Printer Forensics using SVM Techniques

Aravind K. Mikkilineni[†], Osman Arslan[†] , Pei-Ju Chiang[‡], Roy M. Kumontoy[†], Jan P. Allebach[†], George T.-C. Chiu[‡], Edward J. Delp[†]; [†]School of Electrical and Computer Engineering, [‡]School of Mechanical Engineering, Purdue University; West Lafayette, Indiana, United States of America

Abstract

In today's digital world securing different forms of content is very important in terms of protecting copyright and verifying authenticity. We have previously described the use of image texture analysis to identify the printer used to print a document. In particular we described a set of features that can be used to provide forensic information describing a document. In this paper we will introduce a printer identification process that uses a support vector machine classifier. We will also examine the effect of font size, font type, paper type, and "printer age".

Introduction

In today's digital world securing different forms of content is very important in terms of protecting copyright and verifying authenticity. [1,2] One example is watermarking of digital audio and images. We believe that a marking scheme analogous to digital watermarking but for documents is very important.[1] Printed material is a direct accessory to many criminal and terrorist acts. Examples include forgery or alteration of documents used for purposes of identity, security, or recording transactions. In addition, printed material may be used in the course of conducting illicit or terrorist activities. In both cases, the ability to identify the device or type of device used to print the material in question would provide a valuable aid for law enforcement and intelligence agencies. We also believe that average users need to be able to print secure documents, for example boarding passes and bank transactions.

There currently exist techniques to secure documents such as bank notes using paper watermarks, security fibers, holograms, or special inks.[3] The problem is that the use of these security techniques can be cost prohibitive. Most of these techniques either require special equipment to embed the security features, or are simply too expensive for an average consumer. Additionally, there are a number of applications in which it is desirable to be able to identify the technology, manufacturer, model, or even specific unit that was used to print a given document.

We propose to develop two strategies for printer identification based on examining a printed document. The first strategy is passive. It involves characterizing the printer by finding intrinsic features in the printed document that are characteristic of that particular printer, model, or manufacturer's products. We shall refer to this as the *intrinsic signature*. The intrinsic signature requires an understanding and modeling of the printer mechanism, and the development of analysis tools for the detection of the signature in a printed page with arbitrary content.

The second strategy is active. We embed an *extrinsic signature* in a printed page. This signature is generated by modulating the process parameters in the printer mechanism to encode identifying information such as the printer serial number and date of printing. To detect the extrinsic signature we use the tools developed for intrinsic signature detection. We have successfully been able to embed information into a document with electrophotographic (EP) printers by modulating an intrinsic feature known as *banding*. This work is discussed in [4].

We have previously reported techniques that use the print quality defect known as *banding* in electrophotographic (EP) printers as an intrinsic signature to identify the model and manufacturer of the printer.[5,6] However, it is difficult to detect the banding signal in text. One solution which we have reported in [7] is to model the print quality defects as a texture in the printed areas of the document. To classify the document we used grayscale co-occurrence texture features. These features can be measured over small regions of the document such as individual text characters. Using these features we demonstrated the ability to process a page of printed text and correctly identify the printer that created it.

In our prior work, we did not account for several variables in our printer identification process. The type of paper, font type, font size, printer age, and other variables can affect the performance of our proposed classifier. We will examine the effects of these variables in this paper. We will also introduce a modified system using a support vector machine (SVM) classifier which provides better generalization than the nearest neighbor classifier previously used.

This research was supported by a grant from the National Science Foundation, under Award Number 0219893. Address all correspondence to E. J. Delp at ace@ecn.purdue.edu

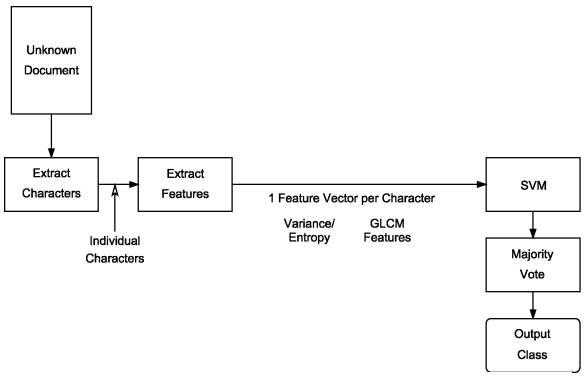

Figure 1. Process diagram for printer identification

Table 1: Percent correct classification for varying font type

Manufacturer	Model	DPI
Hewlett-Packard	LaserJet 5M	600
Hewlett-Packard	LaserJet 6MP	600
Hewlett-Packard	LaserJet 1000	600
Hewlett-Packard	LaserJet 1200	600
Lexmark	E320	1200
Samsung	ML-1430	600
Samsung	ML-1450	600
Brother	HL-1440	1200
Minolta-QMS	1250W	1200
Okidata	14e	600

System Overview

Figure 1 shows the block diagram of our printer identification scheme. Given a document with an unknown source, referred to as the *unknown document*, we want to identify the printer that created it.

The first step is to scan the document at 2400 dpi with 8 bits/pixel (grayscale). Next all the letter "e"s in the document are extracted. The reason for this is that "e" is the most frequently occurring character in the English language.

A set of features are extracted from each character forming a feature vector for each letter "e" in the document. These features are obtained from simple pixel level statistics and from the graylevel co-occurence matrix (GLCM) as described in [7]. Each feature vector is then individually classified using an SVM.

The SVM classifier is trained using 5000 known feature vectors. The training set consists of 500 feature vectors from each of the 10 printers listed in Table 1. Each of these feature vectors are independent of one another.

Let Φ be the set of all printers $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ (in our work these are the 10 printers shown in Table 1) For any $\varphi \epsilon \Phi$, let $c(\varphi)$ be the number of "e"s classified as being printed by printer φ. The final classification is decided by choosing φ such that $c(\varphi)$ is maximum. In other words, a majority vote is performed on the resulting classifications from the SVM.

SVM Classifier

In our previous work we used a 5-Nearest-Neighbor (5NN) classifier in place of the SVM [7]. The reason for investigating an SVM based classifier is that the 5NN classifier does not generalize well when the ratio of training vectors to dimension is relatively low. The SVM is able to provide better generalization in this scenario.

An SVM classifier maps input vectors into a high dimensional space through a nonlinear mapping. Optimal separating hyperplanes are then constructed in the high dimensional space.[8,11] The decision function which realizes this system is given by

$$f(x)=sign(\Sigma_{Support\ Vectors}\ y_i\alpha_i K(x_i,x)-b) \qquad (1)$$

where $K(x_i,x)$, the kernel function, performs the scalar product on its arguments in the higher dimensional space.[9] In our experiments we chose $K(x_i,x)$ as the radial basis function (RBF)

$$K(x_i,x)=exp\{-\gamma||x_i-x||^2\} \qquad (2)$$

The method we used for SVM training and classification is described in [10]. Using this procedure we compared the SVM and 5NN classifier using a “12pt Times” training and testing data sets.

Using the 5NN classifier, 9 out of 10 printers are correctly classified after the majority vote. The printer that was not correctly classified was the *lj1200*. Classification of the *lj1200* was ambiguous because the majority vote had to choose between two equally weighted classes, *lj1000* and *lj1200*. This can be explained by the fact that these two printers seem to share the same print engine. The classification accuracy before the majority vote was only 52.4% using this classifier.

Using the SVM, all 10 printers were correctly classified after the majority vote and the classification accuracy before the majority vote is 93.0%. This implies that we can expect less ambiguity with the majority vote.

Test Variables and Procedure

Four variables are considered in our experiment. These variables are listed in Table 2. In our previous work we considered only 12pt Times text printed using one type of paper.

We would like to know whether our printer identification technique works for other font sizes, font types, paper types, and age difference between training and testing data sets.

Four cases will be explored. In each case the training set will consist of 500 "e"s and the test set will consist of 300 "e"s. As described in [7], using our Forensic Monkey Text Generator (FMTG) we estimated that testing using 300 "e"s is representative of a typical page of printed English text.

The first case considered is where the printer identification system is trained using data of font size fs_{train} and tested using data of font size fs_{test} with all other variables held constant (*ft*=Times; *pt*=PP-0001). It is assumed that printing the training and test data immediately after one another holds age constant in this case.

The second case is where the system is trained using data of font type ft_{train} and tested using data of font type ft_{test} with all other variables held constant (*fs*=12pt; *pt*=PP-0001).

In the third case the system is trained using data of paper type pt_{train} and tested using data of paper type pt_{test} with all other variables held constant (*fs*=12pt; *ft*=Times).

Finally we consider the case where the system is trained on "old" data and tested on "new." We used testing and training data sets printed 5 months apart. 10 sub-cases are considered by testing and training using data from the sets {fs_x,Times,PP-0001} and {12pt,ft_x,PP-0001}. This is representative of a forensic scenario where the printing device that created a suspect document needs to be identified given only the document in question and newly generated test and training data from the printer.

Results

The results for case 1 are shown in Table 3. The rows of the table correspond to the value of fs_{train} during training, and the columns correspond to the value of fs_{test} . Each entry contains two values. The first value is the percent correct classification of the system (i.e. the percentage of printers classified correctly from those listed in Table 1). The second value, surrounded by parentheses, is the percent correct classification of the individual feature vectors immediately after the SVM. From the table we find that when the font sizes of the training and testing data are within 2 points of each other, at least 9 out of 10 printers are correctly classified.

The results for case 2 are shown in Table 4. These results show that our current feature set is very font dependent. If $ft_{train}=ft_{test}$ we can classify 9 out of 10 printers correctly. At most 7 out of 10 printers are classified correctly if

Table 2: Four variables considered in our experiments

Category	Sub-Types
Font size (*fs*)	08pt
	10pt
(eeeee)	12pt
	14pt
	16pt
Font type (*ft*)	Arial
	Courier
(eeeee)	Garamond
	Impact
	Times
Paper type (*pt*)	PP-0001: 20lb, 84brt
	PP-0006: 28lb, 97brt
	PP-0008: 32lb, 100% cotton
Age (consumables)	-

$ft_{train}\neq ft_{test}$. Even though the font size was 12pt for each font type, the height of the "e" in each instance was different as seen in Table 2. It is possible that this implicit font size difference partly caused the low classification rates for different font types. The Times "e" and Courier "e" are approximately the same height and the classification rate for training on Times and testing on Courier is shown to be 70%.

The results for different paper types, case 3, are shown in Table 5. We obtain 100% correct classification if both the training and testing sets use the same paper type. If we train using paper type PP-0001 or PP-0006, and test on PP-0001 or PP-0006, then at least 9 out of 10 printers are classified correctly. The same is not true for paper type PP-0008. Paper types PP-0001 and PP-0006 are both visually similar except that PP-0006 appears slightly smoother and brighter. Paper type 8 has a visually rougher texture than the other two paper types, possibly due to the 100% cotton content. The features we use might be affected by the paper texture as well as textures from the printer itself.

Table 6 shows the results for the fourth case, training with new data and testing with old. At least 7 out of 10 printers are correctly identified in each sub-case. The individual SVM classifications (which are not shown due to space restrictions) show that in each of these sub-cases, the *lj1200* was classified as an *lj1000*. We observed this behavior in previous work and attribute it to the fact that the two printers appear to have the same print engine.

Conclusion

From our results we find that our printer identification technique works for various font sizes, font types, paper types, and printer age when those variables are held constant. In the case where font size or font type varies between the

Table 3: Percent correct classification for varying font size (% after SVM)

		Test				
		8pt	10pt	12pt	14pt	16pt
Train	8pt	100 (87.6)	90 (82.9)	80 (61.0)	50 (43.0)	40 (35.1)
	10pt	100 (78.3)	100 (95.3)	90 (72.9)	70 (56.3)	50 (47.9)
	12pt	80 (58.3)	90 (73.3)	100 (93.0)	100 (84.1)	80 (66.0)
	14pt	50 (43.6)	70 (62.7)	100 (88.9)	90 (89.7)	90 (81.2)
	16pt	40 (37.6)	50 (48.1)	80 (74.4)	90 (84.2)	90 (89.5)

Table 4: Percent correct classification for varying font type (% after SVM)

		Test				
		arial	courier	garamond	impact	times
Train	arial	90 (84.1)	40 (35.0)	40 (26.0)	20 (17.8)	40 (34.7)
	courier	20 (23.0)	90 (86.8)	50 (43.8)	0 (2.6)	50 (49.3)
	garamond	10 (12.4)	40 (43.2)	90 (82.3)	10 (11.9)	20 (27.8)
	impact	10 (16.8)	10 (10.4)	10 (11.4)	90 (82.9)	10 (17.9)
	times	20 (30.1)	70 (57.0)	40 (33.0)	10 (6.6)	90 (84.0)

Table 5: Percent correct classification for varying paper type (% after SVM)

		Test		
		PP-0001	PP-0006	PP-0008
Train	PP-0001	100 (93.0)	90 (83.3)	60 (47.2)
	PP-0006	90 (75.2)	100 (93.2)	40 (32.4)
	PP-0008	50 (40.4)	30 (28.1)	100 (93.0)

Table 6: Percent correct classification for varying age (testing data 5 months older than training data)

$fs_{train}fs_{test}$	08pt	10pt	12pt	14pt	16pt
%system %SVM	90 (66.0)	90 (76.3)	90 (72.3)	80 (66.9)	80 (67.8)

$ft_{train}ft_{test}$	Arial	Courier	Garamond	Impact	Times
%system %SVM	70 (64.6)	70 (62.6)	80 (67.3)	80 (67.0)	80 (58.5)

testing and training set, further study can be done to understand the effects those variable have on the GLCM features used for classification. It might be possible to "normalize" the features given prior knowledge of the font size and type.

Also from a forensics viewpoint the results from Table 6 are promising. The underlying SVM classification rates are low compared to those corresponding to equivalent system classification rates shown in Table 3 and 4. Some of the same issues mentioned for further study for font size and type could also be used to improve the underlying classification results in this case.

References

1. E. J. Delp, "Is your document safe: An overview of document and print security," *Proceedings of the IS&T International Conference on Non-Impact Printing*, San Diego, California, September 2002.
2. A. M. Eskicioglu and E. J. Delp, "An overview of multimedia content protection in consumer electronics devices," *Signal Processing:Image Communication*, vol. 16, pp. 681–699, 2001.
3. R. L. Renesse, Optical Document Security. Boston: Artech House, 1998.
4. P.-J. Chiang, G. N. Ali, A. K.Mikkilineni, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, "Extrinsic signatures embedding using exposure modulation for information hiding and secure printing in electrophotographic devices," *Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies*, vol. 20, Salt Lake City, UT, October/November 2004, pp. 295–300.
5. 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," *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VI*, vol. 5306, San Jose, CA, January 2004, pp. 455–466.
6. 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," *Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies*, vol. 20, Salt Lake City, UT, October/November 2004, pp. 301–305.
7. A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T. Chiu, J. P. Allebach, E. J. Delp, " Printer identification based on graylevel co-occurrence features for security and forensic applications," *Proceedings of the SPIE International conference on Security, Steganography, and Watermarking of Multimedia Contents VII*, vol. 5681, pp. 430-440, March 2005.
8. Vladimir N. Vapnik, The Nature of Statistical Signal Processing. New York, NY: Springer-Verlag, 1995.
9. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181-202, March 2001.
10. C.-W. Hsu, C.-C. Chang, C.-J. Lin, "A Practical Guide to Support Vector Classification," http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2005.
11. N. Cristianini, J. S. Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press, 2000.