pp. 2290–2308. IEEE (2017)

[22] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE transactions on pattern analysis and machine intelligence 41(1), 162–175 (2017)