Comparing manual and automated gesture analysis methods for naturalistic video recordings

Christeena Thomas & Melanie J. Bell
Anglia Ruskin University, Cambridge, UK
ct842@pgr.aru.ac.uk, melanie.bell@aru.ac.uk

In recent years, gesture research has progressed beyond traditional qualitative approaches by incorporating gesture kinematics using motion-capture and video-based technologies to facilitate both qualitative and quantitative analyses (Ienaga et al. 2022). Advancements in pose estimation technologies, such as OpenPose, offer automated alternatives to labour-intensive manual annotation. However, questions remain regarding their reliability and accuracy in capturing fine-grained kinematic and temporal features of gesture. This paper presents a comparative analysis, investigating to what extent OpenPose's automated gesture detection aligns with manual annotations of naturalistic video recordings using ELAN.

The study uses the *Ellen* corpus (Tong 2023), a series of 1:1 chat-show interviews between a host (Ellen) and various guests. In ELAN, we manually annotated these videos for all hand movements produced by all speakers, identifying the onset and offset of gesture movements and coding them for semiotic type after McNeill (1992). Inter-rater reliability for gesture type was high ($k > 0.88$). We then used the MULTIDATA pipeline (MULTIDATA 2024) to extract positions for all key body points generated per frame by OpenPose; hand movements were detected using Euclidean distance and velocity calculations implemented in Python. Finally, using the manual coding as a benchmark, we calculated OpenPose detection accuracy across hand movements for different speakers, and across semiotic types.

Preliminary findings from performance metrics reveal higher precision than recall. In other words, while gestures were correctly identified, many instances were missed, suggesting that the velocity thresholds used for movement detection may have been suboptimal. Self-adaptors and metaphoric gestures were detected more reliably than the other types; we will present the results of ongoing analyses examining whether this reflects gesture trajectory variations or other kinematic features characteristic of different semiotic types. Overall, our findings indicate that while OpenPose shows promise, further refinements are needed to fully capture the phonetics of gesture, possibly using machine learning classifiers that use both motion and handshape features as input for better gesture recognition.

References • Ienaga, N., Cravotta, A., Terayama, K., Scotney, B. W., Saito, H., & M. G. Busà (2022). Semi-automation of gesture annotation by machine learning and human collaboration. *Language Resources and Evaluation*, 56(3), 673–700. • McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press. • MULTIDATA (2024). *Skills and resources for the multimodal turn: Unlocking audiovisual datasets for research and learning*. • Tong, Y. (2023). *Embodiment of Concrete and Abstract Concepts: The Role of Gesture* [PhD, Vrije Universiteit Amsterdam].