

Hybrid CNN-GRU Architecture for Efficient Emotion Recognition in Immersive Augmented Reality Systems

Saloni Joshi* and Priyanka Sharma

Faculty of Computer Science and Application, PAHER University, Udaipur, India

*Corresponding author: salonithakar67@gmail.com

Abstract. This paper presents a hybrid deep learning architecture that takes gaze data as input and recognizes emotions efficiently and in real time in resource-constrained augmented reality (AR) environments. Unlike conventional systems, which typically rely on facial expressions or speech signals, the proposed system uses eye tracking and gaze dynamics as the main indicators of emotional state, which offers better privacy, stability, and usability in practical AR applications. The model combines a convolutional neural network (CNN), which extracts spatial representations from eye-region images and gaze heatmaps, with a gated recurrent unit (GRU), which models time-varying gaze behavior. The model was evaluated on a self-curated dataset of 3,500 temporally sequenced samples annotated with six emotion classes. The proposed model achieved an accuracy of 91.2%, outperforming the baseline long short-term memory (LSTM) and standalone CNN models while remaining efficient. The GRU-based architecture was also more sensitive to temporal patterns and had a lower computational cost, with an average inference latency of 14.5 ms per sample, which meets real-time processing requirements.

Keywords. Emotion recognition; augmented reality; eye tracking; CNN; GRU; deep learning; affective computing

I. INTRODUCTION

Emotion recognition has become an important research topic in affective computing [1], human-machine interaction, and intelligent tutoring systems [2], where the user's emotional state can be used to adapt and personalize system behavior. Conventional emotion recognition methods rely on facial expressions [3], [4], speech signals [5], and physiological responses [6], [7], but these approaches often face limitations in augmented reality (AR) environments because of privacy concerns, environmental variability, and hardware constraints. This paper addresses these issues with a lightweight hybrid model that combines a convolutional neural network (CNN) and a gated recurrent unit (GRU). The CNN-GRU model captures the spatial and temporal features of gaze data to recognize emotions in AR in real time.

II. REVIEW OF LITERATURE

Recent studies have shown emotion recognition to be a central feature of affective computing, intelligent interaction systems, and immersive environments. Eye-gaze behavior has proven effective for recognizing emotional and anxiety-related states, especially in educational contexts [8]-[10]. Deep learning-based gaze analysis methods can identify spatial and temporal behavioral patterns and characteristics [11], [12]. Multimodal emotion recognition approaches that combine gaze with audiovisual and physiological signals have also achieved higher accuracy [13]-[16]. Emotion recognition has been applied in virtual reality (VR) [17],

human-robot interaction (HRI) [18], and personalized learning systems [19].

III. PERFORMANCE METRICS AND MATHEMATICAL FORMULATION

Let the dataset contain N samples and C emotion classes, where each input x_i has a corresponding label y_i . Accuracy measures overall correctness, (1); precision measures the reliability of predictions, (2); recall (sensitivity) measures detection capability, (3); the F1-score is the harmonic mean of precision and recall, (4); and inference latency is the time required per prediction, (5). Here TP , FP , and FN denote true positives, false positives, and false negatives.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}}, \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (3)$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (4)$$

$$\text{Latency} = T_{\text{infer}}. \quad (5)$$

IV. METHODOLOGY

The method uses eye-tracking and gaze data obtained with a HoloLens 2 headset equipped with the Tobii eye-tracking software development kit (SDK).

A. Gaze Feature Representation

Each input sample X is a sequence of eye-region images $I = [I_1, I_2, \dots, I_T]$, with $I_t \in \mathbb{R}^{h \times w \times c}$, and of gaze vectors $G = [g_1, g_2, \dots, g_T]$, with $g_t \in \mathbb{R}^2$ (for example, the $[x, y]$ gaze coordinates on the screen). The gaze vectors are encoded into gaze heatmaps $H_t \in \mathbb{R}^{h \times w}$ with a Gaussian kernel centered at each gaze point:

$$H_t(x, y) = \exp\left(-\frac{(x - g_t^x)^2 + (y - g_t^y)^2}{2\sigma^2}\right), \quad (6)$$

where σ controls the spread of the heatmap. Each frame F_t is represented as the composite image $\text{concat}(I_t, H_t)$, with $F_t \in \mathbb{R}^{h \times w \times (c+1)}$.

B. CNN Module for Spatial Feature Extraction

The CNN processes each frame F_t to extract spatial features $f_t \in \mathbb{R}^d$. A typical CNN stack computes

$$f_t = \text{CNN}(F_t) = \phi(W^{(c)} * F_t + b^{(c)}), \quad (7)$$

where $*$ denotes convolution, ϕ is the rectified linear unit (ReLU) activation, and $W^{(c)}$ and $b^{(c)}$ are the learnable weights and biases. Tables 1 and 2 show the successive layers of the convolutional pipeline that extract spatial features from the eye-region images. In these tables, X1 and Y1 are the spatial dimensions after the Conv2D layer, and X2 and Y2 are the dimensions after max pooling.

V. HYBRID ARCHITECTURE

The CNN takes these inputs and derives spatial features, while the other gaze features are processed by a dense layer. The joint features are then classified by a softmax layer into six emotion categories: happiness, sadness, anger, fear, surprise, and neutral. Fig. 1 depicts this hybrid deep learning architecture for identifying emotions from eye-tracking data in AR settings.

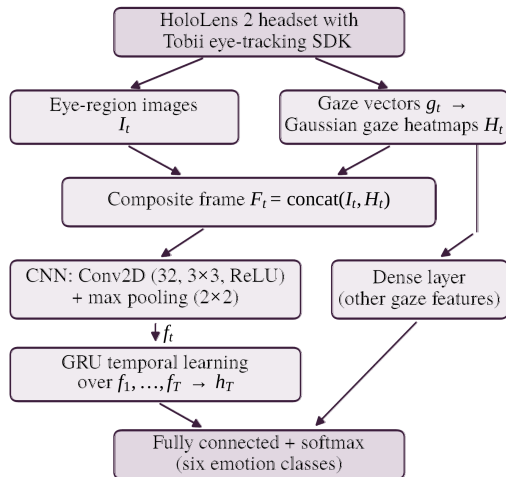


Fig. 1. System pipeline of the proposed hybrid CNN-GRU architecture.

TABLE 1. CONV2D LAYER (32 FILTERS, 3×3) FOR FEATURE EXTRACTION

Set	Filters	Kernel	Activation	Output shape	Trainable params.
Training	32	3×3	ReLU	(None, X1, Y1)	896
Test	32	3×3	ReLU	(None, X1, Y1)	896

TABLE 2. MAX POOLING LAYER FOR DIMENSIONALITY REDUCTION

Set	Pool	Stride	Input shape	Output shape	Purpose
Training	2×2	2	(X1, Y1, 32)	(X2, Y2, 32)	Reduces spatial dimension
Test	2×2	2	(X1, Y1, 32)	(X2, Y2, 32)	Avoids overfitting

A. GRU for Temporal Learning

To model the sequence $F = [f_1, f_2, \dots, f_T]$, a GRU is used; it suits real-time AR systems because it has fewer parameters than a long short-term memory (LSTM) network. The GRU cell is updated as follows:

$$z_t = \sigma(W_z f_t + U_z h_{t-1} + b_z), \quad (8)$$

$$r_t = \sigma(W_r f_t + U_r h_{t-1} + b_r), \quad (9)$$

$$\tilde{h}_t = \tanh(W_h f_t + U_h(r_t \odot h_{t-1}) + b_h), \quad (10)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \quad (11)$$

where z_t is the update gate, r_t is the reset gate, $\tilde{h}_t$ is the candidate hidden state, h_t is the hidden state, $\odot$ denotes element-wise multiplication, and σ is the sigmoid activation. The final hidden state h_T is passed to a fully connected layer.

B. LSTM for Temporal Feature Learning

LSTM networks address the vanishing gradient problem when deep or very long dependencies exist in sequential data, and they can therefore capture long-term dependencies. For emotion recognition from eye-tracking sequences, the LSTM allows the model to capture the temporal nature of gaze movements and the shifts across multiple frames.

C. Output Layer

The final prediction is

$$\hat{y} = \text{softmax}(W_o h_T + b_o), \quad (12)$$

where $\hat{y} \in \mathbb{R}^C$ is the probability distribution over the emotion classes and $C = 6$ (happiness, sadness, anger, fear, surprise, and neutral). The model is trained with the categorical cross-entropy loss:

$$L = - \sum_{i=1}^C y_i \log(\hat{y}_i). \quad (13)$$

VI. RESULT ANALYSIS AND DISCUSSION

The GRU-based and LSTM-based emotion recognition models were evaluated on temporally sequenced gaze data and eye-region images captured in AR environments. A total of 3,500 eye-tracking sequences were recorded during controlled AR interaction sessions with a HoloLens 2 headset and the Tobii eye-tracking SDK. The models were trained on an 80/20 train-test split with a batch size of 32 for 100 epochs to ensure convergence and adequate generalization. The GRU model consistently outperformed the LSTM model on all evaluation metrics.

Tables 3 to 7 and Fig. 2 compare the GRU and LSTM models on five key metrics under the same protocol (80/20 train-test split, batch size of 32, and 100 training epochs).

VII. COMPARATIVE PERFORMANCE

Because computational latency is a key factor in real-time AR, the GRU and LSTM models were compared

specifically to assess temporal learning efficiency under real-time AR constraints. Table 8 and Fig. 3 summarize the comparison.

TABLE 3. ACCURACY COMPARISON

Model	Accuracy (%)	Epochs	Batch size	Optimizer	Split
GRU	91.2	100	32	Adam	80/20
LSTM	89.6	100	32	Adam	80/20

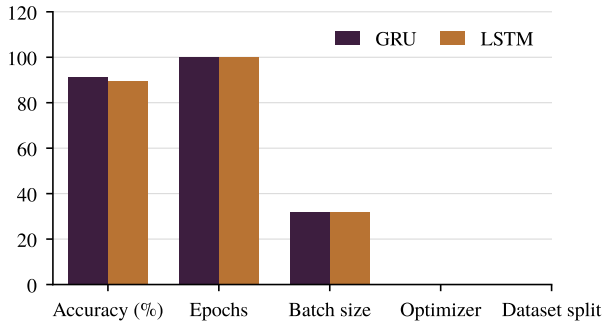


Fig. 2. Accuracy and training settings of the GRU and LSTM models.

TABLE 4. PRECISION COMPARISON

Model	Precision (%)	Class-wise mean	Split	Notes
GRU	90.1	High for happiness and fear	80/20	Slightly better than LSTM
LSTM	87.8	Average across classes	80/20	More confusion between anger and sadness

TABLE 5. RECALL COMPARISON

Model	Recall (%)	Split	Observations
GRU	92.4	80/20	Captures temporal cues well
LSTM	88.9	80/20	Missed fast-changing emotions

TABLE 6. F1-SCORE COMPARISON

Model	F1-score (%)	Split	Interpretation
GRU	91.2	80/20	Balanced precision and recall
LSTM	88.3	80/20	Slightly less balanced performance

TABLE 7. INFERENCE LATENCY COMPARISON

Model	Inference time (ms/sample)	Batch size	Notes
GRU	14.5	32	Lightweight, faster
LSTM	19.3	32	Heavier, with more gates

TABLE 8. PERFORMANCE MATRIX

Metric	GRU	LSTM
Accuracy	91.2%	89.6%
Precision	90.1%	87.8%
Recall	92.4%	88.9%
F1-score	91.2%	88.3%
Latency	14.5 ms	19.3 ms

The GRU model outperformed the LSTM model on all measured metrics: it was more precise, generalized better, and had lower latency. The GRU is therefore the preferable option for emotion recognition systems in immersive AR

and VR settings, where delay, latency, and accuracy are particularly critical.

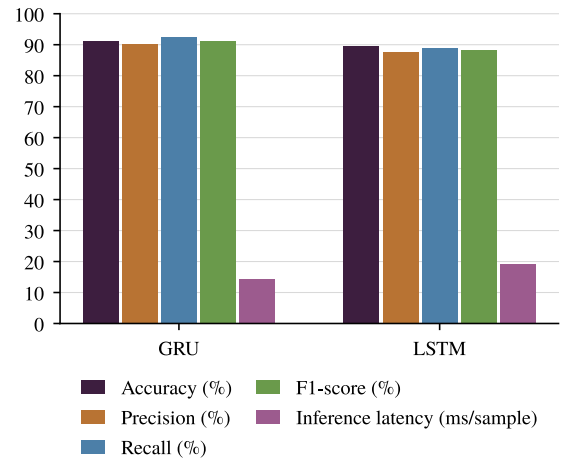


Fig. 3. Comparison of the GRU and LSTM models.

Table 9 compares the proposed CNN-GRU framework with other model configurations. It outperforms the other configurations on accuracy, precision, recall, and F1-score, with lower inference latency than the CNN-LSTM configuration, which makes it suitable for real-time AR-based emotion recognition systems.

TABLE 9. ABLATION ANALYSIS OF DIFFERENT MODEL CONFIGURATIONS

Configuration	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Latency (ms)
CNN only	84.1	83.5	82.8	83.1	11.2
GRU only	86.7	85.9	86.2	86.0	13.8
CNN-LSTM	89.6	87.8	88.9	88.3	19.3
CNN-GRU (proposed)	91.2	90.1	92.4	91.2	14.5

VIII. CONCLUSION

The CNN-GRU model achieved an accuracy of 91.2%, precision of 90.1%, recall of 92.4%, and F1-score of 91.2% (Table 9). The model is highly accurate yet has low computational complexity and latency, which makes it suitable for wearable AR devices. The findings confirm that lightweight hybrid deep learning networks can capture emotional patterns in real time.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] Philipp V. Rouast, Marc T. P. Adam, and Raymond Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524-543, 2021, doi: 10.1109/TAFFC.2018.2890471.
- [2] Caitlin Mills, Julie Gregg, Robert Bixler, and Sidney K. D'Mello, "Eye-Mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering," *Hum.-Comput. Interact.*, vol. 36, no. 4, pp. 306-332, 2021, doi: 10.1080/07370024.2020.1716762.
- [3] Premananda R., Sunil S. Harakannanavar, Sapna Kumari C., and Veena I. Puranikmath, "Development of real time emotion detection in faces using deep learning approach," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4, pp. 2621-2633, 2024. [Online]. Available: <https://www.ijisae.org/index.php/IJISAE/article/view/6689>
- [4] Byoung Chul Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, Art. no. 401, 2018, doi: 10.3390/s18020401.
- [5] Moatay El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572-587, 2011, doi: 10.1016/j.patcog.2010.09.020.
- [6] Usman Alhaji Abdurrahman, Lirong Zheng, Abdulrauf Garba Sharifai, and Ishola Dada Muraina, "Heart rate and pupil dilation as reliable measures of neuro-cognitive load classification," in *Proc. 2022 Int. Conf. Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, 2022, pp. 1-7, doi: 10.1109/ASSIC55218.2022.10088296.
- [7] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42-55, 2012, doi: 10.1109/T-AFFC.2011.25.
- [8] Khafidurrohman Agustianto, Hendra Yufit Riskiawan, Dwi Putro Sarwo Setyohadi, I. Gede Wiryawan, Andi Besse Firdausiah Mansur, and Ahmad Hoirul Basori, "Eye gaze based model for anxiety detection of engineering students," in *Emerging Technology Trends in Internet of Things and Computing*, ser. Commun. Comput. Inf. Sci., vol. 1548, 2022, pp. 195-205, doi: 10.1007/978-3-030-97255-4_14.
- [9] Pradeep Raj Krishnappa Babu and Uttama Lahiri, "Classification approach for understanding implications of emotions using eye-gaze," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 7, pp. 2701-2713, 2020, doi: 10.1007/s12652-019-01329-8.
- [10] Jia Zheng Lim, James Mountstephens, and Jason Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, Art. no. 2384, 2020, doi: 10.3390/s20082384.
- [11] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 7509-7528, 2024, doi: 10.1109/TPAMI.2024.3393571.
- [12] Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji, "Automatic gaze analysis: A survey of deep learning based approaches," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 61-84, 2024, doi: 10.1109/TPAMI.2023.3321337.
- [13] Jindi Bao, Xiaomei Tao, and Yinghui Zhou, "An emotion recognition method based on eye movement and audiovisual features in MOOC learning environment," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 1, pp. 171-183, 2024, doi: 10.1109/TCSS.2022.3221128.
- [14] Magdiel Jiménez-Guarneros and Gibran Fuentes-Pineda, "CFDA-CSF: A multi-modal domain adaptation method for cross-subject emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1502-1513, 2024, doi: 10.1109/TAFFC.2024.3357656.
- [15] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 1170-1176. [Online]. Available: <https://www.ijcai.org/Proceedings/15/Papers/169.pdf>
- [16] Jing Zhang, Sung Park, Ayoung Cho, and Mincheol Whang, "Recognition of emotion by brain connectivity and eye movement," *Sensors*, vol. 22, no. 18, Art. no. 6736, 2022, doi: 10.3390/s22186736.
- [17] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jitrapol Intarasirisawat, Maxine Glancy, and Chee Siang Ang, "VREED: Virtual reality emotion recognition dataset using eye tracking & physiological measures," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 1-20, 2021, doi: 10.1145/3495002.
- [18] Jain Joseph, R. P. Aneesh, and Joseph Zacharias, "Deep learning based emotion recognition in human-robot interaction with multi-modal data," *AIP Conf. Proc.*, vol. 3122, Art. no. 030007, 2024, doi: 10.1063/5.0216000.
- [19] Vijaya Tulsani, Prashant Sahatiya, Jignasha Parmar, and Sohail Parmar, "Real-time facial emotion recognition for AI-enhanced personalized learning," *J. Inf. Syst. Eng. Manage.*, vol. 10, no. 40s, pp. 427-436, 2025, doi: 10.52783/jisem.v10i40s.7313.