ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

CYBERNETICS

DOI 10.51885/1561-4212_2021_1_101
МРНТИ 28.17.33

S. Kondratiuk
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
***E-mail**: sergey.kondrat1990@gmail.com*

ҚИМЫЛ ТІЛІН ТАНУҒА АРНАЛҒАН ҮШ ӨЛШЕМДІ ТҮЙІНДЕР

ТРЕХМЕРНЫЕ СВЕРТКИ ДЛЯ РАСПОЗНАВАНИЯ ЯЗЫКА ЗНАКОВ

THREE-DIMENSIONAL CONVOLUTIONS FOR SIGN LANGUAGE RECOGNITION

Abstract. *The technology is proposed for recognition of gesture units (fingerspelling alphabet) of sign language. Implemented technology performs recognition of dactyl items from camera input using trained on collected training dataset set convolutional neural network, based on the MobileNetv2 architecture. Multiple configurations of layers and hyper-parameters were used for experiments, and based on their results, optimal configuration in terms of complexity and quality was selected. Also, dataset is collected using various conditions of environment and parameters of hand. On the collected test dataset accuracy of over 98 % is achieved.*
Keywords: *sing language, recognition, convolutional neural network, mobilenetv2.*

Abstract. *The technology is proposed for recognition of gesture units (fingerspelling alphabet) of sign language. Implemented technology performs recognition of dactyl items from camera input using trained on collected training dataset set convolutional neural network, based on the MobileNetv2 architecture. Multiple configurations of layers and hyper-parameters were used for experiments, and based on their results, optimal configuration in terms of complexity and quality was selected. Also, dataset is collected using various conditions of environment and parameters of hand. On the collected test dataset accuracy of over 98 % is achieved.*
Keywords: *sing language, recognition, convolutional neural network, mobilenetv2.*

Аңдатпа. *Ымдау тілінің белгілік бірліктерін (саусақ іздері алфавиті) тану технологиясы ұсынылды. Іске асырылған технология MobileNetv2 архитектурасы негізінде жинақталған оқу деректер жиынтығында оқытылған конволюциялық нейрондық желіні қолданып, камерадан кіріс деректеріндегі саусақ іздерін тануды жүзеге асырады. Тәжірибелер үшін бірнеше қабатты конфигурациялар мен гиперпараметрлер қолданылды, олардың нәтижелері күрделілігі мен сапасы жағынан оңтайлы конфигурацияны таңдауға пайдаланылды. Деректер жиынтығы әртүрлі қоршаған орта жағдайлары мен қол параметрлерін қолдана отырып жинақталады. Жиналған тестілік деректер жиынтығы 98 %-дан жоғары дәлдікке жетеді.*
Түйін сөздер: *ымдау тілі, тану, конволюциялық жүйке жүйесі, mobilenetv2.*

Аннотация. *Предложена технология для распознавания жестовых единиц (дактилоскопического алфавита) жестового языка. Реализованная технология выполняет распознавание дактильных элементов на входных данных из камеры с использованием обученной на собранном наборе обучающих данных сверточной нейронной сети, основанной на архитектуре MobileNetv2. Для экспериментов использовалось несколько конфигураций слоев и гиперпараметров, по результатам которых была выбрана оптимальная по сложности и качеству конфигурация. Также набор данных собран с использованием различных условий окружающей среды и параметров руки. На собранном тестовом наборе данных достигается точность более 98 %.*
Ключевые слова: *язык жестов, распознавание, сверточная нейронная сеть, mobilenetv2.*

Introduction. Sign language is one of the most common ways of communication for people with inclusive needs. People with hearing disabilities could use additional software that would

make it easier for them to communicate with society and within their own community. This information technology should consist of a gesture recognition module, which would form the technology of learning the dactyl alphabet. In recent years, smartphones have become one of the most common devices with an operating system, along with personal computers and laptops. This makes cross-platform an important aspect of the technology, as it allows to develop and run the technology without changing the code, thus providing the user with a unified experience on a wide range of platforms, both mobile and low-resource and powerful and stationary. The use of simulation and gesture recognition is more widespread in many areas related to communication, human-computer interfaces, and so on.

Distributed computing technologies [1] and cross-platform development [2] give an approach to beat the issue of platform diversity. Cross-platform development can be utilized instead of virtual-machines [3] or a lot of mono-platforms development.

The paper is devoted to recognition of sign language gesture using machine learning algorithms and neural networks and development of appropriate modules cross-platform technology to run on a variety of modern platforms. The sing (gesture) recognition is a part of a single gesture communication technology and this paper is a further development of author's previous works [4], [5].

Existing approaches for recognition of sign language. Detection of hand gestures can be considered as a type of task of object detection, which has a set of mature and novel approaches in both classic computer vision and deep learning, with convolutions neural networks specifically.

As bigger datasets with recorded activities were released (AlexNet [6], Sports-1M [7], Kinetics [8], Jester [9]), convolutional neural networks with 3-dimensional convolutions because successful. The size of the dataset allowed to train the model without overfitting[10].

Gestures of sign language were detected using different approaches based on classic computer vision with hand-crafted features such as orientation of histograms [11], histogram of oriented gradients (HOG) [12] or bag-of- features [13]. Although the state of the art hand gesture recognition architectures are based on CNNs [14, 15, 16], similar to other computer vision tasks.

Research of existing approaches among CNN [17, 18] was accented on lightweight architectures which show satisfying performance on mobile cpus, such as SqueezeNet[19], MobileNet [20], MobileNetV2 [21], ShuffleNet [22] and ShuffleNetV2 [23], MobileNetV3 [24] which aim to reduce computational cost but still keep the accuracy high. In our work, we have used the 2D and 3D versions of MobileNetV2.

Problem statement. The proposed technology should consist of sign language gesture recognition module. Module should be able to run without codebase modification on multiple platforms and should be developed using cross-platform tools.

Gesture recognition module should consist of a model which is able to detect and identify the gesture, specified by the user, from a camera input. Set of gestures is limited by the Ukrainian dactyl language, but can be extended further. An appropriate dataset of Ukrainian dactyl language should be collected for testing the model performance.

The gesture recognition module should utilize the model which show robust and state-of-the-art performance along with high efficiency in terms of computational resources in order to achieve high accuracy and FPS-rate on various platforms, using cross-platform technologies.

Proposed approach. To developed a technology for Ukrainian dactyl language recognition, which can run on multiple platforms, without changing the codebase, an approach based on cross-platform tools is proposed. To develop a gesture recognition module, a cross-platform framework Tensorflow [27] is proposed. This approach based on cross-platform framework for machine learning allows to developed and train a gesture recognition model once, and then deploy it on multiple platforms (mobile, desktop and web) without any modifications to the model

or the code for training. Altogether, the proposed technology novelty is that it's a unified cross-platform technology for Ukrainian dactyl language recognition, with improved MobileNet architecture for improved recognition of the Ukrainian dactyl alphabet.

Gesture recognition. Gesture recognition, as a part of cross-platform technology for Ukrainian dactyl language recognition, should be implemented using cross-platform tools. Some approaches, for example, Ong et al. [28] proposes Sequential Pattern Mining in order to detect signs based on the tree structures.

Convolutional Neural Networks (CNN) is a class of deep neural networks which are regularized versions of multilayer perceptrons, most commonly applied to analyzing images and videos. CNNs are especially good at analyzing images due to ability to take into account locality reference of the data in the image (typically nearby samples at some input data are not related, which is not true in case of an image). Therefore, CNN show state-of-the-art results in image classification and recognition tasks [29], [30].

The task of dynamic gesture recognition involves modeling the temporal aspect of gestures in addition to identifying features. The paper considers two strategies of dynamic temporal gesture modeling:

• creation of descriptors that carry spatial and temporal information;

• recognition of sequences of spatial descriptors or images using space-time classifiers.

The second approach was used in the work. As the input data of the neural network, one can submit both single gesture images and a sequence of images, if there is a need to analyze the video stream. Thus, several improvements have been made compared to existing gesture recognition approaches:

• analysis of several neighboring images simultaneously allows to train the network, taking into account the temporal aspect, i.e. the dynamics of change of gestures in several images, for greater resistance to change in such a dynamic object as the hand;

• bad gesture recognition can be smooth out on a single image in the video stream sequence. One of the images may be with artifacts, excessive or insufficient lighting, or blurred, or the subject will be obstructed – all of these problems can be smoothed out by gesture recognition using adjacent frames in sequence.

To increase the efficiency of this approach, the use of a temporary floating window was proposed and implemented. To do this, it is proposed to divide the input sequence into n subsequences with a minimum length m, which intersect (into a certain part, from 10 % to 50 % of the subsequence length).

When obtaining video sequences and individual frames, the same fingerprints may differ both in the external features of the hand (the task of recognizing such differences rests on the recognition model) and in the parameters of the data (size, quality, focal length, lighting, background, artifacts, blur and etc.). For further calculations within the selected recognition model, a unified data processing procedure was developed to reduce them to a general form, both at the stage of training the model and at the stage of recognition.

The paper adapts the neural network to the space-time format of the input data. To make better use of spatiotemporal features from the input data, it was proposed to improve the architecture of the convolutional neural network with three-dimensional convolutions. Thus, a three-dimensional convolutional neural network is able to perform a convolution operation not only in the image space but also in time.

Data processing consists of three steps:

• normalization;

• noise reduction;
• reduction to a single size.

MobileNetV2 architecture (Figure 1) is a new mobile architecture, development of the MobileNet model. MobileNetV2 extends its predecessor with 2 main ideas. Residual blocks connect the beginning and end of a convolutional block with a skip connection. By adding these two states the network has the opportunity of accessing earlier activations that weren't modified in the convolutional block.

Input	Operator	exp size	$\#out$	SE	NL	s
$224^2 \times 3$	conv2d, 3x3	-	16	-	HS	2
$112^2 \times 16$	bneck, 3x3	16	16	✓	RE	2
$56^2 \times 16$	bneck, 3x3	72	24	-	RE	2
$28^2 \times 24$	bneck, 3x3	88	24	-	RE	1
$28^2 \times 24$	bneck, 5x5	96	40	✓	HS	2
$14^2 \times 40$	bneck, 5x5	240	40	✓	HS	1
$14^2 \times 40$	bneck, 5x5	240	40	✓	HS	1
$14^2 \times 40$	bneck, 5x5	120	48	✓	HS	1
$14^2 \times 48$	bneck, 5x5	144	48	✓	HS	1
$14^2 \times 48$	bneck, 5x5	288	96	✓	HS	2
$7^2 \times 96$	bneck, 5x5	576	96	✓	HS	1
$7^2 \times 96$	bneck, 5x5	576	96	✓	HS	1
$7^2 \times 96$	conv2d, 1x1	-	576	✓	HS	1
$7^2 \times 576$	pool, 7x7	-	-	-	-	1
$1^2 \times 576$	conv2d 1x1, NBN	-	1024	-	HS	1
$1^2 \times 1024$	conv2d 1x1, NBN	-	k	-	-	1

Figure 1. Architecture of MobileNetv2

Network Improvements have been made in two ways:

• Layer removal
• swish non-linearity

The paper proposes, as a further development of recognition technologies, a mechanism for smoothing anomalous recognition results due to the implemented method of accumulation of probabilities from previous subsequences.

Because subsequences are formed on the principle of a floating window with intersections, the recognition results should gradually reduce the maximum probability of the current gesture, and over time (i.e., subsequences) should increase the maximum probability of the next gesture.

The model proposed and implemented in the work is to accumulate predictions from previous subsequences and update the current recognition result only when the accumulated sum of probabilities exceeds a certain threshold, which can be presented as follows:

$$\overline{\sum_{t=t-n}^{t+n} \sum_{i=t-k}^{t+k} p_i > thresh} \quad (1)$$

where:

• p_i - the probability of a gesture in the frame,
• i - frame number on the current subsequence,
• t - number of the current subsequence,
• k - the size of the subsequence in both directions,
• n - number of accumulated subsequences.

Done by projection layer from the last layer from the previous block.

Thus we can remove that projection layer and the filtering layer from the previous bottleneck layer(block).

Ukrainian dactyl alphabet dataset collections for recognition with MobileNet

An analysis of the collected educational data set, collected for the first time in such quantity and diversity by people and environment for the Ukrainian dactyl alphabet (Fig. 5). Different lighting conditions were used (with distribution: 20% of images in low light conditions, 30% in low light conditions and 50% in high quality lighting). About 10% of the images were distorted by noise and blur. In total, ~ 50,000 original images were collected as a training data set. After applying additional data magnification techniques (such as rotation, random cropping, mirroring, etc.), the final data set was about 150,000 images. 10% of the data set was selected for testing, resulting in a final set of 135,000 images and a final test set of 15,000 images.

Figure 2. Dataset example

Data augmentation is performed in order to increase the size of the data set without manually creating new images. Augmentation is performed using a number of techniques that increase the number of images in the data set several times, diversify them and, importantly, reduce the neural network's ability to overfit features that are present in the original collected data set, thus making the model more robust to changes in input data. Approaches to image modification can be combined to further distort the original data set. Thus, it is possible to change the environments in which the trained model is tested. Among the operations carried out as part of the data increase:

- Gaussian noise;
- affine transformation;
- trimming + shift;
- reflection;
- distortion of perspective;
- blurring

Gesture recognition experiment.

Experimental tests of algorithms covered each of the parts of the software implementation of dactyl recognition of the Ukrainian dactyl alphabet.

During the Convolutional Neural Network training process, based on the MobileNetv2 archi-

tecture, which demonstrates high quality and performance on mobile and devices with limited computing power, several modifications of the architecture were created to find the best trade-off between layer number and accuracy.

The chosen architecture at the training stage can be configured with hyperparameters, which are selected for each training individually (learning rate, batch size, number of epochs), and the architecture itself, i.e. the number and configuration of repeating the same type of layers.

This set of configurations and possible values of hyperparameters forms a grid within which a set of neural networks is trained and compared on a single test set.

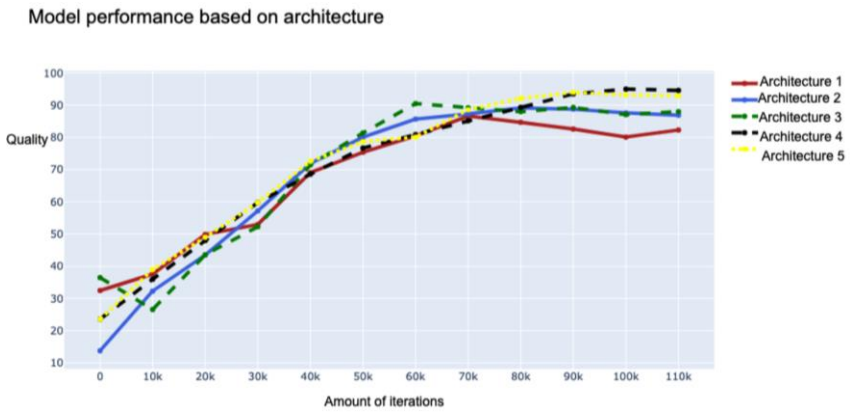

Figure 3. Model quality related to architecture and number of iterations

Architecture 1	Architecture 2	Architecture 3	Architecture 4	Architecture 5
Conv / s2	Conv / s2	Conv / s2	Conv / s2	Conv / s2
Conv dw / s1	Conv dw / s1	Conv dw / s1	Conv dw / s1	Conv dw / s1
Conv / s1	Conv / s1	Conv / s1	Conv / s1	Conv / s1
Conv dw / s2	Conv dw / s2	Conv dw / s2	Conv dw / s2	Conv dw / s2
Conv / s1	Conv / s1	Conv / s1	Conv / s1	Conv / s1
Conv dw / s1	Conv dw / s1	Conv dw / s1	Conv dw / s1	Conv dw / s1
Conv / s1	Conv / s1	Conv / s1	Conv / s1	Conv / s1
Conv dw / s2	Conv dw / s2	Conv dw / s2	Conv dw / s2	Conv dw / s2
Conv / s1	Conv / s1	Conv / s1	Conv / s1	Conv / s1
Conv dw / s1	Conv dw / s1	Conv dw / s1	Conv dw / s1	Conv dw / s1
Conv / s1	Conv / s1	Conv / s1	Conv / s1	Conv / s1
Conv dw / s1	2 x Conv dw / s1	3 x Conv dw / s1	Conv dw / s2	Conv dw / s2
Conv / s1	2 x Conv / s1	3 x Conv / s1	Conv / s1	Conv / s1
Conv dw / s2	Conv dw / s2	Conv dw / s2	4 x Conv dw / s1	5 x Conv dw / s1
Conv / s1	Conv / s1	Conv / s1	4 x Conv / s1	5 x Conv / s1
Avg Pool / s1	Avg Pool / s1	Avg Pool / s1	Conv dw / s2	Conv dw / s2
FC / s1	FC / s1	FC / s1	Conv / s1	Conv / s1
Softmax / s1	Softmax / s1	Softmax / s1	Avg Pool / s1	Conv dw / s2
			FC / s1	Conv / s1
			Softmax / s1	Avg Pool / s1
				FC / s1
				Softmax / s1

Table 1. Architectures considered

The developed technology generated 5 configurations of neural network architecture with different number of layers and number of parameters, which allowed to find a balanced neural network architecture with limited size and high performance on the test data set.

Over time, the accuracy of the trained model stopped growing, as shown in Figure 3, so the architecture №4 (Figure 4) was considered optimal in terms of the smallest architecture with the best accuracy (average macrobal f1).

Standard techniques of fighting overfitting of the neural network were applied on each training.

Conclusions. The proposed technology consists of a main gesture recognition module, which use the database with gestures specifications stored in YAML format in a PostgreSQL [31] database.

Proven data processing algorithms have shown that gesture recognition in a sequence of images (video) requires a special model architecture and data preparation procedures to improve results.

The results of experiments with different model architectures and different data sets demonstrated the improvement of recognition quality using the advanced MobileNetv2 architecture with three-dimensional convolution. The architecture selection process demonstrated the optimal relationship between the complexity of the model and its efficiency in recognition on a given data set. The quality of the model was achieved at 0.97 macroballs f1 on a given test data set.

As part of the proposed implementation, a data set of 50,000 images with all the dactyls of the Ukrainian dactyl alphabet, demonstrated by 50 different people, was collected for the first time, and up to 150,000 images were augmented. The proposed gesture communication technology can be further augmented with other gestures and languages and with other cross-platform modules.

References

1. Peter Mell and Timothy Grance (September 2011). The NIST Definition of Cloud Computing (Technical report). National Institute of Standards and Technology: U.S. Department of Commerce. doi:10.6028/NIST.SP.800-145. Special publication 800-145.
2. The Linux Information Project, Cross-platform Definition.
3. Smith, James; Nair, Ravi (2005). «The Architecture of Virtual Machines». Computer. IEEE Computer Society. 38 (5): 32-38.
4. I. Krak, S. Kondratiuk (2017). Cross-platform software for the development of sign communication system: Dactyl language modelling, Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 1, pp. 167-170. DOI: 10.1109/STC-CSIT.2017.8098760.
5. Yu.V.Krak, Yu.V. Barchukova, B.A. Trotsenko (2011). Human hand motion parametrization for dactylemes modeling, Journal of Automation and Information Sciences, 43 (12). – Pp. 1-11.
6. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097-1105, 2012.
7. A. Karpathy,G. Toderici, S. Shetty,T. Leung,R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convo- lutional neural networks. In Proceedings of the IEEE con- ference on Computer Vision and Pattern Recognition, pages 1725-1732, 2014.
8. J. Carreira and A. Zisserman. Quovadis, actionrecognition? a new model and the kinetics dataset. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 4724-4733. IEEE, 2017.
9. T. B. N. GmbH. The 20bn-jester dataset v1. https://20bn.com/datasets/jester, 2019.
10. K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pages 18-22, 2018.
11. W.T. Freeman and M. Roth. Orientation histogramsforh and gesture recognition. In International workshop on automatic face and gesture recognition, volume 12, pages 296-301, 1995.
12. L. Prasuhn, Y. Oyamada, Y. Mochizuki, and H. Ishikawa. A hog-based hand gesture recognition system on a mobile device. In 2014 IEEE International Conference on Image Processing (ICIP), pages 3973-3977. IEEE, 2014.
13. N.H. Dardas and N.D. Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instru- mentation and measurement, 60(11):3592-3607, 2011.
14. O. Ko ̈pu ̄klu ̈, N. Ko ̈se, and G. Rigoll. Motion fused frames: Data level fusion strategy for hand gesture recognition. arXiv preprint arXiv:1804.07187, 2018.

15. P. Molchanov, S. Gupta, K. Kim, and K. Pulli. Multi-sensor system for driver's hand-gesture recognition. In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE Inter- national Conference and Workshops on, volume 1, pages 1-8. IEEE, 2015.
16. P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Handgesture recognition with 3d convolutional neural networks. In The IEEE Conference on Computer Vision and Pattern Recogni- tion (CVPR) Workshops, June 2015.
17. J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation net- works. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132-7141, 2018.
18. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learn- ing for image recognition. In Proceedings of the IEEE con- ference on computer vision and pattern recognition, pages 770-778, 2016.
19. F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360, 2016.
20. A.G. Howard, M.Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Effi- cient convolutional neural networks for mobile vision appli- cations. arXiv preprint arXiv: 1704.04861, 2017.
21. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottle- necks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4510-4520. IEEE, 2018.
22. X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6848-6856. IEEE, 2018.
23. N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. arXiv preprint arXiv:1807.11164, 5, 2018.
24. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang: Searching for MobileNetV3. axXiv: 1905.02244, 5, 2019.
25. ASL Sing language dictionary [http://www.signasl.org/sign/model]
26. Unity3D framework [https://unity3d.com/]
27. Tensorflow framework documentation [https://www.tensorflow.org/api/]
28. Eng-Jon Ong et al. Sign language recognition using sequential pattern trees. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE. 2012. – Pp. 2200-2207.
29. American Sign language: Real-time American Sign Language Recognition with Convolutional Neural Networks Brandon Garcia Stanford University Stanford, CA, 2015.
30. Hand gesture recognition using neural network based techniques, Vladislav Bobic, School of Electrical Engineering, University of Belgrade, 2016.
31. PostgreSQL official web site [https://www.postgresql.org/]