P-82

Non-Intrusive Appearance-Based Gaze Estimation for Real-World Clinical and Telehealth Applications

Se-Young Bak[1,2], Sardor Abdirayimov Odil Ugli[1], Yubin Kim[1,2], Eun-Hye Chung[1,2], Heegoo Kim[1,2], Eunyoung Cho[2,3], Miri Suh[2,3], Seyoung Shin[1,2,3], HyeongMin Jeon[1,2,3], and MinYoung Kim[1,2,3†]

[1]Digital Therapeutics Research Team, Department of Research, CHA Bundang Medical Center, Seongnam 13520
[2]Department of Rehabilitation Medicine, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam 13496
[3]Graduate School, CHA University, Pocheon-si, Gyeonggi-do 11160

Background and Purpose

- Model-based gaze estimation systems (e.g., Tobii) have been widely utilized in research and clinical settings; however, their reliance on specialized hardware and constrained experimental environments limits broader applicability.
- To overcome these limitations, the Bundang CHA Rehabilitation Research Team developed CHA-Gaze, an appearance-based gaze estimation algorithm that operates using only standard camera systems.
- CHA-Gaze is built upon the Adaptive Feature Fusion Network (AFF-Net), incorporating an additional head pose prediction mechanism into the conventional architecture. This modification was motivated by the premise that integrating supplementary information can enhance model performance. Specifically, the head pose vector is defined as $\nu = [x, y, z]$, where $x = \cos\theta \cdot \sin\psi$, $y = \sin\theta$, and $z = \cos\theta \cdot \cos\psi$.
- The present study aimed to evaluate whether CHA-Gaze demonstrates superior estimation accuracy compared with AFF-Net, a widely recognized deep learning model for appearance-based gaze estimation, under within-domain conditions.

Methods

- CHA-Gaze was evaluated using a unified validation protocol on the MPIIFaceGaze dataset, which comprises 37,590 images from 15 participants collected under semi-natural conditions.
- For each individual dataset, 80% of the images were randomly allocated to the training set and the remaining 20% to the testing set (Figure 1).
- The training and evaluation procedure was repeated 15 times per participant and mean Euclidean gaze estimation error was calculated across repetitions.
- Performance was compared among AFF-Net, CHA-Gaze with a small network size, and CHA-Gaze with a large network size.

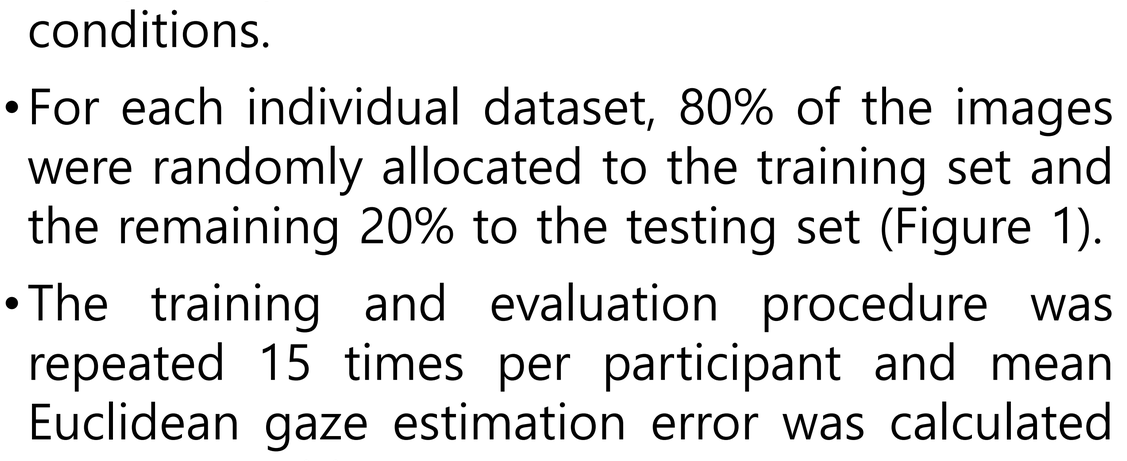

Figure 1. Schematic processes of training and testing gaze estimation models.

Results

- The results demonstrated that CHA-Gaze achieved a significantly lower mean Euclidean error (1.88 cm) than the baseline AFF-Net (2.59 cm; $p < 0.001$; Figure 2), despite comparable model complexity (Table 1).
- CHA-Gaze models with small (1.94cm) and large (2.33cm) network size also showed larger mean Euclidean error than original CHA-Gaze.

Attributes	CHA-Gaze	AFF-Net
GFLOPS	4.84	4.84
Latency (GPU) in seconds	0.008	0.007
Latency (CPU) in seconds	0.023	0.025
Training duration for one epoch over 37,000 images seconds	110	109
Size of model (assuming float 32) in MB	7.33	7.43
Number of parameters in millions	1.91	1.94
Number of epochs/ Number of backprop steps	25/12050	25/12050

Table 1. Comparison of model attributes between CHA-Gaze and AFF-Net.
GFLOPS, GPU floating point operations per second; GPU, graphics processing unit; CPU, central processing unit; MB, megabyte.

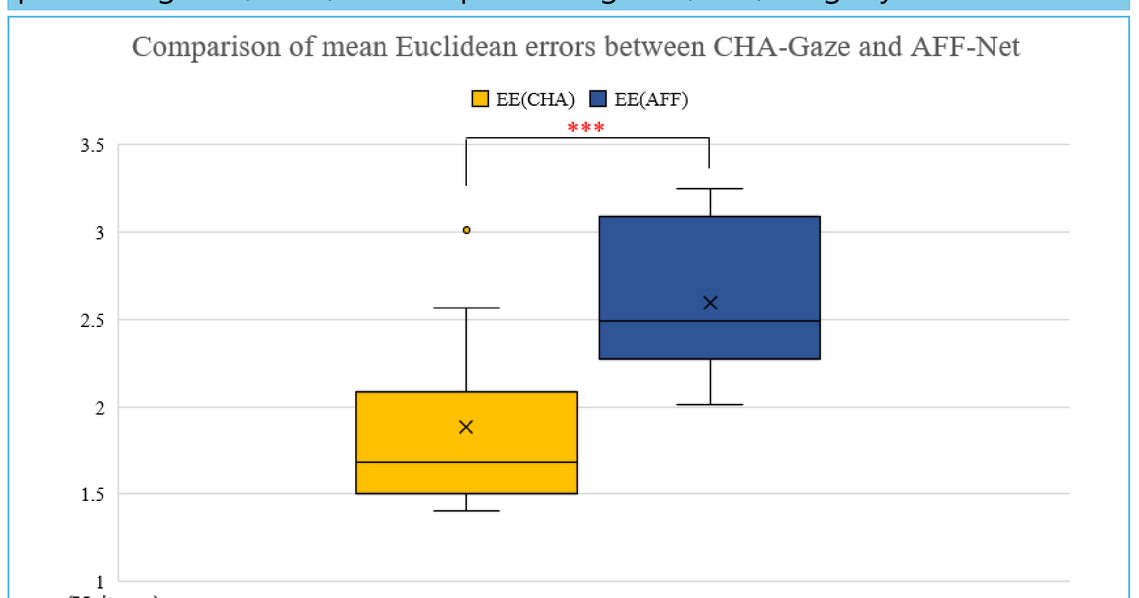

Figure 2. Comparison of mean Euclidean errors between CHA-Gaze and AFF-Net.
***, p <0.001; EE, Euclidean error; CHA, CHA-Gaze; AFF, AFF-Net

Conclusion

- Despite maintaining comparable model complexity to AFF-Net in terms of parameter count and computational cost (GFLOPs), CHA-Gaze achieves consistently superior performance, suggesting that architectural design choices contribute more substantially to accuracy gains than model scale alone.
- These findings underscore the significance of architectural refinement in appearance-based gaze estimation and substantiate the viability of CHA-Gaze for real-world deployment across digital therapeutics, telehealth, and accessibility domains, offering a scalable and non-intrusive solution operable on standard webcams.

Acknowledgement

This research was supported by the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (HR22C1605) and the Korea Planning and Evaluation Institute of Industrial Technology (KEIT), funded by the Ministry of Trade, Industry and Energy(RS-2025-08762968).