A Workflow for Real-time Visualization and Data Analysis of Gesture using Motion Capture

Jean-François Jégo, Vincent Meyrueis, Dominique Boutet

A Workflow for Real-time Visualization and Data Analysis of Gesture using Motion Capture

Jean-François Jégo
INREV-AIAC Laboratory
Université Paris 8, France
jean-francois.jego03@univ-paris8.fr

Vincent Meyrueis
LCPI Laboratory
ENSAM, France
vincent.meyrueis@gmail.com

Dominique Boutet
DYLIS Laboratory
Université de Rouen, France
dominique.boutet@univ-rouen.fr

ABSTRACT

In this paper, we investigate new ways to understand and to analyze human gesture in a research context applied on co-verbal gesture across language. The research project focuses on the quality of the movement and consider the gesture "pulse of effort." We propose a workflow for real-time gesture analysis to visualize gesture kinematics features (Velocity, Acceleration, Jerk) from heterogeneous data (Video, Motion Capture and Gesture Annotations) at the same time base. The tools designed here provide immersive and interactive explorations of data: users can test hypotheses and embody gesture visualization and descriptors adopting different Frames of Reference using augmented reality. We have conducted an evaluation protocol in the field of linguistics that compares 496 annotated gestures to benchmark the workflow.

CCS CONCEPTS

• **Human-centered computing** → Visualization • **Applied computing** → **Arts and humanities** → Language translation

KEYWORDS

Gesture, Visualization, Motion Capture, Embodiment, Workflow, Real-time, Augmented Reality

1 INTRODUCTION AND MOTIVATION

We propose to investigate new ways to understand human gesture—from the perception to the analysis—in a research context applied on co-verbal gesture across language. More specifically, we wish to study and measure aspectuality of gestures of two locutors comparing three languages (Russian, German and French). The research project focuses on the quality of the movement and consider the gesture Boundedness (bounded or unbounded) in terms of kinetics "pulse of effort" as defined by Boutet et al. [1]. In the first study, the gestures were coded and analyzed visually using videos with ELAN. We needed to evaluate the "pulse of effort" which is based on gestures kinematics and it appeared video was limited to extract features. Indeed, videos only allow recording "2D" point of view defined by the initial position of the camera. Thus, obtaining different frames of reference of the participants' gestures requires several cameras. In this study we wish to explore the different levels of complexity of the "pulse of efforts" and its boundedness, and this raises two main questions:

—How to efficiently visualize gesture with different frames of reference?

—How to investigate new gesture descriptors hypothesis with a quicker data process?

We expect real-time 3D technologies could help in solving these questions from the acquisition to the analysis. We propose first to look at previous works to explore this hypothesis.

2 PREVIOUS WORKS

2.1 Frame of Reference

The ways the languages conceptualize the space through events or in reference to the objects are called Frames of Reference (FoR). Levinson et al. [2] categorize the FoR in three main types: absolute, relative and intrinsic. In the absolute FoR, objects or events are located according to a geographic location or a direction (north, south, west, east). The second type of FoR is relative to the orientation given by an observer such as the egocentric point of view [3].

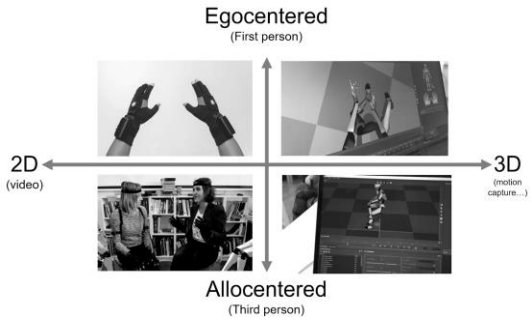

Figure 1: ego or allocentric point of views regarding 2D video or 3D Mocap data collections

The third FoR—named intrinsic—locates an entity according to the proper orientation of an object.

Applied to Motion Capture data (Mocap), the three FoR are declined, in an absolute way through the reference of the lab, or in a relative way—mainly an egocentric one—where the movement is going up or down, to the left or the right. The ultra-dominant analyses for the gestures or for the Sign Languages are made in this egocentric FoR [4]—[6]. The third category deconstructs each segment in an intrinsic FoR, for which the degrees of freedom and the range of motion associated establish the direction of the possible location of the segments regardless of the location of an observer [7]. We note many annotators' works are using video producing an allocentric FoR but in the meanwhile, they also imitate the gestures, embodying them with an egocentric FoR. Thus, we wish to investigate new strategies one could adopt to better understand the gesture produced by the performer.

2.2 Gesture Analysis Using Motion Capture

Mocap is more and more present in gesture studies in addition to traditional video annotation but it produces many data which is challenging to understand or to visualize. Indeed, the tools available are mostly based on video analysis paradigms offering video preview and timeline creation for annotation such as in software like ELAN or ANVIL [8], [9]. The video has the advantage to be a ground truth but the point of view is still limited (number of possible observation angles, field of views and FoR). We note ELAN allows to import data tables represented as curves in the timeline but data is limited to 1 dimension per curve [8]. In ANVIL we can import Mocap but tools of data visualization and process in real-time are still limited [9].

Several platforms such as MOVA [10] helps in dealing with gesture descriptors in an interactive and web page in real-time. The online tool shows relevant descriptors such as velocity/acceleration/jerk or based on Labanotation. We note Mocap and data are presented in 2D timelines and it is not yet possible to observe and to process 3D Mocap in the web browser and to export data. Maestre et al. [11] also developed an online hosting platform to extract descriptors and annotate multimodal string quartet performance data (audio, video, Mocap, and derived signals). The authors provided first-person outlook on the technical challenges involved in the workflow design applied to music performance but doesn't propose an immersive visualization tool in an embodied Frame of Reference.

2.3 Embodying Gesture Analysis

Regarding cognitive science and recent theories, embodiment could help to harness embodied cognition, empathy state or emotions [12]. In the Fig. 1, we propose to formalize the different FoR that could take place during gesture annotation. Then, we make the hypothesis immersive tools such as augmented reality could help analysis, annotating gesture showing hidden qualitative and quantitative descriptors of movement and switching quickly between FoR, going for instance back and forth between egocentric and allocentric point of views. Hypotheses will be explored with a use case and prototypes, then we'll discuss what is changing in annotation workflow.

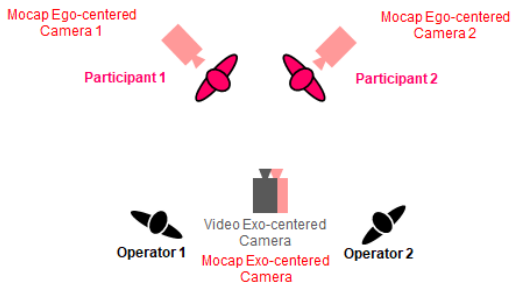

Figure 2: data collection of two participants using video camera (in gray) and Mocap "virtual" cameras (in red)

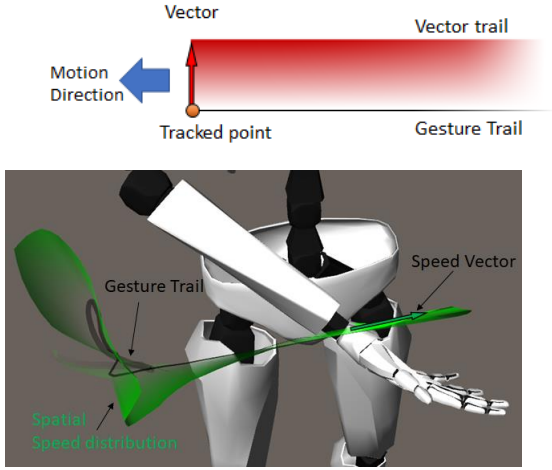

Figure 3: (top) fading rules according to the 3D vector representation on gesture trail; (bottom) velocity vector representation following the right-hand gesture in Unity

3 WORKFLOW DESIGN AND RESULTS

3.1 Experimental Setup

The protocol involving Mocap is slightly different from a traditional data collection using a single video camera. Not only does Mocap allow one point of view on data, but as many as required since we can use virtual camera to study gesture from the "skeleton" of the participant (Fig. 1). We can also take advantage of virtual cameras to render for instance a video with a wider or a closer point of view. An egocentric point of view of the participant can also be adopted to visualize and embody the gesture which is difficult to achieve with a classic setup using a video camera (Fig. 2).

Regarding the Mocap systems to use, many exist (mechanical, optical, inertial…), with different use cases, accuracy and price. In our study, we opt for an inertial system (Perception Neuron) which appears to be more convenient in terms of accuracy (millimeters) and frequency (from 60 Hz to 120 Hz). This type suit also solves occlusion problems compared with optical systems. However, experiment conditions have to be considered since inertial motion-capture system uses IMUs (Inertial Measurement units) which are sensitive to the magnetism of the environment. During the recording phase of the data collection, the participant has to adopt a specific posture called T-Pose to set up then the virtual body in a 3D engine. To synchronize the Mocap with the video and the audio recordings, we asked the participants to clap their hands once right after the T-Pose and one more time at the end of the recordings. It helps operators to do a quick evaluation of the accuracy and check if data are not missing or drifting. We took care of data formatting to be compatible with traditional analytical tools like ELAN where it is possible to annotate a video analysis in a primary step. Then, we exported the gesture list to the 3D real-time platform we designed to explore Mocap in a more immersive approach.

3.2 Gesture Data Visualization

In order to analyze and compare the gestures, we develop a tool to extract features and visualize them. It is based on the real-time computation of the velocity, the acceleration and the jerk vectors expressed in any Cartesian coordinates such as the room or the user's body (Fig. 3). We realize displaying too many motion curves could disturb the analysis, then we implement a real-time selection that allows isolating a specific gesture from other body movements. To assess the question of embodying gesture, we use a real-time game engine (Unity) which offers immersive and interactive capabilities for gesture analyses. This approach is distinguished from the traditional gesture analysis based on planar projections [8], [9]: immersive technologies allow keeping the richness of 3D for gesture descriptors allowing manipulating data with a more natural three-dimensional perception of information; interaction allows adopting dynamic point of views depending on gestures. Thus, the user is free to switch between point of views in real-time for instance to analyze the gesture at the desired speed (Fig. 4). It makes the gesture analysis free from

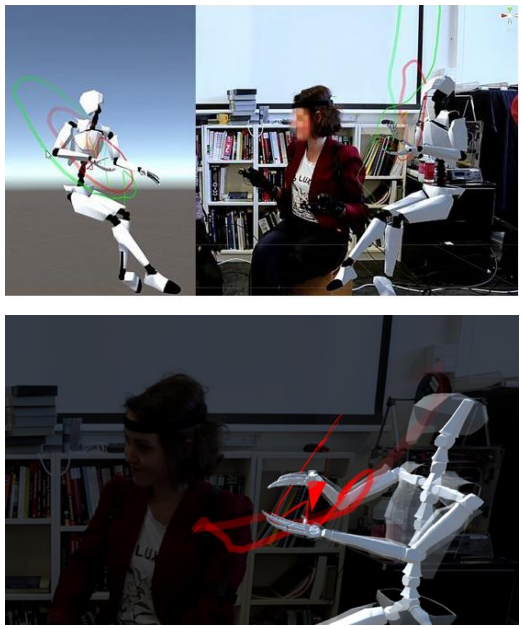

Figure 4: (top) Motion Capture and video superimposed and synchronized, in red: acceleration, in green: jerk; (bottom) "Augmented video" showing acceleration curve

time-space constraints that enhance gesture reading. To assist the 3D analysis with visual clues, we design the kinematic gesture vector analysis (velocity, acceleration and jerk) as a ribbon (Fig. 3, 4). This shape representation is inspired by the mechanical force distribution for structure analysis usually used in engineering. The more the surface is thick, the more the value of the vector is high at this time (Fig. 3). It allows showing the distribution according to the gesture trail: we are able then to perceive more efficiently the magnitude and the vector direction on the gesture evolution geometry. Since we compute vectors from Mocap data stream in real-time, we can obtain an augmented video (Fig. 4) and we can record and export them to external software such as ELAN to complete a traditional video analysis including Mocap data expressed in any FoR needed.

3.3 Implementation and Benchmark

In order to test our workflow and visualization tool, we run a study on the normalized distributions of the vector's magnitudes of velocity, acceleration and jerk based on the normalized gesture execution time with 50 samples (Fig. 5). By combining all these values, we can compute an average distribution that allows us to characterize the "gesture dynamic" according to unbounded and bounded video coding and the populations studied. To do so, kinematic analyses were conducted with 30 participants. The analysis of the Mocap is based on the video coding data tagged by three different experts using the ELAN software. We focus on the analysis of gestures with the best triple consensus tagging which represents 348 gestures noted bounded and 148 gestures noted unbounded for the whole user panel.

The Mocap analysis shows that both hands have the same behaviors, the bounded gestures have an increasing velocity, acceleration and jerk distribution profiles due to the nature of their shape (Fig. 5, top). Regarding the unbounded gestures, their velocity, acceleration and jerk distribution profiles are less pronounced than the bounded gesture ones due to less affirmed gesture dynamics (Fig. 5, bottom). In both analyses, given the number of samples, the standard deviation remains constant on each sample. Overall, there is a significant difference between the curves that makes possible to distinguish bounded and unbounded gesture based on the gesture analysis approach detailed in this paper [13]. This first result is a first benchmark that shows the workflow allows extracting comparative curves between gestures data processed and synchronized at the same time base in real-time.

4 CONCLUSION & OUTCOMES

In this research, we proposed a workflow for gesture analysis that allows perceiving and annotate gesture in real-time to test new descriptors and hypotheses. We developed a 3D gesture visualization tool based on Mocap associated with real-time technologies (Fig. 4). It allows visualizing gesture kinematics features (velocity, acceleration, jerk) as ribbons in space. The workflow allows synchronizing heterogeneous data (Video, Mocap and ELAN annotations) in a same time base to extract comparative curves between gestures.

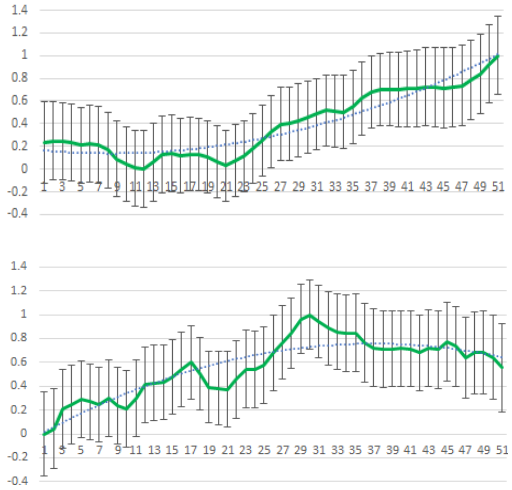

Figure 5: normalized distributions of the velocity for 50 samples of bounded gestures (top) and unbounded gestures (bottom)

In order to benchmark the tools designed, we run an evaluation in the field of linguistics that compares more than 496 annotated gestures. The workflow also allows extracting offline data to complete an analysis of the "pulse of efforts" of gesture. This raises questions regarding traditional offline and online analysis tools and paradigms based on video: real-time technologies, immersion and interaction now offer to embody gesture visualization and descriptors.

To go further, we could imagine using stereoscopy and head motion tracking using for instance virtual reality headsets. This could help to understand the movement without making a mental effort in the resolution of the information related to the movement usually projected on a plane. The current tool allows adopting different Frames of Reference to better understand the performed gestures and we plan to explore further capabilities of real-time technologies to conduct new analyses regarding emotion and empathy.

ACKNOWLEDGMENTS

We wish to thank all the participants. The research was carried out at Moscow State Linguistic University and supported by Russian Science Foundation (project No. 14-48-00067П).

REFERENCES

[1] D. Boutet, A. Morgenstern, et A. Cienki, « Grammatical Aspect and Gesture in French: A kinesiological approach», *Russ. J. Linguist.*, vol. 20, nº 3, p. 132-151, dec. 2016.

[2] S. C. Levinson, « Frames of reference and Molyneux's question: Crosslinguistic evidence», *Lang. Space*, vol. 109, p. 169, 1996.

[3] J. Dokic et E. Pacherie, « On the very idea of a frame of reference », *Typol. Stud. Lang.*, vol. 66, p. 259, 2006.

[4] D. Brentari, *A prosodic model of sign language phonology*. Mit Press, 1998.

[5] S. K. Liddell, *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.

[6] D. Mcneill, *Hand and mind: what gestures reveal about thought*. Chicago ; London: University of Chicago press, 1992.

[7] D. Boutet, « Structuration physiologique de la gestuelle : modèle et tests», *Lidil*, nº 42, p. 77-96, nov. 2010.

[8] H. Sloetjes et P. Wittenburg, « Annotation by category-ELAN and ISO DCR», in *6th international Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[9] M. Kipp, L. F. von Hollen, M. C. Hrstka, et F. Zamponi, « Single-Person and Multi-Party 3D Visualizations for Nonverbal Communication Analysis. », in *LREC*, 2014, p. 3393–3397.

[10] O. Alemi, P. Pasquier, et C. Shaw, « Mova: Interactive movement analytics platform», in *Proceedings of the 2014 International Workshop on Movement and Computing*, 2014, p. 37.

[11] E. Maestre *et al.*, « Enriched multimodal representations of music performances: Online access and visualization», *Ieee Multimed.*, vol. 24, nº 1, p. 24–34, 2017.

[12] A. Berthoz, *The brain's sense of movement*, vol. 10. Harvard University Press, 2000.

[13] D. Boutet, J.-F. Jégo, et V. Meyrueis, «POLIMOD Pipeline: documentation. Motion Capture, Visualization & Data Analysis for gesture studies», Université de Rouen, Université Paris 8, Moscow State Linguistic University, Research Report, dec. 2018.