Italic or Roman : Word Style Recognition Without A Priori Knowledge for Old Printed Documents

Loris Eynard, Hubert Emptoz
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205,
F-69621, France
{loris.eynard, hubert.emptoz}@insa-lyon.fr

Abstract

This paper presents an Italic/Roman word type recognition system without a priori knowledge on the characters' font. This method aims at analyzing old documents in which character segmentation is not trivial. Therefore our approach segments the document into words and analyse the text word per word. To define the word style, we combine three criteria which are based on the visual differences between a word and a slanted version of the same word. These criteria are defined thanks to features computed from the vertical projection profile of the word. Because we do not assume a specific slant angle, we compute these measures on a whole range of possible slant angles and then sum the obtained scores. Our results show a ratio of 100 % recognition for Italic words and 97.2 % for Roman words.

1. Introduction

Our work was designed to find Italic words in historical newspaper of the 18th century, more specifically in the images of the *Gazette de Leyde* dataset.
Commercial software for character recognition are not efficient for recognizing the character in old documents. On the *Gazette de Leyde* dataset we obtain a ratio of 88 % of character recognition [1] when 99 % is considered to be efficient. Of course on well conserved part of the documents the results are very good but on damaged parts this ratio decrease sharply. The Italic words, due to the thickness of the Italic character, are more sensitive to the time damages than the Roman word and so they are badly recognized by commercial OCR.
Those Italic words represent particular nouns (patronymic or location names) and so are very interesting for researchers in Human Sciences. The final aim is to extract words typeset in Italic in paragraphs typeset in Roman in several years of Gazette which represents thousands of pages.

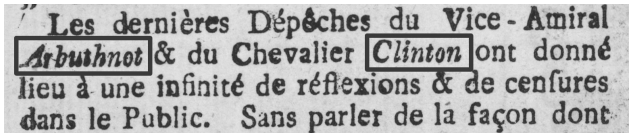
Les dernières Dépêches du Vice-Amiral *Arbuthnot* & du Chevalier *Clinton* ont donné lieu à une infinité de réflexions & de censures dans le Public. Sans parler de la façon dont

Figure 1. Italic style nouns in the middle of a Roman style paragraph

Because of this large amount of documents, our method must be able to decide very precisely if a word is either in Italic or Roman style. The difficulties we encountered are due to the the conservation of the old documents and their digitization, which results in several types of degradations, e.g ink bleed-through, holes, ink fading, etc . These artefacts create links between characters (and even more between Italic characters which are closer than Roman ones) which makes them harder to segment. For this reason we propose a characters' segmentation-free method. We base ourself on the visual characteristics of the characters of a word that can be interpreted by analysing the vertical projection profile of the word image. Those analysis based on three visual features would give us scores to decide whether a word is in Italic or Roman style. As we do not suppose a specific slant angle of the Italic type, which may vary significantly across the document we are processing, we test a range of angles rather than only considering one slant angle.
This paper is organized as follows : first we recall the state of the art on this problem. Then we describe our method and conclude with the results.

[1] with ABBYY Fine Reader 8 Pro

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.176

2. Previous Works

A large amount of works exists on font recognition but not that much on the recognition of the text style and even less specifically on Italic recognition.
Two types of approaches can be identified : One is based on characters segmentation and features extracted on the character. The other considers words as textures and analyzes them statistically or in the frequency domain.
The first hypothesis assumed by the character-based methods is that the character segmentation is correct. In cases where the character segmentation was possible, Chauduri *et al.* [1] obtained good results on recent documents by computing the slant angle of each character. Assuming that there is always a black line going from the bottom of the character to the top of it, they search for the angle that this line makes with the base line to define the character as Italic type. This method works mostly on recent and well slanted characters. Ma *et al.* in [5] assume that OCR results are available. They use those results to select features extracted from the characters in order to classify the characters. A Gaussian model is used to create clusters of characters that are classified between styles. They decide the word's style by summing the characters style. Fan *et al.* use structural informations from strokes extracted from the characters to classify them in three types. Italic characters are detected depending of their type using either gradient information, curvatures of strokes or angle with the horizontal line and then rectified ([2]).
We can also cite the works of Li *et al.* who separate Italic touching characters in [4]. The interesting fact is that they don't suppose an a-priori knowledge on the slant angle. Instead they cover a range of possible angles rotating the word's image. Once they have obtained the correct angle they separate the characters using cut paths.
All those works are mostly efficient for modern documents and suppose the document to be well conserved and digitized, which is not the case for the eighteenth century's documents we are dealing with here. An other possibility, when no characters segmentation is possible, is to process with the full image of the word.
For handwritten documents some interesting works have been presented by Kavallieratou *et al.* in [3] who use the Wigner-Ville Distribution on the vertical histogram to define and correct a slanted word. In [8] , Zhang *et al.* resort to statistical analysis of stroke patterns obtained from a wavelet decomposition of the word image to detect Italic or bold word. Finally in [7], Sun *et al.* define a method to straighten documents by computing the histogram of gradient orientations. They recommend this method for Italic style's recognition and correction by comparing the orientation of the characters with the lines orientation but no results are shown.

3. Our Proposal

Our proposal is based on the differences existing between the original word and a skewed version of it, by attempting to straighten an Italic word. If the word actually was in Italic style then the skewed version would be Roman-like, and if the original word was Roman style then the slanted one will look like an inverted Italic word (see figure 2). Our main idea is to translate the visual obvious differences between the two versions of the word with values computed by analysing the vertical projection profiles. Those differences will give us a Roman style score and an Italic style score for each word with a given slant angle, resulting in a decision for this particular angle. By summing the decisions for each possible slant angle of the range we obtain a final decision.
First of all, we binarize the word images. The type of binarization depends of the documents. We use a threshold applied on the lightness of the image. A good survey of various binarization methods is done by Pamarkos *et al.* in [6].
In the next section we describe our slanting method. Then we introduce the three factors we use for Italic style's decision.

3.1. Shear transform

In this section we briefly describe the method that we used to perform a shear transform on the original word, for obtaining the word version of comparison. This method is inspired by the works of Sun *et al.* in [7].

Figure 2. The original word (right) and the reverse slanted one (left)

The main idea of this transform is that only the top of the character is moved, shifted to the left in our case. The base of the character is not shifted. First we compute the needed difference of width Δ between the two words for a slant angle α as follows: $\Delta = |h * \tan(\alpha)|$ where h is the height of the word's image. Here the value of α is 0 for the original word, negative to straighten an Italic word and positive to futher slant the word. We vertically cut the word's image into Δ stripes having the width of the word but $1/\Delta$ of its height. The first bottom stripe won't be slid, the next one would be slid by one pixel, and so on to obtain a slant of Δ degrees. We obtain the figure 2. Note that the result (on the

left) seems visually correct.
In the next three sections we describe the three criteria we use to define the Italic style. They are based on the differences between a word in Italic style and the same word in Roman style.

3.2. First criterion : Vertical black column

The main difference between an Italic and Roman version of a word is the presence of long vertical strokes in the case of the Roman version, which are represented as peaks in the vertical projection profile.

Figure 3. highest black column of histogram, see that maximum of histogram differs from original and slanted images

Slanting a vertical stroke in the image translates into flattening the corresponding peak in the projection profile (see figure 3). The difference between a Roman style word and an Italic style version of it is given by comparing the maxima of their respective projection profiles.

We obtained two values MH_o for the original word and MH_s for the slanted one. The first criterion, called C_1 is valued at 1 if$MH_s > MH_o$ and 0 otherwise.

3.3. Second Criterion : Overlapping of the characters

Considering an Italic style word, we observe that the top of a slanted character often overlap the bottom of the following. This observation will be the same for an inverted-Italic style word.

Even if there is not a real overlapping then the white space between the two characters is noticeably reduced. This criterion shows this difference between the white spaces in the Roman version and in the Italic version of a word (see fig. 4).

This feature is translated in the vertical projection profile by analysing the white space between black sections. Each black section represents a character (or more if there

Figure 4. Overlapping of characters in slanted word (white space are marked by dark vertical lines for the slanted word and by light vertical lines for the original word)

is overlapping) and the white space represents the space between characters. In the projection profile, we search all the white spaces between two black pixels. By deduction the more white-between-black pixel there is, the more the word is Roman style. Let's call W_o the total width of white space in the original image and W_s in the slanted word. We give the value 1 to the second criterion C_2 if $W_o > W_s$ and 0 otherwise.

3.4. Third criterion : Variation of the slops of the Vertical Histogram

The last criterion that we used is the variation of slope in the vertical projection profile of the word. For each word, the characters are represented as peaks of black pixels in the vertical projection profile. If this word is Roman style, then those peaks would have values of slope certainly higher than for a slanted word. Moreover these slopes will vary much more suddenly than for an Italic word. This could be explained because an Italic style character appears more spread in the vertical projection profile than a well straight Roman one. These differences of slope translate the more horizontal concentration of the black pixel for a non-slanted character.

To compute the variation of slope of the projection profile we compute the second derivative of the vertical projection profile. As for the first criterion it would make no sense to consider only the maximum of these variations. That is why we compute an average of the ten maximum variations of slope of the projection profile. If we consider VS_o the average maxima variation of slope for the original word and VS_s for the slanted image. Then this criterion, C_3 get the value 1 if $VS_o > VS_s$ and 0 if not.

3.5. Final Style Decision

We obtain three binary criteria C_1, C_2 and C_3 giving us indications on the style of the word. Considering only those binary criteria will give us a too arbitrary decision for the word's style. To specify this decision, according to the ground truth, we define a weight for each criterion. By

combining the criterion and its associated weight we obtain a score to decide the word's style. These weights represent a ratio of words verifying the criteria according to their style. For example w_1^{ro} represent the percentage of Roman style words verifying the first criterion. These percentage are computed on only 40 words of each style. The computed weight values are shown in table 1.

		Style	
		Italic	Roman
criteria	C_1	0.15	0.8
	C_2	1	0.5
	C_3	1	0.4

Table 1. Weights values

Thanks to the criteria and their associated weights we can define decision terms for each style. We compute these decision terms for each angle α as follow :

$$T_\alpha^{Roman} = \sum_{i=1}^{3} (w_i^{ro}.C_i + (1 - w_i^{ro}).(1 - C_i))$$

$$T_\alpha^{Italic} = \sum_{i=1}^{3} (w_i^{it}.C_i + (1 - w_i^{it}).(1 - C_i))$$

.

T_α^{Roman} is the Roman style decision term for a word for the skew angle α. T_α^{Italic} is the Italic style decision term of the same word for the same slant angle α. We compare the values of these terms to decide if the word is Italic or Roman style for the specific slant angle α. If $T_\alpha^{Roman} > T_\alpha^{Italic}$ then the word is characterized as Roman style for the slant angle α and vice versa. According to this, we call D_α the decision for the angle α described as follow:

$$D_\alpha = \begin{cases} Roman & if \quad T_\alpha^{Roman} > T_\alpha^{Italic} \\ Italic & if \quad not \end{cases}$$

In old documents, the slant angle of the Italic style is not fixed like in recent's one. We can not assume a specific skew angle but we suppose Italic style to be slanted between 5 and 20 degrees. We test all these angles before choosing the word's style. We assume that if a word is characterized Roman style in most of the 15 angles that we test, then it would be fixed as Roman style. We call D the final decision for a word's style. To define this decision D we adapt the Kronecker Delta noted δ_{s,D_α} to :

$$\delta_{s,D_\alpha} = \begin{cases} 1 & if \quad s = D_\alpha \\ 0 & if \quad not \end{cases}$$

The final decision D gives us the style for which the sum for each angle of this adaptated Kronecker Delta is maximized. Then the final decision D is defined as follow :

$$D = \arg \max_{s=Roman,Italic} \sum_{\alpha=5}^{20} \delta_{s,D_\alpha}$$

If we call intermediate decision for the angle α the D_α. The result of the function D is the style which gives the maximum of intermediate decisions by summing the decision for all fifteen possible slant angles. D result in a two choices decision for the word to be either Roman or Italic style.

4. Results and Discussion

Our approach was designed to detect Italic style words such as proper nouns in the *Gazette of Leyde* dataset. Those nouns are supposed to be patronymics or toponyms which are mostly large words of 4 letters or more. For this reason we don't expect good results on short words (words containing less than 3 letters). Moreover, our method takes into account overlapping characters. For a two letters word there is only one possible overlapping and two for a three letters word. This observation decrease significantly the influence of the criterion C_2 on the final decision. Table 2 shows the word's length histogram for the dataset of 1358 words we used in our experiments.

	words size(in letters)			
	2	3	≥ 3	Total
Italic	15	14	176	205
Roman	269	145	739	1153

Table 2. Number of testing words

This distribution naturally arose from the dataset and has not been influenced by us. The small number of Italic words is related to the specific typesetting of the documents.

	words size(in letters)			
	2	3	≥ 3	total
Italic	100	100	100	100
Roman	89.5	97.2	99.99	97.2

Table 3. Recognition Rates(%)

We obtain really good results for deciding whether a word is Italic style or Roman style. As expected, our results are lower for short words (2 or 3 letters) but still very good. As this method is character segmentation-free, it can recognize word's styles on old or blurred documents, as well as documents containing words with touching characters We expect this method to work for any class of documents and with any character style.

4.1. Acknowledgments

This work take part of a *Cluster Culture, Patrimoine et Création* which is a regional research cluster of the Region Rhone-Alpes, France.

References

[1] B. B. Chaudhuri and U. Garain. Automatic detection of italic, bold and all-capital words in document images. In *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition-Volume 1*, page 610, Washington, DC, USA, 1998. IEEE Computer Society.

[2] K.-C. Fan and C. H. Huang. Italic detection and rectification. *J. Inf. Sci. Eng.*, 23(2):403–419, 2007.

[3] E. Kavallieratou, N. Fakotakis, and G. Kokkinakis. Slant estimation algorithm for ocr system. *Pattern Recognition*, 34:2515–2522, 2001.

[4] Y. Li, S. Naoi, M. Cheriet, and C. Y. Suen. A segmentation method for touching italic characters. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2*, pages 594–597, Washington, DC, USA, 2004. IEEE Computer Society.

[5] H. Ma and D. Doermann. Adaptive word style classification using a gaussian mixture model. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2*, pages 606–609, Washington, DC, USA, 2004. IEEE Computer Society.

[6] E. K. Pavlos Stathis and N. Papamarkos. An evaluation survey of binarization algorithms on historical documents. *ICPR '08: Proceedings of the 19th International Conference on Pattern Recognition*.

[7] C. Sun and D. Si. Skew and slant correction for document images using gradient direction. In *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition*, pages 142–146, Washington, DC, USA, 1997. IEEE Computer Society.

[8] L. Zhang, Y. Lu, and C. L. Tan. Italic font recognition using stroke pattern analysis on wavelet decomposed word images. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*, pages 835–838, Washington, DC, USA, 2004. IEEE Computer Society.