Quantifying sign language movement kinematics from video

Aline Normoyle, Samuel J. Rothstein, and Norman I. Badler

Abstract

We describe our system for extracting kinematic metrics for sign language gestures from a video. The goals of these analyses are to provide insights into how factors related to physical movements, such as principles of least effort and ease of articulation, might affect how sign languages evolve or differ across individuals.

Motivation

Evidence suggests that psycho-physical factors influence sign languages.

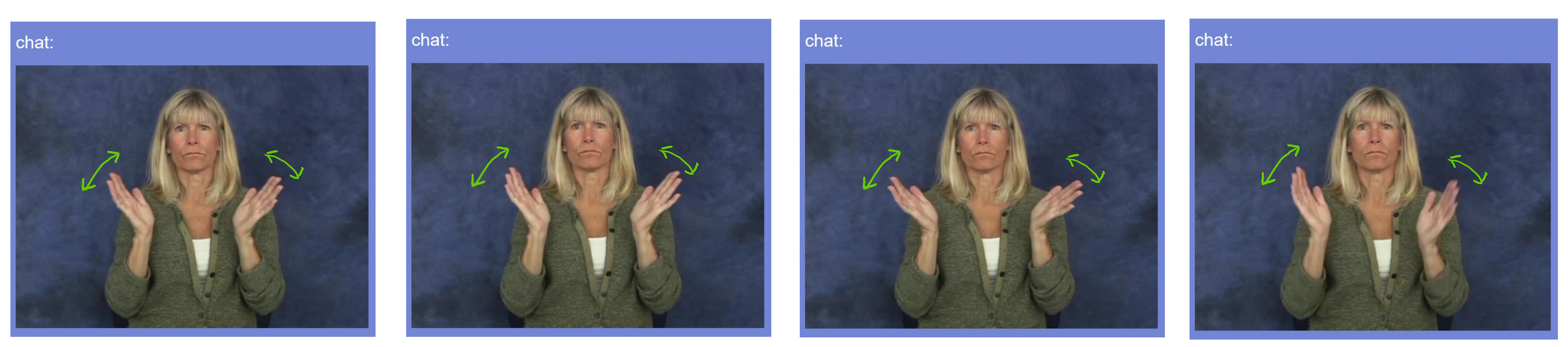
- Signers might use a one-handed version of a sign in casual conversations, and a two-handed version in formal settings. [Zimmer 2000]
- Signers might also reduce the number of repetitions in a sign, or cut short the articulation of a sign, or replace translational symmetry with reflective symmetry. [Napoli et. al 2014]
- Signs may evolve to move away from peripheral locations and closer towards the face [Frishberg 1975]

Such shortcuts are analogous to co-articulations, contractions, and acronyms used in spoken English. We can use kinematic metrics to identify how sign languages change based on context, differ across individuals, or change to make expressions easier and faster to convey.

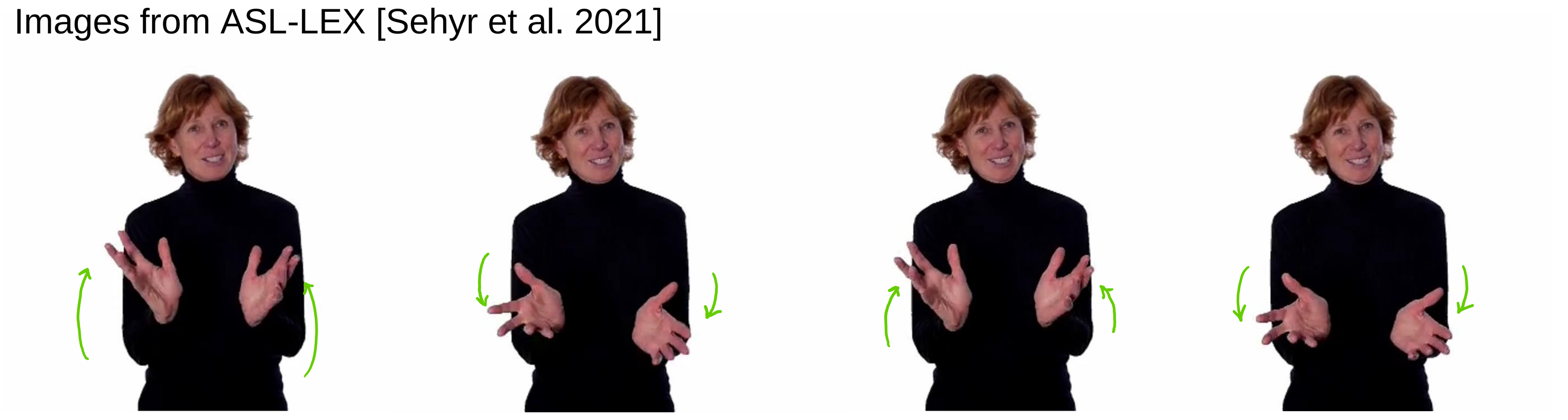
Most sign language datasets are based on video. Usually, fluent speakers need to transcribe and annotate these data sets manually. However, recent advances in computer vision give us the opportunity to automate parts of this process. Furthermore, we can measure quantities that would have been prohibitively difficult to measure before.

ASL movement: Saying *chat*

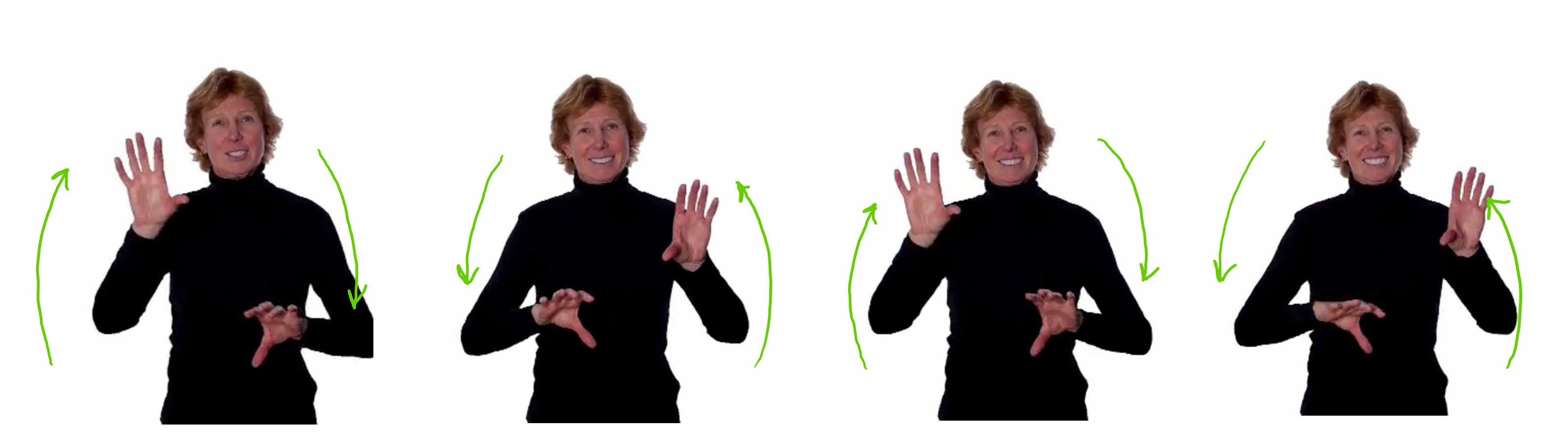
A primary goal of our pipeline is to quantify the *articulatory effort* needed to make different signs. In the examples below, we show several variants for signing the word *chat*, as in "Many people chatted at the cafe". Some variants require the speaker to move the hands much more than others.



Chat in ASL, signed with the hands near the face using a quick repetitive movement. Images from ASL-LEX [Sehyr et al. 2021]



Chat in ASL, signed with the hands below the face using a quick repetitive movement. This gesture indicates a short chat. Images from signingsavvy.com



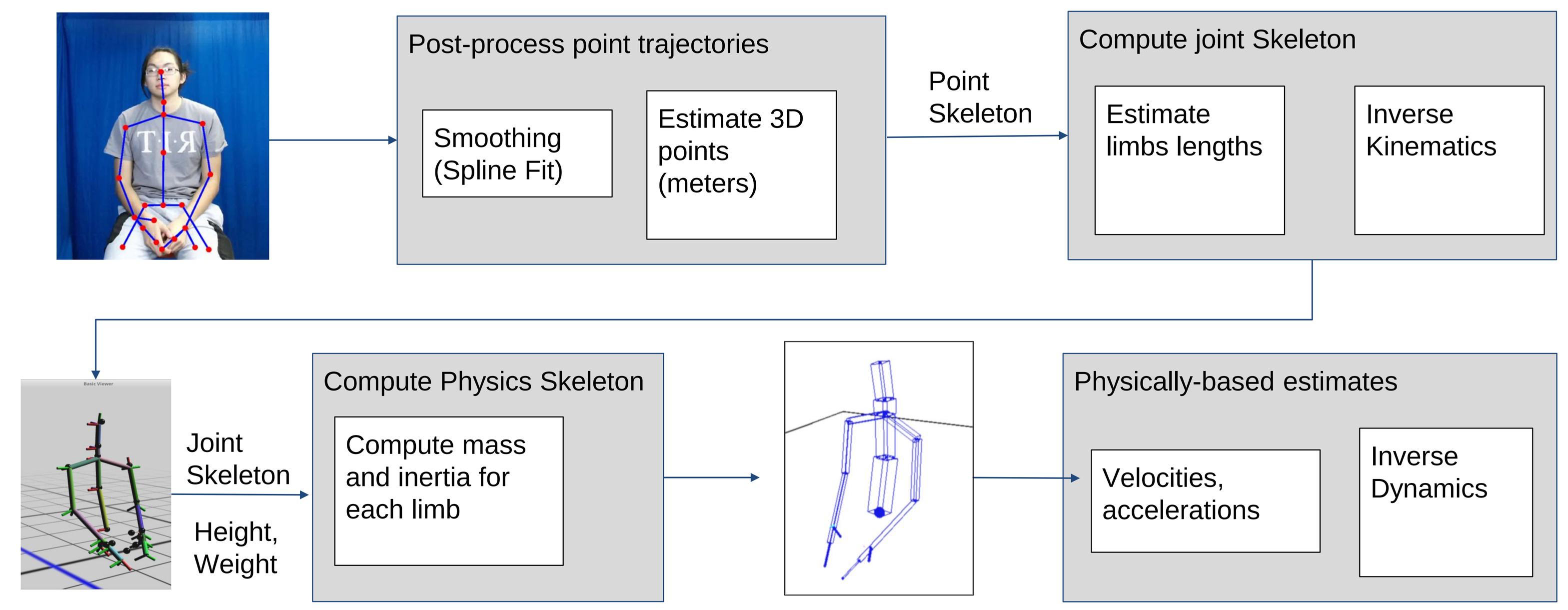
Chat in ASL, signed with the hands alternating up and down with mirrored symmetry. This gesture indicates a long chat. Images from signingsavvy.com

Acknowledgements

Many thanks go to our collaborators: Matthew Dye (RIT), Naomi Caselli (BU), Corrine Occhino (RIT), and Ann Senghas (Barnard) for their expertise in linguistics and sign languages; Bruno Miranda Artacho and Andreas Savakis (RIT) for their work to extract poses from the video; and Matt Huenerfauth for providing the RGB-D data set. Thanks also to our reviewer for their helpful feedback and comments as well as to Effie Liu for her early work on this project. This work is supported by NSF Grant BCS 1749376 and REUs.

Pipeline

From image-based to physics-based poses



We take poses of the upper-body extracted from video [Artacho and Savakis 2020] and then fit a physically-based, hierarchical skeleton to it. The examples here use a data set based on RGB-D images recorded with the Kinect [Hassan et al. 2020]. To estimate joint masses, we use an anthropometrics model based on [Winter 2009] which outlines how weight and center of mass is typically distributed across the body, given a subject's height and weight.

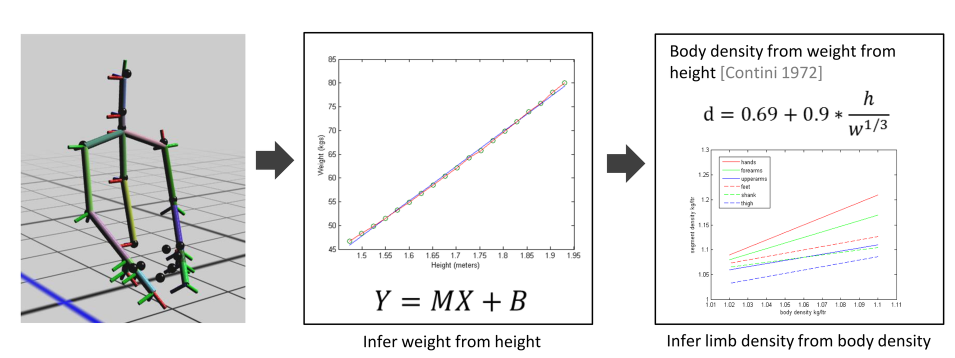
Model

Data set and skeleton properties

Motion is 24 fps

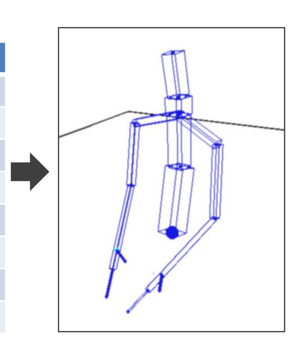
The skeleton consists of 17 points
- 3 for the spine, 1 for the neck, 1 for the head
- 3 for each hand (palm, index, thumb)
- 3 for each arm (shoulder, elbow, wrist)

Computing mass and inertia for the physics skeleton



Segment	Mass	COM
Hand	0.006	0.506
Forearm	0.016	0.430
Upper arm	0.028	0.436
Foot	0.0145	0.5*
Shank	0.0465	0.433
Thigh	0.1	0.433
Head	0.081	0.5*
Trunk	0.497	0.5*

Distribute mass and COM between joints.
* Estimates not based on empirical data

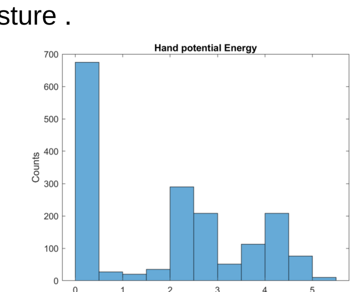


Mass distribution shown in the dart physics simulator

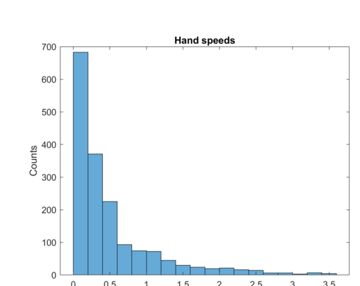
Metrics

Our current metrics explore measures of articulatory effort using different complexity of model. Some metrics, such as distance to a rest posture, only rely on the points of the skeleton. Other metrics, such as kinetic energy and torques, require full mass and inertia estimates for the entire skeleton. We compute velocities and accelerations using 5-point central differencing and torques using inverse dynamics. These metrics are lowest when both hands are in the lap and highest when both arms are accelerating quickly.

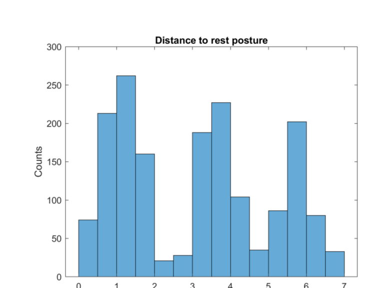
Distance to rest posture: Measures the difference (in meters) between the current pose and a manually chosen rest posture .



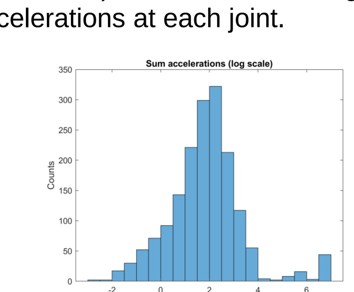
Hand speed: Hand speed (m/s^2) is the sum of the speeds of both wrists.



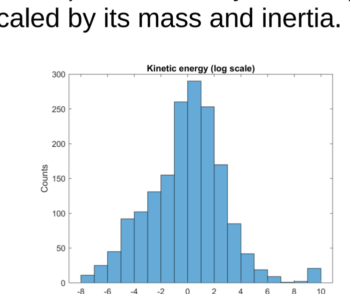
Hand potential energy: Hand potential energy (Joules) quantifies the effect of gravity on the hands



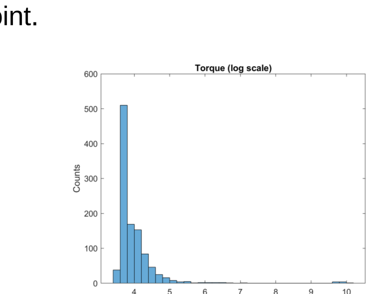
Accelerations: Sum of accelerations ($radians/s^2$) summarizes the angular accelerations at each joint.



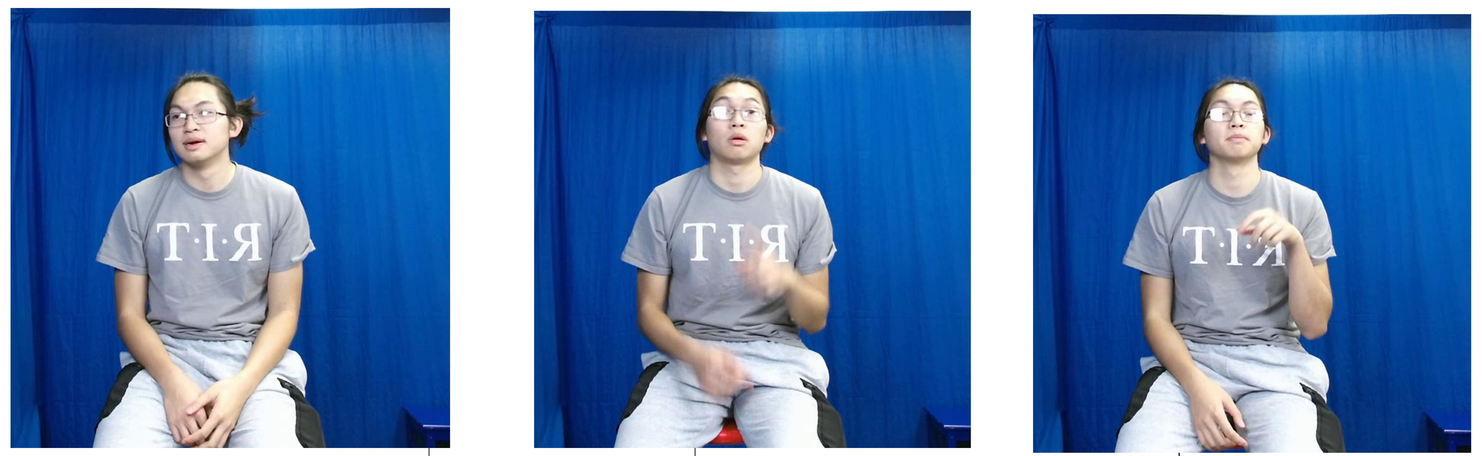
Kinetic energy: The kinetic energy (Joules) is the velocity of each joint scaled by its mass and inertia.

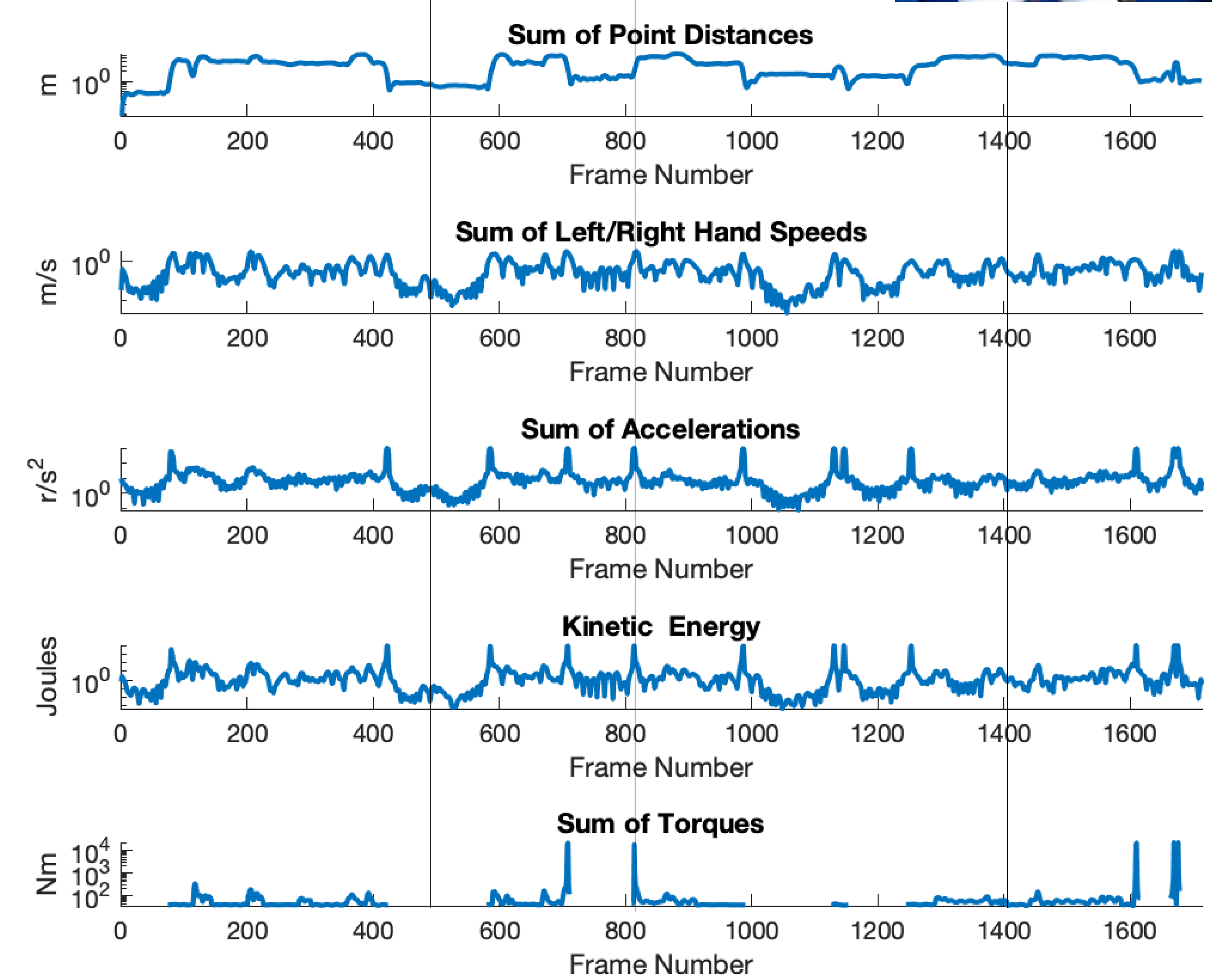


Torques: The sum of torques (Nm) summarizes the forces at each joint.



Example: Introduction





Time series showing all metrics. The x-axis shows the frame number. The y-axis corresponds to the units for each metric. All metrics are on a log scale (different limits). Current metrics have their lowest values when both hands are in the lap (frame 500). High values correspond to frames with large velocities and hands raised (frames 815 and 1424).

Ongoing work

Our current work is exploring additional metrics for measuring symmetry, size, and repetitions of movement in the trajectories of the hands. We are also designing experiments to validate such metrics.

Several challenges need to be addressed to complete this system. The body model currently used to estimate mass and inertia does not take into account different body sizes or gender. Furthermore, the distribution of mass across the body requires the weight and height of the individual, which is difficult to discern from a video.

References

1. Bruno Artacho and Andreas Savakis. 2020. UniPose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7035–7044.
2. Frishberg, N. (1975) Arbitrariness and iconicity: historical change in American Sign Language. *Language* 51, 696-719.
3. Saad Hassan, Larwan Berke, Elahe Vahdani, Longlong Jing, Yingli Tian, and Matt Huenerfauth. 2020. An Isolated-Signing RGB-D Dataset of 100 American Sign Language Signs Produced by Fluent ASL Signers. In *Proceedings of the LREC 2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. 89–94.
4. Donna Jo Napoli, Nathan Sanders, and Rebecca Wright. 2014. On the linguistic effects of articulatory ease, with a focus on sign languages. *Language* (2014), 424–456.
5. Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education* 26, 2 (2021), 263–277.
6. David A. Winter. 2009. *Biomechanics and Motor Control of Human Movement, Fourth Edition*. John Wiley & Sons, Inc.
7. June Zimmer. 1989. Toward a description of register variation in American Sign Language. In *The sociolinguistics of the Deaf community*. Elsevier, 253–272