

Typographical Analyses and Classes in Optical Character Recognition

정민철*

Min-chul Jung

Abstract This paper presents a typographical analyses and classes. Typographical analysis is an indispensable tool for machine-printed character recognition in English. This analysis is a preliminary step for character segmentation in OCR. This paper is divided into two parts. In the first part, word typographical classes from words are defined by the word typographical analysis. In the second part, character typographical classes from connected components are defined by the character typographical analysis. The character typographical classes are used in the character segmentation.

1. INTRODUCTION

Character recognition, also known as optical character recognition (OCR), is concerned with the automatic conversion of an image of a character, or of characters in running text, into the corresponding symbolic form, which is then accessible for any information processing system. We recognize words from their shapes the patterns of ascenders and descenders, the counters - Counter is the space enclosed within letters like: a, b, d, e, g and etc. - and spaces in and between letters, rather than identifying each letter in sequence and assembling it into a word [1]. This paper presents a typographical analyses and classes as a preprocessing step of OCR. The typographical analyses and classes are divided into two parts, word typographical classes and character typographical classes. The most important step in the OCR systems is character segmentation. Character segmentation step breaks touching characters into single characters. Most recognition errors result from character segmentation errors. These two types of typographical classes which are proposed in this paper are used to the character segmentation step and the character recognition step to improve their performances.

2. WORD TYPOGRAPHICAL ANALYSIS

The objective of word typographical analysis is to assign each word to a word typographical class.

2.1 Typographical structure of a word

One of the most distinctive and inherent characteristics of European languages is the existence of ascenders and descenders. In Fig. 1, a word image is composed of three typographical zones: the ascender zone, the x-height zone, and the descender zone, which are delimited by four virtual typographical lines, the ascender, the x-height, the base, and the descender lines. While the ascender zone and the descender zone depend on the text content, the x-height zone is always occupied regardless of characters. The distance between the x-height line and the base line is called x-height.

2.2 Projection profile

When image function $I(x, y)$ takes on two values (say, black and white), the projection profile is obtained by counting the black pixels in the vertical (Pv) or horizontal direction (Ph).

■ Vertical Projection Profile: Sum of black pixels



Fig. 1. Typographical structure of a word

*상명대학교 컴퓨터시스템공학과
Department of Computer System Engineering,
Sangmyung University

perpendicular to the x -axis

$$P_v(x) = \sum_y I(x, y) \quad (1)$$

■ Horizontal Projection Profile: Sum of black pixels perpendicular to the y -axis.

$$P_h(y) = \sum_x I(x, y) \quad (2)$$

Fig. 2 shows horizontal and vertical projection profiles of a word. The vertical projection profile is slant invariant but the horizontal projection profile is slant variant [2].

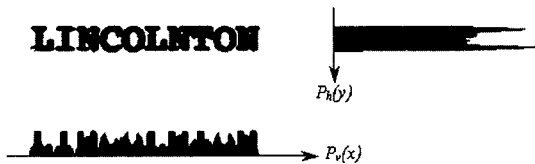


Fig. 2. Horizontal and vertical projection profiles of a word

2.3 Word structure from the horizontal projection profile

Typographical structure of a word is obtained by the horizontal projection profile. As shown in Fig. 3(a), the presence of ascenders in a word (including small over-hanging dots as in 'i' and 'j') makes a small peak on upper end of the horizontal projection profile. Similarly, the presence of descenders in a word makes a small peak on the lower end of the horizontal projection profile. Long black runs generally occur near both the x -height line and the base line, resulting in two significant maxima at these positions. A word consisting of all capitals, all numerals, or all x -height words such as area, come, or essence

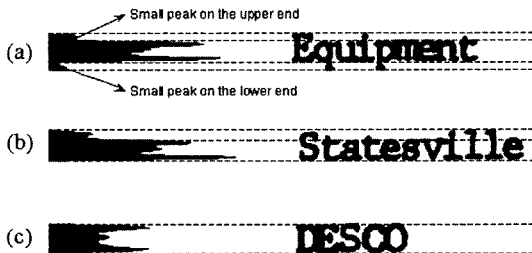


Fig. 3. Word structure from the horizontal projection profile

leads the absence of both small peaks (Fig. 3(c)).

2.4 Typographical line estimation of a word

Typographical line estimation is useful to know a word structure. The typographical lines are simply estimated from the horizontal projection profile as follows:

The slope at the two neighboring pixel points $(P_h(y_i), y_i)$ and $(P_h(y_{i+1}), y_{i+1})$ to the y -axis can be calculated from

$$\theta = \tan^{-1} \frac{\Delta x}{\Delta y} = \tan^{-1} (P_h(y_{i+1}) - P_h(y_i)) \quad (3)$$

$$\frac{\Delta x}{\Delta y} = \frac{P_h(y_{i+1}) - P_h(y_i)}{y_{i+1} - y_i} = P_h(y_{i+1}) - P_h(y_i) \quad (4)$$

Fig. 4 illustrates a horizontal projection profile that is rotated to -90 degrees. In the figure, the slope θ , from the equation (3) is close to $\pi/2$ at the x -height line and to $\pi/2$ at the base line. The start point ($y = y_1$) and the end point ($y = y_i$) of black pixels in the horizontal projection profile are the ascender line and the descender line, respectively in this example. If the slope θ is close to $\pi/2$ at the start point ($y = y_1$), the ascender line does not exist. That implies the word does not have ascenders.

Every word has both the x -height line and the base line. The three zone ratio is used to verify the estimation results, i.e., ascender zone : x -height

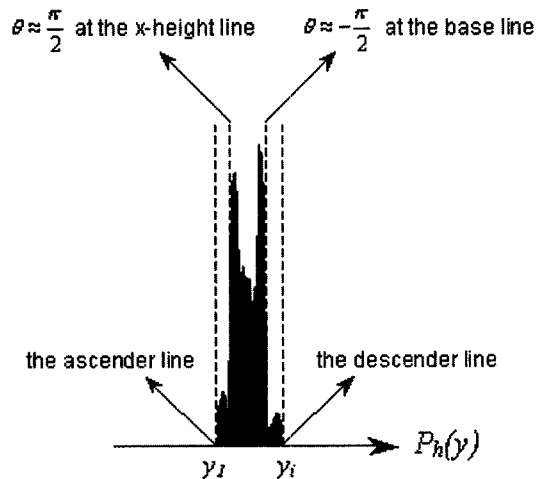


Fig. 4. Typographical line estimation using horizontal projection profile (rotated -90 degrees).

zone : descender zone 1 : 2 : 1.

2.5 Word concatenation

The above typographical line estimation method is accurate for long words, but is failure-prone for very short words. Stable estimates of the typographical lines can be achieved from the horizontal projection profile of an entire text line. However, a longer text is more sensitive to the skew. Therefore, the method usually cannot be used in a whole line without the skew correction. M. Kim [3] stated, "Skew detection and correction needs a lot of processing time and it may distort the character on the document".

It was applied for two-phase estimations of typographical lines to solve the problem of the horizontal projection profile of a short word. The first phase is based on a word. Average English word length is 4-5 characters[4]. In most of the cases, the first phase is enough to estimate typographical lines. If the first phase fails, the second phase is activated. That is, the failed word and its neighboring word are concatenated into a new word of longer length. Fig. 5 illustrates that two short words are concatenated into one long word. This two-phase strategy makes the estimation method more accurate and effective.

2.6 Word typographical classes

Word typographical analysis of the horizontal projection profile of a word assigns it to one of three classes shown in Table 1.

In Table 1, a word in a one-zone class consists of either all capitals or all x-height characters. A word in a one-zone class has neither ascenders nor descenders. That is, without character recognition,

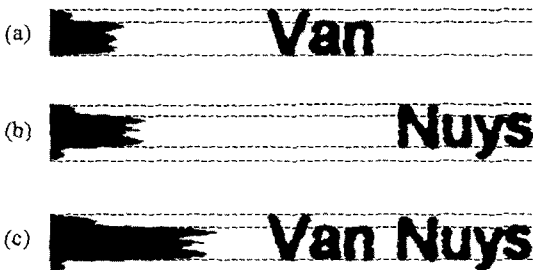


Fig. 5. Word concatenation for a new entity of longer length : Two short words (a) and (b) are concatenated into one long word (c) for more accurate estimation of typographical lines

Table 1. Word Typographical Classes

Word Typographical Classes	Description
One-zone Class	The presence of the x-height zone
Two-zone Class	The presence of two zones Type 1 : the ascender and the x-height zones Type 2 : the x-height and the descender zones
Three-zone Class	The presence of the ascender, the x-height, the descender zones

we can estimate whether a word has ascenders or descenders by the word typographical classes. For example, Fig. 3(a) belongs to a three-zone class, 3(b) to a two-zone class: type1, and 3(c) to a one-zone class.

3. Character typographical analysis

Characters occurring in a text line are generally related to typographical lines by their location. The objective of character typographical analysis is to assign each connected component to a character typographical class.

3.1 Connected components

A line adjacency graph (LAG) is obtained from a run-length representation of a binary image [5] ; its nodes correspond to runs of object pixels and its edges correspond to adjacent runs. 'Blobs' are found as connected components of the LAG. Each component defined as a connected area in a word image, identifies a single character or touching characters except for 'i' and 'j'.

3.2 Character typographical classes

All characters and numerals can be classified into four character typographical classes. Table 2 shows character typographical classes [6, 7].

Character typographical classes are useful to solve the following confusion characters:



Fig. 6. Bounding-Boxed Connected Components

Table 2. Character Typographical Classes

Character Typographical Classes	Description
Ascender Class	All Capitals, Numerals, and b, d, f, h, k, l, t, i
Descender Class	g, p, q, y
x-height Class	a, c, e, m, n, o, r, s, u, v, w, x, z
Full-height Class	j, J and Q (in some fonts)

- 9 (nine)/g (gee), or 8 (eight)/g(gee)
- C/c, O/o, P/p, S/s, U/u, V/v, W/w, X/x, Z/z

The size normalization - a transformation of an input image of arbitrary size into an output image of a fixed pre-specified size, while attempting to preserve structural details - makes it impossible to distinguish between the above pairs [8]. Most of us cannot distinguish between the upper and the lower case version of those characters if the characters are presented in isolation. S. Liang classified only 'j' to a full-height class [6, 7]. However, this should be extended because a few more characters belong to the full-height class according to a font. The classification should be defined by the location of a character, not by the content. In this paper, it has been classified 'J' (in Palatino) and 'Q' (in Bookman, Courier, New Century, Palatino, and Times) to the full-height class. One connected component as a merged character can be classified to a character typographical class. For example, a merged character, "ag" is classified to a descender class, and a merged character, "bg" to a full-height class. This classification is also defined by the location of a connected component, not by the content.

3.3 Character typographical class estimation

When two connected components are neighboring, the character typographical classes of two connected components can be estimated by the relative location of one with respect to the other. Table 3 shows every possible combination of two neighboring connected components.

If it is considered only the scalar of the typographical line displacements of two connected components, those 16 combinations in Table 3 are

Table 3. Character Typographical Class Combinations of Two Connected Components : 'A' stands for an ascender class, 'X' for an x-height class, 'D' for a descender class, and 'F' for a full-height class

AA	AX	AD	AF
XA	XX	XD	XF
DA	DX	DD	DF
FA	FX	FD	FF

simplified to 7 combinations as follows:

- an ascender and an x-height classes (AX/XA)
- an ascender and a descender classes (AD/DX)
- an ascender and a full-height classes (AF/FA)
- an x-height and a descender classes (XD/DX)
- an x-height and a full-height classes (XF/FX)
- a descender and a full-height classes (DF/FD)
- the combinations of the same classes (AA/XX/DD/FF)

Fig. 7 shows two connected components. The first connected component, "Pl" belongs to an ascender class and the second connected component, "easan" to an x-height class.



Fig. 7. Character Typographical Class Estimation

Let the minimum (y_1) be top_w and the maximum (y_4) be $bottom_w$ among four y-coordinates. Then, let y_2 be top_c and let y_3 be $bottom_c$. The difference between $bottom_w$ and top_w is $height_w$, and the difference between $bottom_c$ and top_c is $height_c$. The ratios top_r , $height_r$, and $bottom_r$ are defined as follows :

$$top_r = \frac{top_c - top_w}{height_w} \quad (5)$$

$$height_r = \frac{height_c}{height_w} \quad (6)$$

$$bottom_r = \frac{bottom_w - bottom_c}{height_w} \quad (7)$$

$$top_r + height_r + bottom_r = 1 \quad (8)$$

Equation (8) shows the sum of top_r , $height_r$, and $bottom_r$ is one. Table 4 shows the calculated values

of top_r , $height_r$, and $bottom_r$ for 7 combinations of character typographical classes.

Table 4. Typographical Zone Ratios of Two Connected Components

top_r	$height_r$	$bottom_r$	Class Combinations
0.3	0.7	0.0	AX/DF
0.2	0.6	0.2	AD/XF
0.0	0.7	0.3	AF/XD
0.0	1.0	0.0	AA/XX/DD/FF

Using this table, we can estimate a combination of two character typographical classes. In the example of Fig. 7, $top_r : height_r : bottom_r = 0.3 : 0.7 : 0.0$. From Table 4, we can estimate that the combination is either an ascender class and an x-height class or a descender class and a full-height class. If the word typographical class of the two connected components is a two-zone class, two connected components are composed of an ascender class and an x-height class.

Two connected components that consist of a descender class and a full-height class belong to a three-zone class among word typographical classes.

4. Conclusion

The performance of OCR systems can be improved using the contextual information. Word and Character typographical classes can be used as the contextual information. For example, the character typographical contextual information is used to verify the segmented patterns in character segmentation. If a bounding box of touching characters belongs to an ascender class, the segmented patterns

should belong to either an ascender class or an x-height class. Similarly, if a bounding box of touching characters belongs to an x-height class, the segmented patterns should belong to an x-height class. This information of typographical classes is useful to improve the performance of the character segmentation, the character classification, the post-processing, etc.

5. References

- [1] D. Collier, *Colliers rules for desktop design and typography*, Addison-Wesley, 1991.
- [2] φ. D. Trier and A. K. Jain and T. Taxt, Feature extraction methods for character recognition a survey, *Pattern Recognition*, Vol. 29, No. 4, pp. 641-662, 1996.
- [3] M. K. Kim and Y. B. Kwon, Multi-font and multi-size character recognition based on the sampling and quantization of an unwrapped contour, *International Conference on Pattern Recognition*, pp. 170-174, 1996.
- [4] S. Mori, C. Y. Suen and K. Yamamoto, Historical review of OCR research and development, *Proceedings of the IEEE*, Vol. 80, No. 7, pp. 1029-1058, 1992.
- [5] T. Pavlidis, *Algorithms for Graphics and Image Processing*, Computer Science Press, 1982.
- [6] S. Liang and M. Ahmadi and M. Shridhar, Segmentation of Touching Characters in Printed Document Recognition, *2nd International Conference on Document Analysis and Recognition*, pp. 569-572, 1993.
- [7] S. Liang, M. Shridhar and M. Ahmadi, Segmentation of touching characters in printed document recognition, *Pattern Recognition*, Vol. 27, No. 6, pp. 825-840, 1994.
- [8] J. Wang and J. Jean, Resolving multifont character confusion with neural networks, *Pattern Recognition*, Vol. 26, No. 1, pp. 175-187, 1993.