

3D-LEX v1.0

3D Lexicons for American Sign Language and Sign Language of the Netherlands

O. Ranum¹ , G. Otterspeer¹ , J.I. Andersen¹ ,
R.G. Belleman² , F. Roelofsen¹ 

University of Amsterdam

¹: Institute for Logic, Language and Computation, University of Amsterdam

²: Computational Science Lab, Informatics Institute, University of Amsterdam

oline.ranum@student.uva.nl, {g.otterspeer, j.andersen, r.g.belleman, f.roelofsen}@uva.nl

Abstract

In this work, we present an efficient approach for capturing sign language in 3D, introduce the 3D-LEX v1.0 dataset, and detail a method for semi-automatic annotation of phonetic properties. Our procedure integrates three motion capture techniques encompassing high-resolution 3D poses, 3D handshapes, and depth-aware facial features, and attains an average sampling rate of one sign every 10 seconds. This includes the time for presenting a sign example, performing and recording the sign, and archiving the capture. The 3D-LEX dataset includes 1,000 signs from American Sign Language and an additional 1,000 signs from the Sign Language of the Netherlands. We showcase the dataset utility by presenting a simple method for generating handshape annotations directly from 3D-LEX. We produce handshape labels for 1,000 signs from American Sign Language and evaluate the labels in a sign recognition task. The labels enhance gloss recognition accuracy by 5% over using no handshape annotations, and by 1% over expert annotations. Our motion capture data supports in-depth analysis of sign features and facilitates the generation of 2D projections from any viewpoint. The 3D-LEX collection has been aligned with existing sign language benchmarks and linguistic resources, to support studies in 3D-aware sign language processing.

Keywords: Sign Language, Computer Vision, Datasets

1. Introduction

Sign language processing (SLP) is a dynamic research area concerned with advancing computational methods for sign languages (SL). This multidisciplinary field encompasses tasks such as the automatic understanding, recognition, translation and production of sign language, contributing to a more inclusive future in language technology.

Despite receiving increased attention across computer sciences (Koller, 2020; Rastgoo et al., 2021), SLP remains less developed compared to other areas within Natural Language Processing (Yin et al., 2021). A significant factor contributing to this disparity is the lack of large-scale, high-quality, and publicly accessible sign language corpora (Bragg et al., 2019). Notably, the majority of these datasets are recorded with cameras that view signers from a single, (near-)frontal perspective (Ali et al., 2022). This scarcity of data impedes modern machine-learning algorithms from learning robust sign representations grounded in the three-dimensional nature of sign languages.

Literature supports that depth-awareness and viewing angle matters in both human (Watkins et al., 2024) and machine (Gao et al., 2023; Rastgoo et al., 2020) SL understanding. This implies that repre-

sentations should reflect a degree of 3D awareness, or risk reduced accuracy under normal real-world conditions, such as non-frontal viewpoints.

While systems such as OpenPose (Cao et al., 2021) enable the estimation of 3D poses from video

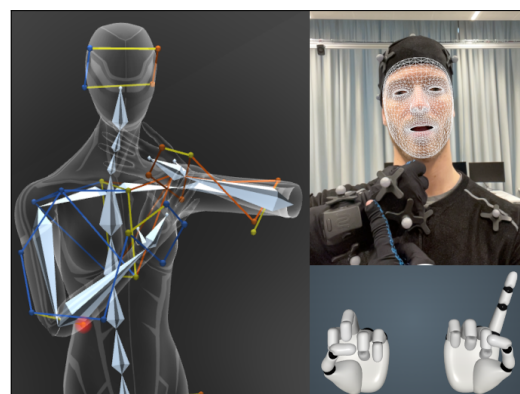


Figure 1: **Motion capture techniques:** The NGT sign *'mango'* captured with the three collection techniques. Left: Pose data captured with Vicon Motion Capture displayed in Shogun Live; Top right: face features captured with Live Link Face (Epic Games); Bottom right: handshapes captured with gloves displayed in Hand Engine (StretchSense).

footage, the precision of such reconstructions is in principle lower than the accuracy achieved through direct 3D motion capture techniques (Jedlička et al., 2020). Navigating imperfectly reconstructed 3D representations can pose significant challenges for downstream SLP tasks.

Providing a 3D ground truth to existing datasets could significantly improve the feasibility of many SLP tasks. Against this backdrop, we introduce 3D Lexicons (3D-LEX) for American Sign Language (ASL) and the Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT). The 3D-LEX datasets include 1,000 isolated signs from each language recorded with three distinct motion capture techniques, as illustrated in Figure 1. The vocabularies have been aligned with existing SL resources, including the WLASL (Li et al., 2020) and SEMLEX (Kezar et al., 2023) benchmarks for isolated sign recognition, the ASL-LEX 2.0 (Sehyr et al., 2021) lexicon and the SignBank NGT (SB NGT) lexicon (Crasborn et al., 2020). The 3D-LEX dataset facilitates the generation of 2D projections from any viewpoint and supports in-depth analysis of sign language features, offering several key advantages:

Automatic recognition of phonetic properties: High-resolution 3D data allows for detailed studies of sign language features, including handshapes, place of articulation, and orientations.

Multi-view SL recognition: Ground truth 3D representations facilitate the rendering of synthetic multi-view 2D data from any angle and translation. This data can be used to train models that are capable of *multi-view SL recognition*, a task that has received little attention in the SLP literature so far.

SL production for XR applications: Current work on SL production focusing on 2D outputs, such as synthetic photorealistic videos or 2D skeleton animations, are not directly suitable for Extended Reality (XR) applications. While reconstructing 3D motion from multiple 2D views is an area of active research, leveraging 3D data to produce 3D animations currently still offers a more effective and accurate approach.

The 3D-LEX v1.0 dataset was developed during our initial exploration of motion capture equipment for capturing three-dimensional sign representations. We acknowledge that the methodology outlined in Section 3 presents significant opportunities for improvement. Specifically, ensuring consistency in data quality will be a primary objective in our future efforts. Nevertheless, even in this nascent stage of development, we could demonstrate the utility of the 3D-LEX data. In Section 4 we showcase how the dataset can be leveraged to produce semi-automatic annotations of handshapes. Evaluating the annotations in a downstream isolated sign

recognition (ISR) task demonstrates that the labels achieved parallel benefits to leveraging annotations provided by linguists. We discuss several observed limitations and prospects for improvement in Section 5, and Section 6 highlights some ethical considerations.

2. Background

2.1. Sign Language

Sign languages are visual, complete, and natural languages, each with a distinct structure, grammar, and lexicon. They employ a combination of manual markers (e.g. handshapes, hand location, palm orientation and movements) and non-manual markers (e.g. mouthings, facial expressions, gaze) to convey meaning (Stokoe). Sign languages serve as the primary language in Deaf communities.

2.2. Sign Language Datasets

The majority of publicly available resources demonstrating sign language are captured in video. These datasets consist of either isolated signs (e.g. Sehyr et al., 2021; Athitsos et al., 2008; Kezar et al., 2023; Joze and Koller, 2019; Li et al., 2020) or continuous sign sentences (e.g. von Agris and Kraiss, 2010; Schembri et al., 2013). Key distinguishing features between the collections include the source language, signer variability, data scope, linguistic domain, and the availability and quality of annotations.

Most datasets comprise RGB video formats, but they may also include depth estimations or skeletal poses generated from joint approximations. While these datasets usually feature a single, (near-)frontal viewpoint, there is a growing trend in lab-curated datasets to provide multiple viewing angles (e.g. Duarte et al., 2020; Mopidevi et al., 2023; Rastgoo et al., 2020; Gao et al., 2023). Depth cameras have been used to capture 3D positioning, for example using the Kinect depth sensor (e.g. Oszust and Wysocki, 2013; Cooper et al., 2012; Huang et al., 2018). For an extensive summary of sign language datasets, refer to Kopf et al. (2022).

Datasets facilitating 3D awareness in sign representations either leverage depth estimations or 3D reconstruction techniques. For the creation of more precise 3D representations, numerous motion capture datasets have been curated (e.g. Lu and Huenerfauth, 2010; Heloir et al., 2006; Benchiheub et al., 2016), typically to generate signing avatars (Bragg et al., 2019) or for exploring automatic synthesis of sign language utterances (e.g. Jedlička et al., 2020; Gibet, 2018).

3. The 3D-LEX Dataset

3.1. Data Scope

The 3D-LEX v1.0 dataset includes lexical datasets sampled from ASL and NGT, where the scope was defined to ensure integration with existing benchmarks. A total of 1,000 signs are selected from each language, and recorded with two data collection techniques to capture manual markers and one technique to capture non-manual markers¹. We release three distinct data formats corresponding to the different capturing techniques, and one component integrating handshapes and body pose data.

Handshape Data The handshape(s) of each sign is captured with the StretchSense Pro Fidelity Motion Capture Gloves². The gloves measure the splay and bend of the fingers, alongside the relative rotation of each joint within the hand. The available data include the stretch sensor readings and exported FBX³ files. Detailed guidance on interpreting and assessing StretchSense data can be found in the project’s Git repository for data evaluation⁴.

Body Pose Data The place of articulation, movement, and body pose of each sign is captured using a Vicon (V) Motion Capture setup with optical markers. The raw marker location data is published, alongside processed FBX data, which has been exported via Shogun Post.

Face Blendshape Data Facial features are captured as blendshapes with the Live Link Face⁵ (LLF) application and ARKit on iPhone.

Retargeted Animation Data For sign language production and animation purposes, we release FBX files containing the body pose data and the handshapes.

3.2. Production Method

To efficiently capture the lexicons, we have developed a recording pipeline that achieves an average capture time of 10 seconds per sign. This includes the time for sign demonstration, performance, recording, and storage of the captured sign,

though it varies with the sign’s length. Setup preparations, which involve fitting the suit, positioning markers, and calibrations, require approximately 1 hour with our current method.

3.2.1. Recording Setup and Procedure

Our studio setup includes a designated detection zone for the Vicon cameras, an iPhone equipped with Live Link Face mounted on a tripod, one screen to display glosses and reference videos, and a second screen to showcase the recordings for immediate evaluation.

A triple-foot pedal system facilitates the remote operation of the motion capture control system. Each pedal is configured for a distinct function: The left pedal triggers the start and stop of recordings across all three motion capture systems simultaneously; the middle pedal stores the latest recording and issues a request to the SignCollect platform to display the next gloss in the vocabulary; and the right pedal is used to proceed to the next sign without saving any data. Signcollect is a platform developed to enable the efficient processing of glosses, providing a studio interface managed by gesture or pedal control. For details on the SignCollect platform consult [Otterspeer et al. \(2024\)](#).

The capture process for a single sign involves the following steps: First, the signer assumes an upright posture, with arms relaxed at their sides in a neutral position. By pressing the right pedal, a sign is prompted from the SignCollect platform, and the sign’s gloss and a reference video are displayed on one of the screens. A recording is started by pressing the left pedal, and the signer performs the sign and returns to the neutral stance before the recording is ended with another press of the left pedal. The recorded data is automatically exported to SignCollect and visualized on an avatar rendered with Unreal Engine v5.3, allowing the signer to immediately review the quality of the data. If the data’s quality is satisfactory, the signer can advance to the next gloss by pressing the right pedal, which saves the preceding recording. Should the sign’s execution be deemed inadequate, the signer can repeat the recording by pressing the left pedal again or proceed by pressing the right pedal. For visualizing the sign we created an avatar in *Ready Player Me Studio*, a cross-platform avatar generator that allows users to build avatars for general purposes.

A total of five signers contributed to capturing the ASL and NGT vocabularies. The signers were given two options to operate the pedal. Either they could control the pedal and capture process themselves, or they could delegate the pedal control to a team member. Preferences varied, with three signers opting for controlling the pedal themselves and two preferring assistance to concentrate on signing. Details regarding the number of words

¹The data is available under a [CC BY 4.0 license](#) at osf.io/g7u9c/?view_only=8090319e12aa4fd991d81e369a1cbd88

²stretchsense.com/mocap-pro-fidelity-glove-2/

³A 3D model file facilitating the transfer of animation data between various modeling applications including Maya, Blender, and Unreal Engine.

⁴github.com/OlineRanum/SAPA

⁵apps.apple.com/us/app/live-link-face/id1495370836

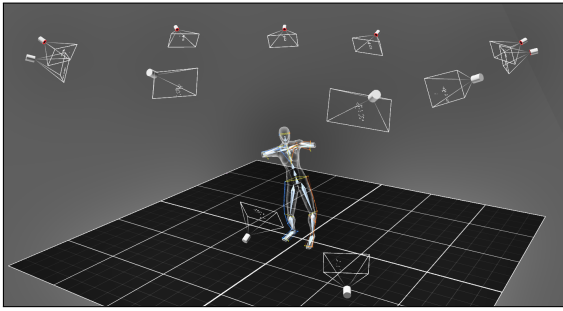


Figure 2: **Setup of the Vicon detection zone:** The illustration indicates the placement of the Vero Cameras on the rig and in front of the signer.

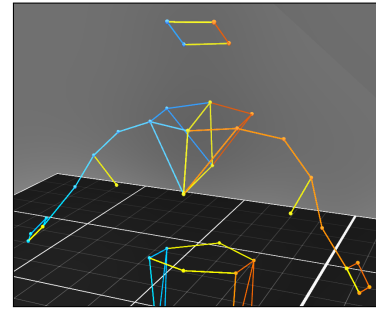


Figure 3: **Marker layout for the Vicon system:** Layout according to FrontWaist 53-marker set template, displayed on signer in Shogun Live.

recorded by each signer per language and pedal control preferences are provided in table 1.

The control system and comprehensive details about the pipeline are available on GitHub⁶. In the following paragraphs we describe each motion capture component in greater detail.

3.2.2. Vicon Motion Capture System

Setup: A Vicon rig is affixed to the ceiling, equipped with ten Vicon Vero v2.2 optical motion capture cameras⁷, as detailed in Figure 2. To mitigate occlusions, particularly those caused by the lower hands of the signer, an additional two Vicon Vero cameras are placed on the floor in front of the signer.

The markers are placed on the signer following the standard Vicon FrontWaist 53-marker set template⁸, as displayed in Figure 3. Shogun Post is used to make a retarget for the motion capture data, which is used during recording to stream the data to Unreal Engine from Shogun Live.

Calibration: For calibrating the Vicon camera system, we adhere to the built-in calibration protocol provided by Vicon. To ensure consistency in the calibration and that the origin remains approximately in the same position across multiple recording sessions, we place masking tape on the floor. This tape serves a dual purpose: one set of markings indicates the precise location for positioning the calibration wand during each calibration process. Another set of tape strips marks the designated spot where the signer is to stand during recordings.

Software Specifications: To manage the Vicon camera system, we utilize Shogun Live 1.11, and to perform the retarget of the motion capture data we use Shogun Post 1.11.

⁶github.com/OlineRanum/GLEX_Controller

⁷vicon.com/hardware/cameras/vero/

⁸docs.vicon.com/display/Shogun18/Create+subjects#Createsubjects-PlaceMarkersPlacemarkersonaperformer

3.2.3. StretchSense Gloves

Setup: The StretchSense Pro Fidelity gloves interface with Hand Engine Pro through two USB dongles, which are docked on a separate Dell Universal Dock (UD22) to ensure adequate power supply. Hand Engine is configured to receive remote triggering from Shogun Live, and to retarget animation data directly to Unreal Engine.

Calibration: The StretchSense Pro Fidelity gloves are calibrated using the calibration functionality of the Hand Engine software, which involves capturing pre-defined hand poses, to match the recorded output to an individual's hand. Our procedure combines general-purpose poses with specialized ones to customize the glove's fit for each user to capture sign language.

- i. **Express Calibration Poses:** Our general-purpose hand pose set corresponds to the express calibration poses available in Hand Engine, which comprises five common hand-shapes.
- ii. **Advanced Calibration Poses:** A more detailed hand pose library was developed, incorporating the most commonly occurring hand-shapes found in the 3D-LEX NGT (20 poses) and ASL (25 poses) vocabulary, as labeled by linguists in the aligned resources. The advanced pose libraries have been made accessible on GitHub.

We employ the training functionality of Hand Engine to fit the gloves' output data specifically to the signer. We configure all calibration poses to the blend pose mode, a Hand Engine feature that uses the calibration poses as landmarks in a continuous motion space, and interpolates between these poses to yield continuous outputs. The gloves are calibrated and retrained each time a signer puts them on to maintain accuracy.

Following initial consultations with StretchSense about employing the Pro Fidelity gloves for sign

language capture, we developed the specific number of poses and this calibration scheme. However, throughout the creation of 3D-LEX and subsequent discussions, it became evident that the calibration scheme was not ideal. We acknowledge this shortcoming and will reevaluate the calibration process in future works. For a discussion of these limitations and suggestions for potential improvements, please see Section 5.

Software Specifications: The StretchSense Pro Fidelity gloves are operated with the Hand Engine Pro software, version 3.0.6.

3.2.4. Live Link Face

Setup: An iPhone is mounted on a tripod, which is placed directly in front of the signer. Recordings are started, stopped, and saved automatically by the remote triggers.

Calibration: Live Link Face was not calibrated per signer. However, this functionality is available in the Live Link Face application and should be explored in a later version of the dataset.

Hard- and software Specifications: We use an iPhone 13 Pro and run Live Link Face version 1.3.2 with iPhone AR Kit.

3.3. Dataset Characteristics

The recording procedure introduces several recurring patterns into the raw data. Notably, the initial and final arm and hand positions often adopt a neutral stance, with the handshape closely resembling a ‘5’ handshape (refer to Figure 7). This results in, for instance, parts of the handshape recordings capturing signals that are not characteristic of a particular sign. This includes handshapes observed during the transition from a neutral state to the sign’s active posture, or when a sign involves a series of distinct handshapes, resulting in recordings that capture multiple pose signals within a single sign. An illustration of a typical temporal series according to the Euclidean distance is provided in Figure 5.

Data captured using LLF presents a non-uniform sampling rate, as frames are only recorded upon detected changes in the current state of the sensor. Conversely, the body poses captured with the Vicon system and handshapes captured with StretchSense are sampled uniformly.

The lexicons include a variety of handshapes. Figure 4.a showcases the distribution of handshapes in the ASL Lexicon, annotated by sign language linguists in the ASL-LEX resource.

Signer characteristics All participants are native signers, who acquired sign language from an early age. Details about each signer’s primary language,

along with their preferences for operating the pedal, are provided in Table 1.

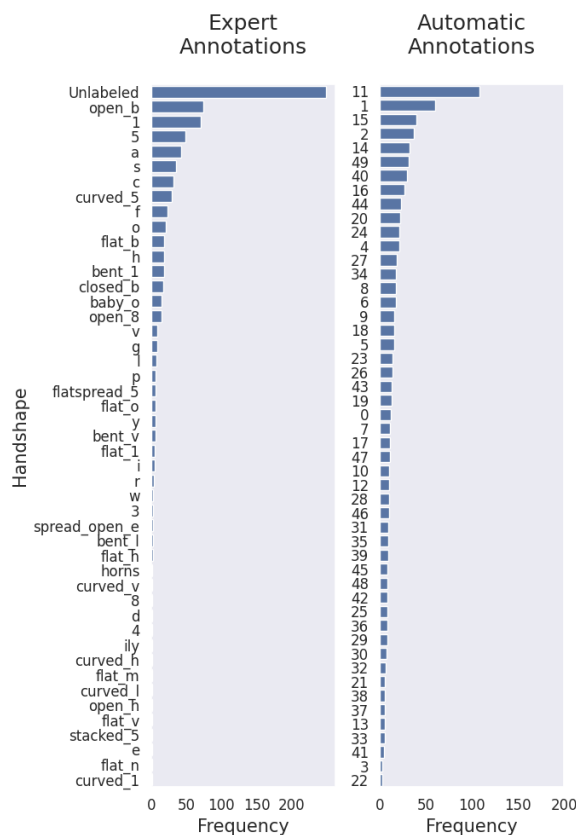


Figure 4: **Distributions of handshapes in the 3D-LEX vocabulary:** the distribution of handshapes as identified by (a) human experts and (b) the automated annotation process described in Section 4.1. The automatic annotations assign arbitrary cluster IDs to different groups of handshapes determined through a K-means clustering method. It’s important to note that these handshape cluster IDs may not directly correspond to the linguistic labels used by human experts in Subfigure 4.a.

Signer ID	01	02	03	04	05
Native Language	NGT	NGT	NGT	NGT	ASL
NGT Signs	10	400	590	0	0
ASL Signs	155	12	0	644	189
Pedal Control	YES	YES	YES	NO	NO

Table 1: **Signer Characteristics:** Native background of each signer and preference for operating (YES) or delegating (NO) the control of the pedal.

Alignment with existing SL resources The vocabularies of 3D-LEX have been aligned with existing SL resources to promote research integrating 3D data with datasets comprised of video data and linguistic databases. Table 2 lists the number of glosses in 3D-LEX overlapping with the vocabularies of the aligned resources, the number of sign pose estimations from example videos available for the glosses in the datasets, and the number of glosses that have been provided with expert human annotations for the dominant hand.

The 3D-LEX ASL vocabulary was selected to ensure that a minimum of five reference videos per sign are available in each ASL dataset. Currently, no dataset with multiple reference videos per gloss exists for NGT, but we anticipate that this situation will change in the future. Currently, the SB NGT lexicon (Klomp et al., 2024; Crasborn et al., 2020) provides one reference video for each gloss in the 3D-LEX NGT vocabulary.

	SEMLEX	WLASL	SB NGT
	ASL	ASL	NGT
Vocabulary	1,000	1,000	1,000
Reference Videos	49,274	12,051	1,000
Expert HS	921	695	888

Table 2: **Alignment with other datasets:** The vocabulary overlap, the number of available reference videos, and the number of available expert handshape annotations for the 3D-LEX vocabulary in the SEMLEX, WLASL, and SB NGT datasets.

4. Evaluation

To demonstrate the utility of 3D sign data we turn to one of the envisioned benefits mentioned initially: the facilitation of automatic phonetic labeling. In particular, we present a baseline method for semi-automatic handshape annotation. The efficacy of the annotations is evaluated in an ISR task, through comparison with labels provided by linguists and against scenarios devoid of any labels.

While we expect that the data can be used to label other phonetic properties (*e.g.* hand location, movement, orientations, eyebrow position) we here zoom in on the handshapes. This is an intentional choice, as we consider the use of StretchSense gloves to be the most experimental data acquisition technique for sign language capture. The development of semi-automatic annotation methods benefits both linguistic research and various SLP tasks, including recognition and production.

4.1. Semi-Automatic Handshape Annotations

In this section, we demonstrate one simple approach for generating phonetic annotations derived from the 3D-LEX handshape data. Due to the absence of an NGT benchmark for isolated sign recognition, we only generate and assess labels derived from the 3D-LEX ASL vocabulary.

Our approach is designed to produce labels that resemble the handshape annotations typically found in ISR benchmarks, facilitating a meaningful comparison between automated and expert annotations. The glosses in ISR benchmarks are commonly assigned a single handshape label, based on the dominant handshape observed in a single reference video. We ensure that the number of possible label classes in our estimations corresponds approximately to the set of classes identified in the video-data benchmark WLASL. For the implementation and instructions on how to replicate our findings, please refer to the GitHub repository.

Temporal segmentation To differentiate characteristic handshape signals from any resting or transitional poses, we construct a temporal segmentation method by calculating the Euclidean distance to each frame relative to the calibration poses. This method enables us to perform a first-order discrimination of signals within a recording.

We estimate and segment the poses of both hands to take into consideration that the signer may not strictly enforce the use of their dominant hand. Subsequently, we calculate the frequency of each observed handshape and select the handshape with the highest frame count. As the typically most frequent signal is the resting pose '5', we only select the '5' handshape if it is detected in more than 90% of the frames, otherwise, we select the second most frequently occurring class. The frames where the dominant handshape was detected are then selected as candidate frames for downstream analysis. Figure 5 showcases the output of a Euclidean distance handshape classification approach on frames from the captured sign 'zero'. Here, the handshape 'o' was identified as the characteristic handshape of the sign.

Semi-automatic labeling The Euclidean distance labeling technique limits the identification of handshapes to those poses used during the glove calibration phase. This is suboptimal, as the calibration methodology of stretch sensors for capturing sign language is still in a nascent stage. Specifically, the calibration poses may not cover the full range of handshapes present in the lexicons.

To enable a more flexible identification of handshapes, we applied k-means clustering on the average poses of the frames selected during the temporal segmentation. We selected $k=50$, which is

Temporal Segmentation of EDs Poses

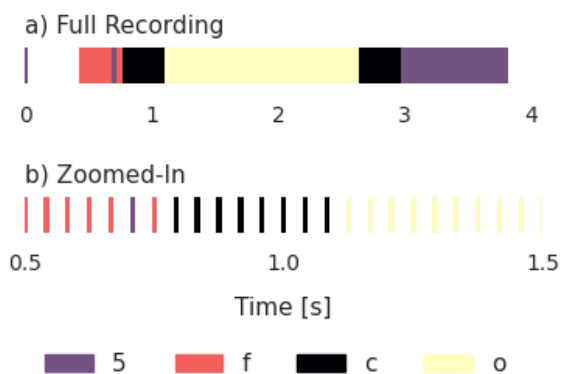


Figure 5: **Time-series visualization of handshape classification:** Classification of the ASL sign 'zero', labeled by experts with the handshape 'o'. Frames are captured and displayed as bars, and each bar's color indicates the handshape, determined by applying the Euclidean distance method frame-by-frame. White space indicates that no data was recorded at that time. The timeline, marked on the x-axis, spans four seconds for this sign. A detailed view at the 1-second mark is provided in the lower row for closer inspection. Our segmentation pipeline identifies the handshapes '5', 'f', 'c', and 'o', selecting frames corresponding to 'o' as the characteristic signal of 'zero'.

approximately the number of handshapes identified in ASL-LEX for the 3D-LEX vocabulary. We assign a new handshape label to each sign in 3D-LEX ASL, corresponding to the arbitrary cluster IDs assigned while clustering the high-dimensional features.

Figure 6 presents a t-SNE projection into two dimensions of the average hand poses, demonstrating that the high-dimensional features cluster. This implies that the signals from the gloves carry sufficient information to distinguish between different handshapes in sign language, revealing distinct characteristics for clusters of signs.

Evaluation of annotations To evaluate the efficacy of our annotations, we employed the OpenHands framework (Selvaraj et al., 2022). More precisely, we adopted the framework's adaptation as implemented by Kezar et al. (Kezar et al., 2023), which facilitates gloss recognition supported by phonetic properties. Their foundational work demonstrated that training with phonetic labels enhances gloss recognition accuracy, by merging the WLASL benchmark with the expert linguistic descriptions provided by the ASL-LEX dataset.

In our evaluation process, we trained an SL-GCN (Jiang et al., 2021) architecture to predict glosses within the WLASL dataset, where we use the subset

t-SNE Projection of Handshapes

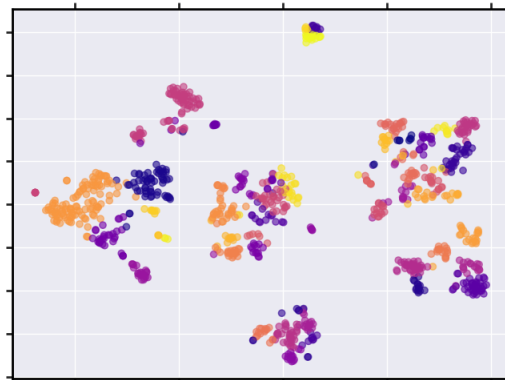


Figure 6: **t-SNE projection:** A t-SNE projection of average hand poses into two dimensions, where the poses were averaged across temporal segments of each sign determined by the Euclidean segmentation method. The projection space lacks units and aims solely to illustrate how high-dimensional 3D-LEX handshape features cluster, highlighting distinguishable signals. Each color represents one of 50 k-means cluster IDs, serving merely to aid visual differentiation of the clusters.

of the WLASL data which overlaps with the 3D-LEX vocabulary. Training persists until validation accuracy ceases to improve for 30 consecutive epochs. The distribution of files across the training, validation, and test splits utilized in our experiments is detailed in Table 3.

To provide a baseline for comparison, we trained the SL-GCN to predict glosses both with and without leveraging handshape labels from the ASL-LEX. Subsequently, we substituted the ASL-LEX handshape labels with our semi-automatic annotations and retrained the models to undertake gloss recognition supported by our annotations. This approach facilitates a comparison of our semi-automatic annotation method against human expert annotations, in terms of their ability to support learning in a downstream ISR task.

Train	Val	Test
8209	2174	1668

Table 3: **Train-Val-Test splits:** Number of examples in the Train-Val-Test splits for the WLASL benchmark experiments.

Results The outcome of our isolated sign recognition experiment using semi-automatic handshape labels is presented in Table 4. We provide the top-1 recognition accuracy on the test set, meaning the ratio of how often the model predicted the correct

gloss as the most likely label for a video amongst 1,000 classes. As can be observed, the automatic annotations perform on par with annotations provided by linguistic experts. This is an indication that high-resolution 3D data can offer to reduce the costs associated with linguistic annotation of signs in video datasets and that StretchSense signals are adequate to capture essential handshape features in signs.

a_1^N	a_1^E	a_1^A
0.44 \pm 0.01	0.48 \pm 0.01	0.49\pm0.01

Table 4: **Top-1 recognition accuracy:** Accuracy using no (N) handshape labels, expert (E) labels and automatic (A) labels. The accuracies are averaged across 8 runs, and the standard deviation across measurements is provided in the subscripts.

5. Limitations and Prospects

In the process of capturing our data, we have observed many potential areas for improvement. In this section, we highlight some of the current limitations in our methodology, and our intent for addressing them in future work.

Like numerous datasets in sign language research, a significant limitation of 3D-LEX is signer diversity. A dataset comprising a single example for each sign, and which contains only five signers, is insufficient for representing the diversity and rich prosody inherent to sign languages. It is as such not possible to use 3D-LEX in isolation to learn representations useful in sign applications. Consequently, 3D-LEX can primarily serve for limited feature studies or to support video datasets by either providing a 3D ground truth or synthesizing multi-view 2D data from one signer. Future work should consider exploring 3D data which includes both multiple examples per signer and multiple signers per gloss.

While all participants were native signers, it is critical to highlight that only one had ASL as their primary language. As a result, a significant segment of the 3D-LEX ASL dataset was produced by signers whose primary language is NGT but who were proficient in ASL. The impact of employing signers whose primary sign language differs from the captured target language, on the quality and authenticity of lexical sign data remains an area for future research. This concern is recognized as a limitation in v1.0 of 3D-LEX.

The dataset has a limited scope, which comprises a non-exhaustive set of phonological features and vocabularies from the complete languages. However, our method facilitates the pro-

duction of larger vocabularies and data for additional sign languages.

We observed several limitations in our current pipeline. While experimenting with the data acquisition control we noticed varying preferences among signers for operating the pedal. The choice of operator resulted in the emergence of several distinct patterns within the data. When signers themselves operate the pedal, it's generally more efficient but introduces a signal from foot movement at the start and end of each sign. Conversely, using an external operator can result in greater variability in the timing of recordings, affecting the consistency of the recorded time window around each sign. Efforts to streamline these production elements are anticipated in future work.

While our system has been designed with a focus on efficiency, we have identified several limitations concerning the hardware. To the best of our knowledge, 3D-LEX is the first publicly available dataset using the StretchSense gloves to conduct statistical analysis on handshapes in sign language. These gloves were initially developed to generate animation data, which typically does not require the same degree of accuracy in capturing detailed, varied and intricate movements of fingers and hands. Therefore, employing these gloves to provide detailed studies of handshapes in sign language represents a novel and experimental approach. Although the gloves have shown promising capabilities, their performance has presented several challenges.

Notably, the precision of the gloves' measurements is closely tied to how well they fit the signers' hands and the length of time they are worn. A snugger fit typically leads to higher accuracy. However, prolonged usage has been observed to decrease accuracy, likely due to the glove's position shifting on the hand, thereby deviating from its calibrated stance. Shifts can occur for example when hands swell from accumulated heat and from natural movements during wear. Larger gloves relative to the hand size are more prone to positional shifts, exacerbating this issue.

The Hand Engine software is prone to overfitting the sensor data to the calibration poses, a tendency that amplifies when training involves an extensive calibration pose set. Currently, the calibration process utilizes either 20 or 25 poses. We observed that such a detailed pose repertoire complicates Hand Engine's ability to accurately replicate more complex poses and distinguish between poses where the shift in stretching values are relatively small. Figure 7 illustrates a series of poses that exhibit substantial differentiation challenges for the gloves under our calibration framework. With the current version of Hand Engine, future research may gain advantages from employing a smaller set of calibration poses. Ideally, these selected hand-

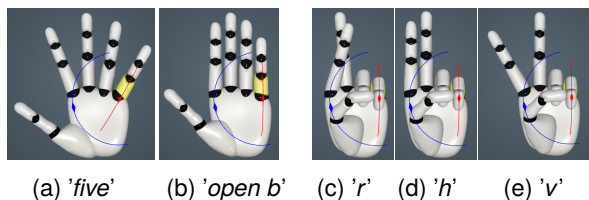


Figure 7: **Failure modes using the StretchSense gloves:** Example handshapes that are challenging to discern for the gloves, conditioned on our calibration scheme. The gloves struggled to differentiate between the handshapes 'five' and 'open b', and between the handshapes 'r', 'h' and 'v'.

shapes should not only be representative of those within the dataset but also exhibit maximum distinction from each other within the calibration set.

An in-depth assessment of calibration methods to address overfitting issues warrants further exploration. This becomes especially critical in capturing continuous signing, where the range of anticipated handshapes is far more variable and unpredictable than in lexical datasets. The 3D-LEX team is actively engaging with StretchSense to enhance glove calibration for sign languages, focusing on better support for continuous signing and capturing a broader spectrum of handshapes. The gloves' ability to accurately represent signing is contingent upon the calibration process, however, as this is a software concern, we expect the conditions for continuous signing to improve in later versions of the Hand Engine software.

Upon assessing the Vicon data, we identified several artifacts occasionally occurring in recordings. For example, we observed random hand orientation flips, which can be attributed to occlusions, where the cameras lost clear line-of-sight to the hands. In such instances the markers may be mistaken for each other, causing the palm to rotate when displayed on an avatar. To mitigate this issue, one can attempt to optimize the positions of the cameras standing on the floor or apply post-processing techniques, such as the filter and gap solver functionalities available in Shogun Post, or by re-labeling the swapped markers.

Moreover, due to limited time, we could not assess the data generated by the LLF application in detail. However, we observed considerable variation in the use of markers like mouthing cues and gaze among participants. In our future research, we aim to delve into these patterns and thoroughly evaluate the quality of the facial feature data.

In our evaluations of 3D-LEX, we presented a basic approach to deriving annotations. However, we emphasize that signs are complex and may contain transitions or oscillate between multiple characteristic handshapes throughout the execution of a sign. While our method approximates the dominant hand-

shape, there are potential benefits in deriving more sophisticated annotation strategies, which consider these transitions and oscillations, and potentially provide multiple phonetic properties for the handshape per sign. However, it is noteworthy that, even in the nascent stages of developing the 3D-LEX production methodology, our automatic annotations yield benefits comparable to those derived from leveraging annotations provided by experts.

6. Privacy and Ethical Considerations

The success of machine learning methods has led to large increases in requests for data. While this implies heightened concerns for privacy across computational sciences, it is important to recognize that data collection from minority language communities is at particular risk: Both because a status as deaf classifies as sensitive information, but also because data collection from small populations limits anonymity (Bragg et al., 2020). Additionally, certain sign language datasets that are publicly accessible were compiled without obtaining informed consent from the individuals featured, particularly those datasets that gather information from platforms such as YouTube. All signers contributing to the production of 3D-LEX gave informed consent and received compensation. Moreover, the anonymity of contributors is enhanced compared to typical video datasets, since the motion capture recordings do not visually reveal the signers. To further protect signer anonymity, each participant has been assigned a unique signer ID.

7. Conclusion

In this paper, we introduce a new and efficient method for collecting 3D sign language data, resulting in the 3D-LEX dataset, and describe a semi-automatic approach for producing phonetic annotations. The 3D-LEX dataset was produced leveraging three distinct motion capture systems, with two collection techniques to capture manual markers and one technique to capture non-manual markers. Although our approach shows considerable room for improvement, we highlight its potential by automatically generating handshape labels for 1,000 ASL signs. Our initial evaluations of the labels on a downstream ISR task reveal that the semi-automatic annotations offer benefits parallel to those of expert annotations. In conclusion, the 3D-LEX v1.0 demonstrates considerable potential even in its early stages of development. We anticipate that future research using 3D-LEX will investigate synthesizing multi-view data from the 3D ground truths to support tasks such as multi-view SLR, and develop approaches annotating additional phonetic classes.

8. Acknowledgments

The team behind 3D-LEX consisted of both Deaf and hearing researchers, whose participation in the project was made possible through financial support from the Platform Digital Infrastructure for the Social Sciences and the Humanities (PDI-SSH) and the Netherlands Organization for Scientific Research (NWO). We extend our gratitude to all external participants who assisted in gathering the data, and to the [Visualisation Lab](#) and [SignLab](#) at the University of Amsterdam for generously providing us with their facilities and equipment. In addition, we wish to thank the company ProCare⁹, who performed the setup of our Vicon rig.

9. Bibliographical References

- SK. Ashraf Ali, M. V. D. Prasad, P. Praveen Kumar, and P. V. V. Kishore. 2022. [Deep multi view spatio temporal spectral feature embedding on skeletal sign language videos for recognition](#). *International Journal of Advanced Computer Science and Applications*, 13(4).
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Brafport, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. [Sign language recognition, generation, and translation: An interdisciplinary perspective](#). In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.
- Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. [Exploring collection of sign language datasets: Privacy, participation, and model performance](#). In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA. Association for Computing Machinery.
- Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. 2021. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186.
- Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3413–3423.
- Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023. [Improving sign recognition with phonology](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia. Association for Computational Linguistics.
- Oscar Koller. 2020. [Quantitative survey of the state of the art in sign language recognition](#). *ArXiv*, abs/2008.09918.
- Gomèr Otterspeer, Ulrika Klomp, and Floris Roelofsen. 2024. Signcollect - a 'touchless' pipeline for constructing large-scale sign language repositories. In *LREC-COLING 2024 - 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy.
- Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3451–3461.
- Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. [OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.
- William C. Stokoe. [Sign language structure: An outline of the visual communication systems of the american deaf](#). *The Journal of Deaf Studies and Deaf Education*, 10(1):3–37.
- Freya Watkins, Diar Abdikarim, Bodo Winter, and Robin L. Thompson. 2024. [Viewing angle matters in british sign language processing](#). *Scientific Reports*, 14(1).
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

⁹ProCare BV, The Netherlands

10. Language Resource References

Language Resources

- Athitsos, Vassilis and Neidle, Carol and Sclaroff, Stan and Nash, Joan and Stefan, Alexandra and Quan Yuan and Thangali, Ashwin. 2008. *The American Sign Language Lexicon Video Dataset*. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.
- Benchiheub, Mohamed-El-Fatah and Berret, Bastien and Braffort, Annelies. 2016. *Collecting and Analysing a Motion-Capture Corpus of French Sign Language*. Workshop on the Representation and Processing of Sign Languages.
- Helen Cooper and Eng-Jon Ong and Nicolas Pugeault and Richard Bowden. 2012. *Sign Language Recognition using Sub-Units*. Journal of Machine Learning Research.
- Crasborn, Onno and Bank, Richard and Zwitserlood, Inge and van der Kooij, Els and Ormel, Ellen and Ros, Johan and Schüller, Anique and de Meijer, Anne and van Zuilen, Merel and Nauta, Yassine Ellen and van Winsum, Frouke and Vonk, Max. 2020. *NGT dataset in Global Signbank*. Radboud University, Centre for Language Studies, ISLRN 976-021-358-388-6.
- Amanda Cardoso Duarte and Shruti Palaskar and Deepti Ghadiyaram and Kenneth DeHaan and Florian Metze and Jordi Torres and Xavier Giró-i-Nieto. 2020. *How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language*. ISLRN 583-408-694-292-6.
- Gao, Liqing and Zhu, Lei and Xue, Senhua and Wan, Liang and Li, Ping and Feng, Wei. 2023. *Multi-View Fusion for Sign Language Recognition through Knowledge Transfer Learning*. Association for Computing Machinery, VRCAI '22.
- Gibet, Sylvie. 2018. *Building French Sign Language Motion Capture Corpora for Signing Avatars*.
- Heloir, Alexis and Gibet, Sylvie and Multon, Franck and Courty, Nicolas. 2006. *Captured Motion Data Processing for Real Time Synthesis of Sign Language*. Springer Berlin Heidelberg.
- Huang, Jie and Zhou, Wengang and Zhang, Qilin and Li, Houqiang and Li, Weiping. 2018. *Video-based sign language recognition without temporal segmentation*. AAAI Press, AAAI'18/IAAI'18/EAAI'18.
- Jedlička, Pavel and Krňoul, Zdeněk and Kanis, Jakub and Železný, Miloš. 2020. *Sign Language Motion Capture Dataset for Data-driven Synthesis*. European Language Resources Association (ELRA).
- Hamid Reza Vaezi Joze and Oscar Koller. 2019. *MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language*. BMVA Press.
- Kezar, Lee and Thomason, Jesse and Caselli, Naomi and Sehyr, Zed and Pontecorvo, Elana. 2023. *The Sem-Lex Benchmark: Modeling ASL Signs and their Phonemes*. ACM, ASSETS '23.
- Klomp, Ulrika and Gierman, Lisa and Nauta, Ellen and Otterspeer, Gomèr and Pelupessay, Ray and Stern, Galya and Wubbolts, Casper and Oomen, Marloes and Roelofsen, Floris. 2024. *An extension of the NGT dataset in Global Signbank*. SignBank.
- Kopf, Maria and Schulder, Marc and Hanke, Thomas. 2022. *The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources*. European Language Resources Association.
- Li, Dongxu and Rodriguez, Cristian and Yu, Xin and Li, Hongdong. 2020. *Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison*. The IEEE Winter Conference on Applications of Computer Vision.
- Lu, Pengfei and Huenerfauth, Matt. 2010. *Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research*. Association for Computational Linguistics.
- Mopidevi, Suneetha and Prasad, M.V.D. and Kishore, P.V.V. 2023. *Multiview meta-metric learning for sign language recognition using triplet loss embeddings*. Pattern Analysis and Applications.
- Mariusz Oszust and Marian Wysocki. 2013. *Polish sign language words recognition with Kinect*. 2013 6th International Conference on Human System Interactions (HSI).
- Razieh Rastgoo and Kourosh Kiani and Sergio Escalera. 2020. *Hand sign language recognition using multi-view hand skeleton*. Expert Syst. Appl.
- Adam C. Schembri and Jordan B Fenlon and Ramas Rentelis and Sally Reynolds and Kearsy Cormier. 2013. *Building the British Sign Language Corpus*. University of Hawaii Press.
- Sehyr, Zed Sevcikova and Caselli, Naomi and Cohen-Goldberg, Ariel M and Emmorey, Karen. 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.

von Agris, Ulrich and Kraiss, Karl-Friedrich. 2010. *SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition.* European Language Resources Association (ELRA).