Identification of Inkjet Printers for Forensic Application

Osman Arslan, Roy Kumontoy, Aravind K. Mikkilineni, Jan P. Allebach, and Edward J. Delp, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
Pei-Ju Chiang and George T. Chiu, School of Mechanical Engineering, Purdue University, West Lafayette, Indiana, USA

Abstract

In forensic applications, identification of the source of a printed document can be very critical. However, this task is very challenging, because there are a variety of factors, such as the media type, the age of the printer, and the amount of colorant left in the cartridge that can affect the characteristics of a printed document. For inkjet printers, the availability of different print modes, and the use of variety of ink types adds even more complexity to this process. In this paper, we investigate different texture features of the characters printed by inkjet printers for classification based on text-only documents. We check the in model stability of these features by using a cost function. Finally, we perform stepwise discriminant analysis to reduce the feature set.[8]

Introduction

Identification of the source of a printed document can provide valuable information in forensic applications. For this purpose, we need to identify some features of the printer by investigating the contents of a printed document. These contents may include images, charts, or text. In this project, we are interested in text-only documents. There are two basic type of printers in the market that are widely used by an average customer: electrophotographic (EP) and inkjet printers. Wolin *et al* gave an overview of the characteristics of different printing techniques and suggested different ways to identify these printers in his paper.[1] Oliver *et al* used variety of image features and print quality metrics to discriminate one inkjet printer model from another.[2] In previous work with laser EP printers, Ali *et al* used the banding artifact as an intrinsic signature for printer identification.[3,4] They applied principal component analysis and Gaussian mixture models for classification of the printers. Mikkilineni *et al* used texture features of the printed characters to identify the model of the EP printer that is used to print the document in question.[5,6] More specifically, they focused on the texture features introduced by Haralick *et al*.[7] In this paper, we investigate similar texture features to be able to identify different models of inkjet printers. We search appropriate features for printer classification by changing the parameters of the texture features. We first choose the texture features that are stable within the same printer model. Finally, among these features, we select the most significant ones for printer classification by performing stepwise discriminant analysis.

Harlick's Texture Features

The features that are used to identify different printer models have to be robust to certain variations in the inkjet printers. For example, if we use the average gray level of a character as a feature, this feature may heavily depend on the amount of colorant left in the cartridge. Also the features should not depend directly on the size or type of font of the character. For example, if we use the length or width of a character as our feature, this will directly depend on the type and size of the font used in that document. To be able use this feature, the font type and size of the character has to be detected first, which will make the process too complicated. For these reasons, we decided to use Haralick's texture features as our feature set.[7]

Haralick's features are based on gray-tone spatial dependencies of the image. The gray-level co-occurrence matrices (GLCM) are used to calculate these features. This matrix is an estimate of the second order probability density function of the pixels in the character image. Suppose that the image has N_x pixels in the horizontal direction and N_y pixels in the vertical direction, and there are N_g gray levels for each pixel. We can define $L_x = 1, 2, \ldots, N_x$ as the horizontal spatial domain and $L_y = 1, 2, \ldots, N_y$ as the vertical spatial domain. Then $L_x \times L_y$ will be the set of pixels ordered by their row-column designations. The texture-context information in the character image is contained in the overall spatial relationship which the gray tones in the image have with one another.

An entry of a GLCM $P(i, j, d, \theta)$ represents the frequency of two neighboring pixels in an image I separated by a distance d at an angle of θ with gray-level values i and j. For the angle θ, four angles of 0°, 45°, 90°, and 135° are considered. Formally, we can define the frequencies as follows:

$$
\begin{aligned}
&P(i, j, d, 0°) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times \\
&(L_y \times L_x), |k - m = 0, |l - n| = d, \\
&I(k, l) = i, I(m, n) = j\}, \\
&P(i, j, d, 45°) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times \\
&(L_y \times L_x), | (k - m = d, l - n = -d), \\
&or\ (k - m = -d, l - n = d), \\
&I(k, l) = i, I(m, n) = j\}, \\
&P(i, j, d, 90°) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times \\
&(L_y \times L_x), ||k - m| = d, l - n = 0, \\
&I(k, l) = i, I(m, n) = j\}, \\
&P(i, j, d, 135°) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times \\
&(L_y \times Lx), | (k - m = d, l - n = d), \\
&or(k - m = -d, l - n = -d), \\
&I(k, l) = i, I(m, n) = j\}. \qquad (1)
\end{aligned}
$$

where # denotes the number of elements in the set. Note that these matrices are symmetric, $P(i, j, d, \theta) = P(j, i, d, \theta)$. Given a GLCM, 14 different textural features can be calculated. The following equations define these features.

Notation

$p(i,j)$	(i,j)th entry in a normalized gray-tone spatial-dependence matrix, $p(i,j) = P(i,j)/R$, where R is the number of pixel pairs.
$p_x(i)$	ith entry in the marginal probability matrix obtained by summing the rows of $p(i,j)$, $p_x(i) = \sum_{j=1}^{N_g} P(i,j)$
N_g	Number of distinct gray levels in the image
$\sum_i$ and $\sum_j$	$\sum_{i=1}^{N_g}$ and $\sum_{j=1}^{N_g}$, respectively

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j).$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{\substack{j=1 \\ i+j=k}}^{N_g} p(i,j),\ k = 2,3,...2N_g.$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{\substack{j=1 \\ |i-j|=k}}^{N_g} p(i,j),\ k = 0,1,...N_g - 1.$$

Textural Features

1. Angular Second Moment:

$$f_1 = \sum_i \sum_j \{p(i,j)\}^2.$$

2. Contrast:

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^{N_g}\sum_{j=1}^{N_g} p(i,j) \right\}.$$

3. Correlation:

$$f_3 = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$$

where μ_x, μ_y, σ_x, and σ_y are the means and the standard deviations of p_x and p_y.

4. Sum of Squares: Variance

$$f_4 = \sum_i \sum_j (i-\mu)^2 p(i,j).$$

5. Inverse Difference Moment:

$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j).$$

6. Sum Average:

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i).$$

7. Sum Variance:

$$f_7 = \sum_{i=2}^{2N_g} (i-f_8)^2 p_{x+y}(i).$$

8. Sum Entropy:

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i)\log\{p_{x+y}(i)\}.$$

9. Entropy:

$$f_9 = -\sum_i \sum_j p(i,j)\log(p(i,j)).$$

10. Difference Variance:

f_{10} = variance of p_{x-y}.

11. Difference Entropy:

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i)\log\{p_{x-y}(i)\}.$$

12. First Information Measure of Correlation:

$$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}}.$$

Table 1: Printers installed in the printer bank.

Maker	Printer Model	Print Mode
HP	3420	Normal
HP	3650	Normal
HP	psc1315	Normal
Lexmark	Z25	Better
Lexmark	Z2250	Normal
Canon	S330	Standard

13. Second Information Measure of Correlation:

$f_{13} = (1 - \exp[-2.0\,(HXY2 - HXY)])^{1/2}$

where HX and HY are entropies of p_x and p_y, and

$$HXY = -\sum_i \sum_j p(i,j)\log(p(i,j)),$$

$$HXY1 = -\sum_i \sum_j p(i,j)\log(p_x(i)p_y(j)),$$

$$HXY = -\sum_i \sum_j p_x(i)p_y(j)\log(p_x(i)p_y(j)),$$

14. Maximal Correlation Coefficient:

f_{14} = (Second largest eigenvalue of Q)$^{1/2}$ where

$$Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}.$$

Some of the measures that are defined above have intuitive meanings, but for some of them, it is difficult to identify which special textural characteristic they represent. The mean and the range of these measures can be used as features. An average of the four different directions θ of these measures can be used as textural features. Another option is to use just one direction, depending on the application.

Methods

We investigated six different inkjet printer models in our printer bank. These printers are shown in Table 1. There are two identical units for each printer model. Inkjet printers have different print modes depending on their printing speed. These modes affect the characteristics of the printed document significantly. So we used the default modes for each printer we investigated. The only exception was Lexmark Z25. As the print quality was not acceptable for the default mode, we selected the next better mode for this printer.

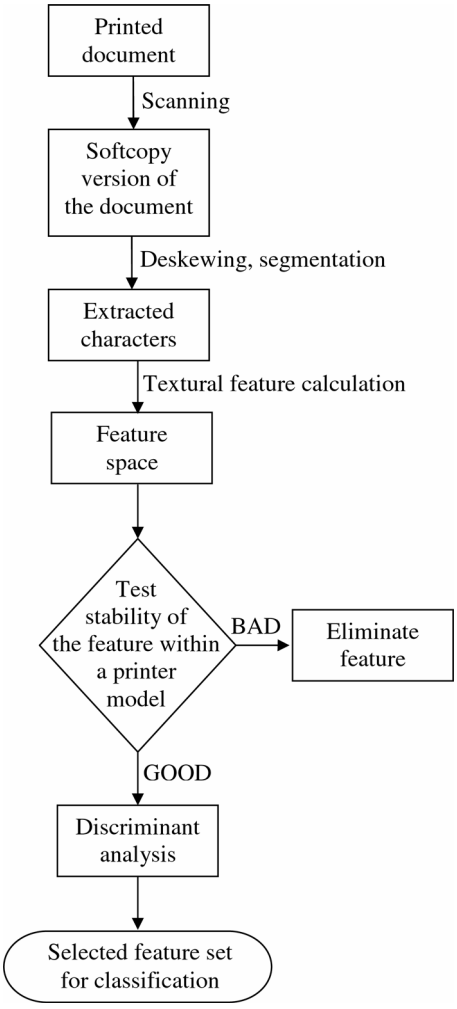

Figure 1. Steps followed in textural feature selection for printer classification.

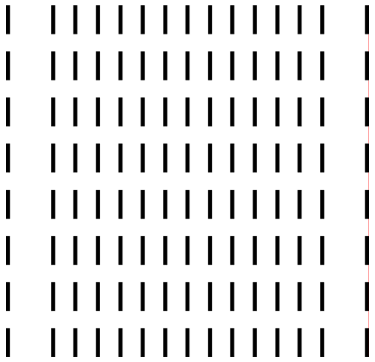

Figure 2. Character set printed on the test page.

Figure 1 shows the steps followed in textural feature selection for printer classification. The character set printed on the test page we used for printer identification is shown in Fig. 2. Because we observed instability in the characters on the border areas, the three left-most and right-most columns as well as the upper-most and lower-most rows of the character set were not included in the data set. The characters consist of 12 point Arial capital 'I' letters. We extracted 42 characters for each specific printer model.

The document was scanned at 2400 dpi with 256 gray levels. The skew of the scanned pages were checked and if the skew was not acceptable, the scanning process was repeated. Next, the characters were extracted for each printer. The average size of each extracted character is about 268 × 35 pixels. We calculated the textural features of each character by using Haralick et al's methods.[7] For the angle θ, we used values of 0° and 90°. For the distance d parameter of the GLCM, we used values of (1,2,4,...,128) for angle 90°, and we used values of (1,2,4,8,16) for angle 0° because of the size limitations of the images. After the features were calculated, we first checked the stability of each feature within a printer model. This is achieved by comparing the features of the two identical units of each printer model. Assume that we would like to check the in-class stability of feature f for printer model PrA where we have two identical units PrA_1 and PrA_2. Then our cost function Ψ for in-class stability will be:

$$\Psi = \frac{\left|\mu_{f_{PrA_1}} - \mu_{f_{PrA_2}}\right|}{\sqrt{\sigma^2_{f_{PrA_1}} + \sigma^2_{f_{PrA_2}}}} \frac{\max(\sigma_{f_{PrA_1}}, \sigma_{f_{PrA_2}})}{\min(\sigma_{f_{PrA_1}}, \sigma_{f_{PrA_2}})}$$

This cost function compares the variances and the normalized separation of the means of the features. We eliminated features that have a cost value of greater than 2.0. We also observed correlation between the textural features, so we selected a subset of them by looking at their stability among same models. This subset consisted of features f_2, f_3, f_9, f_{10}, and f_{14}.

After we refined our feature set by checking in-model stability, we applied stepwise discriminant analysis (SDA) to the refined feature set.[8] The objective of the SDA is to find the best features within a feature set that separate different classes from one another while at the same time keeping each cluster as tightly packed as possible.

Specifically, SDA starts with the full set of all features, calculates the significance level of each feature and removes the feature with the lowest significance. This process continues until all the features are above a predefined significance.

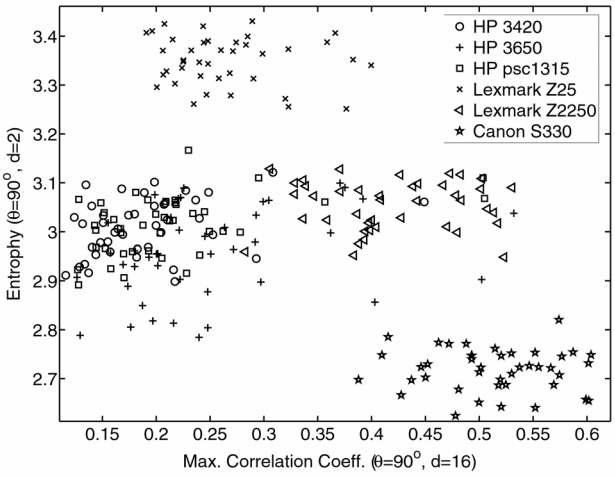

Figure 3. Scatter plot of the data for the six inkjet printers with respect to the entropy and maximal correlation coefficient features.

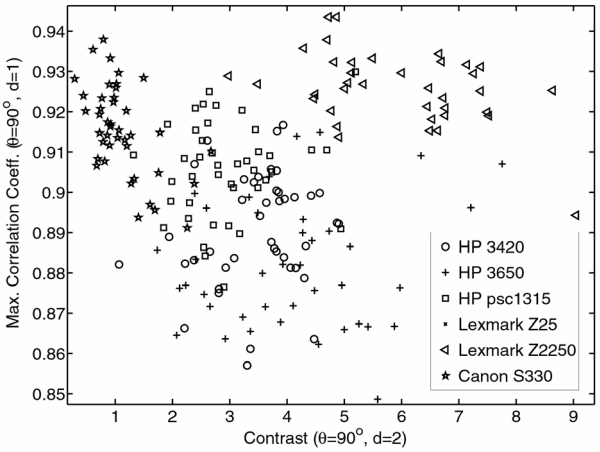

Figure 4. Scatter plot of the data for the six inkjet printers with respect to the contrast and maximal correlation coefficient features.

Preliminary Experimental Results

We used the methodology described in the previous section to select four features from the textural feature set. The most significant features obtained from the discriminant analysis for classification of the printers are entropy ($\theta = 90°$, $d = 2$), contrast ($\theta = 90°$, $d = 2$), and maximal correlation coefficient ($\theta = 90°$, $d = 1$, and $\theta = 90°$, $d = 16$). In Fig. 3, we show the scatter plot of the data points plotted with respect to two of the features. In this case, there are four clusters. The three HP printers fall onto each other, this may be because a similar printer technology was used to manufacture these printers. But all of the other printers were clustered nicely. We show the scatter plot of the data for the remaining two features in Fig. 4. This is the case where the three HP printers are separated the most. Even in this case, there is a considerable overlap between printer model HP3420 and the other two HP printers. This supports the idea that these printers have similar technologies.

Conclusion

We investigated the applicability of textural features for classification of inkjet printers. Specifically, we focused on textural features introduced by Haralick *et al.*[7] We produced a feature pool by varying different parameters of these features and selected most significant features that perform best classification. This is done by first eliminating the unstable features within a printer model, and then applying stepwise discriminant analysis. We found four features which are the most significant for classification of the inkjet printers in our printer bank.

References

§ This research is supported by a grant from National Science Foundation, under award number 0219893.

1. D. Wolin, Document Verification and Traceability through Image Quality Analysis, IS&T's NIP18: International Conference on Digital Printing Technologies, pg. 214. (2002).
2. John F. Oliver and Joyce X. Chen, Use of Signature Analysis to Discriminate Digital Printing Technologies, IS&T's NIP18: International Conference on Digital Printing Technologies, pg. 218. (2002).
3. G. N. Ali, P. C. Chiang, A. K. Mikkilineni, J. P. Allebach, G. T. Chiu, and E. J. Delp, Intrinsic and Extrinsic Signatures for Information Hiding and Secure Printing with Electrophotographic Devices, IS&T's NIP19: International Conference on Digital Printing Technologies, pg. 511. (2003).
4. G. N. Ali, P. C. Chiang, A. K. Mikkilineni, G. T. Chiu, E. J. Delp, and J. P. Allebach, Application of Principal Components Analysis and Gaussian Mixture Models to Printer Identification, IS&T's NIP20: International Conference on Digital Printing Technologies, pg. 301. (2004).
5. A. K. Mikkilineni, G. N. Ali, P. C. Chiang, G. T. Chiu, J. P. Allebach, and E. J. Delp, Signiture-embedding in Printed Documents for Security and Forensic Applications, SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VI, pg. 455. (2004).
6. A. K. Mikkilineni, P. C. Chiang, G. N. Ali, G. T. Chiu, J. P. Allebach, and E. J. Delp, Printer Identification based on Textural Features, IS&T's NIP20: International Conference on Digital Printing Technologies, pg. 306. (2004).
7. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification", IEEE transactions on Systems, Man, and Cybernetics, SMC-3, 610 (1973).
8. K. Huang, M. Velliste, and R. F. Murphy, Feature Reduction for Improved Recognition of Subcellular Location Patterns in Fluorescence Microscope Images, Proceedings of the SPIE, pg. 48. (2003).

Author Biography

Osman Arslan received his B.S. degree in Electrical Engineering from Bilkent University, Ankara, Turkey, in 1997. He worked in the Turkish Scientific Research Center for one year as a research engineer in 1998. He earned his M.S. degree from Purdue University, West Lafayette, IN, in 1999. He is currently pursuing his Ph.D. degree in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. He has been working in the Electronic Imaging Systems Laboratory at Purdue University since 2000. His research interests include electronic imaging systems, image quality, human perception, and secure printing.