Printer Profiling for Forensics and Ballistics

Eric Kee
Department of Computer Science
Dartmouth College
Hanover, NH 03755
erickee@cs.dartmouth.edu

Hany Farid
Department of Computer Science
Dartmouth College
Hanover, NH 03755
farid@cs.dartmouth.edu

ABSTRACT

We describe a technique for authenticating printed and scanned text documents. This technique works by modeling the degradation in a document caused by printing. The resulting printer profile is then used to detect inconsistencies across a document, and for ballistic purposes – that of linking a document to a printer.

Categories and Subject Descriptors

I.4 [**Image Processing**]: Miscellaneous

General Terms

Security

Keywords

Digital Tampering, Digital Forensics

1. INTRODUCTION

In early September of 2004, the CBS news show *60 Minutes* obtained documents critical of then U.S. Presidential hopeful George W. Bush's service in the National Guard. The documents were said to come from Bush's commander Colonel Jerry B. Killian's personal files. Shortly after the news story aired, on September 8th, questions began to emerge as to the authenticity of the documents. CBS correspondent and news anchor Dan Rather initially defended the authenticity of the documents. Approximately two weeks later, Andrew Heyward, President of CBS News said, "Based on what we now know, CBS News cannot prove that the documents are authentic, which is the only acceptable journalistic standard to justify using them in the report. We should not have used them. That was a mistake, which we deeply regret."

In an increasingly digital age, printed and scanned documents are becoming more common. As the above example illustrates, new tools are required to authenticate such documents. Previous work in this area employed watermarks [1, 8, 12] or visible security marks [16] that are applied at the time of printing. In order to contend with the majority of cases in which such secure markings are not available, we focus on passive techniques that do not require any watermarks or signatures. Several such passive techniques have previously been proposed [11, 2, 10, 9, 15, 7]. In [11], the authors use line width and raggedness, dot roundness and other features to identify printers by their make and model (i.e., printer identification). Delp and colleagues [2, 10, 9] have exploited printer banding artifacts for printer identification. And in [15], the authors employ invariant moments for printer identification.

Here we describe a complementary technique for the passive authentication of printed and scanned text documents. This technique works by modeling the geometric degradation in a document caused by printing. Instead of explicitly modeling the degradation caused by a printer, our printer profile consists of a linear basis generated from a set of degraded characters (e.g. the letter *e*). This basis representation embodies the printer degradation. Unlike previous work, we exploit this printer profile both for printer identification and to detect local tampering in a document.

2. METHODS

We describe the construction of a printer profile and its use in forensics and ballistics. This technique can be divided into three main parts: (1) a series of fairly standard pre-processing steps to prepare the printed documents for analysis; (2) the construction of a printer profile which embodies the distortions introduced by a printer; and (3) the use of the printer profile for forensics and ballistics.

2.1 Pre-processing

A printed document is first digitally scanned and saved in an uncompressed format. Each page of a document is then processed in the same manner, as described below.

In this first stage, multiple copies of the same character are located in a scanned document. To do so, a user first selects a bounding box around a character of interest to serve as a template. As described in detail in Appendix A, a correlation-based approach is taken to extract the spatial location of matching characters.

In order to minimize the effect of luminance variations across printers, the intensity histograms of the characters are matched as follows. A random set of characters is selected, and their intensity histograms averaged to create a reference histogram. Each character's intensity histogram is then

MM & Sec '08 Oxford, UK

matched to this reference histogram [1]

A single character is then selected as the reference character. Each character is brought into spatial alignment with this reference character using a coarse-to-fine differential registration technique. This transformation is limited to a rigid-body transformation consisting of translation, scaling, and rotation. In order to contend with any large-scale translations, each character's center of mass is first brought into alignment with the reference character's center of mass. This transformation consists of an integer-based translation only. Then, assuming that the characters have the same intensity profile, the rigid-body transformation between each character, $f(\cdot)$, and the reference, $f_r(\cdot)$, is given by:

$$f_r(x,y) \quad = \quad f(\hat{x},\hat{y}), \tag{1}$$

where:

$$\begin{pmatrix}\hat{x}\\ \hat{y}\end{pmatrix} \quad = \quad \begin{pmatrix} m_1 & m_2 \\ -m_2 & m_1 \end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix} + \begin{pmatrix}m_3\\ m_4\end{pmatrix}, \tag{2}$$

and where m_1 and m_2 embody the scaling and rotation parameters, and m_3 and m_4 are the translation parameters. These parameters can be solved for using standard differential techniques [6, 4, 14]. Briefly, a quadratic error function is established in terms of m_1–m_4, that embodies the relationship of Equation (2). This error function is rewritten in terms of a truncated Taylor series expansion in order to linearize the error function in terms of m_1–m_4. Standard least-squares estimation is then used to solve for these transformation parameters. This entire process is embedded within a coarse-to-fine architecture in order to contend with the limited scope of the differential operators.

2.2 Printer Profile

With all of the characters properly aligned, we seek to construct a profile of the degradation introduced by the printer. Because the nature of this degradation can be highly complex [3], we take a data-driven approach to characterizing this degradation. Specifically, a principal components analysis (PCA) [5] is applied to the aligned characters to create a new linear basis that embodies the printer degradation. Briefly, each of m zero-meaned characters of size $n \times n$ are packed into the columns of a $n^2 \times m$ matrix D. An eigen-decomposition is performed on the covariance matrix $C = DD^T$, from which the top eigenvalue eigenvectors, $\vec{e}_i$ are extracted to form the desired linear basis. From a practical point of view, when $m < n^2$, it is computationally more efficient to compute the eigenvectors of $C' = D^T D$, from which the desired principal components are given by $D\vec{e'}_i$, where $\vec{e'}_i$ are the eigenvectors of C'. Although this approach limits us to a linear basis, this degradation model is easy to compute and is able to capture fairly complex degradations that are not easily embodied by a low-dimensional parametric model.

The final profile consists of both the mean character, $\vec{\mu}$, (subtracted off prior to the PCA) and the top p eigenvalue eigenvectors $\vec{e}_i$, $i \in [1,p]$. Note that this printer profile is constructed on a per character basis, e.g., for the letter e of the same font and size.

[1] An image I_1 is histogram matched to an image I_2 by passing I_1 through a look-up table consisting of the cumulative distribution function of I_2.

2.3 Forensics

In a forensics setting, we are interested in determining if part of a document has been manipulated by, for example, splicing in portions from a different document, or digitally editing a previously printed and scanned document and then printing the result.

A printed and scanned document is processed as described in the previous two sections to construct a printer profile $\mathcal{P} = \{\vec{\mu}, \vec{e}_1, ..., \vec{e}_p\}$. Each character $\vec{c}_j$ is then projected onto each basis vector:

$$\alpha_{ji} \quad = \quad (\vec{c}_j - \vec{\mu})^T \vec{e}_i, \tag{3}$$

where a character $\vec{c}_j$ is a 2-D grayscale image reshaped into vector form. The basis weights for each character $\vec{c}_j$ are denoted as $\vec{\alpha}_j = \begin{pmatrix}\alpha_{j1} & \alpha_{j2} & \dots & \alpha_{jp}\end{pmatrix}$. With the assumption that tampering will disturb the basis representation (the weights), we subject the weights $\vec{\alpha}_j$ to a normalized graph cut partitioning [13] to determine if they form distinct clusters.

Briefly, a weighted undirected graph $G = (V, E)$ is constructed with vertices V and edges E. Each vertex corresponds to a character $\vec{c}_j$ with $j \in [1, m]$, and the weight on each edge connecting vertices k and l is given by:

$$w(k,l) \quad = \quad \exp\left(-\frac{d_\alpha^2(\vec{\alpha}_k, \vec{\alpha}_l)}{\sigma_\alpha^2}\right) \cdot \exp\left(-\frac{d_c^2(k,l)}{\sigma_c^2}\right), \tag{4}$$

where $d_\alpha(\cdot)$ is the Mahalanobis distance defined as:

$$d_\alpha(\vec{x}, \vec{y}) \quad = \quad \sqrt{(\vec{x}-\vec{y})\Sigma^{-1}(\vec{x}-\vec{y})}, \tag{5}$$

and where Σ is the covariance matrix. The second term in the weighting function, $d_c(\cdot)$, is the distance between two characters, defined as the linear distance in scan line order (i.e., in the order in which the text is read, left to right, top to bottom). This additional term makes it more likely for characters in close proximity to be grouped together.

The cost of splitting the graph G into two disjoint subgraphs A and B is the sum of the weights between all vertices in A and B, termed the cut:

$$\text{cut}(A,B) \quad = \quad \sum_{a \in A}\sum_{b \in B} w(a,b) \tag{6}$$

When minimizing this cost function, there is a natural tendency to simply cut a small number of low-cost edges. A normalized cut was introduced to remove this bias [13]. This cost function normalizes the cut by the total cost of all edges in the entire graph. As a result, small partitions are penalized. Solving for the optimal normalized cut is NP-complete. Formulation as a real-valued problem, however, yields an efficient and approximate discrete-valued solution [13]. Define W to be a $m \times m$ weighting matrix such that $W_{k,l} = w(k,l)$, and D to be a $m \times m$ diagonal matrix whose k^{th} element on the diagonal is $\sum_l w(k,l)$. Solve the generalized eigenvector problem $(D - W)\vec{e} = \lambda D\vec{e}$, for the eigenvector $\vec{e}$ with the second smallest eigenvalue λ. Let the sign of each component of $\vec{e}$ (corresponding to each vertex of G) define the membership of that vertex into one of two sets, A or B – for example, vertices with corresponding negative components are assigned to A and vertices with corresponding positive components are assigned to B.

This approach provides both a partitioning of the characters and the cost associated with this partitioning. If the partitioning cost is low, then it is likely that the characters

are distinct and hence a subset of characters are inconsistent with the printer profile.

2.4 Ballistics

In a ballistics setting, we are interested in determining if a document was printed from a specific printer. A printer profile is generated from a printer to determine if the document in question was printed from this printer. We assume that the printer profile is constructed from the same font family and size as the document to be analyzed. The printer profile is generated as described above to yield $\mathcal{P} = \{\vec{\mu}, \vec{e}_1, ..., \vec{e}_p\}$. Each character, $\vec{c}_j$, in the document is first aligned to the reference character used to construct the printer profile, Section 2.1. Each aligned character is then projected onto each basis vector:

$$\alpha_{ji} = (\vec{c}_j - \vec{\mu})^T \vec{e}_i, \quad (7)$$

where a character $\vec{c}_j$ is a 2-D grayscale image reshaped into vector form. The basis weights for each character $\vec{c}_j$ are denoted as $\vec{\alpha}_j = \begin{pmatrix} \alpha_{j1} & \alpha_{j2} & \dots & \alpha_{jp} \end{pmatrix}$. If the document originated from the printer with profile $\mathcal{P}$, then we expect that the profile will afford an accurate representation of the characters $\vec{c}_j$. As such, the reconstruction error between the actual character $\vec{c}_j$ and the new basis representation is computed to determine the suitability of the printer profile. Specifically, the reconstructed character is given by:

$$\vec{r}_j = \vec{\mu} + \sum_{i=1}^{p} \alpha_{ji} \vec{e}_i, \quad (8)$$

and the reconstruction error is given by:

$$E_j = \sqrt{(\vec{c}_j - \vec{r}_j)^T (\vec{c}_j - \vec{r}_j)}. \quad (9)$$

As described below, this reconstruction error is then used to determine if a document originated from a printer of a specific make and model.

3. RESULTS

We show the efficacy of the printer profile described above in both a forensic and ballistic setting.

3.1 Forensics

A simple fake document was created by printing a page of 12-pt Courier *e*'s on a HP LaserJet 4350 and a Xerox Phaser 5500DN. Each of these documents were then scanned at 600dpi and combined, with the top half composed of the HP document and the bottom half composed of the Xerox document. This document was then printed on a HP LaserJet 4300 and re-scanned at 300dpi. Shown in Figure 1 is a magnified view of a portion of this fake document (the top row was printed on the HP printer, and the bottom row on the Xerox printer). A printer profile was constructed from this document (2280 copies of the letter *e*), as described in Section 2.2. The printer profile consisted of the mean $\vec{\mu}$ and the maximal eigenvalue eigenvector, $\vec{e}_1$. As described in Section 2.3, the characters were then subjected to the clustering based on their profile representation, Equation (3). The parameters for the weighting functions, Equation (4), were $\sigma_\alpha = 0.5$ and σ_c equal to 100 times the width of a line of text (in pixels). These parameters were held fixed for each example described below. Shown in Figure 1 are the classification results that clearly reveal the differences between the top and bottom halves of the document. The cost of this clustering was 0.08.

A second similar fake was constructed, where this time each of the original documents were printed on different Xerox Phaser 5500DN printers. These documents were scanned and combined as before and printed on a HP LaserJet 4300 printer. Shown in Figure 2 is a magnified view of a portion of this fake document (the top and bottom row were each printed on different printers). A printer profile was constructed from this document, and used to classify the entire document. Shown in Figure 2 are the classification results that reveal the inability of the printer profile to discriminate a forgery constructed from different printers of the same make and model. The cost of this clustering was 0.80, an order of magnitude larger than the previous example. This high cost reveals that this clustering is not the result of tampering.

Shown in the left panel of Figure 3 is a page from *The Tale of Two Cities*, where the top half was printed on a HP LaserJet 4350 and the bottom half was printed on a Xerox Phaser 5500DN. These documents were scanned and combined as described above and printed on a HP LaserJet 4300 printer. A printer profile was created from 200 copies of the letter *a*. Shown in Figure 3 are the classification results, which correctly classify the top and bottom halves of the document as originating from different printers. The cost of this clustering was 0.05. Shown in the right panel of Figure 3 is a similar example, where only the second paragraph was printed on the HP printer. In this example, 201 copies of the letter *a* were used in the construction of the profile (the correlation-based character extraction returns slightly different results depending on the template character). The cost of this clustering was 0.03. In this example, we see that even relatively small regions of tampering can be detected. And lastly, shown in Figure 4 is an example where the first two lines, and the last sentence (starting with "It is likely enough" and ending with "of the Revolution") originated from a different printer than the rest of the document. Only one letter on the third line ("was") was mis-classified. The cost of this clustering was 0.28, slightly higher than the previous costs (due to the distance between the doctored regions), but still much less than the results in which no tampering were present.

In each of the above examples, the printer profile consisted of the mean $\vec{\mu}$ and only the maximal eigenvalue eigenvector, $\vec{e}_1$. We have found that a printer profile consisting of the top two or three eigenvectors contributes little to the overall accuracy. We have also found that performance degrades significantly with a profile consisting of four or more eigenvectors, suggesting that the higher-order terms in the PCA are not representative of distinct printer degradation features.

Combined, these results suggest that the printer profile is effective in detecting fakes composed of parts initially printed on different printers (in terms of make and model), but not sufficiently descriptive so as to discriminate between printers of the same make and model.

3.2 Ballistics

We constructed a printer profile for each of eight printers, Figure 5. In addition to choosing printers of different makes and models, we chose several printers of the same make and

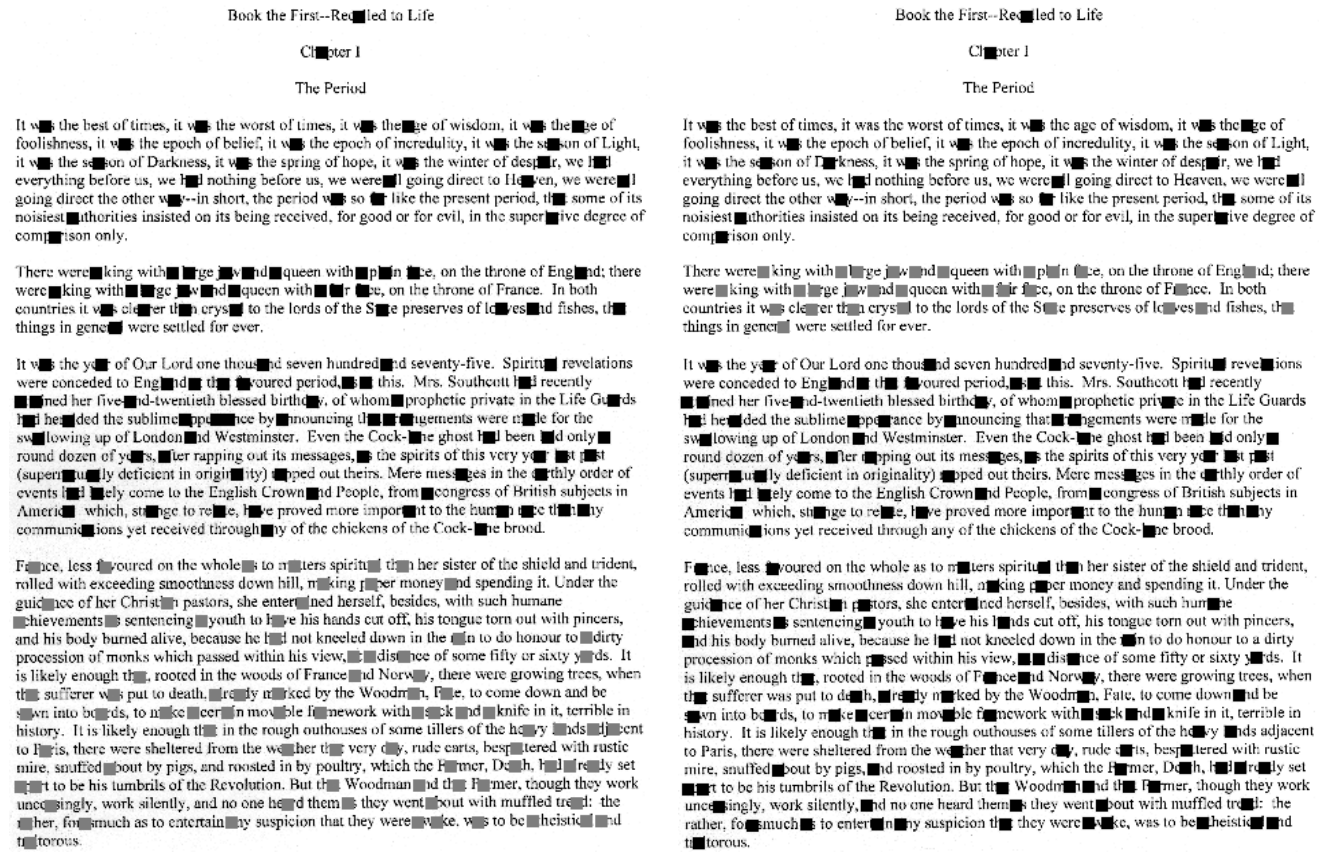


Figure 3: Shown are the clustering results for a printed page of *The Tale of Two Cities*. Each small square region denotes a single letter a, and the color coding (gray/black) denotes the cluster assignment. In the left panel, the top and bottom halves were printed on different printers, and in the right panel, the second paragraph was printed on a different printer from the rest of the page. The cost of these clusterings was 0.05 **(left) and** 0.03 **(right).**

model. These printers allowed us to test the efficacy of our printer profile across and within printers of different and the same make and model.

For each printer, 10 pages consisting entirely of 12-pt Courier e's were printed and scanned at 300 dpi. Each of the resulting $22,400$ characters per printer was of size 90×90 pixels. Each character was processed as described in Section 2.1. A printer profile was constructed for each printer, Section 2.2, from a random sampling of $2,000$ characters per printer (200 characters per page). The printer profile consisted of the mean $\vec{\mu}$ and the maximal eigenvalue eigenvector, $\vec{e}_1$. The reconstruction error, Equation (9), was then computed for each character against each printer profile, Figure 6.

To test the efficacy of each profile in identifying a document, the remaining $20,400$ characters per printer were randomly partitioned into 408 subsets of 50 characters each. A maximum a posteriori estimator (MAP) was then used to classify each set of 50 characters. The MAP estimator maximizes the posterior probability:

$$P(\mathcal{P}_i \mid \vec{E}) \quad = \quad \frac{P(\vec{E} \mid \mathcal{P}_i) P(\mathcal{P}_i)}{\sum_j P(\vec{E} \mid \mathcal{P}_j)}, \tag{10}$$

where $\vec{E}$ is the mean reconstruction error, Equation (9), of a character set against each of eight printer profiles, $\mathcal{P}_1, \ldots, \mathcal{P}_8$. One half of the 408 character sets were used to estimate the likelihood probability, $P(\vec{E} \mid \mathcal{P}_i)$, the prior probability $P(\mathcal{P}_i)$ was assumed to be uniform, and it was assumed that the reconstruction errors are independent, that is:

$$P(\vec{E} \mid \mathcal{P}_i) = P(E_1 \mid \mathcal{P}_i) P(E_2 \mid \mathcal{P}_i) ... P(E_8 \mid \mathcal{P}_i). \tag{11}$$

Shown in Figure 7 are the classification results for the remaining 204 character sets. Note that in every case, the classification is nearly perfect.

It may appear that our printer profile is sufficiently descriptive to discriminate between printers of the same make and model. Note, however, that the printers differed in their toner levels, Figure 5. We wondered if these toner levels could be responsible for the nearly perfect classification. To this end, we constructed printer profiles for the Xerox Phaser 5500DN (printers 4, 5, and 6 in Figure 5) with different toner levels (81%, 75%, and 25%, respectively). We will refer to these printers as $4a$, $5a$, and $6a$. Shown in Figure 8 are the classification accuracies for these three printers. Note that in this case, there is confusion between printers of the same make and model, and that this confusion roughly follows the toner levels. Specifically, printer $4a$ with toner level 81% is

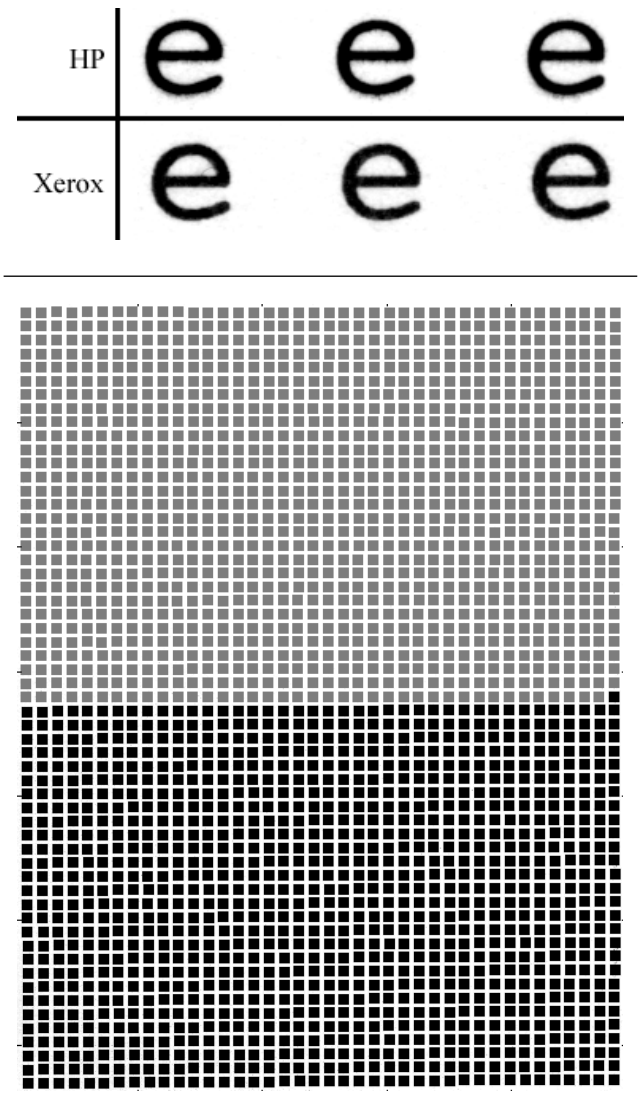


Figure 1: Shown on top is a magnified view of a portion of a document printed on a HP LaserJet (top row) and a Xerox Phaser (bottom row). Shown below is the clustering results for the entire document. Each small square region denotes a single character, and the color coding (gray/black) denotes the cluster assignment. Note that the top- and bottom-half of this document are correctly classified as originating from different printers. The cost of this clustering was 0.08.

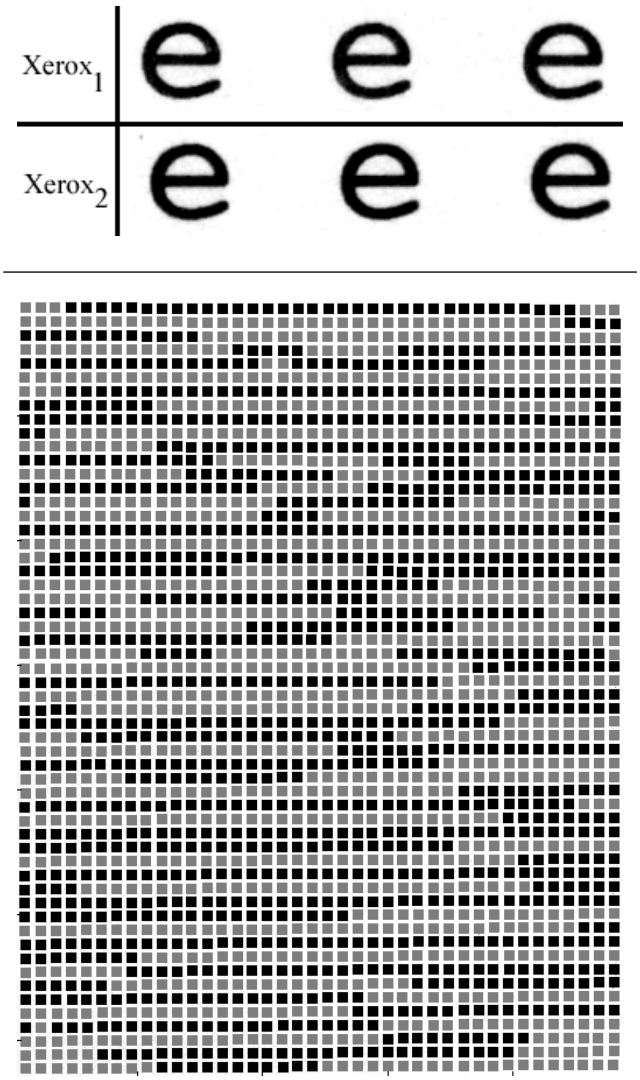
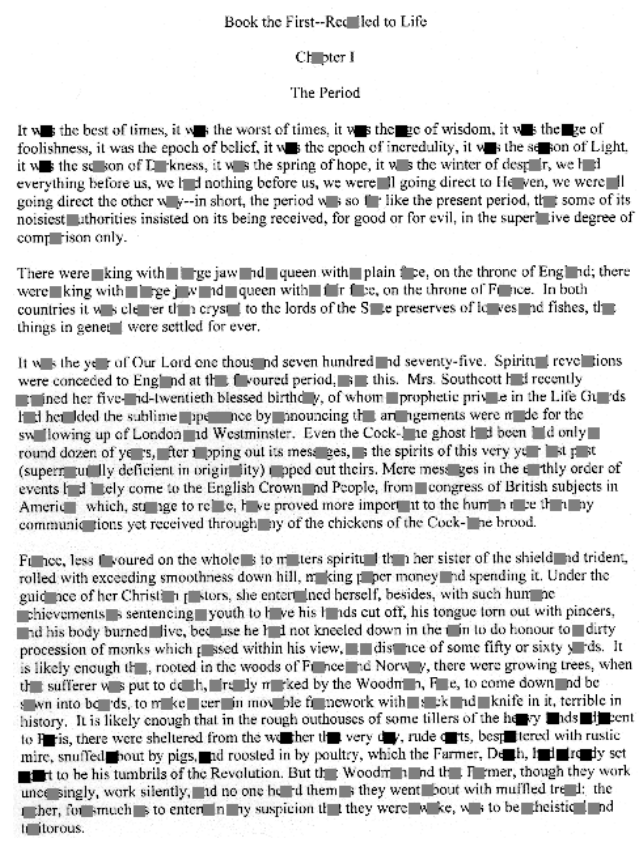


Figure 2: Shown on top is a magnified view of a portion of a document printed on a Xerox Phaser (top row) and a different printer of the same make/model (bottom row). Shown below is the clustering results for the entire document. Each small square region denotes a single character, and the color coding (gray/black) denotes the cluster assignment. Note that in this case, we are unable to detect a fake composed of text originating from printers of the same make and model. The cost of this clustering was 0.80.



Book the First--Recalled to Life

Chapter I

The Period

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way--in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face, on the throne of England; there were a king with a large jaw and a queen with a fair face, on the throne of France. In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes, that things in general were settled for ever.

It was the year of Our Lord one thousand seven hundred and seventy-five. Spiritual revelations were conceded to England at that favoured period, as at this. Mrs. Southcott had recently attained her five-and-twentieth blessed birthday, of whom a prophetic privateer in the Life Guards had heralded the sublime appearance by announcing that arrangements were made for the swallowing up of London and Westminster. Even the Cock-lane ghost had been laid only a round dozen of years, after rapping out its messages, as the spirits of this very year last past (supernaturally deficient in originality) rapped out theirs. Mere messages in the earthly order of events had lately come to the English Crown and People, from a congress of British subjects in America: which, strange to relate, have proved more important to the human race than any communications yet received through any of the chickens of the Cock-lane brood.

France, less favoured on the whole as to matters spiritual than her sister of the shield and trident, rolled with exceeding smoothness down hill, making paper money and spending it. Under the guidance of her Christian pastors, she entertained herself, besides, with such humane achievements as sentencing a youth to have his hands cut off, his tongue torn out with pincers, and his body burned alive, because he had not kneeled down in the rain to do honour to a dirty procession of monks which passed within his view, at a distance of some fifty or sixty yards. It is likely enough that, rooted in the woods of France and Norway, there were growing trees, when the sufferer was put to death, already marked by the Woodman, Fate, to come down and be sawn into boards, to make a certain movable framework with a sack and knife in it, terrible in history. It is likely enough that in the rough outhouses of some tillers of the heavy lands adjacent to Paris, there were sheltered from the weather that very day, rude carts, bespattered with rustic mire, snuffed about by pigs, and roosted in by poultry, which the Farmer, Death, had already set apart to be his tumbrils of the Revolution. But that Woodman and that Farmer, though they work unceasingly, work silently, and no one hears them as they go about with muffled tread: the more, forasmuch as to entertain any suspicion that they are awake, was to be hideously and fatally unfavourable.

Figure 4: Shown are the clustering results for a printed page of *The Tale of Two Cities*. Each small square region denotes a single letter *a*, and the color coding (gray/black) denotes the cluster assignment. The first two lines, and the last sentence originated from a different printer than the rest of the document. The cost of this clustering was 0.28**.**

characterized as printer 4 (94%) or printer 6 (60%), but not printer 5 with significantly lower toner (31%). Similarly, printer 6*a* with toner level 25% is characterized as printer 5 with a similar toner level of 31%. Nevertheless, even with the dependence on toner level, the three printers ($4a - 6a$) were correctly classified in terms of their make and model.

In each of the above examples, the printer profile consisted of the mean $\vec{\mu}$ and only the maximal eigenvalue eigenvector, $\vec{e}_1$. We wondered if a profile consisting of the mean and the second or third eigenvector would afford better discriminability between printers of the same make and model or be less sensitive to toner level. The intuition being that the first eigenvector captures general degradations and the higher-order terms capture more idiosyncratic properties of the printer. Printer profiles based on the second eigenvector $\vec{e}_2$, however produced nearly identical results to those based on the first eigenvector. It remains to be seen if even higher-order terms will afford some discrimination, but based on our forensic results, we think this unlikely.

Combined, these results suggests that our printer profile is sufficiently descriptive to discriminate between printers of different make and model. Because the printer profile depends on toner level, a profile will have to be built for different toner levels.

$\mathcal{P}$	Make/Model	Toner
1	HP LaserJet 4300	72%
2	HP LaserJet 4350	26%
3	HP LaserJet 4350	2%
4	Xerox Phaser 5500DN	94%
5	Xerox Phaser 5500DN	31%
6	Xerox Phaser 5500DN	60%
7	Xerox Phaser 8550DP	-
8	Xerox Phaser 8550DP	-

Figure 5: Eight printers with corresponding toner levels (the last two printers employ solid-ink, as opposed to dry toner, technology).

4. DISCUSSION

We have described a technique for modeling geometric degradations caused by a printer. These degradations, measured on a per character basis, are modeled by building a linear basis from a set of commonly occurring letters on a printed page. In a forensics setting, the representation in this new basis is used to detect inconsistencies in printer degradations across a page. For the purposes of ballistics, the accuracy with which a document can be modeled with a given printer's basis representation is used to determine provenance. This approach requires no specialized scanning hardware – a 300 dpi scan provides sufficient resolution.

We have shown the efficacy of our approach in both forensic and ballistic applications. In each case, our technique can distinguish between printers of different make and model, but not between printers of the same make and model. In a ballistic setting, the printer profiles depend on the printer toner levels, and hence multiple profiles will have to be constructed for different toner levels. We are currently investigating how to remove this dependency and how or if this approach can be made more sensitive so as to distinguish between printers of the same make and model. We are also studying if a printer profile is stable over time, or if any changes in the profile can be modeled (which would allow us to determine when a document was printed).

Acknowledgment

This work was supported by a gift from Adobe Systems, Inc., a gift from Microsoft, Inc., a grant from the National Science Foundation (CNS-0708209), a grant from the U.S. Air Force (FA8750-06-C-0011), and by the Institute for Security Technology Studies at Dartmouth College under grants from the Bureau of Justice Assistance (2005-DD-BX-1091) and the U.S. Department of Homeland Security (2006-CS-001-000001). Points of view or opinions in this document are those of the author and do not represent the official position or policies of the U.S. Department of Justice, the U.S. Department of Homeland Security, or any other sponsor.

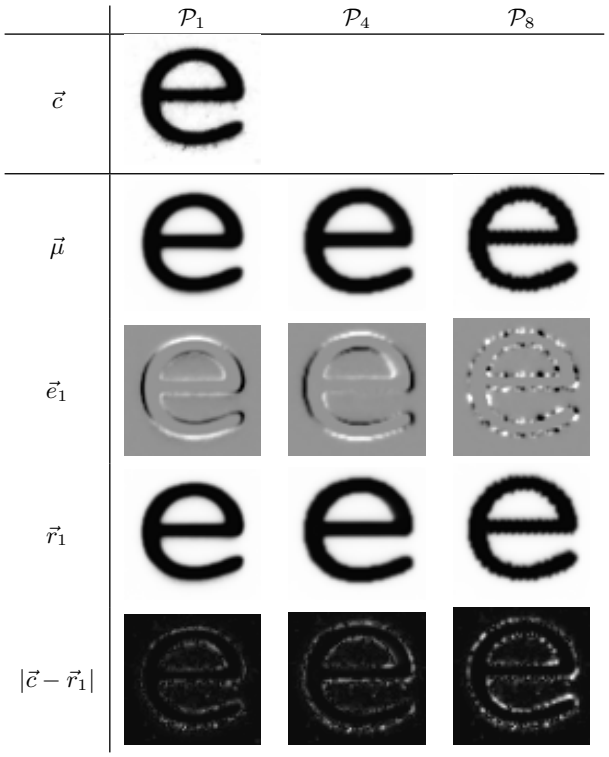


Figure 6: Shown in the top row is the letter *e* printed on printer $\mathcal{P}_1$ (see Figure 5). Shown in second and third rows are the printer profiles for printers $\mathcal{P}_1$, $\mathcal{P}_4$, and $\mathcal{P}_8$, consisting of the mean $\vec{\mu}$ and maximal eigenvalue eigenvector $\vec{e}_1$. Shown in the fourth row is the reconstructed character from each printer profile. And shown in the last row is the reconstruction error, which in this case is minimal for printer $\mathcal{P}_1$.

	1	2	3	4	5	6	7	8
1	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	99.0	1.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.5	99.0	0.5	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100

Figure 7: Ballistic classification results presented as a confusion matrix. Each row/column corresponds to a printer (see Figure 5). A value of 100% on the diagonal corresponds to perfect classification, and non-zero values off the diagonal correspond to misclassification. See also Figure 8.

	1	2	3	4	5	6	7	8
4a	0.0	0.0	0.0	87.0	0.0	13.0	0.0	0.0
5a	0.0	0.0	0.0	0.0	7.2	92.8	0.0	0.0
6a	0.0	0.0	0.0	0.0	91.9	8.1	0.0	0.0

Figure 8: Ballistic classification results presented as a confusion matrix. Each column corresponds to a printer (see Figure 5), and each row corresponds to a printer with different toner levels. See also Figure 7.

5. REFERENCES

[1] G. Ali, P.-J. Chiang, A. K. Mikkilineni, J. P. Allebach, G. T. Chiu, and E. J. Delp. Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices. In *IS&T's NIP19: International Conference on Digital Printing Technologies*, pages 511–515, 2003.

[2] G. N. Ali, P.-J. Chiang, A. K. Mikkilineni, G. T.-C. Chiu, E. J. Delp, and J. P. Allebach. Application of principal components analysis and gaussian mixture models to printer identification. In *International Conference on Digital Printing Technologies*, pages 301–305, 2004.

[3] H. Baird. The state of the art of document image degradation modeling. In *International Workshop on Document Analysis Systems*, Rio de Janeiro, 2000.

[4] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, Feb. 1994.

[5] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[6] B. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[7] N. Khanna, A. Mikkilineni, A. Martone, G. N. Ali, G.-C. Chiu, J. P. Allebach, and E. Delp. A survey of forensic characterization methods for physical devices. *Digital Investigation*, 3S:S17–S28, 2006.

[8] A. K. Mikkilineni, G. N. Ali, P.-J. Chiang, G. T. Chiu, J. P. Allebach, and E. J. Delp. Signature-embedding in printed documents for security and forensic applications. In *SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*, pages 455–466, 2004.

[9] A. K. Mikkilineni, O. Arslan, P.-J. Chiang, R. M. Kumontoy, J. P. Allebach, G. T.-C. Chiu, and E. J. Delp. Printer forensics using svm techniques. In *IS&T's NIP21: International Conference on Digital Printing Technologies*, pages 223–226, 2005.

[10] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp. Printer identification based on graylevel co-occurrence features for security and forensic applications. In *Security, Steganography, and Watermarking of Multimedia Contents*, pages 430–440, 2005.

[11] J. Oliver and J. Chen. Use of signature analysis to discriminate digital printing technologies. In *International Conference on Digital Printing Technologies*, 2002.

[12] J. Picard. Digital authentication with copy detection patterns. In *SPIE International Conference on Optical Security and Counterfeit Deterrence Techniques*, 2004.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[14] E. Simoncelli. *Handbook of Computer Vision and Applications*, chapter Bayesian Multi-scale Differential Optical Flow, pages 397–420. Academic Press, 1999.

[15] V. Talbot, P. Perrot, and C. Murie. Inkjet printing discrimination based on invariant moments. In *International Conference on Digital Printing Technologies*, 2006.

[16] B. Zhu, J. Wu, and M. Kankanhalli. Print signatures for document authentication. In *Conference on Computer and Communications Security*, Washington, DC, 2003.

Appendix A

Shown below is pseudo-code for the correlation-based extraction of characters from a scanned document. The function THRESHOLD converts a grayscale image into a binary image, where the threshold is selected to minimize the within region variance of the corresponding black and white image regions. The function OVERLAP returns true if the bounding box centered at position (i, j) overlaps with any other bounding box in the set of tuples C.

EXTRACT-CHARACTERS($f(x, y)$)

```
1  ▷ find all instances of template t(·) in image f(·)
2  ▷ f(x, y): scanned image
3  ▷ t(x, y): template character
4  ▷ N_x: width of f(x, y)
5  ▷ N_y: height of f(x, y)
6  ▷ τ: correlation threshold
7  f'(x, y) ← −2 × THRESHOLD(f(x, y)) + 1
8  t'(x, y) ← −2 × THRESHOLD(t(x, y)) + 1
9  d(x, y) ← f'(x, y) ⋆ t'(−x, −y) ▷ 2-D convolution
10 c = 1
11 for i = 1 : N_x
12       do for j = 1 : N_y
13             do if ( d(i, j) > τ N_x N_y & !OVERLAP(i, j, C) )
14                   do C(c, :) = [i, j]
15                      c = c + 1
16 return C ▷ coordinates of extracted characters
```