Journal of Visual Communication and Image Representation 53 (2018) 257–272

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Connecting the dots: Toward accountable machine-learning printer attribution methods☆

Luiz C. Navarro[a,*], Alexandre K.W. Navarro[b], Anderson Rocha[a], Ricardo Dahab[a]

[a] Institute of Computing – University of Campinas (Unicamp), Campinas, SP, Brazil
[b] Engineering Department – University of Cambridge Cambridge, UK

ARTICLE INFO

Keywords:
Accountable machine learning
Digital forensics
Source printer attribution
Feature back-projection
Feature mapping
Feature importance

ABSTRACT

Digital forensics is rapidly evolving as a direct consequence of the adoption of machine-learning methods allied with ever-growing amounts of data. Despite the fact that these methods yield more consistent and accurate results, they may face adoption hindrances in practice if their produced results are absent in a human-interpretable form. In this paper, we exemplify how human-interpretable (a.k.a., accountable) extensions can enhance existing algorithms to aid human experts, by introducing a new method for the source printer attribution problem. We leverage the recently proposed Convolutional Texture Gradient Filter (CTGF) algorithm's ability to capture local printing imperfections to introduce a new method that maps and highlights important attribution features directly onto the investigated printed document. Supported by Random Forest classifiers, we isolate and rank features that are pivotal for differentiating a printer from others, and back-project those features onto the investigated document, giving analysts further evidence about the attribution process.

1. Introduction & related work

Over the past decade, machine-learning methods have attained substantial importance in the field of digital forensics. Such techniques have been successfully applied in image and video forgery detection [1], predatory conversation identification in social media [2], face expression recognition [3], attribution of documents to their source printers [4], among others. Of particular interest, attributing printers to documents is a problem of practical relevance to forensic analysts when connecting printers and available evidence (such as forged documents, fraudulent reports, and fake bills) apprehended in search-and-seizure operations. Despite advances in electronic documenting and digital signature algorithms, integrity enforcement, authentication and non-repudiation methods, our society continues to produce printed and physically-signed documents for official purposes, posing a constant need for source attribution techniques of questioned documents.

Traditional methods for printer attribution often use physical properties of the paper and ink to determine the association between printers and printed documents. Techniques may use, for example, an infra-red spectrometer equipped with a microscope [5], or reactive dyes, chemical assays, and microscopy [6]. Other works [7] rely on Fourier transform spectroscopy to perform spectral discrimination and detect counterfeit documents. The costs involved in these methods are substantial as they often require expensive made-to-order equipment and specialized personnel. Moreover, methods such as those involving chemical analyses can lead to unintended consequences such as damaging or destroying apprehended evidence.

An alternative to these approaches is to focus on printer defects of malfunctioning, as captured on the scanned images of a printed document and use image processing techniques to identify the document's source. Such methods are based on intrinsic signatures extracted from the document's image: texture characterization methods, as described by Chiang et al. [8,9], and geometric distortions on printed pages, as investigated by Shang et al. [10].

The banding effect, which encompasses cyclical space variations on the halftones distance and ink density, has also been the subject of investigation. Deviations produce such effects due to mechanical tolerances and defects of printer components such as axis eccentricity and motor drift. As such, they result in unique features for attributing a document to its printer. This technique has also eased cost-related concerns. Whereas previous experiments [11,12] required high-resolution document scanning for precise measurements, ranging from 1200 up to 8000 DPIs, more recent studies [13–15] have used 600-DPI

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: luiz.navarro@students.ic.unicamp.br (L.C. Navarro), akwn2@cam.ac.uk (A.K.W. Navarro), anderson.rocha@ic.unicamp.br (A. Rocha), rdahab@ic.unicamp.br (R. Dahab).

https://doi.org/10.1016/j.jvcir.2018.04.002
Received 6 June 2017; Received in revised form 20 March 2018; Accepted 11 April 2018
Available online 13 April 2018
1047-3203/

documents, which are less costly and easier to find in typical commercial scanners.

Another line of investigation for printer attribution considers geometric distortion methods to measure and correlate linear geometric distortions between the actual printed image and an expected ideal image, looking at characters extracted by OCR [16] or using estimated centroid variations from halftones [17,18].

Due to its power to represent intrinsic details of a printed document, textures have also been subject to research when attributing documents to their printers. Texture-based methods rely upon patterns across neighboring pixels created by imperfections such as toner ink melting and fixation problems; toner spread around letters and knurled contours. These methods benefit from image processing descriptors and machine-learning algorithms for the identification of discriminant patterns, further correlating them to the source printer of a document. One of the first authors to exploit texture features, Mikkilineni et al. [19] introduced the use of Gray-Level Co-occurrence Matrices (GLCMs) image descriptors over images of letters "e" extracted from 2400-DPI scanned documents allied with a simple KNN classifier for source printer attribution. The authors further improved their results by using Support Vector Machines (SVM) classifiers [20].

Ferreira et al. [4] experimented with 600-DPI images of scanned documents of 10 laser printers and created one of the first public standardized datasets in the area [21], which we also adopt in this work. The authors investigated the use of letters "e" extracted from documents as a whole and also rectangular non-overlapping regions cropped from documents usually containing several characters at once, referred to as frames. Various image descriptors including GLCM, Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HoG) were investigated. The authors also introduced one description method tailored for the attribution problem and referred to as Convolution Textures Gradient Filter (CTGF). Tsai et al. [22] extracted microscopy images using 300× magnification of characters and applied descriptors such as LBP and GLCM (among others) in four different alphabets (English, Arabic, Chinese, and Japanese).

Despite the fact that image-processing and machine-learning techniques outperform traditional chemical-based methods and are more appropriate in some setups, these algorithms often do not provide clear explanations as for why and how each document is attributed to a printer. The lack of human-interpretable evidence is particularly troublesome in forensic science, where decisions made by forensic analysts inform investigations and therefore may lead to legal implications. Moreover, the use of machine-learning methods that do not provide human-interpretable outputs in the forensic analysis may be legally inadmissible shortly. An example of the emergence of such restrictions, the European Union has voted in 2016 a resolution to be implemented by mid-2018 regarding the rights to human-interpretable explanations when decisions that can significantly affect its citizens are founded on machine-learning algorithms [23,24]. Other countries may shortly follow this example, and it hallmarks the need for digital forensics researchers to devise and develop human-interpretable machine-learning methods and hold the algorithms accountable.

Drawing on these insights, in this paper, we highlight how human-interpretable machine-learning methods can be derived from non-interpretable ones. We extend upon the Convolutional Texture Gradient Filter (CTGF) algorithm introduced by Ferreira et al. [4] to analyze, isolate and produce visible and interpretable features, giving rise to the CTGF-Map algorithm. The proposed method investigates the source of a document by finding the most discriminant features for attribution and by back-projecting those features directly onto the document, showing analysts the most relevant regions used in the process. Finally, we also discuss empirical results for this new method, tradeoffs between interpretable and non-interpretable counterparts of CTGF and the use of these techniques alongside other forensic processes.

2. Background concepts

The following sections present key concepts and methods used in this paper and discuss some properties of the techniques.

2.1. Accountable machine learning and explainability

In the last few years, there has been increasing concern and interest in accountable machine learning. There are new dedicated conferences and workshops on the subject such as the ones promoted by FAT/ML organization [25], the International Conference on Machine Learning (ICML) Workshop on Human Interpretability in Machine Learning (WHI) [26] and the AAAI W11 Workshop on Human-Aware Artificial Intelligence [27]. These events have been focusing on *fairness*, *accountability*, and *transparency* concepts for machine learning algorithms and applications. Governments and Non-governmental organizations are issuing policies, directives and best practices concerning the use of technology, and recently focusing on consequences and human rights related to decisions made by algorithms. Some examples include European Parliament recently approved General Data Protection Regulation (GDPR) [24,23] as well as studies from the Center for Democracy & Technology (CDT) [28]. Most concerns are due to the misuse and ethics of machine learning applications and the human rights of data privacy, but also on how to demonstrate decision results in a way that humans can understand. Most publications in this area address one or more principles exposed by a recent MIT Technology Review study [29] in MIT Technology: *Responsibility*, *Explainability*, *Accuracy*, *Auditability*, and *Fairness*. Explainability is defined in FAT/ML organization Principles for Accountable Algorithms [30] as: "[To] ensure that algorithmic decisions, as well as any data driving those decisions, can be explained to end-users and other stakeholders in non-technical terms".

With this backdrop, the primary objective of our work herein is to provide forensic analysts with a printer attribution method that fits in the explainability concept above. Questioned documents source attribution is usually part of proof and evidence presented in court trials by experts to court members, who are not familiar with machine-learning methods, but they can more easily understand physical explanations with visual evidence presentation. Handwriting analysis, for example, is an old and routine technique to visually present in courts for manually-written letters, memos, and most common for signatures. In our case, we aim at providing experts with a machine-learning method, which can classify questioned documents with high accuracy and precision, and also can show, **visually**, which regions on the image were used by the algorithms to identify the source of the document in the decision-making process thereof.

2.2. CTGF image descriptor

Convolutional Texture Gradient Filter (CTGF) [4] is a computational forensics method for describing documents regarding features that can be used by machine-learning algorithms to attribute a document to its source printer. The original presentation for the CTGF method [4] followed an empirical standpoint. In this section, we review this method while providing an alternative explanation based on the underlying physical principles of laser printers as well as a probabilistic interpretation of the descriptor.

The core physical principles used by laser printers to generate documents are electrostatics, photonics, and thermal curing. Electrostatics is used in the first stage of printing a document when electrical charges ink powder is placed on the printer optical charged drum (OPC) that will carry out the printing (see the OPC drum and process steps in Fig. 1). OPC drum is uniformly charged in steps A and B; then for the printer to differentiate from blank and printed areas, the OPC drum needs to be anisotropically charged. This anisotropy is