Enhancing Appearance-Based Gaze Estimation Through Attention-Based Convolutional Neural Networks

Rawdha Karmi[1,3,*], Ines Rahmany[2,3] and Nawres Khlifa[3]
[1]*National School of Electronics and Telecoms of Sfax, University of Sfax, Tunisia*
[2]*Faculty of Sciences and Techniques of Sidi Bouzid, University of Kairouan, Tunisia*
[3]*Research Laboratory of Biophysics and Medical Technologies, Higher Institute of Medical Technologies of Tunis, University of Tunis El Manar, 1006 Tunis, Tunisia*

Keywords: Appearance-Based, Gaze Estimation, Convolutional Neural Network, Attention Mechanism.

Abstract: Appearance-based gaze estimation is crucial for applications like assistive technology and human-computer interaction, but high accuracy is challenging due to complex gaze patterns and individual appearance variations. This paper proposes an Attention-Enhanced Convolutional Neural Network (AE-CNN) to address these challenges. By integrating attention submodules, AE-CNN improves feature extraction by focusing on the most relevant regions of input data. We evaluate AE-CNN using the ColumbiaGaze dataset and show that it surpasses previous methods, achieving a remarkable accuracy of 99.98%. This work advances gaze estimation by leveraging attention mechanisms to improve performance.

1 INTRODUCTION

One of the most effective ways to read someone's attention, interest, and even emotions is through their stare. By enabling robots to dynamically adjust their behavior in response to user requirements and reactions, the ability to capture and analyze these visual signals might greatly improve how machines and people interact. However, there are several technical issues involved in effectively predicting gaze from still images or video streams, especially given the variety of facial looks, different lighting situations, and small eye movements.

Within this framework, new avenues for the study of gaze estimation are made possible by developments in artificial intelligence, particularly the introduction of convolutional neural networks (CNN) and attention mechanisms. These advanced models are especially well-suited to the challenging task of interpreting visual cues from human gaze because they can learn extremely abstract representations from raw visual data.

Deep learning (DL) models' attention mechanisms are modeled after how the human brain handles information, especially regarding gaze estimation. For example, when we look at an image, our brain automatically focuses on the most crucial aspects and filters out irrelevant information, facilitating our ability to comprehend and analyze the image more quickly. Along the same lines, the attention processes of DL models facilitate the automatic identification of the most crucial elements in an input, this can enhance the model's functionality for a variety of applications. By adding gaze estimation into this process, models can also be trained to identify important facial and ocular regions for accurate gaze direction estimation.

Depending on the kind of model and the particular objective, different attention strategies are implemented in different deep learning models. Typically, this method is implemented in attention-based CNNs by adding a layer that sets the weights for each input feature. These weights improve the model's performance on the given task by enabling the CNN to concentrate on the most pertinent elements in the input.

Attention-based CNNs have demonstrated strong performance on a range of tasks, such as as image classification (Wang et al., 2018), natural language processing (Vaswani, 2017) and object recognition (Hu et al., 2018). These networks are very skilled at teaching themselves to concentrate on particular elements or objects within the provided photos, hence increasing the classification accuracy of the model. When applied for gaze estimation tasks, attention-based CNNs can learn to prioritize significant facial

*Corresponding author

Karmi, R., Rahmany, I. and Khlifa, N.
Enhancing Appearance-Based Gaze Estimation Through Attention-Based Convolutional Neural Networks.
DOI: 10.5220/0013055500003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 15-23
ISBN: 978-989-758-737-5; ISSN: 2184-433X

characteristics and eye regions linked to head attitude and gaze direction, hence improving the gaze estimation models' accuracy.

In particular, our research examines the use of these novel technologies in an appearance-based gaze estimate method. We use facial feature analysis to determine the subject's gaze direction instead of only identifying eye positions in an image. A CNN, a specific kind of neural network intended for pattern identification in complicated input, like photographs, is used to make this feasible. Our method improves gaze estimation accuracy by permitting the model to concentrate on the facial regions that are most pertinent through the integration of attention mechanisms.

Our study's main goal is to determine whether adding attention mechanisms to convolutional neural networks may improve appearance-based gaze estimation's accuracy over alternative techniques. We take on a number of challenges to address this question: managing the wide range of facial features.

The structure of this paper is as follows: Section 2 presents the literature review. Section 3 outlines the proposed methodology. Section 4 reports the experimental results and their analysis. Finally, Section 5 concludes the paper and offers suggestions for future research.

2 LITERATURE REVIEW

In recent years, the literature has reported numerous remote gaze estimation algorithms, which fall into two categories: appearance-based and model-based techniques. Appearance-based methods focus on analyzing the visual appearance of the eyes and surrounding regions to infer gaze direction, often leveraging machine learning algorithms like convolutional neural networks (CNNs). Conversely, model-based approaches rely on geometric models of the eye, head, and scene geometry to estimate gaze direction, requiring precise calibration and modeling of eye and head movements. This classification framework is commonly used to categorize gaze estimation techniques. Each category offers unique strengths and limitations, with appearance-based methods proving robust to variations in lighting and head movements, while model-based techniques can provide more accurate estimates under controlled conditions. By understanding and exploring both categories, we aim to contribute to the advancement of remote gaze tracking technology.

2.1 Appearance-Based Techniques for Gaze Estimation

CNNs, or convolutional neural networks, were employed by (Choi et al., 2016) in order to estimate head posture with driver gaze area categorization (the left window, the center rearview mirror, and the right and left sections of the windshield). Their system attains a 95% accuracy rate, and They've created a unique dataset comprising both male and female drivers, even in scenarios where they are donning spectacles.

The gaze tracking study by (Konrad et al., 2016) was conducted in a highly restricted setting, with the camera positioned 51 cm away from the subject's face. They created a data set consisting of five subjects' images and used it to devolop their CNN neural networks. Although the results show promise, to properly train the CNN, a substantial amount of data is needed.

An altered form of the Viola-Jones formula is used by George and Routray's algorithm (George and Routray, 2016) to identify faces in an image. Facial landmarks and geometric relations are used to identify the rough eye region. Next, the classification of gaze direction is performed using a convolutional neural network. This algorithm performs well in terms of computational complexity when tested on the Eye Chimera data set, making it a viable option for smart devices.

To enhance appearance-based gaze estimate,(Chen and Shi, 2018) suggested using dilated convolutions. When compared to traditional networks, the results demonstrate notable improvements in accuracy. On the MPIIGaze and Columbia Gaze datasets, the study uses Dilated-Nets to attain cutting-edge performance. This development could enhance human-machine interaction by using eye tracking to identify user intentions.

A method for detecting eye contact was proposed by (Omori and Shima, 2020) in 2020. It involved using both an SVM and a CNN that had been trained beforehand for both eye area images. Furthermore, instead of creating eye images, tests revealed that a CNN pre-trained on object photo datasets could be utilized as an extractor of features for both eye regions. Utilized was the Cave dataset, which includes 5880 images of 56 individuals. According to experiments, picture augmentation can enhance the precision of the two-class eye contact classification. Additionally, the statistics show that the peak 86% accuracy rate with glasses was 5% lower than the detection accuracy of 91.04 percent without spectacles.

(Ewaisha et al., 2020) proposed a multitasking convolutional neural network to enhance gaze region

accuracy. Realizing the value of recording the underlying distance between gaze regions, they introduced regression-based forecasting of gaze yaw and pitch angles. Furthermore, their model used multitask learning to predict head posture angles and gaze simultaneously. An evaluation using the database Columbia Gaze, which has 5880 high-resolution images of 56 participants, showed remarkable accuracy of 78.2% in between-subjects tests and 95.8% on the test set, proving that the model is generally applicable and robustness over head attitude variations.

In 2024, (Karmi et al., 2024) developed a new neural network architecture named "CoGaze-Net", which exploits the concepts of cascade and bilinearity to improve both the originality and efficiency of the results. Their innovative architecture is composed of multiple cascading processing layers, each dedicated to performing a specific transformation on the input data. By integrating cascading, bilinearity and their lightweight and efficient architecture, they achieved exceptional results, reaching 96% accuracy, which capture complex information in the input data. The scientists used the Columbia Gaze database, comprising 5,880 photos.

Critical Analysis of Existing Methods and Gaps. Existing appearance-based gaze estimation methods, while robust to variations in lighting and head poses, often struggle with accurately capturing subtle eye movements. Furthermore, these methods face challenges in generalizing across diverse facial appearances and environmental conditions, due to limited exposure to varied training scenarios. Model-based techniques, on the other hand, demand precise calibration and are highly sensitive to imaging conditions, limiting their practical applicability. An often-overlooked aspect in prior works is the strategic use of data augmentation, which can significantly enhance model robustness without introducing distortions that compromise gaze estimation accuracy.

Our proposed AE-CNN addresses these critical issues by combining appearance-based learning with tailored data augmentation techniques. Unlike prior methods, our approach emphasizes the following:

- **Targeted Data Augmentation.** By simulating realistic variations in zoom, cropping, and brightness, our method exposes the model to a broader range of conditions, enhancing generalization while avoiding artificial distortions such as rotations that can misrepresent natural gaze orientations.
- **Incorporation of Attention Mechanisms.** Attention modules within the CNN architecture focus on the most relevant eye and facial features, ensuring precise capture of gaze cues even under challenging conditions.
- **Improved Robustness and Generalization.** By addressing the diversity of facial appearances and environmental conditions, our method overcomes the limitations of previous approaches, providing superior performance for practical applications.

Through this combination of innovations, our method effectively bridges the gap between the robustness of appearance-based techniques and the need for fine-grained gaze estimation, setting a new benchmark in gaze estimation research.

2.2 Model-Based Techniques for Gaze Estimation

Model-based approaches to gaze estimation require accurate detection of facial features like eye centers or corners,Poulopoulos and Psarakis presented DeepPupil Net, a fully convolutional neural network (FCN) designed to accurately localize eye centers. Using an encoder-decoder architecture, the model performs image-to-heatmap regression to map eye regions onto heat maps corresponding to eye center positions. A novel loss function is introduced to penalize inaccurate localizations and improve accuracy. The model achieves real-time performance and outperforms existing techniques in eye center localization accuracy. Evaluations on three public databases show significant improvements, making it a promising solution for low-cost eye tracking devices (Poulopoulos and Psarakis, 2022), and they rely on fitting a geometric model of the eye to the image of the eye; Park et al (Park et al., 2018) presented a novel gaze estimation method suitable for real-world, unconstrained environments. Their approach relies on a machine learning model that accurately localizes landmarks in the eye region using synthetic data for training. This model outperforms existing methods in terms of iris localization and eye shape registration on real images. The detected landmarks are then used as the basis for lightweight, iterative model-based gaze estimation methods. This method outperforms traditional and appearance-based methods, even in the presence of variations in postures, facial expressions, and lighting. (Wang and Ji, 2018) presented a novel model-based 3D gaze estimation method that overcomes the limitations of personal calibration techniques. Their approach uses four natural constraints for implicit calibration, without requiring intrusive calibration. The constraints are: (1) consistency between two different gaze estimation methods, (2) the principle of the center of the screen where the majority of fixations are concentrated, (3) the concentration of gaze in the

screen region for console-based interactions, and (4) the anatomical limits of eye parameters. These constraints are embedded in an unsupervised regression problem solved by an iterative hard-EM algorithm. Experiments conducted on everyday interactions such as web browsing and video watching demonstrate the effectiveness of this implicit calibration method. However, image resolution and lighting play a major role in their accuracy, which can lead to subpar performance in practical scenarios. Numerous geometric techniques involved either directly fitting a 3D face-eye deformable model (Chen and Ji, 2008), in 2015, Sun et al.(Sun et al., 2015) presented a novel, low-cost, non-intrusive, and easy-to-implement eye tracking system using a consumer-grade depth camera (Kinect) instead of infrared lights and high-quality cameras. Using a 3D eye model and a parametric iris model, this system can estimate gaze direction with an accuracy of 1.4 to 2.7 degrees, even in the presence of natural head movements. Two real-time applications, a chess game and text input, demonstrate the robustness and feasibility of the system. In the future, the goal is to develop a version using a simple webcam and explore new applications, (Rahmany et al., 2018),or anchoring the result to a stationary point on the face, Wang and Ji (Wang and Ji, 2017) proposed a novel 3D model-based gaze direction estimation method, enabling accurate and real-time eye tracking using a single webcam. Using a deformable 3D model of the eye and face, the system can efficiently estimate gaze direction from 2D facial landmarks. The model is pre-trained offline using data from multiple subjects, facilitating fast and efficient calibration for each user without requiring additional hardware. Experimental results show that this method outperforms existing approaches in terms of accuracy, while providing simplified setup and the ability to achieve natural head movements. Some techniques ((Venkateswarlu et al., 2003), (Wood and Bulling, 2014), (Aboudi et al., 2023)), fit an ellipse to the observed iris and compute the position to obtain the gaze. We have showcased the most advanced methods currently available for estimating gaze. Most gaze estimate techniques rely on one of two approaches: appearance-based techniques or model-based techniques. Table 1 shows the prior research done on estimating glance direction.

3 APPROACH

Using labeled data, our approach seeks to precisely identify and categorize an individual's gaze direction. Fig. 1 illustrates how the model incorporates an attention submodule along with basic CNN pipelines. Targeting areas of high human activity enhances their influence while removing irrelevant and potentially disruptive information from other portions of the sensor data. This is an advantage of the integrated attention mechanisms. When fully labeled data is used for supervised learning, this method works especially well. Our approach is specifically tailored to reliably identify gaze direction from labeled data, which is frequently encountered in practical applications. It is able to reduce classification errors and effectively filter out background noise signals by concentrating attention on pertinent components of the sensor data sequence. Additionally, we present a novel method for converting compatibility densities from compatibility scores. This approach is customized to the unique properties of sensor data, which are structurally and characteristically different from image data. Our approach enables precise gaze direction localization by converting compatibility scores into compatibility densities, which enhances the model's overall performance in practical circumstances.

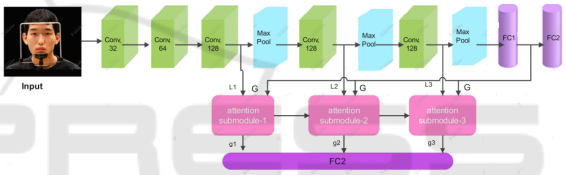

Figure 1: Our AE-CNN architecture.

Our approach aims to accurately estimate an individual's gaze direction using labeled data. To provide a clear overview of our methodology, Figure 2 illustrates the research framework, detailing all the steps involved from input data to model evaluation.

The key steps in our methodology are as follows:

1. **Input Data.** We utilize labeled images from the publicly available **Columbia Gaze** dataset, which provides detailed annotations specifically designed for gaze estimation tasks.
2. **Data Preprocessing.**

- **Normalization.** The images are resized and standardized to ensure uniform and consistent input for the model.
- **Data Augmentation.** To enhance the robustness of our model, we applied data augmentation techniques such as zooming, cropping, and brightness adjustment, which simulate realistic variations in shooting conditions. However, we avoided transformations like rotation, as they can distort the natural alignment of the eyes and face, potentially introducing errors, especially in models that rely on facial geometry and gaze direction.

Table 1: Review of the literature on gaze estimation methods.

Authors	Year	Method	Database	Result
Choi et al. (Choi et al., 2016)	2016	CNNs	Own dataset created	95%
Konrad et al. (Konrad et al., 2016)	2016	fundamental CNN model (based on the LeNet design)	Tablet Gaze Calibration Eye Dataset	6.7°
George and Routray (George and Routray, 2016)	2016	Viola-Jones modification for fast face detection, facial cues to localize eye region, and CNN to classify gaze direction.	Eye Chimera Dataset	—
Chen and Shi (Chen and Shi, 2018)	2018	Dilated-CNN	Columbia Gaze	62%
Omori and Shima (Omori and Shima, 2020)	2020	SVM, CNN	Columbia Gaze	91.04%
Ewaisha et al. (Ewaisha et al., 2020)	2020	Multitask Learning	Columbia Gaze	95.8%
Karmi et al. (Karmi et al., 2024)	2024	CoGaze-Net	Columbia Gaze	96.88%
Poulopoulos and Psarakis (Poulopoulos and Psarakis, 2022)	2022	DeepPupil Net	MUCT, (Milborrow et al., 2010) BioID (Jesorsky et al., 2001) and Gi4E (Villanueva et al., 2013)	8.5°
Wang and Ji (Wang and Ji, 2018)	2018	hard-Expectation Maximization (hard-EM)	Real data and synthetic data	1.3°
Park et al. (Park et al., 2018)	2018	CNN and Support Vector Regression (SVR) based on landmarks	Columbia and EYEDIAP	7.1°

3. **Attention-Enhanced Convolutional Neural Network (AE-CNN).** Our primary model, **AE-CNN**, leverages attention mechanisms to enhance the network's ability to focus on critical regions of interest for gaze estimation.
4. **Attention Mechanisms.** Attention mechanisms allow the model to automatically identify relevant regions within the images, such as the eyes and other key facial features, thereby improving prediction accuracy.
5. **Feature Extraction.** The **AE-CNN** extracts hierarchical representations of facial and ocular regions. These features are used to model the relationship between the input data and the gaze angles.
6. **Gaze Estimation.** Using the extracted features, the model predicts both gaze angles and head orientation. These predictions are made within a supervised learning framework using labeled data.
7. **Model Evaluation.**

- The model's performance is assessed using standard metrics such as the mean angular error.
- We compare the performance of the **AE-CNN** with other state-of-the-art methods to demonstrate its superiority.

This research framework provides a comprehensive and structured overview of our methodology. It highlights the innovative integration of attention mechanisms into the **AE-CNN**, which significantly enhances the accuracy of gaze estimation.

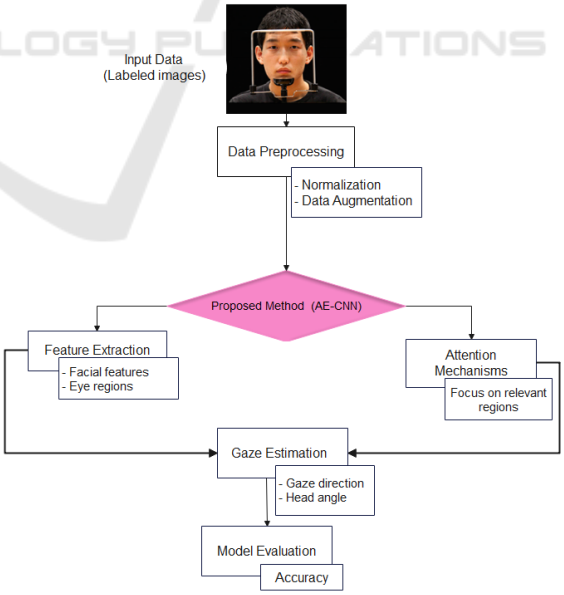

Figure 2: Flow Diagram of the AE-CNN Gaze Estimation Method.

3.1 Attention Submodule

Our attention method, which is shown in Fig. 1, primarily emphasizes carrying out a compatibility calculation connecting the locally generated feature vectors at intermediate CNN pipeline layers to the vector of global features that was previously sent into the pipeline's final levels of linear classification (Jetley et al., 2018). The collection of extracted feature vectors at a specific convolutional layer $t \in \{1,2,\ldots,T\}$ is denoted by $L^t = \{l_1^t, l_2^t, l_3^t, \ldots\}$, where in the region's feature vector L^t, l_j^t represents the j-th feature array of m total spatial places.

Previously supplied into the last set of fully linked layers to obtain classification results, a compatibility function C now combines local characteristic vectors L^t with the general feature vectors H produced by data input that navigates whole CNN streams. There are several methods to define the compatibility function C as stated in (Bahdanau, 2014), (Xu et al., 2015), we can apply a measure of the weight vector's point-product across U and $l_j^t + H$ to concatenate the H and l_j^t:

$$c_j^t = \langle U, l_j^t \rangle + H, \quad j \in \{1,\ldots,m\} \tag{1}$$

where c_j^t is the compatibility score and U represents the weight vector. In this case, U can help learn the universal collection of characteristics that are pertinent to the several categories of activity within the sensor's data. Furthermore, we may evaluate the compatibility of H and l_j^t using the dot product:

$$c_j^t = \langle l_j^t, H \rangle, \quad j \in \{1,\ldots,m\} \tag{2}$$

Following the computational process, we obtain a collection of compatibility scores $C(L^t, H) = \{c_1^t, c_2^t, \ldots, c_3^t\}$, which are subsequently standardized into which are subsequently standardized into $A^t = \{a_1^t, a_2^t, a_3^t\}$ by a softmax function.

$$a_j^t = \frac{\exp(c_j^t)}{\sum_{i=1}^m \exp(c_i^t)} \tag{3}$$

Or the tanh function :

$$a_j^t = \frac{\exp(c_j^t) - \exp(-c_j^t)}{\exp(c_j^t) + \exp(-c_j^t)} \tag{4}$$

The A^t standardized compatibility scores are utilized in order to generate a unique vector h^t through element-wise weighted averaging for every layers.

$$h^t = \sum_{j=1}^{m} a_j^t \cdot l_j^t \tag{5}$$

Most importantly, globally descriptive feature of the input data H of the incoming information can now be substituted with the h^t. The newly created global vectors, are provided as input to the classification phase after being connected to form a single vector $h = [h^1, h^2, \ldots, h^t]$.

3.2 Basic CNN

Our core CNN, which is the foundation upon which the attention submodules are built, consists of fully connected, pooling, and convolutional layers. The summary is as follows, as shown in Fig.1 : Conv(32); Conv(64); Conv(128); Pol ; Conv(128); Pol; Conv(128); Pol; FC(128) ; softmax, in which $Conv(L^t)$ stands for a L^t feature map-equipped convolutional layer s, Maximum Pooling Layer Pol, A layer of m units that is fully connected is denoted by FC(m) and a softmax classifier by softmax. In addition, a ReLU activation function is used to alter each layer's output.

In order to guarantee an improved resolution for the local feature maps extracted from the three convolutional layers Conv(128) that are used to estimate attention, After the initial two convolutional layers, there is neither pooling layer. In order to prevent mapping distinct feature vector dimensionalities to the same one, which could result in additional processing costs, we purposefully designed the global features H and the local features L^t (both of which are 128) to have the identical dimensions.

Finally, we employed the Location Function. In a basic CNN design, the class prediction probabilities are typically generated by a fully connected layer after a global characteristic description H has been calculated from the input dataset. For the classes to be able to be divided from one another linearly, the information needs to be translated into a high-dimensional space in order to clarify H with discrete dimensions that reflect significant higher-order data concepts. The attentional method adds filters early in the CNN pipeline so that it can discover mappings that are comparable and work with the original design to produce H. The compatibility rating Conv(L^t, H) ought to only be elevated when the fraction of the dominant data category is present in the related patch, as a result of our model's attention mechanism.

4 EXPERIMENT AND RESULTS

The experiments conducted in this study aim to evaluate the effectiveness of the proposed AE-CNN model for appearance-based gaze estimation. By leveraging the Columbia Gaze dataset, we assess the performance of our approach in comparison to prior state-

of-the-art methods. The evaluation process is structured into three key components: the dataset characteristics and preprocessing steps, the architecture and experimental setup of the proposed model, and the analysis of the obtained results. This section presents a detailed breakdown of these elements to demonstrate the robustness and accuracy of the proposed approach.

4.1 Data SET for Gaze Estimation

The Columbia Gaze dataset (Smith et al., 2013) is one of the most varied datasets that was collected in a controlled setting and uses degrees to describe pitch and yaw. A comprehensive set of eye gaze data with varying gaze directions and head positions is provided by the 5,880 photographs of 56 individuals. This dataset outperforms other eye gazing datasets that were available to the public at the time of release in terms of the quantity of individuals. Almost half of the sample's participants wear glasses, which is an essential characteristic given their diverse ethnic backgrounds.

This dataset's high-quality images all have a resolution of 5184 x 3456 pixels, which is excellent for forecasting the dataset itself but not indicative of the typical camera used for these kinds of images, thus it doesn't enable the training model to predict data more accurately under different situations. An additional limitation of this dataset is that the lighting is nearly constant across the images, which reduces the robustness of the trained models in varying lighting conditions. Fig. 3 shows a sample of the dataset with varying head positions.

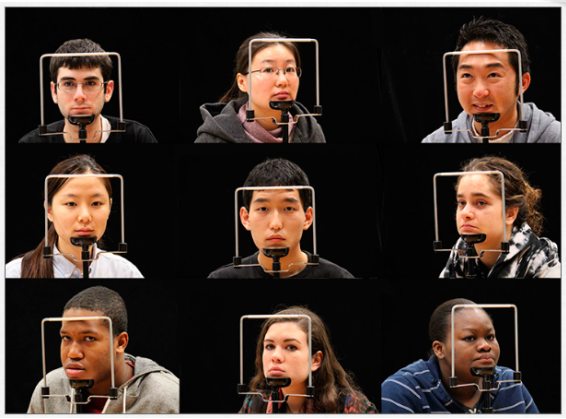

Figure 3: An excerpt from the Columbia Gaze dataset (Smith et al., 2013).

We selected the Columbia Gaze dataset as our primary dataset due to its diversity in individual participants and head positions, which makes it well-suited for testing appearance-based gaze estimation models.

4.2 Model and Experimental Setup

In our study on gaze tracking, we adopted an empirical method to explore the effectiveness of convolutional neural networks (CNNs), by integrating an attention mechanism. The inclusion of attention mechanisms aims to improve the model's ability to focus on relevant features during gaze estimation.

Python is used to implement our model, and the Tensor Flow machine learning technology is used. We used the "fit()" function from the Keras library to train our models, setting hyperparameters such as the "adam" optimizer the categorical-crossentropy Loss function. The rate of learning was set to 0.001 and the input batch size was 128, the number of epochs was set to 10. To evaluate our approch, we split our data into validation and training sets, with 20% being used for validation and 80% being used for training.

4.3 Performance Comparison

Our appearance-based gaze estimation method performs better than previous methods, as Table 2 illustrates. To assess the effectiveness of the AE-CNN model, we conducted a one-tailed z-test on the Columbia Gaze dataset. This statistical test was chosen as it is well-suited for evaluating performance improvements over established benchmarks.

When we compare the outcomes, we see that performance has significantly improved over time. Although the Dilated-CNN method used by Chen and Shi (Chen and Shi, 2018) demonstrated a very modest accuracy of 62%, significant increases were noted with subsequent methods. While Omori and Shima (Omori and Shima, 2020) used SVM and CNN to reach an accuracy of 91.04%, Ewaisha et al. (Ewaisha et al., 2020) included multi-task learning and achieved an amazing accuracy of 95.8%. With the introduction of the CoGaze-Net by Karmi et al. (Karmi et al., 2024), the accuracy was raised to 96.88%. However, our suggested method,a CNN based on attention for gaze estimation, surpassed all prior methods with an exceptional 99.98% accuracy.

In terms of relative improvement, the AE-CNN model outperformed the Dilated-CNN by 37.98%, Multitask Learning by 4.18%, SVM-CNN by 8.94%, and CoGaze-Net by 3.1%. These improvements are statistically significant, with p-values of < 0.01 and < 0.05 for the comparisons with the previous methods. This demonstrates that the observed performance gains are not due to random chance but result from the superior feature extraction capability of the attention-based CNN architecture.

This development suggests that adding attention

Table 2: Evaluation of the suggested method's outcomes in relation to earlier approaches using the "columbia Gaze" data set.

Author	Method	Batch size	Epochs	Optimizer	Classificateur	Accuracy
Chen and Shi (2018)	Dilated -CNN	64	8000	Adam	Softmax	62%
Ewaisha et al. (2020)	Multitask Learning	-	-	-	-	95.8%
Omori and Shima, (2020)	SVM, CNN	-	-	SGD	-	91.04%
Karmi et al. 2024	CoGaze -Net	128	50	Adam	Softmax	96.88%
Proposed method	AE -CNN	128	10	Adam	Softmax	99.98%

mechanisms to CNNs has a significant potential to enhance appearance-based gaze estimating systems' functionality, marking a significant advancement in the field. As shown in Fig.4, the outcomes are also evaluated by comparing the loss functions from the validation and learning phases, as well as by tracking the evolution of the validation precision values in relation to the training values.

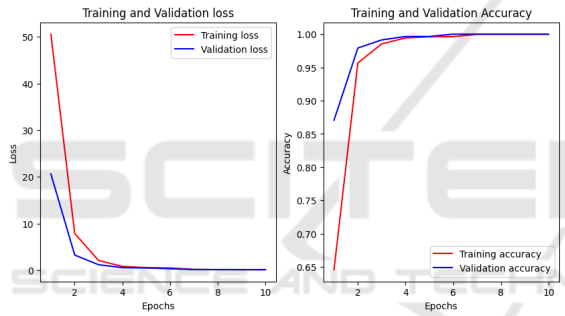

Figure 4: The two example graphs showing the accuracy and loss function evolution for the suggested technique.

5 CONCLUSION

Our exploration of an attention-based CNN for gaze direction estimation image classification was fruitful. The experimental result we obtained clearly demonstrates that our proposed attention model results in a clear improvement in accuracy compared to previous approaches, as we observed on the ColumbiaGaze dataset. This finding highlights the effectiveness of our attention mechanisms regarding appearance-based gaze estimate. In conclusion, this study reinforces the credibility of our attention-based convolutional neural networks as powerful tools to increase precision and reliability of estimation of gaze systems, thus paving the way for further advances in fields like human-computer interaction and visual recognition.

For future work, several promising avenues are open to improve appearance-based gaze direction estimation. A first approach would be to refine our model by integrating more advanced learning techniques, such as more complex attention mechanisms or hybrid neural network architectures, which could more accurately capture subtle variations in eye and facial appearance under various lighting conditions and angles.

Furthermore, it would be interesting to extend our research to other datasets, particularly those containing a wider variety of subjects, facial expressions, and environmental contexts, to assess the robustness and generalization ability of our approach. This could include using data from real-world scenarios, such as videos captured in dynamic environments or mobile applications, to test and improve the model's performance in real-world situations.

Furthermore, we envision generating novel methods by combining appearance-based gaze direction estimation with transfer learning or meta-learning techniques, which would reduce the need for annotated data and improve the model's effectiveness on new domains. Creating models that can quickly adapt to new users or changing conditions, without requiring manual recalibration, would represent a significant advance in the field.

Finally, exploring new data sources, such as those synthesized by generative neural networks, could also provide opportunities to increase the diversity of training data and improve model performance in challenging scenarios where real-world data are limited.

REFERENCES

Aboudi, N., Khachnaoui, H., Moussa, O., and Khlifa, N. (2023). Bilinear pooling for thyroid nodule classification in ultrasound imaging. *Arabian Journal for Science and Engineering*, 48(8):10563–10573.

Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chen, J. and Ji, Q. (2008). 3d gaze estimation with a single

camera without ir illumination. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.

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

Choi, I.-H., Tran, T. B. H., and Kim, Y.-G. (2016). Real-time categorization of driver's gaze zone and head pose using the convolutional neural network. In *Proceedings of HCI Korea*, pages 417–422.

Ewaisha, M., El Shawarby, M., Abbas, H., and Sobh, I. (2020). End-to-end multitask learning for driver gaze and head pose estimation. *Electronic Imaging*, 32:1–6.

George, A. and Routray, A. (2016). Real-time eye gaze direction classification using convolutional neural network. In *2016 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Jesorsky, O., Kirchberg, K. J., and Frischholz, R. W. (2001). Robust face detection using the hausdorff distance. In *Audio-and Video-Based Biometric Person Authentication: Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6–8, 2001 Proceedings 3*, pages 90–95. Springer.

Jetley, S., Lord, N. A., Lee, N., and Torr, P. H. (2018). Learn to pay attention. *arXiv preprint arXiv:1804.02391*.

Karmi, R., Rahmany, I., and Khlifa, N. (2024). Gaze estimation using convolutional neural networks. *Signal, Image and Video Processing*, 18(1):389–398.

Konrad, R., Shrestha, S., and Varma, P. (2016). Near-Eye Display Gaze Tracking Via Convolutional Neural Networks. Standford University: Standford, CA. *Technical report*, USA, Tech. Rep.

Milborrow, S., Morkel, J., and Nicolls, F. (2010). The muct landmarked face database. *Pattern recognition association of South Africa*, 201(0):535.

Omori, Y. and Shima, Y. (2020). Image augmentation for eye contact detection based on combination of pre-trained alex-net cnn and svm. *J. Comput.*, 15(3):85–97.

Park, S., Zhang, X., Bulling, A., and Hilliges, O. (2018). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–10.

Poulopoulos, N. and Psarakis, E. Z. (2022). Deeppupil net: Deep residual network for precise pupil center localization. In *VISIGRAPP (5: VISAPP)*, pages 297–304.

Rahmany, I., Guetari, R., and Khlifa, N. (2018). A fully automatic based deep learning approach for aneurysm detection in dsa images. In *2018 IEEE international conference on image processing, applications and systems (IPAS)*, pages 303–307. IEEE.

Smith, B. A., Yin, Q., Feiner, S. K., and Nayar, S. K. (2013). Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280.

Sun, L., Liu, Z., and Sun, M.-T. (2015). Real time gaze estimation with a consumer depth camera. *Information Sciences*, 320:346–360.

Vaswani, A. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Venkateswarlu, R. et al. (2003). Eye gaze estimation from a single image of one eye. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 136–143. IEEE.

Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S., and Cabeza, R. (2013). Hybrid method based on topography for robust detection of iris center and eye corners. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(4):1–20.

Wang, K. and Ji, Q. (2017). Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1003–1011.

Wang, K. and Ji, Q. (2018). 3d gaze estimation without explicit personal calibration. *Pattern Recognition*, 79:216–227.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.

Wood, E. and Bulling, A. (2014). Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the symposium on eye tracking research and applications*, pages 207–210.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.