

Russian Character Recognition using Self-Organizing Map

D Gunawan^{1*}, D Arisandi¹, F M Ginting¹, R F Rahmat¹ and A Amalia²

¹Department of Information Technology, Faculty of Computer Science and Information Technology University of Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

²Department of Computer Science, Faculty of Computer Science and Information Technology University of Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

*Email: danigunawan@usu.ac.id

Abstract. The World Tourism Organization (UNWTO) in 2014 released that there are 28 million visitors who visit Russia. Most of the visitors might have problem in typing Russian word when using digital dictionary. This is caused by the letters, called Cyrillic that used by the Russian and the countries around it, have different shape than Latin letters. The visitors might not familiar with Cyrillic. This research proposes an alternative way to input the Cyrillic words. Instead of typing the Cyrillic words directly, camera can be used to capture image of the words as input. The captured image is cropped, then several pre-processing steps are applied such as noise filtering, binary image processing, segmentation and thinning. Next, the feature extraction process is applied to the image. Cyrillic letters recognition in the image is done by utilizing Self-Organizing Map (SOM) algorithm. SOM successfully recognizes 89.09% Cyrillic letters from the computer-generated images. On the other hand, SOM successfully recognizes 88.89% Cyrillic letters from the image captured by the smartphone's camera. For the word recognition, SOM successfully recognized 292 words and partially recognized 58 words from the image captured by the smartphone's camera. Therefore, the accuracy of the word recognition using SOM is 83.42%

1. Introduction

Typing Cyrillic letters to search for Russian word translation in smartphone's digital dictionary might be as difficult as searching by using a dictionary book. As well as its other neighbor countries, Russia uses letters called Cyrillic instead of Latin. Most of the people who visit Russia are not familiar with the Cyrillic letters. This problem will affect the ability of the visitors to translate the Russian word.

From the statement mentioned earlier, the problem in translating Russian word is how to provide the input word to the digital dictionary. It is difficult to search the Russian word in dictionary book or type it in digital dictionary either. According to those problems, how if we change the way users input the word to the dictionary? Image that is captured by smartphone's camera can be the alternative way to input the word to the digital dictionary. With this approach, we expect the application will help 28 million visitors who visit Russia [1] to translate the Russian word easier.

However, using the image from the smartphone's camera emerges a new problem. As the computer cannot read the text directly from the image, it needs an image recognition technique to read the text in the image. Quraishi provided a new approach for image recognition using Artificial Neural Networks [2]. It calculated total average error on generated data matrix networks on test data matrix to check



whether proper identification can be made or not. The observation result that the value of average error is less than that of the test image without application of artificial neural network. In image character recognizing, Bai proposed a Shared-Hidden-Layer Deep Convolutional Neural Network (SHL-CNN) [3]. SHL-CNN can reduce recognition error by 16-30% relatively compared with conventional CNN. This result verified the effectiveness of the learned SHL-CNN on both English and Chinese image character. Another research that utilizes Self-Organizing Map (SOM) to recognize speech has been conducted [4]. It obtains maximum feature Intensity as 98.17% for Mean-SOM performance and 98.54% for Median-SOM performance.

This paper is organized as follows. In section 2, we describe the research methodology to recognize the Russian character. In section 3, we discuss the result of the research. Finally, we draw some conclusions section 4.

2. Research Methodology

Generally, as shown in figure 1, there are several steps to recognize Russian character such as image input, cropping preferred word, image pre-processing, feature extraction and image recognition using Self-Organizing Map algorithm [5]. The client provides image input either from smartphone's gallery or directly from camera. As the image input may contain more than one word or other objects, it should be cropped to separate it with unnecessary objects in the image. As the result, the image input will contain only the preferred word.

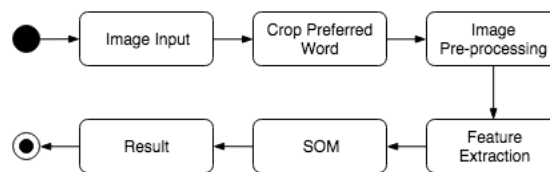


Figure 1. Russian character recognition method

To enhance the image input, some image pre-processing such as noise filtering, binary image processing, segmentation, scaling and thinning are applied. The first step of image pre-processing is noise filtering. Usually images taken by the digital camera will contain noise. In this research, application utilizes median filter which is much better at preserving straight edge structure than Gaussian smoothing [6] to reduce noise. Application applies median filter with neighbor size 3x3. The process will be started from the first pixel in coordinate 0,0 to the last pixel. Neighbor size 3x3 means the process picks the neighbor coordinate of the current coordinate starts from x-1, y-1 to x+1, y+1. Notation x and y represent current coordinate (x, y). The result will be assembled to a matrix. From the matrix, the color values are sorted and then the median value of red, green and blue (RGB) are calculated.

The next step after noise filtering is binary image process. Binary image means the image only consists of two colors, black and white. In this step, binary image is yielded by implementing Otsu algorithm [7]. This algorithm only accepts grayscale image as input. Therefore, grayscale process is applied to the image before implementing Otsu algorithm. This algorithm needs a threshold value. Pixel color which exceeds threshold value will be converted to white, otherwise it will be converted to black. The expected result of this step is the characters in the image will be black and the background will be white.

After binary image processing, the next step is segmentation. In this research, Self-Organizing Map is designed to recognize a character. The word should be separated into characters which is represented in black pixel. The process involved filtering out the white pixel and keeping the black pixel. The segmentation process will consider two stick objects in the image as one object. Therefore, this system will not recognize handwriting font. As the result, segmentation process yields images which contain each character from the word.

The images from the segmentation process will have different size. The size is scaled to 20x25 pixels to uniform the image size for thinning process. The thinning process is done to obtain character's skeleton. The character is thinned down to a skeleton of unitary thickness, one pixel. This research utilizes Zhang-Suen algorithm to obtain the character's skeleton [8]. This process should be done because the thickness of the same character of different fonts are not equal. Usually the same character of different fonts will share the similar character's skeleton.

At the end of image pre-processing, the result is array of images that contain a character for each image. Every character in the image will be extracted with pixel mapping method. Value 1 will be given to the black pixel, and 0 will be given to the white pixel. The extracted value will be saved in a string. Because the size of each image is 20x25 pixels, then the size of the string is $20 \times 25 = 500$ characters. Value extracted from pixel in coordinate (0,0) and coordinate (0,1) will be put in string index 0 and 1 respectively, and so forth. String from the feature extraction process will be sent to server for character recognition process.

Character recognition is done by utilizing Self-Organizing Map algorithm, one of the neural network branch. Training is required to be done to the system in order to utilize this algorithm for character recognition. Training uses unsupervised method with learning rate 0.1. The system is trained with 11 random fonts which support Cyrillic letters. Every font consists of 33 lowercase and 33 uppercase letters.

Three methods are used to test the accuracy of Self-Organizing Map implementation. The first method generates images to represent each character of 15 random fonts that support Cyrillic character. Computer will generate 33 images of lowercase letters and 33 images of uppercase letters for each font. These images will be tested against the earlier training dataset. The second method, three fonts are randomly chosen. All the lowercase and uppercase letters of the chosen fonts are printed in white background paper. Those characters will be captured by using smartphone's camera in a room. The camera resolution is 8 Megapixel. Third method, the test is done by capturing the Cyrillic word from posters, leaflets, and newspapers using the smartphone's camera. The accuracy of testing dataset will be calculated with the formula in (1). All photos are taken twice, and the best picture will be taken to be the dataset input.

$$Accuracy (\%) = \frac{Recognized\ Pattern}{Sum\ of\ Testing\ Pattern} \times 100\% \quad (1)$$

3. Result

The process of character recognition is divided into two separate parts, user's smartphone and server. User's smartphone is used as image input and image enhancement. Server is used to do the character recognition process.

Some image pre-processing such as noise filtering, binary, segmentation and scaling are done before applying Self-Organizing Map to a cropped image. For example, we will use the cropped image as shown in figure 2.

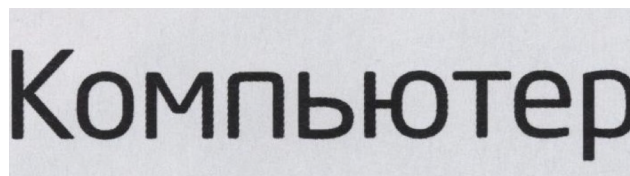


Figure 2. Cropped image as input

The first image pre-processing step is noise filtering. As mentioned above, in this research median filter is used to reduce noise in an image. The result of the noise filtering is shown in figure 3.



Figure 3. Image after noise filter being applied

The next step is converting the image to black and white image. This step is called binary process. The color which exceed the threshold will be treated as white pixel, otherwise will be treated as black pixel. The result after binary process is shown in figure 4.



Figure 4. Image after binary image process

Self-Organizing Map will not recognize the whole word. It will recognize a character only. To support that requirement, character separation from the word is applied to the image. This action is called segmentation. As the result of segmentation process, there are new images for each character as shown in figure 5. In this process, failure to separate character might happen. As shown in figure 6, this failure is caused by more than one character are too close or stick with the character next to it.

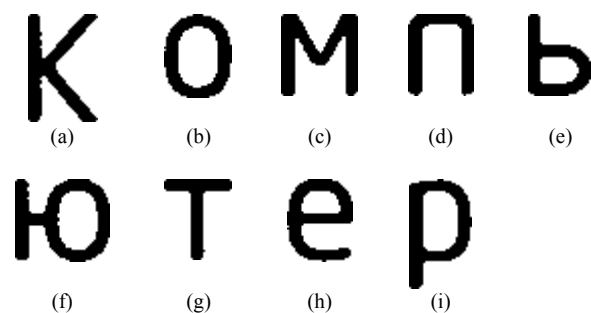


Figure 5. New images (a-i) after segmentation



Figure 6. Segmentation failure

Each character has its own image after segmentation process. The next process is scaling. The image is scaled to 20x25 pixels. The result is shown in figure 7.

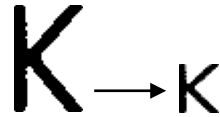


Figure 7. Image after scaled down to 20x25 pixels

After being scaled, all new images are thinned down to yield character's skeleton image. The result of the thinning process is shown in figure 8.

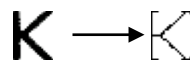


Figure 8. Image after thinning process

After thinning process, the image is ready to be extracted. Extracted feature string size is 500 characters long, based on image width and height.

For training dataset, the system has been trained with 11 random fonts that support Cyrillic. They are, Arial, Arial Black, Batang, Calibri, Cambria, Comic Sans, Courier New, Lucida Sans, Tahoma, Times New Roman, and Verdana.

To test the character and word recognition, three testing methods has been conducted. The first method, computer generates images of each character from 15 random fonts. The character is represented by the black pixel, and the background is white. As shown in table 1, Copperplate Gothic and Corbel font are fully recognized by the system. After being tested against training data, the accuracy of the data testing is 89.09%.

Table 1. Data Testing for Method 1

Font type	Recognized	Not Recognized
Arial Narrow	65	1
Book Antiqua	54	12
Lucida Console	63	3
Calisto MT	57	9
Candara	65	1
Century Gothic	55	11
Consolas	63	3
Constantia	53	13
Copperplate Gothic	66	0
Corbel	66	0
Georgia	55	11
Impact	46	20
MS Sans Serif	64	2
Palatino Linotype	46	20
Segoe UI	64	2

Second testing method is similar with the first method, unless all of the letters are printed and then they are captured by the smartphone's camera. Also, in this method, only three fonts are used. They are

Arial Narrow, Book Antiqua and Lucida Console as shown in Table 2. If we compare to the first method, the character recognition is slightly drop to 88.89% when images were taken by the smartphone's camera.

Table 2. Data Testing for Method 2

Font type	Method 1		Method 2	
	Recognize	Fail	Recognize	Fail
Arial Narrow	65	1	64	2
Book Antiqua	54	12	64	2
Lucida Console	63	3	48	18

The third method, testing is done to recognize a full word. The word image is taken by using smartphone's camera. Source of the image is posters, leaflets and newspapers. They are taken directly in outdoor at noon. From 350 photos of different words taken, the system is successfully recognized 292 words, while 58 other words were partially recognized. Some letters were not successfully recognized. The accuracy for the word recognition is 83.42%.

4. Conclusion

Character recognition by Self-Organizing Map (SOM) from images that generated by computer or captured by smartphone's camera has slightly different accuracy. This means the image source does not have much influence to the character recognition accuracy.

Partial word recognition is caused by the failure of recognizing some letters, not the whole letters in a word. For future works, there are some approaches that can be taken to increase word recognition accuracy. First, enhance the segmentation process in order to obtain clearly separate images for closed characters in a source image. Second, apply a prediction algorithm to predict the word based on the partially recognized character.

References

- [1] UNWTO 2014 UNWTO Tourism Highlight, 2014 (World Tourism Organization)
- [2] Quraishi M I, Choudhury J P and De M 2012 Image recognition and processing using Artificial Neural Network *Recent Advances in Information Technology (RAIT), 2012 1st Intl. Conf. on* 95-100 (IEEE)
- [3] Bai J, Chen Z, Feng B and Xu B 2014 Image character recognition using deep convolutional neural network learned from different languages *Image Processing (ICIP), 2014 IEEE Intl. Conf. on* 2560-2564 (IEEE)
- [4] Venkateswarlu R L K and Kumari R V 2011 Novel approach for speech recognition by using self — Organized maps *Emerging Trends in Networks and Computer Communications (ETNCC), 2011 Intl. Conf. on* 215-222 (IEEE)
- [5] Kohonen T 1990 The self-organizing map *Proc. of the IEEE* 1464-1480 (IEEE)
- [6] Smith S M and Brady J M 1997 SUSAN— A New Approach to Low Level Image Processing *International Journal of Computer Vision* **23** 45-78
- [7] Otsu N 1979 A Threshold Selection Method from Gray-Level Histograms *Systems, Man and Cybernetics, IEEE Transactions on* **9** 62-66 (IEEE)
- [8] Zhang T Y and Suen C Y 1984 A fast parallel algorithm for thinning digital patterns *Communications of the ACM* **27** 236-239 (ACM)