29th International Conference
ENGINEERING MECHANICS 2023
Milovy, Czech Republic, May 9 – 11, 2023

MEDIAPIPE AND ITS SUITABILITY FOR SIGN LANGUAGE RECOGNITION

Šnajder J.*, Krejsa J.**

Abstract: *The paper presents the framework MediaPipe as a tool to extract pose features for the task of word-level isolated sign language recognition. It tests the framework's suitability on the state-of-the-art sign language dataset AUTSL. Extracted sequences of pose features are classified by the Long Short-Term Memory recurrent neural network constructed with the TensorFlow computational library. The paper describes the proposed method flow, preprocessing of the extracted features, and training. Obtained results are then validated on test datasets, and the impact of face landmarks is evaluated. The top-1 accuracy with face landmarks is 49.89 %, while 53.21 % without them.*

Keywords: Sign language recognition, MediaPipe, Long Short-Term Memory, neural network, classification.

1. Introduction

Sign language is a very complex means of communication; it joins together not only hand movements and gestures but also body language and facial expressions. It is a natural language for the hearing-impaired, and besides being their way of communication, it is also their way of thinking. Unlike its spoken counterpart, sign language heavily relies on the visual sensory system. However, the two also have similarities; both types of languages could be understood as word-based, i.e., sign language gestures can be roughly interpreted as words.

Sign language recognition (SLR) has been an object of many research papers. Early approaches were inspired by successes in speech recognition and were based on traditional image processing methods to obtain features and hidden Markov model-based classifiers. However, the feature extraction by per-image methods lacked robustness. This changed with the rise of convolutional neural networks (CNN), which proved suitable for this task.

Nowadays, most SLRs use CNNs for feature extraction, and they can be split into appearance-based and pose-based methods. The appearance-based methods use a neural network in an end-to-end fashion. The first part is CNN to extract the features, and the second classifies the features in their raw form. The advantage is that the feature extractor is tailored for the SLR, and the model can differentiate between similar signs; however, these approaches require an extensive training set. The pose-based methods extract the features as a per-image pose of the signer. The features are, in this case, defined key points in the pose, which lowers the impact of appearance-based factors, such as different environments, clothes, etc. Furthermore, posed-based methods are usually less computationally demanding and thus more suitable for mobile devices. Their disadvantage is that some appearance-based factors are lost, making it more difficult for the classifier to distinguish between similar gestures.

Since our previous work on Czech sign language alphabet classification (Šnajder & Bednařík, 2022) (Šnajder & Krejsa, 2022) was based on pose extractor MediaPipe (Lugaresi et al., 2019), we propose a

* Ing. Jan Šnajder: Institute of Solid Mechanics, Mechatronics and Biomechanics, Brno University of Technology, Technická 2896; 616 69, Brno; CZ, jan.snajder@vutbr.cz

** doc. Ing. Jiří Krejsa PhD.: Institute of Solid Mechanics, Mechatronics and Biomechanics, Brno University of Technology, Technická 2896; 616 69, Brno; CZ, krejsa@fme.vutbr.cz

ISSN: 1805-8256, doi: 10.21495/em2023-251

pose-based method. Its advantages align with our aim to provide a real-time solution for mobile devices. This paper focuses on the framework MediaPipe, and we inspect its suitability for the task of SLR. The feature extractor is tested on Ankara University Turkish Sign Language (AUTSL) (Sincan & Keles, 2020), the state-of-the-art dataset for SLR, which was used for the CVPR21 competition (Sincan et al., 2021). The proposed method uses Long Short-Term Memory (Hochreiter. & Schmidhuber, 1997) neural network to classify the pose-based features.

In the following, we describe the core parts of the method in detail before discussing the datasets and their preprocessing. Finally, we present and analyze obtained results.

2. Methods

The proposed method is shown in Fig. 1. To extract pose-based features from the images, the algorithm uses MediaPipe Hands (MPH) (Zhang et al., 2020) and MediaPipe Pose (MPP) (Bazarevsky et al., 2020). Both these tools are part of the MediaPipe framework for building cross-platform machine-learning solutions. Both are based on CNNs, which are trained to extract the coordinates of multiple key points. While MPH generates 21 hand landmarks in essential parts of the palm and fingers, the MPP produces 32 landmarks all over the body. Since sign language uses primarily only the upper body, we omit landmarks from the lower body. MPP also creates landmarks in the face; although they are only a few and not detailed, we decided to inspect its impact on the final classification. Using these two frameworks, we obtain coordinates of necessary arms and face landmarks together with landmarks of both hands – altogether, 65 pairs of x and y coordinates.

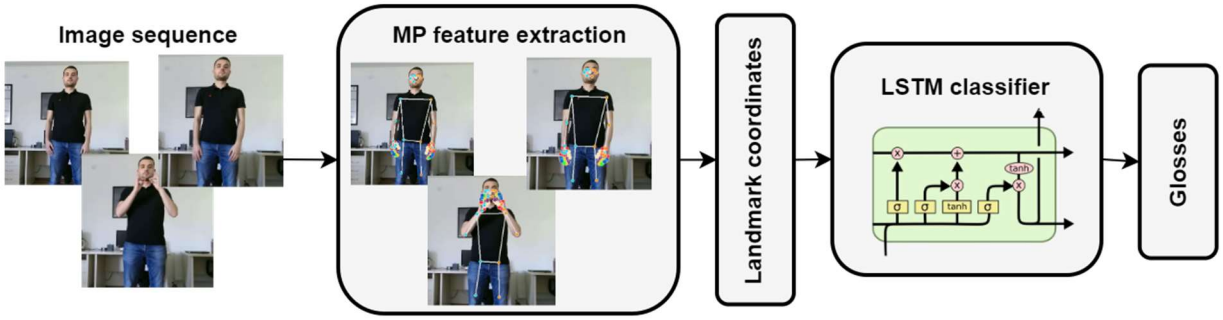

Fig. 1: The proposed method.

Sequences of these coordinates are then used as input to the LSTM network, where each coordinate represents one feature. Long Short-Term Memory networks are a type of recurrent neural network (RNN). Compared to their predecessors, they are designed to avoid the vanishing gradient problem, and thus, they can learn long-term dependencies. The core idea of the LSTM cell is the so-called *cell state* which acts as maintained memory. The *cell state* is interfaced by three *gates*: *forget*, *input*, and *output*. These *gates* decide what to remove, add and propagate from the *cell state*. *Gates* consist of neural network layers, and their parameters are trainable. The proposed algorithm uses LSTM for classifying sequences of pose coordinates. The details of the topology itself are described in the chapter Training.

2.1. Data Acquisition

The proposed algorithm was tested on the AUTSL dataset. It focuses on repeating fewer signs in a controlled environment by professional and amateur signers. The dataset captures 43 signers, including hearing impaired, Turkish sign language (TSL) translators, instructors, and students. Ten of these signers are men, thirty-three are women, and two of the signers are left-handed. The dataset contains depth information; however, since our goal is to use only a monocular camera, the depth data was not used. The statistical summary of the dataset is shown in Table 1.

Tab. 1: Summary of AUTSL dataset.

Glosses	Signers	Samples	Backgrounds	Mean	Resolution	FPS
226	43	38,336	20	169.6	512x515	30

2.2. Preprocessing

Preprocessing consists of MediaPipe annotating the coordinates of pose key points. These coordinates were then transformed to be relative to one of its parts; namely, for hands, it was the wrist coordinate; for arms, it was the shoulder coordinates; and for face, the nose coordinates were used. This step ensured that the inputs were invariant to the signer location in the image.

All videos were annotated twice, with and without face landmarks. Although facial expressions are essential to sign language, we were unsure about the impact of plain face landmarks generated by MPP. If the effect is meaningful, there should also be consideration using MediaPipe Face Mesh (Kartynnik et al., 2019).

2.3 Training

Several topologies were tested, varying in the number of cells in each layer, dropout setup, etc. The topology which exhibited the most promising behavior is shown in Table 2. The TensorFlow library (Abadi et al., 2015) in Python was used to create such a neural network. The batch size for the training was set to 256 sequences, the stochastic optimization algorithm Adam was used as the training algorithm, and sparse categorical cross-entropy was used as the loss function. The number of epochs was set to 100.

Tab. 2: Selected network topology (without/with face landmarks).

Layer	Size	Function	Output	Params
Masking	-	-	Sequences	-
LSTM	512	TanH	Sequences	1 271 808 / 1 316 864
Dropout	10 %	-	Sequences	-
LSTM	256	TanH	Sequences	787 456
Dropout	10%	-	Sequences	-
LSTM	128	TanH	Values	197 120
Full	226	SoftMax	Values	29 154

2.4 Results

For the performance evaluation of the model, we use the top-K absolute accuracy. This metric means that any of our model *K* highest probability answers must match the expected response, e.g., top-1 absolute accuracy is the conventional accuracy – the model answer must be exactly the expected answer. To align with the authors of the AUTSL dataset, besides the top-1, we also consider the top-3 and top-5. The number of different signs in the dataset and the similarities between signs justify these additional metrics.

The overall results for all three metrics are shown in Table 3. The first row displays results without including face landmarks; the second row comprises face landmarks. The results show that the face landmarks provided by MPP do not bring additional information; on the contrary, it makes the results less accurate. This could be caused by the shallow representation of the face landmarks; it was observed that these landmarks are relatively inaccurate and just a complementary part of MPP.

From the metrics point of view, there is roughly a 25 % difference between top-1 and top-5 absolute accuracy. This confirms that some signs are very similar. We compared our results to those presented by the dataset authors (Sincan & Keles, 2020) as they use the same LSTM classifier. The main difference between our approaches is the feature extractor, where the paper uses an appearance-based approach with CNN. Their baseline results with the LSTM classifier were 23.00 %, 37.03 %, and 43.66 % for top-1, top-3, and top-5 metrics, respectively. They further improved the feature extraction and classifier, however even with the approach combining CNN, Feature Pooling Module (FPM), Bidirectional LSTM, and Attention model, they obtained 49.22 %, 68.89 %, and 75.78 % for the three metrics, which is comparable with results presented in our paper. This indicates that the posed-based approach has enormous potential as it is both computationally efficient and relatively accurate.

Tab. 3: Overall results for various metrics.

Metric	**Top-1**	**Top-3**	**Top-5**
Hands + arms	53.21 %	71.78 %	78.81 %
Hands + arms + face	49.89 %	69.59 %	76.19 %

3. Conclusions

The paper shows a successful application of the framework MediaPipe on word-level isolated sign language recognition. By joining landmarks extracted from the MediaPipe Hands and Pose with the classifier in the form of Long Short-Term Memory, we obtained an overall success rate of around 53 %. Compared to methods with similar classifiers, our approach generates better results and shows that the MediaPipe framework is suitable for feature extraction in this task.

The result analysis shows several important insights:

- Face landmarks generated by MPP do not provide any improvement. To capture the facial expression of the signer, a dedicated tool for extracting facial features should be used.
- There is a vast difference between top-1 and top-5 absolute accuracy (around 25 %). This suggests that the classifier has a good grasp of the sign's meaning; however, the differences between the signs are relatively small. This information can be used at higher application levels, and natural language processing methods can eliminate these misclassifications.
- The framework is still under active development (version 0.8.11 was used), and future updates may bring both optimizations and preciseness.

Future work will explore the different classifiers options and fine-tune the feature preprocessing. Furthermore, the results of this paper could be used to develop a real-time solution applicable to mobile devices.

Acknowledgment

This work was supported by FME Brno University of Technology under the project FSI-S-20-6407 "Research and development of modern methods for simulations, modelling and machine learning in mechatronics".

References

Abadi, M. et al. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems, Software available from tensorflow.org

Bazarevsky, V. et al. (2020) BlazePose: On-device Real-time Body Pose tracking.

Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory. Neural computation.

Kartynnik, Y. et al. (2019) Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs

Lugaresi, C. et al. (2019) MediaPipe: A Framework for Building Perception Pipelines.

Sincan, O. M. et al. (2021) ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research.

Sincan, O. M. & Keles, H. Y. (2020) AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods.

Šnajder, J. & Bednařík, J. (2022) Czech Sign Language Single Hand Alphabet Classification with MediaPipe. 27/28th International Conference – Engineering Mechanics 2022.

Šnajder, J. & Krejsa, J. (2022) Classification of Czech Sign Language Alphabet Diacritics via LSTM. 20th International Conference of Mechatronics – Mechatronika 2022.

Zhang, F. et al. (2020) MediaPipe Hands: On-device Real-time Hand Tracking.