Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37731

Research on Emotion-Aware Multimodal Input Adaptation Strategies in Augmented Reality Environments

Ziteng Feng

University of Birmingham, Birmingham, UK
fengziteng02520@outlook.com

Abstract. This study investigates how emotion-aware multimodal input adaptation can improve interaction stability in augmented reality environments. In AR tasks, users rely on gesture, gaze, speech, and touch-based input, but these modalities become less reliable under stress, confusion, or fatigue. To address this issue, an emotion-aware adaptation framework is developed, integrating multimodal sensing, emotion recognition, and real-time input adjustment. The prototype is implemented using Microsoft HoloLens 2, Unity 2022.3 LTS, MRTK 2.8, Python 3.10, PyTorch 2.1, Flask, and PostgreSQL 15. User data include gaze behaviour, hand gestures, speech features, heart rate, electrodermal activity, skin temperature, and interaction logs. A late-fusion BiLSTM model classifies users into neutral, stress, confusion, and fatigue states. Based on recognition results, the system dynamically adjusts input priority, confirmation rules, visual cues, menu complexity, and task guidance. Experimental results show that recognition accuracy reaches 88.42%, task completion time is reduced by about 14%, input error rate decreases by about 39%, and perceived workload is significantly lowered under adaptive input conditions.

Keywords: augmented reality, multimodal input, emotion recognition, adaptive interaction, human-computer interaction

1. Introduction

Augmented reality has become an important interaction environment for education, industrial training, medical simulation, exhibition guidance, maintenance support, and collaborative design. Unlike conventional screen-based systems, AR interfaces require users to interact with virtual objects through embodied and spatial input, including gesture, gaze, speech, and touch-based selection. Recent studies on AR notification and multimodal interaction show that gaze, hand pointing, speech, and touch can support flexible task execution, but their effectiveness depends heavily on task context, user attention, and the stability of input channels [1]. Gaze, gesture, and speech can also be configured together to support more natural interaction, yet each modality has its own error sources and timing constraints [2]. Gesture and speech elicitation studies further indicate that users often combine modalities differently when object manipulation becomes more complex [3].

A key problem is that most AR systems still use fixed input rules. The system assumes that users remain cognitively stable during the whole task, but real AR interaction is often influenced by stress,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37731

confusion, fatigue, hesitation, and increased cognitive load. Physiological studies of AR cognitive load show that heart activity, electrodermal activity, eye behaviour, and other bodily signals can reflect changes in user state during immersive tasks [4]. When emotion and workload are not considered, users may produce inaccurate gestures, delayed voice commands, excessive gaze wandering, repeated selections, or interrupted task flow.

This paper proposes an emotion-aware multimodal input adaptation framework for AR environments. The study focuses on four user input types: gesture, gaze, speech, and touch. It detects four emotional states: neutral, stress, confusion, and fatigue. Based on real-time recognition results, the system adjusts input priority, feedback strength, confirmation requirements, menu depth, and task guidance. The contribution of the study lies in connecting emotion recognition with concrete AR input adaptation strategies rather than treating affective computing as a separate recognition problem.

2. Literature review

2.1. Multimodal input in augmented reality

Multimodal input allows AR users to combine body movement, visual attention, speech command, and spatial selection. Gesture input is effective for object manipulation because it matches the embodied nature of AR tasks. Gaze input supports fast target localization but may cause accidental selection when the user only looks at an object without intending to activate it. Speech input is suitable for command execution, especially when hands are occupied, but it becomes unreliable in noisy environments or when users hesitate. Touch or indirect confirmation can improve precision, but it may reduce immersion. Existing research shows that multimodal AR interaction is more robust when modalities are coordinated according to user intention, task complexity, and environmental constraints [1-3].

2.2. Emotion recognition in interactive systems

Emotion recognition in interactive systems often depends on behavioural, physiological, acoustic, and eye-tracking signals. VR and AR studies have shown that eye movement and physiological signals can support emotion classification because fixation duration, blink rate, gaze deviation, heart-rate variability, electrodermal activity, and skin temperature are sensitive to arousal, attention, and fatigue [4, 5]. Deep learning and explainable feature-selection methods have also been used to improve emotion recognition in immersive environments, especially when eye-tracking and physiological data are combined [6]. Recent VR-based emotion studies further confirm that physiological signals can classify affective states with practical value for adaptive interfaces [7].

2.3. Adaptive interaction strategies

Adaptive interaction means that the system modifies interface structure, input mode, feedback, and guidance according to user state. In AR environments, adaptation is especially important because immersive interfaces impose perceptual, motor, and cognitive demands at the same time. Adaptive feedback research suggests that physiological monitoring and interface refinement can reduce cognitive load during AR tasks [8]. Inclusive AR and VR research also emphasizes that interaction systems should be flexible enough to accommodate different abilities, attention levels, and situational limitations [9]. Hands-free head-worn interaction studies further show that alternative

DOI: 10.54254/2755-2721/2026.37731

input channels can reduce dependence on unstable modalities when speech, gaze, or manual control becomes unsuitable [10]. Therefore, emotion-aware adaptation should be understood as a practical HCI strategy for maintaining interaction reliability.

3. Method

3.1. AR prototype environment

The AR prototype is implemented on Microsoft HoloLens 2 using Unity 2022.3 LTS and MRTK 2.8. The system operates on a spatial coordinate framework defined by HoloLens world anchors, where each virtual component is registered in a persistent spatial mapping mesh. Object interaction is governed by a state machine with discrete states {idle, hover, selected, manipulated, confirmed}, and transitions are triggered by multimodal input events with temporal constraints (e.g., gesture dwell > 300 ms or gaze fixation > 500 ms). Mesh assets imported from Blender are optimized through polygon decimation (<15k triangles per object) and collider simplification to reduce physics computation overhead. Interaction precision is controlled through raycasting thresholds (angular tolerance < 2.5° for gaze selection) and hand joint tracking confidence (>0.8). Unity scripts implement event-driven logging with a sampling frequency of 60 Hz for spatial data and 10 Hz for interaction events. Temporal synchronization is maintained using a monotonic clock to avoid drift between rendering and sensor pipelines.

3.2. Multimodal sensing configuration

The sensing pipeline integrates heterogeneous data streams with different sampling rates and noise characteristics. Gaze vectors are sampled at 30 Hz and smoothed using a second-order Butterworth low-pass filter (cutoff frequency 5 Hz) to reduce jitter. Hand tracking data are processed using joint position interpolation and velocity thresholding to distinguish intentional gestures from micro-movements. Gesture recognition relies on a rule-based classifier using joint angle constraints and temporal consistency windows (≥200 ms). Speech signals are sampled at 16 kHz and segmented using voice activity detection (VAD) based on energy thresholds and zero-crossing rate. Feature extraction includes 13-dimensional MFCCs with delta and delta-delta coefficients, computed over 25 ms frames with 10 ms overlap. Physiological signals from Empatica E4 are resampled to a unified 4 Hz frequency. HRV is computed using RMSSD over sliding windows of 30 seconds, while EDA signals are decomposed into tonic and phasic components using continuous decomposition analysis (CDA). All modalities are aligned through timestamp interpolation and buffered in a sliding window of 5 seconds with 50% overlap. Missing data are handled using forward filling for short gaps (<1 s) and masked during model input for longer interruptions.

3.3. Emotion recognition model

The model processes multimodal time-series data using a late-fusion BiLSTM architecture designed to preserve modality-specific temporal dependencies. Each modality branch consists of a two-layer BiLSTM with hidden size 128, followed by layer normalization and dropout (p = 0.3). Input sequences are segmented into fixed-length windows (T = 50 time steps), and each branch outputs a hidden representation $h_{m,t}$.

Attention weights are computed using a scaled dot-product mechanism, where query vectors are derived from a learned global context embedding. This allows dynamic weighting of modalities

DOI: 10.54254/2755-2721/2026.37731

based on their reliability at each time step. To address modality imbalance, feature normalization is applied per modality using z-score normalization computed over the training set.

The multimodal fusion process can be formally expressed as shown in Equation (1)-(2):

$$zt = \sum m = 1^{M} \alpha_{m,t} \cdot \tanh\left(W_m h_{m,t} + b_m\right) \tag{1}$$

$$\alpha_{m,t} = \frac{exp\left(q_t^{\top} K_m h_{m,t}/\sqrt{d}\right)}{\sum_{k=1}^{M} exp\left(q_t^{\top} K_k h_{k,t}/\sqrt{d}\right)} \tag{2}$$

where $\mathbf{z}t$ is the fused representation at time step t, $\alpha_{m,t}$ denotes the attention weight for modality m, $q_t^{\top}$ is the query vector, and K_m is the modality-specific projection matrix.

The loss function incorporates class weighting to compensate for skewed emotion distribution (stress and fatigue underrepresented). Temporal smoothness regularization penalizes abrupt changes in attention weights, effectively reducing sensitivity to transient noise spikes in physiological signals. Gradient clipping (threshold = 5.0) is applied during training to stabilize recurrent updates.

3.4. Multimodal input adaptation strategy

The adaptation system is designed as a control layer that dynamically changes interaction parameters in real time based on the emotional state inference. The system does not rely on rule switching but dynamically changes continuous parameters within certain limits. To illustrate, the gesture sensitivity is controlled by changing the threshold of minimum distance Δxmin relative to the stress level, and the gaze dwell time threshold is dynamically adjusted upwards in cases of fatigue in order to reduce false positives.

The decision about the adaptation is made via the hysteresis process in order to avoid oscillations between states. In other words, the transition into a new state occurs only if the probability prediction is higher than 0.70 for two consecutive windows and lower than 0.55 before the return to previous state. This design adds temporal stability to the adaptation behavior.

The intensity of visual feedback is controlled by adjusting shader parameters (for example, emission intensity multiplied by attention weight), and the simplicity of the user interface is increased by reducing unnecessary UI elements on the basis of task relevance. All procedures are carried out during 16 ms frame period.

3.5. System communication architecture

The system follows the design of asynchronous client-server model with non-blocking communications in order to reduce the rendering delay. Unity sends batches of feature vectors in form of JSON payloads using HTTP POST requests with an interval of 2 seconds. Network overhead is reduced by reducing feature vectors to float16 data type before sending them. The backend of the Flask server uses the processing model of queue where every request is assigned a unique session ID for independent processing. Model inference delay is maintained below 50 milliseconds using GPU acceleration in PyTorch with CUDA support. The output data is sent back as light JSONs containing probabilities of emotions and adaptation parameters. In Unity, double buffering is used to ensure that received adaptation commands will not interrupt ongoing interaction

DOI: 10.54254/2755-2721/2026.37731

cycles. The designed system allows the processes of sensing, inference, and adaptation to run simultaneously without blocking the rendering process and maintaining 60 Hz frame rate. Figure 1 shows the general architecture of the system.

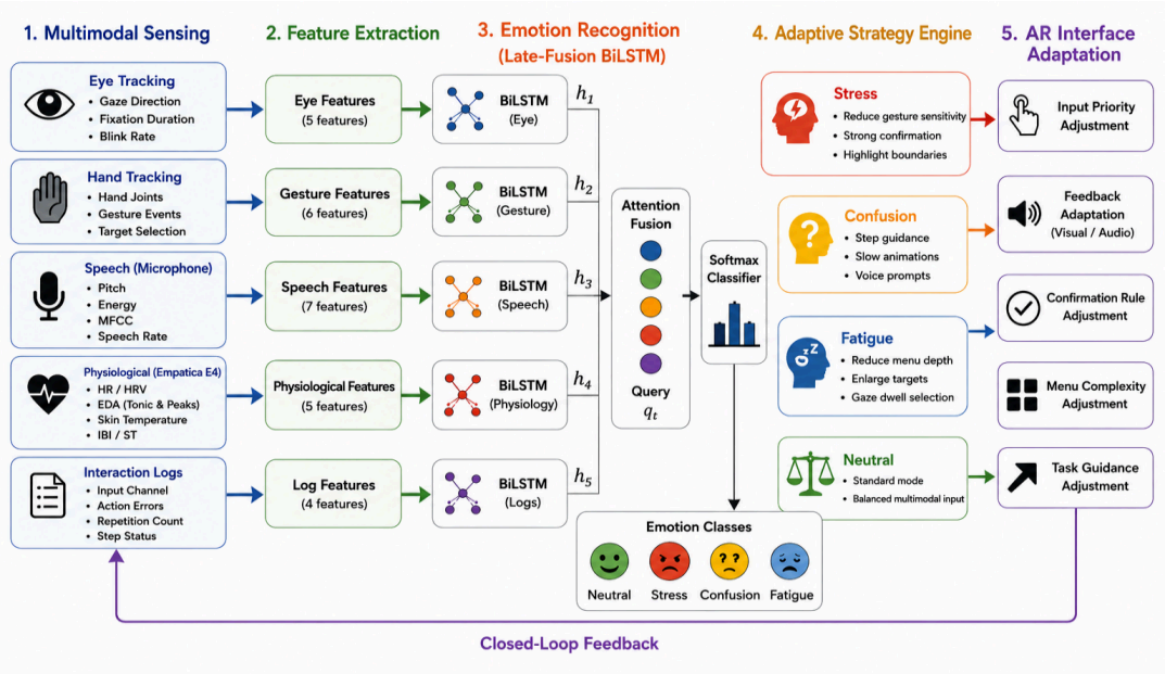


Figure 1. Overall architecture of the emotion-aware multimodal input adaptation framework in AR environments

4. Experimental procedure

4.1. Participant recruitment and task design

The experiment involved 36 university students with normal or corrected eyesight who had some prior experience using digital gadgets. Each student performed three augmented reality (AR) assembly tests. Test 1 consisted of 8 stages, Test 2 of 12 stages, and Test 3 of 16 stages. The participants performed the tests in two different modes: conventional mode of AR input and emotion-based adaptive mode of AR input.

4.2. Data collection and annotation

In every experiment, the HoloLens 2 device is capturing eye-gaze, gesture, voice command, and operational data. At the same time, physiological measurements from the Empatica E4 are being recorded. After each experiment, the participants fill out the Self-Assessment Manikin, the NASA-TLX workload measurement, and an emotion checklist. The labels of the emotions are obtained by combining the questionnaire answers, experiment events, and logs of interactions reviewed by two annotators.

4.3. Evaluation setting

In the evaluation of the emotion recognition model, the metrics used included accuracy, macro-F1, recall, precision, and the confusion matrix. Adaptive interaction impact was evaluated through the

DOI: 10.54254/2755-2721/2026.37731

use of task completion time, input error rate, command repeat rate, success select rate, NASA-TLX rating, and System Usability Scale score. T-tests for paired samples were conducted in SPSS 29. Data from Unity logs and PostgreSQL was imported to Python for data processing.

5. Results and discussion

5.1. Emotion recognition performance across multimodal signals

BiLSTM model with late fusion had better results compared to the unimodal models, achieving 88.42% accuracy and macro-F1 of 0.873. The performance of each modality separately was lower than the performance of multimodal approach, which proves the effectiveness of using multimodality. It should be mentioned that the biggest improvements were shown for stress and fatigue recognition when combining physiology and behavior features. Still, there was class confusion due to similar hesitation patterns (Table 1).

Table 1. Emotion recognition performance across modality settings

Model setting	Accuracy Mean ± SD	Macro-F1 Mean ± SD	Stress recall	Confusion recall	Fatigue recall
Eye features only	73.18 ± 3.91	0.711 ± 0.039	0.742	0.681	0.704
Speech features only	69.44 ± 4.25	0.684 ± 0.046	0.697	0.662	0.653
Physiological features only	79.63 ± 3.14	0.781 ± 0.032	0.826	0.713	0.804
Early fusion BiLSTM	84.06 ± 2.87	0.831 ± 0.030	0.861	0.792	0.846
Late-fusion BiLSTM	88.42 ± 2.36	0.873 ± 0.028	0.903	0.821	0.887

5.2. Effects of adaptive input strategies on AR task performance

Adaptive input significantly improved performance. Task completion time, error rate, and command repetition decreased, while selection accuracy increased. NASA-TLX scores also dropped, indicating reduced workload. These results confirm that emotion-aware adaptation enhances efficiency and interaction reliability (see Table 2).

Table 2. Comparison between standard AR input and emotion-aware adaptive input

Indicator	Standard AR input Mean ± SD	Adaptive AR input Mean ± SD	Test statistic	Significance
Task completion time	412.37 ± 66.51 s	354.82 ± 58.74 s	t = 6.18	p < .001***
Input error rate	14.63 ± 4.28%	8.91 ± 3.17%	t = 7.04	p < .001***
Command repetition rate	2.84 ± 0.91	1.63 ± 0.72	t = 6.72	p < .001***
Successful selection rate	86.52 ± 5.39%	93.47 ± 3.88%	t = -6.31	p < .001***

DOI: 10.54254/2755-2721/2026.37731

Table 2. (continued)

NASA-TLX score	58.76 ± 9.84	43.29 ± 8.67	t = 7.46	p < .001***
SUS score	71.38 ± 7.52	82.64 ± 6.93	t = -6.85	p < .001***

5.3. Discussion of usability improvement and interaction stability

Emotion-adaptive techniques improved the stability of interactions by adapting the input methods based on the emotions of the user. In this regard, stress resulted in the use of rigorous confirmation processes, while confusion led to the provision of guidance and fatigue resulted in simplified interactions. These results demonstrate that incorporating emotion detection in adaptive input design reduces cognitive load in AR interfaces.

6. Conclusion

This study developed and evaluated an emotion-aware multimodal input adaptation framework for AR environments. By combining gaze, gesture, speech, physiological signals, and interaction logs, the system classified neutral, stress, confusion, and fatigue states and adjusted input strategies in near real time. Experimental results showed improved recognition performance, reduced input errors, shorter task completion time, lower workload, and higher usability. The study demonstrates that emotion-aware adaptation can make AR interaction more stable, flexible, and user-sensitive.

References

[1] Plabst, L., Niebling, F., Oberdörfer, S., & Ortega, F. (2025). Order up! Multimodal interaction techniques for notifications in augmented reality. *IEEE Transactions on Visualization and Computer Graphics, 31*(5), 2258–2267.
[2] Wang, Z., Wang, H., Yu, H., & Lu, F. (2021). Interaction with gaze, gesture and speech in a flexibly configurable augmented reality system. *IEEE Transactions on Human-Machine Systems, 51*(5), 524–534.
[3] Zhou, X., Williams, A. S., & Ortega, F. R. (2022). Eliciting multimodal gesture+speech interactions in a multi-object augmented reality environment. *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*.
[4] Suzuki, Y., Wild, F., & Scanlon, E. (2024). Measuring cognitive load in augmented reality with physiological methods: A systematic review. *Journal of Computer Assisted Learning, 40*(2), 375–393.
[5] Tabbaa, L., Searle, R., Bafti, S. M., Hossain, M. M., Intarasisrisawat, J., Glancy, M., & Ang, C. S. (2021). VREED: Virtual reality emotion recognition dataset using eye tracking and physiological measures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5*(4), Article 178.
[6] Alharbi, H. (2024). Explainable feature selection and deep learning based emotion recognition in virtual reality using eye tracker and physiological data. *Frontiers in Medicine, 11*, Article 1438720.
[7] Polo, E. M., Iacomi, F., Valdes Rey, A., Ferraris, D., Paglialonga, A., & Barbieri, R. (2025). Advancing emotion recognition with virtual reality: A multimodal approach using physiological signals and machine learning. *Computers in Biology and Medicine, 193*, Article 110310.
[8] Sun, J., & Liao, T. (2025). Designing adaptive feedback systems for managing cognitive load in augmented reality. *Design Science, 11*, Article e49.
[9] Creed, C., Al-Kalbani, M., Theil, A., Sarcar, S., & Williams, I. (2024). Inclusive augmented and virtual reality: A research agenda. *International Journal of Human-Computer Interaction, 40*(20), 6200–6219.
[10] Gemicioglu, T., Winters, R. M., Wang, Y.-T., Gable, T. M., Paradiso, A., & Tashev, I. J. (2023). Gaze & tongue: A subtle, hands-free interaction for head-worn devices. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*.