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
Warandelaan 1, 5037AB Tilburg, The Netherlands

Abstract

Despite the rapid advances in AI and its impact on machine translation (MT), when it comes to sign language (SL) processing and MT, there is a big bottleneck – the lack of substantial quantities of quality signed data suitable for developing SLMT models. Marker-based motion capturing (MoCap) is a technique for tracing and recording the body movements (including hands and figures) in 3D space with high precision and has been widely used in SL research. MoCap data is of high representative accuracy, making it very suitable for analysing movement patterns and articulatory features. However, it is also very complex – a recording of a single sign may contain more than 240 entries over 156 features making it difficult for processing. In this paper we analyse MoCap data aiming to understand which captured features are of high importance. Consecutively, we optimise the MoCap data representation, reducing the number of features, and assess how this feature-reduced data impacts sign classification task. We organise MoCap features based on their importance and show how models trained on feature-reduced representations outperform those developed on the complete feature set.

Keywords: sign language NLP, motion capturing, data optimisation, sign classification

1. Introduction and background

The rapid technological and methodological advances in deep learning, and in AI in general, have not only improved language modeling and machine translation (MT), recognition of image, video and audio signals, the synthesis of life-like 3D avatars, etc., but have also led to the fusion of interdisciplinary research innovations. These advancements have created a fruitful environment for impactful research in the field of sign language processing and translation. Initiatives such as the SignON [1] and EASIER [2] consortia, the sign language MT shared tasks in the WMT conference (Müller et al., 2022; Müller et al., 2023) and, in general, the growth of number of publications in the field testify for this trend. Despite this spur of activities, one significant bottleneck remains. In particular, this bottleneck is data – high volumes quality data are required for training models of sufficient performance. Unfortunately, sign language (SL) data are scarce and scattered making training MT systems for SLs impractical. Furthermore, SL data have not been collected with MT (and the underlying pipeline components) in mind (Vandeghinste et al., 2024). This challenge can be addressed by either collecting new data and / or processing existing data to fit the requirements of an MT (or LLM) framework.

When processing existing video data, it is common to use a pose estimator, such as OpenPose (Cao et al., 2018) or MediaPipe (Lugaresi et al., 2019) for feature extraction. This facilitates the processing of input videos by reducing the complexity in the video stream, retaining information relevant to the signer (Rastgoo et al., 2021). However, these methods suffer from some limitations when it comes to SL processing. First, they have limited capacities with estimating depth and with occlusion. Second, they are not specifically trained on SL data or developed for SL research use-cases. As De Coster et al. (2023) shows, the lack of domain-specificity leads to errors in the (recognition). Existing SL specific methods (e.g. Forte et al. (2023); Ivashechkin et al. (2023)) have a better ability to process SL data, but still suffer from limitations related to depth accuracy and occlusions (Andersen et al., 2025). Pose estimators are an efficient solution to processing existing data and despite their limitations they do provide a mechanism to convert a signed video into a numerical format that can be processed with existing ML and DL tools.

However, when it comes to collecting new data, an alternative to first recording video and then processing the video data (with pose estimators) is to use *motion capturing*. In recent years, data collection via marker-based motion capture techniques has gained increasing attention for SL linguistics and SL technology purposes (Andersen et al., 2025). Given their ability to provide fine-grained 3D information, marker-based motion capture methods generate highly accurate representations of what is being signed. In addition, collecting

[1] https://signon-project.eu
[2] www.project-easier.eu

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 120–128
16 May 2026.

data in a 3D setup prevents a number of limitations, such as those concerning depth accuracy or occlusions previously mentioned, especially when this setup is tailored towards sign language production. The recent work by Andersen et al. (2025) and earlier work by Jedlička et al. (2020) present such tailored setups outlining alternatives and highlighting limitations and decision points.

Similar to those of pose estimators, MoCap data are very elaborate – a sequence of spatial data in 3D, i.e. three numerical values for a keypoint, and a number of keypoints for every captured frame, typically in a 120 frame-per-second recording. These data are quite useful when it comes to sign language synthesis, i.e. the task of generating an artificial 3D signer (often referred to as avatar) that can convey a signed utterance (Bernhard et al., 2022; Jedlička et al., 2020; Naert et al., 2020).

While the majority of work using MoCap data focuses on data-driven synthesis (Naert et al., 2020) and (Bernhard et al., 2022; Jedlička et al., 2020; Naert et al., 2020), in this work we analyse the MoCap data through the lens of sign language recognition (SLR), the task of processing an input video (containing a sign or a signed utterance) and generating a format suitable for a specific downstream task. SLR is typically the first step in the sign language machine translation pipeline (Shterionov et al., 2024). We explore the structure of MoCap data with the goal of identifying important features and reducing complexity. MoCap generates highly detailed and close to realistic motion data of a signer's movements and articulations. By analysing MoCap data, in contrast to video recordings (potentially processed with a pose estimator) aiming at a better understand of which human features have highest impact on classification of signs.

We analyse a dataset of 44 isolated signs, each produced by 3 signers (i.e. a total of 132 signs). The language is Sign Language of the Netherlands (*Nederlands Gebarentaal* – NGT). These 44 signs were selected because they capture the whole phonological inventory of NGT. Our investigation involves feature importance analysis and classification of individual signs based on BioVision Hierarchy (BVH) files. This format is widely used for motion capturing data and for its usability with avatar-generating softwares (e.g. Animics[3]).

Feature importance analysis for sign language has been previously conducted in the context of signer identification (among others (Bigand et al., 2020, 2021b)). Bigand et al. (2020) and 2021b also employ classification (and use dimensionality reduction) however their task differs from ours – we aim at classification of signs rather than of sing**ers**. Dimensionality reduction has also been used in the context of spontaneous sign language decomposition into elementary movements Bigand et al. (2021a); however, our work specifically looks into the features native for MoCap and MediaPipe (expressed in BVH format) rather then of their principle components.

[3] https://github.com/upf-gti/animics

2. Dataset Creation

The dataset we used is a newly recorded set of 44 signs from the Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT). The data collection took place as part of the CoCoS project – small-scale project funded by Tilburg University and executed in collaboration with the Nederlands Gebarencentrum (NGT). Within this project, we organised a co-creative workshop which, among others, allowed us to record MoCap data.[4] Three people participated in the data collection – two female and one male participants; one of them with a dominant left hand.

We used a setup with six OptiTrack Flex 13 motion capture cameras[5] to record full-body motion data. All cameras were synchronised with the OptiTrack Motive software[6] (version. 3.3). System calibration was performed prior to the capture sessions achieving <0.3 mm mean residual error. Reflective markers (14 mm diameter) were placed on an OptiTrack suit; these markers were distributed on the anatomical regions of the upper body in accordance to a standard rigidbody marker set, with particular emphasis on wrists, elbows, shoulders, and torso. Hand marker placement was in accordance with the Manus Quantum Metaglove alignment.[7]

The captured data-three signers signing 44 individual signs-was exported in BVH format for post-processing and analysis. Our data processing code is available at: https://github.com/dimitarsh1/CoCoS_dataprocessing.

3. Experiments and results

We used the CoCoS MoCap data (Sec. 2) to train various classifiers and to assess their performance.

3.1. Objectives

Our experiments aimed to assess the impact of various features on the performance of various temporal-spacial architectures on the task of classification of signs based on inputs in BVH format. The following model families were trained:

[4] In a separate session we also recorded videos of the same 44 signs to be used with a body / pose estimation framework. For the scope of this paper, we only focus on the MoCap data.

[5] https://optitrack.com

[6] https://optitrack.com/software/motive

[7] https://manus-meta.com

LSTM, BiLSTM, CNN1D, CNN+LSTM hybrid, Transformer Encoder, Temporal Convolutional Network (TCN), and Graph Convolutional Network (GCN). In particular, we aim to answer the following research question:

Which features are of high importance (for MoCap-based inputs) and how they impact the predictive performance of the models and how they impact model performance?

3.2. Data preprocessing

The MoCap data from the three participants were concatenated together to form a larger data set. This raw data was organised at frame level as a tabular feature matrix X, and a frame-wise label vector y. To enable sequence-based models, the frame stream was converted into fixed-length sequences using a sliding-window segmentation approach. Sequences were generated using a window length of 60 frames and a stride of 20 frames. Formally: $X[t_k : t_k + 60, :]$ where $t_k \in \{0, 20, 40, \ldots\}$.

Importantly, individual annotated sign segments were not treated as single input sequences. Instead, the fixed-length sliding window was applied over the continuous frame stream irrespective of sign boundaries. Consequently, a single annotated gesture could contribute to multiple overlapping windows, and sliding windows may cross sign-segment boundaries, including transitions among different gestures or between gesture and non-gesture frames. Therefore, for a given window $X[t_k : t_k + 60, :]$, the corresponding frame-level labels y_t may contain multiple class identities present in the same window.

Because labels were available per frame, a single class label had to be assigned to each 60-frame window. For each window, the final sequence label was computed using a majority vote over the 60 constituent frame labels: $\hat{y}^{(k)} = \arg\max_c \sum_{t=t_k}^{t_k+59} \mathbf{1}[y_t = c]$. This choice gives precedence to the class most dominant within the window and decreases sensitivity to short label noise or transitional frames.

The resulting dataset of sequences $(2248 \times 60 \times 156)$ was split into train (72.25% of all data), validation (12.75% of the data), and test (15%) segments. This validation split was used for early stopping and model selection.

For the dataset format, sequences were wrapped in a PyTorch Dataset and loaded using DataLoaders with batch size 64. Shuffling was only enabled for training to reduce ordering bias.

For graph-based models, the feature vector at each time step was assumed to consist of concatenated joint coordinates, to ensure structural compatibility, the feature dimension F was required to be divisible by the number of coordinate channels per joint C, allowing inference of the number of joints $J = F/C$. The input was therefore interpreted as a structured tensor of shape $T \times J \times C$.

A chain-structured adjacency matrix was constructed to define local spatial connectivity between neighboring joints. This simplified topology yields a computationally streamlined and reproducible baseline graph setup.

3.3. Models and model setup

We experimented with the following architectures and hyperparameters using the aforementioned same data preprocessing (see Section 3.2) pipeline:

- **LSTM:** with $input\ size = F$, $hidden\ size = 128$, $num\ layers = 1$, $cell = "lstm"$, and $bidirectional = False$.
- **BiLSTM:** identical to the LSTM configuration, but with $bidirectional = True$.
- **CNN1D:** instantiated with $n\ features = F$ and $num\ classes = C$ (convolutional block details follow the repository implementation).
- **TCN:** instantiated with $n\ features = F$ and $num\ classes = C$ (dilation schedule, kernel sizes, and residual block settings follow the repository implementation).
- **Transformer encoder:** instantiated with $num\ classes = C$ (number of heads, encoder depth, and feed-forward dimensions follow the repository implementation).
- **CNN+LSTM hybrid:** with $input\ size = F$, $cnn\ channels = (64, 128)$, $lstm\ hidden = 128$, $lstm\ layers = 1$, and $bidirectional = False$.
- **GCN:** with $num\ joints$, $coords\ per\ joint$, $num\ classes = C$, and $adj\ matrix = adjacency\_full$ (graph layer configuration follows the repository implementation).

Training and evaluation settings are as follows: the "Adam" optimizer from PyTorch, 5-fold cross-validation, maximum epochs of 100, learning rate 10^{-3}, and early stopping patience of 3 epochs without validation improvement (Bergman et al., 2024; Goodfellow et al., 2016). Cross-validation was used to obtain a solid estimate of model performance and reduce dependence on a single train/validation split, given the small dataset. Early stopping was applied to limit overfitting, given that overlapping windows may increase sample similarity.

4. Results

4.1. Model performance

Table 1 shows the mean F1 scores on the validation set across the seven models, averaged over all

cross-validation folds, For completeness, the post-hoc test F1-scores on the held-out test set are also reported for comparison, although model selection was based exclusively on validation performance. The TCN model achieved the highest performance (best validation F1 = 0.605), followed by the Transformer model (best validation F1 = 0.558). The recurrent models (LSTM, BiLSTM) and the convolution–recurrent hybrids performed significantly worse.

Model	Val F1	Test F1
TCN	0.605	0.632
Transformer	0.558	0.568
CNN1D	0.558	0.580
GCN	0.271	0.273
BiLSTM	0.268	0.270
CNN+LSTM	0.268	0.270
LSTM	0.268	0.270

Table 1: *Validation F1 (mean over folds) and Test F1. Test score is reported post-hoc and was not used for model selection.*

In the final evaluation on the held-out test set, the TCN achieved a test F1-score of 0.57 on the tenth epoch, confirming its robustness and generalizability of the model Figure 1.

To demonstrate the stability of the TCN model, we also show the TCN's training F1 score during the different epochs in Figure 1. The model exhibited consistent improvement in discriminative performance (train F1 from 0.264 to 0.681 over 10 epochs), and the steady increase in training F1-score indicates stable optimisation behaviour across epochs.

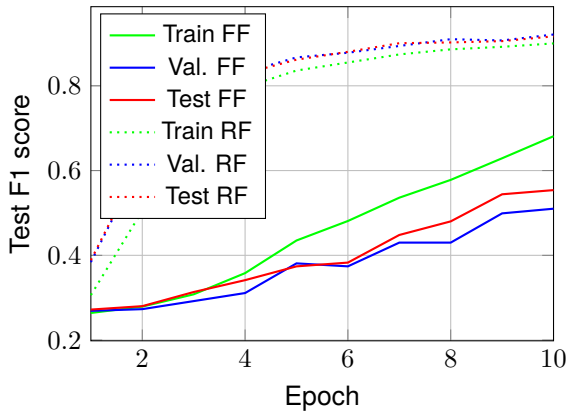

Figure 1: Post-hoc test F1 scores for the best model trained on the full-feature (FF) set, and the reduced-feature (RF) set.

4.2. Feature importance

According to Tsinganos et al. (2022), gesture discrimination is mainly influenced by features related to the forearm and hand. This is also shown in our feature importance analysis of the TCN model (Table 2). The 30 features with highest F1-score drop are mainly associated with right-hand articulators, particularly the right forearm and right hand. Given that most of our participants has primarily a dominant right hand, we assert that these results should reflect this feature (hand dominance) rather than be interpreted as absolute.

We assessed feature importance through a model-agnostic permutation importance process. After training, the baseline weighted F1-Score was computed on a held-out set. For each feature, its values were randomly permuted across samples while preserving the temporal structure within each sequence. The model was re-evaluated on the permuted data, and feature importance was defined as the decrease in F1-score relative to the baseline.

This method measures the extent to which the trained models rely on each feature for classification. Because permutation was applied across samples, the analysis captures sequence-level discriminative contribution rather than time-step-specific saliency. To control computational cost, evaluation was performed on a subset of the validation data.

Table 2 shows an F1 drop of 0.212 for the `RightForeArm_Zrotation` (most significant absolute reduction) to 0.100 for the `LeftForeArm_Yrotation` (list significant absolute reduction). This reduction in performance signals a notable degradation in classification performance when temporal patterns were disrupted. Left-sided joints (e.g., `LeftThumb3`, `LeftForeArm`, `LeftPinky2`) also contribute, but overall feature importance is clearly dominated by right-sided joints.[8]

5. Unsupervised Structure: Clustering in PCA Space

To further examine the latent structure of the motion feature space, K-means clustering was performed on PCA-reduced representations of the motion features. The optimal number of clusters was determined using silhouette analysis. The silhouette analysis indicated $k = 3$, which produced the highest average silhouette coefficient of 0.558 (Figure 2). Although this value reflects only moderate cluster separability, these results show that signs retain a low-dimensional structure that remains separable even in the absence of label supervision (Rousseeuw, 1987).

[8] As noted above, $\frac{2}{3}$ of our participants use their right hand as dominant. We stress that as hand dominance impacts signing, so would these results be impacted in case the majority of our participants were left-hand-dominant signers. This falls under our limitations and opens the potential for future work.

Feature idx	Feature name	F1 drop	F1 perm mean	F1 base
81	RightForeArm_Zrotation	0.212	0.504	0.716
84	RightHand_Zrotation	0.189	0.527	0.716
27	LeftHand_Zrotation	0.170	0.545	0.716
86	RightHand_Yrotation	0.154	0.562	0.716
82	RightForeArm_Xrotation	0.148	0.567	0.716
36	LeftHandThumb3_Zrotation	0.141	0.575	0.716
29	LeftHand_Yrotation	0.117	0.599	0.716
83	RightForeArm_Yrotation	0.116	0.600	0.716
24	LeftForeArm_Zrotation	0.102	0.614	0.716
26	LeftForeArm_Yrotation	0.100	0.616	0.716

Table 2: *Top 10 Permutation-based feature importance ranked by F1 score drop for the TCN model.*

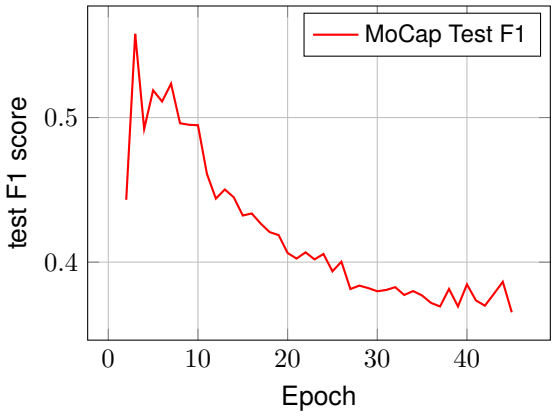

Figure 2: Silhouette score

Correlation matrices revealed very high correlations (> 0.95) between anatomically adjacent joints, including: `RightHandRing2 X_rotation` and `RightHandRing3_Xrotation`; `LeftHandMiddle2_Xrotation` and `LeftHandMiddle3_Xrotation` as well as `Spine_Zrotation` and `Spine1_Zrotation`.

This is expected in articulated-body motion capture: rotations of physically connected joints propagate through the kinematic chain, creating mechanical interdependence. This redundancy partially explains why convolutional models, which leverage local temporal and spatial coherence, outperform recurrent models on this task.

6. Models with reduced feature set

By modifying the experimental arrangement to reduce the number of input features, the models were retrained using only the 30 most influential features identified through permutation importance. As reported in Table 3, this feature reduction led to a marked increase in validation F1-scores across all evaluated architectures. For the TCN model, the validation F1-score improved from 0.605 (see Table 1) with the full 156-feature representation to 0.936 using the reduced feature set (see Table 3). This suggests that removing redundant or less significant features may facilitate more successful learning.

Because the feature ranking was derived prior to splitting, these results may be optimistic owing to potential information leakage; we therefore interpret them as indicative rather than conclusive; we leave the detailed analysis for future work.

Model	Val F1	Test F1
CNN+LSTM	0.936	0.933
TCN	0.931	0.925
CNN1D	0.925	0.921
BiLSTM	0.920	0.918
Transformer	0.913	0.911
LSTM	0.906	0.904
GCN	0.823	0.807

Table 3: *Validation F1-scores and Test F1. cross deep learning architectures using top 30 features. Test score is reported post-hoc and was not used for model selection.*

The superior performance of the Temporal Convolutional Network (TCN) compared to the other evaluated architectures can be interpreted in light of the structural properties of the motion capture data and the modeling assumptions underlying each architecture.

6.1. Temporal Convolutions and Local Motion Structure

TCNs employ dilated temporal convolutions, which allow the model to detect both short-range and long-distance dependencies within sequential data (Bai et al., 2018). Through stacked dilation layers, the receptive field increases exponentially, empowering the model to integrate information across multiple temporal scales. In the context of sign languages, this design is particularly relevant. Individual signs consist of fine-grained articulatory movements, such as finger and wrist rotations, embedded inside broader motion trajectories of the arm and upper body. The dilated convolutional

structure may therefore facilitate the simultaneous modeling of micro-level kinematic variations and macro-level movement patterns.

Given the large dimensionality and arranged nature of the BVH feature encoding, the convolutional approach appears well-suited to capturing locally coherent movement sequences while maintaining sensitivity to longer temporal connections.

6.2. Limitations of Recurrent Models for High-Dimensional Data

In contrast, recurrent models such as LSTMs and BiLSTMs process each time step sequentially and operate on flattened feature vectors. While recurrent models are intended to capture temporal connections, they do not clearly incorporate spatial relationships among joints unless additional framework constraints are introduced. Although CNN+LSTM combines spatial feature extraction with temporal recurrence, the added architectural complexity does not necessarily yield better generalization when the temporal structure is already well captured by expanded convolutions (Jafari and Jafari, 2026). Prior work has shown that explicitly modeling skeletal topology can improve performance in motion recognition tasks (Yan et al., 2018).

Furthermore, the comparatively lower performance of recurrent models may be attributed to their limited ability to exploit spatially structured correlations among joint rotations. Given the strong interdependencies observed between anatomically adjacent joints, architectures that leverage local feature coherence appear better suited to this representation.

6.3. Transformer Performance and Data Requirements

Transformer-based models rely on self-attention modules to model long-range dependencies across sequences (Vaswani et al., 2017). In principle, this allows for flexible modeling of long-range interactions between time steps. However, Transformers benefit from large-scale training data to obtain robust focus patterns and minimise overfitting.

As shown in both Table 1 and Table 3, the Transformer achieved competitive but slightly lower performance compared to the TCN. This may reflect insufficient data for the attention mechanism to fully exploit its modeling capacity; we ought to stress that our dataset is relatively small and acknowledge the need for further experimentation with a much larger dataset. Nevertheless, the strong performance of the Transformer suggests that attention-based architectures remain an encouraging path for larger motion capture datasets.

7. Analysis

Permutation importance evaluation showed that perturbing right-hand rotational channels yielded the largest decreases in weighted F1-score, indicating their central role in classification performance. If we consider that two of the signers who were recorded had their right hand as dominant, we might see this finding as consistent with phonological descriptions of sign languages, in which the dominant hand typically carries the primary articulatory load (Brentari, 1998; Sandler and Lillo-Martin, 2012). However, at the current stage we cannot be certain of the role played by signer's handedness. Further data collection and experiments should be performed to test this hypothesis.While the present study does not aim to validate linguistic theory, the correspondence between model-derived feature significance and phonological structure strengthens the understandability of the results.

Unsupervised clustering further supports this observation. The emergence of moderately separable clusters in PCA-reduced space suggests that sign instances retain structured organization even within lower-dimensional representations. The fact that the most influential PCA components correspond predominantly to right-hand rotations reinforces the importance of distal articulators in discriminating signs.

7.1. Redundancy and correlated joint movements

Correlation analysis showed very high correlations between anatomically adjacent joints, commonly exceeding 0.95. This superfluity is consistent with the mechanical dependencies intrinsic in articulated-body kinematics. Rotational changes in one segment of the kinematic chain naturally propagate to connected segments, resulting in highly correlated feature channels (Morasso, 2025).

From a machine learning perspective, such redundancy increases feature dimensionality without proportionally increasing discriminative information. This phenomenon likely contributes to the comparatively lower performance recorded when models are trained on the full 156-feature dataset. The substantial performance increases observed in the reduced-feature experiment suggest that excluding redundant or weakly informative channels can simplify the learning task and improve generalization.

7.2. Architectural implications

The comparative performance of the evaluated models provides insight into how temporal-spatial inductive biases interact with skeletal motion data.

The excellent performance of the TCN suggests that convolutional modelling of time sequences is

notably well-suited to motion-capture–based sign classification. Dilated temporal convolutions allow the model to integrate fine-grained articulatory movements with wider motion trajectories, while preserving local integration in feature space. This inductive bias appears advantageous in representations characterised by strong local correlations and structured redundancy.

In contrast, recurrent models process flattened feature vectors sequentially and do not explicitly encode spatial structure unless further mechanisms are introduced. Given the high degree of inter-joint correlation observed in the dataset, this lack of overt spatial modelling may limit the effectiveness of these models. Furthermore, LSTM (and BiLSTM) suffer from forgetting issues, especially with long sequences. Recurrent neural networks (RNNs) suffer from the vanishing gradient descent problem; Long-short term memory (LSTM) units were proposed to address the aforementioned forgetting issue, however they still struggle with long sequences. The work of (Bahdanau et al., 2015; Luong et al., 2015) proposes the attention mechanism for RNNs / LSMTs, reducing significantly the forgetting issue in such networks. The observed results suggest the presence of the forgetting issue and, consecutively, that the encoded sequences perhaps contain significant spatio-temporal variations.

Transformer-based models achieved competitive performance, indicating that attention components can successfully capture long-range dependencies in motion sequences. Nevertheless, their slightly lower performance relative to the TCN may reflect the limited dataset size, as attention-based architectures typically benefit from larger training corpora (Abdullayeva and Alishzade, 2025).

8. Conclusions

This work investigated the role of feature importance in motion capture representations for deep learning-based sign classification. Specifically, BioVision Hierarchy (BVH) motion-capture data where evaluated across multiple temporal models, including recurrent, convolutional, transformer-based, as well as graph-based models. The objective was to examine how feature saliency patterns relate to model performance and to assess whether high-dimensional skeletal representations are fully required for accurate sign recognition.

Taken together, the results show that motion-capture representations of isolated signs contain substantial structured redundancy, with distinctive information concentrated in a relatively small subset of distal articulatory features. This observation suggests that extremely high-dimensional skeletal representations may not be strictly indispensable for effective sign classification. Instead, targeted feature selection or structured dimensionality minimization may improve computational effectiveness while preserving, or even enhancing, predictive performance. Furthermore, the comparative results across architectures indicate that models that exploit local temporal–spatial coherence are particularly well suited to BVH-based sign modeling.

Limitations and future work At the same time, the aforementioned conclusions must be interpreted in light of the study's limitations. The small dataset consists (three participants signed 44 signs),[9], a limited number of participants (only 3) and was collected under monitored conditions, which restricts the generalizability of the findings. Validation on larger and more diverse sign-language corpora would be necessary to determine whether the observed architectural advantages and feature-importance patterns extend beyond the present dataset. Forthcoming experiments should includes a larger group of left-hand- and right-hand-dominant signers to allow for the thorough investigation of handedness. Furthermore, future studies should include analyses of by-signer attributes to get a more fine-grained understanding of feature importance linked to individual signers.

Bibliographical References

Gulchin Abdullayeva and Nigar Alishzade. 2025. A study on recurrent, attention-based, and hybrid neural architectures for sign language recognition. *Informatics and Control Problems*.

Jari Ivar Andersen, Gomer Otterspeer, Robert Belleman, and Floris Roelofsen. 2025. Designing a marker based motion capture setup for sign language research. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv*.

[9] Ideally this would imply 132 sign MoCap recordings; however, as not all signs were signed by all participants, our dataset includes 116 sign MoCap recordings

Edward Bergman, Lennart Purucker, and Frank Hutter. 2024. Don't waste your time: Early stopping cross-validation. *arXiv*.

Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. 2022. Towards automated sign language production: A pipeline for creating inclusive virtual humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 260–268, New York, NY, USA. Association for Computing Machinery.

Félix Bigand, Elise Prigent, and Annelies Braffort. 2020. Person identification based on sign language motion: Insights from human perception and computational modeling. In *Proceedings of the 7th International Conference on Movement and Computing*, MOCO '20, New York, NY, USA. Association for Computing Machinery.

Félix Bigand, Elise Prigent, Bastien Berret, and Annelies Braffort. 2021a. Decomposing spontaneous sign language into elementary movements: A principal component analysis-based approach. *PLOS ONE*, 16(10):1–18.

Félix Bigand, Elise Prigent, Bastien Berret, and Annelies Braffort. 2021b. Machine learning of motion statistics reveals the kinematic signature of the identity of a person in sign language. *Frontiers in Bioengineering and Biotechnology*, Volume 9 - 2021.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. Mit Press.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv*, abs/1812.08008.

Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. 2023. Towards the extraction of robust sign embeddings for low resource sign language recognition.

Maria-Paola Forte, Peter Kulits, Chun-Hao Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. 2023. Reconstructing signing avatars from video using linguistic priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12791–12801.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. 2023. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Alireza Jafari and Fatemeh Jafari. 2026. How much temporal modeling is enough? a systematic study of hybrid cnn-rnn architectures for multi-label ecg classification. *arXiv*.

Pavel Jedlička, Zdeněk Krňoul, Jakub Kanis, and Miloš Železný. 2020. Sign language motion capture dataset for data-driven synthesis. 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*, pages 101–106, Marseille, France. European Language Resources Association (ELRA).

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Pietro Morasso. 2025. Coordinating the redundant dofs of humanoid robots. *Actuators*, 14(7). 354.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 744–772. Association for Computational Linguistics.

Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2020. LSF-ANIMAL: A motion capture corpus in French Sign Language designed for the animation of signing avatars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6008–6017, Marseille, France. European Language Resources Association.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Syst. Appl.*, 164:113794.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Wendy Sandler and Diane Lillo-Martin. 2012. *Sign Language and Linguistic Universals*. Cambridge University Press.

Dimitar Shterionov, Lorraine Leeson, and Andy Way. 2024. The pipeline of sign language machine translation. In Andy Way, Lorraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*, pages 1–25. Springer Nature Switzerland, Cham.

Panagiotis Tsinganos, Bart Jansen, Jan Cornelis, and Athanassios Skodras. 2022. Real-time analysis of hand gesture recognition with temporal convolutional networks. *Sensors*, 22(5). 1694.

Vincent Vandeghinste, Mirella De Sisto, Santiago Egea Gómez, and Mathieu De Coster. 2024. Challenges with sign language datasets. In Andy Way, Lorraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*, pages 117–139. Springer Nature Switzerland, Cham.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Red Hook, NY, USA.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.