Isolated Sign Language Recognition Using Deep Learning

Barbora Ľapinová[*,†], Ľubomír Antoni[*,†] and Šimon Horvát[*,†]

Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 040 01 Košice, Slovakia

Abstract

Individuals with hearing or speech impairments rely on sign language as their main form of communication, yet a communication barrier between this community and rest of the population persists. Recognizing AI and Deep Learning's role in aiding communication for the deaf and hard of hearing, this paper investigates deep learning methods for isolated sign language recognition using a complex video dataset. A Convolutional Neural Network (CNN) is employed to classify signs, and a thorough analysis of the model's performance is conducted to uncover common misclassification patterns and particularly challenging sign categories. This research outlines future work integrating hand pose data to potentially enhance model robustness and accuracy. a Approach presented in this paper aims to improve sign language recognition systems.

Keywords

Sign language, Deep learning, Video classification

1. Introduction

Communication, the exchange of information, ideas, and emotions, is fundamental to everyday human interaction. While spoken language serves as the primary mode of communication for the majority, individuals with hearing or speech impairments rely on sign language as their main form of expression. Sign languages are natural, fully developed languages with independent lexicons and grammatical structures, not derived from spoken language. They are conveyed through a bodily-visual modality, involving hand shapes, movements, facial expressions, and body postures, making them fundamentally different from spoken languages, which rely on the vocal-auditory channel [1, 2, 3].

However, the bodily-visual nature of sign languages presents unique challenges for computational processing. Unlike speech, which can be recorded and analysed as linear audio, sign language is spatially and temporally rich, involving simultaneous visual cues across multiple articulators. This complexity creates a communication barrier between the signing community and the majority who use spoken language. Often, human interpreters are required to bridge this gap, but they are not always readily available. With the advent of intelligent systems capable of interpreting human actions, new possibilities have emerged—particularly the use of artificial intelligence (AI) to recognize and translate sign language automatically, thereby improving accessibility and societal inclusion.

Recent breakthroughs in AI and Deep Learning (DL) have revolutionized numerous fields—from healthcare and autonomous vehicles to personalized recommendation engines and voice assistants. Importantly, these technologies also offer transformative potential for improving the quality of life for marginalized communities, such as those who rely on sign language for communication [1, 2, 3].

Sign Language Recognition (SLR) has thus emerged as a crucial interdisciplinary research area, integrating computer vision, gesture recognition, and sequence modelling. Its primary aim is to enable automatic interpretation of sign language, either through isolated word recognition or continuous sentence-level translation [4, 5, 6, 7]. Despite progress over the past decade—particularly with the rise of deep learning and advanced neural architectures—SLR remains a complex and evolving task. The large number of signs, their subtle variations, and grammar-specific intricacies all contribute to the difficulty of developing robust recognition systems.

ITAT'25: Information technologies—Applications and Theory, September 26–30, 2025, Telgárt, Slovakia

*Corresponding author.

†These authors contributed equally.

barbora.lapinova@student.upjs.sk (B. Ľapinová); lubomir.antoni@upjs.sk (Ľ. Antoni); simon.horvat@upjs.sk (Š. Horvát)

This paper aims to contribute to the growing field of SLR by designing and evaluating a deep learning-based approach for isolated sign language recognition using a complex, diverse video dataset. We propose a custom Convolutional Neural Network (CNN) model and assess its performance through experiments on different subsets of this challenging dataset. A comprehensive error analysis is conducted using confusion matrices, identifying the worst-classified categories and the impact of label duplication on model learning. Finally, we outline future research directions, with a focus on integrating additional features such as hand skeleton tracking to capture fine-grained spatial information and enhance recognition accuracy.

2. Background and Related Work

2.1. Sign Language Structure and Characteristics

Sign languages are fully natural, richly expressive languages conveyed via the bodily-visual modality. They consist of *manual* elements—hand shape, orientation, movement, and location—and *non-manual* components—facial expressions, head movement, and body posture—which together form a sign and represent a *gloss*. Glosses serve as proxies for signs in annotation, yet lack a universal written form and vary widely across languages (e.g., ASL, BSL, CSL, GSL), complicating computational modeling [2, 8, 9].

2.2. Applications of AI in Sign Language Processing

Artificial Intelligence (AI) research in sign language has evolved into three interrelated domains:

- **Sign Language Recognition (SLR)**: identifying signs from video or sensor input [10, 11].
- **Sign Language Translation (SLT)**: mapping from sign inputs to grammatically correct spoken or written output [12, 13].
- **Sign Language Generation (SLG)**: synthesizing sign output (e.g., via avatars) [14, 15].

This paper focuses on *isolated SLR*, where each video depicts a single sign. Many studies emphasize combining computer vision, gesture recognition, and linguistic insights, though genuine datasets capturing varied lighting, signer appearance, and real-world contexts remain limited [16, 8]. Solving this requires interdisciplinary collaboration and richer, real-world data sources.

2.3. Deep Learning for Video Classification

Deep learning has become foundational in SLR due to its capacity to model complex visual and temporal patterns. Key architectures include:

Feed-forward Neural Networks (FNNs), which are well-suited for static data, but fall short in modelling spatial or temporal dependencies.

Convolutional Neural Networks (CNNs), that exploit spatial hierarchies via convolution and pooling, and models such as VGG, ResNet, and MobileNet are popular backbones [17, 16].

For video, these models have evolved into:

- **2D CNNs** applied per-frame with temporal pooling or fed into sequence models like LSTMs;
- **3D CNNs**, performing spatiotemporal convolutions across frames;
- **Hybrid CNN–RNN architectures**, e.g., CNN + attention-based LSTM, which have achieved accuracies over 84% on a challenging WLASL [18] dataset [16];
- **(2+1)D CNNs**, mixing spatial and temporal convolutions, sometimes in novel fused architectures [9].

Recent advances include transformer models like SHuBERT for self-supervised representation learning in ASL [19], and hybrid CNN–Transformer networks for isolated Chinese SLR [20]. Techniques emphasizing spatial-temporal trajectory awareness, such as CorrNet+, have shown state-of-the-art performance in continuous recognition and translation tasks [21].

These developments illustrate the rapid progress in applying modern deep architectures to SLR tasks, motivating our work with deep learning-based approaches.

3. Methodology

As mentioned earlier, this paper addresses the task of isolated sign language recognition which can be seen as a machine learning classification problem with the input in the form of a video on which the sign is performed, and the target output being the corresponding gloss.

Our methodology began with identifying and selecting a suitable video dataset for this task, followed by its preprocessing to prepare the data for model training. Next, we designed and implemented a neural network architecture specifically for this task. Due to the heuristic nature of neural network design, this process, along with its training and testing, was performed iteratively across several cycles to optimize performance.

3.1. Dataset Selection and Preprocessing

Choosing an appropriate dataset is one of the most crucial steps in solving any machine learning task. The most common form of input data in SLR is a video, which is represented by a sequence of images—video frames. A number of research groups, especially in the past, have used data obtained using various sensors. An example of such a sensor can be the data glove [22, 23], but nowadays, data in the form of a video that can be easily captured on a smartphone is far more practical for real-time SLR systems.

Datasets for isolated and continuous sign language recognition are not as abundant as, for example, image datasets for sign language alphabets, especially if we emphasize a sufficient number of glosses in the lexicon of the dataset, an adequate number of video samples in the dataset for these glosses, or the diversity of the individual videos.

In our work we chose the *Greek sign language* (GSL) dataset [2]. This dataset contains 40826 videos depicting signs corresponding to 310 unique glosses. However, one of the classes in this dataset contains only one video sample and was excluded from further processing—the resulting number of videos and classes was 40825 and 309, respectively. Seven signers are featured in the videos, and the individual signs may be considered common in communication of the users of sign language in healthcare or public administration.

Although the average number of video samples per class is approximately 132—suggesting a seemingly adequate dataset size—the distribution of samples across classes reveals significant imbalance. A more detailed analysis yielded the following insights:

- The **median** and **25th percentile** are both 35, indicating that at least 50% of the classes have **35 or fewer samples**, which is far below the average.
- The **75th percentile** is 105, meaning only 25% of the classes have more than 105 samples.
- The **minimum** number of samples in a class is 15, while the **maximum** is 2693, demonstrating a **severe skew** in the class distribution.

For clarity, these key statistics are summarized in the following table:

Due to the range of possible values in the number of videos per class which indicates an imbalanced dataset, the average is somewhat skewed, and more accurate information about the representation of the number of video samples is provided by the median, the 25% quantile and the 75% quantile.

The individual videos in this dataset can be found in 525 folders, with the name of each folder specifying one of the areas covered by the signs (those that may occur in communication at the police

Statistic	Value
Minimum number of samples per class	15
25th percentile (Q1)	35
Median (50th percentile)	35
75th percentile (Q3)	105
Maximum number of samples per class	2693
Mean (average) number of samples per class	~132

Table 1
Summary of video sample counts per class.

station, in healthcare, etc.), which sign language performer is featured in these videos, and also the repetition order of the recording of the particular sign by the respective sign language performer, as each of the signs is recorded multiple times for each sign language performer. An example of a folder name might be `health1_signer1_rep1_glosses`—this folder contains a portion of the healthcare signs, the videos feature the first signer, and it is the first repetition of the recordings of the given signs. In each of these folders there are several other folders—one for each video.

In the folder for one video, there are images in the `.jpg` format, which represent the video already divided into video frames, discarding those frames where the sign perfomer was idle (e.g., the beggining and the end of the video), since the authors of this dataset have performed part of the preprocessing (splitting the video into video frames, selecting appropriate video frames).

Our preprocessing of each video involved changing its height and width to 240×320, and 5 equidistant video frames were selected to represent each video to be fed into the model (e.g., for a video with 80 video frames, video frames at indices 0,20,40,60,80 were selected). The dataset contains 271 videos with a frame count of less than five, which were excluded from further classification, resulting in 40554 final samples. An example of a preprocessed input sequence can be seen in Figure 1.

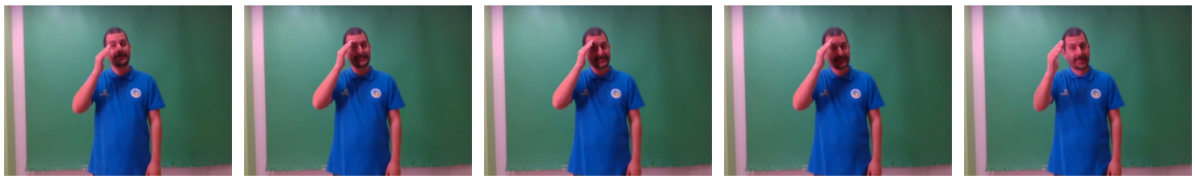

Figure 1: Example of an input sequence from the GSL dataset [2].

In this dataset, in addition to the videos themselves, there are also `.csv` files containing information about which gloss, i.e. class, corresponds to each video, or information about the bounding-box coordinates.

3.2. CNN Model Architecture

Our proposed model is a 3D CNN designed to classify sign language video sequences into 309 distinct classes. The network accepts a video input, represented by a tensor of shape $(5 \times 240 \times 320 \times 3)$, corresponding to five equidistant video frames with a spatial resolution of 240×320 pixels and three color channels.

The architecture begins with a 3D convolutional layer with 16 filters, each of size $(3 \times 3 \times 3)$, with a ReLU activation function. Next is a 3D max-pooling layer with a pooling window size of $(1 \times 2 \times 2)$, which performs spatial downsampling while preserving the temporal dimension. Then, a second 3D convolutional layer with 32 filters of size $(3 \times 3 \times 3)$ is applied, again with ReLU activation. This is followed by another 3D max-pooling layer with the same pooling dimensions.

The resulting feature maps are flattened into a one-dimensional vector of length 144768 and passed through a series of fully connected layers. First is a dense layer with 10000 neurons, followed by

additional two layers with 500 and 128 units. Each layer uses ReLU activation.

Finally, the network outputs a probability distribution over 309 target classes via a dense layer with 309 neurons and softmax activation. The model optimizes a categorical cross-entropy loss function and uses the ADAM optimizer.

The proposed model is the result of experiments with different network architectures and parameter settings. In the first unsuccessful attempts, we also experimented, for example, with different numbers of convolutional-pooling blocks and their hyperparameter settings.

Compared to the first experimental models, a notable improvement of our proposed model's performance occurred after removing the dropout layers, which usually improve the network performance, but in our case their occurrence led to its decrease. Enhancement was also noted after adding two dense layers with 10000 and 500 neurons between the dense layer with 128 units and the output layer. By adding additional dense layers a slower reduction of the vector that is the output of the flattening layer and the input to the output layer was ensured. The flattened vector with 144768 elements is reduced to a 128 element vector, which is input to the output layer gradually, not in a single step. Thus, the neural network has more room to select important features to use in classification.

4. Experiments and Results

4.1. Experimental Setup

In order to validate the proposed model on the GSL dataset, a case study was conducted to evaluate its performance on subsets of the dataset with respect to the number of target classes. In these subsets, we selected samples belonging to three, ten, and then all classes of the dataset.

The selection of categories for the subset of the dataset used in the three-class case study section was essentially arbitrary. However, care was taken to ensure these categories contained similar number of samples to simulate an ideal, balanced dataset. The subset contained the following target classes and their number of samples:

Class	Number of samples
ΒΙΒΛΙΟ (BOOK)	228
ΕΝΤΑΞΕΙ (OK)	244
ΤΑΞΙΔΙ (TRAVEL)	244

Table 2
Number of samples for classes in a three-class subset of the dataset.

To create a dataset with ten output classes, seven additional labels were added to the three classes used previously. The selected categories in this subset of the dataset displayed less uniform representation, yet the differences in number of samples between classes were not significant:

Class	Number of samples
ΤΑΥΤΟΤΗΤΑ (IDENTITY, ID CARD)	244
ΣΦΡΑΓΙΔΑ (SEAL, STAMP)	241
ΓΙΑ (FOR, ABOUT)	210
ΓΕΙΑ (HELLO)	208
ΚΙΝΗΤΟ (MOBILE, MOBILE PHONE)	208
ΕΥΧΑΡΙΣΤΩ (THANK YOU)	175
ΑΥΤΟΣ (HE, SHE, IT, SELF)	169

Table 3
Number of samples for classes in a ten-class subset of the dataset.

Generated subsets were in all cases divided into training and test sets using a stratified hold-out split ensuring that class distributions were preserved, and network performance was evaluated based

Classes	% of Samples for Testing	Train Samples	Test Samples	Batch Size	Epochs
3	15%	608	108	16	25
10	15%	1845	326	32	20
309	20%	32443	8111	128	18

Table 4
Dataset split and training parameters for case study sections.

on accuracy. Further analysis was conducted based on normalized confusion matrices. The network utilized batch learning for all three sections of the case study. The Table 4 provides a summary of the various network parameters including the proportion of samples used for testing the network, the number of epochs for its training, and the batch sizes.

4.2. Performance Analysis on Dataset Subsets

Table 5 displays the performance results of the proposed model, evaluated with respect to the maximum value of accuracy among all epochs.

In the section of the case study that involved training and testing the model's performance on a reduced dataset containing three categories with approximately equal representation, the model achieved a maximum accuracy value of 100.00% during both the training and testing phases. In the training and testing of the CNN on a subset of the dataset with ten categories, a 100.00% accuracy was achieved during training, and the model demonstrated a maximum accuracy value of 97.19% during testing.

In the case of the performance of the proposed model on the largest subset of the dataset containing 309 classes, the maximum accuracy value of 98.17% was achieved during the training phase, and 82.92% in the case of testing of the proposed CNN.

The observed decrease in the accuracy value on the test set when transitioning from training and testing on subsets of the dataset with fewer balanced classes to the full, unbalanced dataset with a larger number of classes may be attributed to several potential factors. The most likely factor that may have influenced this change is the uneven representation of samples across classes. It is possible that the model had more opportunities to learn samples from classes with more instances and fewer opportunities to learn samples from classes with fewer instances. This may have led to the model's poor performance in recognizing samples from the latter mentioned classes during testing. The elevated accuracy metrics observed in the training set, as compared to the testing set, might also be indicative of ineffective generalization by the model.

Classes	Train Accuracy	Test Accuracy
3	100.00%	100.00%
10	100.00%	97.19%
309	98.17%	82.92%

Table 5
Maximal accuracy across epochs in case study experiments.

4.3. Analysis of Model Error (Confusion Matrix)

Given that the results of the testing phase of the proposed model in the part of the case study in which the model was tested on the subset of the dataset with 309 target classes were no longer as optimal as in the previous two sections on smaller balanced subsets, a normalized confusion matrix was constructed for the trained model providing insight into which classes were challenging for the model.

In Table 6, we present the ten classes with the weakest performance in terms of the proportion of correctly classified samples from the class. These classes are sorted primarily by the percentage of samples correctly classified and secondarily by the class name, in ascending order.

As demonstrated in the table, the class EGΩ(3) exhibited the poorest performance, failing to achieve a correct classification for any of its test samples. A more detailed analysis for this class revealed that in 100% of the cases the samples from it were classified as the class ΕΓΩ(1). In fact, the GSL dataset contains three distinct versions of signs for the gloss ΕΓΩ (I, ME). The authors of the dataset in [2], state that these variations may be attributed to regional variations. A systematic search of all the classes in the dataset yielded three additional glosses to which multiple versions of signs belong.

This analysis resulted in the merging of the recurring classes into one. Specifically, the considered classes were ΕΓΩ(1), ΕΓΩ(2), ΕΓΩ(3), ΚΟΚΚΙΝΟ(1), ΚΟΚΚΙΝΟ(2), ΓΙΑΤΡΟΣ(1), ΓΙΑΤΡΟΣ(2), ΚΑΤΩ(1) and ΚΑΤΩ(2) for glosses ΕΓΩ (I, ME), ΚΟΚΚΙΝΟ (RED), ΓΙΑΤΡΟΣ (DOCTOR) and ΚΑΤΩ (DOWN, BELOW, UNDER) respectively. The classes representing the same gloss were merged into a single category, such as ΕΓΩ(1), ΕΓΩ(2), and ΕΓΩ(3) into a common category ΕΓΩ, while the split of the samples into training and test sets was preserved.

Class	% Correctly Classified
ΕΓΩ(3) (I, ME)	0.00%
ΤΡΙΤΟΝ (THIRD)	9.52%
15	14.29%
ΠΡΟΣ (TO, TOWARDS)	14.29%
ΕΓΩ(2) (I, ME)	15.38%
ΚΕΠΑ(Δ.Α.) (DISABILITY CERTIFICATION CENTRE)	23.81%
200	28.57%
ΑΚΟΥΩ_ΜΕΙΩΝΩ (HEARING LOSS)	28.57%
ΑΥΤΗ_ΤΗ_ΣΤΙΓΜΗ (RIGHT NOW)	28.57%
ΓΥΝΑΙΚΟΛΟΓΟΣ (GYNECOLOGIST)	28.57%

Table 6
Ten worst classified classes.

This modification resulted in 304 classes from the previous 309 classes. The same neural network was then trained on this modified data, with the exception of changing the number of output layer neurons from 309 to 304, again at 18 epochs. The performance of the network was evaluated using accuracy, and comparison of the maximum accuracy achieved during training and testing of the model before and after merging the classes is summarized in the Table 7.

The maximal accuracy value achieved during the training phase was 98.87% (previously 98.17%) and during the testing phase was 83.16% (previously 82.92%). Merging the classes thus results in a slight increase in the maximal accuracy value. However, it should be noted that 300 other classes, in addition to the ones that were merged, contribute to these resulting values. Therefore, an analysis of the normalized confusion matrix was conducted again.

Phase	Accuracy before merging	Accuracy after merging
Train	98.17%	**98.87%**
Test	82.92%	**83.16%**

Table 7
Classification accuracy before and after merging classes.

Table 8 shows the number of classes with the correct classification percentage under different thresholds. It can be seen that the number of filtered classes decreased after merging the classes for each threshold. Merging the classes, especially the problem class ΕΓΩ(3), may have given the neural network more room to learn and improve the classification of the remaining classes.

Threshold	Before merging	After merging
80%	155	122
75%	137	108
70%	94	65
65%	91	62
60%	84	59
55%	45	33
50%	40	29

Table 8
Number of categories with correct classification percentage below the selected threshold, before and after merging.

Table 9 shows the ten worst-classified categories after merging the repeating classes. It is sorted in the same way as Table 6. Except for class 15, whose percentage of correctly classified samples increased from 14.29% to 28.57%, the structure of the worst-classified classes changed after merging the repeating categories. For instance, the classification of classes such as ΤΡΙΤΟΝ (THIRD), ΠΡΟΣ (ΤΟ, TOWARDS), and 200 has improved; these classes are no longer among the ten worst. However, after merging, the class ΤΕΤΑΡΤΟΝ (FOURTH), which was not among the ten worst-classified classes before merging, was added. This change may not have been directly caused by merging of the classes since several factors affect classification, such as the initialization of parameters in the network, which was in our case random.

Class	% Correctly Classified
ΜΟΝΟ (ONLY, ALONE)	14.29%
ΤΕΤΑΡΤΟΝ (FOURTH)	21.43%
1000	28.57%
15	28.57%
2	28.57%
ΑΛΗΘΕΙΑ (TRUTH)	28.57%
ΑΝΑΘΕΤΩ (ASSIGN, DELEGATE)	28.57%
ΑΞΙΟΛΟΓΗΣΗ (ASSESSMENT)	28.57%
ΕΚΤΥΠΩΝΩ (PRINT)	28.57%
ΝΟΣΟΚΟΜΕΙΟ (HOSPITAL)	28.57%

Table 9
Ten worst-classified categories after merging repeated categories.

4.4. Comparison with State-of-the-Art Models

In their paper, Adaloglou et al. [2] present the GSL dataset for both isolated and continuous SLR. They also use the models proposed in [24, 25, 26] for the classification tasks on this dataset. These neural network architectures were primarily utilized for continuous SLR; however, their performance was also evaluated on the isolated SLR version of the GSL dataset.

Table 10 presents the results of our proposed model and the models introduced in [2]. Considering these results, it can be concluded that our solution achieves similar results to those of the far more complex models. However, it is important to note that our model was trained and tested using different training and testing sets than those used for the presented models. Therefore, to objectively evaluate our model in the future, we propose training and validating all models using the same subsets of the dataset.

Model	CNN Type	Accuracy
GoogLeNet + TConvs	(2+1)D	86.03%
3D-ResNet + BLSTM	3D	86.23%
I3D + BLSTM	3D	**89.74%**
Ours (before merging repeated classes)	3D	82.92%
Ours (after merging repeated classes)	3D	83.16%

Table 10
Comparison of our model with different architectures on the GSL dataset.

5. Conclusion

5.1. Discussion of Findings

In this paper, we have addressed the problem of isolated sign language recognition. To solve this task using a deep learning approach, we proposed a 3D convolutional neural network model, selecting the complex GSL dataset for the purpose of its training and testing.

Due to the considerable number of classes in the GSL dataset, a case study was performed on the proposed model. The model's performance was evaluated on two smaller, balanced subsets of the dataset, for which optimal results were obtained during training and testing.

In the context of training and testing the network on the full, imbalanced dataset, the testing results did not align with the results achieved during network training. This inconsistency may be attributed to the uneven representation of the classes. A thorough analysis of the confusion matrix revealed that certain glosses were associated with multiple signs, suggesting the presence of regional variations characterized by subtle changes. As a result of this analysis, the sign variations corresponding to the same gloss were merged into a single class. Following this transformation, an improvement in the accuracy value was observed, accompanied by an enhancement in the analysis outcomes for the problematic classes.

The added value of this paper stems from the proposed methodology for addressing duplicate or highly similar categories in sign language datasets. By systematically analyzing the confusion matrix, we identified hard examples (frequently misclassified classes such as regional variants of the same gloss) and soft examples (classes with consistent but subtle overlaps). Based on this analysis, we introduced a principled approach for merging categories, which not only improved the recognition accuracy of our CNN model but also provided a more realistic treatment of sign variation in computational systems. This methodology represents a practical contribution toward building more robust SLR pipelines, especially for datasets where gloss definitions are not strictly standardized or where regional sign variants coexist. Beyond the scope of this study, such an approach could be generalized to other sign languages and multimodal datasets, thereby reducing annotation noise while still respecting linguistic diversity.

In consideration of the results obtained, it is possible to conclude that the proposed model can be considered an effective tool for isolated sign language recognition systems with smaller, thematically focused lexicons. In a real-world setting, the potential applications of this model could include the recognition of signs in healthcare communication, where a limited, specifically defined repertoire of signs is typically used.

The proposed model is limited in two aspects. First, the training process is time-consuming. Second, the model receives a limited number of video frames as input. In the context of training and testing, almost entire GSL dataset was used, excluding videos with less than 5 frames and a video with no other samples in its class with a training set consisting of 32433 examples and a test set containing 8111 examples. The duration of one epoch was found to be approximately five hours. Therefore, another possible future task is to optimize it with respect to computation time, as well as to experiment with accepting a higher number of video frames as input.

5.2. Future Work

In a series of papers, including [27, 28, 29, 30, 31], supplementary information such as depth or joint information is utilized to enhance the sign dynamics information in the neural network, yielding a multimodal input stream.

Following our work, initial steps were taken for the incorporation of joint information for hands—hand skeletons. The creation of hand skeletons has already been explored through experimentation with the Hands model from the MediaPipe library [32]. The result of the experimental hand skeleton creation for a sequence of video frames can be seen in Figure 2.

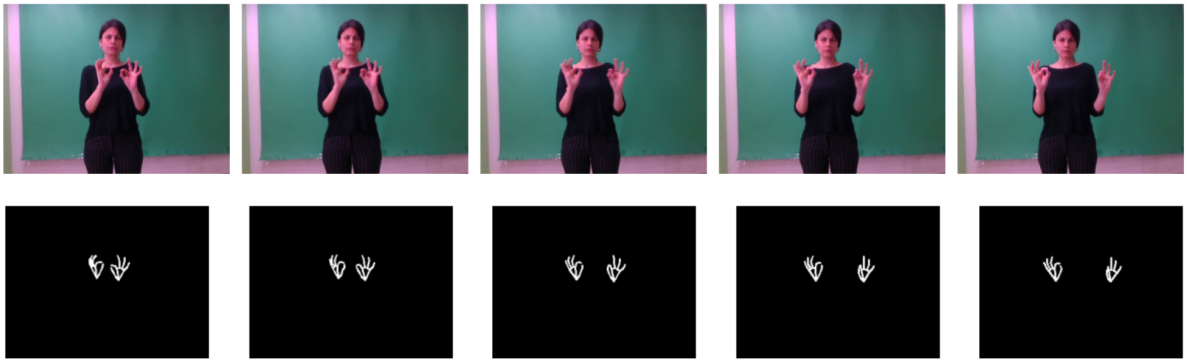

Figure 2: Example of a generated hand skeleton sequence for a video from the GSL dataset [2].

It is essential to acknowledge that the sequence of hand skeletons serves merely as a supplementary piece of information, and it cannot entirely replace the input sequence of video frames. This is due to the fact that the skeleton images neglect some characteristics of the signs, such as the facial expression or the overall posture of the signer, and only emphasize the most basic parts of the signs—the shape of the hands and the movement they perform.

The integration of such information could potentially enhance the performance of the model by providing an additional perspective on the signing process, particularly in capturing temporal patterns and articulations that might otherwise be overlooked. However, the reliability and practical usage of such an approach remains to be tested.

Beyond the integration of hand pose information, another promising research direction lies in exploiting recent advances in large multimodal language models (LLMs). For example, the **SignCLIP** model leverages contrastive pretraining on paired sign-language videos and spoken-language text, learning joint video–text embeddings that support both video-to-text and few-shot recognition tasks across sign languages [33]. Leveraging such automatically generated textual representations could provide complementary supervision signals for sign recognition models, potentially improving generalization across domains and signers. Moreover, pretraining on large-scale gesture-to-text corpora may help bridge the gap between visual sign representations and linguistic meaning, opening pathways toward richer sign language translation systems. We thus see the integration of multimodal representation learning—not only skeletal features but also language-guided supervision—as a highly relevant direction for future research in sign language recognition.

Another limitation of the present study is that we restricted our evaluation to a custom 3D CNN model. While we compared our results to architectures reported in the literature, those models were often trained and validated using different experimental setups, making direct comparison less reliable. For a fairer benchmark, it will be necessary to re-implement and train alternative architectures such as CNN–LSTM hybrids or Transformer-based models (e.g., vision transformers or multimodal transformers) on the same dataset splits. Recent works have shown that such architectures can effectively capture long-range temporal dependencies and multimodal context in sign language videos [12, 34]. We consider systematic evaluation across these model families, under unified conditions, as an essential direction for future work.

Acknowledgments

This article was supported by the *Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic* under contract VEGA 1/0539/25.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. I. Stamoulis, Sign Language Detection, Master's thesis, University of West Attica, 2023.
[2] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, P. Daras, A comprehensive study on deep learning-based methods for sign language recognition, IEEE Transactions on Multimedia 24 (2021) 1750–1762.
[3] D. M. Perlmutter, What is sign language, Linguistic Society of America 1325 (2011) 20036–6501.
[4] T. Tao, Y. Zhao, T. Liu, J. Zhu, Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges, IEEE Access (2024).
[5] M. J. Cheok, Z. Omar, M. H. Jaward, A review of hand gesture and sign language recognition techniques, International Journal of Machine Learning and Cybernetics 10 (2019) 131–153.
[6] R. Elakkiya, Retracted article: Machine learning based sign language recognition: a review and its research frontier, Journal of Ambient Intelligence and Humanized Computing 12 (2021) 7205–7224.
[7] J.-H. Kim, C. Ko, M. Huerta-Enochian, S. Y. Ko, Shedding light on the underexplored: Tackling the minor sign language research topics, in: Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, 2024, pp. 147–158.
[8] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, B. F. Alkhamees, Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs, Mathematics 11 (2023) 3729.
[9] A. M. Sultan, W. M. M. Zaki, M. Kayed, A. M. A. Ali, Multiple Sign Language Identification Using Deep Learning Techniques, Sci. J. Circuits Syst. Signal Process. 11 (2023) 1–11. doi:10.11648/j.cssp.20231101.11.
[10] O. Koller, J. Forster, H. Ney, Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers, in: Computer Vision and Image Understanding, volume 141, Elsevier, 2015, pp. 108–125.
[11] J. Huang, W. Zhou, Q. Wu, H. Li, Video-based sign language recognition without temporal segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[12] N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10023–10033.
[13] J. Ko, S. Cho, Neural sign language translation based on human keypoint estimation, Applied Sciences 9 (2019) 2683.
[14] B. Zhou, L. Shi, Y. Cui, X. Wang, J. Cheng, H. Lu, Spatio-temporal graph convolutional network for skeleton-based sign language recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 600–608.
[15] S. Stoll, N. C. Camgoz, S. Hadfield, R. Bowden, Text2sign: Towards sign language production using neural machine translation and generative adversarial networks, in: International Journal of Computer Vision, volume 128, Springer, 2020, p. 2515–2530.
[16] D. Kumari, R. S. Anand, Isolated Video-Based Sign Language Recognition Using a Hybrid

CNN-LSTM Framework Based on Attention Mechanism, Electronics 13 (2024) 1229. doi:10.3390/electronics13071229.

[17] J. Sharma, K. S. Gill, M. Kumar, R. Rawat, Deep Learning for Sign Language Recognition: Exploring VGG16 and ResNet50 Capabilities (2024). doi:10.56155/978-81-955020-9-7-13.

[18] D. Li, C. Rodriguez, X. Yu, H. Li, Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1459–1469.

[19] S. Gueuwou, X. Du, G. Shakhnarovich, K. Livescu, A. H. Liu, SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction, arXiv preprint arXiv:2411.16765 (2024).

[20] S. Jing, G. Wang, H. Zhai, Q. Tao, J. Yang, B. Wang, P. Jin, Dual-view Spatio-Temporal Feature Fusion with CNN-Transformer Hybrid Network for Chinese Isolated Sign Language Recognition, arXiv preprint arXiv:2506.06966 (2025).

[21] L. Hu, W. Feng, L. Gao, Z. Liu, L. Wan, CorrNet+: Sign Language Recognition and Translation via Spatial-Temporal Correlation, arXiv preprint arXiv:2404.11111 (2024).

[22] D. L. Quam, G. B. Williams, J. R. Agnew, P. C. Browne, An experimental determination of human hand accuracy with a dataglove, in: Proceedings of the Human Factors Society Annual Meeting, volume 33, SAGE Publications Sage CA: Los Angeles, CA, 1989, pp. 315–319.

[23] A. Z. Shukor, M. F. Miskon, M. H. Jamaluddin, F. bin Ali, M. F. Asyraf, M. B. bin Bahar, et al., A new data glove approach for malaysian sign language detection, Procedia Computer Science 76 (2015) 60–67.

[24] R. Cui, H. Liu, C. Zhang, A deep neural framework for continuous sign language recognition by iterative training, IEEE Transactions on Multimedia 21 (2019) 1880–1891.

[25] J. Pu, W. Zhou, H. Li, Iterative alignment network for continuous sign language recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4165–4174.

[26] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[27] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3413–3423.

[28] J. Huang, W. Zhou, H. Li, W. Li, Sign language recognition using 3d convolutional neural networks, in: 2015 IEEE international conference on multimedia and expo (ICME), IEEE, 2015, pp. 1–6.

[29] J. Zhang, Q. Wang, Q. Wang, Z. Zheng, Multimodal fusion framework based on statistical attention and contrastive attention for sign language recognition, IEEE Transactions on Mobile Computing 23 (2023) 1431–1443.

[30] D. Laines, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, G. Bejarano, Isolated sign language recognition based on tree structure skeleton images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 276–284.

[31] A. S. M. Miah, M. A. M. Hasan, S.-W. Jang, H.-S. Lee, J. Shin, Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition, Electronics 12 (2023) 2841.

[32] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, M. Grundmann, Mediapipe hands: On-device real-time hand tracking, arXiv preprint arXiv:2006.10214 (2020).

[33] Z. Jiang, G. Sant, A. Moryossef, M. Müller, R. Sennrich, S. Ebling, Signclip: Connecting text and sign language by contrastive learning (2024). arXiv:2407.01264.

[34] A. Brettmann, J. Grävinghoff, M. Rüschoff, M. Westhues, Breaking the barriers: Video vision transformers for word-level sign language recognition (2025). ArXiv preprint.