26th International Conference
ENGINEERING MECHANICS 2020
Brno, Czech Republic, November 24 – 25, 2020

CLASSIFICATION OF CZECH SIGN LANGUAGE ALPHABET LETTERS USING CNN – PRELIMINARY STUDY

Krejsa J.*, Vechet S.*

Abstract: *The paper deals with the classification of Czech sign language single hand alphabet letters from static images using convolutional neural network (CNN). Proposed CNN architecture exhibits about 71 % successful rate of classifying the letters signed by the person not included in the training data set.*

Keywords: Classification, Sign language, Convolutional neural networks.

1. Introduction

Sign language is language used by hearing impaired for communication. Sign languages are natural languages with own grammar and lexicon (Sandler, 2006). Sign language are not universal; however, they use similar set of manual articulations. Its conversion is a difficult task not yet successfully solved. A number of methods has been proposed, using special instrumentation such as accelerometers on gloves/hands of the gesturer, see e.g. (González, 2018), or 3D sensing technology (Dong, 2015, Ma, 2016), which makes it impractical for real world use. Using monocular camera images as the only source of input is substantially cheaper and easier to use.

Sign language gestures detection can be used in general gesture detection applications. This paper is focused onto the classification of single hand Czech alphabet letters from static images. Czech alphabet has certain differences compared to international sign language alphabet, mainly by the presence of letter "CH", that is unique among alphabets. Furthermore, it has letters with diacritics (accents). Those are expressed by the motion of the corresponding sign and were not considered in this study. Another difficulty in the task arises from the fact, that some people use randomly reversed position of the hand with certain letters (typically letter Y), see examples in Fig. 1.

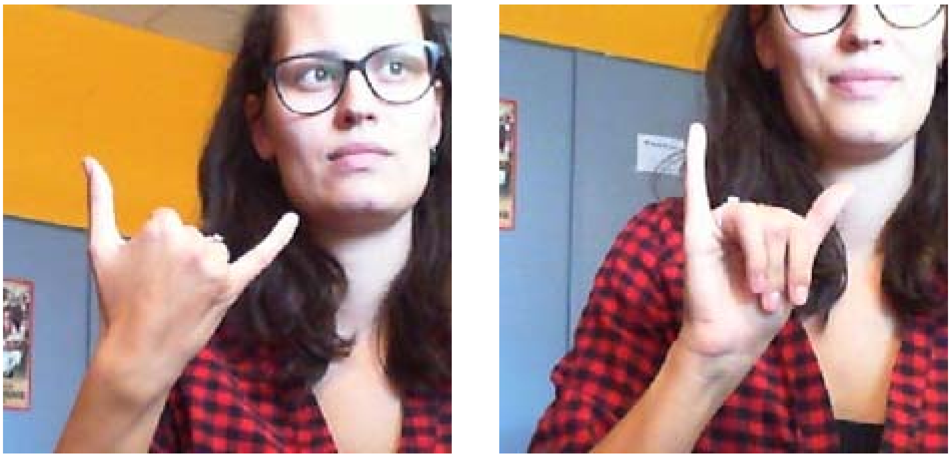

Fig. 1: Letter Y gesture by the same person during a single recording session.

The classification further described in this paper is based on convolution neural network (CNN) approach, with static image being the input to the network and probabilities of particular letter occurrence being the output.

* Assoc. Prof. Ing. Jiří Krejsa PhD., Assoc. Prof. Ing. Stanislav Věchet, PhD., Institute of Thermomechanics AS CR v.v.i. Brno department, Czech Republic, krejsa@fme.vutbr.cz

doi: 10.21495/5896-3-310

2. Methods

Due to vast possible variance in hand and fingers physiology, position and image background it is close to impossible solve the classification problem using traditional image processing techniques. On the other hand, deep learning techniques proved to present encouraging results in this type of task and were therefore chosen as the approach used.

2.1. Convolutional neural networks

Convolutional neural networks are multilayer perceptron type of neural networks that use convolution in some of its layers, typically in the first few layers in order to build and learn a set of filters to recognize automatically the features in the input, followed by common flatten fully connected layer(s) that handles the classification itself. Computations in this paper were all performed using TensorFlow library (Abadi, 2015).

2.2. Data recording and preprocessing

Images for training and validation were recorded using custom written application, that allows automatic labeling. The person performing the sign alphabet sees what letter should be signed and images recorded are labelled accordingly. Both hearing impaired persons and professional sign language interpreters recorded the images. Up to date data from 22 people were recorded. Only right hand letters were taken into account.

Individual images were further preprocessed using data augmentation. In particular the translation, rotation in range of -3 to +3 degrees, uniform scaling and non-uniform scaling in both vertical and horizontal direction. This way 30 additional images were made out of each source image. Augmentation was performed prior to the training to speed up the process. Images size was fixed 224 x 224 pixels using 3 channels (RGB).

Data were divided into training and validation sets so the validation set contained the data from one individual whose images were not included in the training set at all. In total 434 861 images were used for training and 43 697 images were used for validation.

2.3. CNN structure and training

A number of network architectures was trained and validated. The parameters of promising network are shown in Tab. 1.

Tab. 1: Structure of CNN.

Layer type	Depth	Activation	Output size	Number of parameters
Convolution 5 x 5	12	Relu	220 x 220 x 12	912
Pooling 2 x 2 (max)			110 x 110 x 12	
Convolution 5 x 5	12	Relu	106 x 106 x 12	3612
Pooling 2 x 2 (max)			53 x 53 x 12	
Convolution 5 x 5	12	Relu	49 x 49 x 12	3612
Pooling 2 x 2 (max)			24 x 24 x 12	
Convolution 3 x 3	12	Relu	22 x 22 x 12	1308
Pooling 2 x 2 (max)			11 x 11 x 12	
Convolution 3 x 3	14	Relu	9 x 9 x 14	1526
Pooling 2 x 2 (max)			4 x 4 x 14	
Fully connected	27	Softmax	224 x 27	6075

Total number of parameters in the network was 17045, all parameters were trainable. The training was performed until no significant improvement was seen in validation data. This usually did not exceed 20 epochs. Validation data contained augmented images to get better insight in cases where unusual physiological hand proportions occurred, however, the validation on fully augmented test data and source images only do not show significant differences, as will be illustrated in Results section.

3. Results

The network with structure shown in Tab. 1 reached 75.8 % accuracy on training data and 71.6 % and 71.3 % accuracy on validation data (augmented / source only). The comparison of the success rates for augmented and source data for particular letters in the alphabet are shown in Fig. 2. As one can see, there are small differences, but the general performance is the same. The only significant difference can be seen in letter M, where augmented data performance is degraded by the augmentation due to similar features of M and N letters.

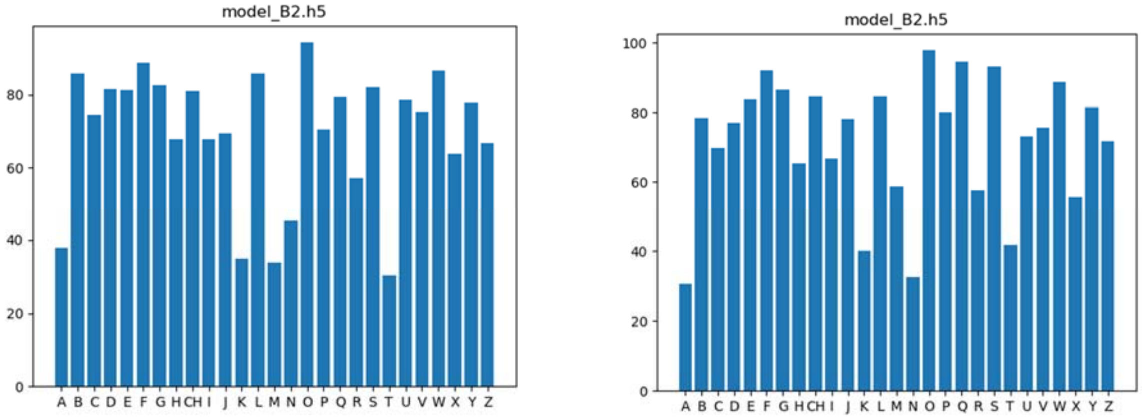

Fig. 2: Accuracy of the network on validation sets (augmented – left, source only – right) for particular letters in Czech single hand sign language alphabet.

We can furthermore take a look at the individual letters, as shown in Figs. 3 and 4, where the numerosity of detection of each letter for given input image are shown. Figs. 3 and 4. shows results for source images only.

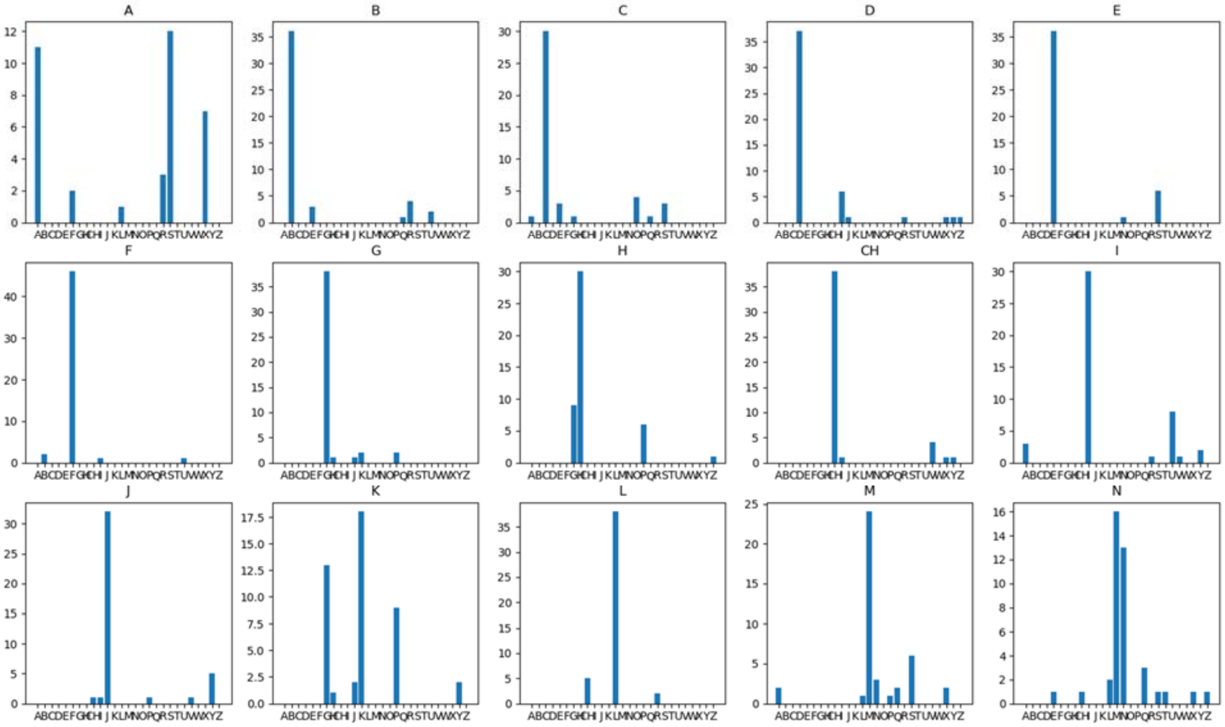

Fig. 3: Individual letters classification results A - N.

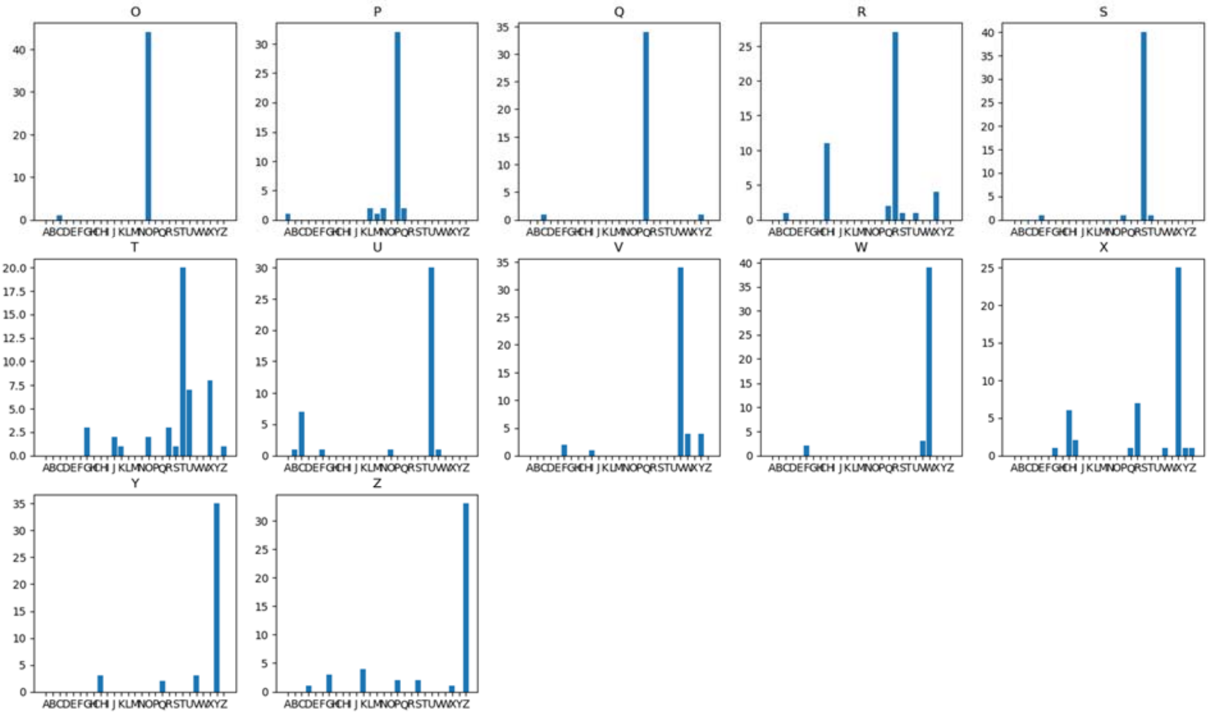

Fig. 4: Individual letters classification results O - Z.

We can see that letters with similar features exhibit the clear degradation of the results. This is particularly clear on letter N, with letter M being detected with higher frequency. However, such results are encouraging in our view, as it could be handled by other means, compared to the case when misdetection would be spread on the full alphabet.

4. Conclusions

Results shown are preliminary, as a number of techniques can be incorporated to improve the classification accuracy, such as techniques that prevent overfitting (regularization, dropout). Those techniques are useful for the larger nets with higher number of parameters that would tend to overfitting, but should increase the accuracy of trained model.

Acknowledgement

The results were obtained with institutional support RVO 61388998 of the Institute of Thermomechanics AS CR v.v.i. and with support of Mobility Plus Projects with Ministry of Science and Technology of AS CR, v.v.i., project no. MOST-20-06. The authors would like to thank T. Zbavitel for his help in recording the training data.

References

Sandler, W., Lillo-Martin, D. (2006) Sign Language and Linguistic Universals. Cambridge: Cambridge University Press.

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

González, G. S. et al. (2018) Recognition and Classification of Sign Language for Spanish, Computación y Sistemas, vol. 22, no. 1, pp. 271-277.

Dong, C., Leu, M., Yin, Z. (2015) American sign language alphabet recognition using microsoft kinect, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 44-52.

Ma, L., Huang, W. (2016) A static hand gesture recognition method based on the depth information, in: Intelligent Human-Machine Systems and Cybernetics (IHMSC), 8th International Conference on, vol. 2, IEEE, pp. 136-139.