Printer Identification Based on Texture Features

Aravind K. Mikkilineni, Pei-Ju Chiang, Gazi N. Ali
George T.-C. Chiu, Jan P. Allebach, Edward J. Delp
Purdue University
West Lafayette, Indiana, USA

Abstract

In today's digital world securing different forms of content is very important in terms of protecting copyright and verifying authenticity. Many techniques have been developed to protect audio, video, digital documents, images, and programs (executable code). One example is watermarking of digital audio and images. We believe that a similar type of protection for printed documents is very important. In this paper we describe the use of image texture analysis to identify the printer used to print a document. In particular we will describe a set of features that can be used to provide forensic information about a document. We will demonstrate our methods using 10 EP printers.

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

Introduction

In today's digital world securing different forms of content is very important in terms of protecting copyright and verifying authenticity.[1, 2, 3] 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.[4] 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. Examples include instruction manuals, team rosters, meeting notes, and correspondence. 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.[5, 6] 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 modelling 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 in is discussed in [7].

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.[8, 9, 10] We showed that different printers have different sets of *banding frequencies* which are dependent upon brand and model. This feature is relatively easy to estimate from documents with large midtone regions. However, it is difficult to estimate the banding frequencies from text. The reason for this is that the banding feature is present in only the process direction and in printed areas. The text acts as a high energy noise source upon which the low energy banding signal is added.

One solution is to find a feature or set of features which

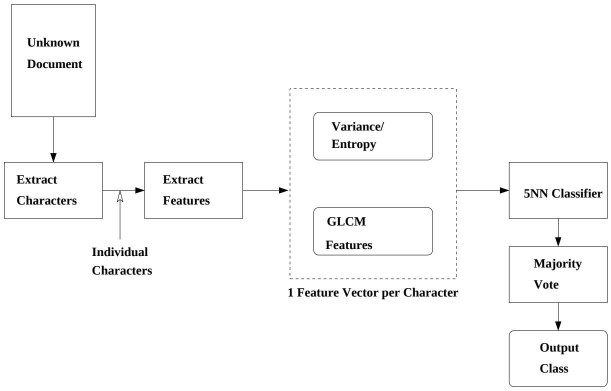

Figure 1: *System diagram of printer identification scheme*

Manufacturer	Model	DPI
Brother	hl1440	1200
HP	lj4050	600
Lexmark	e320	1200
HP	lj1000	600
HP	lj1200	600
HP	lj5M	600
HP	lj6MP	600
Minolta	1250W	1200
Okidata	14e	600
Samsung	ml1430	600

Table 1: *Printers used for classification.*

can be measured over smaller regions of the document such as individual text characters. If the banding signal is modelled as a texture in the printed areas of the document then texture features can be used to classify the document. These types of features can be more easily measured over small areas such as inside a text character.

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 be able to identify the printer that created it. For our test purposes we will use our Forensic Monkey Text Generator (FMTG) described in [8] to create random documents with known statistics to be classified.

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. Each feature vector is then classified individually using a 5-Nearest-Neighbor (5NN) classifier as described in [11].

The 5NN classifier is trained with 5000 known feature vectors. The training set is made up of 500 feature vectors from each of 10 printers listed in Table 1. Each of these feature vectors are independent of one another. To classify an unknown feature vector X, the euclidean distances between X and all the known feature vectors are obtained. A majority vote among the 5 smallest distances provides the classification result.

Let Ψ 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 $\phi\epsilon\Psi$, let $c(\phi)$ be the number of "e"s classified as being printed by printer ϕ. The final classification is decided by choosing ϕ such that $c(\phi)$ is maximum. In other words, a majority vote is performed on the resulting classifications from the 5NN classifier.

Graylevel Co-Occurrence Texture Features

We want to be able to determine a set of features that can be used to describe each printer uniquely by observing an example of the output of the printer. We will treat the output scanned document as an "image" and use image analysis tools to determine the features that characterize the printer. We will accomplish this by extracting features from individual printed characters, in particular an "e". Each character is very small, about 180x160 pixels and is non-convex, so it is difficult to perform any meaningful filtering operations in either the pixel or transform domain if we are interested only in the printed region of each character. The banding phenomenon introduces a periodic graylevel fluctuation in the printed character in the process direction. These variations can be modelled as textures in the character.[12] To model the texture we used graylevel co-occurrence texture features as described in [12, 13] as well as two pixel based features.

Graylevel co-occurrence texture features assume that the texture information in an image is contained in the overall spatial relationships among the pixels in the image.[12] This is done by first determining the Graylevel Co-occurrence Matrix (GLCM). This is an estimate of the second order probability density function of the pixels in the image. The features are then statistics obtained from the GLCM.

We assume that the texture in a document is predominantly in the process direction due to the banding signal. Figure 2 shows an idealized character, $Img(i, j)$, from which features are extracted. The region of interest (ROI) is the set of all pixels within the printed area of the character. The determination of this region involves morphological filtering and is discussed in [8].

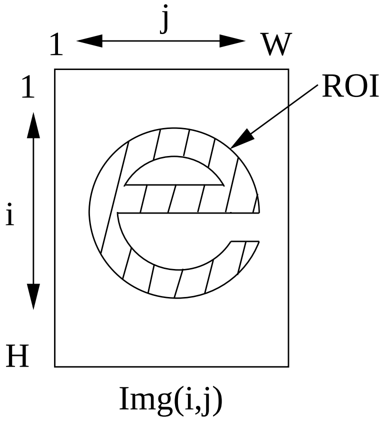

Figure 2: *Idealized character*

We define the number of pixels in the ROI to be

$$R = \sum_{(i,j)\epsilon ROI} 1. \qquad (1)$$

We then estimate the Gray-Level Co-occurrence Matrix (GLCM). This matrix, defined in Equation 2, has entries $glcm(n,m)$ which are equal to the number of occurrences of pixels with graylevels n and m respectively with a separation of (dr,dc) pixels (see Figure 3). The number of pixels over which this estimate is obtained is given by Equation 3. If the GLCM is normalized with respect to R_{glcm}, its entries then represent the probability of occurrence of pixel pairs with graylevels n and m with separation (dr,dc). We will choose $dc = 0$ and vary dr between 1 and 10 since we believe the textural variation is only in the process direction.

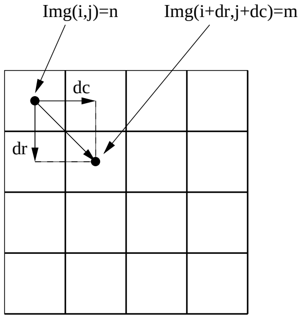

Figure 3: *Generation of* $glcm(n,m)$

$$glcm(n,m) = \sum_{(i,j),(i+dr,j+dc)\epsilon ROI} 1_{\{Img(i,j)=n, Img(i+dr,j+dc)=m\}} \qquad (2)$$

$$R_{glcm} = \sum_{(i,j),(i+dr,j+dc)\epsilon ROI} 1 \qquad (3)$$

$$p_{glcm}(n,m) = \frac{1}{R_{glcm}} glcm(n,m) \qquad (4)$$

Twenty features are obtained from the GLCM. The first four are the marginal means and variances defined by Equations 7-10 which are estimated from the marginal probability densities defined by Equations 5 and 6.

$$p_r(n) = \sum_{m=0}^{255} p_{glcm}(n,m) \qquad (5)$$

$$p_c(n) = \sum_{m=0}^{255} p_{glcm}(n,m) \qquad (6)$$

$$\mu_r = \sum_{n=0}^{255} p_r(n) \qquad (7)$$

$$\mu_c = \sum_{m=0}^{255} p_c(m) \qquad (8)$$

$$\sigma_r^2 = \sum_{n=0}^{255} n^2 p_r(n) - \mu_r^2 \qquad (9)$$

$$\sigma_c^2 = \sum_{m=0}^{255} m^2 p_c(m) - \mu_c^2 \qquad (10)$$

The next seven features are the energy of the normalized GLCM, three entropy measurements, the maximum entry in the GLCM, and two correlation metrics. These are defined by Equations 11- 17.

$$Energy = \sum_{n=0}^{255}\sum_{m=0}^{255} p_{glcm}^2(n,m) \qquad (11)$$

$$hxy1 = -\sum_{n=0}^{255}\sum_{m=0}^{255} p_{glcm}(n,m)\log_2(p_r(n)p_c(m)) \qquad (12)$$

$$hxy2 = -\sum_{n=0}^{255}\sum_{m=0}^{255} p_r(n)p_c(m)\log_2(p_r(n)p_c(m)) \qquad (13)$$

$$h_{glcm} = -\sum_{n=0}^{255}\sum_{m=0}^{255} p_{glcm}(n,m)\log_2 p_{glcm}(n,m) \qquad (14)$$

$$MaxProb = \max_{n,m}\{p_{glcm}(n,m)\} \qquad (15)$$

$$\rho_{nm} = \sum_{n=0}^{255}\sum_{m=0}^{255} \frac{(n-\mu_r)(m-\mu c)p_{glcm}(n,m)}{\sigma_r\sigma_c} \qquad (16)$$

$$diagcorr = \sum_{n=0}^{255}\sum_{m=0}^{255} |n-m|(n+m-\mu_r-\mu_c)p_{glcm}(n,m) \tag{17}$$

Four features, Equations 19- 22, are obtained from the difference histogram defined by Equation 18. They are the energy, entropy, inertia, and local homogeneity of $D(k)$ respectively.

$$D(k) = \sum_{\substack{0\leq n\leq 255\\ 0\leq m\leq 255\\ |n-m|=k}} p_{glcm}(n,m) \tag{18}$$

$$Denergy = \sum_{k=0}^{255} D(k) \tag{19}$$

$$h_D = -\sum_{k=0}^{255} D(k)\log_2 D(k) \tag{20}$$

$$I_D = \sum_{k=0}^{255} k^2 D(k) \tag{21}$$

$$h_D = \sum_{k=0}^{255} \frac{D(k)}{1+k^2} \tag{22}$$

The last five features, Equations 25- 29, are obtained from the sum histogram defined by Equation 23. They are the energy, entropy, variance, cluster shade, and cluster prominence of $S(k)$ respectively.

$$S(k) = \sum_{\substack{0\leq n\leq 255\\ 0\leq m\leq 255\\ n+m=k}} p_{glcm}(n,m) \tag{23}$$

$$\mu_S = \sum_{k=0}^{510} kS(k) \tag{24}$$

$$Senergy = \sum_{k=0}^{510} S(k) \tag{25}$$

$$h_S = -\sum_{k=0}^{510} S(k)\log_2 S(k) \tag{26}$$

$$\sigma_S^2 = \sum_{k=0}^{510}(k-\mu_S)^2 S(k) \tag{27}$$

$$A_D = \sum_{k=0}^{510} \frac{(k-\mu_r-\mu_c)^3 S(k)}{(\sigma_r^2-\sigma_c^2+2r\sigma_r\sigma_c)^{\frac{3}{2}}} \tag{28}$$

$$B_D = \sum_{k=0}^{510} \frac{k-\mu_r-\mu_c)^4 S(k)}{(\sigma_r^2-\sigma_c^2+2r\sigma_r\sigma_c)^2} \tag{29}$$

In addition to the 20 graylevel features above, two simple features are also included and defined in Equations 31 and 33. These are the variance and entropy of the pixel values in the ROI.

$$\mu_{Img} = \frac{1}{R}\sum_{(i,j)\epsilon ROI} Img(i,j) \tag{30}$$

$$\sigma_{Img}^2 = \frac{1}{R}\sum_{(i,j)\epsilon ROI} (Img(i,j)-\mu_{Img})^2 \tag{31}$$

$$p_{Img}(\alpha) = \frac{1}{R}\sum_{(i,j)\epsilon ROI} 1_{\{Img(i,j)=\alpha\}} \tag{32}$$

$$h_{Img} = -\sum_{\alpha=0}^{255} p_{Img}(\alpha)\log_2 p_{Img}(\alpha) \tag{33}$$

Results

Using our FMTG, we estimated that in a page of English text printed at 12 point font there are on average 458 “e”s per page. For a 10 point font there are on average 630 “e”s per page. To test our classification technique we use a test document containing 300 “e”s. This document is printed on each of our test printers and scanned at 2400dpi. Each scanned document is then classified using values of dr between 1 and 10, providing 10 classification results for each document.

Using all 22 features described above, we find that the dr which provides the best classification results is $dr = 5$. The classification matrix for this choice of dr is shown in Figure 6. Each entry of the matrix is the number of “e”s out of the 300 in the test document which were classified as the printer listed at the heading of its column. For example, examining the first row we see that 197 “e”s, printed from the Brother HL-1440 were classified correctly. The second highest number of “e”s were classified as being printed by the Minolta 1250W. A majority vote indicates that this document was most likely printed by the HL-1440, which is correct.

Again let $c(\phi)$ be equal to the number of “e”s from any one unknown document classified as being printed by printer ϕ. Furthermore let $c(\phi_1)$ be the greatest among all classes, and $c(\phi_2)$ be the second greatest. Then we say that the final classification resulting from the majority vote among the $c(\phi)$ has a higher confidence if the ratio between $c(\phi_1)$ and $c(\phi_2)$ is larger. In this case, all the printers are classified correctly and with a relatively high confidence with the exception of the Okidata 14e, which is classified as being a Minolta 1250W.

The classification is repeated using 4 manually chosen features. These features are σ_{Img}^2, h_{Img}, μ_r, and $Energy$.

Scatter plots of these features which show good separation between all ten printes are shown in Figures 4 and 5. The HP LaserJet 4050 is a higher quality printer compared to the others in our test set, and the graylevel variance of its output is very low, moving all of its data points to the far left of Figure 4 off the graph.

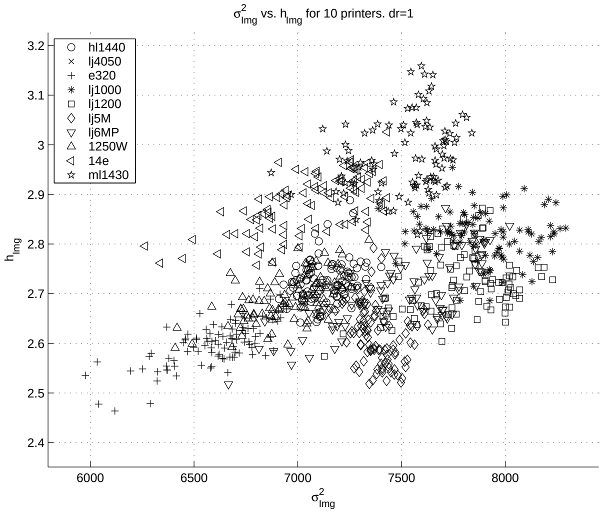

Figure 4: σ^2_{Img} *vs* h_{Img}

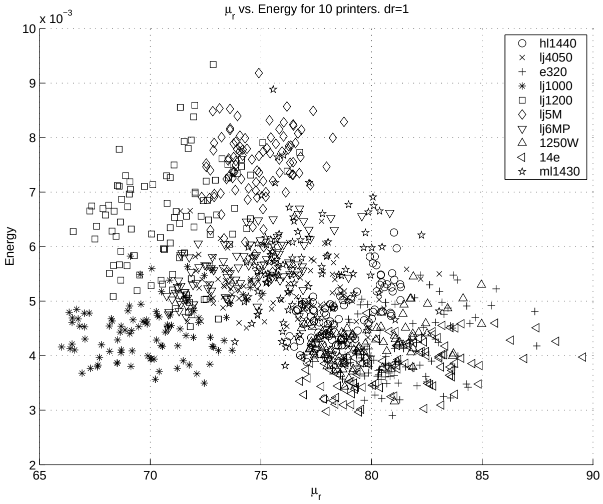

Figure 5: μ_r *vs* $Energy$

Using these four features, we find that $dr = 9$ provides the best classification results which are shown in Figure 7. All the printers are classified correctly with the exception of the HP LaserJet 1200 which is classified as being an HP LaserJet 1000. This is reasonable since these two printers we believe use the same print engine.

We can not say that the results obtained using 4 features are the same or better than when using all 22 features. The reason is that the confidence in the classification results using 4 features is in general lower than when using all the features. For example, the ratio between $c(\phi_1)$ and $c(\phi_2)$ for the Minolta 1250W when using 22 features is 3.69, while in the 4 feature case it is only 1.06. Similarly the HP LaserJet 6MP, Minolta 1250W, and Okidata 14e all have lower confidence in their classification results despite the fact that they are classified correctly when using only 4 features.

Conclusion

Printer identification using graylevel co-occurence texture features shows promising results. Although in Figure 6 we demonstrated good results, we would like to go further and reduce the number of features needed. Clearly this is possible given the results in Figure 7 using only 4 of the 22 original features. Our preliminary results indicate that some of the features may be redundant. Further work needs to be done in constructing an optimal feature set.

It is important to note that the technique presented requires that prior information about the printers in question be known. If the unknown document was printed by a printer which is not included in the classifier training data set, then it will be mistakenly classified as one of the known printers.

It will also be important to extend this technique to work with multiple font sizes, font types, and also different characters. Using our current system to classify a document using all the characters contained in it, not just “e”s, would require a separate 5NN classifier block for each character. This increases the complexity of the classifier. Ideally we would like a subset of features, or techniques for estimating the GLCM which are character independent. The same holds true for different font sizes and font types.

References

1. Mauro Barni, Christine I. Podilchuk, Franco Bartolini and Edward J. Delp, Watermark embedding: hiding a signal within a cover image, IEEE Communications Magazine, 39, 102 (2001).
2. R. W. Wolfgang, C. I. Podilchuk and E. J. Delp, Perceptual watermarks for digital images and video, in Proceedings of the IEEE, pp. 1108–1126 (1999).
3. C. I. Podilchuk and E. J. Delp, Digital watermarking: Algorithms and applications, IEEE Signal Processing Magazine, 18, 33 (2001).
4. Edward J. Delp, Is your document safe: An overview of document and print security, in Proceedings of the IS&T’s NIP18: International Conference on Digital Printing Technologies (2002).
5. Rudolf L. van Renesse, Paper based document security - a review, IEE European Conference on Security and Detection, 51, 75 (1997).
6. Rudolf L. Renesse, Optical Document Security, Artech House, Boston, MA (1998).

In\Out	hl1440	lj4050	e320	lj1000	lj1200	lj5M	lj6MP	1250W	14e	ml1430	Majority Vote
hl1440	**197**	0	1	1	0	11	6	***57***	21	6	**hl1440**
lj4050	0	**300**	0	0	0	0	0	0	0	0	**lj4050**
e320	0	0	**248**	0	2	0	0	***36***	13	1	**e320**
lj1000	4	0	0	**152**	***66***	5	11	7	4	51	**lj1000**
lj1200	3	0	0	***99***	**130**	14	11	13	1	29	**lj1200**
lj5M	***60***	0	1	1	7	**165**	29	30	5	2	**lj5M**
lj6MP	***30***	0	14	11	6	28	**153**	29	9	20	**lj6MP**
1250W	33	0	***49***	2	1	7	4	**181**	20	3	**1250W**
14e	***74***	0	25	1	2	2	3	**128**	62	3	**1250W**
ml1430	10	0	9	***61***	15	21	30	13	17	**124**	**ml1430**

123 = Correctly Classified

123 = Incorrectly Classified

123 = 2nd highest classification

Figure 6: *Classification Results using 22 features and* $dr = 5$

In\Out	hl1440	lj4050	e320	lj1000	lj1200	lj5M	lj6MP	1250W	14e	ml1430	Majority Vote
hl1440	**141**	0	0	3	2	24	12	***69***	41	8	**hl1440**
lj4050	0	**300**	0	0	0	0	0	0	0	0	**lj4050**
e320	0	0	**283**	0	0	1	0	***12***	4	0	**e320**
lj1000	7	0	0	**152**	***80***	23	27	8	0	3	**lj1000**
lj1200	13	0	1	**140**	***93***	26	20	4	0	3	**lj1000**
lj5M	***53***	0	1	6	7	**187**	21	25	0	0	**lj5M**
lj6MP	33	0	25	***51***	44	39	**66**	16	0	26	**lj6MP**
1250W	40	0	***103***	0	0	31	12	**109**	5	0	**1250W**
14e	***98***	0	33	1	0	1	1	36	**115**	15	**14e**
ml1430	40	0	1	15	15	1	40	11	***46***	**131**	**ml1430**

123 = Correctly Classified

123 = Incorrectly Classified

123 = 2nd highest classification

Figure 7: *Classification Results using 4 features and* $dr = 9$

7. Pei-Ju Chiang, Gazi N. Ali, Aravind K. Mikkilineni, George T.-C. Chiu, Jan P. Allebach and Edward J. Delp, Extrinsic signatures embedding using exposure modulation for information hiding and secure printing in electrophotographic devices, in Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies (2004).
8. Aravind K. Mikkilineni, Gazi N. Ali, Pei-Ju Chiang, George T.-C. Chiu, Jan P. Allebach and Edward J. Delp, Signature-embedding in printed documents for security and forensic applications, in Proceedings of the SPIE/IS&T Conference on Security, Steganography and Watermarking of Multimedia Contents, pp. 455–466 (2004).
9. Gazi N. Ali, Pei-Ju Chiang, Aravind K. Mikkilineni, Jan P. Allebach, George T. Chiu and Edward J. Delp, Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices, in Proceedings of the IS&T's NIP19: International Conference on Digital Printing Technologies, pp. 511–515 (2003).
10. Gazi N. Ali, Pei-Ju Chiang, Aravind K. Mikkilineni, George T.-C. Chiu, Edward J. Delp and Jan P. Allebach, Application of principal components analysis and gaussian mixture models to printer identification, in Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies (2004).
11. Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, San Diego, CA (1990).
12. Robert M. Haralick, K. Shanmugam and Its'Hak Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics, SMC-3, 610 (1973).
13. Richard W. Conners, Mohan M. Trivedi and Charles A. Harlow, Segmentation of a high-resolution urban scene using texture operators, Computer Vision, Graphics, and Image Processing, 25, 273 (1984).