Document Identification To Discourage Illicit Copying

Steven. H. Low Aleta M. Lapone Nicholas F. Maxemchuk
AT&T Bell Laboratories, Murray Hill NJ 07974
{slow, amg, nfm@research.att.com}

Abstract

An important application of future communications networks will be electronic publishing and digital library, provided copyright can be protected. A way to discourage illicit copying and distribution of documents is to uniquely mark each document copy by shifting certain lines or words slightly so that the original registered recipient can be identified from an illicit copy by detecting its mark. In this paper we present two techniques for reliable document identification, the centroid and the correlation detection. By analyzing the noise characteristics, we obtain the maximum likelihood detectors for both methods and their probabilities of error. We have applied these results to implement a marking and identification strategy proposed earlier, which marks a line both vertically by line shift and horizontally by word shift to make the marking robust against distortions in either direction. Preliminary experimental results are presented.

1 Introduction

An important application of future communications networks will be electronic publishing and digital library provided copyright can be protected. Currently, there is little means to inhibit people from copying and distributing copyrighted or sensitive documents. We have implemented a document marking and identification system that discourages illicit copying and distribution [1]. The system automatically puts a unique and indiscernible mark on each document copy and registers its recipient. If an illicit copy is recovered the system detects the mark from the copy, identifying the original recipient. It can be used to protect copyright in electronic publishing even when the electronic document is printed, photocopied or faxed. In this paper we describe the identification subsystem and presents experimental results.

To mark a page certain lines are shifted slightly up or down from their normal positions or certain words are shifted slightly left or right. The shifting pattern is different on different copies. To detect the mark the horizontal profiles of lines and vertical profiles of words are compiled from a digitized image of the page. We have experimented extensively with two detection methods. The first method measures spacing between profile centroids. The second method treats a profile as a discrete time signal and chooses the direction of shift that is most likely to account for the observed corrupted signal.

In §2 we define formally a profile and propose a simple noise model to model how a horizontal or vertical profile is corrupted by printing, photocopying, scanning and other processing. Our main result is presented in §3, which shows that the noise that corrupts the centroid of a profile is approximately zero-mean Gaussian whose variance is computable from the original uncorrupted profile.

Based on this approximation, we derive in §4 the maximum likelihood detectors for both methods and their probabilities of detection error. A maximum likelihood detector minimizes the average probability of error when all marking patterns are equally likely a priori.

Finally, we describe in §5 the implementation of a marking and identification strategy proposed earlier in [2]. The strategy takes advantage of the possibility that the vertical and horizontal profiles can be distorted to different degrees. A line is marked both vertically using line shifting and horizontally using word shifting. To detect the marking the probability of detection error on horizontal and vertical profiles are estimated after common distortions have been compensated for. Detection is then made in the less noisy direction. Our system uses the centroid detector for line shifts and the correlation detector for word shifts. A set of detection results on photocopies and fax-copies are presented.

Advances in computing, storage and communication technologies have made electronic publishing imminent. In [3] a cryptographic system for the secure distribution of electronic documents is described. In

0-7803-2509-5/95 US$4.00

[4] the approach to indiscernibly mark each document copy by varying the line or word spacing or by varying certain character features slightly is proposed. In [2] an experiment is reported which reveals that a document can be distorted much more severely in one direction than the other, and a marking and identification strategy that is robust against severe distortion in either direction is described. Along the same line but for different medium, schemes are described in [5, 6] to embed unique marking on images to deter illicit copying. In [7] a copyright management testbed is outlined for deposit and registration of electronic copyright materials and for on-line clearance of rights. In [8] several ways to assign unique identifiers to copies of digital data are studied that are secure against collusion among recipients to detect and remove the marking.

Throughout the paper $h(y)$ denotes an original unmarked and uncorrupted profile and $g(y)$ denotes its corrupted copy, marked or unmarked.

2 Model

2.1 Profiles and marking

Upon digitization the image of a page is represented by a function

$$f(x,y) \;=\; 0 \text{ or } 1, \quad x = 0,1,\ldots,W, \; y = 0,1,\ldots,L$$

where W and L, whose values depend on the scanning resolution, are the width and length of the page, respectively. The image of a text line is simply the function restricted to the region of the text line:

$$f(x,y) \;=\; 0 \text{ or } 1, \quad x = 0,1,\ldots,W, \; y = t,t+1,\ldots,b$$

where t and b are the top and bottom 'boundaries' of the text line, respectively. For instance, we may take t or b to be the mid-point of the interline spacing. The *horizontal profile* of the text line

$$h(y) \;=\; \sum_{x=0}^{W} f(x,y), \qquad y = t,t+1,\ldots,b$$

is the total number of 1's along the horizontal scan-lines y. The *vertical profile* of the text line

$$v(x) \;=\; \sum_{y=t}^{b} f(x,y), \qquad x = 0,1,\ldots,W$$

is the total number of 1's along the vertical scan-lines x. Figure 1 shows a typical horizontal profile of three text lines and a typical vertical profile of six words.

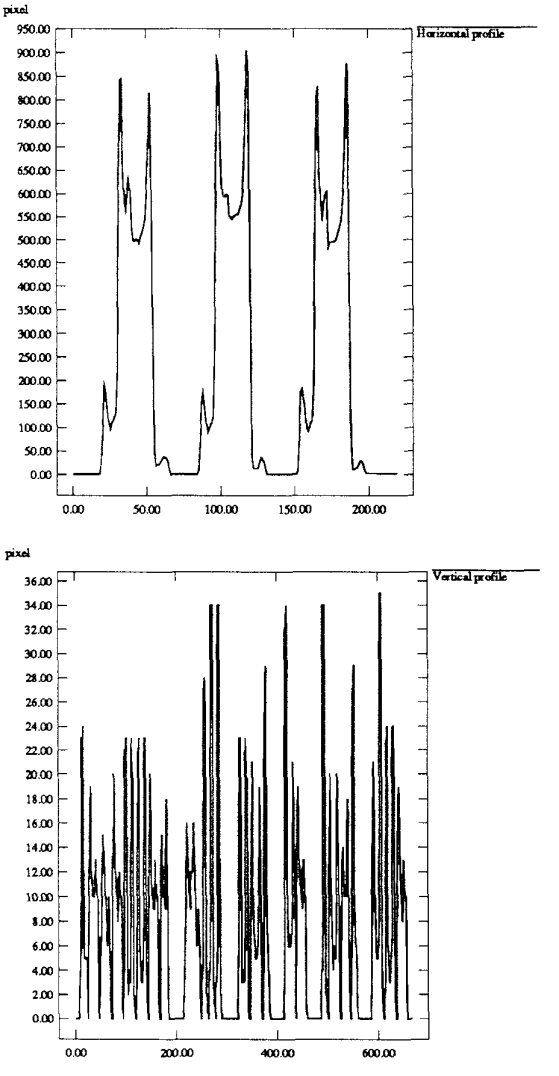

Figure 1: Horizontal and vertical profile (resolution = 300 dots-per-inch)

For simplicity we assume that $f(x, y)$, and hence the profiles $h(y)$ and $v(x)$, take continuous values.

A text line can be marked vertically by shifting it slightly up or down from its normal position to carry one bit of the identifier unique to the document. Its neighboring lines, called the control lines, are not marked. Alternatively a line can be marked horizontally by shifting certain words slightly left or right from their normal positions. The line is divided into some odd number of groups of words such that each group contains a sufficient number of characters. Each even group is then shifted while each odd group, called the control group, remains stationary.

Both line-shift and word-shift marking can be considered within the same model where we have a profile, denoted by $h(y)$, that covers three 'blocks'. For line shifting each block is the horizontal profile of a text line. For word shifting each block is the vertical profile of a group of words. The middle block is shifted slightly while the other two blocks, called the control blocks, are stationary.

2.2 Noise

When the marked original is printed, photocopied, and then scanned and processed, the text may be distorted by translation, scaling, and other random distortions. These are compensated for as in [2]. The remaining distortion is modeled by a white Gaussian noise. That is, we assume that a profile $h(y)$ on some interval $[b, e]$ is distorted only by additive noise $N(y)$ to become

$$g(y) = h(y) + N(y), \quad y = b, \ldots, e. \tag{1}$$

where $N(y)$ are i.i.d. Gaussian random variables with mean 0 and variance σ^2.

3 Centroid noise

Consider a block of profile $h(y)$, $y = b, \ldots, e$, corrupted by white Gaussian noise $N(y)$ as in (1). The *centroid* of the uncorrupted profile $h(y)$ is defined as

$$c = \frac{\sum_{y=b}^{e} y\, h(y)}{\sum_{y=b}^{e} h(y)}.$$

The noise $N(y)$ distorts the centroid of $g(y)$ randomly to $c + V$, i.e.,

$$V = \frac{\sum_{y=b}^{e} y\, g(y)}{\sum_{y=b}^{e} g(y)} - c.$$

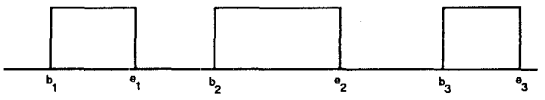

Figure 2: Profile h(y)

Our key result shows that V is approximately Gaussian (See [9]).

Theorem 1 *The centroid noise V is approximately a zero mean Gaussian random variable with variance*

$$(\sigma_1/\mu)^2 = \frac{\sigma^2 w}{H^2}\left(\delta^2 + \frac{1}{12}(w^2 - 1)\right) \tag{2}$$

where $H = \sum_b^e h(y)$ is the total weight of the profile $h(y)$, $w = e - b + 1$ is its width, and $\delta = c - \frac{e+b}{2}$ is the deviation of the uncorrupted centroid from the center.

Hence the noise variance depends on the original profile $h(y)$ only through three parameters, its weight H, width w, and deviation δ of centroid from the center.

4 Detection and performance

Suppose we are given a profile covering three blocks. As noted earlier each block is a horizontal profile in line-shift detection and is a vertical profile of a group of words in word-shift detection. Based on the noise model developed in the last section we present two detection methods. The first method measures the distance between the centroid of adjacent blocks and decides whether the middle block has been shifted left or right. The second method treats the profile as a discrete time signal and decides whether it originated from a signal corresponding to a left or right shifted middle block. For each method we present the maximum likelihood decision rule and its probability of error.

4.1 Centroid detection

This method works well only if each block of the given profile can be accurately delineated. Assume this has been done and we have a profile $h(y)$ and three intervals $[b_1, e_1]$, $[b_2, e_2]$, and $[b_3, e_3]$ that define the three blocks in $h(y)$, as shown in Figure 2. This is the original unmarked profile. The centroid of block i, $i = 1, 2, 3$, is

$$c_i = \frac{\sum_{b_i}^{e_i} y h(y)}{\sum_{b_i}^{e_i} h(y)}.$$

We also have the profile

$$g(y) = h(y) + N(y), \qquad y = b_1, b_1+1, \ldots, e_3$$

of the marked copy on the same interval $[b_1, e_3]$. It is corrupted by additive white zero-mean Gaussian noise $N(y)$ whose variance $\mathrm{Var}(N(y)) = \sigma^2$. The control blocks have centroids

$$U_1 = c_1 + V_1 \qquad \text{and} \qquad U_3 = c_3 + V_3.$$

The middle block has been shifted by a size $\epsilon > 0$ so that its centroid is

$$U_2 = c_2 + V_2 - \epsilon$$

if it is left shifted, and

$$U_2 = c_2 + V_2 + \epsilon$$

if it is right shifted. In view of the last section V_i, $i = 1, 2, 3$, are (approximately) independent zero-mean Gaussian with variance ν_i^2 given by

$$\nu_i^2 = \frac{\sigma^2 w_i}{H_i^2}\,(\delta_i^2 + (w_i^2 - 1)/12) \tag{3}$$

$$H_i = \sum_{b_i}^{e_i} h(y) \tag{4}$$

$$w_i = e_i - b_i + 1 \tag{5}$$

$$\delta_i = c_i - \frac{e_i + b_i}{2}. \tag{6}$$

To eliminate the effect of translation we base our detection on the distance $U_i - U_{i-1}$ between adjacent centroids instead of centroid U_2 of the middle block. It is convenient to use as decision variable the differences

$$\Gamma_l := (U_2 - U_1) - (c_2 - c_1)$$
$$\Gamma_r := (U_3 - U_2) - (c_3 - c_2)$$

of the corrupted centroid separations and the uncorrupted separations.

Proposition 1 *1. The maximum likelihood detector, when the observed value of (Γ_l, Γ_r) is (γ_l, γ_r), is*

decide left shift	*if* $\gamma_l/\nu_1^2 \le \gamma_r/\nu_3^2$
decide right shift	*otherwise*

where ν_1^2 and ν_3^2 are the centroid noise variances of the left and right control blocks, respectively.

2. The error probability of the maximum likelihood detector is

$$P_E = erf\Big(-\epsilon \sqrt{\frac{\nu_1^2 + \nu_3^2}{\nu_1^2\nu_3^2 + \nu_2^2(\nu_1^2 + \nu_3^2)}}\Big)$$

where $erf(x) := \int_{-\infty}^{x} e^{-y^2/2}\, dy$.

Note that the decision does not depend on the middle block, except through Γ_l and Γ_r. The test in the proposition does not require measurement of the profile noise variance σ^2 since it appears in both ν_1^2 and ν_3^2 (see (3)). Only the three parameters H_i, w_i, δ_i of each uncorrupted control block are required.

4.2 Correlation detection

In this subsection we present the correlation detector. We again have a profile $h(y)$ and three intervals $[b_1, e_1]$, $[b_2, e_2]$, and $[b_3, e_3]$ as shown in Figure 2. We assume that $h(y) = 0$ between these intervals. Let $h^l(y)$ be the resultant profile when the middle block is left shifted by $\epsilon > 0$:

$$h^l(y) = \begin{cases} h(y), & y < b_2 - \epsilon \text{ or } y > e_2 \\ h(y+s), & b_2 - \epsilon \le y \le e_2 - \epsilon \\ 0, & e_2 - \epsilon \le y < e_2 \end{cases} \tag{7}$$

and $h^r(y)$ be that when the middle block is right shifted:

$$h^r(y) = \begin{cases} h(y), & y < b_2 \text{ or } y > e_2 + \epsilon \\ 0, & b_2 \le y < b_2 + \epsilon \\ h(y-s), & b_2 + \epsilon \le y \le e_2 + \epsilon \end{cases} \tag{8}$$

The profile $g(y)$ compiled from the illicit copy and after distortion compensation is corrupted by additive white Gaussian noise such that

$$g(y) = h^l(y) + N(y), \qquad y = b_1, \ldots, e_3 \tag{9}$$

if the middle block is left shifted, and

$$g(y) = h^r(y) + N(y), \qquad y = b_1, \ldots, e_3 \tag{10}$$

if it is right shifted.

Proposition 2 *1. The maximum likelihood detector given the observed profile $g(y)$ is*

decide left shift	*if* $\sum_{b_1}^{e_3} g(y)\,(h^l(y) - h^r(y)) \ge 0$
decide right shift	*otherwise*

where h^l and h^r are computed from the original profile $h(y)$ according to (7) and (8), respectively.

2. The error probability of the maximum likelihood detector is

$$P_E = erf\Big(-\sqrt{\frac{\sum h^2(y) - \sum h^l(y)h^r(y)}{2\sigma^2}}\Big)$$

where $erf(x) := \int_{-\infty}^{x} e^{-y^2/2}\, dy$.

5 Experiments

In this section we describe an application of earlier results to build a document marking and identification prototype and present experimental results.

5.1 Prototype

An experiment in [2] reveals that depending on what type of process(es) a document goes through, the noise on the profiles after distortion compensation can be much more severe in one direction than the other. For example when a document is photocopied, depending on the copier and the copying option selected, the distortion can be more severe in the horizontal direction of the text. To take advantage of this possibility we propose in [2] a marking and identification strategy in which a line is marked both vertically using line shift and horizontally using word shift. To detect the marking the probability of detection error on horizontal and vertical profiles are estimated after common distortions have been compensated for. Detection is then made in the less noisy direction. This strategy has been implemented in a software prototype.

The marking subsystem takes as input the postscript file of a document and a list of its intended recipients. Each recipient is assigned a unique binary identifier. It marks a line vertically by shifting it slightly, e.g., 1/150 inch, up or down from its normal position. The line is also divided into some odd number of groups of words. Each even group is then shifted slightly, e.g., 1/150 inch, left or right while the odd groups remain stationary. Instead of using multiple groups to carry multiple bits per line, we use them to carry just one bit with redundancy to combat noise. The system automatically marks the document, stores the identifier with the corresponding recipient in a database, and prints a copy for each recipient.

When an illicit copy is discovered a marked page is scanned to produce a bitmap image, the image is processed to correct for excessive skewing (rotation of text) and to remove 'salt–and–pepper' noise [10, 11], and the horizontal and vertical profiles are then compiled. The system then estimates and compensates for translation and scaling of the profiles.

Using the maximum likelihood detectors in Propositions 1 and 2, the prototype detects line or word shift according as the horizontal or vertical profile is less noisy. We have found empirically that centroid detection works well for line shifts and correlation detection works well for word shifts.

5.2 Experimental results

We present 2 sets of experimental results. A two-page document, the first page being a title page, were marked by both line and word shifting. They were printed on a laser printer Hewlett–Packard LaserJet IIISi. In the first experiment, repeated copies were made on a Xerox 5052 plain paper copier to create successively more degraded copies. In the second experiment, the document was transmitted through a Xerox Telecopier 7033 fax machine and received at another machine of the same model. The copies were scanned using a Ricoh FS2 Apunix scanner to produce bitmap images. These images were processed to generate vertical and horizontal profiles. Marking was detected from these profiles.

Page 1 of the document contained 8 marked lines and page 2 contained 11, giving a total of 19 marked lines. Each of these lines was shifted vertically . It was also divided into groups of words and each even group was shifted horizontally.

For the first experiment, the size of both the line and word shift was 2 pixels, or 1/150 inch. Photocopies were made in both portrait and the orthogonal, or landscape, orientations. In landscape orientation, all 8 marked lines fitted into page 1 but only 10 fitted into page 2. Each landscape copy thus contained only 18 marked lines, except the 10th copy for which page 1 was skewed so much that profiles could not be properly generated. Hence it only had a total of 10 marked lines. Denote the laser printed copy as copy 0; the ith copy, $i = 1, \ldots, 10$, is obtained by photocopying copy $i - 1$. The detection results are in Table 1. With a shift of 2 pixels, line-shift marking was robust

	Portrait copies		Landscape copies	
Copy	Line shift	Word shift	Line shift	Word shift
0	0/19 error	0/19 error	0/18 error	0/18 error
1	0/19	0/19	0/18	0/18
2	0/19	0/19	0/18	0/18
3	0/19	6/19	0/18	0/18
4	0/19	5/19	0/18	0/18
5	0/19	4/19	0/18	0/18
6	0/19	3/19	0/18	0/18
7	0/19	9/19	0/18	0/18
8	0/19	11/19	0/18	0/18
9	0/19	10/19	0/18	0/18
10	0/19	6/19	0/10	0/10

Table 1: Detection on photocopies (shift = 1/150 inch)

enough to survive both types of copying. In landscape orientation word-shift detection was so accurate that not only were all marked lines detected correctly as shown in the table, in fact, all groups (43 of them per copy) were detected correctly for all copies except one group in the 10th copy. In the portrait orientation the word shift detection results were not monotonic because each experiment involved photocopying and scanning that introduced randomness. Table 1 is a sample value of this random process.

Table 2 shows the detection result on the fax copy. In fact all the 43 groups of word shift were detected

Copy	Line shift	Word shift
fax	0/19 error	0/19 error

Table 2: Detection on fax copy (shift = 1/150 inch)

correctly. In this experiment the paper document is converted to electronic medium and subjected to lossy compression before being detected.

References

[1] A. M. Lepone, S. H. Low, and N. F. Maxemchuk. Document marking and identification techniques, Part II: Implementation. Preprint, July 1995.

[2] S. H. Low, N. F. Maxemchuk, J. T. Brassil, and L. O'Gorman. Document marking and identification using both line and word shifting. *Proceedings of Infocom'95*, April 1995.

[3] A. K. Choudhury, N. F. Maxemchuk, S. Paul, and H. G. Schulzrinne. Copyright protection for electronic publishing over computer networks. Technical Memo BL011382-940428-75TM, AT&T Bell Laboratories, April 1994. Submitted for publication.

[4] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman. Electronic marking and identification techniques to discourage document copying. *Proceedings of Infocom'94*, pages 1278–1287, June 1994.

[5] K. Tanaka, Y. Nakamura, and K. Matsui. Embedding secret information into a dithered multi-level image. *Proceedings of the 1990 IEEE Military Communications Conference*, pages 216–220, September 1990.

[6] Germano Caronni. Assuring ownership rights for digital images. Submitted to ASIACRYPT'94, 1994.

[7] Robert E. Kahn. Deposit, registration and recordation in an electronic copyright management system. *IMA Intellectual Property Project Proceedings*, 1(1):111–119, January 1994.

[8] Dan Boneh and James Shaw. Collusion secure fingerprinting for digital data. Technical Report CS–TR–468-94, Princeton Computer Science Department, 1994.

[9] S. H. Low, A. M. Lepone, and N. F. Maxemchuk. Document marking and identification techniques. Submitted for publication, 1995.

[10] L. O'Gorman. Image and document processing techniques for the RightPages Electronic Library System. *Int. Conf. on Pattern Recognition (ICPR)*, pages 260–263, September 1992.

[11] L. O'Gorman. The document spectrum for structural page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11), November 1993.