27/28th International Conference
ENGINEERING MECHANICS 2022
Milovy, Czech Republic, May 9 –12, 2022
Paper **#42,** pp. 381–384, doi: 10.21495/51-2-381

CZECH SIGN LANGUAGE SINGLE HAND ALPHABET CLASSIFICATION WITH MEDIAPIPE

Šnajder J.*, Bednařík J.**

Abstract: *The paper presents the classification of static images of the single-handed Czech sign language alphabet. It uses the framework MediaPipe for annotation, and the classification is performed by a neural network using the TensorFlow computational library. The flow of the proposed method, data acquisition, preprocessing, and training are described in the paper. Obtained results consist of the classification success rate of the validation dataset for various MediaPipe configurations. The overall success rate was around 94%.*

Keywords: Czech sign language, Fingerspelling, Classification, Mediapipe, Neural network.

1. Introduction

Sign language is a way of communicating using hand gestures, movements, body language, and facial expression instead of spoken words. It is used primarily by hearing-impaired to effectively fulfill the same social and mental functions as spoken language. And like its verbal counterpart, there is not just one universal sign language, but rather the way of signing varies all over the world. Fingerspelling, a means to sign single letters, is an indispensable part of every sign language. It is often used to spell out names, places, and words that have no specific sign yet. It is also used during sign language teaching as it is an easy way to communicate if one does not know or cannot remember a specific sign. This paper introduces the classification of the Czech sign language alphabet letters from static images.

In the last decade, automatic sign language detection and classification have been an object of many research papers with different approaches to the problem. We propose an approach using monocular camera images as the only input source. The main aim of our method is a real-time classification of input images on smartphones, so it could find utilization in real-world applications and could help with the inclusion of hearing-impaired. We propose an algorithm that uses artificial neural networks (ANN) for classification as they can be used in a real-time application and show promising results in such tasks. Furthermore, with the usage of modern tools which implement ANN, such as TensorFlow (Abadi et al., 2015) and MediaPipe (Lugaresi et al., 2019), our solutions could be easily deployed to smartphones.

In the following, we describe the proposed classification method using the framework MediaPipe, before discussing the dataset and preprocessing. Finally, we present and analyze obtained results.

2. Methods

The flow of the proposed algorithm is shown in Fig. 1. The core of this method is MediaPipe Hands (MPH) (Zhang et al., 2020), which is part of the MediaPipe framework for building cross-platform machine learning solutions. MPH is high precision hand-landmark localization provided by Google, and it has been open-sourced to encourage researchers to develop production-ready machine learning applications. MPH consists of two convolution neural networks. The first one is trained to detect the hand region in an image, and the latter detects 21 hand landmarks in the detected hand region. MPH returns coordinates of these 21 landmarks and the handedness of observed hands (i.e., left or right hand). Coordinates consist of x and y values normalized to intervals from zero to one. The third coordinate z is relative to the wrist landmark.

* Ing. Jan Šnajder: Institute of Solid Mechanics, Mechatronice and Biomechanics, Brno University of Technology, Technická 2896; 616 69, Brno; CZ, 171291@vutbr.cz

** Ing. Josef Bednařík: Institute of Solid Mechanics, Mechatronice and Biomechanics, Brno University of Technology, Technická 2896; 616 69, Brno; CZ, 170217@vutbr.cz

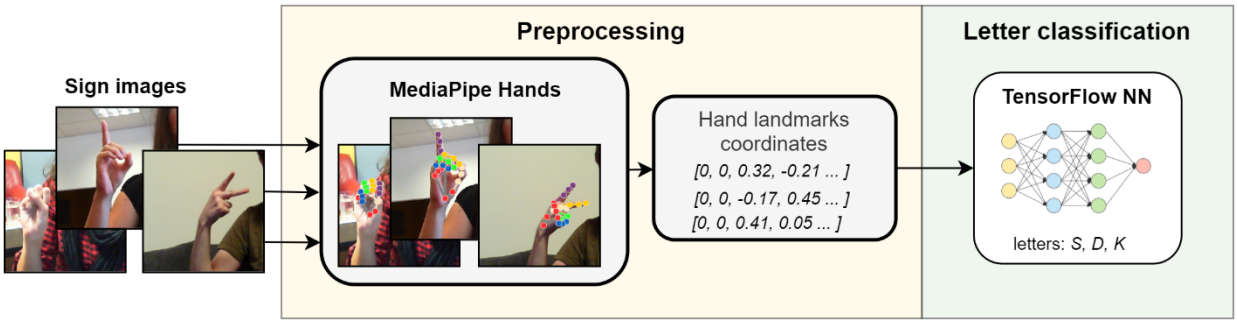

Fig. 1: The flow of the proposed algorithm.

The proposed algorithm uses MPH to annotate input images. Annotation consists of absolute x and y hand landmark coordinates, which are then transformed to be relative to the wrist coordinates. This transformation makes the classification indifferent to the location of the hand in the image.

An array of 42 or 63 coordinates (depending on whether the z coordinate is taken into consideration) is then used in a feedforward neural network, where each coordinate represents one feature. Used layers and the topology itself are described in chapter Training. The output of this network is one of the 27 classes, where each class represents one letter.

Data Acquisition

The data used for network training was introduced by Krejsa and Vechet (2020). It consists of sign images from 31 right-handed adult gesturers, both hearing impaired and professional Czech sign language interpreters. All images have a fixed size of 224x224 pixels, where each pixel has three channels (RGB). As the dataset already includes augmented images, no additional operations were done.

As stated in paper (Krejsa and Vechet 2020), the dataset contains training and validations sets. There are three validation sets; however, this paper uses only one denoted as PG. This validation set is the most comprehensive of the three as it consists of 1.181 images and is from an individual not included in the training set.

Preprocessing

According to the algorithm flow shown in Fig. 1, preprocessing consists of data annotation by MPH and coordinate transformation. Since MPH is the algorithm's core, its configuration greatly impacted obtained results. We preprocessed the training and validation sets with several configurations to find the most suitable one.

The first parameter, which impact was observed, is the number of dimensions. As stated earlier, MPH returns three coordinates. However, the z coordinate, which indicates depth, is mainly estimated, and we were unsure of its precision. The dataset was preprocessed by counting both 2D and 3D coordinates.

The second parameter is so-called *detection confidence.* It ranges from zero to one, indicating the confidence for the detection to be considered successful. *Detection confidence* default value is set to 0,50. However, we observed that higher values lead to more precise hand landmarks position estimation. The whole dataset was preprocessed with four *detection confidence* values. The number of successfully annotated images is shown in Fig. 2.

The graph shown in Fig. 2 gave us a few critical insights. The average number of images for each letter was around 21.000, and this number rapidly decreases with increasing *detection confidence*. For the highest inspected *detection confidence* 0,80, the average number of annotated images dropped to 12.536. Furthermore, the dataset lost its uniform distribution as MPH had severe problems with annotating several letters, especially letters "M", "N", and "Q". On the other hand, *detection confidence* 0.50 and 0.60 showed relatively high numbers of annotated images (20.046 and 18.979 respectively); however, portions of the annotations were incorrect due to MPH's inaccurate hand localization.

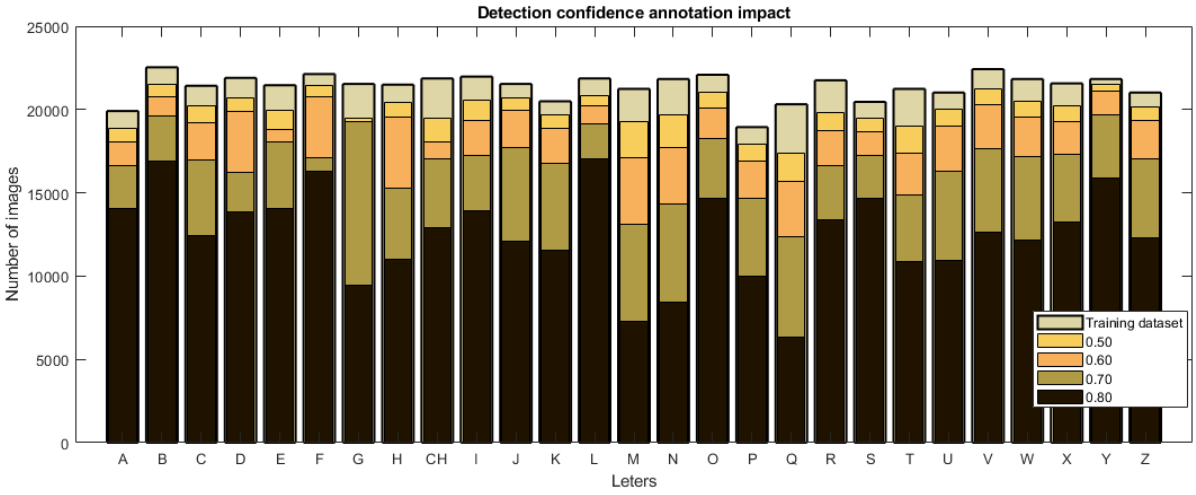

Fig. 2: The dependence of successfully annotated images on the detection confidence configuration

All annotated datasets were used in training further to examine the impact of *detection confidence* configuration on classification, although all insights mentioned above were considered in conclusion.

Training

Several topologies were tested, varying in size and number of fully connected layers and rate of dropout layers. The final topology of the network, which is later presented, consists of three fully connected layers with 512, 256, and 128 neurons, respectively. Each layer uses a hyperbolic tangent as an activation function followed by a 10% dropout layer. A fully connected classification layer with SoftMax activation function and 27 neurons (each representing a single letter) is used as the output layer.

The described network was created using the TensorFlow library. The network was trained with stochastic optimization algorithm Adam and a batch size of 4096 images. The course of training showed rapid growth of accuracy during the first few epochs. The training was stopped after 20 epochs as it did not lead to further improvements. All experiments were done on an AMD Ryzen based PC with 32 GB RAM running a 64-bit Windows 10 operating system. Actual computations were performed by two GeForce RTX 3060 Ti GPUs with 8 GB GDDR6 RAM.

Results

The test results for the PG validation data set are shown in Table 1. The second and fourth rows show the average successful classification using 2D and 3D coordinates, respectively. These results indicate that higher *detection confidence* means a higher classification success rate with 97,83 % for *detection confidence* 0,80. However, suppose we take successful MPH annotation into consideration. In that case, the classification success rate rapidly decreases for higher *detection confidence* configurations as configuration with 0,80 *detection confidence* successfully annotated only around 66% of validation images. Results with this consideration are shown on the third row for 2D and the fifth for 3D coordinates.

Tab. 1: Obtained results for various detection confidence configurations.

Detection confidence	**0,50**	**0,60**	**0,70**	**0,80**
x, *y* classification	96,35 %	96,51 %	97,47 %	**97,83 %**
x, *y* classification + annotations	**93,98 %**	91,44 %	84,92 %	64,94 %
x, *y*, *z* classification	96,79 %	97,41 %	97,38 %	**97,83 %**
x, *y*, *z* classification + annotations	**94,41 %**	92,29 %	84,85 %	64,94 %

There is not any significant difference in results between 2D and 3D coordinates. Although 3D coordinates show slightly better results, the estimation of the *z* coordinate is unpredictable, and its usage must be verified in real-time applications.

From a particular letters classification point of view, the classification success rate is mostly around 100 %. Exceptions are letters “M”, “N” and “Q”, which are not only very similar, as shown in Fig. 4, but also the shape and orientation of the hand during these signs make them challenging for MPH to annotate successfully. A similar situation also applies to letters “K” and “Z”, which, for some gesturers, slightly differs only in thumb position. These signs are also shown in Fig. 4.

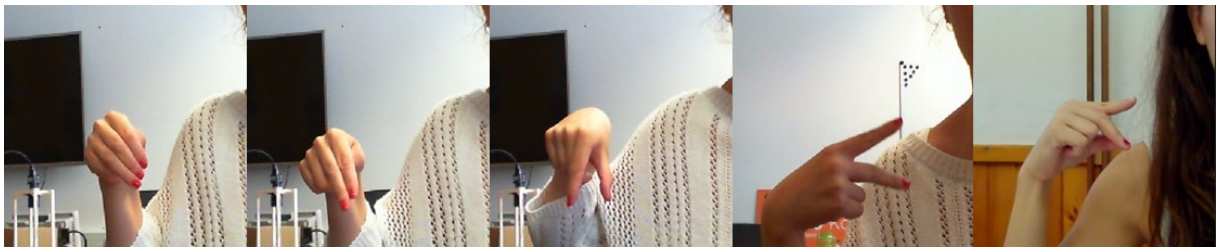

Fig. 3: Signs from the PG validation set for a particular letter, from left: “M”, “N”, “Q”, “K”, “Z”.

Obtained results prove that the role of a particular gesturer is not crucial as images from the gesturer of the PG validation set were not included in the training set. On the other hand, image quality is essential because it helps MPH successfully annotate a given image. Considering insights gathered from preprocessing and Table 1, configurations with *detection confidence* 0,50 or 0,60 seem most suitable for real-time usage. Although setups with *detection confidence* 0,70 and 0,80 showed higher classification success rates, they do not outweigh the negative that around one-third of the validation set was left unannotated.

3. Conclusions

The paper shows a successful application of the framework MediaPipe Hands to classify Czech sign language single hand alphabet letters. The average success rate of classification on all images from the validation data set is around 94 % for configuration with *detection confidence* set to 0,50.

Result analysis shows several important insights:

Although higher *detection confidence* means a higher classification success rate, it also means a lower annotation success rate.

Most misclassifications are concentrated around a few letters. This can be eliminated on a higher application level as we have the knowledge about similarities between signs.

Since MediaPipe Hands returns handedness, our approach can be used for left-handed gesturers as the only operation, which must be done, is flipping x coordinates.

The future work will be focused on implementing a real-time application, which would use results from this paper to decode fingerspelled words and sentences. Furthermore, we would like to classify diacritics as an important part of the Czech language. Both focuses will require the processing of image sequences and natural language processing.

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

Krejsa, J. and Vechet, S. (2020) Czech Sign Language Single Hand Alphabet Letters Classification. *19th International Conference on Mechatronics – Mechatronika (ME)*. https://doi.org/10.1109/ME49197.2020.9286667

Lugaresi, C. et al. (2019) MediaPipe: A Framework for Building Perception Pipelines. http://arxiv.org/pdf/1906.08172v1

Zhang, F. et al. (2020) MediaPipe Hands: On-device Real-time Hand Tracking. http://arxiv.org/pdf/2006.10214v1