



Preprocessing MediaPipe Joint Annotation for Sign Language Similarity Analysis

Kehina Manseri

Université de Lorraine, CNRS, Inria,
LORIA
F-54000 Nancy, France
kehina.manseri@inria.fr

Sam Bigeard

Université de Lorraine, CNRS, Inria,
LORIA
F-54000 Nancy, France
esbig Beard@gmail.com

Slim Ouni

Université de Lorraine, CNRS, Inria,
LORIA
F-54000 Nancy, France
Slim.Ouni@loria.fr

Abstract

This paper introduces a preprocessing pipeline for keypoints extracted using MediaPipe, aiming to improve pose annotation consistency in sign language datasets. We evaluate its effectiveness using a sign similarity task based on phonological features, without relying on gloss annotations. Similarity is measured using Dynamic Time Warping (DTW) across videos from sign language dictionaries. Although such similarity analyses can support various sign language processing applications - such as lexical search, clustering, and data enrichment - the main contribution of this work is to standardise pose features across heterogeneous sources, including different signers and backgrounds. Experiments on two dictionary datasets show that our pipeline significantly improves similarity measurements, with promising benefits for other sign language processing tasks.

Keywords

Sign language, Similarity, MediaPipe, Pose estimation, Dynamic Time Warping, Principal Component Analysis, Uniform Manifold Approximation and Projection, Dimensionality Reduction, Normalisation, Handedness

ACM Reference Format:

Kehina Manseri, Sam Bigeard, and Slim Ouni. 2025. Preprocessing MediaPipe Joint Annotation for Sign Language Similarity Analysis. In *ACM International Conference on Intelligent Virtual Agents (IVA Adjunct '25)*, September 16–19, 2025, Berlin, Germany. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3742886.3756716>

1 Introduction

Sign Language (SL) processing faces a persistent challenge: the scarcity of large, high-quality annotated datasets [7, 15]. This limitation hampers the generalisation capabilities of models, especially when confronted with variation across signers, body types, signing styles and recording conditions [17]. As a result, careful preprocessing is crucial to diversify the possible uses of the available data and to reduce linguistically irrelevant variability.

Sign identification - the task of matching a given sign video to a reference entry in a database - is commonly approached through supervised learning, using annotated datasets where glosses or subtitles guide representation learning across multiple examples [1].

These annotated instances often feature consistent labels but differ in signer identity, camera viewpoint, or sentence context. In contrast, our work focuses on sign identification within SL dictionaries, where such annotations are either minimal or unavailable.

We use the Multilingual Sign Language WordNet [25], a meta-dataset of several European SL dictionaries [3, 6, 8, 9, 19, 28]. Our objective is to identify phonetically identical signs across languages, independent of their meaning. This excludes the use of glosses, keywords, or any metadata that could help identify the sign in any way. Consequently, supervised learning techniques are not applicable, and we instead emphasise preprocessing to standardise visual input.

To guide and validate our approach, we use dimensionality reduction and 2D visualisation techniques [18]. These tools allow us to observe the structure of the feature space and evaluate whether phonetically similar signs cluster more coherently as preprocessing is refined. This qualitative feedback is very useful in revealing dataset biases and inconsistencies that affect identification performance.

Our contributions are threefold. First, we demonstrate the role of 2D visualisations for diagnosing and refining SL data preprocessing pipelines. Second, we propose a preprocessing strategy that standardises keypoint-based pose representations with the aim to enable signer- and context-independent sign comparison, without requiring annotated training data or language-specific adaptation. Finally, we evaluate our pipeline on a sign similarity task across two SL dictionaries, using Dynamic Time Warping (DTW) [21] over 2D keypoints extracted by MediaPipe [16], a lightweight framework for real-time pose estimation. Each experiment uses 50 keypoints per frame, including 8 upper-body joints (nose, neck, shoulders, elbows, wrists) and 21 landmarks per hand capturing detailed articulatory features relevant to sign production.

2 Related Work

2.1 Features for sign language processing

In action recognition, two primary types of features are commonly used: RGB-based features and pose estimation annotations [23]. RGB features capture information related to colour, texture, and shape, while pose estimation models extract skeletal keypoints from the body. Similar feature types are used in Sign Language Recognition (SLR), where hand keypoints have been shown to significantly improve performance, as they encode essential information for SL production [20, 26]. Moreover, pose-based representations are generally easier to normalise and transfer across signers, making them particularly well-suited for unsupervised or weakly supervised settings.



This work is licensed under a Creative Commons Attribution 4.0 International License. *IVA Adjunct '25, Berlin, Germany*

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1996-7/25/09
<https://doi.org/10.1145/3742886.3756716>

2.2 Limitations of pose estimation models

Despite their usefulness, pose estimation models are not without limitations. MediaPipe [16] often produces noisy or inconsistent keypoint annotations, especially for the hands, which are highly articulated and subject to occlusion, motion blur, and low video resolution [24, 27]. Studies have proposed correction techniques to counter these issues, including temporal interpolation across frames [24] and backpropagation-based refinement using limb proportions and joint angle initialisation [27]. Although pose estimation models are lighter than RGB-based approaches and are robust to variations in background or lighting, they still preserve certain biases from the input videos, such as signer-camera distance, framing, or individual anatomy. Normalisation techniques are therefore widely used to improve consistency. For instance, Fragkiadakis et al. [10] apply anchor-based normalisation to adjust body proportions. They also applied referential transformations to reduce camera-dependent effects and emphasise internal configurations. Boháček and Hrůz [2] proposed normalising hand keypoints relative to the wrist, rather than to the full body, using bounding boxes. Their method improved SLR accuracy by 14% demonstrating the impact of well-designed normalisation strategies.

2.3 Dimensionality Reduction for Sign Language

Previous studies have applied dimensionality reduction techniques to visualise signs and enhance SLR performance. For instance, Gao et al. [11] and Haque et al. [12] used Principal Component Analysis (PCA) on RGB features extracted from images of finger-spelled signs. PCA finds features within the data that contribute the most to the overall variance of the sample. The data is then projected onto the axes exhibiting the most variance, allowing for it to be visualised in 2 or 3 dimensions.

In Fragkiadakis et al. [10], both PCA and UMAP [18] are employed to assess sign similarity based on keypoints extracted using OpenPose [4]. Each video is represented as a flattened array, created by concatenating the coordinates from each frame, which serves as a row in the input matrix for dimensionality reduction. This representation sidelines the temporal aspect of the videos, focusing instead on global relationships within the sign. The results in Fragkiadakis et al. [10] show that UMAP outperforms PCA for similarity analyses. UMAP projects each video onto a two-dimensional space by iteratively relocating points to preserve the structure of the original high-dimensional data. The resulting graphs demonstrate that signs are clustered according to their overall shape, effectively grouping visually similar signs together.

2.4 Sign Language Similarity Analyses

Regarding the evaluation task, previous studies have approached the measurement of sign phonological similarity using RGB features and labels. In Bilge et al. [1], unseen signs are associated with sign classes based on similarity between the textual embeddings of their phonological descriptions. CNN, 3D-CNN and LSTM were used to encode spatio-temporal information, while BERT was used for textual embeddings. Jungsoo Shin [14] proposed a pipeline to group signs based on their hand configurations using image-based

RGB representations. Clustering algorithms were used to classify samples.

In regard to metrics, various studies used DTW [21], an alignment algorithm allowing for time series to be locally shifted, contracted, and stretched. DTW searches for the temporal alignment that minimises the most distances between series. It is therefore adapted to video and action comparison tasks where length may vary across all data points. It has been previously used and proved to act as a reliable metric to assess similarity between extracted keypoints in SL videos [10, 13]. Euclidean distance has been used in other studies as a similarity metric. In Fragkiadakis et al. [10], distances between sign videos projected onto a 2D space have been shown to perform as well as DTW, and sometimes better when used on specific body parts. Claassen [5] used embeddings from the SL transformer architecture SPOTER [2], based on MediaPipe keypoints, to cluster signs using a silhouette score.

However, most of these studies use annotated datasets for similarity analyses and evaluation. Few studies, notably those involving inverse search technologies, use non-annotated datasets. Inverse search dictionaries allow users to submit a query video and return the most similar signs based on phonological features. These studies notably highlighted limitations in existing dictionaries that require users to manually specify their handedness or manual parameters involved in the query sign [10, 13].

3 Datasets

3.1 Multilingual Sign Language WordNet



Figure 1: Samples from CDPSL and Nederlandse Gebarentaal

Experiments and analyses were conducted on a subset of the Multilingual Sign Language WordNet Dataset [25], which is a meta-dataset linking dictionary-style entries from six dictionaries covering different sign languages¹: BSL SignBank (British Sign Language, or BSL) [9], Noema dictionary (Greek Sign Language, or GSL) [8], Nederlandse Gebarentaal (Sign Language of the Netherlands, or NGT) [6], Svenskt teckenspråkslexikon (Swedish Sign Language, or STS) [19], the Corpus-based Dictionary of Polish Sign Language (Polish Sign Language, or PJM) [28] and Gebärdensprach-Datenbank (Swiss-German Sign Language, or DSGS) [3]. Combined,

¹Only the datasets for which we gained authorisation are considered.

these datasets comprise a total of 10 321 videos. Figure 1 shows screenshots of typical entries in these dictionaries. With the exception of Noema and Gebärdensprach-Datenbank, each individual dataset includes multiple signers and a variety of backgrounds. While some datasets provide various annotations, for example, hand configuration, the diversity and heterogeneity of the sources make it impossible to obtain consistent or complete phonological annotations across the meta-dataset.

3.2 ASLLVD



Figure 2: Samples from ASLLVD

To generalise and evaluate preprocessing methods, we also made use of the ASLLVD dataset [22], whose characteristics are well-suited to our study. Like the WordNet dataset, ASLLVD consists of dictionary-style videos with similar length, signer variation, and framing. Moreover, the dataset includes annotations regarding hand configurations and the number of arms involved in the sign; that is to say, if the sign is one-handed or two-handed. All videos were cropped to preserve a frontal view, and the final dataset contains 7,963 videos. Figure 2 shows screenshots from typical entries in this dataset.

4 Bias Identification

Performance of sign language processing tasks, including the one addressed in this paper, can be affected by various factors throughout the processing pipeline. Previous studies [24, 27] have reported issues not only with pose estimation frameworks, but also with the quality of the input videos, as well as the diversity of interpreters and recording conditions. To better understand these limitations and potential biases, we applied dimensionality reduction techniques to project videos onto a 2D space. This allowed us to quickly identify clusters and investigate their causes. Such factors must be taken into account to standardise the dataset and ensure that our comparison methods focus on the linguistic properties of signs, rather than source-dependent attributes such as signer variation or inaccuracies in keypoint detection.

4.1 Sign Visualisation

We applied UMAP to 511 videos from the WordNet dataset using joint variances as input. Using a global descriptor for each video and

each joint allowed to apply UMAP on videos of different lengths, preventing sampling or padding. Each row in the matrix is of length 100, twice the amount of extracted keypoints (x and y axes). The result is shown in figure 3.

4.2 General Observations

Unlike PCA, UMAP does not use axes to represent interpretable components, clusters must therefore be analysed individually.

We observed that while the clusters are not strictly separated, several groups emerge that seem to reflect distinct properties of the signs. For instance, signs positioned on the extreme left section of the graph correspond to one-handed signs where the hand moves above or beside the head. The videos isolated from the main clusters are typically videos with missing keypoints in multiple frames. These keypoints' coordinates default to zero, artificially putting them far away from normal positions, thus isolating them.

4.3 Pose differences

Colours in the graph represent different data sources. We observe that some clusters consist exclusively of videos from a single dataset.

4.3.1 Handedness.

For instance, red points are grouped in two distinct clusters in the lower half of the graph. These are exclusively composed of one-handed sign videos from the Noema dataset, which only includes a left-handed signer. All of the other videos include right-handed signers. In sign language, each signer has a dominant hand. For one-handed signs, only the dominant hand is active. In two-handed signs, the dominant hand typically performs the more intense, faster, or semantically meaningful motion. Left-handed individuals tend to use their left hand as the dominant one, and right-handed individuals, their right. Since UMAP calculates distances by comparing corresponding values in each vector, and all vectors are structured with the left hand's coordinates first, distances are sometimes computed using hands that are dominant for one signer and not for the other, creating noise in the data and preventing relevant comparisons.

4.3.2 Non-dominant arm.

In addition to biases linked to handedness, most one-handed signs from the Gebärdensprach-Datenbank (dark blue), Nederlandse Gebarentaal (purple) and Svenskt teckenspråkslexikon (green) datasets are separated based on their source dataset. When looking at the videos in those clusters, we can notice a difference in the way passive arms are positioned. In the Gebärdensprach-Datenbank dataset, the signer has her left hand offscreen. All of its landmarks are therefore set to 0 during extraction. Both Nederlandse Gebarentaal and Svenskt teckenspråkslexikon interpreters keep their passive arms in a curved position, resting on their stomach. These differences lead UMAP to group videos with similar non-dominant arm configurations together, overshadowing the arm and the hand actually involved in the sign.

4.3.3 Morphology and Framing.

Finally, videos also appear to be grouped according to the interpreters' morphologies and the framing of the shot.

For example, three videos from the Nederlandse Gebarentaal dataset, which feature an extremely close-up shot compared to the

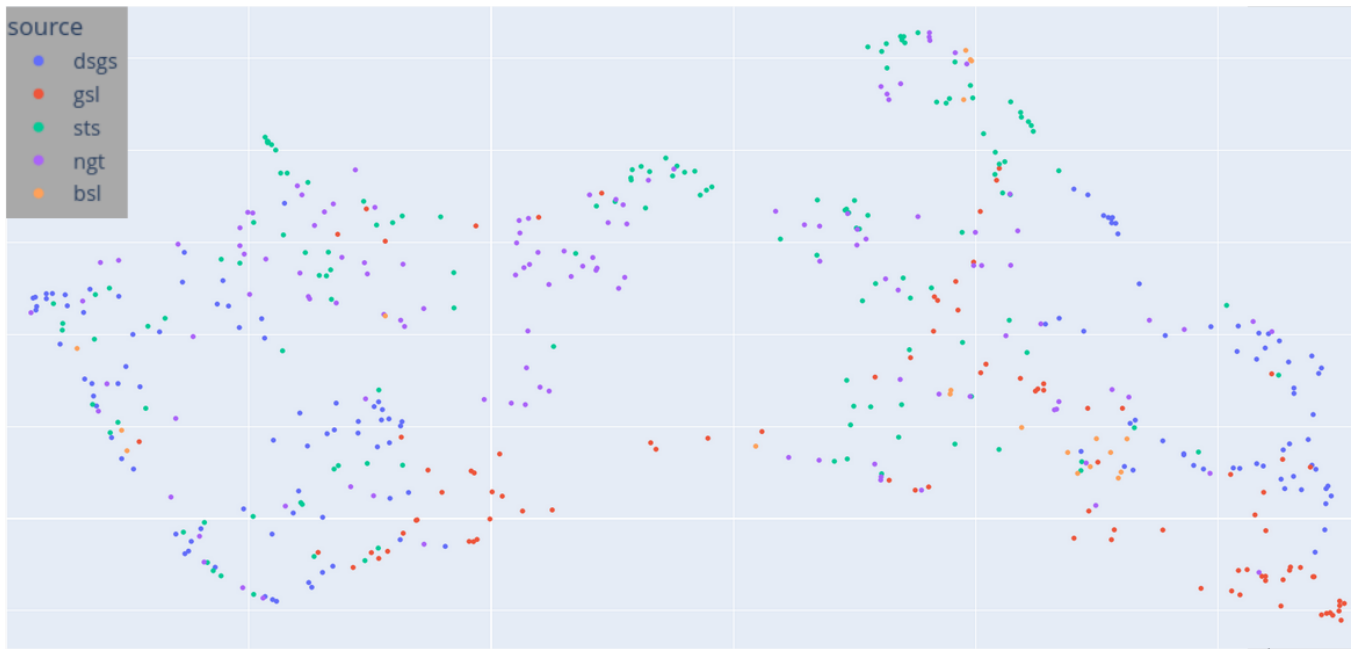


Figure 3: 2D visualisation of WordNet dataset using UMAP

rest of the dataset, are each other’s closest neighbours in the graph. Additionally, male signers in the Svenskt teckenspråkslexikon dataset tend to cluster together more frequently than with their female counterparts. This may be attributed to similar morphological traits, such as broader shoulders.

These observations suggest that morphology and framing introduce a bias in the clustering, as they influence global distances more strongly than subtle elements like hand configurations.

5 Bias correction

After identifying biases and how they could impact performances of sign similarity analyses, we applied processing steps on both ASLLVD and WordNet datasets to improve annotation consistency on different levels.

5.1 Interpolating missing keypoints

We previously showed that MediaPipe could fail to detect keypoints in certain frames, leading to biases in similarity matching. To avoid keeping zero values in some frames, we applied the interpolation method proposed in Zelinka and Kanis [27] to both datasets after landmark detection. This approach estimates missing keypoints by computing a weighted average of the corresponding keypoint coordinates from neighbouring frames, taking into account the confidence scores provided by MediaPipe.

While this method allows for the disappearance of null values, it raises a new issue: hands that have been wrongly identified by MediaPipe in a reduced number of frames now propagate throughout the entire video. Immobile or robotic hands are therefore present alongside the real one(s).

To prevent this problem, we implemented a condition before interpolation: for each detected hand, we check if it appears in

more than 10% of the frames. If not, the hand is flagged as non-real. Additionally, if a hand is perfectly immobile for the entirety of the video, it is flagged as non-real.



Figure 4: Frames with detected and non-detected right hand

5.2 Identifying the dominant arm

During landmarks retrieval, MediaPipe annotates each hand as the left or the right one with a confidence score. We have shown previously that handedness acts as a dominant discriminative feature in dimensionality reduction, overshadowing other aspects such as movement or hand configuration. Horizontally flipping all videos to account for handedness would double computation time and prevent the possibility of focusing on specific body parts during dimensionality reduction. For instance, to identify identical signs, Fragkiadakis et al. [10] applied UMAP to the dominant hand keypoints, resulting in an accuracy of 95. This score remains the highest, surpassing performances obtained by using keypoints from both the hands and the upper body. Solely comparing keypoints between dominant hands requires identifying them beforehand.

5.2.1 Identifying the dominant hand.

In Fragkiadakis et al. [10], the velocity of the hand is mentioned as being used to identify the dominant hand. We include two conditions in our pipeline. We start by computing the velocity of each hand in a video between all frames using the wrist coordinates and the number of frames per second. We average those measures to get an average velocity for each hand. If the left hand is on average faster than the right one, it is considered dominant. If the right hand is the fastest, the maximum height (y-axis) across all frames is used as a fallback, and the higher hand is considered dominant. Adding this fallback improves identification performance.

5.2.2 Evaluation of dominant hand identification.

To evaluate the performance of this method, 1000 one-handed and 1000 two-handed videos were randomly chosen in each dataset, with half of the WordNet videos showing a left-handed signer. (Not possible for the ASLLVD videos, as all signers are right-handed) For two-handed signs, only non-symmetrical signs have been counted as wrong predictions.

Table 1: Dominant Hand Identification Accuracy

Type of sign	WordNet	ASLLVD
One-handed	0.99	0.97
Two-handed	0.93	0.90
All	0.96	0.94

This combined method yields an accuracy of 0.96 over 2000 videos on the WordNet dataset. The total accuracy is slightly lower on the ASLLVD one. Most errors stem from inaccurate MediaPipe annotations, such as wrongly identified arms. Other errors, for two-handed signs, concern rapid non-dominant hands sometimes sharing identical configurations with the dominant ones but with different orientations/movements; see figure 5.

Finally, another common element leading to flawed handedness identification concerns signers moving the non-dominant hand during one-handed signs. For instance, in the Noema and BSL Sign-Bank datasets, signers sometimes move their non-dominant hand towards the bottom of the screen at the beginning of the video, and keep on slightly moving it until the end of the sign. Using y-coordinates as a primary condition for handedness identification has been tested to counter such cases; however, various signs



Figure 5: Signers wrongly flagged as left-handed

are located near the stomach or the thighs, leading to more errors overall.

5.2.3 Vertical flip.

If the dominant hand is identified as being the left one, all x-coordinates are flipped vertically ($x' = 1 - x$), ensuring that the dominant hand is always represented on the left side of the screen; see figure 6. The position of the left body parts and left hand are furthermore switched with the right body parts and right hand in the final pose embedding. These steps allow for more consistency across and within both datasets. Post handedness identification, left-handed signer clusters in figure 3 mostly disappeared from UMAP visualisation, showing that identifying handedness does indeed add consistency to the data, and allows for signs performed with the left hand to be associated with signs performed with the right hand.

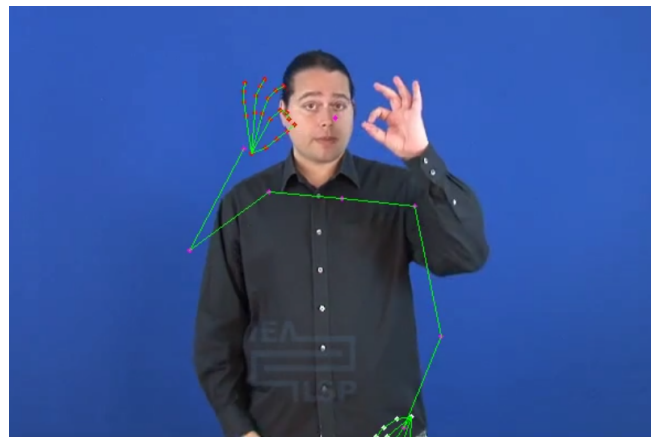


Figure 6: Vertical flipping of a left-handed signer

5.3 Normalising the passive arm

As stated previously, passive arms, those not involved in a one-hand sign, can be interpreted as being part of signs as they vary massively in positions across both datasets. Discrepancies in non-dominant arm positions influence distance computations and can therefore prevent grouping similar signs together. With the aim to improve consistency, these passive arms have been identified and reconstructed towards a unique fixed position.

5.3.1 Passive arm identification.

Passive arms mostly appear in two positions: straight and resting against the hip, or curved and resting on the stomach. Both types share a low elbow angle variance throughout the entire video. These angles are computed using the distances between the shoulder and the elbow (SE), the elbow and the body wrist (EW), and the body wrist and the shoulder (WS). The cosine law formula used to retrieve the value of the elbow angle is:

$$C = \arccos\left(\frac{SE^2 + EW^2 - WS^2}{2 * SE * EW}\right) \quad (1)$$

After evaluating variances of elbow angles in passive arms across 1000 videos from the WordNet dataset, 98.50% of those were under 80°. For all of the straight passive arms, angle values were between 170° and 180°.

These numbers allow us to identify if a non-dominant arm is passive in new videos based on the following conditions: we start by checking if the non-dominant arm has an elbow angle variance below 80°, or if all of its angles are between 170° and 180°. If these conditions are met, the arm is flagged as passive.

5.3.2 Reconstructing the passive arm.



Figure 7: Passive arms reconstruction

All passive arms are then reconstructed based on a fixed reference position and scaled to match the signer’s body proportions. They are regenerated starting from the neck, joint by joint, using

the average limb lengths measured on the opposite side. Joint angles are arbitrarily defined to mimic a natural resting position. The formula used to convert polar coordinates to Cartesian ones is:

$$x_1 = x_0 + \text{distance} * (-\cos(\text{angle})) \quad (2)$$

$$y_1 = y_0 + \text{distance} * \sin(\text{angle}) \quad (3)$$

5.3.3 Evaluation.

To evaluate passive arm identification and reconstruction, previously stated conditions and processes were applied to 1000 one-handed signs from the ASLLVD dataset. The dataset annotates each video as showing a one-handed or two-handed sign, allowing for an automatic evaluation. Passive arms were correctly identified in 95.20% of the videos.

6 Normalisation

As observed with UMAP visualisation, variability in signers’ morphologies and distances with the camera seem to influence matching performances; in certain cases, clusters are formed around signers’ characteristics rather than phonological features. To reduce the impact of these variations, we apply classic 2D pose normalisation [2, 10, 24] on our data.

6.1 Keypoints translation

The first step in the normalisation process is the translation of keypoints, meaning their coordinates are re-evaluated relatively to a new origin. The neck is chosen as the origin for the body coordinate system, while the wrists are chosen as the origin of their corresponding hands. Using the wrist as a reference point removes the absolute position of the hand joints, a position already conveyed by the wrist in the body system, and allows the focus to shift to hand configurations [2].

6.2 Anchor based normalisation

After translating each joint relative to the neck or to the wrist, differences in body morphologies can still result in variations in values of a same keypoint across the dataset. Someone with broad shoulders will have shoulder coordinates with higher values than those of someone with narrow shoulders. To address this, we apply a scaling step based on an anchor distance, a stable and representative measure of body proportions. By dividing all joint coordinates by this anchor distance, we aim to reduce variability in limb lengths across the dataset. In other words, other body distances (length between limbs) should vary in proportion to changes in the anchor distance.

In Fragkiadakis et al. [10], the authors proposed using the nose-neck distance as the anchor for body normalisation, arguing that it showed less variance than the shoulder-shoulder distance frequently used in sign language recognition tasks. After comparing the average variance of both these distances on all of the WordNet dataset, the nose-neck distance did appear less variant than the shoulder-shoulder one, although the difference remains extremely small (difference of 0.0003). To evaluate which anchor yields better proportional consistency, we conducted a secondary analysis: for each anchor (nose-neck vs. shoulder-shoulder), we applied the

corresponding normalisation to 500 videos, then measured the variance of other limb lengths (including shoulder-elbow, elbow-wrist, and shoulder-wrist distances) across the resulting normalised data.

Our aim was to minimise variance while preserving relative body proportions. In this context, higher variance in limb lengths after normalisation indicates more diversity in morphologies across signers. However, the purpose of normalisation is precisely to reduce such variation.

Although the nose-neck anchor yielded slightly lower variance in its own value, it led to greater variances in upper-body limb proportions post-normalisation when used on 500 test videos. In contrast, using the shoulder-shoulder distance as an anchor preserved these proportions more effectively, resulting in more stable and consistent body shapes across different signers.

We therefore chose to use the shoulder-shoulder distance as our anchor distance and applied the following final normalisation formula to both datasets, where (x_0, y_0) represents the coordinates of the neck for body keypoints and the wrist for hand keypoints:

$$(x'_k, y'_k) = \frac{(x_k, y_k) - (x_0, y_0)}{\|(x_{\text{shoulder1}}, y_{\text{shoulder1}}) - (x_{\text{shoulder2}}, y_{\text{shoulder2}})\|} \quad (4)$$

7 Similarity Analyses

After applying the preprocessing steps described above to MediaPipe joint annotations across both WordNet and ASLLVD datasets, we decided on establishing similarity scores among videos using both the initial and corrected keypoints to observe the beneficial effects of our methods.

7.1 Implementation

Preprocessed keypoints are extracted from each video and stored in 2D arrays of shape (NumberOfFrames, 50, 2), with 50 being the number of joints and 2 being the number of dimensions (x and y axes). We start by reshaping each array to (NumberOfFrames, 100) and use DTW to measure the similarity between pairs of videos.

7.2 Evaluation

The lack of compatible annotations in multilingual datasets makes evaluating similarity measures extremely difficult. Most papers, as in Fragkiadakis et al. [10] and Jangyodsuk et al. [13] look at the top-k returned signs for a query video, and count the result as a true positive if at least one of the returned signs is identical to the query sign.

7.2.1 WordNet.

We manually selected 511 videos from the WordNet dataset, so that each of these videos presents the exact same sign as at least one other video in the subset, forming matching groups. Although manual parameters are identical in all of these groups, non-manual ones, such as facial expressions, can differ. Groups show different signers and can involve two-handed and one-handed signs. For each query video, one true positive was counted if at least one sign in the top-k returned results was identical, that is to say, if at least one of the matching signs was present.

7.2.2 ASLLVD.

Compared to the WordNet dataset, the ASLLVD dataset provides information on hand configurations for each sign. The evaluation for this dataset is based solely on hand configuration: while not representing the sign in its entirety, this allows a completely automatic evaluation. 500 videos were randomly selected from the dataset, with half representing one-handed signs and the other half two-handed signs. For each query video, one true positive was counted if at least one sign in the top-k returned results shared the same hand configuration (an identical configuration for one-handed signs, and two shared configurations for two-handed signs).

8 Results

DTW was applied on every possible pair of each set using both raw MediaPipe coordinates and the ones resulting from the preprocessing pipeline. Results are shown in tables 2 and 3.

Table 2: ASLLVD: Top-k accuracy for raw and preprocessed MediaPipe annotations

	Top-5	Top-10	Top-15	Top-20
Raw	0.15	0.23	0.34	0.39
preprocessed	0.29	0.41	0.49	0.53

Table 3: WordNet: Top-k accuracy for raw and preprocessed MediaPipe annotations

	Top-5	Top-10	Top-15	Top-20
Raw	0.07	0.09	0.12	0.14
preprocessed	0.24	0.32	0.37	0.41

We observe accuracy increases up to 15 points in the best cases for the ASLLVD dataset and up to 31 for WordNet. These results demonstrate that preprocessing keypoints significantly enhances the ability to match signs based on hand configurations and other manual parameters.

The ASLLVD scores surpass those of WordNet, likely due to the evaluation focusing on a single manual parameter, making the task less complex.

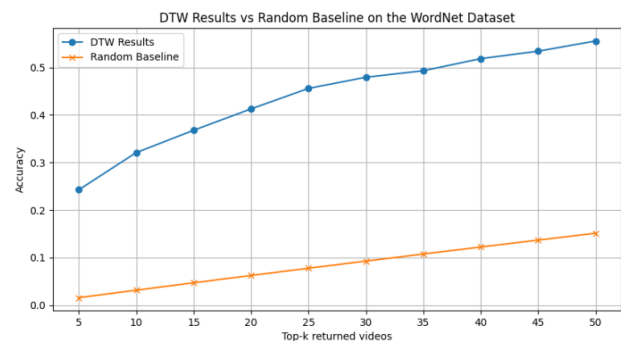


Figure 8: Preprocessed WordNet data: DTW results vs. Random baseline

For the sake of completeness, for each preprocessed query video in the WordNet subset, we computed the expected accuracy under random top-k retrieval and compared it to our actual results; see figure 8. The difference in values between this random baseline and the actual accuracies shows how DTW performances are not based on randomness, and illustrates the difficulty of the task.

9 Conclusion

Our work highlights the value of dimensionality reduction techniques to uncover structural biases in SL corpora, and introduces a preprocessing pipeline for pose-based features that significantly enhances the accuracy of sign similarity assessments.

We showed that applying dimensionality reduction techniques to pose features, such as body and hand keypoints, preserves enough information to identify meaningful patterns in a 2D space. This visualisation approach helped identify issues related to handedness and the representation of passive arms. By addressing these issues, through passive arm reconstruction, mirroring left-handed signers, and pose normalisation, we observed substantial improvements in similarity scores, as measured by DTW on both raw and preprocessed MediaPipe keypoints. Further linguistic analysis is needed to better understand how phonological features drive the grouping of non-identical signs. The proposed pipeline is well suited for inverse search applications, especially in datasets involving diverse signers and recording conditions. All code developed in this study, along with a curated subset of identical signs from the WordNet dataset, will be released publicly. Future work will include additional similarity analyses using methods by Fragkiadakis et al. [10], Boháček and Hruží [2], and Jungsoo Shin [14]. Our aim is to enrich the WordNet dataset, by enabling connections between signs not only based on meaning, but also on phonological similarity, thereby improving the accessibility and linguistic utility of SL dictionaries.

Acknowledgments

This work is part of Défi Inria COLaF, which was financed by Plan National de Recherche en Intelligence Artificielle.

References

- [1] Yunus Can Bilge, Ramazan Gokberk Cinbis, and Nazli Ikişler-Cinbis. 2023. Towards Zero-Shot Sign Language Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (Jan. 2023), 1217–1232. doi:10.1109/tpami.2022.3143074
- [2] Matyáš Boháček and Marek Hružík. 2022. Sign Pose-Based Transformer for Word-Level Sign Language Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. placeholder, placeholder, 182–191.
- [3] Penny Braem. 2001. A multimedia bilingual database for the lexicon of Swiss German Sign Language. *Sign Language & Linguistics* 4, 1-2 (December 2001), 133–143. doi:10.1075/sll.4.12.10booy
- [4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008 [cs.CV]. <https://arxiv.org/abs/1812.08008>
- [5] Mathias Claassen. 2023. Building a zero-shot Sign Pose Embedding model. <https://blog.xmartlabs.com/blog/machine-learning-sign-language-recognition/>.
- [6] Onno Crasborn, Richard Bank, Inge Zwitterlood, Els van der Kooij, Ellen Ormel, Johan Ros, Anique Schüller, Thierry and Goggi, Sara and Isahara, Hitoshi and Maegaard, Bente and Mariani, Joseph and Mazo, Hélène and Odijk, Jan and Piperidis, Stelios (Ed.). European Language Resources Association (ELRA), placeholder, 2478–2487. <http://www.lrec-conf.org/proceedings/lrec2022/index.html>
- [8] Eleni Efthimiou, Kiki Vasilaki, Stavroula-Evita Fotinea, Anna Vacalopoulou, Theodore Goulas, and Athanasia-Lida Dimou. 2018. The POLYTROPON Parallel Corpus. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, Proceedings of the LREC 2018 Workshop*, Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, Johanna Mesch, and Yutaka Osugi (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan, 39–44. ISBN: 979-10-95546-01-6. EAN: 9791095546016.
- [9] Jordan Fenlon, Kearsy Cormier, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. 2014. BSL SignBank: A lexical database of British Sign Language (First Edition). <https://www.bsllcorpproject.org/bsl-signbank/>.
- [10] Manolis Fragkiadakis, Victoria Nyst, and Peter van der Putten. 2020. Signing as Input for a Dictionary Query: Matching Signs Based on Joint Positions of the Dominant Hand. In *Proceedings of the LREC2020 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*, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch (Eds.). European Language Resources Association (ELRA), Marseille, France, 69–74. <https://aclanthology.org/2020.signlang-1.11/>
- [11] Yalan Gao, Cheng Xue, Ran Wang, and Xianwei Jiang. 2020. pca fingerspelling recognition via gray-level co-occurrence matrix and fuzzy support vector machine. *EAI Endorsed Transactions on e-Learning* 7, 20 (10 2020), placeholder pages. doi:10.4108/eai.12-10-2020.166554
- [12] Promila Haque, Badhon Das, and Nazmun Nahar Kaspary. 2019. Two-Handed Bangla Sign Language Recognition Using Principal Component Analysis (PCA) And KNN Algorithm. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. placeholder, placeholder, 1–4. doi:10.1109/ECCE.2019.8679185
- [13] Pat Jangyodsuk, Christopher Conly, and Vassilis Athitsos. 2014. Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features. In *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments (Rhodes, Greece) (PETRA '14)*. Association for Computing Machinery, New York, NY, USA, Article 50, 6 pages. doi:10.1145/2674396.2674421
- [14] Jaehae Jung Jungsoo Shin. 2023. ASL Recognition by the Layered Learning Model Using Clustered Groups. *Computer Systems Science and Engineering* 45, 1 (2023), 51–68. doi:10.32604/csse.2023.030647
- [15] Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Marc Schulder (Eds.). European Language Resources Association, Marseille, France, 102–109. <https://aclanthology.org/2022.signlang-1.16/>
- [16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC]. <https://arxiv.org/abs/1906.08172>
- [17] M. Madhwaran and Partha Pratim Roy. 2022. A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets. arXiv:2204.03328 [cs.CV]. <https://arxiv.org/abs/2204.03328>
- [18] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML]. <https://arxiv.org/abs/1802.03426>
- [19] Johanna Mesch, Thomas Björkstrand, Eira Balkstam, Patrick Hansson, and Nikolaus Riemer Kankkonen. 2024. Swedish Sign Language Resources from a User's Perspective. In *Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages*. ELRA Language Resources Association, Marseille, France, 254–261. CC BY-NC 4.0.
- [20] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021. Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition. arXiv:2104.10166 [cs.CL]. <https://arxiv.org/abs/2104.10166>
- [21] Meinard Müller. 2007. Dynamic time warping. *Information Retrieval for Music and Motion* 2 (01 2007), 69–84. doi:10.1007/978-3-540-74048-3_4
- [22] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. ASL Video Corpora and Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP). arXiv:2201.07899 [cs.CL]. <https://arxiv.org/abs/2201.07899>
- [23] Erdefi Rakun, Mirna Andriani, I Wayan Wiprayoga, Ken Danniswara, and Andros Tjandra. 2013. Combining depth image and skeleton data from Kinect for recognizing words in the sign system for Indonesian language (SIBI [Sistem Isyarat Bahasa Indonesia]). In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. placeholder, placeholder, 387–392. doi:10.1109/ICACSIS.2013.6761606
- [24] Kyunggeun Roh, Huije Lee, Eui Jun Hwang, Sukmin Cho, and Jong C. Park. 2024. Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Johanna Mesch, and Marc Schulder (Eds.). ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL), Torino, Italy, 400–411. <https://www.sign-lang.uni-hamburg.de/lrec/pub/24052.pdf>
- [25] Marc Schulder, Sam Bigeard, Maria Kopf, Thomas Hanke, Anna Kuder, Joanna Wójcicka, Johanna Mesch, Thomas Björkstrand, Anna Vacalopoulou, Kyriaki Vasilaki, Theodoros Goulas, Stavroula-Evita Fotinea, and Eleni Efthimiou. 2024. Signs and Synonymity: Continuing Development of the Multilingual Sign Language Wordnet. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Johanna Mesch, and Marc Schulder (Eds.). ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL), Torino, Italy, 281–291. <https://www.sign-lang.uni-hamburg.de/lrec/pub/24034.pdf>
- [26] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. 2021. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation* 76 (2021), 103055. doi:10.1016/j.jvcir.2021.103055
- [27] Jan Zelinka and Jakub Kanis. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. In *Proceedings of the Conference*. Placeholder, Conference Location, 3384–3392. doi:10.1109/WACV45572.2020.9093516
- [28] Joanna Lacheta, Małgorzata Czajkowska-Kisil, Jadwiga Linde-Usiekiewicz, and Paweł Rutkowski (Eds.). 2016. *Korpusowy słownik polskiego języka migowego / Corpus-based Dictionary of Polish Sign Language*. Faculty of Polish Studies, University of Warsaw, Warsaw. <http://www.slownikpjm.uw.edu.pl> Online publication.