

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, extracting some properties from the characters of printed and scanned documents, and relate this properties through a co-occurrence matrix. This matrix can be used to create a "fingerprint" of characters (and related printers), that allows identify a specific printer device that print 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 documents. The changes included unlimited credit line, 0 percent interest rates and no fees. When the bank decided to terminate man's credit card, because overdue payments, he sued they for more than 24 million rubles (US\$ 727.000). How the bank could prove the falsification?

Although we are living in a digital era, printed documents still are significant part of our daily. 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 character's

texture analysis.

Our approach for the analysis of character's texture made up 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 an 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

Here goes the state-of-the-art research (talk about prior work for solving the same problem).

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.

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 [2].

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

*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

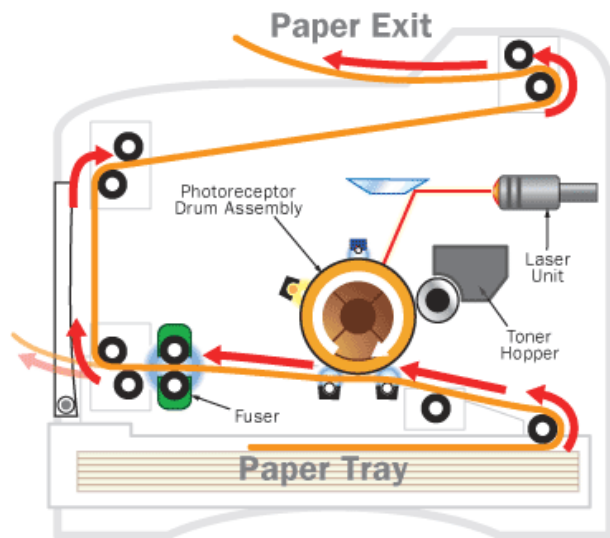


Figure 1. Default page printer schema [3].

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

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 sheet of paper. 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 1 shows a default page printer schema.

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] 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. 1

- [3] Forensic Document Examination Services. Letter frequencies in the english language. http://www.forensicdocumentexaminer.co.uk/Document_Analysis.html, last access in November 17, 2013. 2