Open access • Proceedings Article • DOI:10.1109/ICDAR.1999.791797

Multifont classification using typographical attributes — Source link

Min-Chul Jung, Yong-Chul Shin, Sargur N. Srihari

Institutions: State University of New York System

Published on: 20 Sep 1999 - International Conference on Document Analysis and Recognition

Topics: Optical character recognition, Font, Contextual image classification and Character encoding

Multifont Classification using Typographical Attributes

Min-Chul Jung, Yong-Chul Shin and Sargur N. Srihari
Center of Excellence for Document Analysis and Recognition
State University of New York at Buffalo, Buffalo, New York 14260 U.S.A.
{mjung,ycshin,srihari}@cedar.buffalo.edu

Abstract

This paper introduces a multifont classification scheme to help recognition of multifont and multisize characters. It uses typographical attributes such as ascenders, descenders and serifs obtained from a word image. The attributes are used as an input to a neural network classifier to produce the multifont classification results. It can classify 7 commonly used fonts for all point sizes from 7 to 18. The approach developed in this scheme can handle a wide range of image quality even with severely touching characters. The detection of the font can improve the character segmentation as well as the character recognition because the identification of the font provides information on the structure and the typographical design of characters. Therefore, this multifont classification algorithm can be used in maintaining good recognition rates of a machine printed OCR system regardless of fonts and sizes. Experiments have shown font classification accuracies reach high performance levels of about 95 percent even with severely touching characters. The technique developed for the selected 7 fonts in this paper can be applied to any other fonts.

Keywords : *Font Classification, Multifont, Typographical Attributes, OCR.*

1 Introduction

In recognition of machine printed characters on printed documents, a high accuracy can be obtained due to *a priori* regularity within a limited range of fonts known to an OCR system. In practical applications such as IRS tax forms processing system [1], however, the recognition accuracy often drops significantly when a document printed in a different font is encountered because of its irregularities. It is still challenging to develop a recognition system which can maintain the high recognition rate regardless of the irregularities such as the quality of the input documents and the character fonts. As an effort to overcome the difficulties in recognition of characters with a variety of fonts, an omnifont OCR system has been introduced utilizing an abstraction of the font that generalizes the individual differences of various fonts [2].Contrasting to the abstraction of the font, an identification of the font can give details on the structural and the typographical design of characters. Furthermore, with the font information it is possible to make the OCR system handle a document with a confined effort for an identified font. In other words, an OCR system consisting of a various mono-font segmentation tools and recognizers can perform a font-specific processes. The multifont classification system presented in this paper can classify the font from a word image as an input. Font classification provides information such as presence or absence of serif, its shape and the representation of special characters like ‘**a**’ and ‘ɑ’, ‘**g**’ and ‘g’, and the ligatures like ‘fi’ and ‘fi’, ‘fl’ and ‘fl’, etc. The information is useful in the subsequent character segmentation tool and the recognizer.

2 Approach to the Multifont Classification

In this study, 7 fonts have been selected for the 26 uppercases and the 26 lowercases of the English alphabet. The fonts are Avant Garde, Bookman, Courier, Helvetica, New Century Schoolbook, Palatino and Times. Figure 1 shows some of the 364 target characters.

	Characters with Ascenders							Characters with Descenders				Characters with x heights only					
Avant Garde	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x
Bookman	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x
Courier	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x
Helvetica	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x
New Century Schlbk	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x
Palatino	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x
Times	b	d	f	k	h	l	t	g	p	q	y	a	m	r	u	w	x

Figure 1: The sample of 7 fonts

Many gross distinctions among those typefaces are easily made by an untrained observer. e.g. serif vs. sans-serif faces and variable pitch designs vs. fixed pitch ones. Bookman, New Century Schoolbook, Palatino and Times are serifed fonts that have the small decorative strokes at the ends of main vertical strokes. Courier, the original typewriter font has a heavy square serif (i.e. Slab Serif) and is printed at a fixed pitch regardless of its actual width. Avant Garde and Helvetica are sans-serif fonts – ‘sans’ means without – that have no serifs and no contrast between thick and thin strokes, i.e. fixed stroke width. Figure 2 shows the location of a multifont classifier in an OCR system. The multifont classifier is located between the word segmentation and the character segmentation. i.e. The multifont classifier is independent of the character segmentation and the character recognition. This approach is *a priori* font classification that identifies the font without any knowledge of the characters, compared to *a posteriori* font classification approach that classifies the font using the knowledge of the characters. The multifont classifier were tested on the name and address block on tax forms of the U.S. Internal Revenue Service [1] for learning and classification. In such documents, only a few known fonts are used and only one font is usually used in one document. Besides, the slanted style such as oblique, italic is rarely used. Especially at person name line on the tax form, the lexicon based postprocessing is hard to realize because the number of last name in the U.S. is 1.7 millions aside from first name. The long name in the constrained form space causes tight character spacing and when scanned even at a moderate threshold, most characters can be joined (See Figure 4). More than half recognition errors come from the touching characters. The segmentation method for a fixed pitch font and a variable pitch one should be different. If the word in Figure 4 is known to be in Courier font, meaning a fixed pitch font, it is easy to segment such severe touching characters as the characters are printed at a fixed pitch.

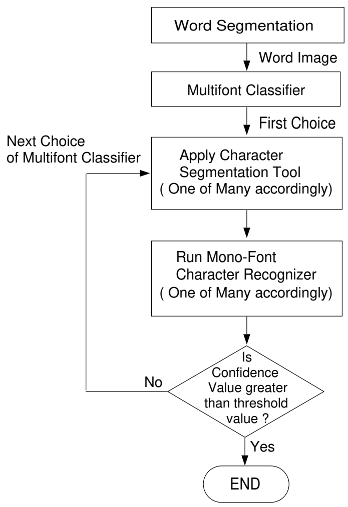

Figure 2: *A priori* font classification

3 Methodology

3.1 Typographical Zones

As shown in Figure 4, text line images are composed of three typographical zones: the ascender, the x height and the descender zones, which are delimited by four virtual horizontal lines, ascender, x height, base and descender lines [3]. The lowercases with ascenders that extend above the x height line are b, d, f, h, k, l, i, and t. They occupy the ascender and the x height zones. The lowercases with descenders that extend below the base line are g, p, q, and y. They occupy the descender and the x height zones. The centered lowercases having neither ascenders nor descenders are a, c, e, m, n, o, r, s, u, v, w, x, and z. They occupy the x height zone only. The lowercase j that has both an ascender and a descender occupies all three zones. All uppercase characters and numerals take up the ascender and the x height zones. From the vertical projection profile of a word image after the correction of skew angle, the typographical structure can be estimated. The analysis of the vertical projection profile shows that a word image has one of four types [4],

- Type I: The presence of all three zones,
- Type II: The presence of the ascender and the x height zones,
- Type III: The presence of the x height and the descender zones,
- Type IV: The presence of the x height zone only.

In the case of Type I, the word must have both lowercases with descenders and lowercases with ascenders. In the case of Type II, the word must be either in lowercases with ascenders or in all uppercases. The word in this case can not have a lowercase with a descender. Since the shapes of the vertical projection profile in the case of 'lowercases with ascenders' and in the case of 'all uppercases' are totally different, it is able to know which case it is. Likewise, Type III has lowercases with descenders and Type IV has lowercases without ascenders and descenders. The cases of Type III and Type IV are rarely happened in the name and address block as the proper nouns are capitalized such as a person name, a street name and a city name. In this paper the cases of Type I and Type II are considered.

3.2 Font Size

If the scanning resolution is known, the font size can be extracted from the vertical pixel count, i.e. the line number of the vertical projection profile. The formula is

$$\frac{\Sigma VerticalPixelNumber}{ScanningResolution} \times 72.27 = FontSize$$

because one inch is equal to 72.27 points [5] and the scanning resolution is measured by dpi (dot per inch). 'Σ Vertical Pixel Number' is the sum of three zone heights. By the vertical pixel number of the x height zone alone, the font size can be obtained using Table 1.

Table 1. New Century Schoolbook Font Size vs. the Line Number of Each Zone when scanned at 200 dpi

Font Size	7	9	10	12	14	18
Ascender	5	7	8	9	10	14
x Height	10	12	13	17	19	24
Descender	4	6	6	7	8	10
Total	19	25	27	33	37	48
Calculated	19.4	24.9	27.6	33.2	38.7	49.8

Table 1 has been obtained by counting the vertical pixel number at each zone of the vertical projection profile using the scanned images printed in New Century Schoolbook font at 200 dpi. The total values are the sum of three zone heights and the calculated values come from $\frac{FontSize}{72.27} \times 200$. It is observed that the total number and the calculated values are very close to each other. The vertical pixel number at each zone is roughly the same at one font size regardless of the fonts. Courier font is about 12 percent smaller than the calculated value. It is the smallest among 7 fonts in the same point size.

3.3 Analysis of the Typographical Attributes

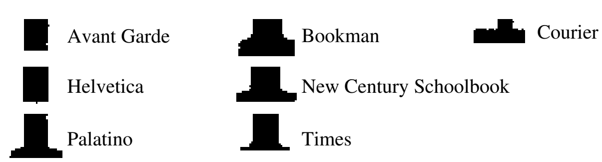

Figure 3: The various serifs of 7 fonts in the descender zone

Serifs are the most obvious features dividing typefaces. They have various shapes that may be named as Hairline serif, Wedge serif, Slab serif and Bracketed serif [5]. However, serifs have the same type of shape in the same font. Figure 3 shows the various serifs of 7 fonts in the descender zone. It is noted that *the serif heights in the descender zone* as well as its shapes are different. The serif height in the descender zone of Courier font is the shortest. The lowercase characters may have ascenders, descenders as well as serifs while the uppercase characters may have serifs only. The details of the analysis are presented as follows. It has been implemented in a computer program that handles the cases of 'Capitalized Word' and 'All Uppercases Word'.

Case 1 : Capitalized Word

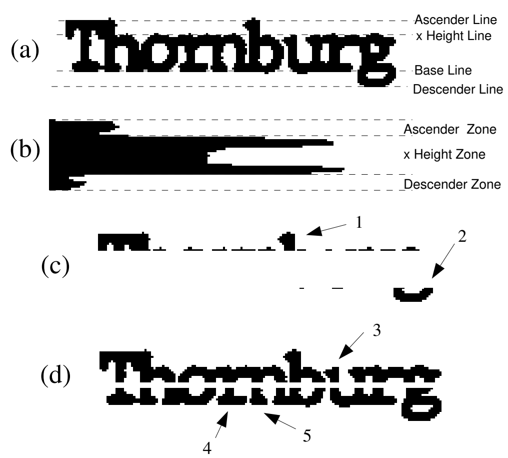

Figure 4: (a) Capitalized word (b) The vertical projection profile of (a). (c) The residual component blobs after deleting the x height zone. (d) The residual component blobs after deleting the middle zone of the x height zone.

The capitalized words can be in the Type I or the Type II. Since each zone of a word is known by the analysis of the vertical projection profile, the ascenders and the descenders are easily extracted from the ascender and the descender zones (Figure 4b and 4c). The characters b, d, h, k, and l(ell) have the same shape of the ascenders. The characters p and q have the same shape of the descenders. Those ascenders and descenders may have serifs. The characters 't, f' and 'g, j, y' have unique ascenders and descenders according to the font, respectively. The ascenders of characters 'i' and 'j' are dots. The dot is discarded from our discussion as it is not contributing to the font classification. In an uppercase character at the first (e.g T in Figure 4a), the part in the ascender zone is discarded because of its wide width. All ascenders' widths are narrow enough to compare with the character size. Therefore, there are 3 ascenders ('bdhkl', 'f' and 't') and 4 descenders ('g', 'j', 'pq' and 'y') per font. After checking all the ascenders and the descenders in each zone, the serifs in the x height zone are examined. The x height zone is divided into 3 zones, the upper, the middle and the lower zones similarly and checked for both the presence of serif and its shape at the upper and the lower zones. The height ratio of the upper, the middle and the lower zones is 2 : 1 : 2. There are some residual component blobs in the upper and the lower zones after deleting the middle zone of the x height (Figure 4d). The algorithm checks the heights and the widths of the blobs and discards the blobs that are wider or higher than the threshold values. The threshold values are determined as follows. Since it is trying to check the serif at the top and the bottom parts of the vertical strokes, the width of the blobs should be smaller than a half of the character width. The character width and height are roughly estimated from the typographical zones. The wide ones that may come from the touching characters or none-vertical strokes of a character are discarded. Therefore, the algorithm has the immunity against the touching characters. However, a severe touching can affect the result of this selection algorithm. The height of the blob should also be less than a half of the x height because the ones taller than a half of the x height have already been checked as the ascenders or the descenders by the previous procedure. In the example of Figure 4a, 5 components (numbered 1 to 5 in the Figure) are finally selected (Figure 4c and 4d).

Case 2 : All Uppercases Word

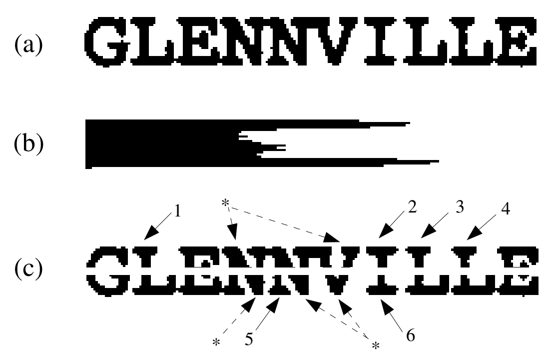

Figure 5: (a) All uppercases word. (b) The vertical projection profile of (a). (c) The residual component blobs after deleting the middle zone of the capital letter.

As mentioned earlier, it is able to know whether a word consists of all uppercases by the shape of the vertical projection profile (Figure 5b). The word in Figure 5 has touching characters. All uppercases words need to be divided into 3 horizontal zones to extract serifs. Likewise the previous method to check the x height zone, it checks some residual component blobs in the upper and the lower zones after deleting the middle zone of the word (Figure 5c). The algorithm checks only the widths of the blobs at this time and discards the blobs that are wider than a half of the character width. Only the vertical stokes that may have serifs are taken and the blobs with a slant side ($\star$ marked in Figure 5c) are discarded. The slope detection algorithm [3] is used to detect the blobs with a slant side. It is observed from the horizontal projection profile of a candidate blob that the vertical stroke is characterized by an upright stem while the one with a slant side is characterized by a rounded peak. In the example of Figure 5a, 6 components (numbered 1 to 6 in Figure 5c) are finally selected.

3.4 Various Serif Types

The selected serifs may have slightly different shapes within the same font, although they are still considered as a single type of serifs such as Slab serif or Hairline serif. Same type of serifs at the top of a vertical stroke (e.g. blob 2 in Figure 5c) and the ones at the bottom (e.g. blob 6 in Figure 5c) are considered differently. For some fonts such as Bookman, New Century Schoolbook, Palatino and Times, the vertical stroke thickness can be either thick or thin. For other fonts such as Courier, the serif can be either narrow or wide (e.g. blobs 5 and 6 in Figure 5c). The serif numbered as 3 in Figure 4d is yet another type of serif. As a summary, Courier font has 7 different serifs while Avant Garde and Helvetica fonts have no serifs and the rest of them have 5 different serifs.

4 Feature Extraction and Neural Network Implementation

A directional run-length based feature extractor has been developed for the application of the font classification. The features are calculated from the selected blobs and are used as the input to a neural network classifier. The images are normalized to 9 rows and 9 columns. A 3×3 grid is used to partition the image into 9 equal-sized regions. The number of pixels with a particular slope, 0, 45, 90 or 135 degree in each region is determined. Since there are 4 values for the slope in each region, there is a set of 36 features that constitutes the input vector to the neural network classifier.

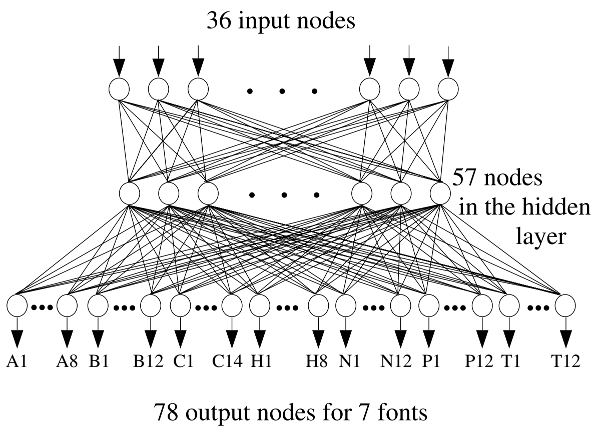

Figure 6: The Neural Network Configuration

The classifier used in the implementation is a 2 layer back propagation network with 36 input nodes, 57 nodes in the hidden layer and 78 output nodes. The number of classes in the output is different according to fonts. For example, Courier font has 3 ascenders, 4 descenders and 7 different serifs. The classes are named as C1, C2, ..., C14. Figure 6 shows the neural network configuration. 'A' stands for 'Avant Garde', 'B' stands for 'Bookman' and so on.

5 Experiments and Results

This section presents the results of classification experiments and discusses the strength and weakness of the approach presented in this paper. Word images have been arbitrarily taken from real documents scanned at 200 dpi. Table 2 lists the average classification rates of 7 fonts with 6 point sizes from 7 to 18, respectively. The classification has been applied to 2520 word images that are different from training images. The classification results show a high performance with an overall accuracy of 95 percent. For each font and size, 60 word images were used.

Table 2. Average classification rates of 7 fonts with 6 point sizes from 7 to 18.

		7	9	10	12	14	18
Serifed Font	Bookman	0.851	0.902	0.912	0.954	0.978	0.998
	New Century Schoolbook	0.848	0.917	0.935	0.968	0.976	0.997
	Palatino	0.859	0.899	0.911	0.939	0.964	0.997
	Times	0.852	0.917	0.933	0.960	0.971	0.998
Sans-Serifed Font	Avant Garde	0.927	0.958	0.966	0.973	0.985	0.998
	Helvetica	0.933	0.954	0.961	0.969	0.978	0.998
Typewriter	Courier	0.916	0.962	0.966	0.974	0.979	0.998

At point sizes 7 and 9, the classification rates are low because typographical attributes, especially serifs are degraded from noises. The small serifs are significantly influenced by noises. Serifed fonts do worse than sans-serifed fonts because serifed fonts have more complex shapes of serifs. It is observed that the classification rates are improved with point sizes increasing.

6 Summary

This paper has proposed a font classification method based on the definition of typographical attributes such as ascenders, descenders and serifs, and the use of a neural network classifier. It can classify with a word image for 7 commonly used fonts with various point sizes from 7 to 18. The experiments have been tested on real images. The results demonstrated the validity of the algorithm and showed a satisfying performance level. The technique developed for the 7 selected fonts in this paper can be applied to any other fonts.

References

[1] Sargur N. Srihari, Yong-Chul Shin and *et al.*, "A System to Read Names and Addresses on Tax Forms," *Proceedings of the IEEE*, vol. 84, no. 7, pp. 1038-1049, Jul 1996.

[2] S. Kahan, T. Pavlidis and H. S. Baird, "On the Recognition of Printed Characters of Any Font and Size," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 2, pp. 274-288, 1987.

[3] Abdelwahab Zramdini and Rolf Ingold, "Optical Font Recognition Using Typographical Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 877-882, Aug 1998.

[4] Abdelwahab Zramdini, *Study of Optical Font Recognition based on Global Typographical Features, Ph. D. Dissertation, University of Fribourg, Switzerland*, pp. 68-69, 1995.

[5] Richard Rubinstein, *Digital Typography: An Introduction to Type and Composition for Computer System Design*, *Addison-Wesley*, 1988.