Semantic Gesture Retrieval for real-time co-speech gestures

Yassine Machta, Marius Le Chapelier, Oussama Silem and Justine Cassell

Introduction

Most existing systems rely on data-driven generation, trained on either:

- **Motion capture [3]**: actor-biased and limited in variety
- **2D-to-3D estimations [4]:** cheaper but noisy due to pose tracking artifacts

Pose estimation

These systems work, but tend to favor generic **beat gestures** and struggle to produce **semantically rich** movements.

Multiple works ([1],[2]) fill the semantic gap but lose their real-time abilities by relying on LLM retrieval for the semantic capability.

Our motivation is to propose a **gesture retrieval framework** that seeks to replace this LLM Bottleneck while ensuring:

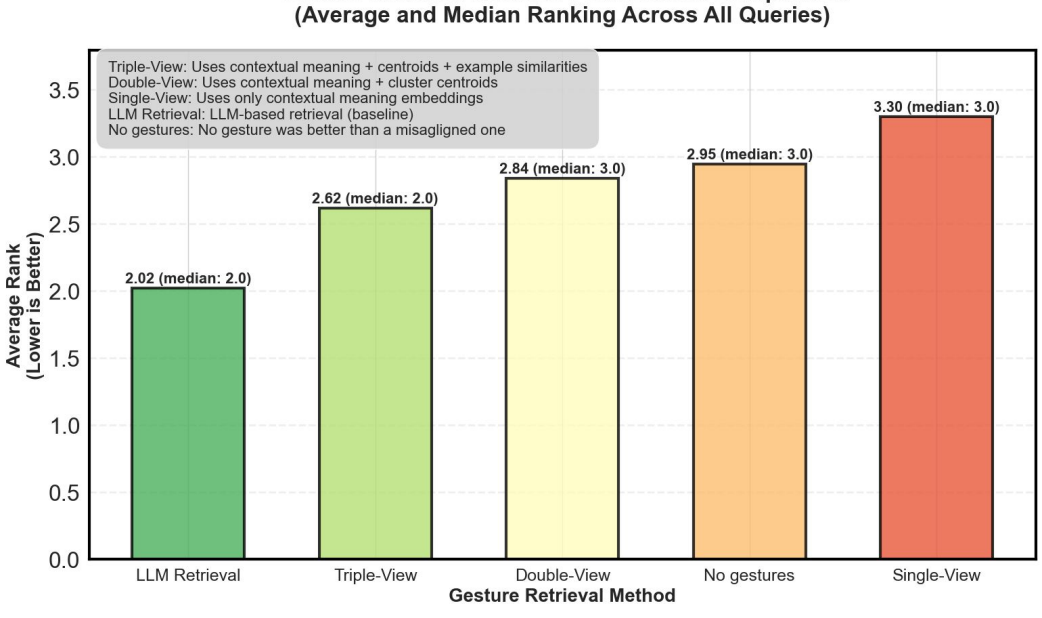
- **Speed**, for real-time gesture selection
- **Interpretability**, by linking gestures to explicit semantic embeddings and tokens
- **Semantically aware**, by choosing gestures that fit the the context

Results

We evaluated our approach using a **human-ranked dataset** of 500 synthesized utterances.

Gesture Retrieval: Compared **Single-View**, **Double-View**, **Triple-View retrievals**, **LLM Retrieval**, and **No gestures** (baseline). (information on retrieval view in the graph below)

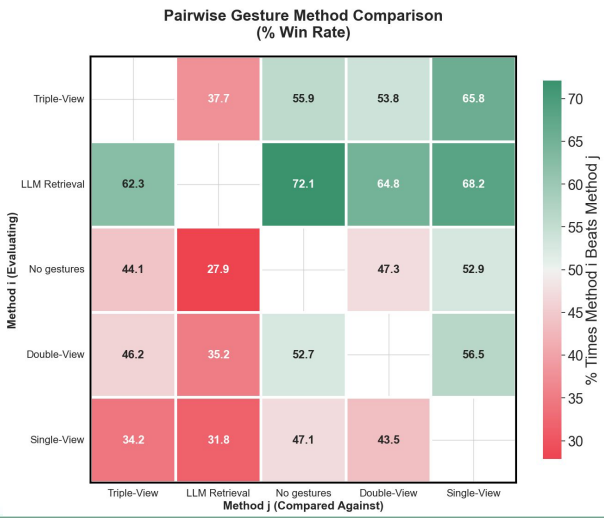
Human Ranking: Evaluators ranked gestures by semantic relevance (1-5)



LLM Retrieval is still the most effective method by a margin, **but our Triple-View is a strong and faster alternative.**

We also were able to prove the **impact of augmentation** as single view retrieval was so misaligned that "No gestures" were preferred over it on average.

The main attraction of our method comes from its **speed and it's satisfying performance**

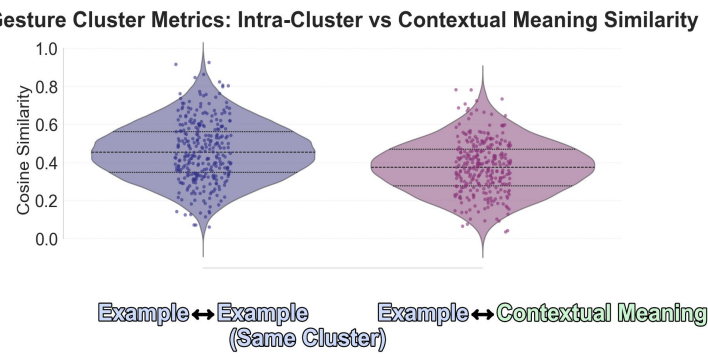


Methods

Initial SeG [1] : Gesture Name+description | Contextual Meaning

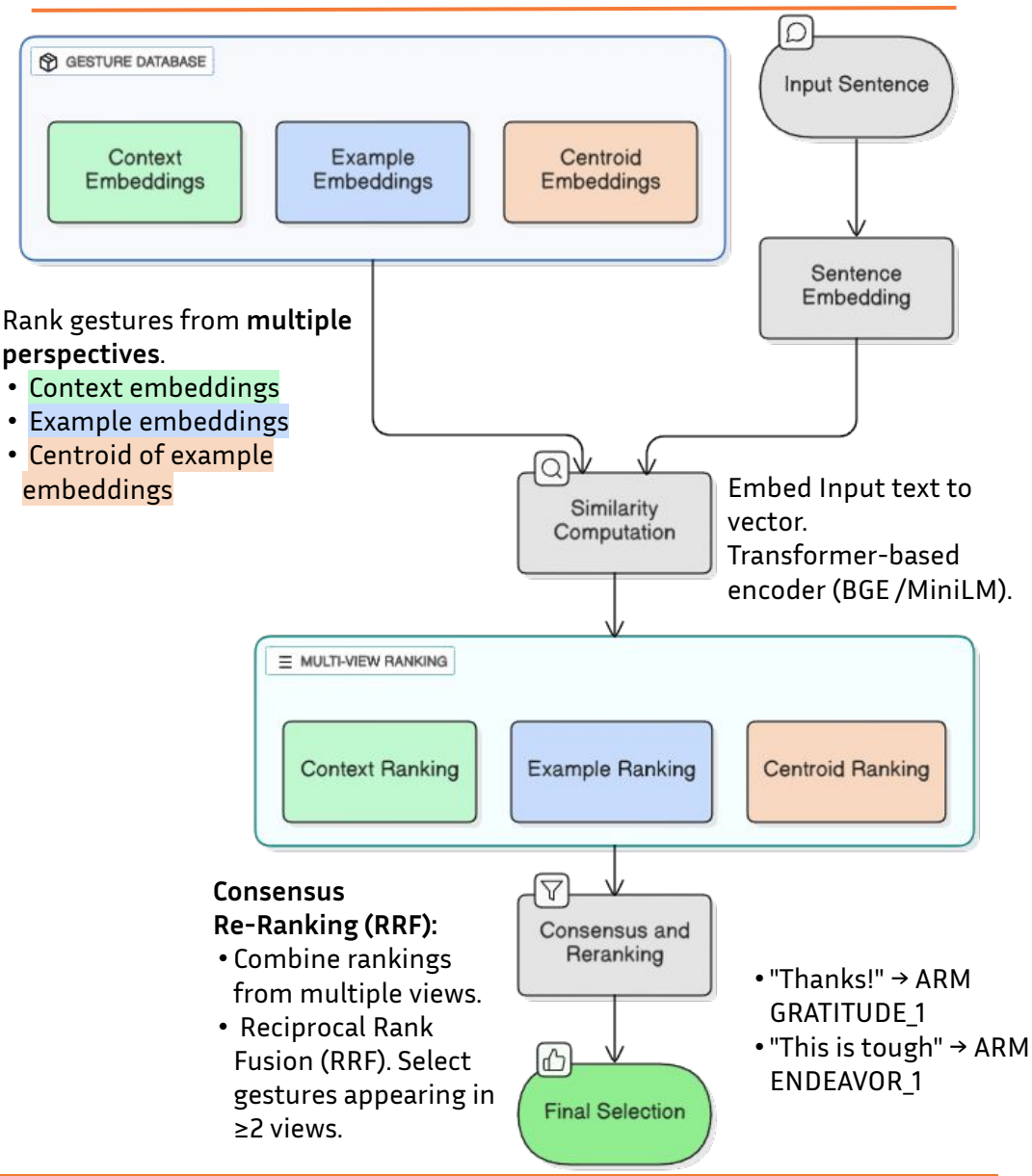
BELLY PAT + "The belly is patted gently with the hand" | "Appreciation"/ "Contentment"

We generate 15 example sentences per gesture → broaden semantic coverage while retaining gesture meaning.
We also compute the centroid of these examples



This augmentation expands the number of documents available that match the same gesture but have different vector similarities **with each other** and with **the original** contextual meaning

Augmented dataset : Gesture | Contextual Meaning | Centroid | Examples



Rank gestures from **multiple perspectives**.

- Context embeddings
- Example embeddings
- Centroid of example embeddings

Embed Input text to vector.
Transformer-based encoder (BGE /MiniLM).

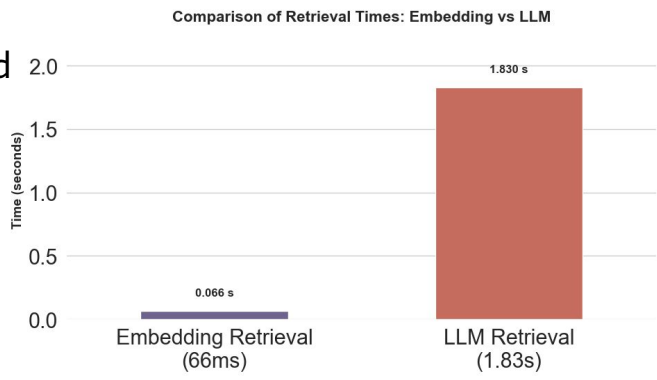
Consensus Re-Ranking (RRF):

- Combine rankings from multiple views.
- Reciprocal Rank Fusion (RRF). Select gestures appearing in ≥2 views.

- "Thanks!" → ARM GRATITUDE_1
- "This is tough" → ARM ENDEAVOR_1

Discussions and conclusions

This method is a **real-time simple lightweight** method to retrieve a semantic gesture for a given utterance.
It is slightly inferior to LLM retrieval but **significantly faster and cheaper** !



Limitations: Lack of flexibility, dependent on semantic embedding model.

Future work: More accurate synthetic documents (examples) to also encode intent and tone and fill the rest of the gap with the LLM. Integration in data-driven frameworks

[1] Z. Zhang et al, **Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis**

[2] MH. Mughal et al, **Retrieving Semantics from the Deep: an RAG Solution for Gesture Synthesis**

[3] H. Liu et al, **EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling**

[4] V. Argawal et al,**Seamless Interaction: Dyadic Audiovisual Motion Modeling and Large-Scale Dataset**