Determining printer and scanner resolution dependency of text classification for digital image forensics

Jason S. Aronoff & Steven J. Simske; Hewlett-Packard Labs; Fort Collins, CO, U.S.A

Abstract

Forensic identification of the hardware used during printing and image scanning is a technology of value for security printing, inspection and even criminal investigations. The more familiar image forensics are concerned with determining the operations that have been performed on a digital image—usually for identifying the camera model used to capture the image. When an image is both printed and scanned, however, the forensics task is more complicated, since the print-scan (PS) cycle introduces less specific effects on the images. In order to identify the printer used to produce and read an image, classification must be performed. In this paper, we use a multi-class Adaboost classifier to determine which of 6 printers, representing 3 inkjet and 2 laserjet models, was used to produce a later-scanned image. Our results, investigating 6 different English characters, show that classification accuracy continues to increase with scanning resolution up to 1200 pixels/inch. The results are character-dependent, suggesting that different characters may be used for different forensic purposes—printer model, cartridge and individual printer identification as examples.

Introduction

Counterfeiting of printed materials—whether currency, legal documents, or packages and labels—continues to be a growing problem globally. High resolution printers, copiers, scanners and photo editing software are ubiquitous, placing powerful counterfeiting tools within reach of any would-be forger. Many advances have been made in the area of print forensics, but considerable work still remains to be done in the field.

In this paper, we investigate the effect of scan resolution and character shape on the accuracy of source printer identification. Specifically, given a set of text documents of known printer origin, if the prints are scanned at different resolutions and different characters are selected for use, we investigate what effects this has on the ability to classify characters and identify a document of unknown origin.

The ability to distinguish documents from printers is due to the fact that printers exhibit unique signatures. There are several reasons that this occurs. First, due to differences in drivers, print engines, motors, rollers, and other design aspects of printers, different makes and models will deposit ink or toner in different ways. Second, due to differences in the design of the cartridges and chemical composition of inks and toners, there will be differences in the appearance of a print. Lastly, due to variances in the manufacturing process of the printer itself and possibly the printer cartridges, each printer may exhibit a unique signature which can differentiate printers of the same make and model.

With regard to the first two aspects, other research groups have found these statements to be true [1-4], and have developed methods to identify the make and model of the source printer. However, with regard to identifying document origins when looking at multiple instances of the same make and model printer, advancements in the field have not been as successful.

To the best of our knowledge, little research has been performed in the area of print forensics sensitivity analysis with the exception of [1]. In this work the authors examined the effects of changing font, font size and substrate on classification accuracy. Each of these tests was performed independently of the other and all tests were performed on only the letter e at a scan resolution of 2400 dpi. The works undertaken in [2-4], were also conducted using a single scan resolution.

The primary goal of this paper is to examine sensitivity analysis for printer forensics. We will examine what the effects of resolution are on classification accuracy as well as whether different shapes are better at eliciting these unique printer signatures. As has been done by other research groups [1-3], we will use characters from the western alphabet as the carrier for detecting unique printer signatures.

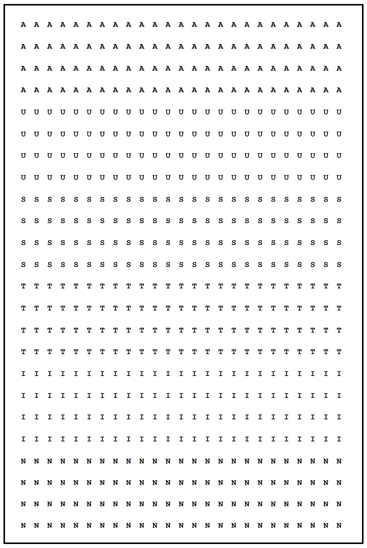
The rest of this paper is structured as follows. In the next section, “Methods”, we will overview our approach to performing sensitivity analysis and classification. We will then discuss the experiments conducted and the results in the “Experiment Results” section. Analysis of the results is reviewed in the “Discussion” section and we end with final remarks and an overview of future work in the “Conclusion” section.

Methods

In our approach, classification and analysis of printed documents can be divided into four steps: (1) printed sample generation; (2) scanning and extraction of individual letters via image processing; (3) feature extraction; and (4) classification of the letters. In the remainder of this section we provide details on each of these steps.

All sample sheets used in our experiments were printed from a single digital raster. The digital raster was generated at 600 ppi and contained 6 letters; {A, U, S, T, I, N}. The selected letters were chosen based on the location of this year’s conference and the fact that the shapes of each character provide a level of uniqueness in curvature, edges and angles. We believe other characters in the western alphabet could work equally as well (e.g. the letter Z could have been selected instead of N).

For each letter in the set, 100 instances of the letter were printed on the page by generating 4 rows of 25 characters each. All characters were generated using the Courier New font at size 10. Figure 1 depicts a scaled down version of the digital raster. Five pages were then printed at 600 dpi for each printer used in the experiments.



```
A A A A A A A A A A A A A A A A A A A A A A A A A
A A A A A A A A A A A A A A A A A A A A A A A A A
A A A A A A A A A A A A A A A A A A A A A A A A A
A A A A A A A A A A A A A A A A A A A A A A A A A
U U U U U U U U U U U U U U U U U U U U U U U U U
U U U U U U U U U U U U U U U U U U U U U U U U U
U U U U U U U U U U U U U U U U U U U U U U U U U
U U U U U U U U U U U U U U U U U U U U U U U U U
S S S S S S S S S S S S S S S S S S S S S S S S S
S S S S S S S S S S S S S S S S S S S S S S S S S
S S S S S S S S S S S S S S S S S S S S S S S S S
S S S S S S S S S S S S S S S S S S S S S S S S S
T T T T T T T T T T T T T T T T T T T T T T T T T
T T T T T T T T T T T T T T T T T T T T T T T T T
T T T T T T T T T T T T T T T T T T T T T T T T T
T T T T T T T T T T T T T T T T T T T T T T T T T
I I I I I I I I I I I I I I I I I I I I I I I I I
I I I I I I I I I I I I I I I I I I I I I I I I I
I I I I I I I I I I I I I I I I I I I I I I I I I
I I I I I I I I I I I I I I I I I I I I I I I I I
N N N N N N N N N N N N N N N N N N N N N N N N N
N N N N N N N N N N N N N N N N N N N N N N N N N
N N N N N N N N N N N N N N N N N N N N N N N N N
N N N N N N N N N N N N N N N N N N N N N N N N N
```

Figure 1 *Scaled version of digital raster used for test sheet generation. Actual size is 8.5" by 11".*

Each printed test sheet was then scanned at multiple resolutions on an HP Scanjet 8350 flatbed scanner with an automatic document feeder (ADF). Scans were performed at 75, 150, 300, 600, and 1200 ppi at 24-bit RGB resolution. All scans were saved in a lossless PNG format. Using in-house image processing software, each scanned page was analyzed to extract each individual letter from the page.

After shape extraction, each letter was analyzed to extract 36 feature descriptors of the shape. Table 1 lists the features extracted from each shape. Since the features themselves are not a primary focus of this body of work, we refer the reader to [5] and [6] for expanded descriptions and equations for calculating the feature descriptors. While it may appear that some of these features are redundant, rather than manually removing features we made the decision to keep all of them and allow the classifier to determine which features provided the best information to discriminate between classes.

For classification we selected the Adaboost algorithm developed by Freund and Schapire in [7], and modified it to handle multi-class classification problems by using the approach described in [8]. We briefly describe the Adaboost algorithm and multi-class approach below and refer the reader to [7] and [8] for additional details.

The Adaboost algorithm is a simple yet effective classifier which utilizes the concept of boosting to take weak classifiers and linearly combine them into an overall strong classifier. Formally, the strong classifier can be defined as

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{1}$$

In Equation (1), $h_t(x)$ represents the weak hypothesis of the weak classifier t, and α_t is the weighting of the hypothesis. In the two class approach the indicator variable y is defined as $y_i \in Y = \{-1,+1\}$. The final hypothesis of the classifier, or predicted class, as expressed in Equation (1) is then the sign of $f(x)$.

To train the Adaboost classifier, one specifies the number of weak learners in parameter T and for each weak learner iterates over the training set to determine the weightings. Given a training set of m samples, h_t can be defined as

$$h_t = \arg\min_{h_j \in H} \varepsilon_j = \sum_{i=1}^{m} D_t(i)\left[y_i \neq h_i(x_i)\right] \tag{2}$$

If the error value ε is greater than 0.5 then stop. The weight of the weak learner can then be updated by the function

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{3}$$

Lastly, the distribution D_t is updated by the function

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{4}$$

where Z_t is a normalizing constant.

To expand this to a multi-class problem requires determining how to reduce the multiclass problem to a set of binary problems so that it can be integrated into the Adaboost algorithm. A number of approaches have been suggested by others such as one-against-many or all-pairs. Since the latter requires significantly more computational overhead due to the number of combinations, we opted to utilize the one-against-all approach. Using this approach, as described in [8], given a set of k classes, the multi-class problem is broken up into k binary tests where each class is compared against the remaining classes which are treated as a single second class.

The first part of this approach is to define an encoding matrix ***M*** of size k by k which represents the class labels. Using the same values for indicator variables as defined in the original Adaboost approach [7], the diagonal of the matrix will contain +1 and all other elements will be -1.For example, in a 3 class problem class-1 is defined as [+1,-1,-1], and class-3 is defined as [-1,-1,+1].

With the addition of multiple classes, the Adaboost algorithm must now take a new error function to encompass the one-against-all approach. To do this, one must perform an additional iteration over the classes which is nested inside the iterations over the weak learners. This allows the function to compute the weights over all the classes.

$$\varepsilon_t = \sum_{i=1}^{m} \sum_{s=1}^{k} D_t(i,s)\left[M(y_i,s) \neq h_t(x_i,s)\right] \tag{5}$$

In Equation (5), the distribution is defined over the training samples and each class. This is computed in a manner similar to the two class approach in Equation (4) as seen in the following equation:

$$D_{t+1}(i,s) = \frac{D_t(i,s)\exp(-\alpha_t M(y_i,s) h_t(x_i,s))}{Z_t} \tag{6}$$

As with Equation (4), the value Z_t is a normalizing constant.

The last step is to redefine Equation (1) to handle multiple classes. Using the concept of loss based decoding from [8], the loss function can be defined as

$$\frac{1}{mk}\sum_{i=1}^{m}\sum_{s=1}^{k}L(M(y_i,s)f_s(x_i)) \tag{7}$$

In Equation (7), the hypothesis f_s, which represents the hypothesis that sample x_i belongs to class s, is compared to the encoded value of each training point from the encoding matrix M to give the average loss over all training points and hypotheses.

The final hypothesis of the multi-class problem can then be determined by looking at the vector of class hypotheses ***f(x)***. Using loss-based decoding as a distance measure, the predicted class is determined by computing the distance of the prediction to each class and assigning the sample to the class with the shortest distance or loss. Formally this is defined as

$$d_L(M(r),f(x))=\sum_{s=1}^{k}L(M(r,s),f_s(x)) \tag{8}$$

where r is the row of the label which is closest to ***f(x)***.

Experiment Results

To look at whether scan resolution or shape has an effect on classification accuracy, three sets of experiments were performed. For the first two experiments we used 6 printers. For the third experiment we used only 2 of the 6 printers. The printers used in all of the experiments are listed in Table 2. As described in the Methods section, 5 pages of the digital raster were printed for each printer at 600 dpi. All prints were performed using the *normal print quality* settings and for all color printers the settings were modified so that prints were done using only black ink or toner. All prints were done on the same brand of office paper.

For each of the three experiments, classification tests were broken out by letter and resolution. With 6 letters and 5 resolutions, 30 separate classification tests were performed. Since our goal was forensic based, training and testing was performed by a leave-one-out approach. For each classification test, the classifier was trained on samples from 4 of the 5 printed pages for each of the printers used. The fifth page was then used as the test set. For example, looking at the letter "A" at a scan resolution of 600 ppi, if all 6 printers are used then there are 2400 letters used for training (6 printers * 4 pages per printer * 100 A's per page), and 600 letters used for testing (6 printers * 1 page per printer * 100 A's per page). Similarly to k-fold cross-validation tests, the classification tests were conducted so that each of the 5 pages was rotated through as a test page and the remaining 4 were used for training.

In the first experiment, all 6 printers were treated as individual classes. To capture the general trends of classification accuracy we distilled the results into two graphs, one for the training set and the other for the test set, shown in Figures 2 and 3. In each of these figures, the plot depicts the mean accuracy for each letter with respect to scan resolution. However, it is important to point out that these plots do not fully represent the data. The confusion matrices for the results of the test sets-where page 1 is the test sheet are listed in Table 3. Due to space limitations and the large number of classification tests performed, it is not possible to list the results of each test or the confusion matrices of the training data.

In the second experiment, the two instances of the 6940 Deskjet printer were treated as a single class; thereby reducing the problem to a 5-class classification test. This doubled the number of samples used for testing and training of the 6940 class, but all other aspects of the experiment were conducted the same as the first. The mean classifier accuracy of the train and test sets is plotted in Figures 4 and 5 and the confusion matrices for the test results using the first page are in Table 4.

For the third experiment, we ran a two class problem where only the two instances of the 6940 were analyzed. Table 5 shows the confusion matrices for all letters and resolutions from the results of the first test page.

Discussion

Based upon the results presented in Tables 3, 4, and 5, several trends can be identified. First, it is clear that as resolution increases, the overall accuracy of classification increases as well. Figures 2 through 5 substantiate this claim as do the confusion matrices in Tables 3-5. But, it is important to point out that in our approach the improvement in accuracy from 600 ppi to 1200 ppi is marginal. Depending on what the final uses of the results are from classification, one may be able to use resolutions lower than 600 ppi. For example, if one were to use a simple majority vote of the classified characters on a test page to identify the source printer, using the results from the letter N, one could scan as low as 150 ppi and still correctly identify all source printers in the 6-class problem. If one looks a bit deeper into the data in Table 3, it is apparent that this approach will not work for the letters A or S. In these two tests the two 6940 TIJ printers are not distinguishable.

If the classification problem is changed into a 5-class problem, to only identify make and model, then at 150 ppi we are able to correctly identify every printer using majority vote regardless of which letter is selected. This represents a typical real-world workflow in which a document as a set of printed marks is forensically analyzed.

The results of the most interest to us, however, were those of the 2-class problem in which the objective was to distinguish one 6940 from the other. Looking again at the results in Table 5, with a scan resolution as low as 150 ppi, our approach is able to correctly identify the source printer using majority vote regardless of letter. However, it is worth pointing out that of the letters used, the letter N yielded the highest classification accuracy at each resolution. As with the other experiments, the peak classification accuracy appears to be at 600 ppi. For the letter N, the accuracy is 90% at this resolution but drops slightly to 88% at 1200 ppi. Other letters exhibited a much stronger downward trend from 600 to 1200 ppi. The letter S, for example, is classified with an accuracy of 87.5% at 600 ppi, but drops precipitously to 70% accuracy at 1200 ppi. This trend is only weakly exhibited, if at all, for letter/scan resolution combinations in the first two experiments (multi-class problems). However, as previously stated, with scan resolutions as low as 150 ppi, satisfactory printer source identification results can be attained.

The results seen in the three sets of experiments suggest a possible workflow which can be used to identify source printers in forensic settings. First, identify printer models which are potential source candidates for a document in question. Using the approach of experiment 2, perform a multi-class classification to identify the model in question. Then following the approach of experiment 3, perform a classification analysis of the instances for the make and model in question.

Bounded Binary X Centroid	Second Order Column Moment
Bounded Binary Y Centroid	Second Order Mixed Moment
Bounded Weighted X Centroid	Second Order Row Moment
Bounded Weighted Y Centroid	Std. Dev. Radial Distance
Circularity 1	Texture Contrast
Circularity 2	Texture Correlation
Cropped Height	Texture Energy
Cropped Width	Texture Entropy
Height	Texture Homogeneity
Ink Area (grayscale)	Theta 1
Major Axis Length	Theta 2
Major Axis Orientation	Weighted X Centroid
Mean Radial Distance	Weighted Y Centroid
Minor Axis Length	Width
Minor Axis Orientation	X-axis symmetry
Perimeter Length	Y-axis symmetry
Perimeter Pixel Count	X-centroid
Pixel Area (binary)	Y-centroid

Table 1 *List of feature descriptors used for classification.*

Printer	Print Technology
HP Laserjet 3005d	DEP
HP Laserjet 3600dn	DEP
HP Deskjet 6127	TIJ
HP Photosmart C6280 AiO	TIJ
HP Deskjet 6940 (1)	TIJ
HP Deskjet 6940 (2)	TIJ

Table 2 *Printers used in experiments. Note for the 6940 two instances of the same printer model were used and they are designated as (1) and (2).*

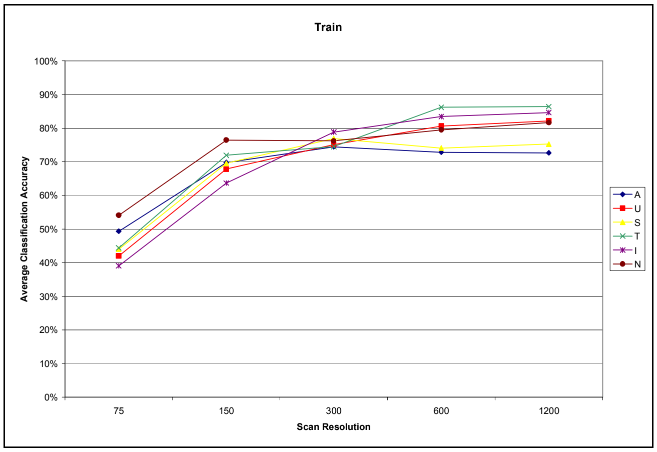


Figure 2 *Mean classification accuracy of the training data from the 6-class leave-one-out experiment.*

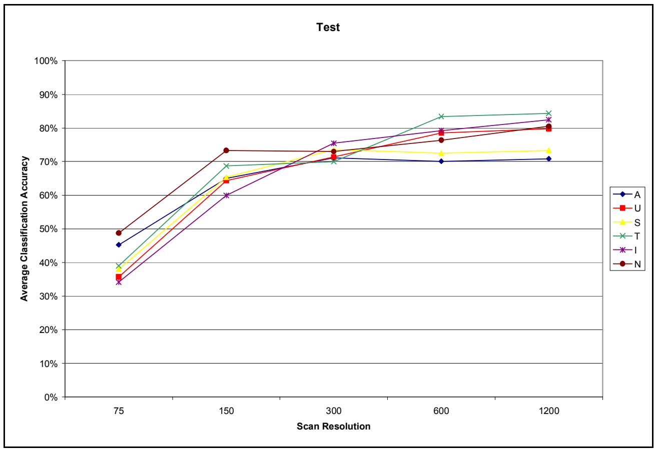


Figure 3 *Mean classification accuracy of the test data from the 6-class leave-one-out experiment.*

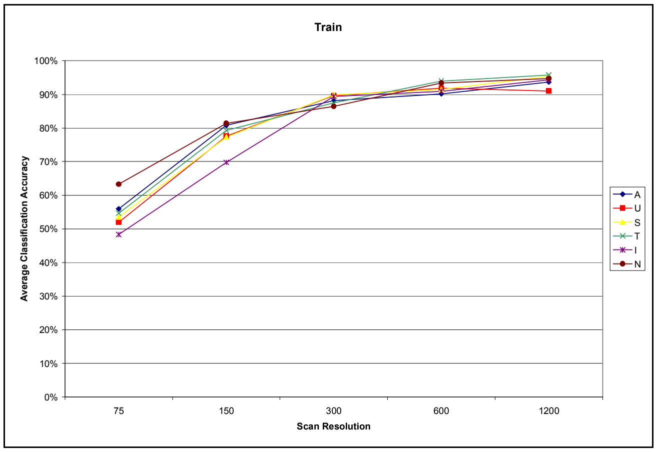


Figure 4 *Mean classification accuracy of the training data from the 5-class leave-one-out experiment.*

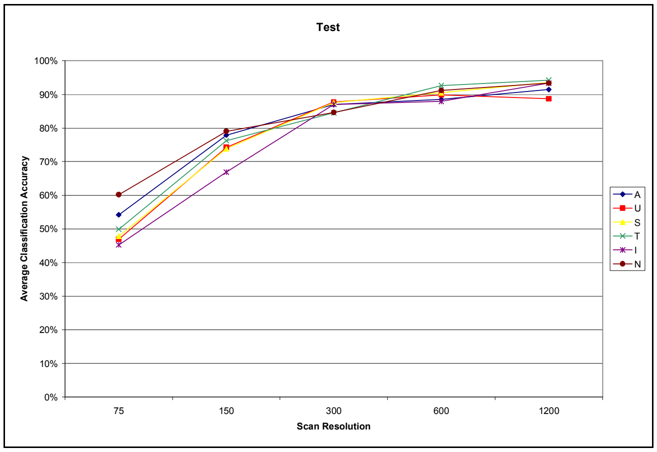


Figure 5 *Mean classification accuracy of the test data from the 5-class leave-one-out experiment.*

Letter		A						U						S						T						I						N					
Resolution		6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005
75	6940(1)	24	28	0	31	5	12	40	20	2	31	2	5	25	24	5	20	7	19	25	29	0	34	0	12	26	18	13	31	11	1	28	39	0	21	1	11
	6940(2)	20	39	0	11	2	28	28	20	8	24	8	12	26	31	6	18	3	16	19	26	1	38	3	13	20	30	13	22	10	5	23	59	1	11	1	5
	6280	0	1	70	0	24	5	0	0	71	0	21	8	1	5	56	0	18	20	0	1	75	0	19	5	5	7	60	0	22	6	0	2	70	0	17	11
	6127	14	19	0	64	1	2	21	25	3	42	5	4	20	26	1	46	3	4	11	17	1	61	1	9	29	13	2	48	6	2	19	19	0	62	0	0
	3600	1	2	34	0	39	24	4	4	57	1	22	12	1	4	43	0	33	19	1	0	58	0	14	27	9	11	38	0	26	16	1	4	39	0	28	28
	3005	8	13	7	3	19	50	11	11	26	15	22	15	12	22	17	2	22	25	14	13	23	10	10	30	22	15	26	15	14	8	11	7	16	1	32	33
150	6940(1)	37	42	0	9	4	8	41	32	0	12	0	15	50	26	0	16	2	6	56	26	0	10	0	8	49	18	0	14	0	19	66	18	0	10	4	2
	6940(2)	24	55	1	3	11	6	26	62	0	4	1	7	45	44	2	3	4	2	27	69	0	0	0	4	23	52	2	5	1	17	13	86	1	0	0	0
	6280	0	0	93	0	7	0	1	2	88	0	9	0	0	2	94	0	3	1	0	2	91	0	4	3	0	0	78	0	15	7	2	3	88	0	6	1
	6127	18	4	0	69	0	9	28	5	0	53	0	14	26	6	0	66	0	2	21	4	0	73	1	1	16	4	0	73	0	7	9	4	0	83	2	2
	3600	2	1	2	0	59	36	1	1	4	2	76	16	0	0	4	0	78	18	0	0	6	0	70	24	0	0	12	0	57	31	0	0	7	0	66	27
	3005	13	1	1	0	14	71	9	4	0	3	18	66	20	5	0	4	9	62	20	7	0	2	14	57	23	6	1	0	19	51	1	3	1	0	27	68
300	6940(1)	43	53	0	1	3	0	69	18	0	6	7	0	51	25	0	3	7	14	65	18	0	8	1	8	70	20	0	3	0	7	82	5	0	6	6	1
	6940(2)	39	59	0	0	2	0	55	40	0	0	5	0	33	64	0	0	2	1	57	39	0	1	0	3	45	52	1	0	2	0	44	54	0	1	1	0
	6280	0	2	94	0	4	0	0	0	99	0	1	0	3	2	92	0	3	0	4	0	94	0	2	0	0	1	98	0	1	0	5	4	82	0	9	0
	6127	17	2	0	75	0	6	18	0	0	72	1	9	8	0	0	80	4	8	18	0	0	78	0	4	19	1	0	79	0	1	2	1	0	89	5	3
	3600	1	0	0	0	97	2	0	0	1	0	95	4	1	0	1	0	82	16	3	0	5	0	60	32	0	0	2	0	85	13	0	0	0	0	85	15
	3005	3	1	0	0	17	79	5	0	0	2	25	68	3	0	0	0	12	85	7	0	0	9	13	71	18	1	1	1	10	69	1	0	0	1	52	46
600	6940(1)	92	7	0	0	0	1	77	16	0	6	0	1	94	4	0	0	0	2	86	9	0	4	0	1	73	12	0	9	0	6	98	1	0	1	0	0
	6940(2)	92	7	0	0	1	0	47	51	0	0	0	2	77	21	0	0	2	0	53	46	0	0	0	1	21	76	0	0	0	3	75	25	0	0	0	0
	6280	0	0	97	0	3	0	0	0	100	0	0	0	2	0	98	0	0	0	0	1	98	0	1	0	1	0	99	0	0	0	5	2	79	0	14	0
	6127	40	0	0	44	4	12	15	1	0	73	1	10	18	0	0	68	8	6	7	0	0	86	1	6	9	0	0	76	0	15	31	0	0	59	0	10
	3600	0	0	23	0	73	4	0	1	1	0	76	22	2	0	6	0	52	40	0	0	2	0	84	14	2	0	8	0	80	10	0	0	0	0	100	0
	3005	0	0	0	0	13	87	2	0	0	2	7	89	3	0	0	0	2	95	3	0	0	3	6	88	8	2	0	3	6	81	0	0	0	0	37	63
1200	6940(1)	90	10	0	0	0	0	88	7	0	3	0	2	84	8	0	0	7	1	72	16	1	3	0	8	86	10	0	3	0	1	93	3	1	1	2	0
	6940(2)	95	5	0	0	0	0	54	43	0	0	0	3	70	27	0	0	3	0	32	66	1	1	0	0	48	51	0	0	0	1	51	48	1	0	0	0
	6280	5	3	92	0	0	0	0	0	99	0	0	1	7	1	92	0	0	0	3	1	96	0	0	0	4	0	95	0	1	0	20	1	75	0	4	0
	6127	20	0	0	72	2	6	6	2	0	76	0	16	7	0	0	69	16	8	12	0	0	84	0	4	0	0	0	87	2	11	13	0	0	84	3	0
	3600	0	0	20	0	77	3	0	2	2	0	85	11	0	0	7	0	74	19	0	0	1	0	92	7	0	0	3	0	96	1	0	0	2	0	89	9
	3005	0	0	0	0	1	99	0	0	0	1	2	97	0	0	0	0	2	98	3	0	0	1	2	94	16	0	1	6	17	60	0	0	1	0	7	92

Table 3 *Confusion matrices of the test data results for experiment 1 (6-class tests). The classifier is trained on the second through fifth pages from each printer and tested using the first page.*

Letter		A					U					S					T					I					N				
Resolution		6940	6280	6127	3600	3005	6940	6280	6127	3600	3005	6940	6280	6127	3600	3005	6940	6280	6127	3600	3005	6940	6280	6127	3600	3005	6940	6280	6127	3600	3005
75	6940	174	0	12	10	4	148	13	27	3	9	131	10	21	9	29	170	1	16	7	6	150	25	21	3	1	165	0	21	4	10
	6280	3	79	0	14	4	1	67	0	25	7	5	49	0	19	27	1	74	0	17	8	13	68	0	14	5	0	66	0	21	13
	6127	73	0	26	1	0	71	2	21	4	2	65	1	30	2	2	73	1	19	1	6	73	0	27	0	0	44	0	56	0	0
	3600	10	44	0	34	12	11	59	3	17	10	6	40	0	36	18	4	55	0	19	22	32	44	0	15	9	2	33	0	33	32
	3005	42	12	1	32	13	37	29	12	12	10	44	14	2	24	16	38	23	3	18	18	60	25	2	5	8	21	15	0	28	36
150	6940	164	1	9	17	9	181	0	5	1	13	171	2	12	5	10	181	0	10	1	8	153	2	18	1	26	193	1	3	2	1
	6280	0	93	0	7	0	3	89	0	8	0	1	95	0	3	1	1	91	0	5	3	0	80	0	13	7	3	90	0	5	2
	6127	21	0	68	0	11	39	0	45	0	16	43	0	55	0	2	34	0	66	0	0	27	0	66	0	7	27	0	69	1	3
	3600	5	2	0	57	36	2	4	2	76	16	1	3	0	79	17	0	5	0	73	22	0	12	0	57	31	0	7	0	66	27
	3005	10	1	3	13	73	18	0	3	19	60	24	0	3	9	64	30	1	2	15	52	34	2	0	18	46	6	1	0	28	65
300	6940	194	0	0	5	1	197	0	2	1	0	191	0	0	4	5	193	0	1	0	6	195	1	1	0	3	193	1	3	2	1
	6280	2	94	0	4	0	0	99	0	1	0	6	91	0	3	0	1	96	0	2	1	1	98	0	1	0	7	85	0	8	0
	6127	14	0	81	1	4	9	0	82	2	7	7	0	80	7	6	11	0	85	0	4	8	0	88	0	4	7	0	84	5	4
	3600	0	0	0	97	3	0	3	0	92	5	0	1	0	92	7	0	4	0	62	34	0	2	0	85	13	0	0	0	85	15
	3005	2	0	0	16	82	0	0	2	26	72	4	1	0	15	80	3	0	7	17	73	5	1	3	10	81	0	0	0	52	48
600	6940	197	0	2	1	0	197	0	3	0	0	195	0	0	1	4	198	0	2	0	0	191	0	3	0	6	198	0	1	1	0
	6280	0	97	0	3	0	0	100	0	0	0	0	99	0	0	1	1	98	0	1	0	1	99	0	0	0	3	85	0	12	0
	6127	17	0	66	3	14	8	0	79	2	11	9	0	81	6	4	6	0	87	0	7	6	0	81	0	13	5	0	84	0	11
	3600	0	18	0	78	4	0	1	0	76	23	0	0	0	62	38	0	2	0	87	11	0	7	0	83	10	0	0	0	99	1
	3005	0	0	0	7	93	0	0	2	3	95	1	0	0	3	96	1	0	4	7	88	8	0	3	6	83	0	0	0	35	65
1200	6940	198	0	2	0	0	194	0	2	0	4	199	0	0	1	0	196	0	2	0	2	198	0	0	0	2	198	2	0	0	0
	6280	6	92	2	0	0	1	98	0	0	1	4	96	0	0	0	4	96	0	0	0	0	99	0	1	0	8	89	0	3	0
	6127	7	0	86	1	6	17	0	69	0	14	6	0	87	1	6	9	0	89	0	2	1	0	88	0	11	5	0	92	1	2
	3600	0	14	0	83	3	0	2	0	86	12	1	0	0	84	15	0	0	0	93	7	0	3	0	94	3	0	0	0	89	11
	3005	0	0	0	1	99	0	0	1	2	97	0	0	0	0	100	2	0	1	1	96	0	0	4	13	83	0	0	0	2	98

Table 4 *Confusion matrices of the test data results for experiment 2 (5-class tests) where the two 6940 printers are combined into a single class. The classifier is trained on the second through fifth pages from each printer and tested using the first page.*

Letter		A		U		S		T		I		N	
Resolution		6940(1)	6940(2)	6940(1)	6940(2)	6940(1)	6940(2)	6940(1)	6940(2)	6940(1)	6940(2)	6940(1)	6940(2)
75	6940(1)	53	47	48	52	39	61	45	55	49	51	59	41
	6940(2)	39	61	47	53	27	73	42	58	34	66	30	70
150	6940(1)	69	31	69	31	77	23	75	25	77	23	86	14
	6940(2)	42	58	44	56	41	59	29	71	40	60	14	86
300	6940(1)	75	25	72	28	81	19	73	27	91	9	92	8
	6940(2)	49	51	50	50	29	71	23	77	35	65	26	74
600	6940(1)	85	15	90	10	91	9	94	6	90	10	93	7
	6940(2)	36	64	46	54	16	84	27	73	12	88	13	87
1200	6940(1)	56	44	60	40	64	36	83	17	84	16	87	13
	6940(2)	36	64	40	60	24	76	26	74	15	85	11	89

Table 5 *Confusion matrices of the test data results for experiment 3 (2-class tests). Only the two 6940 printers were used in this experiment.*

Conclusion

Using a multi-class Adaboost classifier, we performed a sensitivity analysis study to examine how scan resolution and shape affect the ability to correctly classify samples to their original source printers. Our results show that, with a resolution as low as 150 ppi, it is possible to identify the make and model of the source printer. Once this has been achieved, reducing the classification problem to look only at instances of the make/model in question allows one to correctly identify the specific source printer. These results are advancement over the studies and findings in [1-2].

Future Work

Based on the experiments performed and the results obtained, there remain a number of unanswered questions which warrant further research. First, our approach for identifying the specific instance of a make and model printer was limited to two printers. A more in-depth study is required to determine the robustness of our approach if more than two printers of the same make and model were in question. Secondly, with regard to how the printer signatures originate, research needs to be undertaken to investigate whether swapping ink/toner cartridges between printers of the same make and model changes the signatures of the printers or moves with the cartridges. Finally, an in-depth analysis needs to be performed to assess the effectiveness of our approach when using laser printers.

References

[1] A.K. Mikkilineni, et al., Printer Forensics Using SVM Techniques, Proc. 21st Intl. Conf. on Digital Printing Technologies, pp.223-226. (2005).

[2] A.K. Mikkilineni, et al., Printer Identification Based on Graylevel Co-occurrence Features for Security and Forensic Applications, Proc. SPIE-IS&T Electronic Imaging, pp.430-440. (2005).

[3] E. Kee and H. Farid, Printer Profiling for Forensics and Ballistics, Proc. 10th ACM workshop on Multimedia and Security, pp. 3-10. (2008).

[4] O. Bulan, J. Mao, and G. Sharma, Geometric Distortion Signatures for Printer Identification, Proc. ICASSP, pp. 1401-1404. (2009).

[5] R.C. Gonzalez and R.E. Woods, Digital Image Processing, 3rd ed., (Pearson Education, Inc., Upper Saddle River, NJ, 2008).

[6] L.G. Shapiro and G.C. Stockman, Computer Vision, (Prentice-Hall, Inc., Upper Saddle River, NJ, 2001).

[7] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In Proc. of the Second European Conference on Computational Learning Theory. LNCS, pp. 23-27. (1995.)

[8] E.L. Allwein, R.E. Schapire, and Y. Singer. “Reducing multiclass to binary: a unifying approach for margin classifiers,” Journal of Machine Learning Research, 1, pp. 113-141 (2001).

Author Biography

Jason Aronoff received his MS in Computer Science from Colorado State University in 2008. He has been working full time for HP Labs since the beginning of 2007 when he joined what has now become the Security Printing and Imaging group. His work has focused on deterrent qualification and functional printing as applied towards anti-counterfeiting techniques. His is a member of IS&T, IEEE, and ACM.