SocialGesture: Delving into Multi-person Gesture Understanding

Supplementary Material

6. Broader Impacts

We highlight the following social impacts of our work:

1. ***First Multi-person Gesture Dataset:*** By introducing multi-person scenes into gesture recognition, our work can lead to the development of more robust social agents. These models are expected to perform more reliably across diverse real-world scenarios, benefiting human-AI interaction applications such as online education and collaborative virtual environments.
2. ***Bringing Gesture Recognition to the Large Language Model Era:*** We provide a new Visual Question Answering (VQA) setting to evaluate the performance of Vision-Language Models (VLMs) and multimodal Large Language Models (LLMs) in human gesture understanding. Experimental results reveal that gesture understanding remains a challenging task for these models, highlighting areas for future research and improvement.

7. Additional Information for SocialGesture

We leverage two main data resource: YouTube and Ego4D. The data collections and annotations have been approved by the Institutional Review Board (IRB). Ego4D dataset contains the games of One Night Ultimate Werewolf and The Resistance: Avalon. To fullfill the Fair Use Policy of public YouTube data, we choose to use Creative Commons Attribution Non Commercial 3.0 (cc-by-nc-3.0) to release the annotations. This will provide free and complete access to the research community, excluding only commercial use.

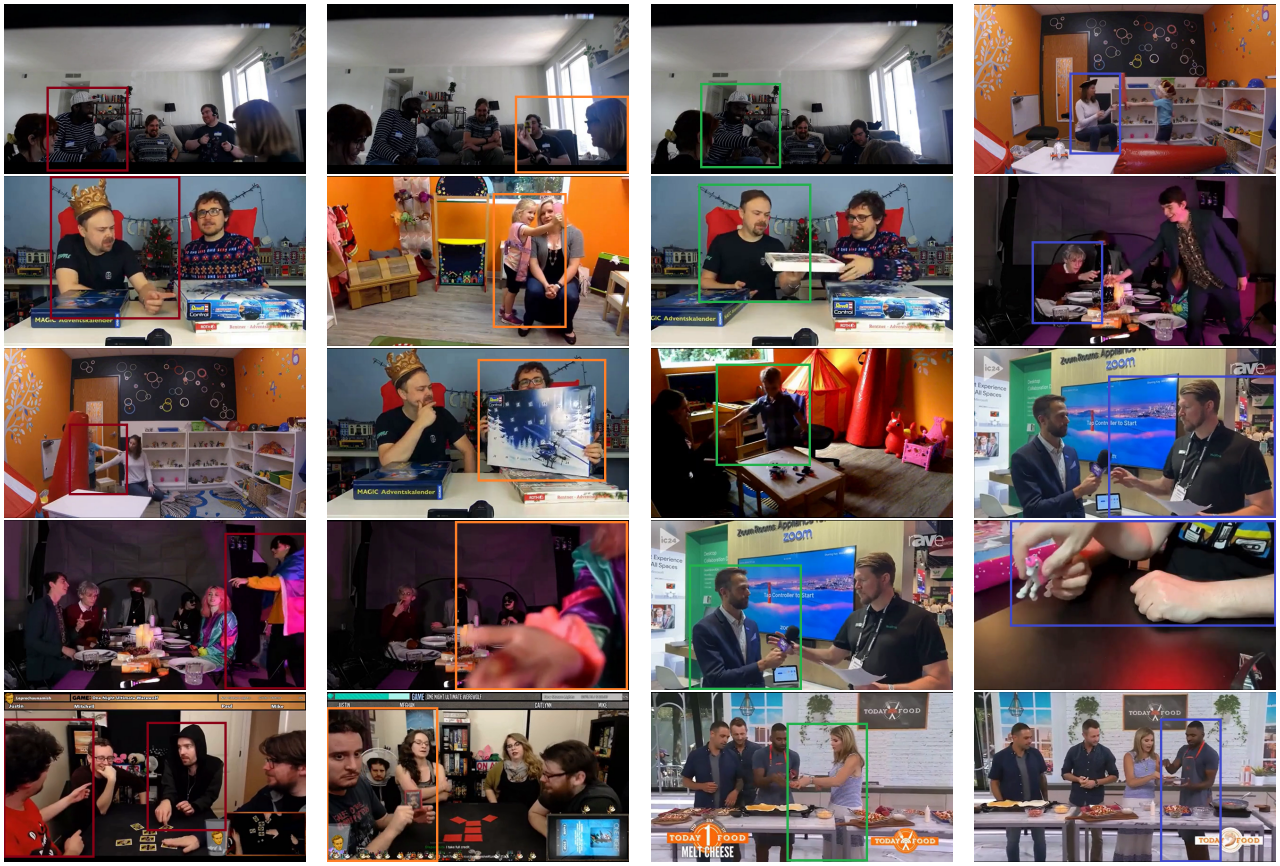
12 annotators (4 male, 8 female) labeled the SocialGesture dataset, all of whom held at least a bachelor's degree. The annotators had an average age of 39 years (ranging from 22 to 56). We employed a structured three-step annotation methodology consisting of annotation, verification, and quality assurance. Annotation consistency was evaluated using inter-annotator agreement and Cohen's kappa. For ambiguous cases and final decisions, we adopted a three-person consensus approach to ensure accuracy and reliability.

As shown in Table 1, our dataset is diverse and includes many different social scenes. Figure 6 presents examples of social gestures in various settings. In SocialGesture, we observe that the pointing gesture constitutes a higher proportion compared to other gestures. The frequencies of showing, giving, and reaching gestures are approximately equal.

Moreover, the manifestation and proportion of gestures vary significantly across different scenes. For instance, in social game scenes such as *One Night Ultimate Werewolf*, the majority of social gestures are pointing gestures. This is because *One Night Ultimate Werewolf* is a game where participants need to vote for the werewolf during gameplay, making pointing gestures highly prevalent. In contrast, giving and reaching gestures have a higher proportion in collaborative cooking videos and interview settings. All four gesture types occur with similar frequency in children education videos. This is because gestures play a crucial role for children in expressing their unspoken thoughts to listeners, as young children have not yet fully developed strong spoken language abilities.

8. Discussion

1. ***Imbalanced Distribution of Social Gestures:***
   Since the videos in our dataset are from natural scenes, the distribution of social gesture is imbalanced. The proportions of gesture types in our dataset are (pointing:reaching:giving:showing = 13.03:1.23:1.22:1.00). It reflects the natural distribution of gestures in context. Standard datasets, such as MS-COCO, also exhibit such imbalances (e.g., 20 times more people than bicycles), which are an inherent property of natural scenes. Our dataset is also similarly naturally imbalanced, e.g. the gestures reaching, giving, and showing are rarer than pointing because they require object interactions. We address imbalance in our multi-class classification experiments by creating a balanced subset.
2. ***Granularity of Categorization:***
   While we have categorized social gestures into four types based on David McNeill's theory of deictic gestures, this classification is coarse considering the wide range of such gestures. It's possible that within each category, there are more subtle, fine-grained gesture types not explicitly accounted for. For example, the showing gesture can be divided into showing to a person on-screen and showing to an off-screen audience. Similar distinctions can occur based on different video perspectives, such as egocentric (first-person) or exocentric (third-person) views. Future work could involve developing more granular classifications and corresponding evaluation metrics to provide a deeper understanding of social gestures.
3. ***Multi-person Gesture and Gaze:*** Both gaze and gesture serve as crucial non-verbal social cues and share several common characteristics. Notably, recent Transformer-based approaches for gaze target estimation [45, 47] have potential to be adapted for detecting the targets of pointing gesture. To facilitate further research, we will release the pointing gesture subset of SocialGesture, providing a benchmark for these methods and enabling exploration of co-occurrence scenarios of gaze and pointing gestures.



Pointing Gesture | Showing Gesture | Giving Gesture | Reaching Gesture

Figure 6. The diversity of four social gesture in our dataset.