Optical Font Recognition Using Typographical Features

Abdelwahab Zramdini and Rolf Ingold

Abstract—A new statistical approach based on global typographical features is proposed to the widely neglected problem of font recognition. It aims at the identification of the typeface, weight, slope and size of the text from an image block without any knowledge of the content of that text. The recognition is based on a multivariate Bayesian classifier and operates on a given set of known fonts. The effectiveness of the adopted approach has been experimented on a set of 280 fonts. Font recognition accuracies of about 97 percent were reached on high-quality images. In addition, rates higher than 99.9 percent were obtained for weight and slope detection. Experiments have also shown the system robustness to document language and text content and its sensitivity to text length.

Index Terms—Optical font recognition, typographical features, font models, multivariate Bayesian classifier, document analysis, OCR.

1 Introduction

A considerable amount of research has been dedicated to optical character recognition (OCR) of printed texts. Early OCR systems, called monofont reading systems were able to read a single font, sometimes even specific fonts that were designed for optical reading purposes (OCR-A and OCR-B fonts). The tendency of recent developments was oriented toward omnifont recognition methods, which aim at recognizing characters of any font and style [1], [2]. Some of the currently available OCR products are only able to distinguish two or three font styles such as italic, bold, seriffed, or sanserif. Results of such tools are, however, still not very accurate. To our knowledge, there has been no serious study of the optical font recognition (OFR) problem, which can be addressed through two complementary approaches [3]: the a priori approach, in which characters of the analyzed text are not yet known, and the a posteriori approach, where the content of the given text is used to recognize the font.

Only a few works have addressed the automatic typeface recognition, with focus on the identification of some font attributes, such as slope and weight for OCR, document analysis, and image editing purposes [4], [5], [6], [7], [8]. Morris has examined the applicability of human vision models to typeface discrimination; he used Fourier amplitude spectra of images to extract global feature vectors used by a Bayesian classifier [9]. Khoubyari and Hull presented an algorithm that identifies the predominant font in a document [10].

In this paper, we present a novel contribution to the a priori OFR approach. Our goal is to discriminate the font from a given piece of text among a set of several hundred known fonts constituting the so-called font model base. In our system called *ApOFIS* (A priori Optical Font Identification System), global typographical features are extracted from the text image and used by a multivariate Bayesian classifier.

The rest of this paper presents our approach to OFR. In Section 2, features used by the classifier are briefly presented and their power to font discrimination is highlighted. In Section 3, experimental results are discussed. They show the relevance of our approach on printed and then scanned documents. Appendices A and B present formally a classification of connected components and the used features.

- *R. Ingold is with the Institute of Informatics, University of Fribourg, Ch-1700 Fribourg, Switzerland. E-mail: Rolf.Ingold@unifr.ch.*
- *A. Zramdini is with A2i SA, Oron la Ville, Switzerland. E-mail: Abdelwahab.Zramdini@a2i.ch.*

Manuscript received 21 Oct. 1996; revised June 1998. Recommended for acceptance by M. Mohiuddin.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107078.

0162-8828/98/$10.00 © 1998 IEEE

2 The *ApOFIS* Approach to Optical Font Recognition

In this section, typographical notions and the features used in *ApOFIS* are presented.

2.1 Typographical Study and the *ApOFIS* Approach

Features used to model fonts have been derived from global typographical properties of text lines. Subsequently, these properties are presented in order to justify the features used by *ApOFIS*. The notion of font is also specified.

2.1.1 Font Specification and Identification Attributes

Typographically, a font is a particular instantiation of a typeface design, often in a particular size, weight and style [11]. Typefaces are distinguished by their writing style (cursive, typesetter), shape of serifs, x-height proportion, character spacing (fixed, proportional, with/without kerning), and loop axes, etc. [12], [11]. Within *ApOFIS*, a font is fully specified by five attributes: *typeface* (Times, Courier, Helvetica, ...), *weight* (light, regular, demi, bold, heavy), *slope* (roman, italic), *width* (normal, expanded, condensed), and *size*.

2.1.2 Typographical Structure of Text Lines

As shown in Fig. 1, text line images are composed of three typographical zones: the upper, central, and lower zones, which are delimited by four virtual horizontal lines. While the upper and lower zones depend on the text content, the central zone is always occupied regardless of the characters that occur. The height of the central zone is commonly called x-height, and its proportion in the text height differs from one typeface to another.

Within printed Latin texts, characters are separated by two kinds of spaces: character- and word-spaces. The former is an intrinsic aspect of typeface design. We assume that their values depend exclusively on the typeface nature (with proportional/fixed spacing) and character sequences, where negative spaces may occur with italic style or in case of kerning.[1] Character-spaces have, therefore, to be preserved in typeface discriminations. The latter depend exclusively on formatting parameters, such as margins and justification mode. They must be ignored during feature extraction, since they do not provide any relevant information on the font.

2.1.3 The ApOFIS Approach

ApOFIS aims at font identification from images of text lines, which are assumed to be homogeneously typeset, i.e., with the same font. Practically, *ApOFIS* works as a multivariate Bayesian classifier based on feature distributions that are estimated independently of the content, structure, and language of texts, but taking into account the influence of the text length.

2.1 Feature Extraction

The *ApOFIS* prototype uses eight global features extracted from connected components and from horizontal and vertical projection profiles of text lines. Features have been carefully selected in order to discriminate fonts regardless of the text content and structure.

1. Character spacing within words can be adjusted for improved appearance or to fit text to a specific width.

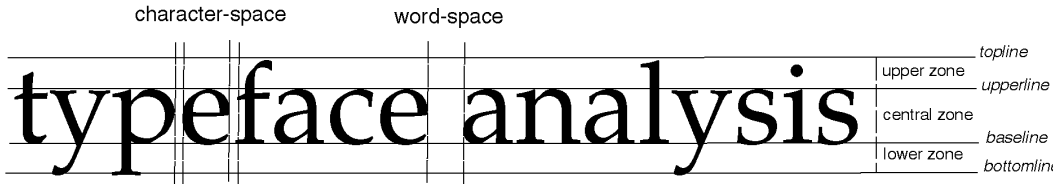

Fig. 1. Typographical structure of text lines.

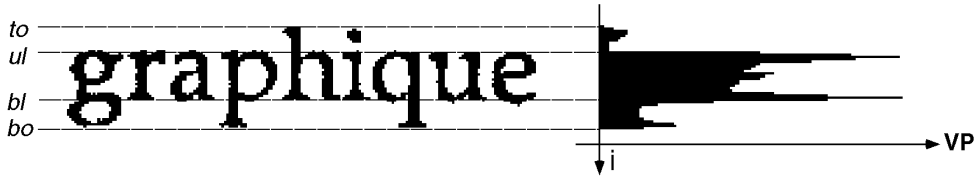

Fig. 2. Four typographical lines from vertical projection profiles.

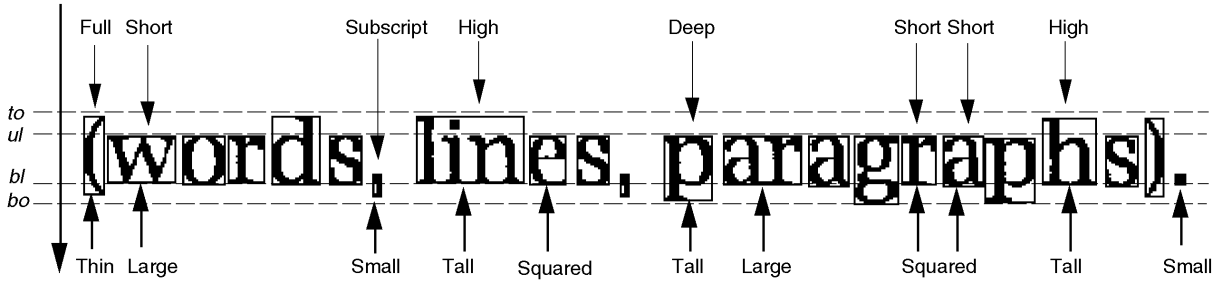

Fig. 3. Typographical and morphological classifications of connected components.

Most of the feature vectors have been shown to follow normal laws [3] $N(u, \Sigma)$, so that learning consists of the evaluation of the multivariate normal density parameters μ and Σ.

The feature extraction process, which assumes skew-free images, performs three steps:

1) determination of the typographical structure of the text line;
2) classification of connected components;
3) calculation of features using the typographical structure of the text and the classified connected components.

2.2.1 Determination of the Typographical Structure

The typographical structure of text lines is used to classify connected components and to delimit the features extraction area. It is determined from the vertical projection profile, VP, as shown in Fig. 2. Each component VP[i] represents the sum of black pixels of the scanline i. The *ul* and *bl* scanlines, which estimate the *upperline* and the *baseline*, correspond to the main peaks of VP, such that:

$$ul = i \text{ if } i \in \left[to, to + \tfrac{1}{2}|bo - to|\right] \,\&\, \max\left(\mathrm{VP}[i+1] - \mathrm{VP}[i]\right);$$

$$bl = i \text{ if } i \in \left]to + \tfrac{1}{2}|bo - to|, bo\right] \,\&\, \max\left(\mathrm{VP}[i-1] - \mathrm{VP}[i]\right).$$

2.2.2 Classification of Connected Components

Within images, characters can be located by the rectangular envelopes of their connected components, which may, however, correspond to linked or broken characters. In order to extract some of the features, typographical and morphological classifications of connected components are performed.

As illustrated in Fig. 3, we use the positions of connected components within the typographical zones of the text line to distinguish between six typographical classes (Full, High, Deep, Short, SuperSc., SubSc.). Similarly, we distinguish between six morphological classes (Wide, Large, Squared, Tall, Thin, Small) using the dimensions of connected components, especially the width to height ratio (see Appendix A).

The assessment of the proposed typographical and morphological classifications, has not been performed through exhaustive tests, but on the basis of empirical measurements. In contrast with OCR applications based on connected components, which assume a nearly 100 percent accurate component preclassification [13], [14], *ApOFIS* can tolerate some misclassifications, since it assumes that features are computed from relatively long strings.

2.2.3 Features and Their Contribution to Font Discrimination

The features selection consists of searching for global aspects of the text, which allow the text reader to distinguish visually between font weights, slopes, sizes, and typefaces. In this section, we only give an intuitive description of the selected features, with a special focus on their power to discriminate font attributes. A detailed and formal description of each feature is given in Appendix B.

Weight and Slope Detection

As shown in Fig. 4 , the font weight is reflected by the density of black pixels in the text line image. One can easily notice that bold texts have a higher density than regular ones. Therefore, the density is taken as weight discrimination feature, at least when the typeface is known.

One can also observe from the horizontal profile that roman texts are characterized by a set of upright and tall peaks. For italic texts, the peaks are less tall, rounded, and boarder. Taking the squared values of the profile derivative has been shown to be relevant for slope discrimination.

Furthermore, vertical stems width and horizontal stems height within characters rely on the typeface design and on the font weight and size. The estimation of the stem's width and height allows us to distinguish not only between regular / bold, but also between roman / italic for the same typeface.

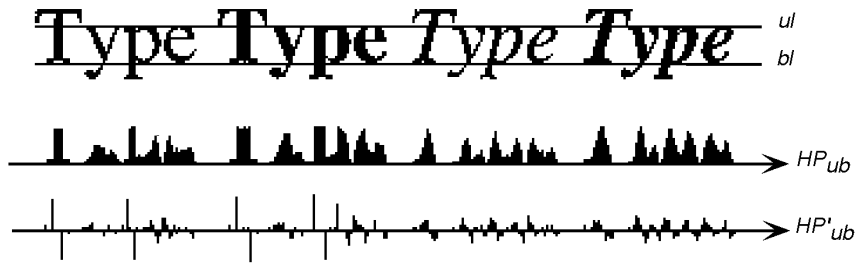

Fig. 4. Effects of font style on the horizontal projection profile and its first derivative.

Fig. 5. Difference between scanlines for seriffed and sanserif texts computed from connected components extremities.

Typeface Detection

Serifs are the most obvious features that distinguish seriffed from sanserif typefaces. They are mainly located at the end of the character main strokes. As shown by Fig. 5, serifs are exhibited by computing the difference between consecutive scanlines, especially from regions around the top and bottom of connected components. The density of the resulting image has shown its relevance in seriffed / sanserif discrimination.

On the other hand, the intercharacter spacing mode (fixed, proportional) is a fundamental aspect of typeface design. The character spacing mode changes from one typeface to another. It is revealed by the average pixel distances between the rectangular envelopes of connected components within words. Character spacing is also significantly influenced by the text slope, with even negative values for italics.

Size Detection

The text size is obviously revealed by character heights and widths. The text height is globally characterized by the x-height, X-height, and the total text height. These measures depend, however, on the text content and structure. A more sophisticated measure was defined and based on a normalized height of the connected components. It uses the typographical class of each component defined in Section 2.1.2.

Similarly, the average width of connected components that are squared and located in the *central* zone, e.g., those corresponding to characters a, c, and u, estimates the character width.

3 Experiments

This section presents the results of various classification experiments and discusses the strengths and weaknesses of the *ApOFIS* approach to font recognition. The classifier uses a font model base (FMB), which presently includes 280 font models representing 10 typefaces combined with seven sizes (*8, 9, 10, 11, 12, 14, 16pt*) and four styles (*regular, bold, italic, bold-italic*).[2] Each font model has been estimated from feature vectors of about 100 text lines of about six cm length each. Texts were arbitrarily taken from English documents, produced by a 300-dpi laser printer and scanned again at 300 dpi.

The test set contained at least 100 French text lines for each font. Images were produced under the same conditions as for learning.

3.1 Classification Results

Table 1 lists the average recognition rates of fonts and font attributes for each typeface. The classification was performed among the 280 font models of the FMB.

For each font, the measured recognition accuracy is expressed as an average rate of the processed lines. The typeface recognition accuracy corresponds to the average of its 28 font rates and is therefore estimated from about 3,000 text lines. We distinguish two kinds of accuracies; one is obtained for fonts considered as a "whole," where each attribute misclassification leads to a recognition error, the other concerns the individual font attributes, i.e. the family, weight, slope, and size.

3.1.1 Attributes Discrimination

When focusing on the "font" recognition, the results show that the classifier achieved a good performance with an overall recognition rate near 97 percent. The hardest fonts to recognize were the seriffed one's (95.6 percent) since they are characterized by very complex character shapes.

Typewriter fonts, for which we obtained an accuracy of 98.2 percent, appear to be the easiest to recognize, probably because of their fixed-pitch aspect. A few recognition failures were registered for the Lucida-Sans and Times fonts.

The classification results show also that the classifier is very robust in slope detection with an overall accuracy of 99.97 percent. It has, however, demonstrated a few failures, especially for some Courier fonts. This can be explained by the small x-height of that typeface, which is reflected by very short vertical stems. The lowest slope recognition rate remains higher than 97 percent.

The identification of the font weight is inherently more complex since typefaces may exist in many weights. The considered typefaces were either *light* or *demi*, or *regular* and *bold* (however, four weights have been used with the Helvetica typeface: *regular, bold, black,* and *heavy*[3]). The system succeeded in weight discrimination regardless of the large size of the FMB, with a lowest rate of

2. This means that 28 fonts have been considered for each typeface.

3. In fact, we considered the Helvetica-Black typeface as a variant of Helvetica.

TABLE 1
AVERAGE RECOGNITION RATES OF FONTS AND FONT ATTRIBUTES WHEN THE CLASSIFICATION IS PERFORMED AMONG THE 280 FONTS

		Font	Typeface	Size	Weight	Slope
seriffed typefaces						
Palatino	PL	97.46	97.82	98.25	99.78	99.86
New-Century-Schlbk	NC	96.92	97.92	97.67	99.63	99.92
Lucida-Bright	LB	95.27	95.81	95.30	99.94	100
Times	TM	92.87	93.72	93.59	99.88	99.94
sanserif typefaces						
Helvetica-Black	HB	99.84	99.99	99.88	99.95	100
Avant-Garde	AG	99.58	99.77	99.66	99.85	99.95
Helvetica	HV	99.44	99.68	99.61	99.97	100
Lucida-Sans	LS	91.26	91.57	99.04	99.95	100
typewriter typefaces						
Courier	CR	99.30	99.97	99.35	100	99.95
Lucida-Sans-Typewriter	LT	97.16	97.20	99.88	100	100
average		96.91	97.35	98.22	99.90	99.97

TABLE 2
TYPEFACE CONFUSION MATRIX WHEN THE CLASSIFICATION IS PERFORMED AMONG THE 280 FONTS

	seriffed				sanserif				typewriter		FER
	LB	NC	PL	TM	AG	HV	HB	LS	LT	CR	
LB	**95.81**	1.86	0.91	1.36	0	0	0	0	0	0	4.19
NC	0.67	**97.92**	0.89	0	0	0	0	0	0	0	2.08
PL	1.09	0.81	**97.82**	0	0	0	0	0	0	0	2.18
TM	2.31	2.95	0.90	**93.72**	0	0	0	0	0	0	6.28
AG	0	0	0	0	**99.77**	0	0		0	0	0.23
HV	0	0	0	0	0	**99.68**	0	0	0	0	0.32
HB	0	0	0	0	0	0	**99.99**	0	0	0	0
LS	0		0	0	0	0	0	**91.57**	8.12	0	8.43
LT	0	0	0	0	0	0	0	2.61	**97.20**	0	2.80
CR	0	0	0	0	0	0	0	0	0	**99.97**	0
ERTCF	4.07	5.62	2.70	1.97	0	0	0	2.61	8.12	0	2.65

97 percent observed for one Lucida-Sans font. In spite of the high accuracy observed for Helvetica, the weight-detection problem remains open when many weights for the same typeface are considered. We can, however, claim that if typeface is correctly identified, then its weight is also. Furthermore, the weight has to be considered as a relative measurement, since the weight of a given bold typeface may be lighter than a regular font of another typeface.

In spite of the presence of small and consecutive sizes, the system shows a good size discrimination power with an average accuracy rate of 98.22 percent. Similarly, except for Lucida-Sans and Times, typefaces are relatively well discriminated, with an accuracy rate of 97.35 percent. In fact, the identification of typeface and size is very tedious since

1) typeface discrimination, especially from short texts, is a domain reserved to skilled typographers and
2) even for a known typeface, one point difference in size is hardly noticeable even by well advised readers.

In addition, recognition rates have shown that size and typeface misclassifications were often tightly coupled. This suggests that the a prior knowledge of the typeface will certainly enhance size recognition and vice-versa. Classification results have shown that the size knowledge has improved the typeface discrimination, with rates increasing from 93.72 percent to 98.77 percent for Times.

3.1.2 Typeface Confusion

Table 2 shows the typeface confusion matrix, which is read as follows. Each [f_i, f_j] entry gives the percentage of effective fonts f_i, which were classified as the top-choice fonts f_j. In the last column, the misclassification rates (FER) are given. In the last row, the error rates for top-choice fonts (ERTCF) are listed. The table indicates the "noisy" typefaces, which affect recognition rates of the other typefaces, e.g. Lucida-Bright and New-Century. Finally, the overall error rate is given in the bottom-right entry of the matrix, which indicates that 2.65 percent of the 30,000 text lines were misclassified.

The matrix shows the power of *ApOFIS* in discriminating between the *seriffed, sanserif,* and *typewriter* typefaces, where misclassifications mainly occur within the same family. A discrimination rate of 99.65 percent was obtained between seriffed and sanserif. The matrix shows, however, that the system failed in discriminating between Lucida-Sans (LS) and Lucida-Sans-Typewriter (LST). The same behavior that originates from the fact that LST is a fixed-width typeface highly stylized to look like LS was also observed in Morris's experiments [9].

3.2 Effects of Text Length

The discussion so far has not considered text length. In the previous experiments, all text entities, used to create the FMB and to assess the classifier, have similar lengths. The study on the influence of text length is of importance because document analysis may require font identification from fragments of various lengths, e.g., words, lines, or even paragraphs.

In the following experiment, the classifier was applied on texts of four different lengths using the original FMB (generated from text lines of a default length L). The recognized lines were either

TABLE 3
EVOLUTION OF RECOGNITION RATES FOR FOUR TEXT LENGTHS USING THE DEFAULT FMB

	$\frac{1}{4}$L	$\frac{1}{2}$L	L	2L	4L
Weight	95.54	98.41	99.90	100	100
Slope	96.16	99.0	99.97	100	100
Size	77.31	90.73	98.22	99.44	99.54
Typeface	75.32	89.11	97.35	98.87	99.51
Font	64.15	84.62	96.91	98.23	99.02

TABLE 4
TYPOGRAPHICAL AND MORPHOLOGICAL CLASSES OF CONNECTED COMPONENTS

T(cc)			M(cc)		
Class	Condition	Samples	Class	Condition	Samples
Full	$\|t(cc) - to\| \le \epsilon$ and $\|bo - b(cc)\| \le \epsilon$	f,[,],{,},(,),/	Wide	$r \ge 1.75$	linked characters
High	$\|t(cc) - to\| \le \epsilon$ and $\|b(cc) - bl\| \le \epsilon$	b,d,f,h,k,l,t	Large	$r \in$]1.25, 1.75]	m, w
Deep	$\|t(cc) - ul\| \le \epsilon$ and $\|bo - b(cc)\| \le \epsilon$	g,p,q,y	Square	$r \in$]0.75, 1.25]	a,c,e,n,o,r,z,*,+,
Short	$\|t(cc) - ul\| \le \epsilon$ and $\|b(cc) - bl\| \le \epsilon$	a,c,e,m,w,x	Tall	$r \in$]0.5, 0.75]	b,d,h,k,&,g,p,q,y
Sup	$\|t(cc) - to\| \le \epsilon$ and $\|ul - b(cc)\| \le \epsilon$	', ‘	Thin	$r < 0.5$	l,t,f,|
Sub	$\|t(cc) - bl\| \le \epsilon$ and $\|bo - b(cc)\| \le \epsilon$	,,.	Small	$h(cc) \le \frac{1}{25}$ *resol.*	. -

broken into smaller entities of length $\frac{1}{4}$L and $\frac{1}{2}$L, or merged[4] together to build new entities of length 2L and 4L. Table 3 shows recognition rates for these text lengths. The results confirm the intuitive prediction that recognition accuracy enhances with the length of the text line. While the weight and slope detection remains robust to short texts, size, and typeface are, as expected, obviously less accurate.

The classifier has also been applied on single words of a limited FMB of 84 fonts. Experiments have confirmed the fact that the size and typeface attributes are the hardest to identify on short texts. A text length modeling was used to automatically adapt the classifier to text length. An improvement of recognition results was observed [3].

4 CONCLUSIONS

This paper has addressed the problem of optical font recognition (OFR), which has been widely ignored by the scientific community so far. The aim of the developed system[5] is to analyze a text line image and to identify the typeface, the font style, and size from a given set of already learned fonts. We have adopted a statistical approach based on the extraction of a few well-selected global features from a medium resolution image of a scanned text.

The method has been extensively tested on more than thirty thousand two-column formatted text lines extracted from scanned documents using a set of 280 distinct fonts. The experimental results are extremely encouraging: the measured overall recognition rate was close to 97 percent. The classifier obtained accuracy rates even higher than 99.9 percent for the more practical problem of identifying the font style of a given typeface. The method can be considered as applicable on short texts of about ten characters.

The accuracy of the results suggests the use of such a tool, not only for logical structure recognition, for which the system was originally planed, but also to improve OCR accuracy by combining it with monofont OCR systems [15]. The latter are known to have better performances than omnifont ones.

4. In case of merging, feature vectors were computed as averages from those of the default lines of length L.

5. *ApOFIS* is available as a C++ library that can be downloaded from our web site http://www-iiuf.unifr.ch/groups/sde/projects/das/apofis.html.

Finally, the reported results have been applied on 300 dpi images of rather good quality obtained by scanning of laser printed pages. Some recent experiments have shown that the method is still applicable on slightly degraded documents such as those obtained by first- or second- generation photocopies, provided that the font model base has been built under the same conditions. In practice, such an assumption may only be hardly satisfied; therefore, we have to consider a more realistic approach, which consists of adapting the font model base automatically according to the type of degradation.

APPENDIX A
CLASSIFICATION OF CONNECTED COMPONENTS

Within images, characters are located by the envelopes of their connected components (cc), which are defined by their top $t(cc)$, bottom $b(cc)$, left $l(cc)$, and right $r(cc)$ coordinates. Their heights are defined as $h(cc) = b(cc) - t(cc)$ and their widths as $w(cc) = r(cc) - l(cc)$.

Typographical classes are practically determined using the positions of $t(cc)$ and $b(cc)$ from the typographical lines. A tolerance factor ϵ, empirically fixed to $\epsilon = \frac{1}{12}|to - bo|$, is introduced to consider position fluctuations within the line. Six typographical classes ($T(cc)$) have been defined as illustrated in Table 4 and Fig. 3.

The morphological classification is based on the cc dimensions with classes distinguished by the ratio $r = w(cc)/h(cc)$ as shown in Fig. 3. Finally, six morphological classes ($M(cc)$) have been defined as illustrated in Table 4. Frontiers between these classes were fixed empirically through a statistical analysis of the ratio r for various fonts.

APPENDIX B
FEATURES DESCRIPTION

Features are extracted from the horizontal projection profile of text images and from connected components. Spaces between words are ignored and replaced by a fixed value corresponding to a median character-space. Each space between two successive connected components bigger than $|bl - ul|$ is assumed to be a word-space and therefore ignored.

Two features are extracted from the horizontal projection profile HP_{ub} computed on the *central* zone [*ul*, *bl*] in order to guarantee

text content and structure independence (see Fig. 4). They are HP_{ub} density (*hpdn*) and the density of squared values of HP_{ub} derivative (*hpdr*):

$$hpdn = \frac{1}{n}\sum_{i=1}^{n} HP_{ub}[i] \text{ and } hpdr = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(HP_{ub}[i+1] - HP_{ub}[i]\right)^2 .$$

where n represents the size of HP_{ub}.

The rest of the features are extracted from connected components. Let us suppose a text entity composed of n connected components horizontally ordered (cc_i) where *Small*, *SuperSc*., and *SubSc*. components are ignored). We define

1) *ahcc* as the average normalized height of *Full*, *High*, *Deep*, and *Short* components. Normalization is based on the proportions of the *upper*, *central*, and *lower* zones:

$$ahcc = \frac{1}{n}\sum_{i=1}^{n} \alpha h(cc_i), \text{ where } \alpha \begin{cases} \frac{|bl-ul|}{|bo-to|} & \text{if } T(cc_i) = Full \\ \frac{|bl-ul|}{|bl-to|} & \text{if } T(cc_i) = High \\ 1 & \text{if } T(cc_i) = Short \\ \frac{|bl-ul|}{|bo-ul|} & \text{if } T(cc_i) = Deep \end{cases}$$

2) *awcc* as the average width of *Squared* and *Short* connected components:

$$awcc = \frac{1}{m}\sum_{i=1}^{m} w(cc_i), \text{ where } m = \#cc_i \text{ such that}$$
$$T(cc_i) = Short \;\&\; M(cc_i = Squared).$$

If such components are missing in a text line, which is very unlikely, a simple average is retained.

3) *ascc* as the average space between cc_i envelopes within words. A Space (ccs_i) separate envelopes cc_i and cc_{i+1}, i.e., $ccs_i = l(cc_{i+1}) - r(cc_i)$:

$$ascc = \frac{1}{m}\sum_{i=1}^{m} ccs_i, \text{ where } m = \#cc_i \text{ such that } ccs_i \le |bl - ul|.$$

4) *awhr* as the average width of horizontal black-runs. Each cc_i is traversed horizontally at its middle scanline $y_c = \frac{1}{2}\left(t(cc_i) + b(cc_i)\right)$ to compute k black-runs hbr_j. From a text line, *awhr* is finally computed as an average of the retained hbr_j (very large runs are ignored):

$$awhr = \frac{1}{m}\sum_{j=1}^{m} hbr_j, \text{ where } m = \#\, hbr_j \text{ such that}$$
$$\frac{1}{8}w(cc_i) < hbr_j < \frac{1}{2}w(cc_i).$$

5) *ahvr* as the average height of vertical black-runs. Each cc_i is traversed vertically at the central scancolumn $x_c = \frac{1}{2}\left(l(cc_i) + r(cc_i)\right)$ to compute black runs vbr_j. From a text line, *ahvr* is defined as the average of the retained vbr_j (high runs are ignored):

$$ahvr = \frac{1}{m}\sum_{j=1}^{m} vbr_j, \text{ where } m = \#\, vbr_j \text{ such that}$$
$$\frac{1}{8}h(cc_i) < vbr_j < \frac{1}{3}h(cc_i)$$

6) *dpld* as the density of the difference between consecutive scanlines. It is computed on regions $R_1 = [t(cc), t(cc) + \delta]$ and $R_2 = [b(cc) - \delta, b(cc)]$, where $\delta = \frac{1}{3}h(cc)$ to focus on serifs:

$$dpld = \frac{1}{n}\sum_{i=1}^{n} \frac{S(cc_i)}{w(cc_i)},$$

where $S(cc_i)$ is # pixels in the difference image in cc_i.

REFERENCES

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

[2] G. Nagy and S. Seth, "Modern Optical Character Recognition," *The Froehlich/Kent Encyclopedia of Telecommunication*, F. Froehlich and A. Kent, eds., vol. 11, pp. 473–531. Marcel Dekker, Inc., 1996.

[3] A. Zramdini, "Study of Optical Font Recognition Based on Global Typographical Features," Ph.D. dissertation, University of Fribourg, 1995.

[4] J.C. Anigbogu, *Reconnaissance de Textes Imprimés Multifontes à l'aide de Modèles Stochastiques et Métriques*, Ph.D. dissertation, Université de Nancy I, 1992.

[5] Y. Chenevoy, *Reconnaissance Structurelle de Documents Imprimés: Études et Réalisations*, Ph.D. dissertation, CRIN-University of Nancy, France, 1993.

[6] G.E. Kopec, "Least-Square Font Metric Estimation From Images," *IEEE Trans. Image Processing*, vol. 2, no. 4, pp. 510–519, Oct. 1993.

[7] B. Cooperman, "Producing Good Font Attribute Determination Using Error-Prone Information," *SPIE*, vol. 3,027, pp. 50–57, 1997.

[8] H. Shi and T. Pavlidis, "Font Recognition and Contextual Processing for More Accurate Text Recognition," *ICDAR'97: Fourth Int'l Conf. Document Analysis and Recognition*, pp. 39–44, Ulm, Germany, Aug. 1997.

[9] R.A. Morris, "Classification of Digital Typefaces Using Spectral Signatures," *Pattern Recognition*, vol. 25, no. 8, pp. 869–876, 1992.

[10] S. Khoubyari and J. Hull, "Font and Function Word Identification in Document Recognition," *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 66–74, 1996.

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

[12] B. Bauermeister, *Manual of Comparative Typography: The PANOSE System*. New York: VNR Company, 1991.

[13] P.G. De Luca and A. Gisotti, "Printed Character Preclassification Based on Word Structure," *Pattern Recognition*, vol. 24, no. 7, pp. 609–615, 1991.

[14] S. Chen, F.Y. Shih, and P.A. Ng, "Fussy Typographical Analysis for Character Preclassification," *TSMC*, vol. 25, no. 10, pp. 1,408–1,413, Oct. 1995.

[15] F. Bapst and R. Ingold, "Using Typography in Document Image Analysis," *RIDT'98: Fourth Int'l Conf. Raster Imaging and Digital Typography*, San Malo, France, Apr. 1998.