International Journal of Advanced Culture Technology Vol.6 No.2 123-128 (2018)
DOI https://doi.org /10.17703/IJACT.2018.6.2.123

IJACT18-2-17

Character Recognition Algorithm using Accumulation Mask

Suk Won Yoo

SeoKyeong Univ., Dept. of Computer Science, Seoul, Korea
swyoo@skuniv.ac.kr

Abstract

Learning data is composed of 100 characters with 10 different fonts, and test data is composed of 10 characters with a new font that is not used for the learning data. In order to consider the variety of learning data with several different fonts, 10 learning masks are constructed by accumulating pixel values of same characters with 10 different fonts. This process eliminates minute difference of characters with different fonts. After finding maximum values of learning masks, test data is expanded by multiplying these maximum values to the test data. The algorithm calculates sum of differences of two corresponding pixel values of the expanded test data and the learning masks. The learning mask with the smallest value among these 10 calculated sums is selected as the result of the recognition process for the test data.

The proposed algorithm can recognize various types of fonts, and the learning data can be modified easily by adding a new font. Also, the recognition process is easy to understand, and the algorithm makes satisfactory results for character recognition.

Keywords: *Image Processing, Character Recognition, Accumulation, Correlation, Mask, Image Analysis, Feature Extraction*

1. Introduction

Due to the development of office automation technology, character recognition technology is used to classify documents and parts in various fields such as delivery service companies, offices, and factories. Algorithms for recognizing characters with a specific font can achieve a higher recognition rate than those using various fonts, but in reality, various fonts are used. In addition, this type of recognition technology has developed into a handwriting recognition technology, and it can be applied to the field of military weapons that tracks specific objects that continuously change in shape captured by the camera over time.

2. The Related Works

In order to distinguish various characters, it is important to find patterns that can determine the characteristics of these characters [1]. Conventional methods extract features by drawing horizontal or vertical lines [2] or extract the skeleton of characters [3]. However, these methods have different results depending on the fonts of characters in the test [4], so they are limited to specific fonts and are not suitable

Manuscript received: June 20, 2018 / revised: June 21, 2018 / Accepted: June 22, 2018
Corresponding Author: swyoo@skuniv.ac.kr
Tel:+82-2-940-7181
SeoKyeong Univ., Dept. of Computer Science, Korea

for various types of fonts [5]. Other studies using neural networks have also been conducted to solve these drawbacks [6]. The new approach using neural network theory [7] can recognize various fonts of characters, but it requires a lot of learning time, and there is a problem in adding new learning data [8]. In this paper, a new algorithm is proposed that recognizes characters in a way that gives a high weight to high frequent parts by accumulating various fonts of learning data.

3. The Main Subject

Because artistic concepts are important nowadays in publishing works, various fonts of characters are preferred instead of metal types. As a result, it has become a more difficult process to recognize artistic characters instead of those with metal types. Conventional character recognition algorithms extracted features constituting characters or analyzed structure of characters. Therefore, only certain fonts of characters were used for character recognition. A typical example is the recognition of license plate. However, this method is limited to specific typefaces and is difficult to apply to various types of fonts.

In this paper, we propose a method to accumulate same characters with different fonts of learning data in order to comprehensively consider variety of learning data with various types of fonts. Through this process, minute differences between various fonts of learning data are eliminated, and test data is recognized as the learning data with minimum error through comparison process with modified learning data.

3.1 The Basic Concept

For 10 characters from 0 to 9, learning data consists of a total of 100 characters with 10 different fonts, and the test data consists of 10 characters with a new font not used by the learning data. The learning data and the test data are black and white images, where the text area is represented by black on a white background.

To comprehensively consider the variety of learning data with 10 different fonts, we add 10 different fonts for each character in pixel unit and store them in the learning mask of the corresponding character. This method eliminates the minute differences between the different fonts. After multiplying the test data by the maximum values of the learning masks, the sum of the differences of the respective pixel values of the extended test data and learning masks is calculated. The learning mask with the minimum value among these calculated values is finally selected as the recognized character.

3.2 Character Recognition Algorithm using Accumulation Mask

Character recognition algorithm using accumulation mask will be described in more detail through the following steps.

Step 1) Learning data is composed of 10 characters from 0 to 9, and each character has 10 different types of fonts. The test data is also composed of 10 characters from 0 to 9, and it is assumed that the test data has a new font not used for the learning data. Thus, the learning data consists of a total number of 100 characters, and the test data has 10 characters.

Step 2) Learning data and test data are black and white images, and the text area is black with white background. Each character of learning data and test data is 17x17 size because of 15x15 size of image itself with border margins of 1 size in the upper, lower, right and left regions. Therefore, the total size of the learning data is 170x170, and the total size of the test data is 170x17.

Step 3) Make 10 learning masks of 170x17 size and initialize them to 0. Since the learning data and the test data are black and white images, white representing the background is regarded as 0, and black representing the character region is regarded as 1. Then, for each of the 10 characters of learning data 0 to 9, 10 different fonts are added in pixel unit and stored in the learning mask of the corresponding character. Because the black color representing the character area in the learning data is represented by 1, and 10 different fonts are added in pixel unit for each of the 10 characters of learning data 0 to 9, each pixel has a value from 0 to 10.

Step 4) For each of the 10 test data,

Step 4-1) For each of the 10 learning masks,

Step 4-1-1) Find the maximum value in the corresponding learning mask of 15x15 size.

Step 4-1-2) The maximum value obtained above is multiplied to the corresponding test data. Let the result be modified test data. Then, the test data composed of 0 and 1 becomes the modified test data composed of 0 and the maximum value.

Step 4-1-3) Sum of the differences of the respective pixel values of the extended test data and the corresponding learning mask is calculated.

Step 4-2) Each test data has 10 results calculated with 10 learning masks. Find the learning mask with the smallest of the 10 result values. This learning mask is selected as the finally recognized character among the 10 learning data characters for the test data.

3.3 The Results

The following figures show the results of execution of the character recognition algorithm using accumulation mask. Figure 1 shows the learning data consisting of 10 characters from 0 to 9. Each of 10 learning data characters has 10 different types of fonts. Therefore, the learning data consists of a total of 100 characters. Figure 2 shows test data consisting of 10 characters from 0 to 9. The 10 test data characters have a new font not used for the learning data.

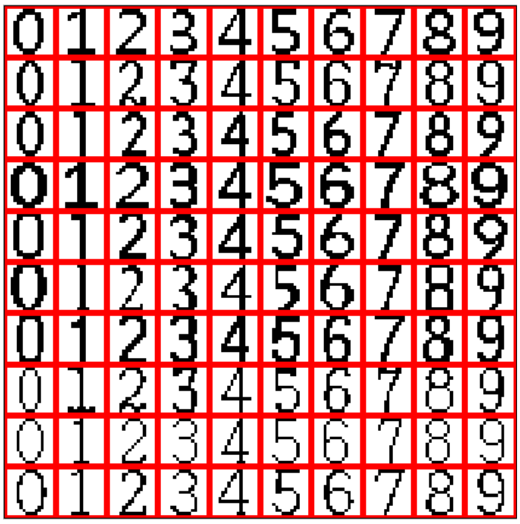

Figure 1. Learning Data Set of 170x170 size

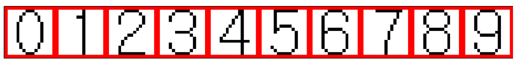

Figure 2. Test Data Set of 170x17 size

The learning data and the test data are black and white images, and the text area is represented by black on a white background. White representing the background of the learning data and test data is regarded as 0, and black representing the character region is regarded as 1. The total size of the learning data is 170x170, and the total size of the test data is 170x17.

Figure 3 shows the result of adding 10 different fonts of character 7 in pixel unit, and storing them in learning mask 7. Since the pixels that constitute learning mask 7 have values in the range from 0 to 10, the maximum value of learning mask 7 is 10. In Figure 3, the rectangular area corresponds to the character area of test data 7. Since the maximum value of the learning mask 7 is 10, the pixels in the rectangular area of the learning mask 7 correspond to the maximum value 10 of the test data 7, and the pixels outside the rectangular area correspond to 0.

0	0	0	6	9	9	9	8	8	8	10	10	4	0	0
0	0	0	5	8	8	7	7	7	6	7	9	3	0	0
0	0	0	2	3	0	0	0	1	2	7	8	0	0	0
0	0	0	2	2	0	0	0	0	5	10	3	0	0	0
0	0	0	0	0	0	0	0	2	7	8	1	0	0	0
0	0	0	0	0	0	0	0	5	8	5	0	0	0	0
0	0	0	0	0	0	0	2	6	8	1	0	0	0	0
0	0	0	0	0	0	0	4	8	5	0	0	0	0	0
0	0	0	0	0	0	2	6	8	3	0	0	0	0	0
0	0	0	0	0	0	3	7	7	0	0	0	0	0	0
0	0	0	0	0	0	4	8	5	0	0	0	0	0	0
0	0	0	0	0	1	6	8	3	0	0	0	0	0	0
0	0	0	0	0	4	8	8	1	0	0	0	0	0	0
0	0	0	0	1	4	8	6	1	0	0	0	0	0	0
0	0	0	0	2	6	8	4	1	0	0	0	0	0	0

Figure 3. Learning MASK #7 and Test Data #7

Figure 4 shows the difference values of the pixels of the learning mask 7 and the corresponding pixels of the test data 7. The sum of these difference values, 260, represents the degree of discordance between the learning mask 7 and the test data 7.

0	0	0	6	1	1	1	2	2	2	0	0	6	0	0
0	0	0	5	8	8	7	7	7	6	7	1	3	0	0
0	0	0	2	3	0	0	0	1	2	7	2	0	0	0
0	0	0	2	2	0	0	0	0	5	0	3	0	0	0
0	0	0	0	0	0	0	0	2	7	2	1	0	0	0
0	0	0	0	0	0	0	0	5	2	5	0	0	0	0
0	0	0	0	0	0	0	2	6	2	1	0	0	0	0
0	0	0	0	0	0	0	4	2	5	0	0	0	0	0
0	0	0	0	0	0	2	6	2	3	0	0	0	0	0
0	0	0	0	0	0	3	7	3	0	0	0	0	0	0
0	0	0	0	0	0	4	2	5	0	0	0	0	0	0
0	0	0	0	0	1	6	2	3	0	0	0	0	0	0
0	0	0	0	0	4	8	2	1	0	0	0	0	0	0
0	0	0	0	1	4	8	4	1	0	0	0	0	0	0
0	0	0	0	2	6	8	6	1	0	0	0	0	0	0

Figure 4. Discordance of Learning MASK #7 and Test Data #7

Table 1 shows the discordance degree, the minimum discordance degree, and the finally selected resulting learning mask that are calculated with 10 learning masks from 0 to 9 for each of the 10 test data characters. For example, test data character 7 has a degree of discordance of 669 with learning mask 0, and discordance of 260 with learning mask 7. As a result, the test data character 7 is recognized as the learning mask 7 because the test data character 7 has the minimum discordance of 260 with the learning mask 7 among 10 learning masks from 0 to 9. As shown in Table 1, the character recognition algorithm using accumulation mask obtains 100% recognition results for 10 test data.

Table 1. Discordance of Test Data

	MASK #0	#1	#2	#3	#4	#5	#6	#7	#8	#9	MASK#	MIN
Test #0	335	564	525	490	671	522	465	558	548	464	0	335
Test #1	709	306	555	582	603	642	681	440	738	682	1	306
Test #2	611	472	391	480	701	588	629	470	570	546	2	391
Test #3	563	510	445	368	675	488	505	504	458	484	3	368
Test #4	625	612	607	582	357	664	653	576	652	660	4	357
Test #5	583	522	547	460	677	368	513	510	518	552	5	368
Test #6	523	568	555	436	681	412	385	570	422	556	6	385
Test #7	669	404	467	506	625	588	685	260	670	634	7	260
Test #8	535	568	521	408	669	418	423	558	388	544	8	388
Test #9	523	568	499	458	741	504	525	562	524	404	9	404

3.4 The Pros and cons of the proposed Character Recognition Algorithm

Similar to the conventional recognition methods, there are advantages and disadvantages in the proposed algorithm. Advantages include: 1) it recognizes various types of fonts. 2) Unlike the method of analyzing the structure of characters, the learning data can be easily modified even if a new font is added. 3) The recognition process is easy to understand, and satisfactory recognition results can be obtained. 4) It requires less computation time compared with existing recognition methods. The disadvantage is that it is hard to recognize characters with abstract fonts especially emphasizing artistic aspects as in the conventional techniques because it is difficult to distinguish minute differences between learning data.

4. Conclusion and Future work

The character recognition algorithm using accumulation mask proposed in this paper comprehensively considers the variety of fonts of learning data. By giving high weights to more frequent parts of the font, it is possible to recognize characters of a new font that has not been learned. Also, the algorithm is easy to implement, and a satisfactory recognition result is obtained. In the future research, we are studying new character recognition algorithm considering the structural modification for the variety of the fonts.

References

[1] C. Leondes, *Image Processing and Pattern Recognition*, Academic Press, pp. 22-35, 1998.

[2] G. Dougherty, *Pattern Recognition and Classification*, Springer, pp. 123-134, 2012.

[3] Y. Zhang, C. Suen, "A Fast Parallel Algorithm for Thinning Digital Patterns," *Comm. ACM*, Vol. 27, No. 3, pp. 236-239, 1984.

[4] Yong Jun Yang, Sang Gu Lee, "An Object Tracking Method for Studio Cameras by OpenCV-based Python Program", *The Journal of Convergence on Culture Technology(JCCT)*, Vol.4, No.1, pp.291-297, Feb 2018.

[5] Nattapong Phanthuna, Fusak Cheevasuwit, "Contrast Image Enhancement Using Multi-Histogram

Equalization", *International Journal of Advanced Culture Technology(IJACT)*, Vol.3, N0.2, pp.161-170, 2015.

[6] D. Rumelhart, J. McClelland, *Pattern Distributed Processing*, MIT Press, pp. 121-127, 1987.

[7] M. Anthony, P. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge Univ. Press, pp. 131-149, 2009.

[8] P. Liu, H. Li, *Fuzzy Neural Network Theory And Application*, World Scientific, pp. 68-94, 2004.