Printer ballistics through character's texture analysis

Adriano Ruggero[*]
Gabriel Rodrigues[†]
Mário Brito[‡]
Maurício Perez[§]
Anderson Rocha[¶]

Abstract

We describe a technique for ballistics of printed documents, that is, link a printed document to a specific printer. The principle of this technique is the analysis of character's texture, by extracting some properties from the characters of scanned images from the printed documents, and relate this properties through a co-occurrence matrix. This matrix can be used to create a "fingerprint" for the characters related to the same printer, what allows us to identify the specific printer device that printed these characters.

1. Introduction

In August of 2013, a russian man wrote his own small print in a credit card contract [1]. The credit card's administrator bank didn't read the amendments made by the client, and just signed and certified the document. The changes included unlimited credit line, 0 percent interest rates and no fees. When the bank decided to terminate the man's credit card, because overdue payments, he sued them for more than 24 million rubles (US$ 727.000). How could the bank prove the falsification?

Altough we are living in a digital era, printed documents still are a significant part of our day by day. Likewise, with the constant reduction in prices and increase in quality of printing equipment, forgeries become increasingly commonplace.

Legal aspects aside, a way to verify if a document, or a part of it, came from a specific device can be through character's texture analysis.

Our approach for the analysis of character's texture is as follows. From printed pages scanned at high resolution, selected characters were extracted. From these characters, we obtained its properties of contrast, correlation, energy and homogeneity, creating with them a co-occurrence matrix. This matrix can be called a "fingerprint" of the character. This "fingerprint" of characters is closely related to the printing device which originated it, and can be used to identify which printer was responsible for printing it.

However, slight imperfections may occur during the printing and/or scanning process of documents. To handle these small errors (or variations), characters were selected from different areas of scanned document, their properties were obtained and then classified using machine learning algorithms.

2. State-of-the-Art

Related work: [2] [3].

3. Proposed Solution

Our solution consist in getting the image of characters selected from scanned documents in grayscale, extract its properties of contrast, correlation, energy and homogeneity - creating a co-occurrence matrix - and cluster them by machine learning algorithms.

3.1. Printers dataset

The documents used for this work were obtained from the Wikipedia site [4], and were written in English, some of them containing pictures, some not. They were printed by printers listed in Table 3.1.

These printed documents were scanned at high resolution and saved as Tiff file format (Tagged Image File Format), forming a database. These files were made available through an FTP site [5]. An example of a typical document

---

[*]Institute of Computing, University of Campinas (Unicamp). **Contact**: `arruggero@lasca.ic.unicamp.br`

[†]Institute of Computing, University of Campinas (Unicamp). **Contact**: `gabriel_rodrigues@aol.com`
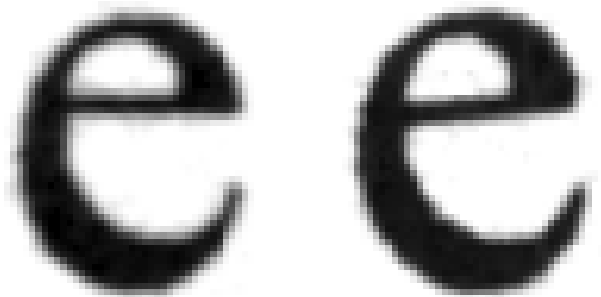
[‡]Institute of Computing, University of Campinas (Unicamp). **Contact**: `britomar@aedu.com`

[§]Institute of Computing, University of Campinas (Unicamp). **Contact**: `mauriciolp84@gmail.com`

[¶]Institute of Computing, University of Campinas (Unicamp). **Contact**: `anderson.rocha@ic.unicamp.br`

Table 1. Printers used in this work

Printer	Documents
Brother-HL4070CDW	28
Canon-D1150	28
Canon-MF3240	28
Canon-MF4370DN	28
HP-CLJ-CP2025A	28
HP-CLJ-CP2025B	28
HP-JL-CP1518	28
Lexmark-E260D	28
OKI-C330	28
Samsung-CLP315	28

Figure 1. Typical document used in this work [4].

Figure 2. Examples of characters: the left one was considered useful; the right one, slightly rotated.

acters printed in normal font, not analyzing characters in bold, italic, underline, strikeout, superscript, subscript, or any other form of letter change.

Were not considered characters from misaligned documents, i.e. characters from apparently obliquely scanned documents. Figure 2 shows examples of two characters obtained from scanned documents, one of them being usable and the other not, being considered rotated.

Subsection 3.4 will be explained in more detail why not use rotated characters.

3.3. Printers

Due to the fact that all scanned documents come from laser printers, it was necessary to take some precautionary measures. Laser printers are known as "page printers", while dot matrix printers and inkjet printers are called "line printers". This is a crucial difference, and should be considered for more careful study.

Line printers print documents line by line from the top of the sheet, keeping a characteristic pattern which periodically repeats for each paper feed. That is, any line printed by this printer model will have basically the same characteristics, regardless of their vertical location in the paper sheet.

Page printers, instead, do not print documents line by line. In this case, an image of the entire page is "printed" on the photoreceptor drum by a laser unit. This image attracts toner particles, and then transfers it to the paper sheet. Finally, a fuser unit heats the paper, so the toner melts and attach it. Figure 3 shows a default page printer schema.

By working in this way, page printers do not maintain a characteristic pattern that is repeated line by line across the printed paper sheet. Due to small imperfections that may exist in the photoreceptor drum, each print area gener-
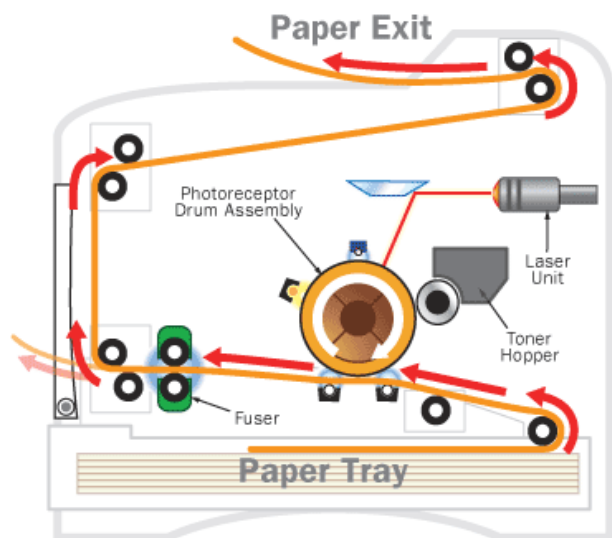
used in this study can be seen in Figure 1.

3.2. Characters

The characters chosen for this work were "e" and "t", both in lowercase, because they are, respectively, the first and second most common letters in texts written in English [6].

To avoid inconsistencies and / or values "off the curve", we used only characters printed in the same font and size, for all printers. With the same objective, we used only char-

Figure 3. Default page printer schema [7].

Figure 4. Division of the scanned document in areas.

ated by this type of device can present different characteristics. Taking into account this fact, it was necessary to obtain characters from different parts of the scanned document.

Basically, each document has roughly divided into three parts: upper, middle, and bottom. The Figure 4 shows an example of a document divided in this way. If this article is being viewed in color, the upper part of the figure is in red, the middle, in green and bottom, in blue.

Despite showing subtle differences between characters located in different areas of the printout, the print device maintains certain intrinsic characteristics unchanged in all printed characters. Such characteristics can be compared to our fingerprints, making it a way to link a printed character to a particular device.

3.4. Gray level co-occurence matrix

Co-occurrence matrix is a tabulation of how many different combinations of intensity values of pixels (grayscale) occur in an image. The primary use of the co-occurrence matrix is characterized texture in an image from a set of statistics for instances of each gray level in different pixels along different directions [8].

Co-occurrence matrix considers the relationship between two pixels at a time, one called as "reference pixel" and another as "neighbor pixel". The neighboring pixel chosen can be neighbor in any direction: east (right), west (left), north (above), south (below), or diagonally, i.e. northeast, northwest, southwest and southeast of each reference pixel. Also the neighborhood need not be exactly one pixel, can be 2, 3, or any value. Each pixel within the image becomes the reference pixel, starting at the top left and proceeding to the lower right. There will be particular cases. For instance,

pixels of the right margin of the image does not have neighboring right.

Co-occurrence, in its most general form, can be specified by a matrix of relative frequencies $P(i, j, d, \theta)$, where two neighboring texture elements separated by a distance d in a direction θ occur in the image, one with the property i and another with property j.

From a co-occurrence matrix can be obtained some properties of the analyzed image. For this work, we used the properties of contrast, correlation, energy and homogeneity to create a fingerprint of the printed characters.

Using an example similar to the rotated character, we can compare the value differences in the properties obtained from the co-occurrence matrix (Figure 5). In this case, the character on the left was obtained directly from the scanned document, the character on the right is the same character, but rotated -4º.

Table 3.4 lists the differences between the properties of the character sample obtained from the co-occurrence matrix.

As shown in the Table 3.4, small changes in the alignment of the print / scan can cause large variations in the properties values obtained from the co-occurrence matrix.
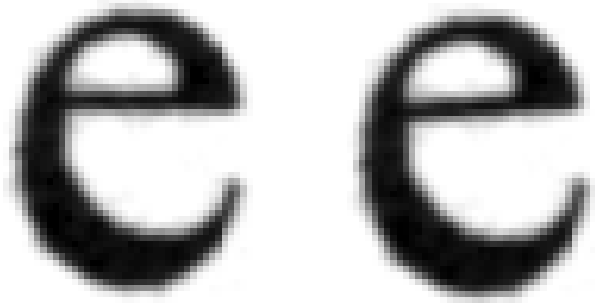
Figure 5. On the left, a character obtained from a scanned document and to the right, the same character, rotated -4°.

Table 2. Differences between an original character and a rotated character.

Property	Original character	Rotated character (-4°)
Contrast	3.2443 - 2.2905	5.1617 - 4.6504
Correlation	0.7869 - 0.8502	0.6967 - 0.7264
Energy	0.1608 - 0.1744	0.1106 - 0.1216
Homogeneity	0.6946 - 0.7462	0.6610 - 0.6995

4. Experiments and Discussion

5. Conclusions and Future Work

References

[1] RT. 700k windfall: Russian man outwits bank with handwritten credit contract. http://rt.com/business/man-outsmarts-banks-wins-court-221/, last access in November 17, 2013. 1

[2] Orhan Bulan, Junwen Mao, and Gaurav Sharma. Geometric distortion signatures for printer identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 1401–1404. IEEE, 2009. 1

[3] Eric Kee and Hany Farid. Printer profiling for forensics and ballistics. In *Proceedings of the 10th ACM Workshop on Multimedia and Security*, MM&Sec '08, pages 3–10, New York, NY, USA, 2008. ACM. 1

[4] Wikipedia. Wikipedia - the free encyclopedia. http://www.wikipedia.org/, last access in November 19, 2013. 1, 2

[5] Giuliano Pinheiro. Index of / giulianorp/printer_dataset. http://www.recod.ic.unicamp.br/~giulianorp/printer_dataset/, last access in November 19, 2013. 1

[6] University of Notre Dame. Letter frequencies in the english language. http://www3.nd.edu/~busiforc/handouts/cryptography/Letter.html, last access in November 17, 2013. 2

[7] Forensic Document Examination Services. Letter frequencies in the english language. http://www.forensicdocumentexaminer.co.uk/Document_Analysis.html, last access in November 17, 2013. 3

[8] Anderson de Rezende Rocha and Neucimar Jerônimo Leite. Classificação de texturas a partir de vetores de atributos e função de distribuição de probabilidades. 3