Differences in the Ability of Humans and Large Models to Recognize Emotions in Composite Faces

Authors: Li Jingting, Zhao Lin, Liu Ye, Su-Jing Wang, Junchi Ma, Li Jingting

Date: 2026-04-13T15:29:07+00:00

Abstract

Faces are important carriers for conveying social information such as emotions. Humans rely on multi-level cognitive mechanisms, including holistic and local processing, to efficiently and accurately recognize basic facial emotions. Although Multimodal Large Language Models (MLLMs) integrate visual encoding components and linguistic reasoning mechanisms, their processing strategies differ significantly in principle from human perceptual processing. Comparing the emotion recognition capabilities of the two helps to understand the differences between them in terms of emotion perception and reasoning strategies. Meanwhile, existing research suggests that text prompts significantly influence the output of MLLMs, but their role in the context of facial emotion recognition [**?**] still lacks systematic examination.

Based on the above reasons, this paper aims to explore the holistic and local feature processing advantages in facial emotion recognition, and further investigate whether this processing pattern is consistent between humans and "virtual participants" generated by MLLMs. The study consists of four experiments, and the results indicate that: MLLMs exhibit a preference for local features that distinguishes them from humans when recognizing composite emotional faces, characterized by low coordination ratios and a tendency to judge images as mutually exclusive; the level of detail in prompts and exemplar images significantly alters the model' s judgment tendencies and coordination ratios. In summary, the research results deepen the understanding of the differences between humans and artificial intelligence in emotion comprehension pathways, and provide new theoretical references for the application of artificial intelligence in the fields of emotion recognition and human-computer interaction.

Full Text

Ability Differences Between Humans and Large Language Models in Recognizing Emotions from Composite Faces

Abstract

The recognition of facial expressions is a fundamental component of social interaction and emotional intelligence. While humans have evolved sophisticated mechanisms for holistic face processing, the rapid advancement of Multimodal Large Language Models (MLLMs) has introduced new questions regarding their ability to interpret complex visual stimuli. This study investigates the performance differences between humans and state-of-the-art MLLMs in identifying emotions from composite faces—images where the upper and lower halves represent different emotional states. Our findings suggest that both humans and MLLMs exhibit specific limitations when processing incongruent emotional information, though the underlying mechanisms of these errors differ significantly between biological and artificial systems. Across several experiments, the results reveal that MLLMs exhibit a distinct preference for local features compared to humans, characterized by low coordination ratios and a tendency to judge images as mutually exclusive. Additionally, the level of detail in prompts and the inclusion of example images significantly alter the models' judgment biases. These findings deepen our understanding of the divergent emotion comprehension pathways between humans and artificial intelligence, offering new theoretical references for AI applications in emotion recognition and human-computer interaction.

1. Introduction

1.1 Human Emotion Recognition Ability Emotion is generally defined as a complex psychological state composed of physiological arousal, subjective experience, and behavioral expression that occurs when an individual is stimulated by external factors [**?**, **?**, **?**]. Facial expressions are typically regarded as vital carriers of social information during human interaction. Numerous studies have demonstrated that humans can accurately recognize basic emotions within a short timeframe; even when an emotional face is presented for only a few hundred milliseconds, observers can rapidly extract useful emotional cues and make categorical judgments [**?**, **?**, **?**].

Electroencephalogram (EEG) studies reveal that emotional faces trigger neural responses related to structural encoding—such as the $N170$ component—suggesting that recognition involves rapid perceptual processing rather than just time-consuming high-level reasoning [**?**, **?**, **?**]. Human emotion recognition depends heavily on holistic processing, where observers form perceptions by integrating the spatial relationships between facial parts [**?**, **?**, **?**]. However, the accuracy and speed of recognition decrease significantly when faces are inverted, segmented, or combined [**?**, **?**, **?**]. Composite emotional faces present information

in a manner that violates natural physiological structures, weakening holistic processing and forcing a reliance on local features.

1.2 Differences Between LLMs and Humans in Emotion Recognition With the development of Large Language Models (LLMs), a significant question has emerged: whether these models possess facial expression recognition capabilities equivalent to humans. Multimodal Large Language Models (MLLMs) can process both text and vision, performing analysis and reasoning based on visual information [**?**, **?**]. While MLLMs demonstrate powerful capabilities in general visual understanding, they often perform poorly in scenarios requiring fine-grained features, such as micro-expressions [**?**, **?**].

Using Marr' s Tri-Level Hypothesis, we can analyze these systems at the representation and algorithm level [**?**, **?**]. Humans integrate visual cues with prior experience, allowing for adaptable emotional reasoning. In contrast, MLLMs rely on associative statistics and pattern-matching from training data. They often fail when stimuli involve ambiguity or conflict, as they tend to select the "most likely label" rather than performing a deep holistic analysis [**?**, **?**].

1.3 Impact of Prompts on MLLM Performance Prompts significantly influence MLLM outputs in linguistic reasoning and retrieval tasks [**?**, **?**]. Structured prompts incorporating facial landmarks or Action Units (AUs) can improve recognition accuracy [**?**, **?**]. Techniques like Chain of Thought (CoT) can provide more detailed reasoning, though their effectiveness in subjective tasks remains debated [**?**, **?**]. This study introduces prompts as a key variable to investigate how textual and visual instructions impact MLLM performance in composite image recognition.

2. Methodology

This study comprises three experiments: Experiment 1 establishes the baseline difference between humans and MLLMs; Experiment 2 tests the stability of these results across different facial identities; and Experiment 3 investigates the influence of visual cues by removing facial division lines.

2.1 Participants **Human Participants:** 100 undergraduate students were recruited ($M_{age} = 22.23$, $SD = 2.40$). All had normal or corrected-to-normal vision and no history of psychiatric disorders. **Virtual Participants:** Generated by MLLMs including Qwen2.5-Instruct, Gemma-2, and GPT-4o. To simulate inter-individual differences, the *temperature* was set to 1.0 with random sampling enabled. Each complete inference cycle of 120 images was treated as one "virtual participant."

2.2 Stimuli Generation Stimuli were sourced from Ekman' s Facial Action Coding System (FACS) database. Composite faces were created by splicing the upper facial region (eyes/forehead) and lower facial region (mouth/jaw). -

Congruent: Both halves represent the same emotion (e.g., Happy Top + Happy Bottom). - **Incongruent:** Halves represent different emotions (e.g., Sad Top + Happy Bottom).

Let E_{top} represent the emotion of the upper face and E_{bottom} represent the emotion of the lower face. A composite face C is defined as:

$$C = \Phi(E_{top}, E_{bottom})$$

where Φ is the spatial alignment function.

[Figure 1: see original paper]

2.3 Experimental Procedure Participants performed five tasks for each image: 1. **Coordination Judgment:** Is the expression "coordinated" (possible on one face) or "mutually exclusive" ? 2. **Coordination Level:** A 5-point scale of intensity. 3. **Emotional Polarity:** Positive or negative. 4. **Emotion Category:** Choosing from the six basic emotions (Anger, Disgust, Fear, Happiness, Sadness, Surprise). 5. **Judgment Basis:** Which region (upper or lower face) was the primary basis for the decision?

3. Results and Analysis

3.1 Coordination Ratio The Coordination Ratio is defined as:

$$\text{Coordination Ratio} = \frac{\text{Number of coordinated choices}}{\text{Total number of images}}$$

ANOVA results revealed a significant main effect for image type ($F(1, \dots) = 876.24, p < .001, \eta_p^2 = .85$) and participant type ($F(3, \dots) = 7.72, p < .001, \eta_p^2 = .84$). Humans showed significantly higher coordination ratios than all MLLMs. For all groups, congruent images had higher coordination ratios than incongruent ones.

3.2 Emotion Confusion and Judgment Basis Confusion matrices showed that humans and MLLMs share some patterns, such as misidentifying disgust as happiness and fear as surprise. However, MLLM errors were highly concentrated (e.g., Gemma misidentifying 100% of disgust as anger), while human errors were more dispersed.

Regarding the judgment basis: - **Happiness:** Primarily identified via the lower face across all groups. - **Surprise:** Primarily identified via the upper face. - **Other emotions:** Criteria varied significantly between models and humans ($F(15, \dots) = 23.40, p < .001, \eta_p^2 = .31$).

3.3 Attention Heatmap Analysis Visualization of attention weights for Qwen2.5 and Gemma revealed stable strategy differences. Qwen2.5 showed a clear upper-face bias (forehead and eyes), while Gemma exhibited a more

distributed structure with enhanced attention toward the mouth region during emotion categorization.

[Figure 1: see original paper] (Attention Heatmaps)

4. Discussion

4.1 Processing Strategy Disparities The results indicate that humans employ a holistic processing strategy, integrating conflicting cues into a rationalized whole based on social experience. MLLMs, however, rely on local feature matching. This is evidenced by the fact that removing the facial dividing line (Experiment 3) improved MLLM coordination ratios but did not close the performance gap with humans. The "concentration" of MLLM errors suggests they form rigid correlations between specific local features (e.g., wide-open eyes) and emotion labels (e.g., surprise).

4.2 Impact of Prompts Experiment 3 demonstrated that "simple text + example images" yielded the best performance for MLLMs. Example images provide concrete references that reduce task ambiguity through few-shot learning. Conversely, overly complex textual instructions (e.g., "ignore the dividing line") increased cognitive load and interfered with judgment, as the models focused more on the features they were told to ignore.

5. Conclusion

This study demonstrates that while MLLMs possess foundational emotion recognition capabilities, they lack the holistic integration and experiential reasoning of humans. MLLMs are sensitive to prompting strategies, but their underlying behavioral patterns are constrained by model architecture and training data. Future development of affective AI should focus on drawing from human cognitive mechanisms to improve multi-cue integration and contextual understanding in complex social stimuli.

Note: Figure translations are in progress. See original paper for figures.