Exploration and Research of Laser Printing Document Traceability Methods in the Context of Big Data

Yifan Tang(✉)

College of Forensic Sciences, Criminal Investigation Police University of China, Shenyang 110854, Liaoning, China
1297075179@qq.com

Abstract. With the optimization of printing technology and the advent of the era of universal printing, the number of cases involving the traceability of laser printed document is increasing and the complexity of cases is by no means the past. Printers and copiers have become an important medium for governments, companies and individuals to exchange information. Printers and photocopiers can produce many documents such as contracts, bills, materials, certificates, etc., and the resulting criminal activities are endless. When authenticating printed document, examiners usually conduct analysis from both morphological and physical and chemical properties. But in principle, there is no rich experience and sufficient samples in actual cases, the success rate of identifying the authenticity of documents, determining the way in which documents are printed and identifying the source of documents is low, and some methods may also damage the documents under investigation. However, there are few integrated studies on traceability methods. Based on this, this paper reviews the existing digital methods and concludes with an analysis of the development trends of document traceability, in an attempt to provide ideas and references for forensic experts.

Keywords: Laser printed document · Forensic examination of printed document · Source printer identification

1 Introduction

Laser printed document uses electronic imaging and laser scanning technology to convert the digital information of a computer into printing information to paper and other media through machines and equipment. Experts use of the information contained in the document to ascertain the source of a printed document. Figure 1 shows a simple view of a typical EP printer. After the electronic imaging of the printed document by charging, exposure, developing, transferring, fixing and cleaning, the character image in the document can reflect the internal characteristics. With the improvement of the printing level, the ability of identification of damage legacy features is weakened, and the inherent traditional inspection and identification method can only be used as a necessary rather than a sufficient condition for document identification. Digital inspection

P. Qi and Z. Chen (Eds.): ICBDIE 2023, AISR 178, pp. 791–795, 2024.
https://doi.org/10.2991/978-94-6463-238-5_103

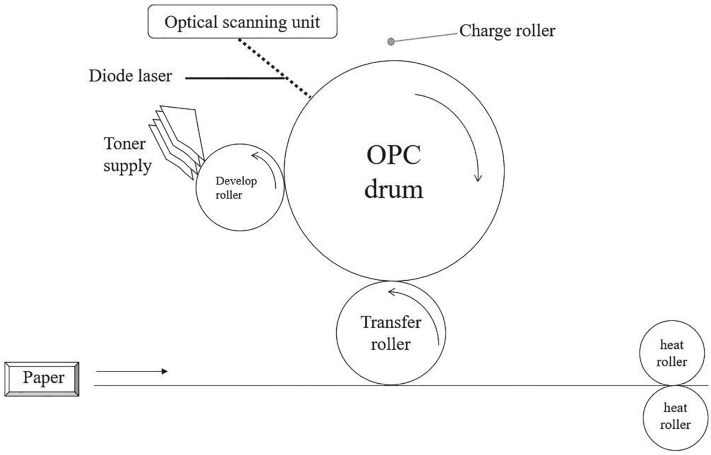

Fig. 1. Diagram of printer identification

is a hot field of attention at home and abroad at present. This paper summarizes the digital traceability methods and collects the research results such as machine learning and computer vision in order to inject new impetus into the future identification work.

2 Digital Methods

With the emergence of traditional comparison vulnerabilities and the continuous development of digital technology, converting printed document into digital formats and applying advanced technologies (machine learning, deep learning and computer vision) to find stable features to identify source printers has attracted wide attention in recent years. Figure 2 shows the block diagram of the printer identification scheme for printed document. This paper is classified accordingly.

2.1 Geometric Distortion

Laser printers have the defects of the polygon mirror rotation speed fluctuation and the paper delivery mechanism which can lead to geometric distortion of the document reflected on the paper that each line of text has a slight slope rather than a complete parallel. Some scholars proposed the line slope (PTLS) and interval (PTLI) as features to calculate the distortion degree of printed lines under different printers [1]. In the latest

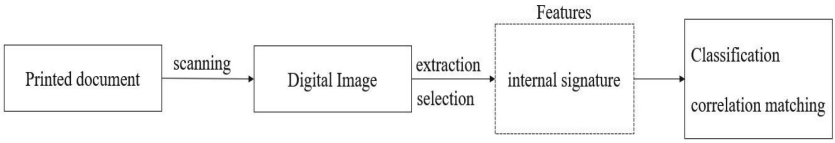

Fig. 2. Diagram of digital printer identification

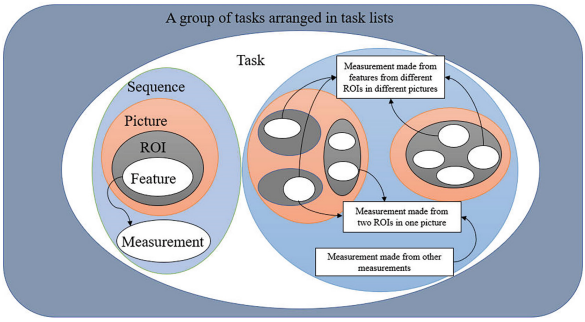

Fig. 3. Diagram of process of ImageXpert

study, Hardik Jain introduces a character-level distortion, using a uniform grid projection to create a recognition system to measure the distortion of a single character. Research shows that the geometric distortion produced by printers varies among different brands and even under the same model [2]. The geometric distortion characteristic is unrelated to the toner, and its performance is not affected by the change of toner, which can be used as the stable characteristics identified by the machine source.

2.2 Banding Frequencies

When the laser printer is working, the rotation speed of the photosensitive drum is not constant: when the rotation speed of the photosensitive drum drops slightly, the gray scale of the corresponding printed area will dim; when the rotation speed increases slightly, the corresponding area will brighten. In both cases, banding is reflected in the character image of the printed document, manifested as transverse uneven open and dark lines. [3] The rotation speed of the photosensitive drum is determined by the gears on both sides and different brands of laser printers own gears with different sizes and rotational speeds [4].

2.3 Image Quality

In the 20th and early 21st centuries, the quantification of digital image quality about office document is lack of study. With the development of an automatic image quality detection system: ImageXpert software, it gradually shows the potential of realizing source printer identification through printing quality analysis. Figure 3 shows the relationship between the elements of the IX system analysis and metrics process [5].

2.4 Texture Features

At the beginning of the century, many researchers examined texture of the same characters to determine whether two or more documents had the same source. The feature extraction algorithms often use classical statistical texture features, such as Gray-Level Co-occurrence Matrix (GLCM) and Discrete Wavelet Transformation (DWT) statistical

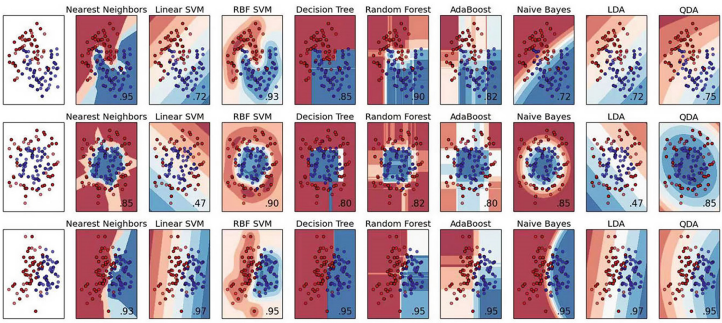

Fig. 4. Diagram of application of Big Data Algorithms for document traceability

texture features, and the average recognition rate of both is more than 90%. Tsai optimized the GLCM and DWT algorithm and SVM to improve the recognition accuracy of the source laser printer [6]. DWT is the decomposition of an image signal into a set of wavelets after shifting and scaling from the original wavelets. Wavelets are known as image microscopes in image processing because of their multi-resolution decomposition ability to strip away image information layer by layer. The means of stripping is through low-pass and high-pass filters. In this paper, the algorithmic model of DWT is described in detail from Tsai's point of view. We use this algorithmic model to decompose and output the image successfully. The three statistical characteristics of sdv (standard deviation), ske (skewness) and kur (kurtosis) are defined by Eqs. 1, 2 and 3.

$$sdv = \sqrt{\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{N}(Img(i,j) - \overline{Img(i,j)})^2} \tag{1}$$

$$ske = \frac{1}{N \cdot sdv^3}\sum_{i=0}^{N}\sum_{j=0}^{N}(Img(i,j)\overline{Img(i,j)})^3 \tag{2}$$

$$kur = \frac{1}{N \cdot sdv^4}\sum_{i=0}^{N}\sum_{j=0}^{N}(Img(i,j)\overline{Img(i,j)})^4 \tag{3}$$

The local binary mode (Local Binary Patterns, LBP) and dual-factor analysis algorithm proposed by scholars later have very high accuracy.

All these features can be borrowed from big data algorithms. As an example, for different research subjects, the accuracy of distinguishing printed documents can be calculated based on the random forest algorithm to rank the importance of feature parameters. Figure 4 shows these findings provide a new and feasible way to examine laser-printed document.

3 Conclusion

Deep learning is a part of machine learning. By learning the internal laws of sample data to represent attribute categories and features, which has achieved remarkable results in text and image recognition. Some scholars have introduced deep learning to document

traceability, which shows a very high accuracy in classification, and the efficiency is not lower than the mode of manual feature extraction. Document traceability can be regarded as the interdisciplinary application of deep learning in text and images. The various uncertainties that affect printing may be converted into stable identifiable features through deep learning model training, which needs further exploration. In-depth application of deep learning will be an important development trend of the future document identification work.

References

1. Hao J, Kong X, Shang S. Printer identification using page geometric distortion on text lines[C]//2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). IEEE, 2015: 856–860.
2. Jain H, Joshi S, Gupta G, et al. Passive classification of source printer using text-line-level geometric distortion signatures from scanned images of printed documents[J]. Multimedia Tools and Applications, 2020, 79(11-12): 7377–7400.
3. Mikkilineni A K, Chiang P J, Ali G N, et al. Printer identification based on graylevel co-occurrence features for security and forensic applications[C]//Security, steganography, and watermarking of multimedia contents vii. SPIE, 2005, 5681: 430–440.
4. Oliver J, Chen J. Use of signature analysis to discriminate digital printing technologies[C]//NIP & Digital Fabrication Conference. Society for Imaging Science and Technology, 2002, 2002(1): 218–222.
5. Li B, Qu Y, Wang C, et al. Studies about quantitative examination of laser printed documents based on image physical metric[J]. Forensic Science International, 2021, 318: 110–119.
6. Tsai M J, Yin J S, Yuadi I, et al. Digital forensics of printed source identification for Chinese characters[J]. Multimedia tools and applications, 2014, 73: 2129–2155.