Stroke Verification with Gray-level Image for Hangul Video Text Recognition

Jinsik Kim, Seonghun Lee, Younghee Kwon, Jin H. Kim
Department of E.E.C.S, Korea Advanced Institute of Science and Technology
{jskim,leesh,kyhee,jkim}@ai.kaist.ac.kr

Abstract

Traditional OCR uses binarization technique, which makes OCR simple. But it makes strokes ambiguous and that causes recognition errors. Main reason of those errors is similar grapheme pair confusing error. It can be reduced by verifying ambiguous area of gray level image. After checking whether there is similar grapheme pair by analyzing traditional OCR result candidates, the base stroke of confused grapheme can be found using the fitness function which reflects the base stroke characteristics. The possibility of confused stroke existence can be measured by analyzing the boundary area of the base stroke. The result is merged with traditional OCR using score-probability converting. We achieved 68.1% error reduction for target grapheme pair errors by the proposed method and it means that 23.1 % total error is reduced.

1. Introduction

Video OCR is the key technology for multimedia database auto-indexing. We built Hangul video OCR based on recent printed Hangul character recognition researches. We will call it traditional OCR on this paper. Most errors of traditional OCR are caused by similar grapheme pair error[1]. In other words some grapheme pairs are frequently confused with each other. It's hard to recognize them with binarized image even with our own eyes because the stroke information is lost on binarization step.

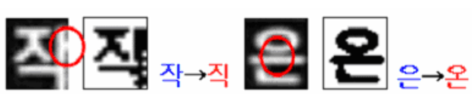

Figure 1. Confusing grapheme error example

So we propose the confused stroke verification method using gray-level image. Section 2 introduces whole system briefly. Section 3 describes traditional OCR and Section 3, 4 explains stroke verification and its merge with traditional OCR. Section 5 represents the recognition results using proposed method and Section 6 makes concluding remarks.

2. System overview

At first traditional OCR results recognition candidates. Based on analyzing those candidates stroke verifier determines which grapheme is confused and measures the possibility of confused stroke existence after finding the base stroke of the grapheme. The scores of traditional OCR and stroke verifier are converted into probabilities and merged on combination step. Figure 2 shows the diagram of the system.

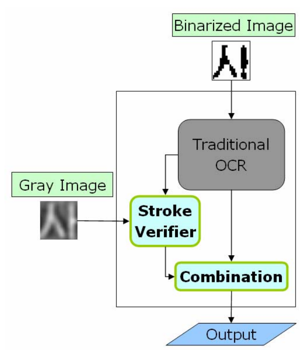

Figure 2. System overview

3. Traditional OCR

To make OCR system with stroke verification, we implemented traditional OCR based on LDA(Linear Discriminant Analysis)[2] and direction angle feature[3]. Number of target class is 327 which contains at least 20 Hangul characters on video text database. Direction angle feature is known as good performed feature on Hangul character recognition.

4. Stoke verifier

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

Stroke verifier determines whether it is used or not based on traditional OCR result. If there seems to be confusing grapheme pair error, it finds the image area to verifty on preparation step. Using edge data from gray-level image it locates the base stroke of the confused grapheme, and measures the possibility of confused stroke existence. Each step will be explained on following sections.

4.1. Preperation for stroke verification

Verification target grapheme pairs are [ㅏ,ㅣ], [ㅓ,ㅣ], [ㅈ,ㅊ], [ㅗ,ㅡ], [ㅜ,ㅡ] which are frequently occurred. Each pair is verified if there exist confused grapheme pair simultaneous on predefined number of recognition candidates. For example if the recognition candidates are [1:가, 2:나, 3:기, 4:키, ..], [ㅏ,ㅣ] grapheme pair is verified because there is [ㅏ,ㅣ] confusing character on 1st and 3nd candidates. And vertical and horizontal edges are extracted from gray-level image using Sobel operator convolution because stroke verification is based on edge information.

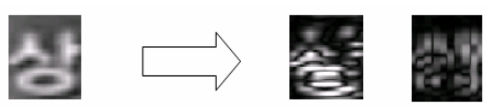

Figure 3. Vertical and horizontal edge from gray level image

4.1. Base stroke extraction

Base stroke is the main stroke of confused grapheme, i.e. [ㅣ] stroke of [ㅏ,ㅣ] pair. Confused stroke is the verification target stroke of the grapheme, which stroke is important to determine the class of the grapheme, i.e. [-] stroke of [ㅏ] on [ㅏ,ㅣ] pair.

Figure 4. Base stroke and confused stroke of [ㅏ]

We defined base stroke fitness function to extract base stroke. It is from energy function of active contour models[4]. It is used to locate object on image by designing the function to have minimum value when locating it successfully. The fitness function of this paper has highest value when most probable stroke is extracted. Base stroke is assumed to be rectangle as seen on figure 4. The fitness function is designed to use the following facts. There is high edge density on the base stroke boundary because there is foreground-background change, and the stroke has proper length and position according to the character class. It is already known that [ㅏ] of [사] is long as height of the image and the base stroke is started on the right top and ended on right bottom of the character.

$$E_{bs} = \log(E_{vb} + E_{hb}) + \alpha_e \log LH_{slen} + \beta_e \log LH_{sloc} \quad (1)$$

E_{vb} , E_{hb} represent vertical and horizontal edge density of the stroke boundary. LH_{slen} , LH_{sloc} are Gaussian weighted function which have value 1 for most probable length and location and have near 0 as they are far from probable one. Each Gaussian weighted function has constant values α_e , β_e for defining the importance of them. These constants are determined experimentally for each grapheme pair and Hangul character type[5]. The detail of fitness function is following.

$$E_{vb} = \frac{1}{(y''-y')}\frac{1}{2w+1}\sum_{i=y'}^{y''}\left(\sum_{j=-w}^{w} D_{ve}(x'+j,i) + \sum_{i=-w}^{w} D_{ve}(x''+j,i)\right)$$
$$E_{hb} = \frac{1}{(x''-x')}\frac{1}{2w+1}\sum_{i=x'}^{x''}\left(\sum_{j=-w}^{w} D_{he}(i,y'+j) + \sum_{i=-w}^{w} D_{he}(i,y''+j)\right)$$
$$LH_{slen} = e^{-\frac{1}{2}\left(\frac{y''-y'-\mu_{len}}{\sigma_{len}}\right)^2}$$
$$LH_{sloc} = e^{-\frac{1}{2}\left(\frac{y'-\mu_{sy}}{\sigma_{sy}}\right)^2} \quad (2)$$

Base stroke : Rectangle with end points $(x', y'), (x'', y'')$
$D_{ve}(x, y)$: Vertical edge density of point (x, y)
$D_{he}(x, y)$: Horizontal edge density of point (x, y)
μ_{len}, σ_{len} : Average and s.t.d value of length of base stroke
μ_{sx}, σ_{sy} : Average and s.t.d value of start point of base stroke
w : Thickness of stroke boundary

Statistical values of length and location of the base stroke are calculated from more than 30 characters for each Hangul character type and grapheme pair.

Using the fitness function, high scored candidate is chosen as base stroke. Base stroke candidates are extracted by 3 steps processing. First step is predicting center of the base stroke using binarized image. Vertical direction projection histogram of binary image and Gaussian weighted function for the center position are used to detect it as seen on left picture of figure 5. Left/right boundaries can be calculated from vertical direction projection histogram of vertical edge. With another Gaussian weighted function for boundary and vertical edge histogram, the left/right maximum values are detected and those points become each side boundary as shown on the right image of figure 5.

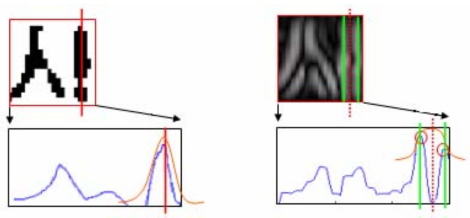

Figure 5. Center of base stroke and left/right boundary

After left/right boundaries are fixed, horizontal direction projection histogram of horizontal edge of that bounded area is seen as figure 6.

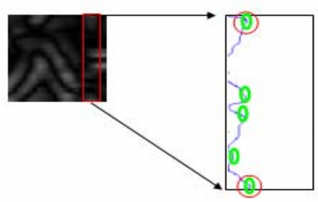

Figure 6. horizontal edge histogram and valley points

Several valley points are detected on the histogram and pair of those points are the candidates of up/down boundaries. So all pairs are calculated through the fitness function and highest scored pair, in other words, the rectangle which boundary is located on high edge area and have proper length/location is selected as base stroke.

4.3. Confused stroke verification

The possibility of confused stroke existence is measured by using two facts in the boundary area of base stroke. First one is that there is edge disappearing phenomenon on boundary area of base stroke if the confused stoke exists. On horizontal direction projection histogram of vertical edge of figure 7, valley point is seen because there is a stroke on that area.

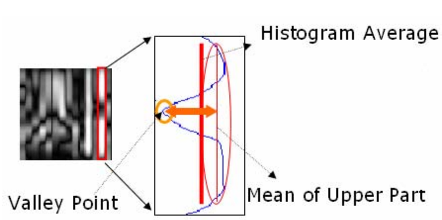

Figure 7. Vertical edge histogram of base stroke boundary

Same phenomenon is happened on horizontal edge histogram. There are ridge points on each side of valley point as seen on figure 8.

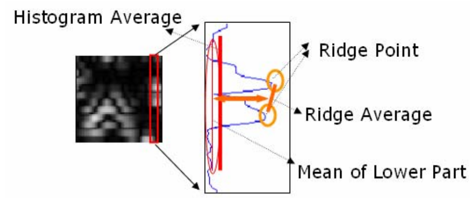

Figure 8. Horizontal edge histogram of base stroke boundary

If the degree of these phenomena are measured, it can be converted into score. In the case of vertical edge of figure 7, mean and s.t.d. values of upper part histogram which is larger than histogram average is defined as μ_h, σ_h. Then the degree of confused stroke existence is defined as following.

$$SD_v = \frac{v - \mu_h}{g_v(\sigma_h)},\ g_v(\sigma_h) = \begin{cases} \alpha_{sd}\mu_h, & if\ \sigma_h < \alpha_{sd}\mu_h \\ \beta_{sd}\mu_h, & if\ \sigma_h > \beta_{sd}\mu_h \\ \sigma_h & ,otherwise \end{cases} \quad (3)$$

On equation (3), v denotes the depth of valley points from the mean of the upper part. The difference between μ_h , v is normalized by σ_h . This normalization is for preventing noisy valley or ridge points from being confused stroke. And there is limitation for minimum and maximum value of normalization parameter. Because if the image is too clean or too noisy, the normalization makes it hard to detect confused stroke. α_{sd} , β_{sd} which are the constant values for these limitations are experimentally determined for each confusing grapheme pair.

In the case of horizontal edge of figure 8, same approach is used. The height of ridge point average is r and the statistics values of lower part which smaller than histogram average is defined as μ_l , σ_l . The degree of confused stroke existence is represented as following.

$$SD_h = \frac{r - \mu_l}{g_h(\sigma_l)}, g_h(\sigma_l) = \begin{cases} \alpha_{sd}\mu_h, & if\ \sigma_l < \alpha_{sd}\mu_h \\ \beta_{sd}\mu_h, & if\ \sigma_l > \beta_{sd}\mu_h \\ \sigma_l & ,otherwise \end{cases} \quad (4)$$

This is also normalized by s.t.d. value of lower part. And the minimum and maximum limitations are same with equation (3). Because those value represents the general tendency of all stroke boundaries of the image.

These two distances are merged with weighted average using following equation.

$$SD = \gamma SD_v + (1-\gamma)SD_h \quad (5)$$

Weight constant value γ is experimentally determined for each confusing grapheme pair and Hangul type. *SD* value is converted into probability and merged into traditional OCR result. And above explanation is for confusing grapheme pair [ㅏ, ㅣ, ㅓ]. For [ㅗ,ㅡ,ㅜ], [ㅈ,ㅊ] pairs it can be handled with same framework if the image is transposed.

5. Stroke verifier merging

Traditional OCR on this paper gives recognition candidates and Mahalanobis distance of each candidate. Stroke verifier gives the degree of stroke existence. Those scores are converted into probability using isotonic regression[6], which is nonparametric converting method for various classifiers. Training data are used for calculating all parameters of isotonic regression. For stroke verifier we separate the training data into two groups which contain confused stroke and doesn't. The regression is used for each confusing grapheme pair and each Hangul type separately.

6. Experimental Result

Hangul video text database is built by collecting news video text of 3 major broadcasting companies in Korea. Totally 327 classes 39,010 samples are used for the experiment. There are more than 327 classes in Hangul characters but the classes which have smaller than 20 characters are removed.

Video text data are splitted into two sets with 2:1 ratio. One set is for training and another set is for testing. Traditional OCR and stroke verifier's parameters are determined by using train data. Before applying stroke verifier, confusing grapheme pair errors are 34.0% of total errors and recognition rate is 95.3%.

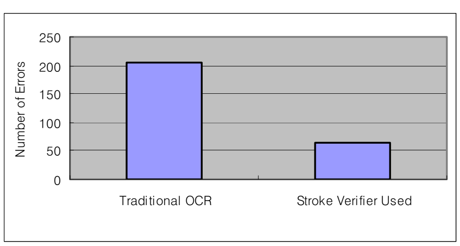

Figure 9. Confusing error reduction by stroke verifier

68.1% of confusing pair errors are reduced by using stroke verifier as seen on figure 9. This means that 23.1% of total errors are reduced because 34.1% of total errors are confusing grapheme pair errors. Recognition rate is improved to 96.4% from 95.3%.

7. Conclusions

We proposed stroke verification method on gray level image. Traditional OCR system uses binarization technique so stroke ambiguity is occurred frequently. The existence of confused stroke can be detected by analyzing traditional OCR candidates. The confused stroke area on gray level character image is found using stroke fitness function and candidate character information. The degree of confused stroke existence is measured by analyzing edge information of that stroke area and the result is merged with traditional OCR result by weighted multiplication of converted probability. We achieved 68.1% error reduction for target confusing grapheme pair errors and this results 23.1% of total error reduction.

8. References

[1] Seung Ick Jang, "Postprocessing of Handwritten Hangul Recognition using Pair-wise Grapheme Discrimination", Master Thesis, Korea Advanced Institute of Science and Technology, 2002, pp.9~14.

[2] Sang-Ho Lee, Ho-Yon Kim, Kil-Tae Lim, Yun-Seok Nam, "Recognition of Handwritten Numerals Based on the Direction Angle Feature", 18th Proc. of Korea Information Science Society, Fall Season, Vol.9, No.2, 2002.

[3] Kyutae Cho, Jinsik Kim, Seonghun Lee, Jahwan Kim, Minsoo Kim, JinHyung Kim, "Recognition for Digitizing Historical Document Pages", Korea Computer Congress of Korea Information Science Society, Vol.32, No.1, 2005, pp. 796~798.

[4] M. Kass, A. Witkin, D. Terzopoulos, "Snake: Active Contour Models", Int J. Computer Vision, Vol4, 1987, pp.321~331

[5] Sung-Bae Cho Jin H. Kim, "Hierarchically Structured Neural Networks for Printed Hangul Chracter Recognition", Proc. Int. Joint Conf. on Neural networks, Vol. 1, San Diego, USA, August 1990, pp. 265-270

[6] Bianca Zadrozny and Charles Elkan, "Transforming Classifier Scores into Accurate Multiclass Probability Estimates", Proc. of the 8th ACM SigKDD Internationcal Conference, 2002

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE