International Journal of Advanced Culture Technology Vol.7 No.1 64-69 (2019)
DOI 10.17703/IJACT.2019.7.1.64

IJACT 19-3-10

Character Recognition using Regional Structure

Suk Won Yoo

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

Abstract

With the advent of the fourth industry, the need for office automation with automatic character recognition capabilities is increasing day by day. Therefore, in this paper, we study a character recognition algorithm that effectively recognizes a new experimental data character by using learning data characters. The proposed algorithm computes the degree of similarity that the structural regions of learning data characters match the corresponding regions of the experimental data character. It has been confirmed that satisfactory results can be obtained by selecting the learning data character with the highest degree of similarity in the matching process as the final recognition result for a given experimental data character.

Keywords: *Regional Structure, Regional Boundary, Image Restoration, Image Processing, Character Recognition, Correlation, Accumulation, Image Analysis, Feature Extraction*

1. Introduction

As science and technology have developed, new fields such as the Fourth Industrial Revolution have appeared. Due to the development of science and technology such as artificial intelligence, Internet of things and virtual reality, positive effects such as improvement of human health and life in medical field, increase of production of goods in industrial field and increase of efficiency of work progress in commercial field are appearing these days. In particular, development in the commercial sector such as office automation has given a lot of free time and convenience to humans, and character recognition is an important part of office automation.

2. The Related Works

Character recognition is a technique to transform document information of image style to new information style that computer systems can process. Typical studies on character recognition are mainly based on a comparison method of values of 2D arrays representing character images [1], a method of expressing the characteristics of character by a vector [2], and a method of comparison of correlation by extracting important components of characters [3]. These methods have high recognition rates for fixed fonts. However, it is not recommended to use these methods for the artfully varied fonts because it is difficult to distinguish the minute differences between fonts. In order to eliminate these differences, new approaches using neural network theory [4] or chaotic theory [5] are also being studied.

Manuscript received: January 15, 2019/revised: February 18, 2019 / Accepted: February 24, 2019
Corresponding Author: swyoo@skuniv.ac.kr
Tel:+82-2-940-7751
SeoKyeong Univ., Dept. of Software, Korea

3. The Main Subject

Nowadays, due to the development of publishing technology, various types of fonts are used, and the recognition of various fonts is an important process in office automation and information system [6]. Conventional recognition methods are based on typical fonts, but these methods are not useful to apply to various types of font recognition [7]. In this paper, we propose an algorithm that recognizes the structure of each character in regional form by using various fonts of learning data characters and hence recognizes characters with a new font not used for the learning data.

3.1 The Basic Concept

Learning data consists of 10 different characters from 0 to 9, and each of learning data character has 10 different types of fonts. For learning data characters, 0 represents the background and 1 represents the text area. We divide each of learning data characters into 25 tile areas of 3x3 size and then examine the number of pixel value 1 belonging to 15 tile areas located in the middle part. These tile regions are classified into meaningful structure regions and meaningless structure regions according to the number of pixel value 1 belonging to 15 tile regions of 10 different fonts of the learning data character. Test data is also composed of 10 different characters from 0 to 9, and it has a new font not used for the learning data. For each of test data characters, we classify its tiles as meaningful tiles and meaningless tiles according to the number of pixel value 1 belonging to 15 tile areas.

For each of test data characters, recognition ratios with 10 learning data characters are calculated by examining the number of occurrence of the following two cases: the first case that the meaningful tile of the test data character corresponds with the meaningful structure area of the learning data character, or the second case that the meaningless tile of the test data character does with the meaningless structure area of the learning data character. After all recognition ratios for the 10 learning data characters are obtained, the learning data character with the highest recognition ratio is selected as the recognition result for the corresponding test data character.

3.2 Character Recognition using Regional Structure

Character recognition algorithm using regional structure will be described in more detail by the following steps.

Step 1) Learning data consists of 10 different characters from 0 to 9, and each learning data character has 10 different fonts. Each learning data character is a black and white image of 15x15 size, where 0 represents background and 1 does text area.

Step 2) Learning data character has empty spaces at left and right sides and has long character area in the vertical direction. Thus, each of the 100 learning data characters is divided into 25 tile regions of 3x3 size, and then the number of pixel value 1 belonging to 15 tile regions located in the center is calculated. Since the size of the tile is 3x3, the number of pixel value 1 belonging to each tile area can be from 0 to 9. From the point of view of forming the structure of the character, a tile with more number of pixel value 1 can be considered to have more weight to form a structure of the corresponding character.

Step 3) Learning data characters have 10 different fonts. For each learning data character,

Step 3-1) For each of 15 tile areas of the learning data character,

Step 3-1-1) If there exists more than 5 number of fonts with the condition that the number of pixel value 1 belonging to the corresponding tile area is greater than or equal to 2, the corresponding tile area is considered as a meaningful structure area. Conversely, if the number of fonts is less than or equal to 5, it is considered as a meaningless structure area.

Step 4) Test data is also composed of 10 different characters from 0 to 9, and it has a new font not used for learning data. For each test data character,
Step 4-1) For each of 15 tile areas of the test data character,
Step 4-1-1) If the number of pixel value 1 belonging to the corresponding tile area is greater than or equal to 2, the tile is classified into meaningful tile representing the structure of the corresponding character. Conversely, if less than or equal to 1, it is classified as meaningless tile.
Step 5) For each of 10 test data characters,
Step 5-1) For each of 10 learning data characters,
Step 5-1-1) Among the 15 tile areas, calculate the number of occurrence of the following two cases: 1) the meaningful tile of the test data character corresponds with the meaningful structure area of the learning data character, or 2) the meaningless tile of the test data character corresponds with the meaningless structure area of the learning data character. This value becomes the recognition ratio that the test data character matches to the learning data character.
Step 5-2) After finding all recognition ratios for the 10 learning data characters, the learning data character with the highest recognition ratio is selected as the recognition result for the test data character.

3.3 The Results

Let's suppose that Learning Data Set [8] is used as a prototype data for the recognition process and Test Data Set [8] is given as experimental data, where the learning data is composed of 100 characters with 10 different fonts and the test data has a new font not used for the learning data. Each of the learning data characters and the test data characters is a 15×15 size black and white image. The background has a pixel value of 0 in white, and the text area does a pixel value of 1 in black.

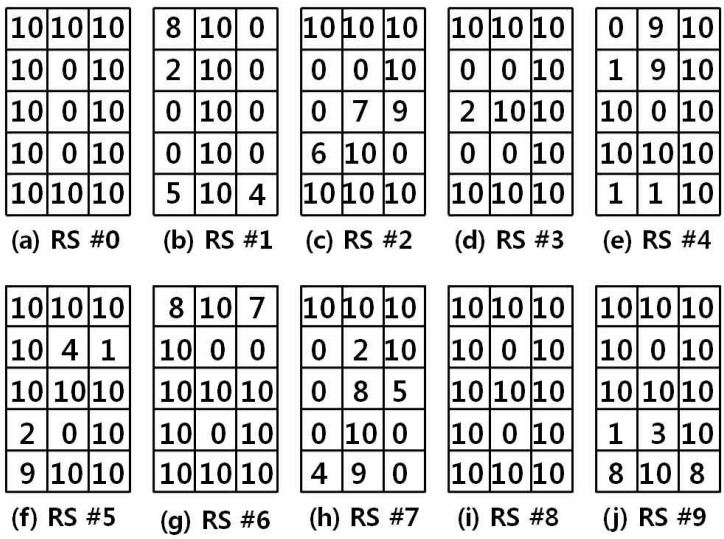

Figure 1. Number of Fonts of 15 tiles of Learning Data Characters

For those 15 tile areas of each learning data character, Figure 1 shows the number of fonts with the condition that the number of pixel value 1 belonging to the tile area is greater than or equal to 2, and Figure 2 does the corresponding meaningful structure area (denoted by O) and meaningless structure area (denoted by X) for the learning data characters.

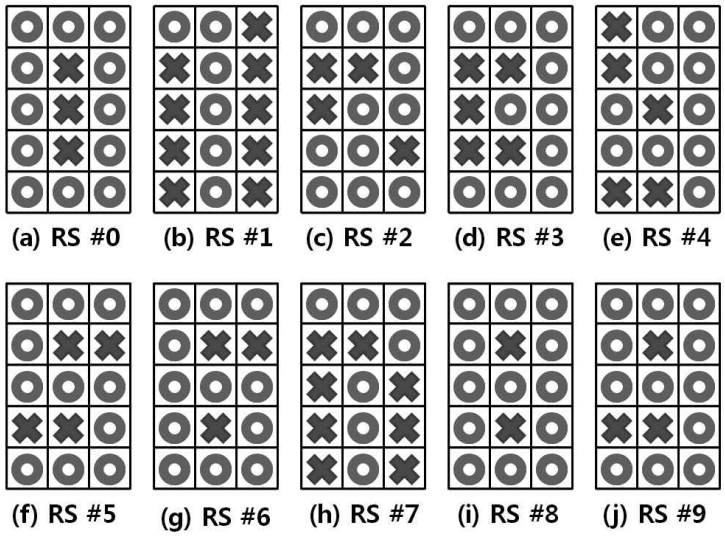

Figure 2. 10 Regional Structures of Learning Data Characters

Learning data characters have 10 different types of fonts. For each of the learning data characters, if there exist more than 5 number of fonts with the condition that the number of pixel value 1 belonging to the corresponding tile area is greater than or equal to 2, the tile area is classified as a meaningful structure area. Here, the reason why the specific tile area is classified as a meaningful structure area is that the tile forms the regional structure of more than half of the 10 different fonts of the corresponding character.

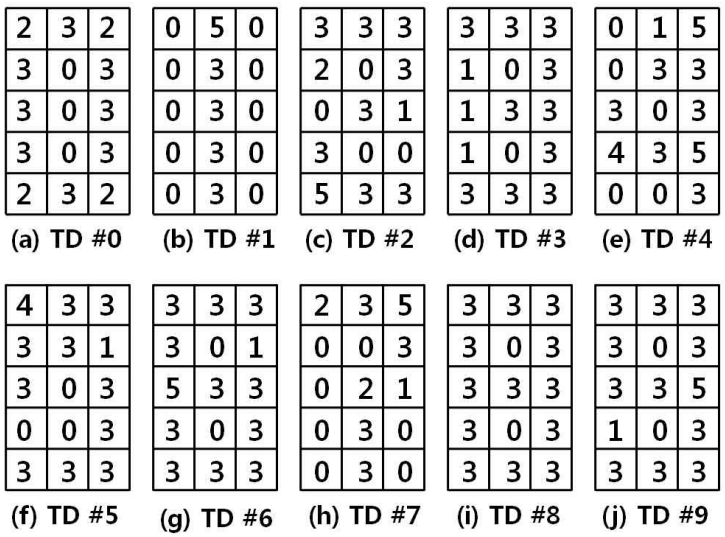

Figure 3. Number of Pixel Value 1 of 15 tiles of Learning Data Characters

For those 15 tile areas of each test data character, Figure 3 shows the number of pixel value 1 belonging to the tile area, and Figure 4 does the meaningful tile (denoted by O) and meaningless tile (denoted by X) that represent the structure of each of 10 test data characters.

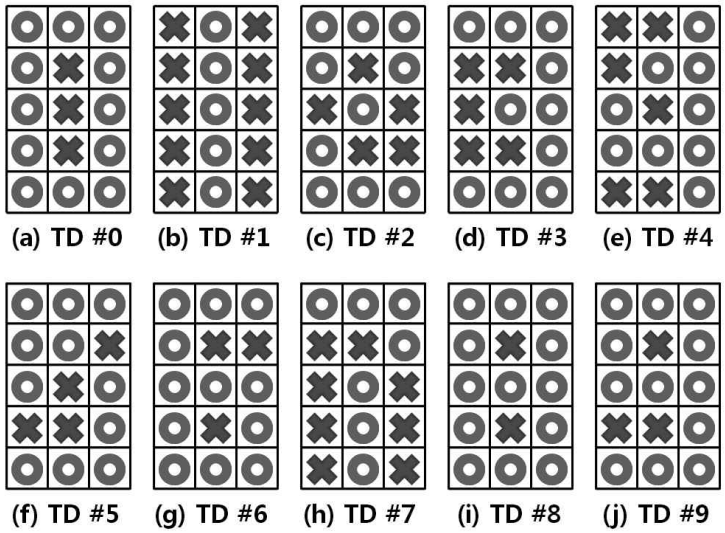

Figure 4. 10 Regional Structures of Test Data Characters

For each of 10 test data characters, Table 1 shows recognition ratios with 10 regional structures of learning data characters, the highest recognition ratio, and the learning data character with the highest recognition ratio. As shown in Table 1, the total recognition rate for all 10 test data characters is 100%, and the duplicated recognition rate is 10%. Test data character 2 is recognized simultaneously as learning data character 2 and 8 with 80% recognition ratio. For each of 10 test data characters, Figure 5 shows matching results with the highest recognition ratio with the regional structures of learning data characters. In Figure 5, the tile area represented by a star symbol (*) is a mismatched area between the test data character and the learning data character with the highest recognition ratio.

Table 1. Recognition of Test Data with Regional Structure of Learning Data

	RS#0	#1	#2	#3	#4	#5	#6	#7	#8	#9	MAX	RS#
TD #0	1.00	0.20	0.66	0.73	0.60	0.80	0.87	0.40	0.93	0.87	1.00	0
TD #1	0.13	0.93	0.47	0.40	0.40	0.33	0.27	0.73	0.20	0.27	0.93	1
TD #2	0.73	0.47	0.80	0.73	0.33	0.67	0.73	0.67	0.80	0.73	0.80	2, 8
TD #3	0.73	0.47	0.80	1.00	0.47	0.80	0.73	0.67	0.80	0.87	1.00	3
TD #4	0.53	0.27	0.47	0.40	0.93	0.33	0.40	0.33	0.47	0.40	0.93	4
TD #5	0.80	0.40	0.47	0.67	0.53	0.87	0.80	0.33	0.73	0.80	0.87	5
TD #6	0.87	0.33	0.67	0.73	0.47	0.93	1.00	0.40	0.93	0.87	1.00	6
TD #7	0.40	0.80	0.73	0.67	0.40	0.47	0.40	1.00	0.47	0.53	1.00	7
TD #8	0.93	0.27	0.73	0.80	0.53	0.87	0.93	0.47	1.00	0.93	1.00	8
TD #9	0.87	0.33	0.67	0.87	0.47	0.93	0.87	0.53	0.93	1.00	1.00	9

3.4 The Pros and cons of the proposed Character Recognition Algorithm

Similar to the conventional character recognition techniques, the proposed method in this paper also has advantages and disadvantages. Advantages include: 1) the image recognition process is easy to understand and easy to implement, 2) satisfactory recognition ratio can be obtained when it is compared with existing recognition methods, and 3) learning data can be easily expanded by adding new fonts. Disadvantages are 1) different recognition results might be obtained according to the font styles of the given learning data characters,

and 2) if the font of the test data is too much artfully deformed, the test data might not be recognized.

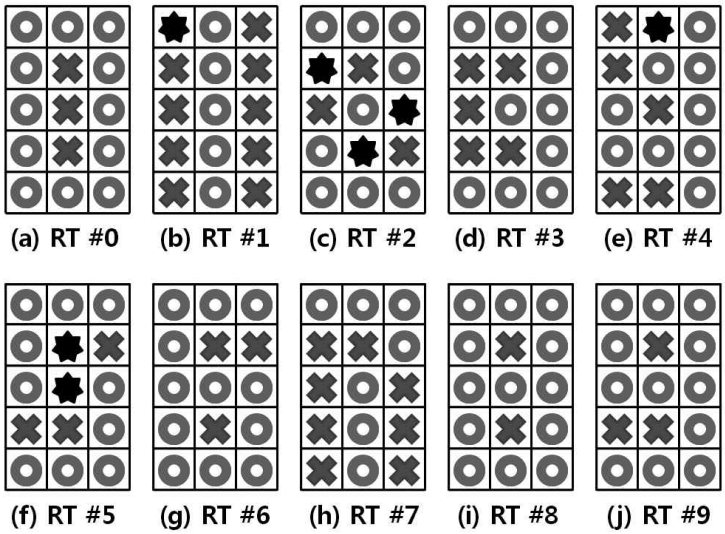

Figure 5. Matching Results of Regional Structures of Learning Data and Test Data

4. Conclusion and Future work

The character recognition algorithm using regional structure in this paper recognizes the structure of each character in an area form by using various fonts of the learning data characters. By using the structural differences of the learning data characters obtained by this process, we can make the following significance of this study that it is possible to recognize characters with a new font not used for learning data. We are now studying how to extract skeletons of characters in future research.

References

[1] A. Das, *Guide to Signals and Patterns in Image Processing*, Springer, pp. 14-38, 2015.
[2] F. Shih, *Image Processing and Mathematical Morphology*, CRC Press, pp. 11-20, 2009.
[3] T. Acharya, A. Ray, *Image Processing*, Wiley, pp. 61-77, 2005.
[4] D. Rumelhart, J. McClelland, *Parallel Distributed Processing*, MIT Press, pp. 121-127, 1987.
[5] A. Crilly, R.Earnshaw, H. Jones, *Fractals and Chaos*, Springer-Verlag, pp. 89-97, 2012.
[6] K. Yoon, Y Chang, "IOT-based SMEs Producing Standardized Information System Model Analysis and Design", Journal of the Convergence on Culture Technology(JCCT), Vol.2, No.1, pp.87-91, 2016.
[7] R. Gonzalez, R. Woods, Digital Image Processing, Prentice Hall, pp. 269-286, 2008.
[8] S. Yoo, "Character Recognition Algorithm using Accumulation Mask", International Journal of Advanced Culture Technology(IJACT), Vol.6, No.2, pp.123-128, 2018.