"Intuition-Ergonomics-Safety": A Triple-Constraint Model for Multimodal Interaction Design in AR-Based Substation Inspection

Yiting Zhang*
School of Creative Design, Dongguan City University, Dongguan, Guangdong, China
**Corresponding Author*

Abstract: Augmented Reality (AR) inspection of substations lacks specific theoretical guidance for natural interaction design. To address this, this paper introduces a safety dimension into the existing "intuition-ergonomics" dual-constraint model, constructing a "intuition-ergonomics-safety" three-constraint design model for multimodal interaction. We first identify the unique constraints of the substation scenario on interaction—safety distance, operating space, attention load, and environmental interference—and provide their quantitative basis. Then, we construct a three-constraint overlapping model, proposing a multimodal collaborative mechanism of "low-energy intention indication + high-confidence action confirmation" and corresponding design principles. This model provides a directly applicable theoretical framework for AR interaction design in high-risk scenarios.

Keywords: Augmented Reality; Substation Inspection; Multimodal Interaction; Three-Constraint Model; Intuition-Ergonomics-Safety

1. Introduction

Substations are key nodes in the power system, and their inspection quality is directly related to the overall safety of the power grid. Traditional inspections rely on paper manuals or tablets, which brings two prominent problems: one is "cognitive break" - the inspectors need to switch their attention repeatedly between equipment, manuals and record sheets; the other is that their hands are completely occupied, making it impossible to operate tools and look up information at the same time. AR technology can make up for these defects: it can overlay digital information such as equipment parameters and operation instructions onto the physical field of vision in real time, and realize contextualized information acquisition [1]. However, after introducing AR technology into substation inspections, the existing interaction methods have exposed obvious limitations. In terms of voice interaction, the noise in the main transformer area of the substation is serious. The measured data shows that 22.2% of the measurement points have noise values ≥80 dB (A) [2], which exceeds the effective working range of most consumer-grade voice recognition systems. In terms of gesture interaction, high-precision data gloves are expensive, and vision-based gesture recognition is easily affected by complex backgrounds and fast movements [3]. Although eye-tracking interaction is fast, it has always faced the classic "Middles touch" problem - the system has difficulty distinguishing between the user's focused gaze and random eye movements [4]. It is evident that a single mode can hardly simultaneously meet the requirements of intuition, ergonomics, and safety in a high-risk, high-interference environment like a substation.

Multimodal interaction provides a new approach for this. By integrating different modalities such as gestures, eye movements, speech, and touch, the complementary advantages of each modality can be used to make up for the defects of a single modality [5]. For example, the head-eye coordination model successfully combines rapid eye pointing with head stabilization confirmation, effectively solving the problem of accidental Midas touch [6]. More directly related, the "intuition-ergonomics" dual-constraint model proposed by Zhang Fan et al. [7] has achieved good results in the inspection of pipe corridors, but it does not consider the safety distance and high voltage risk unique to substations. In other words, substations have rigid requirements for interactive safety, which needs to be included in the design framework as an independent constraint dimension. To this end, this paper constructs the "intuition-ergonomics-safety"

three-constraint model and extends it into a design framework applicable to multimodal interaction.

2. Theoretical Basis and Related Work

2.1 Reality-Based Interaction Theory

The Reality-Based Interaction (RBI) framework proposed by Jacob et al. [8] has a core proposition: interaction design should make full use of users' existing perceptions and motor skills of the physical world, thereby reducing the learning cost. This proposition provides the underlying logic for "intuitive constraints"—gestures should simulate physical operations (such as tapping and swiping), and voice should simulate natural dialogue. In other words, the intuition of the interaction method essentially comes from its faithful mapping of users' daily experience.

2.2 From Dual-Constraint to Triple-Constraint

Zhang Fan et al. [7] proposed an "intuition-efficiency" dual-constraint model for industrial pipe gallery inspection. Among them, the intuition constraint requires that the gestures conform to the user's instinctive mental model. Therefore, they extracted a set of highly consensus gestures through the "Wizard of Oz" experiment. The efficiency constraint advocates the use of "micro-gesture" strategy to avoid the "gorilla arm" effect [9]. This model is effective in general industrial scenarios, but if it is directly transplanted to high-risk environments such as substations (high voltage electric shock risk, narrow operating space, high cognitive load), a key gap will be exposed - it does not consider safety factors such as safe distance and prevention of accidental contact. Safety is not an appendage of intuition or efficiency, but a rigid constraint that needs to be independently evaluated and designed. Therefore, this paper formally introduces the "safety" constraint on the basis of the "intuition-efficiency" dual-constraint model and expands it into a three-constraint model.

2.3 Safety Ergonomics

The principles of minimum risk, fault tolerance, and accessibility safety in safety ergonomics [10] provide a theoretical basis for “safety constraints”: interaction should limit the range of motion, key operations should prevent accidental touches, and feedback should be clear and reliable.

2.4 Multimodal Interaction and Complementary Fusion

Multimodal interaction can integrate two or more input/output modalities and use complementarity to make up for the defects of a single modality [5]. Among them, the head-eye coordination model [6] is a typical example: rapid eye pointing (intent layer) + stable head confirmation (execution layer), which solves the problem of Midas touch while maintaining speed. This mechanism also provides a reference for the subsequent paper: multimodal fusion is an effective path to reconcile the conflict of the three constraints - assigning different modalities to the two levels of intent indication and action execution can take into account both efficiency and accuracy.

3. Substation Scenario Constraint Analysis and Triple-Constraint Model Construction

3.1 Unique Constraints of the Substation Scenario

The substation environment imposes four unique constraints on multimodal interaction in AR inspections. Table 1 provides the quantitative basis and design requirements for these constraints. According to GB 26860-2011 "Electric Power Safety Work Regulations," the required safety distances for equipment at different voltage levels (0.7m for 10kV, 1.0m for 35kV, and 1.5m for 110kV) dictate that the extremities of all limb movements must be limited to within 0.5m of the body's center—modals without limb movement, such as eye movements and voice commands, naturally meet this requirement, while gestures require deliberate convergence. Regarding operating space, the width of high-voltage passageways in substations is typically only 0.8~1.2m, with narrow equipment gaps. Large lateral arm swings or turns easily result in contact with surrounding equipment; therefore, micro-gestures, eye pointing, or voice commands should be prioritized. Noise interference is also a significant concern: noise levels in the main transformer area often exceed 80 dB (A) (GB/T 15190), far exceeding the effective range of consumer-grade voice recognition systems. Simultaneously, outdoor light intensity can reach 100,000 lx while indoor light intensity is less

than 500 lx; these drastic changes in lighting severely impact the visibility of visual feedback. This means that voice can only serve as an auxiliary modality, requiring high-contrast design for visual feedback and the introduction of haptic feedback as a redundant channel. Finally, substation inspection is a high-risk operation, demanding high concentration from operators, which can increase the error rate by 30%–50% (referencing NASA-TLX research). This necessitates that the interaction design prioritize low cognitive load, provides immediate and clear feedback, and employs multimodal redundant confirmation for critical operations. These four constraints collectively constitute the boundary conditions for the multimodal interaction design of substation AR inspection, forming the practical basis for subsequent model construction and principle refinement.

Table 1. Analysis of Unique Constraints of the Substation Scenario on Interaction

Constraint type	Quantitative reference	Requirements for multimodal design
Safety distance	GB 26860-2011:10kV→0.7m,35kV→1.0m,110kV→1.5m	Limb motion end ≤0.5 m from body center; prefer voice/eye tracking
Operational space	High-voltage room corridor width 0.8-1.2 m	Avoid large- amplitude arm swings; micro- gestures or eye tracking
Noise interference	Main transformer area noise >80 dB(A) (GB/T 15190)	Voice only as auxiliary; gesture/eye tracking as primary
Attentional load	Increased error rate in high- risk situations (NASA- TLX studies)	Low cognitive load; immediate feedback; multimodal redundancy
Illumination variation	Outdoor 100 klx, indoor <500 lx	Visual feedback high contrast; haptic redundancy

3.2 Triple-Constraint Design Model

The three constraints of "intuition-ergonomics-safety" constitute a dynamic trade-off design space. As shown in Figure 1, the three circles overlap in pairs to form three types of strategy areas: Intuition ∩ Ergonomics → micro-gestures (such as pinching and sliding), suitable for high-frequency operations; Intuition ∩ Safety → chest tap, suitable for critical operations; Ergonomics ∩ Safety → eye movement/voice, suitable for long-term work. The central triple-overlapping area is the ideal modality, such as "eye-hand coordination".

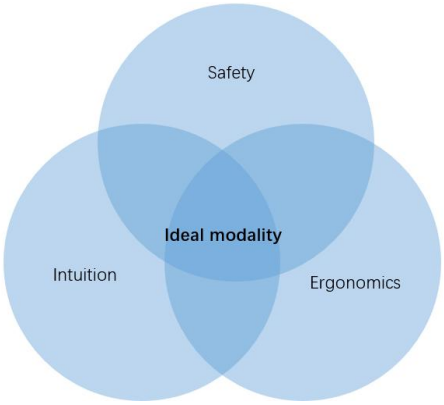

Figure 1. Schematic Diagram of the Triple-Constraint Design Model

In an ideal environment, the three constraints are highly unified, but substations fall under the negative gradient scenario in ergonomics (confined space, high pressure, strong interference), where the three are easily mutually exclusive. Table 2 summarizes three typical conflicts and their multimodal resolution paths. The core solution strategy is to decompose a single mode into "low-energy intention indication" (such as eye-tracking) and "high-confidence action execution" (such as gesture confirmation), thereby approximating the balance of the three constraints in the negative gradient scenario.

Table 2. Three Typical Conflicts and Multimodal Resolution Approaches

Conflict	Manifestation	Multimodal resolution approach
Intuition vs. Safety	Instinctive large-arm-swipe to call menu vs. touching equipment	Replace large- arm- swipe with eye tracking + micro- gesture
Intuition vs. Ergonomics	Large- arm- swipe to scroll pages vs. high- frequency fatigue	Voice input or micro- gesture correction
Ergonomics vs. Safety	Hand naturally down (energy- saving but unrecognizable) vs. need to raise to chest	Eye tracking/voice as limb- free input

3.3 Design Principles

3.3.1 Safety-first principle

Interaction design should prioritize operational safety. Specifically, the extremities of all body movements must be limited to within 0.5m of the body's center to meet the safety distance requirements for energized equipment in substations. Critical operations (such as

emergency shutdowns) must employ at least two modalities for confirmation (e.g., eye-tracking combined with hand gestures, or voice commands combined with hand gestures) to effectively prevent accidental triggering. Furthermore, when the system fails to recognize an instruction or the instruction becomes ambiguous, it should default to a safe state (e.g., no response rather than any action) to avoid safety risks caused by recognition errors.

3.3.2 Intuition-preserving principle

Interaction methods should closely align with users' instinctive mental models and existing experiences. Prioritize interaction methods with high consistency in user research to reduce learning costs. For example, gestures should simulate physical world actions (such as pressing buttons or swiping to turn pages), while voice commands should mimic natural conversation. For operations with significant functional differences, use interaction methods with distinct morphological or modal differences (such as zooming using diagonal dragging with both index fingers, or rotating using circular motions with both index fingers) to prevent semantic confusion and misoperations.

3.3.3 Ergonomics-optimisation principle

To accommodate long-duration inspection work, the interaction design should prioritize minimizing physiological burden. High-frequency operations should employ micro-gestures using finger joints as fulcrums, avoiding large-scale shoulder and elbow movements. Low-frequency operations can tolerate larger ranges of motion, but high-frequency operations must utilize low-energy methods such as micro-gestures, voice commands, or eye movements. Simultaneously, at least 80% of daily operations should be operable with one hand, allowing the other hand to hold the inspection instrument or grip a handrail.

3.3.4 Multimodal collaboration principle

Multimodal interaction should follow a complementary fusion mechanism of "low-energy intent indication + high-confidence action confirmation". The intent layer uses low-energy modalities such as eye movement, head movement, or hovering to quickly point to a target or retrieve a preview. The system only provides intermediate feedback (such as visual highlighting) and does not execute any final command, thus completely avoiding the "midasu touch" problem. The execution layer uses high-confidence modalities such as gesture confirmation, voice commands, or physical buttons, which are consciously triggered by the user and naturally have anti-accidental touch characteristics. The interaction flow strictly follows the temporal relationship of "intent layer → system intermediate feedback → execution layer → final command". Typical collaborative strategies include: eye-hand collaboration combining eye-tracking and gesture confirmation, voice - gesture collaboration combining voice input parameters and gesture confirmation, tactile redundancy using tactile feedback as a supplementary visual or auditory channel, and modal redundancy where emergency operations require confirmation from at least two independent modalities.

3.4 Model Application Workflow

When applying this model for multimodal interaction design, task analysis should be conducted first to clarify the core interaction functions and their frequency and criticality in substation inspection scenarios. Based on this, preliminary interaction schemes (such as gesture sets or modal assignments) should be developed according to general design experience or user survey results. Subsequently, the schemes should be filtered for three constraints, checking whether they meet the three basic requirements of safety, intuitiveness, and ergonomics, and identifying any links that conflict with a particular constraint. For the identified conflicts, the conflict identification and multimodal collaborative solution design phase should begin, reconciling the contradictions between constraints by introducing complementary fusion mechanisms (such as breaking down a single gesture into eye-tracking intention indication and gesture confirmation execution). Finally, the optimized schemes for each task should be integrated to form the final interaction set, completing the entire design process from requirements to feasible solutions.

4. Application Examples of Multimodal Interaction Design

4.1 Example 1: Selection/Confirmation (Gesture)

As the most basic operation, selection and confirmation can be completed using a small gesture of tapping forward with the index finger. The advantages of this gesture are: the amplitude is small, which conforms to the metaphor of

pressing a button in the physical world, and it meets the constraints in terms of intuition, ergonomics, and safety without the need for additional modifications.

4.2 Example 2: Scroll Browsing (Micro-Gesture Correction)

Users instinctively use their palm to swipe up and down to scroll, but this action has two problems: the amplitude is too large, making it easy to touch surrounding devices in confined spaces (violating safety constraints); and high-frequency operation can easily lead to upper limb fatigue (violating ergonomic constraints). Therefore, we modified it to a pinching motion with the thumb and forefinger, sliding up and down, shifting the fulcrum of movement from the shoulder and elbow joints to the interphalangeal joints. The modified micro-gesture satisfies both safety and ergonomic requirements without weakening the "swiping" metaphor.

4.3 Example 3: Rapid Target Localisation (Eye-Hand Coordination, Core Example)

Selecting a target from multiple device icons in the AR field of view is a high-frequency operation in inspection. If only gestures are used, the user needs to repeatedly raise their finger to point at different icons, resulting in a long operation path and low efficiency. This solution follows the principle of complementary fusion and is designed as a combination of eye movement and gestures: First, the user looks at the target icon (intention layer), and the system identifies and highlights the icon through eye movement tracking, providing only intermediate feedback without executing any instructions, thus avoiding accidental touches; then, the user makes a micro-gesture of pressing forward with their index finger (execution layer), and the system confirms and jumps to the details page. This solution performs well in the three-constraint analysis: eye movement pointing conforms to the intuition of "what you see is what you get", and gesture pressing continues the physical metaphor; eye movement does not require limb movement, and the gesture amplitude is very small, significantly reducing fatigue; the dual confirmation mechanism of eye movement only generating a preview and gesture execution triggering an instruction effectively prevents misoperation. This solution draws on the head-eye coordination model of Yi Xinwu et al. [6], but replaces head confirmation with gesture confirmation (micro-gesture), which is more suitable for the requirements of fine operation in industrial scenarios.

4.4 Example 4: Complex Command Input (Voice + Gesture)

For complex tasks such as retrieving detailed historical data of equipment or setting parameters, a collaborative approach can be adopted, primarily using voice input and secondarily using gestures. For example, the user can verbally state "Display the historical temperature data of transformer #3" while simultaneously gesturing to the target device for location assistance. This approach demonstrates a balanced performance in the three-constraint analysis: voice commands conform to natural conversational habits (intuition), require no physical movement throughout the process (efficiency), and voice operation does not require proximity to live equipment (safety).

4.5 Example 5: Emergency Stop (Modality Redundancy)

Emergency stop is the operation with the highest safety requirements. This solution simultaneously employs a "crossed hands" hand gesture, a voice "stop" command, and tactile vibration confirmation via wristbands/gloves, forming a triple modal redundancy. The reason for this is that even if one modality fails due to environmental interference or recognition failure, other modalities can still independently trigger an emergency stop, ensuring that the principle of safety first is rigidly guaranteed.

4.6 Example 6: Status Feedback (Haptic)

For feedback tasks such as operation completion confirmation or abnormal warning, a tactile-based and visual-assisted approach can be adopted. Different vibration patterns are generated through wristbands or gloves (e.g., a short vibration indicates successful operation, and continuous fast vibration indicates abnormal warning). The advantages of this approach are: tactile feedback does not occupy visual attention (efficiency), instant vibration can quickly convey warning information (safety), and the perceptual association between vibration and physical touch conforms to the user's intuitive expectations (intuition). Compared with the light feedback interface used by Chang Yu et al. [11] in agricultural robots, this approach replaces the

visual channel with a tactile channel to better adapt to the strong light interference environment of substations. The above multimodal interaction schemes are summarized in Table 3.

Table 3. Summary of Multimodal Interaction Schemes

Task	Primary modality	Auxiliary modality	Special design
Selection/Confirmation	Gesture	/	Micro- gesture
Scroll browsing	Gesture	/	Micro- gesture correction
Rapid target localisation	Eye tracking + Gesture	/	Intention + execution layers
Complex command input	Voice	Gesture (optional)	Voice + gesture
Emergency stop	Gesture + Voice	Haptic	Modality redundancy
Status feedback	Haptic	Visual	Haptic redundancy

5. Conclusion and Future Work

This paper addresses the lack of specific theoretical guidance for natural interaction design in AR inspection scenarios of substations. Building upon the existing "intuition-ergonomics" dual-constraint model, it introduces safety constraints to construct a "intuition-ergonomics-safety" three-constraint multimodal interaction design model. Through systematic analysis and quantitative analysis of unique constraints in substation scenarios, such as safety distance, operating space, attentional load, and environmental interference, a mechanism for identifying and reconciling conflicts among the three constraints in "human-computer interaction negative gradient scenarios" is proposed. Design principles encompassing safety priority, intuition preservation, ergonomic optimization, and multimodal collaboration are extracted, with the core being a complementary fusion principle of "low-energy intention indication + high-confidence action confirmation." This model not only provides a directly applicable theoretical framework for multimodal interaction design in AR inspection of substations but can also be extended to other high-risk industrial scenarios such as chemical plants, mines, and nuclear power plants. However, this research also has limitations. Based on theoretical derivation and case demonstrations, this model has not yet undergone empirical verification; specific parameters in multimodal collaboration (such as eye-tracking gaze time threshold and gesture confirmation time window) have not been quantified. In the future, we can refer to the eye-tracking experiment method to compare the effectiveness differences between single-modal and multi-modal schemes, and further verify the effectiveness of the model by combining cognitive load assessment indicators.

This work was supported by the Dongguan City University Young Teacher Development Fund Project (Project No. 2025QJ003R), entitled “Research on Interactive Systems for Human-Machine Integrated Intelligent Data Collection”.

References

[1] Li D Y, Yang C, Zhang Y W, et al. Design and Implementation of Intelligent Substation Inspection System Based on AR Technology. Microcomputer Applications, 2020, 36(8): 92-94.

[2] Li H L, Li L, Zhang K, et al. Analysis of Workplace and Individual Noise Levels in 500 kV Substation. Chinese Journal of Industrial Medicine, 2015, 28(5): 376-377.

[3] Cui C, Sunar M S, Su G E. Deep vision-based real-time hand gesture recognition: a review. PeerJ Computer Science, 2025, 11: e2921.

[4] Yang X N, Wang S, Niu H W, et al. Key Technologies of Eye-Tracking Interaction: State of the Art and Prospects. Computer Integrated Manufacturing Systems, 2024, 30(5): 1595-1609.

[5] Tao J H, Wu Y C, Yu C, et al. A Survey on Multimodal Human-Computer Interaction. Journal of Image and Graphics, 2022, 27(6): 1956-1987.

[6] Yi X W, Xue J Y, You Z, et al. A Study on Multimodal Interaction Model for Immersive Virtual Reality. Journal of Jiangxi Normal University (Natural Science Edition), 2024, 48(1): 52-58.

[7] Zhang F, Huang X F, Lou H Z, et al. Designing and Evaluating Gesture Interaction for AR-Based Industrial Inspections. Packaging Engineering, 2026, 47(4): 39-52.

[8] Jacob R J K, Girouard A, Hirshfield L M, et al. Reality-based interaction: a framework for post-WIMP interfaces//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). New York: ACM, 2008: 201-210.

[9] Hincapié-Ramos J D, Guo X, Moghadasian P, et al. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). New York: ACM, 2014: 1063-1072.

[10] Ding Y L. Ergonomics (5th ed.). Beijing: Beijing Institute of Technology Press, 2017.

[11] Chang Y, Li Y. Design and Experiment of Light Interaction Interface for Agricultural Robot End-Effectors. Journal of Agricultural Mechanization Research, 2026, 48(8): 188-198.

