30th International Conference

ENGINEERING MECHANICS 2024

Milovy, Czech Republic, May 14 – 16, 2024

MOTION ANALYSIS OF CZECH SIGN LANGUAGE ALPHABET CNN BASED CLASSIFIER VIA OPTICAL FLOW

Krejsa J.[*], Věchet S.[**], Šnajder J.[***]

Abstract: *Single hand sign language alphabet letters can be successfully detected and classified from a single still image using convolution neural networks (CNN). Czech variant of the alphabet uses specific motions to add diacritics to the letters. The paper deals with determination of such a motion using optical flow analysis of image sequences.*

Keywords: Diacritics detection, sign language, gesture recognition.

1. Introduction

Sign language is fully developed language with its own grammar and lexicon, used as primary communication mean by hearing impaired (Sandler, 2006). Sign languages are not universal nor mutually intelligible, but all are based on manual articulation combined with non-manual markers, such as facial expressions. Sign language alphabet (or fingerspelling) is a subset of sign language used to express words that do not have a specific sign, such as names.

Several methods were proposed to automatically detect and classify particular signs, some 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). With the development of machine learning techniques, the recent focus is on the classification using monocular images as the solely source of input. Czech sign language single hand alphabet was successfully classified by convolution neural network (CNN), see (Krejsa, 2020). Another successful approach by (Šnajder, 2022) combines feature detection using MediaPipe framework with common fully connected multilayer classification neural network.

Czech sign language alphabet has to deal with interesting peculiarity as it contains diacritics (accents) in certain letters. Diacritics is expressed by motion of the hand gesturing the letter and therefore can not be classified from a single image, but from the sequence of images. Czech language uses three types of diacritics. The first one is acute (letter A with acute is written as Á), used for prolonging the length of vowels. Second is caron (letter C with acute is written as Č), this diacritics often completely changes the meaning of the word, eg. "prát" (to wash or to fight) and "přát" (to wish). The last one is diacritic ring that is used only with letter "u" and prolongs it, while it can only be used in middle or end of the word.

When detecting the diacritics from a sequence of images, there are basically two approaches, the first one is to use machine learning tailored to processing the sequences. Such approach was used by (Šnajder, 2023) where Long Short-Term Memory (LSTM) architecture was used successfully applied on the task reaching 87 % accuracy of detection.

[*] Assoc. Prof. Ing. Jiří Krejsa PhD.: Institute of Thermomechanics of the Czech Academy of Sciences, Brno department, Czech Republic, krejsa@fme.vutbr.cz

[*] Assoc. Prof. Ing. Stanislav Věchet, PhD.: Institute of Thermomechanics of the Czech Academy of Sciences, Brno department, Czech Republic, vechet.s@fme.vutbr.cz

[**] Jan Šnajder: Faculty of Mechanical Engineering, Brno University of Technology, Czech Republic, snajder@fme.vutbr.cz

doi: 10.21495/em2024-170

The second approach, further investigated in this paper, is to use CNN for classification of letters and image processing techniques for the analysis of motion and subsequently for accent recognition. In particular, the optical flow is used, as it is proved to be usable for gesture recognition, see (Nagy, 2020) for details.

2. Materials and methods

2.1. Still image classification

The classification of particular letter from a single image (not counting the diacritics) is performed by the convolution neural network. Training data were gathered from both sign language interpreters and hearing impaired. Data were augmented (translation, rotation, both uniform and non-uniform scaling). Network topology consisted of 8 convolution layers interlaced with pooling layers, followed by a single fully connected classification layer. Latest accuracy tests exhibit 93 % successful classification on test set. For the details please refer to (Krejsa, 2020). Mentioned CNN was used for the classification during the experiments.

2.2. Optical flow

Optical flow is a method of motion estimation in a sequence of subsequent images. If we take two subsequent images (called frames) into account, we can observe some regions of the image to change position. Depending on the way the sequence was recorder, it might be the change of the recording device position, or actual motion within the scene. We assume the steady position of the observer (the camera) and static (or slowly changed) lighting conditions, therefore the difference between consecutive frames corresponds to the motion of objects observed. Optical flow method outputs the vector field on the domain of the image (the direction and magnitude of motion for each part of the image). All further described experiments were made using Lucas-Kanade method (Lucas 1981), implemented in OpenCV library.

2.3. Data acquisition

Several sentences were prepared and sign language interpreter gestured them in a single take for each sentence. Several takes of the same sentence were recorded, varying in gesturing speed and amount of motion used for diacritics. The recording was taken by two simultaneously running cameras positioned about 50 cm from each other horizontally to capture the images from different angle. An example of the image sequence for a single letter with caron is shown in Fig. 1 (left).

2.4. Data processing

Whole sentence (approximately 1 000 images) was sequentially fed into CNN that outputs the probabilities of each letter in the alphabet. Two consecutive images were fed into the optical flow implementation, outputting the vector field for given image pair.

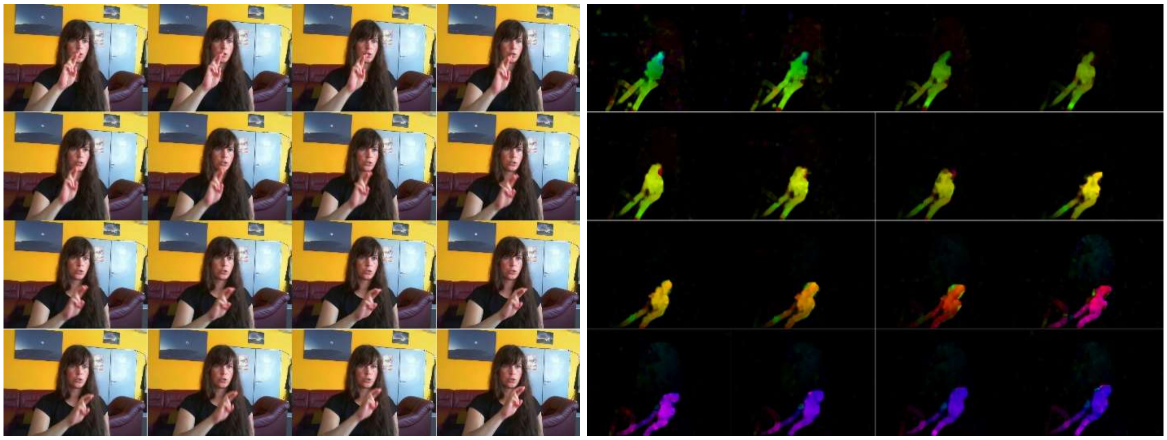

Fig. 1: Composite image of subsequent images of letter Ř (left). The composite is organized in rows, tope left is where the sequence starts, bottom right is where it ends. Corresponding heat map of motion vectors (right).

During the experiments the optical flow was calculated for each pair of images, regardless the CNN output. The vector field is visualized in Fig. 1 (right), where the intensity corresponds to the length of the vector (displacement in the image domain) and the color corresponds to the direction of the vector.

In order to detect the accents, the mean of optical flow vectors for each pair of images was calculated, giving the overall motion between the images. The means were summed up along the whole time sequence, resulting in motion trajectory.

The identification of diacritics works as follows: Each trajectory results in 4 parameters: W = width of bounding box of trajectory, H = height of the bounding box, DX = horizontal distance between the first and the last point in trajectory and DY = vertical distance of the same. For particular diacritics then

$$\text{If } (W>W_{high}) \text{ AND } (H>H_{high} \text{ AND } (DY < H/2) \text{ AND } (DX > W/2) \text{ then CARON} \quad (1)$$

$$\text{If } (W<W_{low}) \text{ AND } (H>H_{high}) \text{ AND } (DY > H/2) \text{ then ACUTE} \quad (2)$$

$$\text{If } (W>W_{high}) \text{ AND } (H>H_{high}) \text{ AND } (DY < H/2) \text{ AND } (DX < W/2) \text{ then RING} \quad (3)$$

The method was tested on 7 sentences with the total of 49 diacritics letters. All of them were tested independently on both sequences from two cameras horizontally distant. For each sentence there were several takes, resulting in 362 diacritics to be identified in total.

3. Results

An example of the motion detected over the whole sentence is shown in Fig. 2. The sentence in particular is “Žlutý papoušek přeletěl keř ve Žďáru nad Sázavou.”, it contains 41 letters, 7 with caron and three with acute. On the figure one can see the motion in both vertical and horizontal axis. Actual letters and diacritics positions are denoted by red and magenta lines (its values on y-axis have no other meaning). Particular trajectories for both letters with/without diacritics are shown in Fig. 3, where blue trajectories correspond to letters without diacritics and black and red are two different instances of the ones with.

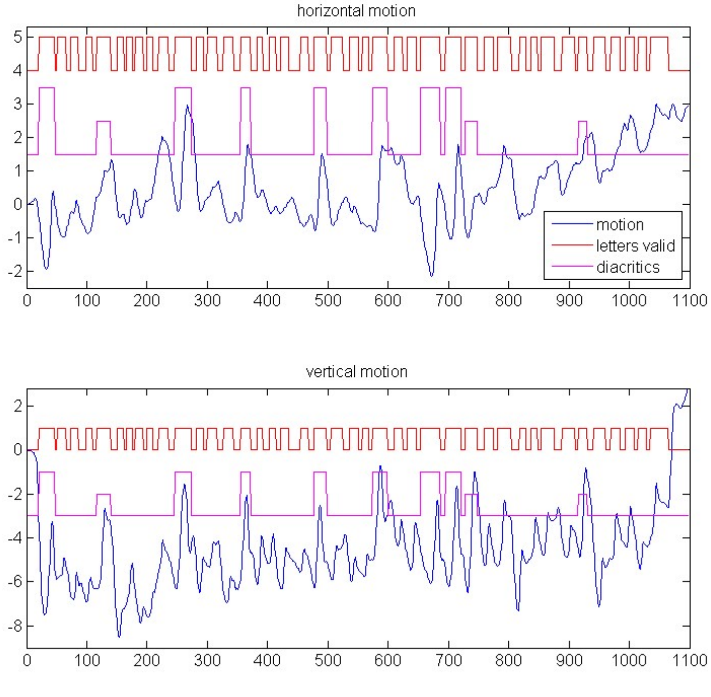

Fig. 2: Detected motion during the sentence. X-axis denotes frame number, Y-axis the amount of motion. Valid letters are shown in red, those with diacritics in magenta.

Out of 362 letters with diacritics the twelve were misclassified by CNN, and in 37 cases the diacritics itself was misjudged by the system, giving over 86 % correct classification rate in total. No significant difference between the values taken by left and right camera were found in diacritics part of the task, however all 12 misclassifications by CNN were made on the data from one of the cameras.

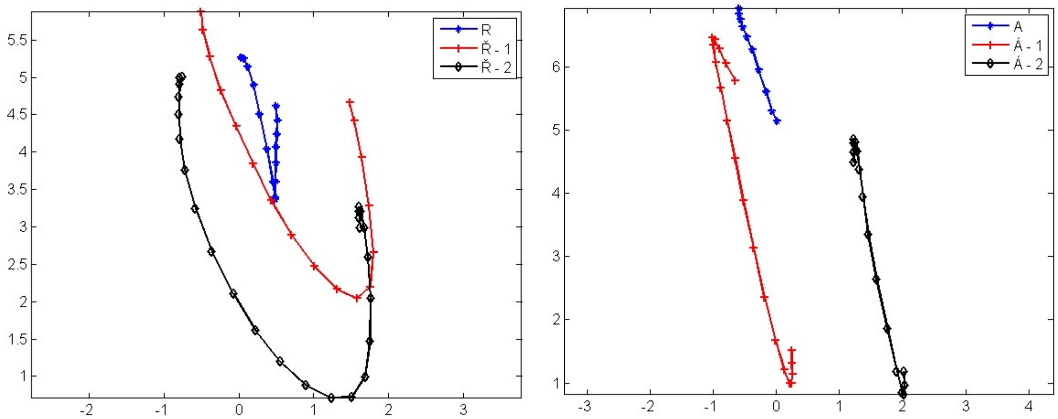

Fig. 3: Detected trajectories for R vs Ř (left) and A vs Á (right).

4. Discussion and conclusions

The main advantage of described approach is that it avoids the retraining of the neural networks used for sequences analysis. Using only the motion detected within the image sequence makes the diacritics detection and classification independent on the gesturer and can easily be extended for any other diacritics in other languages simply by adding different criteria to the trajectory evaluation part. Furthermore, for known letters not containing the diacritics the calculation can be paused, saving the computational means.

The main drawback is the requirement of static observer position and low background motion. The method in its simplest form will inevitably fail when the background has a lot of motion. This could be dealt with using independent background motion analysis and using hand detection techniques for limiting the area the vector field results are used for the diacritics evaluation, however, the authors feel that such extension would increase the number of parameters necessary to fine-tune the method thus decreasing its robustness and contradicting its main advantages.

Future work will be focused on testing the method on larger test data set and further fusion of CNN output sequence with optical flow results.

Acknowledgement

The results were obtained with institutional support RVO 61388998 of the Institute of Thermomechanics AS CR v.v.i. Authors would like to express their gratitude to sign language interpreters.

References

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

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. and Yin, Z. (2015) American sign language alphabet recognition using microsoft Kinect. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 44–52.

Ma, L. and 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*, vol. 2, pp. 136–139.

Krejsa, J. and Vechet, S. (2020) Czech Sign Language Single Hand Alphabet Letters Classification. In: *19th International Conference on Mechatronics - Mechatronika (ME)*, Prague.

Šnajder, J. and Bednařík, J. (2022) Czech Single Hand Alphabet Classification with MediaPipe, In: *Proc. of the 28th Conference Engineering Mechanics*, Milovy, pp. 381–384.

Šnajder, J. and Krejsa, J. (2023) Classification of Czech Sign Language Alphabet Diacritics via LSTM. In: *20th International Conference on Mechatronics - Mechatronika (ME)*, Pilsen.

Nagy, D. Z. and Piller, I. (2020) An Optical Flow-based Gesture Recognition Method, *Acta Marisiensis*, vol. 17, no. 1, pp 22–26.

Lucas, B and, Kanade, T. (1981) An iterative image registration technique with an application to stereo vision. *Proc. of Imaging Understanding Workshop*, pp. 121–130.