IRSTI 28.23.37

M. Kalken[1]
[1]Suleyman Demirel University, Kaskelen, Kazakhstan

HANDWRITTEN OPTICAL CHARACTER RECOGNITION: IMPLEMENTATION FOR KAZAKH LANGUAGE

Abstract. Many documents, including as invoices, taxes, memoranda, and surveys, historical data, and test replies, still require handwriting with the transformation to digital information interchange. Handwritten text recognition (HTR), which is an automatic approach to decode records using a computer, is required in this aspect. For this proposal, I present a study of the implementation of optical recognition algorithms for handwritten text in the Kazakh language, using a recently collected database. The database, called the Kazakh Autonomous Handwritten Text Dataset (KOHTD), contains more than 140,335 segmented images of handwritten exam papers. As an algorithm, I used the proposed model by Harald Scheidl, which consists of several layers of neural networks and an CTC decoder. The trained model by putting an interval of lr = 0.01 and a batch size of 60 showed effective results with indicators of about 85% accuracy.

Keywords: OCR, handwritten text recognition, KOHTD, neural networks, CNN.

***

Аннотация. Многие документы, в том числе счета-фактуры, налоги, меморандумы и опросы, исторические данные и ответы на тесты, по-прежнему требуют рукописного ввода с преобразованием в цифровой обмен информацией. В этом аспекте требуется распознавание рукописного текста, которое представляет собой автоматический подход к декодированию записей с помощью компьютера. Для этого предложения я представляю исследование реализации алгоритмов оптического распознавания рукописного текста на казахском языке с использованием недавно собранной базы данных. База данных, называемая Казахским автономным набором данных рукописного текста (KOHIT), содержит более 140 335 сегментированных изображений рукописных экзаменационных работ. В качестве алгоритма я использовал предложенную Харальдом Шейдлом модель, которая состоит из нескольких слоев нейронных сетей и декодера CTC. Обученная модель с интервалом lr =0,01 и размером партии 60 показала эффективные результаты с показателями точности около 85%.

Ключевые слова: оптическое распознавание, распознавание рукописного текста, KOHTD, нейронные сети, CNN.

***

Аңдатпа. Көптеген құжаттар, соның ішінде шот-фактуралар, салықтар, меморандумдар мен сауалнамалар, тарихи деректер және тест жауаптары әлі де қолжазбаны сандық ақпаратқа түрлендіруді қажет етеді. Бұл жағдай қолмен жазылған мәтінді компьютердің көмегімен танып, автоматты түрде декодтау арқылы сандық ақпаратқа көшіруді қажет етеді. Осы мәселе үшін мен жақында жиналған деректер базасын пайдалана отырып, қазақ тілінде қолжазба мәтінді оптикалық тану алгоритмдерін іске асыруды зерттеуді ұсынамын. Қолжазба мәтінінің қазақ автономиялық деректер жиынтығы деп аталатын деректер қорында қолжазба емтихан жұмыстарының 140 335-тен астам сегменттелген бейнесі бар. Алгоритм ретінде Харальд Шейдл ұсынған модельді қолдандым, ол нейрондық желілердің бірнеше қабаттарынан және CTC декодерінен тұрады. LR = 0.01 аралығымен және 60 партия өлшемімен оқытылған модель 85% дәлдік көрсеткіштерімен тиімді нәтижелер көрсетті.

Түйін сөздер: оптикалық тану, қолжазбаны тану, KOHTD, нейрондық желілер, CNN.

1. Introduction

OCR (optical character recognition) is a technology that transforms text into a machinereadable format [1]. OCR is now used to aid not only digitize handwritten medieval manuscripts, but also to transform typewritten texts into digital form. This has simplified the retrieval of essential information since it eliminates the need to sift through stacks of papers and files in order to find what is needed. Digital preservation of historic data, legislation documents, educational persistence [2], and other requirements are being met by organizations. The handwritten character recognition is one of the fields of OCR and it has achieved significant real-world success in specific applications such as email recognition on mail-pieces for sorting automation and reading of courtesy and legal amounts on bank checks. However, due to the growing of smartphones and tablet devices, where handwriting with a finger or stylus is expected to be a potentially efficient form of input, handwritten text recognition remains a huge topic that is gaining fresh interest as an active area of study [13].

It should be clarified that for machine learning algorithms, it is necessary to take into account the importance of a large database with unique datasets for training and testing machine learning neural networks. One of the most effective and frequently used machine learning algorithms at the moment can include support vector machine (SVM), K-nearest neighbor (KNNS), neural networks (NNS) and convolutional neural networks (CNNs) algorithms [3, 4, 10]. Algorithms of convolutional neural networks effectively use compression and processing of information in the form of digital data that were obtained as a result of converting an image to a digital format containing information about each color area in the RGB representation [5]. For the effective operation of this

algorithm, it is necessary to have a uniquely assembled and consisting of high-quality handwriting images, a voluminous database of manuscripts by different authors. In Cyrillic, unlike Latin handwriting characters, there is no strict common standard database used.

In this article, it is proposed to use recently assembled open access database that contains about 140 thousand unique handwritten images of Cyrillic characters and the algorithm with CNN and recurrent neural networks (RNN) layers. The database consists of 99% of the words of the Kazakh language. Before use, changes will be made to the structure of the database and it is also supposed to pre-process the database using 4 different levels of image preprocessing. After processing, the database is divided into two parts: the first contains 95% of the image that will be used for training the neural network, the second part containing 5%, respectively, for testing and validating the already trained neural network model of the computer vision algorithm and machine learning. The ML algorithm model will be trained only to recognize single words without a phrases and sentences. After checking the model, having studied the result of passing the validation stage, which was visible in the form of a loss value, I noticed that the model shows itself with an effective result in the form of 85% accuracy.

II. Methods

1.1 Binarization

Binarization is the conversion of a color image to black and white (black pixel value =0 and white pixel value =255). This can be done by setting a certain average value of the total pixel range, which is 127.5 out of 0-255. It is logical to assume that when the average value is exceeded, the pixel will be considered white, that is, it will have a value of 255, and with a lower value from the average, to the pixel is given a value of 0, that is, it becomes black.

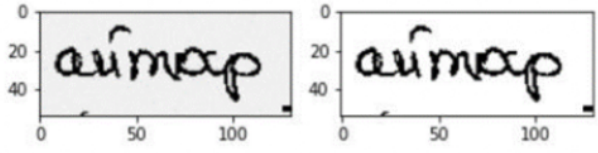

Figure 1. Example of the result after binarization.

1.2. Noise Removal

The main task of noise removal is to smooth the image by removing points whose intensity has a higher value than the rest of the image. Thus, noise removal helps to get rid of unnecessary image distortions when working with computer vision. To do this, the OpenCV library is used with which you can apply this processing to color and binary images.

Figure 2. Example of removing unnecessary noise from an image.

1.3. Thinning and Skeletonization

This processing allows us to reduce or vice versa increase the stroke width if necessary. With printed text, this processing would be unnecessary, since the stroke volume in this state would be the same for the entire text. Since I use a handwritten input database, each author can have a different stroke width. To do this, before using the processing algorithm, it should recognize the text and the maximum stroke width. If it is greater than the expected norm, then the algorithm comes into action, otherwise it skips. As a result, all images come in approximately the same stroke volume, which will be more convenient to work with.

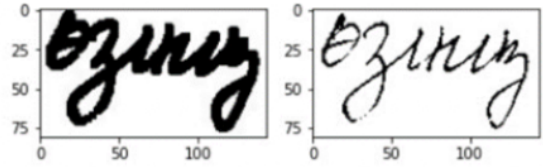

Figure 3. Example of thinning and skeletonization.

After this stage of preparation our database is ready. Also, to identify images, it is needed to prepare a separate file that contains information about each image [12]. That is, when training and recognizing neural networks, it should take information about what this image means. Usually, this type of file is necessarily present in a special database. Since the file in the selected database was collected by manual recognition, it contains information about each image in a separate "json" file. This structure is not suitable for the upcoming algorithm. In this connection, the unification and standardization of the file was carried out in the same way as in the Latin standard of the IAM English database. Which is followed by splitting into columns by separating by commas, which corresponds to the CSV format. The first column contains the name of the file, the second contains its value, except for information on the size and parameters of image processing. As a result, we have all the necessary information contained in one file.

2. Proposed Model

The autonomous handwriting recognition model proposed by Harald Scheidl [9] was used as a recognition and learning model. The model uses 5 CNN layers, 2 recurrent NN (RNN) layers (LSTM) and Connectionist Temporal Classification (CTC) loss. All these layers were implemented using Tensorflow machine learning library.

The figure 4 shows the main 3 stages of learning and recognition. CNN layers accept an incoming image [10]. These layers are designed to extract certain data from an image for subsequent compression. So, three types of operations are implemented in each layer:

1 Convolution operation. Applies a 5×5 filter in the first two layers and 3×3 in the last three layers to the input data.
2 Operation of a nonlinear function. The nonlinear function Rectified Linear Unit (RULE) is used. We can also use Softmax, depending on the case.
3 Operation merge. At the end, the layer combines and summarizes the image areas. At the output we get a reduced version of the data. While the image height is reduced by 2 in each layer, feature maps (channels) are added, so that the output feature map (or sequence) is 32×256 in size.

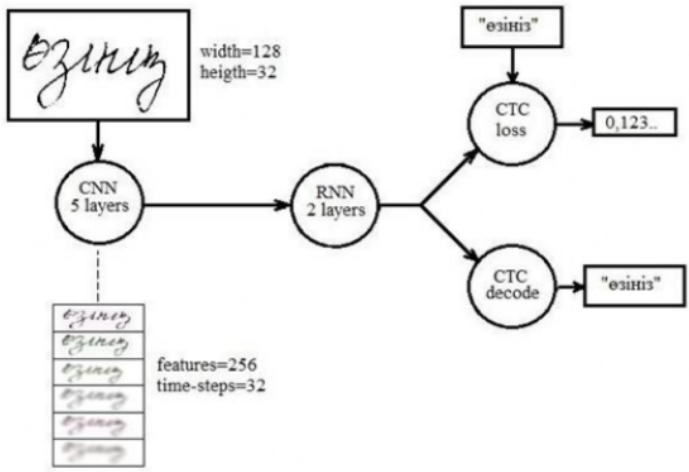

Figure 4. The autonomous handwriting recognition model.

In RNN, two layers of 256 units each are created and smoothed. It was found that the implementation of RNN with long short-term memory (LSTM) spreads information over longer distances and provides a more reliable characteristic of learning [11]. Therefore, instead of a possible vanilla RNN, LSTM is used. The output data from this layer has a representation in the form of a 32x80 matrix. This means that each 32 steps has a possible 80 entries. This is necessary for the next layer of the CTC decoder.

While the neural network is being trained, a matrix from the RNN output is being fed to the

CTC to decrypt the value, as well as to check and calculate losses compared to the true values. The total values should not exceed the size of up to 32 characters. In our case, the "beamsearch" decoder is used [9]. It receives as an

input value all possible letters and symbols, which, according to the level of comparison, determines each character individually. The recognized characters are then combined into one word according to the highest probability value. Figure 5 shows the visualization of the CTC decoder recognition. The word "өзінің" is selected as the input image. From the graph showing the scores for the characters “ө”, “з”, “і”, “н” and the empty CTC label, the text is recognized by selecting the most likely character from each time step.

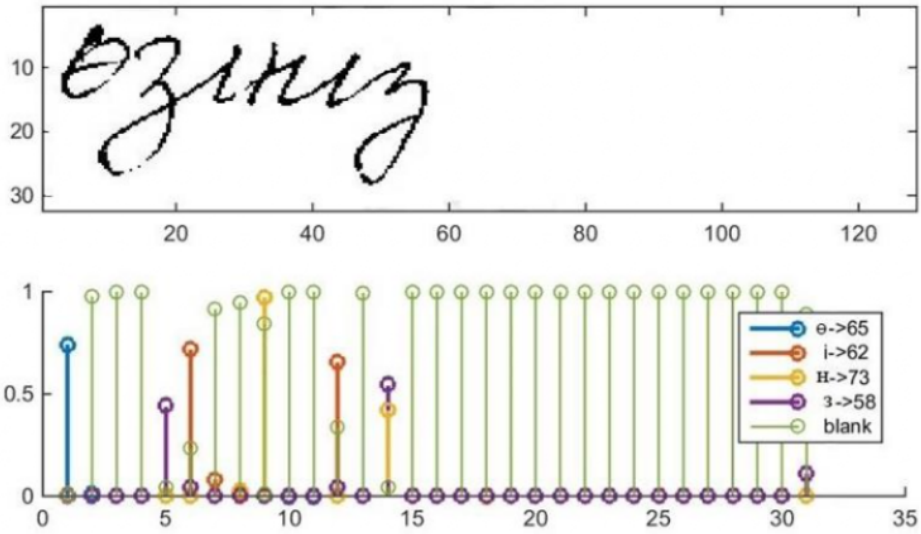

Figure 5. First, input image. Second, probabilities for the characters “ө”, “і”, “н”, “з” and the CTC blank label.

The images from the prepared database do not have exactly the required size, so we need to change its size until it has a width of 128 and a height of 32. To do this, I reduce or increase the image to a height of 32, then fill in the empty spaces in the length with white.

III. Results

The machine in which the tests were conducted has an Intel(R) i5-10300H processor, an NVIDIA GTX 1060Ti graphics card and 16 GB of RAM. The processor was mainly used, but the process could be accelerated 1.2-1.5 times when using a video card. The training parameters were set to values up to 20 epoch and forced stop with the loss value unchanged. Batch has a size of 60 in the image and the values lr=0.01 were set. The training lasted until the value of 30 epoch was reached and was forcibly stopped due to a slight change in the value of loss.

The results of the training can be seen in the figures below (figure 6,7):

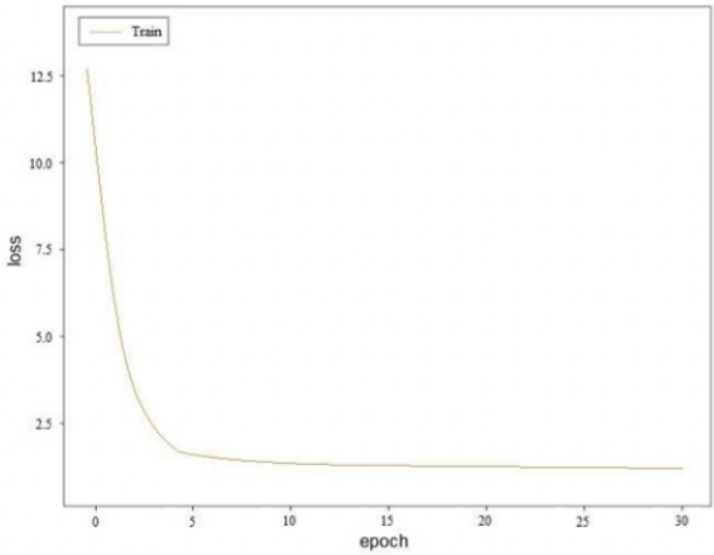

Figure 6. Experiment results: model error.

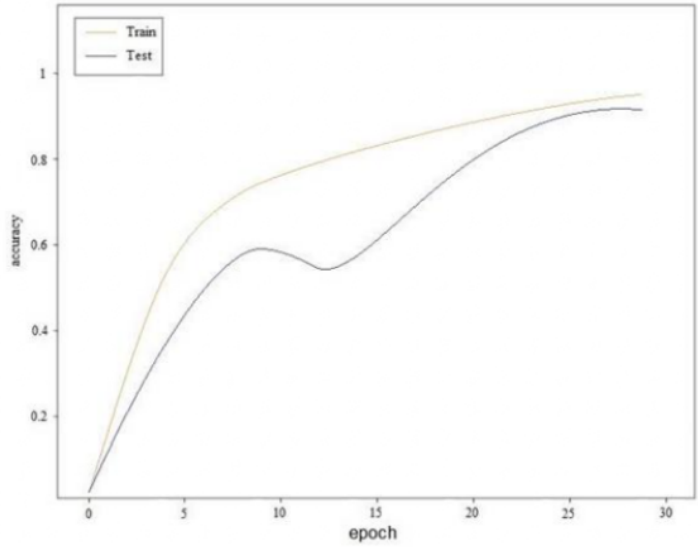

Figure 7. Experiment results: model accuracy.

The figures below show the images that were submitted for input. In them we can see the words "өзіңіз", "айтар". And in the following pictures we can see the recognized word and the probability of correct recognition. For the first image "өзіңіз" with the probability 73.6% and for the second image "айтар" with the probability 85%, respectively.

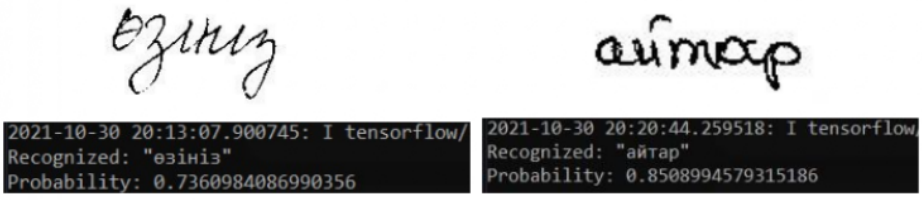

Figure 8. Recognition results.

IV. Conclusion

In this research work, I was among the first to use the Kazakh Autonomous Handwritten Text Dataset (KOHTD), which consists of a large collection of exam papers filled out by students of Satbayev University and Al-Farabi Kazakh National University. Further, an attempt was made to recognize the handwritten Kazakh language and solve problems related to it using the well-known CNN, RNN model. The effectiveness of this model was evaluated with accuracy of 85% and losses of 1.5 -2 loss. The model showed good results and opened the way to possible further research.

In the future, it is possible to conduct research to improve the compiled trained model. This can be done by further correlating the database and increasing the existing database. We can also use other well-known models for training and adding additional libraries to improve the algorithm. After all this, it is possible to develop a model for recognizing sentences.

References

1 Tappert C. C., Suen C. Y., and Wakahara T., The state of the art in online handwriting recognition, IEEE Trans. Pattern Anal. Mach.Intell., 12(1990): pp. 787-808.

2 Zanibbi R., Blostein D., Recognition and retrieval of mathematical expressions, Int. J. Document Anal. Recognit., 15(2012): pp. 331-357.

3 Ahmad I., Fink G.A., Class-based contextual modeling for handwritten Arabic text recognition, 15th International conference on frontiers in handwriting recognition (ICFHR), 2016.

4 Baldominos A., Sáez Y., Isasi P., A survey of handwritten character recognition with mnist and emnist, Applied Sciences, 9(2019).

5 Neha Sharma, Amlan Chakrabarti, Valentina Emilia Balas, Data Management, Analytics and Innovation, Singapore: ICDMAI, 2019. - pp. 299-309

6 Shayahmetov atyndagy til-qazyna ulttyq gylymipraktikalyq ortalygy, Sozdikqor, sozdikqor.com. Last accessed November 25, 2021. URL: https://sozdikqor.kz.

7 Nazgul Toiganbayeva, et al. ,KOHTD: Kazakh Offline Handwritten Text Dataset, Cornell University, 21(2021).
8 Susmith Reddy, Pre-Processing in OCR, Towards Data Science. Last accessed November 26, 2021. URL: https://towardsdatascience.com/pre-processing-in-ocr-fc231c6035a7.
9 Harald Scheidl, Build a Handwritten Text Recognition System using TensorFlow, Towards Data Science. Last accessed November, 26, 2021. URL: https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow2326a3487cd5
10 Sumit Saha, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, Towards Data Science. Last accessed November, 24, 2021. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-theeli5-way-3bd2b1164a53.
11 Harald Sheldi, Handwritten Text Recognition in Historical Documents, Cornell University, 2018.
12 Zamora-Martinez F., Frinken V., Espa˜na-Boquera S., CastroBleda M. J., Fischer A., Bunke H., Neural network language models for off-line handwriting recognition, Pattern Recognition, 2014.
13 Nurseitov D., Bostanbekov K., Kurmankhojayev D., Alimova A., Abdallah A., Tolegenov R., Handwritten kazakh and russian (hkr) database for text recognition, Multimedia Tools and Applications, 80(2021): pp. 375-397.
14 Mahmoud S. A., Ahmad I., Al-Khatib W. G., Alshayeb M., Parvez M. T., An open arabic offline handwritten text database, Pattern Recognition, 47(2014): pp. 1096-1112.
15 Parvez M. T., Mahmoud S. A., Arabic handwriting recognition using structural and syntactic pattern attributes, Pattern Recognition, 46(2013): pp. 141-154.