Journal of Imaging Science and Technology® 55(5): 050602-1–050602-10, 2011.

Effect of Scanner Resolution and Character Selection on Source Printer Identification

Jason S. Aronoff▲ and Steven J. Simske▲
Hewlett-Packard Laboratories, 3404 E. Harmony Rd., Fort Collins, Colorado 80528
E-mail: Jason.S.Aronoff@hp.com

Abstract. *In today's world of highly sophisticated technological crimes, criminals including counterfeiters, document forgers, and other parties interested in altering information have a low barrier of entry. To combat these crimes requires developing a level of forensic analysis to aid law enforcement agencies in tracing the origins of documents or materials in question. In this article, the authors use printer forensics in an effort to understand the effect of resolution and character selection on the accuracy of printer identification. Specifically, they use a multiclass ADABOOST classifier to determine which of six printers, representing several ink jet and laserjet models, were used to produce a subsequently scanned image. Their results, investigating six different English characters, show that classification accuracy continues to increase with scanning resolution up to 1200 pixels/in. 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.* © *2011 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2011.55.5.050602]

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 photograph 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 article, we investigate the effect of scan resolution and character selection 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, if any, this variability has on the ability to classify characters and to 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. Other factors, including substrate and the imaging device(s) used to capture them, may also play a role in eliciting and identifying source printers. However, these factors will not be examined in this work.

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. Advancements in the field have not been as successful in attempts to identify document origins when looking at multiple instances of the same make and model printer.

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

The primary goal of this article is to examine sensitivity analysis for printer forensics. We will examine the effects of resolution on classification accuracy as well as the impact of shape variation for 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. An earlier version of this work is presented at the 2010 IS&T NIP26 conference in Austin, Texas.[5]

The rest of this article is structured as follows. In the "Methods" section, we will overview our approach to sensitivity analysis and classification. We will then discuss the experiments conducted and the outcomes in the "Experimental 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 "Conclusions" section.

METHODS

In our approach, classification and analysis of printed documents can be divided into four steps: (1) printed

▲IS&T Member.

Received Dec. 22, 2010; accepted for publication Jul. 1, 2011; published online Nov. 2, 2011
1062-3701/2011/55(5)/050602/10/$20.00.

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 8.5in. by 11in.

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 six letters; {A, U, S, T, I, N}. The selected letters were chosen based on the fact that the shapes of each character provide a level of uniqueness in curvature, edges and angles (the original version of this work having been targeted for the NIP26 conference, which was located in Austin, Texas). 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 four 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 a resolution of 600 dpi for each printer used in the experiments.

Each printed test sheet was then scanned at multiple resolutions on an HP Scanjet 8350 flatbed scanner with an automatic document feeder (ADF). The test sheets were loaded into the ADF such that the scan occurred down the

Table I. List of feature descriptors used for classification.

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

long axis of the page. Scans were performed at 75, 150, 300, 600, and 1200 ppi at 24-bit RGB resolution. All scans were saved in a lossless portable network graphics (PNG) file format. Using in-house image processing software, each scanned page was analyzed to extract each individual letter from the page. The in-house image processing software, coded in C#, performed extraction in the following manner. Images were binarized using a variant of the method by Kittler and Illingworth,[6] and then connected components were created after an erosion/dilation step to eliminate small ink regions (noise). Each region was encapsulated in a rectangular bounding-box and subsequently copied to an individual file.

After shape extraction, each letter was analyzed to extract 36 feature descriptors of the shape. Table I 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 Refs. 7 and 8, 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 (i.e., classification).

For classification, we selected the ADABOOST algorithm developed by Freund and Schapire,[9] and modified it to handle multiclass classification problems by using the approach described by Allwein et al.[10] We briefly describe the ADABOOST algorithm and multiclass approach below and refer the reader to Freund and Schapire[9] and to Allwein et al.[10] for additional details. For more information on classification approaches we refer the reader to Duda et al.[11]

Table II. 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).

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

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 Eq. (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 Eq. (1) is then the sign of $f(x)$ obtained from the sum of the weighted weak classifiers.

To train the ADABOOST classifier, one specifies the number of weak learners in parameter T, which also corresponds to the number of training iterations performed, and for each weak learner one iterates over the training set to determine the weightings. Given a training set of m samples, the error for the weak classifier at iteration t is defined as

$$\varepsilon_t = \sum_{i=1}^{m} D_t(i)[y_i \neq h_i(x_i)], \tag{2}$$

where D_t represents a distribution of weightings for the training samples. Prior to training, the weight values of D_t are uniformly initialized to $(1/m)$. The weight of the weak learner can then be updated by the function

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

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. As can be seen in Eq. (4), incorrectly classified training samples receive an increase in their weight values proportional to the weighting of the weak classifier used for a given training iteration. This forces the ADABOOST algorithm to focus on the more difficult training examples when building the model.

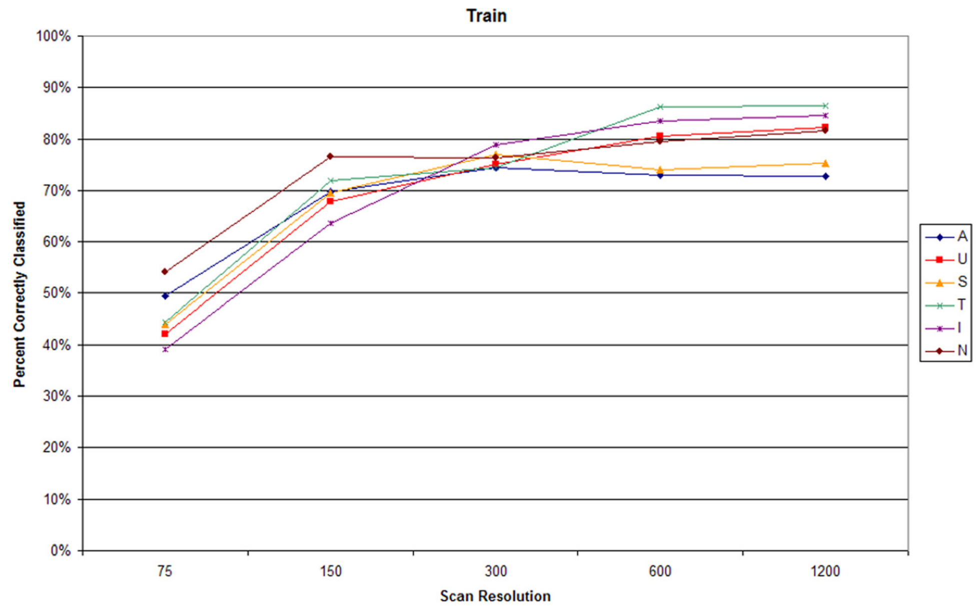

Figure 2. Mean classification accuracy of the training data from the six-class leave-one-out experiment.

To expand this to a multiclass 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,[10] such as one-against-all 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 Ref. 10, given a set of k classes, the multiclass 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 $\boldsymbol{M}$ of size k by k, which represents the class labels. Using the same values for indicator variables as defined in the original ADABOOST approach,[9] the diagonal of the matrix will contain $+1$ and all other elements will be -1. For example, in a three 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, s, which is nested inside the iterations over the weak learners. This allows the function to compute the weights over all classes.

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

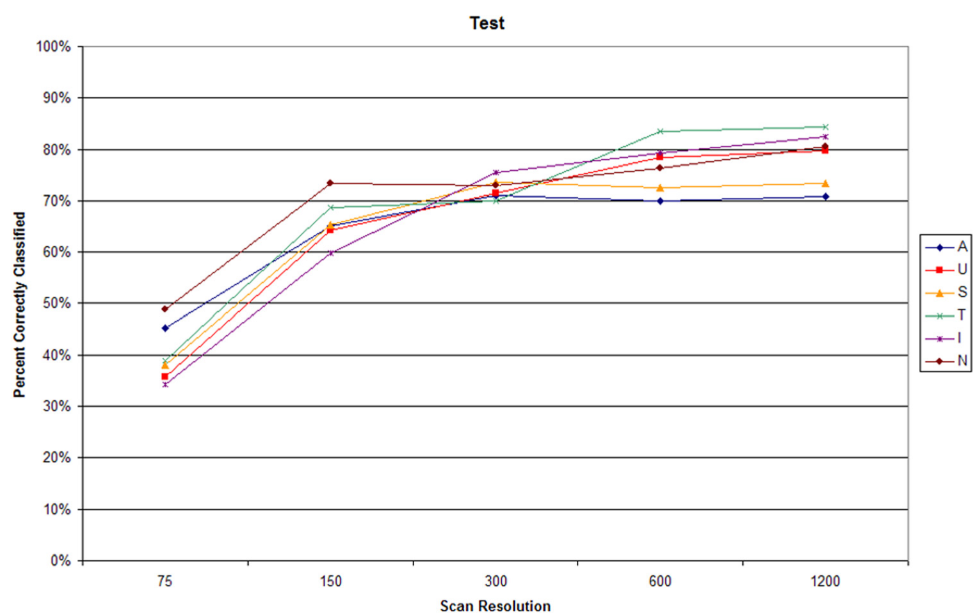

Figure 3. Mean classification accuracy of the test data from the six-class leave-one-out experiment.

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

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

As with Eq. (4), the value Z_t is a normalizing constant defined as $2\sqrt{\varepsilon_t(1-\varepsilon_t)}$. With regard to Eqs. (2) and (5), in the event that the error value, ϵ, exceeds 0.5 the training cycle is terminated since the value of α derived in Eq. (3) would otherwise be negative.

The last step is to redefine Eq. (1) to handle multiple classes. Using the concept of loss-based decoding,[10] the loss function can be defined as

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

In Eq. (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 multiclass 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

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

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

The weak learner selected for this implementation of ADABOOST was the decision stump. The decision stump, a one level decision tree with two leaf nodes, attempts to classify a sample by inspecting the value of a single feature and binning it to a class based on the relation to a split value determined from the training data. For the decision stumps used in this work, split values were determined by finding the value within a feature dimension, which corresponded to the largest information gain. Formally information gain (IG), can be defined as

$$IG(Y|X:\tau) = H(Y) - H(Y|X:\tau), \qquad (9)$$

where the entropy and conditional entropy values are computed as

$$H(Y) = -\sum_{i=1}^{k} p_i \log_2 p_i, \qquad (10)$$

$$\begin{aligned} H(Y|X:\tau) =& H(Y|X<\tau)P(X<\tau) \\ &+ H(Y|X\geq\tau)P(X\geq\tau). \end{aligned} \qquad (11)$$

Table III. 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.

Resolution	Letter	A						U					
		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
	6940(2)	20	39	0	11	2	28	28	20	8	24	8	12
	6280	0	1	70	0	24	5	0	0	71	0	21	8
	6127	14	19	0	64	1	2	21	25	3	42	5	4
	3600	1	2	34	0	39	24	4	4	57	1	22	12
	3005	8	13	7	3	19	50	11	11	26	15	22	15
150	6940(1)	37	42	0	9	4	8	41	32	0	12	0	15
	6940(2)	24	55	1	3	11	6	26	62	0	4	1	7
	6280	0	0	93	0	7	0	1	2	88	0	9	0
	6127	18	4	0	69	0	9	28	5	0	53	0	14
	3600	2	1	2	0	59	36	1	1	4	2	76	16
	3005	13	1	1	0	14	71	9	4	0	3	18	66
300	6940(1)	43	53	0	1	3	0	69	18	0	6	7	0
	6940(2)	39	59	0	0	2	0	55	40	0	0	5	0
	6280	0	2	94	0	4	0	0	0	99	0	1	0
	6127	17	2	0	75	0	6	18	0	0	72	1	9
	3600	1	0	0	0	97	2	0	0	1	0	95	4
	3005	3	1	0	0	17	79	5	0	0	2	25	68
600	6940(1)	92	7	0	0	0	1	77	16	0	6	0	1
	6940(2)	92	7	0	0	1	0	47	51	0	0	0	2
	6280	0	0	97	0	3	0	0	0	100	0	0	0
	6127	40	0	0	44	4	12	15	1	0	73	1	10
	3600	0	0	23	0	73	4	0	1	1	0	76	22
	3005	0	0	0	0	13	87	2	0	0	2	7	89
1200	6940(1)	90	10	0	0	0	0	88	7	0	3	0	2
	6940(2)	95	5	0	0	0	0	54	43	0	0	0	3
	6280	5	3	92	0	0	0	0	0	99	0	0	1
	6127	20	0	0	72	2	6	6	2	0	76	0	16
	3600	0	0	20	0	77	3	0	2	2	0	85	11
	3005	0	0	0	0	1	99	0	0	0	1	2	97

Resolution	Letter	S						T					
		6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005
75	6940(1)	25	24	5	20	7	19	25	29	0	34	0	12
	6940(2)	26	31	6	18	3	16	19	26	1	38	3	13
	6280	1	5	56	0	18	20	0	1	75	0	19	5
	6127	20	26	1	46	3	4	11	17	1	61	1	9
	3600	1	4	43	0	33	19	1	0	58	0	14	27
	3005	12	22	17	2	22	25	14	13	23	10	10	30
150	6940(1)	50	26	0	16	2	6	56	26	0	10	0	8
	6940(2)	45	44	2	3	4	2	27	69	0	0	0	4
	6280	0	2	94	0	3	1	0	2	91	0	4	3
	6127	26	6	0	66	0	2	21	4	0	73	1	1
	3600	0	0	4	0	78	18	0	0	6	0	70	24
	3005	20	5	0	4	9	62	20	7	0	2	14	57
300	6940(1)	51	25	0	3	7	14	65	18	0	8	1	8
	6940(2)	33	64	0	0	2	1	57	39	0	1	0	3
	6280	3	2	92	0	3	0	4	0	94	0	2	0
	6127	8	0	0	80	4	8	18	0	0	78	0	4
	3600	1	0	1	0	82	16	3	0	5	0	60	32
	3005	3	0	0	0	12	85	7	0	0	9	13	71
600	6940(1)	94	4	0	0	0	2	86	9	0	4	0	1
	6940(2)	77	21	0	0	2	0	53	46	0	0	0	1
	6280	2	0	98	0	0	0	0	1	98	0	1	0
	6127	18	0	0	68	8	6	7	0	0	86	1	6
	3600	2	0	6	0	52	40	0	0	2	0	84	14
	3005	3	0	0	0	2	95	3	0	0	3	6	88
1200	6940(1)	84	8	0	0	7	1	72	16	1	3	0	8
	6940(2)	70	27	0	0	3	0	32	66	1	1	0	0
	6280	7	1	92	0	0	0	3	1	96	0	0	0
	6127	7	0	0	69	16	8	12	0	0	84	0	4
	3600	0	0	7	0	74	19	0	0	1	0	92	7
	3005	0	0	0	0	2	98	3	0	0	1	2	94

Resolution	Letter	I						N					
		6940(1)	6940(2)	6280	6127	3600	3005	6940(1)	6940(2)	6280	6127	3600	3005
75	6940(1)	26	18	13	31	11	1	28	39	0	21	1	11
	6940(2)	20	30	13	22	10	5	23	59	1	11	1	5
	6280	5	7	60	0	22	6	0	2	70	0	17	11
	6127	29	13	2	48	6	2	19	19	0	62	0	0
	3600	9	11	38	0	26	16	1	4	39	0	28	28
	3005	22	15	26	15	14	8	11	7	16	1	32	33
150	6940(1)	49	18	0	14	0	19	66	18	0	10	4	2
	6940(2)	23	52	2	5	1	17	13	86	1	0	0	0
	6280	0	0	78	0	15	7	2	3	88	0	6	1
	6127	16	4	0	73	0	7	9	4	0	83	2	2
	3600	0	0	12	0	57	31	0	0	7	0	66	27
	3005	23	6	1	0	19	51	1	3	1	0	27	68
300	6940(1)	70	20	0	3	0	7	82	5	0	6	6	1
	6940(2)	45	52	1	0	2	0	44	54	0	1	1	0
	6280	0	1	98	0	1	0	5	4	82	0	9	0
	6127	19	1	0	79	0	1	2	1	0	89	5	3
	3600	0	0	2	0	85	13	0	0	0	0	85	15
	3005	18	1	1	1	10	69	1	0	0	1	52	46
600	6940(1)	73	12	0	9	0	6	98	1	0	1	0	0
	6940(2)	21	76	0	0	0	3	75	25	0	0	0	0
	6280	1	0	99	0	0	0	5	2	79	0	14	0
	6127	9	0	0	76	0	15	31	0	0	59	0	10
	3600	2	0	8	0	80	10	0	0	0	0	100	0
	3005	8	2	0	3	6	81	0	0	0	0	37	63
1200	6940(1)	86	10	0	3	0	1	93	3	1	1	2	0
	6940(2)	48	51	0	0	0	1	51	48	1	0	0	0
	6280	4	0	95	0	1	0	20	1	75	0	4	0
	6127	0	0	0	87	2	11	13	0	0	84	3	0
	3600	0	0	3	0	96	1	0	0	2	0	89	9
	3005	16	0	1	6	17	60	0	0	1	0	7	92

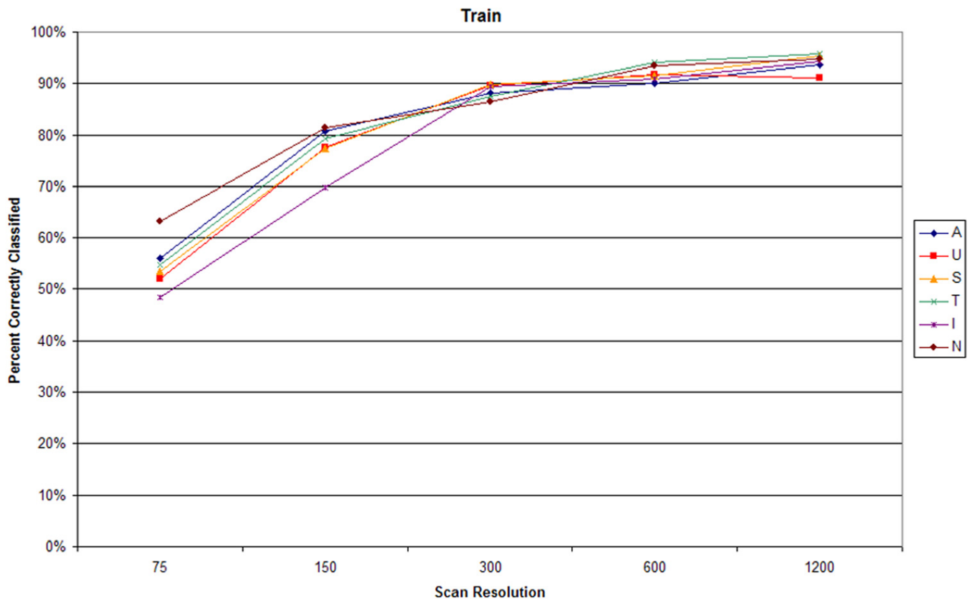

Figure 4. Mean classification accuracy of the training data from the five-class leave-one-out experiment.

In Eqs. (9) and (11), the value τ represents the split point and Eq. (10) is the standard entropy calculation introduced by Shannon[12], where p_i is the probability of being in class *i*. To find the split value corresponding with the largest information gain for a specific feature, values from the training data are sorted in ascending order and the information gain is computed at each unique value. The feature value corresponding to the maximum IG value is then selected as the split point. This process is performed over all feature dimensions for the training data and is precomputed prior to training the ADABOOST model.

EXPERIMENTAL RESULTS

Three sets of experiments were performed to determine the effects of scan resolution and shape on classification accuracy. For the first two experiments, we used six printers. For the third experiment, we used only two of the six printers. The printers used in all of the experiments are listed in Table II. As described in the "Methods" section, five pages of the digital raster were printed for each printer at 600 dpi. All prints were performed using the *normal print quality* settings. For all color printers, the settings were modified so that printing used only black ink or toner. All printing used the same brand of office paper.

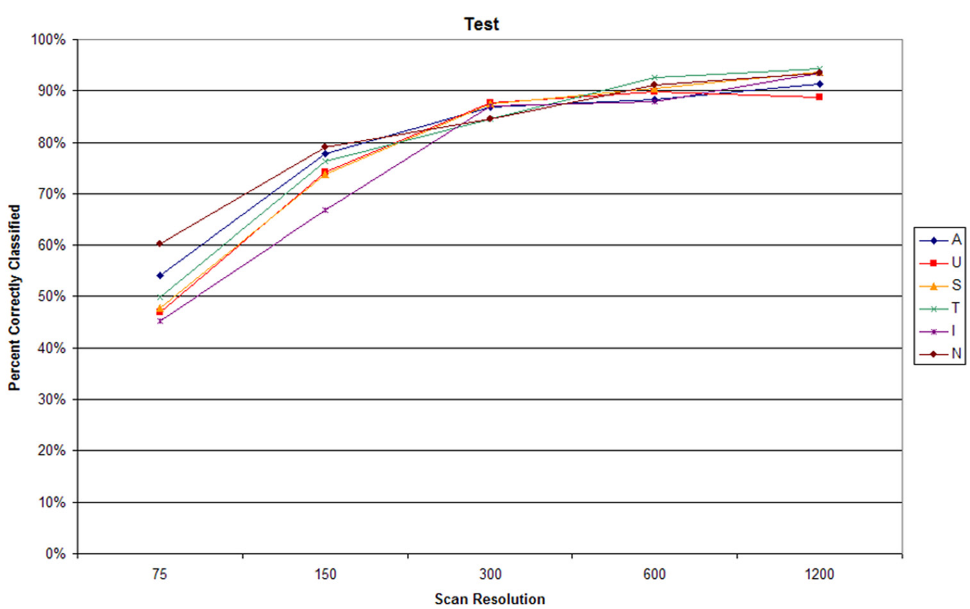

Figure 5. Mean classification accuracy of the test data from the five-class leave-one-out experiment.

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

In the first experiment, all six 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 one is the test sheet are listed in Table III. The actual class of a sample is listed along the left hand columns of the table, while the predicted class is listed across the top of the table. 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 five-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 accuracies of the training and test sets are plotted in Figures 4 and 5 and the confusion matrices for the test results using the first page are in Table IV.

For the third experiment, we ran a two-class problem where only the two instances of the 6940 Deskjet™ printer were analyzed. Table V shows the confusion matrices for all letters and resolutions from the results of the first test page and Figures 6 and 7 depict the mean classification accuracy.

For all of the tests, printer identification was performed by a simple majority vote. Looking at each row of a confusion matrix, the printer selected as the source was identified as the printer having the largest count of classified letters in its column.

DISCUSSION

Based upon the results presented in Tables III and IV, several trends can be identified. First, it is clear that as resolution increases, the overall accuracy of classification

Table IV. 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.

Resolution	Letter	A					U					S				
		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
	6280	3	79	0	14	4	1	67	0	25	7	5	49	0	19	27
	6127	73	0	26	1	0	71	2	21	4	2	65	1	30	2	2
	3600	10	44	0	34	12	11	59	3	17	10	6	40	0	36	18
	3005	42	12	1	32	13	37	29	12	12	10	44	14	2	24	16
150	6940	164	1	9	17	9	181	0	5	1	13	171	2	12	5	10
	6280	0	93	0	7	0	3	89	0	8	0	1	95	0	3	1
	6127	21	0	68	0	11	39	0	45	0	16	43	0	55	0	2
	3600	5	2	0	57	36	2	4	2	76	16	1	3	0	79	17
	3005	10	1	3	13	73	18	0	3	19	60	24	0	3	9	64
300	6940	194	0	0	5	1	197	0	2	1	0	191	0	0	4	5
	6280	2	94	0	4	0	0	99	0	1	0	6	91	0	3	0
	6127	14	0	81	1	4	9	0	82	2	7	7	0	80	7	6
	3600	0	0	0	97	3	0	3	0	92	5	0	1	0	92	7
	3005	2	0	0	16	82	0	0	2	26	72	4	1	0	15	80
600	6940	197	0	2	1	0	197	0	3	0	0	195	0	0	1	4
	6280	0	97	0	3	0	0	100	0	0	0	0	99	0	0	1
	6127	17	0	66	3	14	8	0	79	2	11	9	0	81	6	4
	3600	0	18	0	78	4	0	1	0	76	23	0	0	0	62	38
	3005	0	0	0	7	93	0	0	2	3	95	1	0	0	3	96
1200	6940	198	0	2	0	0	194	0	2	0	4	199	0	0	1	0
	6280	6	92	2	0	0	1	98	0	0	1	4	96	0	0	0
	6127	7	0	86	1	6	17	0	69	0	14	6	0	87	1	6
	3600	0	14	0	83	3	0	2	0	86	12	1	0	0	84	15
	3005	0	0	0	1	99	0	0	1	2	97	0	0	0	0	100

Resolution	Letter	T					I					N				
		6940	6280	6127	3600	3005	6940	6280	6127	3600	3005	6940	6280	6127	3600	3005
75	6940	170	1	16	7	6	150	25	21	3	1	165	0	21	4	10
	6280	1	74	0	17	8	13	68	0	14	5	0	66	0	21	13
	6127	73	1	19	1	6	73	0	27	0	0	44	0	56	0	0
	3600	4	55	0	19	22	32	44	0	15	9	2	33	0	33	32
	3005	38	23	3	18	18	60	25	2	5	8	21	15	0	28	36
150	6940	181	0	10	1	8	153	2	18	1	26	193	1	3	2	1
	6280	1	91	0	5	3	0	80	0	13	7	3	90	0	5	2
	6127	34	0	66	0	0	27	0	66	0	7	27	0	69	1	3
	3600	0	5	0	73	22	0	12	0	57	31	0	7	0	66	27
	3005	30	1	2	15	52	34	2	0	18	46	6	1	0	28	65
300	6940	193	0	1	0	6	195	1	1	0	3	193	1	3	2	1
	6280	1	96	0	2	1	1	98	0	1	0	7	85	0	8	0
	6127	11	0	85	0	4	8	0	88	0	4	7	0	84	5	4
	3600	0	4	0	62	34	0	2	0	85	13	0	0	0	85	15
	3005	3	0	7	17	73	5	1	3	10	81	0	0	0	52	48
600	6940	198	0	2	0	0	191	0	3	0	6	198	0	1	1	0
	6280	1	98	0	1	0	1	99	0	0	0	3	85	0	12	0
	6127	6	0	87	0	7	6	0	81	0	13	5	0	84	0	11
	3600	0	2	0	87	11	0	7	0	83	10	0	0	0	99	1
	3005	1	0	4	7	88	8	0	3	6	83	0	0	0	35	65
1200	6940	196	0	2	0	2	198	0	0	0	2	198	2	0	0	0
	6280	4	96	0	0	0	0	99	0	1	0	8	89	0	3	0
	6127	9	0	89	0	2	1	0	88	0	11	5	0	92	1	2
	3600	0	0	0	93	7	0	3	0	94	3	0	0	0	89	11
	3005	2	0	1	1	96	0	0	4	13	83	0	0	0	2	98

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

Resolution	Letter	A 6940(1)	A 6940(2)	U 6940(1)	U 6940(2)	S 6940(1)	S 6940(2)	T 6940(1)	T 6940(2)	I 6940(1)	I 6940(2)	N 6940(1)	N 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

increases for the six-class and five-class experiments. Fig. 2 through 5 substantiate this claim as do the confusion matrices in Tables III and IV. However, the results for the six-class problem also indicate that in many of the tests, classification accuracy for distinguishing between the two 6940 model printers was low. Looking at Table III, one will notice that by 150 ppi scan resolution, misclassification primarily occurs in two areas. The 6940 printers are misclassified between each other, as are the two laser printers (3005 and 3600). If one was to apply a simple majority test as the means for determining the source printer, it would fail at a number of resolution and letter combinations due to the 6940. As the resolution continues to increase, the laser printers become correctly distinguished from one-another and again a simply majority vote would correctly identify the source printer except in the case of the 6940.

If the classification problem is changed into a five-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. As the resolution increases, the accuracy trends upward and the stronger the vote count for each source printer model. 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 two-class problem in which the objective was to distinguish one 6940 from the other. Looking again at the results in Table V, 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. For these binary tests, the peak classification accuracy appears to be at 600 ppi and this is best seen in Figs. 6 and 7. For the letter N, the mean accuracy of the test data is 94.5% (sample standard deviation $s = 2.9\%$) at this resolution but drops slightly to 91.9% ($s = 3.8\%$) at 1200 ppi. Other letters exhibited a stronger downward trend from 600 to 1200 ppi. The test data for the letter S, for example, is classified with a mean accuracy of 90.5% ($s = 4.4\%$) at 600 ppi, but drops to 83% mean accuracy ($s = 1.1\%$) at 1200 ppi. The letter I on

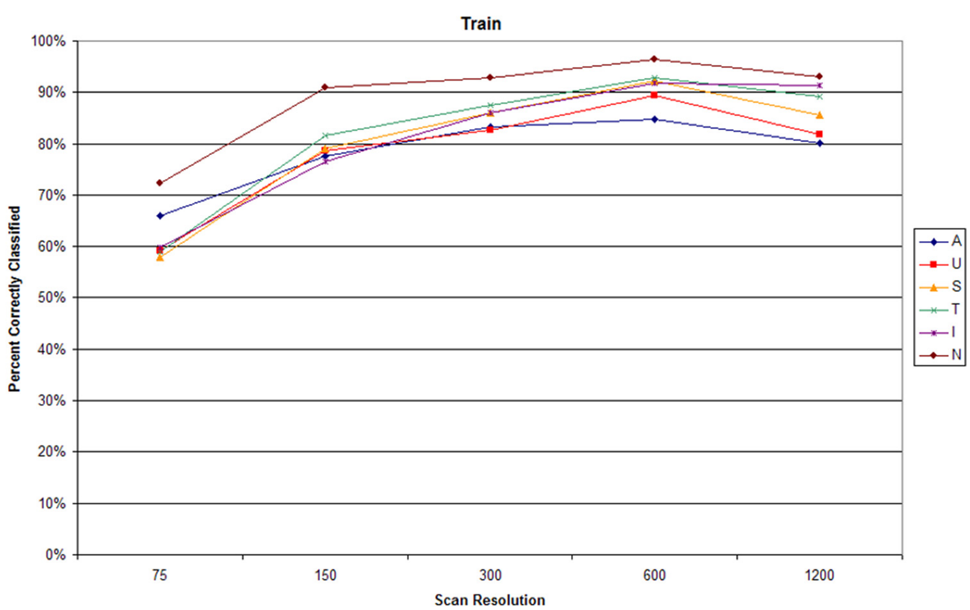

Figure 6. Mean classification accuracy of the training data from the 6940 binary classification experiment.

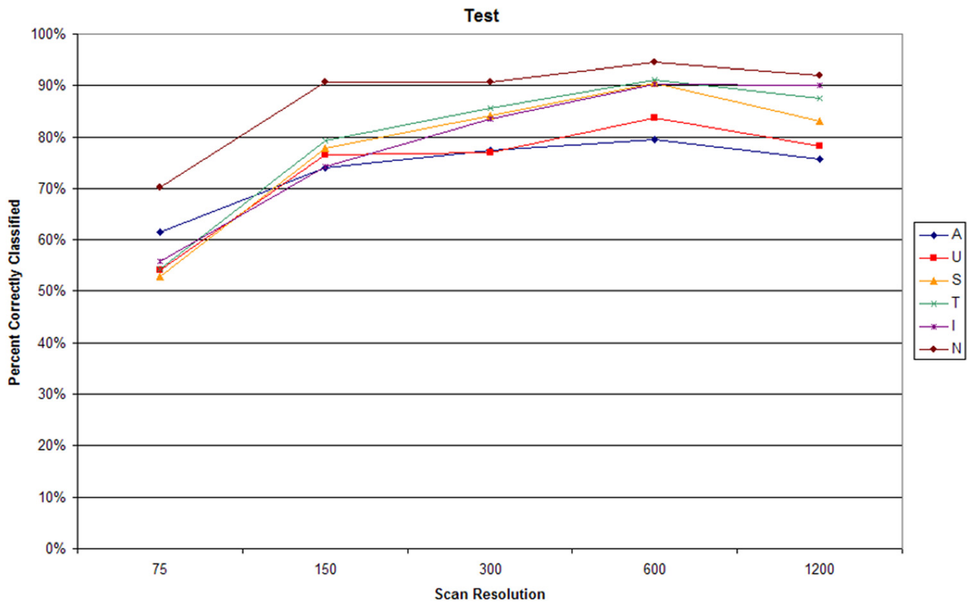

Figure 7. Mean classification accuracy of the test data from the 6940 binary classification experiment.

the other hand dropped marginally from 90.2% ($s=2.4\%$) to 90% ($s=3.6\%$).

Upon closer inspection of the CCD configuration of the scanner, we discovered that different sensors within the CCD were used depending on the specified resolution. For the experiments using scan resolutions between 75 and 600 ppi, a set of 600 ppi native resolution linear CCD arrays were utilized to perform the document scans. For the scans performed at 1200 ppi resolution, a different set of linear CCD arrays with a native resolution of 1200 ppi were utilized. We believe that the drop in classification accuracy between 600 and 1200 ppi can be primarily attributed to the different CCD arrays selected for use by the scanner. However, as previously stated, using our approach with scan resolutions as low as 150 ppi, 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 multiclass 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.

CONCLUSIONS

Using a multiclass ADABOOST classifier, we performed a sensitivity analysis study to examine how scan resolution and shape affect the ability to correctly identify samples with their original source printers. In classification experiments where multiple instances of a printer make and model are treated as a single class, source printer identification using a simple majority vote is successful with resolutions as low as 150 ppi. Performing a binary classification and identification test on two instances of the same make and model printer also resulted in 100% correct printer identification with a resolution as low as 150 dpi. These overall results indicate that, with a resolution as low as 150 ppi, it is possible to identify the source printer by first performing classification and identification on makes and models followed by a focused classification analysis on the make and model instances in question. These results are an advancement over previous studies and findings,[1,2] since we are able to use a much lower resolution to attain printer identification.

FUTURE WORK

Based on our findings, a number of unanswered questions, which warrant further research remain. First, our approach for identifying the specific instance of a make and model printer is 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 are in question. Second, 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. Lastly, an in-depth analysis also needs to be performed to assess the effectiveness of our approach when using laser printers.

REFERENCES

[1] 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," *Proc. IS&T's NIP21: Int. Conference on Digital Printing Technologies* (IS&T, Springfield, VA, 2005), pp. 223–226.

[2] A. K. Mikkilineni, P. J. Chiang, G. M. Ali, G. T. C. Chiu, J. P. Allebach, and E.J. Delp, "Printer identification based on gray level co-occurrence

features for security and forensic applications," Proc. SPIE **5681**, 430–440 (2005).
[3] E. Kee and H. Farid, "Printer profiling for forensics and ballistics," *Proc. 10th ACM Workshop on Multimedia and Security* (ACM, New York, 2008), pp. 3–10.
[4] O. Bulan, J. Mao, and G. Sharma, "Geometric distortion signatures for printer identification," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (IEEE Computer Society, Los Alamitos, CA, 2009), pp. 1401–1404.
[5] J. S. Aronoff and S. J. Simske, "Determining printer and scanner resolution dependency of text classification for digital image forensics," *Proc. IS&T's NIP26: Int. Conference on Digital Printing Technologies* (IS&T, Springfield, VA, 2010), pp. 626–631.
[6] J. Kittler and J. Illingworth, "Minimum error thresholding," J. Pattern Recognit. **19**, 41 (1986).
[7] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. (Prentice-Hall, Inc., Upper Saddle River, NJ, 2008)
[8] L. G. Shapiro and G. C. Stockman, *Computer Vision* (Prentice-Hall, Inc., Upper Saddle River, NJ, 2001).
[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Proc. Second European Conf. on Computational Learning Theory*, Lecture Notes in Computer Science (Springer), Vol. 904, pp. 23–37 (1995).
[10] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," J. Mach. Learn. Res. **1**, 113–141 (2001).
[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (John Wiley & Sons, New York, 2001).
[12] C. E. Shannon, "Prediction and entropy of printed english," Bell Syst. Tech. J. **30**, 50 (1951).