TILBURG UNIVERSITY
Department of Cognitive Science & AI

LREC 2026
Workshop on Representation & Processing of Sign Languages, *16 May 2026, Palma, Spain*

Feature Analysis of MoCap Data for Optimised Sign Language Processing

Yves A. Duppen | Mirella De Sisto | Ifigeneia Mavridou | Phillip Brown | Lisa Lepp | Dimitar Shterionov
{Y.A.Duppen, M.DeSisto, I.Mavridou, P.C.Brown, L.B.Lepp, D.Shterionov}@tilburguniversity.edu — Tilburg University, The Netherlands

Keywords: sign language NLP · motion capture · BVH feature analysis · sign classification · dimensionality reduction

Introduction

Despite rapid advances in AI, sign language (SL) processing faces a **data bottleneck**:

- SL data is scarce, scattered and in different formats across corpora
- SL data have not been collected with MT in mind

Marker-based **motion capture (MoCap)** provides high-precision 3D recordings of body movements and has gained increasing traction for SL linguistics and technology.

A single MoCap recording may contain **more than 240 frames across 156 features**, introducing substantial representational complexity.

Research questions

- Which features are of highest importance for the task of sign classification?
- How does feature reduction impact sign classification?

This work identifies features of highest discriminative importance and shows that **feature-reduced representations outperform full-feature baselines** in sign classification. Sign language under investigation: **Nederlandse Gebarentaal (NGT)**.

Dataset creation

The **CoCoS dataset** was recorded at a co-creative workshop in collaboration with the *Nederlands Gebarencentrum* as part of the CoCoS project.

Recording equipment & setup

- 6 OptiTrack Flex 13 cameras; synchronised via OptiTrack Motive v3.3
- System calibration: $<$0.3 mm mean residual error
- Reflective markers (14 mm) on upper-body anatomical regions
- Hand articulation: Manus Quantum Metaglove alignment

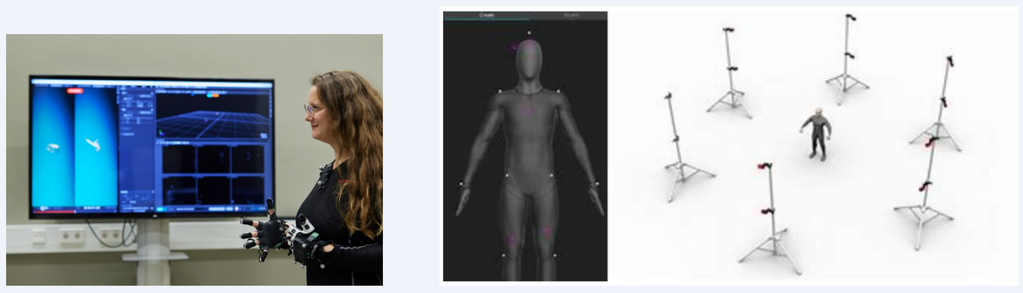

Figure: Left to right: our colleague and co-author in the MoCap suit testing the equipment; the camera setup we used.

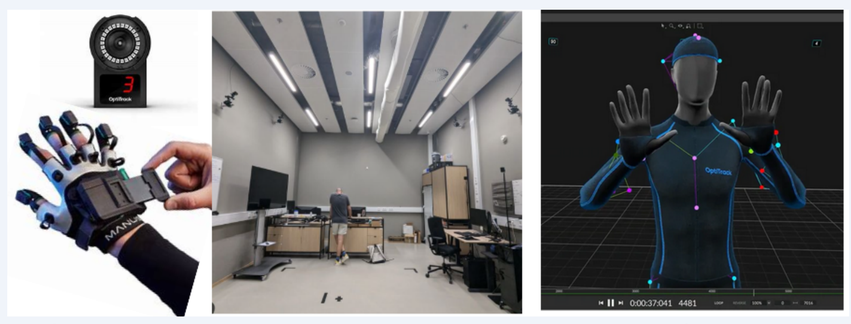

Figure: Left to right: OptiFlex camera (top) and Manus glove (bottom), TiU's VR Lab, Sceletal view through the Motiv software.

Corpus statistics

- 3 signers (2 female, 1 male; 1 left-hand dominant)
- 44 NGT signs covering the full phonological inventory
- 116 MoCap recordings exported in BVH format

```
HIERARCHY
ROOT mixamorig_Hips
{
        OFFSET -0.000329 0.034763 0.013172
        CHANNELS 6 Xposition Yposition Zposition Xrotation Yrotation Zrotation
        JOINT mixamorig_Spine
        {
                OFFSET -0.000531 0.096593 0.001533
                CHANNELS 3 Xrotation Yrotation Zrotation
                JOINT mixamorig_Spine1
                {
                        OFFSET -0.000284 0.099406 -0.006202
                        CHANNELS 3 Xrotation Yrotation Zrotation
                        JOINT mixamorig_Spine2
                        {
                                OFFSET -0.000665 0.100820 -0.005436
                                CHANNELS 3 Xrotation Yrotation Zrotation
                                JOINT mixamorig_Neck
                                {
                                        OFFSET -0.003569 0.181004 -0.017910
                                        CHANNELS 3 Xrotation Yrotation Zrotation
                                        JOINT mixamorig_Head
                                        {
                                                OFFSET -0.000567 0.090675 0.004608
                                                CHANNELS 3 Xrotation Yrotation Zrotation
                                                End Site
                                                {
                                                        OFFSET -0.027755 0.065308 0.087730
                                                }
                                                End Site
                                                {
                                                        OFFSET 0.030266 0.066099 0.089151
                                                }
                                                End Site
                                                {
                                                        OFFSET 0.000889 0.164224 0.030118
                                                }
                                        }
                                }
                                JOINT mixamorig_LeftShoulder
                                {
                                        OFFSET 0.040153 0.123825 -0.018173
                                        CHANNELS 3 Xrotation Yrotation Zrotation
```

Figure: A BVH file snippet representing the MoCap of the NGT sign *vegetable*. The BVH file can directly be loaded in Animics (`animics.gti.upf.edu`) to generate an avatar – the avatar on the right is driven by the BVH file on the left.

Methodology

Data preprocessing

Sliding-window segmentation: window = 60 frames, stride = 20 frames with **Majority-vote labelling** per window.
Dataset shape: 2,248 $\times$ 60 $\times$ 156 (sequences $\times$ frames $\times$ features) with Train, validation and test split: 72.25% / 12.75% / 15%.

Model architectures evaluated

(i) **LSTM & BiLSTM** — sequential recurrent baselines;
(ii) **CNN1D** — 1-dimensional convolutional network;
(iii) **TCN** — Temporal Convolutional Network with dilated convolutions;
(iv) **Transformer Encoder** — self-attention over temporal sequences;
(v) **CNN+LSTM Hybrid** — convolutional feature extraction + recurrence
(vi) **GCN** — Graph Convolutional Network over joint skeleton topology

Feature importance method

Model-agnostic **permutation importance**: each feature's values are randomly permuted across samples while preserving within-sequence temporal structure.
The resulting drop in weighted F1 relative to the unpermuted baseline quantifies each feature's discriminative contribution.

Results: Model Performance

F1-scores for the **full 156-feature** representation and the **reduced top-30 feature** set. Feature reduction yields large, consistent improvements across every architecture.

Model	Val F1 (156 feat.)	Test F1 (156 feat.)	Val F1 (Top 30)	Test F1 (Top 30)
TCN	**0.605**	**0.632**	**0.931**	**0.925**
Transformer	0.558	0.568	0.913	0.911
CNN1D	0.558	0.580	0.925	0.921
CNN+LSTM	0.268	0.270	0.936	0.933
BiLSTM	0.268	0.270	0.920	0.918
LSTM	0.268	0.270	0.906	0.904
GCN	0.271	0.273	0.823	0.807

Table: F1-scores — full vs. reduced feature representations (TCN row highlighted).

Feature Importance Analysis

Feature permutation

We did individual feature permutations and scored the decrease in weighted F1 relative to the unpermuted baseline (**0.716**) measures each feature's discriminative contribution.

#	Feature Name	F1 Drop	F1 Perm.
1	RightForeArm_Zrotation	**0.212**	0.504
2	RightHand_Zrotation	**0.189**	0.527
3	LeftHand_Zrotation	0.170	0.545
4	RightHand_Yrotation	0.154	0.562
5	RightForeArm_Xrotation	0.148	0.567
6	LeftHandThumb3_Zrotation	0.141	0.575
7	LeftHand_Yrotation	0.117	0.599
8	RightForeArm_Yrotation	0.116	0.600
9	LeftForeArm_Zrotation	0.102	0.614
10	LeftForeArm_Yrotation	0.100	0.616

Table: Top-10 features by F1 drop — baseline F1 = 0.716 for all rows.

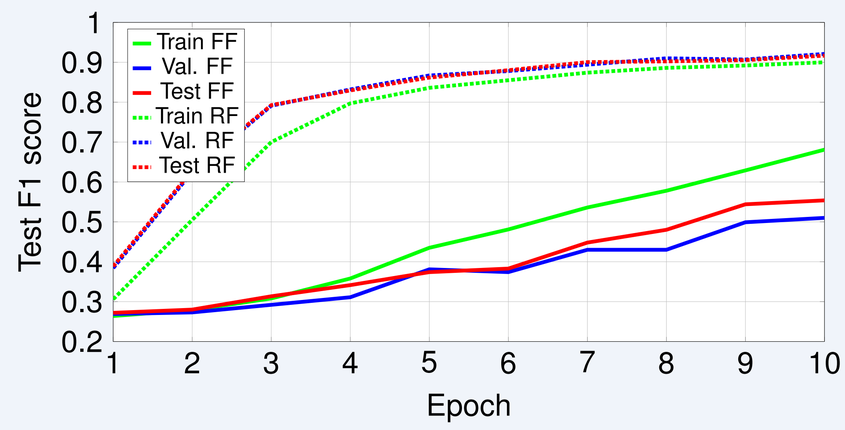

Figure: F1 scores for the best model trained on the full-feature (FF) set, and the reduced-feature (RF) set.

Unsupervised Structure & Redundancy

K-means clustering on PCA-reduced motion features examines latent structure *without label supervision*.

- Silhouette analysis: optimal **k = 3** clusters; average silhouette score = **0.558**
- Signs retain structured low-dimensional organisation even without label supervision, indicating that phonological categories leave detectable traces in the motion data

Correlation analysis

Very high inter-joint correlations ($>$0.95) between anatomically adjacent joints, e.g.:

- RightHandRing2 / RightHandRing3 X_{rot}
- LeftHandMiddle2 / LeftHandMiddle3 X_{rot}
- Spine / Spine1 Z_{rot}

Conclusions

Key Findings

- MoCap sign representations contain **substantial structured redundancy**
- Feature reduction **156** $\rightarrow$ **30 features** raises TCN validation **54%** relatively
- Right-hand rotational channels carry the largest articulatory load, consistent with phonological accounts of dominant-hand primacy
- TCN is best suited to model BVH MoCap data for sign classification
- Strong inter-joint correlations reward architectures exploiting local spatial coherence

Limitations & Future Work

- Small dataset (3 signers, 44 signs, 116 recordings); generalisability to larger, more diverse corpora remains to be established
- Feature ranking derived prior to the train/test split — potential information leakage means results are indicative rather than conclusive
- Handedness imbalance (2/3 right-dominant) precludes definitive claims about left-hand contributions; future work should balance signer groups
- Future work: larger-scale data collection; by-signer attribute analysis; cross-modal validation with video-based pose estimators

Code:
`github.com/dimitarsh1/CoCoS_dataprocessing`

{Y.A.Duppen, M.DeSisto, I.Mavridou, P.C.Brown, L.B.Lepp, D.Shterionov}@tilburguniversity.edu