Open-Source Pipeline for Skeletal Modeling of Sign Language Utterances from 2D Video Sources

Aline Normoyle¹, Bruno Artacho², Andreas Savakis², Ann Senghas³, Norman I. Badler⁴, Corrine Occhino⁵, Samuel J. Rothstein⁶, Matthew W. G. Dye²

¹Bryn Mawr College, 2 Rochester Institute of Technology, 3 Barnard College, 4 University of Pennsylvania, 5 Syracuse University, 6 Swarthmore College

How do the capacities and limitations of human articulatory and perceptual systems shape sign languages? A major challenge for performing research on this question is that most sign language datasets are video-based. Thus, they do not directly provide the head, arm, and finger positions that are needed to estimate locations, distances, velocities, and energy. Here, we propose an open-source pipeline that makes it possible to answer questions about the visual-gestural and articulatory characteristics of sign languages (Figure 1). Our approach leverages recent advances in computer vision to compute three-dimensional estimates of human pose from video [Artacho and Savakis, 2020]. Given these pose estimations, we fit a physically-based, hierarchical skeleton to the data (Figure 2a). This skeletal model incorporates the size and mass of the signer’s limbs and enforces constraints on how the body can move (for example, an elbow can only rotate around a single axis). Once we have the skeletal model, we can extract smooth estimates of distance, velocity, and physics-based quantities such as forces and effort (Figure 2b). Furthermore, we can also import this skeletal data into other open source tools, such as OpenSim [Seth et al., 2018], a biomechanics and muscle simulator, or Dart [Lee et al., 2018], a physics simulator. We demonstrate our pipeline using two data sets. The first consists of RGB-D images recorded with the Kinect [Hassan et al., 2020]. The second is an archive of RGB videos of Nicaraguan Sign Language (NSL) that were recorded decades ago on analog tape and therefore do not encode depth information.

![Diagram of pipeline for extraction of sign-language metrics from video.](image)

**Figure 1:** Pipeline for extraction of sign-language metrics from video.

Traditionally, human coders fluent in a sign language have carried out linguistic analyses by eye, observing the physical properties of signed utterances in order to segment them into individual signs, transcribe meanings, and annotate phonetic properties. Tools such as ELAN [Crasborn & Sloetjes 2008] make this process easier. Our proposed pipeline can complement existing tools, such as ELAN, for annotating videos manually. For example, our pipeline could be used to automatically annotate measures of articulatory effort, such as the energy needed to make different signs, and measures of movement, such as symmetry, onset and offset times, repetitions, production time in specified zones, and ranges of motion.
Figure 2: Preliminary results. (A) Kinematic skeleton fitted to 3D points extracted from decades-old archival video of an NSL signer. (B) Time series showing estimated kinetic energy (log scale) for a signer of ASL, collected using the Kinect (depth-based camera). Kinetic energy is highest when the arms are moving quickly and lowest when the hands are still. Larger joints generate more kinetic energy than smaller ones.

References


