A Study of the Interaction of Paper Substrates on Printed Forensic Imaging

Guy Adams, Stephen Pollard, Steven Simske

HP Laboratories
HPL-2011-171

Keyword(s):
Security; Forensics; Model Matching; Document Fraud; High-resolution Imaging;

Abstract:
At the microscopic level, printing on a substrate exhibits imperfections that can be used as a unique identifier for labels, documents and other printed items. In previous work, we have demonstrated using these minute imperfections around a simple forensic mark such as a single printed character for robust authentication of the character with a low cost (and mobile) system. This approach allows for product authentication even when there is only minimal printing (e.g. on a small label or medallion), supporting a variety of secure document workflows. In this paper, we present an investigation on the influence that the substrate type has on the imperfections of the printing process that are used to derive the character "signature". We also make a comparison between two printing processes, dry electro photographic process (laser) and (thermal) inkjet. Understanding the sensitivity of this approach is important so that we know the limitations of the approach for document forensics.

External Posting Date: October 6, 2011 [Fulltext] Approved for External Publication
Internal Posting Date: October 6, 2011 [Fulltext]

A Study of the Interaction of Paper Substrates on Printed Forensic Imaging

Guy Adams
Hewlett-Packard Labs
Longdown Ave. Stoke Gifford
Bristol UK BS34 8QZ
guy.adams@hp.com

Stephen Pollard
Hewlett-Packard Labs
Longdown Ave. Stoke Gifford
Bristol UK BS34 8QZ
stephen.pollard@hp.com

Steven Simske
Hewlett-Packard Labs
3404 E. Harmony Rd.
Fort Collins CO 80528 USA
steven.simske@hp.com

ABSTRACT
At the microscopic level, printing on a substrate exhibits imperfections that can be used as a unique identifier for labels, documents and other printed items. In previous work, we have demonstrated using these minute imperfections around a simple forensic mark such as a single printed character for robust authentication of the character with a low cost (and mobile) system. This approach allows for product authentication even when there is only minimal printing (e.g. on a small label or medallion), supporting a variety of secure document workflows. In this paper, we present an investigation on the influence that the substrate type has on the imperfections of the printing process that are used to derive the character 'signature'. We also make a comparison between two printing processes, dry electro photographic process (laser) and (thermal) inkjet. Understanding the sensitivity of our methods to these factors is important so that we know the limitations of the approach for document forensics.

Categories and Subject Descriptors
I.4.1 [**Image Processing and Computer Vision**]: Digitization and Image Capture—*Scanning*. I.4.10 [**Image Processing and Computer Vision**]: Image Representation—*Statistical*. K.6.5 [**Management of Computing and Information Systems**]: Security and Protection—*Authentication*.

General Terms
Algorithms, Security

Keywords
Security, Forensics, Model Matching, Document Fraud, High-resolution Imaging.

1. INTRODUCTION
Forensic analysis of printed material including documents, packaging and labels, can be classified into two broad categories: 1) device forensics/ballistics [1]-[3] where a document (or set of documents) is analyzed to see if it was printed on a specific device or class of devices; 2) print forensics [4]-[7] wherein individual printed artifacts are uniquely identified. This second class, which is of interest here, allows the differentiation of individual instances of the same or highly similar documents - including high quality copies. For the majority of printing technologies unique properties result from the unrepeatable statistical aspects of the print process itself and the interaction with the underlying structural properties of the substrate material on which it is printed. Thus a forensic mark can be any form of glyph, character or printed shape of sufficient size to carry information to determine if the forensic mark under investigation is the exact same unique forensic mark that was previously printed [8]. In this way, print is used as a security mechanism preventing the counterfeit and copy of documents and product packaging [9].

We have previously [10] demonstrated a low-cost USB-powered mini-appliance capable of resolving 3.5 microns (7257 dpi) with 1:1 magnification. This is accomplished using a Dyson relay lens which comprises a single refractive surface in series with a mirror and a low cost 3M pixel CMOS image sensor. With a self-contained (white LED) illumination source, this Dyson relay CMOS imaging device (Dr CID) affords the capture of individual typed characters with printing "parasitics"—such as the absorbance of ink into the fibers of the substrate (e.g. paper, cardstock, etc.) along with the droplet "tails" that exhibit micro-random aberrations. The laser process has different characteristics such as toner particle size variation, splatter, charge bleed etc.

In [11]-[13] we adopted a model-based approach that separates the truly random part of the outline of the individual printed character, which we termed a model based signature profile (MBSP), from the shared shape-conveying component. We showed that the MBSP allows forensic (highly statistically significant) levels of authentication to be achieved (even in the case where multiple imaging modalities are employed). Furthermore for many forensic marks the MBSP is extracted in an order fixed by the frame of reference defined by the model. This in turn allowed the introduction (see [12], [13]) of a simple fixed order shape warp descriptor (which is typically less than a few hundred bits long) which is extracted from the MBSP. This is more compact than the MBSP and easier to manipulate and yet retains most of the forensic authentication power of the former.

In this paper, we present an investigation on the influence that the substrate type has on the imperfections of the printing process that are used to derive the character 'signature'. We also make a comparison between two printing processes, dry electro photographic process (laser) and (thermal) inkjet. Understanding the sensitivity of our methods to these factors is important so that we know the limitations of the approach for document forensics. It is also important so that we can determine how many printed characters are required to forensically identify a document. We also go beyond previous work and show that it is possible to base authentication entirely on an analysis of the substrate material in the vicinity of the printed character.

DocEng'11, September 19–22, 2011, Mountain View, California, USA.

2. METHOD

2.1 Model Based Signature Profiles (MBSP)

We define our models simply as a set of N uniformly spaced points (x, y coordinates) defining the outer edge of a character glyph and associated unit normal vectors. Figure 1a shows an illustration of a model of the outer contour of a Times Roman lowercase ‘a’. The process of MBSP extraction is illustrated in the remainder of figure 1 (see also [13] for details). Importantly during the extraction of the profile image 1e it is preferable to low-pass filter the underlying image 1b using a standard Gaussian convolution kernel. This has the effect of removing imaging noise and more importantly avoiding sampling artifacts.

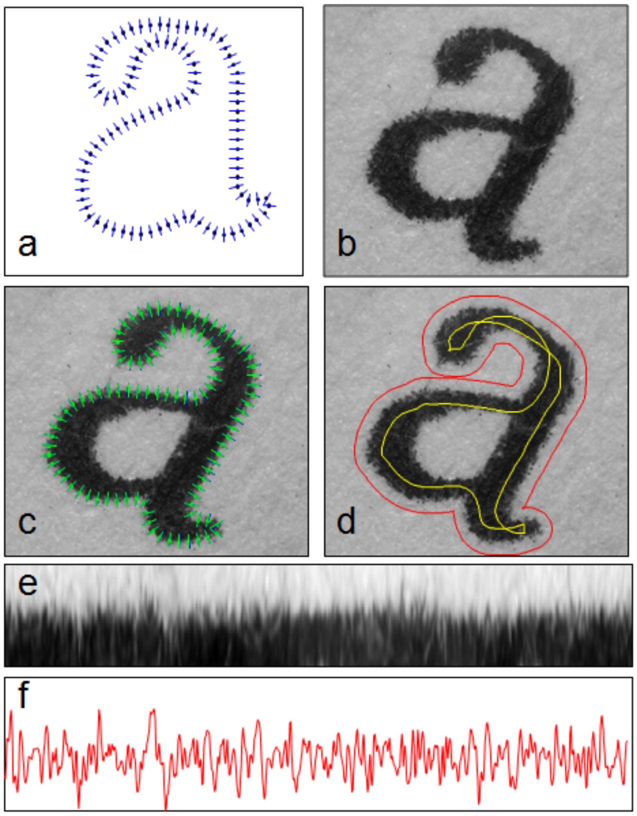

Figure 1. Illustrates, for a Times Roman 12 point ‘a’, the extraction of a MBSP. Where: (a) is a simplified model for the outline of the character composed of 100 feature points shown with associated normals (note that in practice to avoid sampling artifacts models are an order of magnitude more dense than shown in this figure; 2000 model points are used in the experiments presented in this paper); (b) is a 900x800 (wide x tall) image of a 12 point character captured by Dr CID; (c) shows superimposed transformed model; (d) shows the loci of sampled regions for the extracted normal profile images in (e). Each column of (e) corresponds to sampling on a vector between the loci along the normal vector for each individual (x, y) contour point of the model; Finally (f) shows the MBSP extracted from (e).

Many methods can be used to recover the signature profile from the profile image, including simple thresholding or maximum edge detection. We have found the following grayscale edge metric that combines all the data in the profile image to work well. For each column in the profile image the signature profile is defined as:

$$p_i = \sum_j j w_j e_{ij} \Big/ \sum_j w_j \,|\, e_{ij} \,|$$

where e_{ij} is an edge strength corresponding to the digital derivative of the profile image along the column i and w_j is a windowing function (in our case a Gaussian with standard deviation ¼ the column height centered on the mid-point of the column). Dividing by a normalizing sum of windowed absolute edge strength results in a measure that achieves robustness to both scene content and illumination variation.

2.2 Shape Warp Coding (SWC)

Introduced in [12] for micro-color-tile inspection and generalized in [13] for any printed forensic mark (provided a suitably irregular shape for which the matching process recovers a unique model location) the SWC is derived from the MBSP as follows. First we divide the signature profile into N equal length segments. Then for each, compute a sum squared error (SSE) of the residual (which is akin to a local variance):

$$SSE_j = \sum_{p_i \in segment(j)} (p_i - \mu_j)^2$$

where p_i is the signature profile in the segment j and μ_j is the mean value of that profile over the whole of the segment. We then compute the mean value (or alternatively the median) of this error metric, SSE_{mean}, over all N segments. This provides us with an atomic unit of encoding (a “digit”) and it is possible to use it to form an N-position string which is the SWC:

$$SWC(j) = \left\| SSE_j \Big/ SSE_{mean} \right\|$$

where $\|.\|$ is a rounding function. The SDED, for comparing the SWCs of any two forensic marks, is thus defined as:

$$SDED = \sum_j \min\left(| SWC_1(j) - SWC_2(j) |, T_{\max} \right)$$

where T_{max} is an optional threshold to improve robustness. The SDED can be considered a form of modified Hamming Distance where the expected value of SWC(*) is 1 at each digit due to the normalization process described. For example, a pair of SWCs (N=50) extracted from Dr CID data for the same printed ‘a’ and their absolute difference are:

```
SWC1 = 11011111201101111121121112111111011212112111101011 1
SWC2 = 11111111210100101211211121112110111111121111011210
DIFF = 0010000001100101000000000000010000010100000000001101
```

for which the SDED is 11 (or 0.22 when normalized by N).

3. EXPERIMENTS

In this section, we report the results of 2 experiments. The first applies the previously presented methodology (outlined in section 2) to a number of substrate types and includes both laser and thermal inkjet prints where appropriate. The second set of experiments involves substrate-only comparisons using a modified form of the SWC where the SSE is replaced by the variance of the substrate in the top quartile of an extended profile image (chosen to ensure that the region over which the variance is measured is not close to the intentionally inked part of the print). Thus, for the latter experiment, the sole purpose of the model is to provide a unique frame of reference to allow consistent measurement of the properties of the substrate.

After considering a wide range of paper types we chose to test 6 distinct types; 5 for laser (HP 80g office, HP 160g Matte, HP 200g Photo Matte, HP 120g Soft Gloss and Handmade Lokta by Wild Paper) and 3 for inkjet (HP 80g office, HP Premium Photo Glossy and Handmade Lokta) with two in common. These represent a broad cross-section of the paper types available. The laser printer used was an HP CP6015 and the inkjet was an HP K5400 and the printers were configured for the specific paper type being used.

For each print/paper-type combination we printed 40 Times Roman 12 point letter ‘a’s. Each is scanned twice using different Dr CID devices resulting in 640 individual images. Examples of each are shown in figure 2.

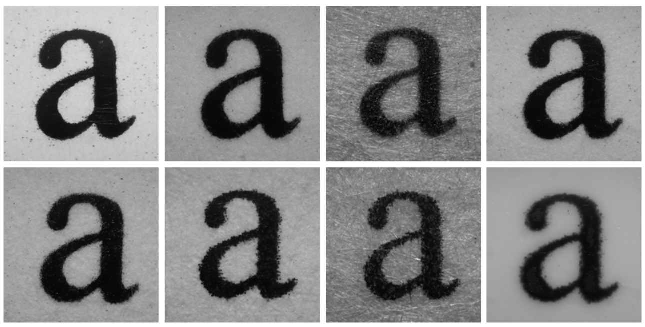

Figure2. Top left to bottom right are Laser HP 6015 on Soft Gloss, Photo Matte, Handmade, Matte and Office, followed by inkjet HP K5400 on Office, Handmade and Premium Photo.

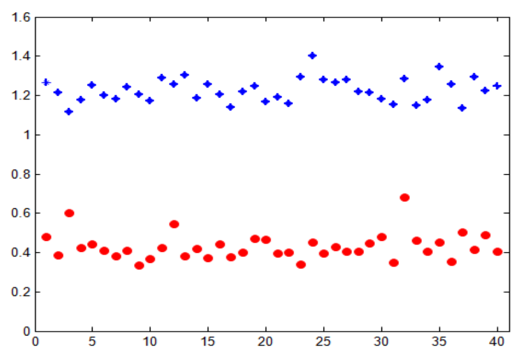

Figure 3. Plots SDED values for 40 veridical matches (red) and 40 random false matches (blue) for CP6015 laser print on 80g office paper. The sigma of Gaussian smoothing was 5.0.

3.1 Substrate Comparison

The first set of experiments use the standard methodology outlined in section 2. In each experiment the 40 SWC’s derived from printed letters captured by one Dr CID device are each compared with an SWC derived from the same printed character captured with the other Dr CID device (veridical match) and with a random incorrectly matching SWC (false match) also captured with the second device (for the same print/substrate combination). The SDED values that result from one such experiment (laser on 80g plain paper) are shown in figure 3. It is clear from this figure that the distribution of SDED values of the veridical matches is well separated from that of the false matches.

Summary statistics (means with standard deviation error bars) of a series of such experiments are shown in figure 4 for our 8 print and paper combinations. Of these only the Handmade Lokta (for both laser and inkjet) shows significantly different distribution statistics to those illustrated in figure 3.

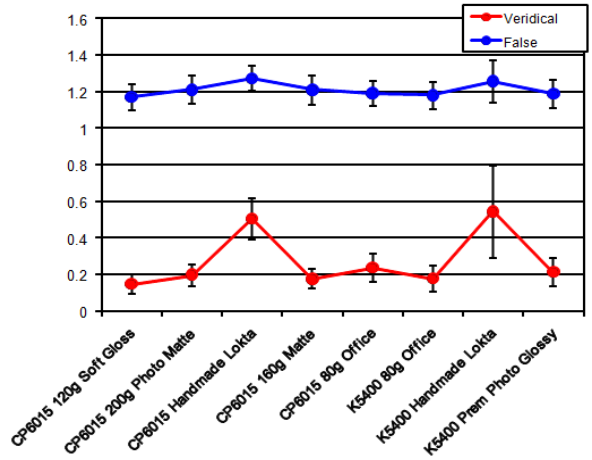

Figure 4. Plots mean SDED values and standard deviation error bars for veridical and false matches for each paper/print combination. The sigma of Gaussian smoothing was 5.0.

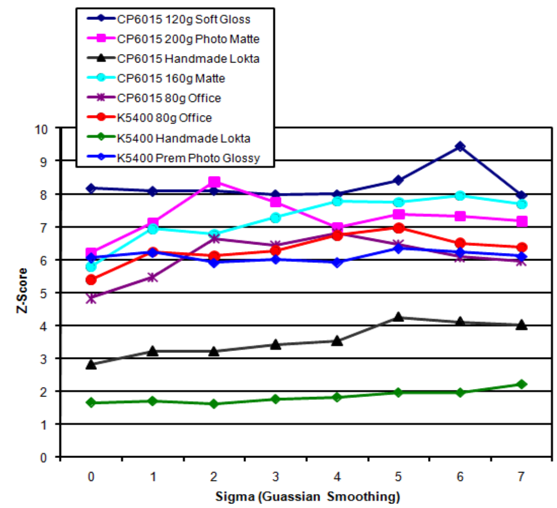

Figure 5. Plots approximate Z-score trend against the Sigma used in low-pass filter for each paper/print combination.

If we assume that the distributions of veridical and false matches are Gaussian (for which we have presented compelling evidence in [13]) then we can use an approximate a Z-score (approximate as these are small sample, rather than population, statistics) to measure the separation of the two populations:

$$Z = \left|\bar{S}_V - \bar{S}_F\right| / \left(\sigma_V + \sigma_F\right)$$

that is the absolute difference of the mean SDED scores for veridical and false matches divided by the sum of their standard deviations. The relationship between Z-score and the probability of false-positive/negative is highly non-linear, while a Z-score of 3 corresponds to a probability of 0.001 a Z-score of 6 relates to a probability of 10^{-9}. Figure 5 shows how approximate Z-scores vary with the degree of low-pass filtering applied to image during the construction of the profile image – illustrating the need to overcome sampling artifacts for effective matching. Note that a sigma of 5.0 results in Z-scores above 6.0 for all but the handmade paper types.

3.2 Substrate Authentication

As discussed in the introduction to this section the second set of experiments investigates the use of substrate variance alone to achieve forensic authentication. Figure 6 shows summary statistics for each of the 8 paper types. Remarkably the results are very similar to those of figure 4 which included both ink/toner and substrate in the region under analysis. Only the Premium Photo Glossy inkjet paper is significantly impaired due to its highly uniform and specular surface properties.

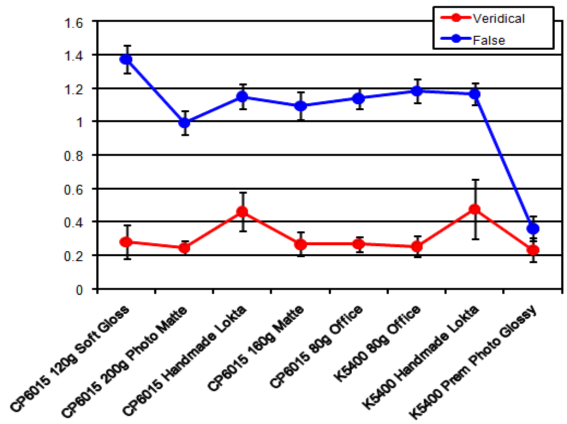

Figure 6. Substrate only plot of mean SDED values and standard deviation error bars for veridical and false matches for each paper/print combination (Sigma 5.0).

Figure 7 shows for the substrate only experiments the Z-score trend with the degree of low-pass filtering. In this case there are more significant relationships that are paper specific. Most surprising is the large difference between the two 80g office papers – given that this experiment uses only the substrate for SDED comparisons we would expect the identical paper types to give identical results. However for all the laser printed documents the substrate will inevitably include flecks of stray toner that add to the microscopic texture that is measured in the analysis of the Dr CID images.

4. DISCUSSION

Our results show that for the majority of print and substrate combinations forensic levels of authentication can be achieved with the analysis of a single Dr CID image of a single printed character whether or not the ink/toner mass of the character is included in the analysis. This is a significant and interesting result that adds to our understanding of print/substrate as a forensic signature of an individual document. In those cases where the statistical significance is reduced it is necessary to use a number of printed characters to achieve forensic-level identification. In general, if the probability of a false positive identification for a given character is p, and the desired forensic-level certainty is F, then n characters are required to achieve forensic-level certainty governed by the equation:

$$p^n = F$$

So, if p=0.022 (as is the case for a Z-score of 2) and $F=10^{-9}$, then we need 6 characters (that is, n=5.4) to achieve forensic-level validation of inkjet print on handmade paper - which while being richly textured is also highly specular and hence a small difference in the capture conditions tends to dominate the image statistics.

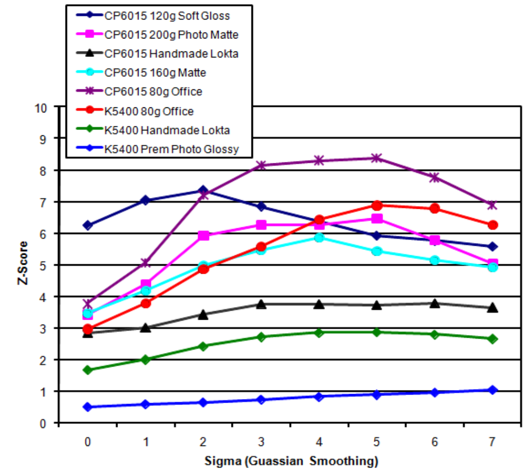

Figure 7. Substrate only plot of Z-score against the Sigma used in low-pass filter for each paper/print combination.

5. REFERENCES

[1] E. Kee and H. Faird, "Printer profiling for forensics and ballistics", *ACM MM&Sec*, 2008.

[2] S.J. Simske, M. Sturgill, P. Everest, and G. Guillory, "A system for forensic analysis of large image sets," *Proc. IEEE WIFS*, 2009.

[3] P-J. Chiang, N. Khanna, A. K. Mikkilineni, M. V. Ortiz Segovia, S. Suh, J. P. Allebach, G. T-C. Chiu & E. J. Delp, "Printer and scanner forensics: examining the security mechanisms for a unique interface", *IEEE Signal Processing Magazine* **72**, 2009.

[4] J. J. Plimmer, "Choosing correct forensic marker(s) in currency, document and product protection", *SPIE-IS&T* 6075, 2008.

[5] B. Zhu, J. WU, M. S. Kankanhalli, "Print Signatures for Document authentication", *CCS ACM,* 2003.

[6] A. Idrissa, T. Fournel & Alain Aubert, "Secure embedded verification of print signatures", *J. Phys.: Conf. Ser.* **206** 012036, 2010.

[7] S.J. Simske and G. Adams, "High-Resolution Glyph-Inspection Based Security System", *IEEE ICASSP*, 2010.

[8] L. Hindus, "Image-Based "Fingerprinting", *Advanced Imaging*, 1998.

[9] D.E. Bicknell and G.M. Laporte, "Forged and Counterfeit Documents", in Wiley Encyclopedia of Forensic Science, 3104, June 2009.

[10] G. Adams, "Hand held Dyson relay lens for anti-counterfeiting", *Proc. IEEE IST*, 2010.

[11] S.B. Pollard, G. Adams and S.J. Simske, "Resolving distortion between linear and area sensors for forensic print inspection", *IEEE ICIP*, 2010.

[12] S.J. Simske, S.B. Pollard & G. Adams, "An imaging system for simultaneous inspection, authentication and forensics", *IEEE IST*, 2010.

[13] S.B. Pollard, S.J. Simske, G.B. Adams, "Model Based Print Signature Profile Extraction For Forensic Analysis of Individual Text," Proc. IEEE WIFS 2010, pp. 1-6, 2010