Marking Text Documents

N. F. Maxemchuk
AT&T Labs - Research
Florham Park, N.J.

S. Low
University of Melbourne
Melbourne, Australia

ABSTRACT

Electronic documents are more easily copied and redistributed than paper documents. This is a major impediment to electronic publishing. Illegal redistribution can be discouraged by placing unique marks in each copy and registering the copy with the original recipient. If an illegal copy is discovered, the original recipient can be identified.

In this work we describe several invisible techniques for encoding information in text documents. We also describe a marking system, for electronic publishing, that is scalable to large numbers of users.

1: Introduction

On June 16,1993 I was at a breakfast meeting with a program director from the National Science Foundation. He stated that electronic publishing was one of the most promising applications of communications but that its use might be limited because electronic documents are too easy to copy and redistribute. He said that he had not been able to identify technical work that might solve this problem, and asked me if I knew of any techniques. I did not, but decided to look for a solution.

On returning to Bell Labs I described the problem to Jack Brassil, Abhijit Choudhury, Steven Low, Larry O'Gorman, Sanjoy Paul, and Henning Schulzrinne. The ideas came quickly. They included:

— marking copies of the documents, so that each copy is unique, and registering the copies, (If an illegal copy of a document is recovered, the original recipient is identified.)
— making it necessary for a recipient to give away personal information, for instance his credit card number, with the document, (It's unlikely that one would give his credit card number to anyone who is willing to receive illegal documents.) and,
— using encryption to make the copy that a recipient can distribute much larger than the copy that the publisher distributes. (For instance, the recipient may obtain an encrypted, Latex version of the document, but the program that decrypts the document converts the Latex to a bit map.)

We have not addressed the problem of retrieving illegal documents.

Most published documents are predominantly text. Therefore, our first application of marking is on text.[1,2,3,4,5] Our objective is to insert marks that do not visibly alter the document. We found that small movements of words or lines, are practically invisible, and are easily implemented in the postscript version of the document.

The first internal report on marking documents was issued on November 3, 1993. Jack Brassil demonstrated that our technique is easily implemented by marking and registering each copy of that report. We challenged the recipients to copy, fax, or otherwise distort the document and then to return it to us for identification. In every case, we identified the original recipient.

As a final demonstration that marking is practical, we arranged to distribute a special issue of the IEEE Journal on Selected Areas in Communications about the Internet, over the Internet[6]. Fortunately, as we were planning this demonstration, the MOSAIC software for browsing the WEB became available. The WEB eliminated the logistics of distributing client side software and operating on a wide variety of computer platforms. On November 23, 1993 Henning completed an initial demonstration of how the WEB could be used to distribute JSAC, and on November 30 the proposal for an electronic publishing trial was approved by the JSAC editorial board.

The JSAC issue was scheduled to be published in September of 1995, and became known as the SEPT issue, for Secure Electronic Publishing Trial. As with the best made plans, the SEPT issue actually appeared in October. There were over 1200 registered users in the first month, and every copy of every paper that was distributed was marked and registered with the recipient[7].

By the summer of 1996 document marking became a cottage industry. It has had its own workshop[8] and the May 1998 issue of JSAC[9] is dedicated to this topic.

Marks, or fingerprints, have been widely used[10] for identifying everything from batches of explosives to versions of maps. Our current application differs from the earlier applications in the number of marks that we are

differentiating. Trace elements that are placed in explosives identify a batch of explosives and intentional errors in maps identify the cartographer. We are using the marks to distinguish between a large numbers of individual copies.

In addition to marking text, it is possible to mark music or voice[11] and pictures[12,13,14]. This paper deals exclusively with text. The applications and problems associated with text marking are unique. The ruggedness of the marking makes it applicable to both paper and electronic publishing and the ease of decoding can provide information and service to the legal recipient as well as the publisher. However, text marks can be modified or destroyed relatively easily, which makes it necessary to consider countermeasures. We discuss the system aspects of marking text in section 2.

In the three years since the invention of this technique, several methods have been found for embedding information in text and for decoding that information. We describe these techniques in section 3.

2: System

Marks that are encoded by repositioning text can be removed or concealed. The marks can always be removed by retyping the document. A large part of this work can be automated by character recognition devices. Marks can be concealed by dithering the positions that contain information. By contrast, marks placed in pictures or speech are relatively indelible.

Text marking is well suited for protecting modestly priced documents, such as newspaper or magazine articles. We assume that if legal and illegal copies are distinguishable, and legal copies are affordable, then most people will not seek out illegal copies. A similar assumption is made by CATV operators who make premium channels unviewable by scrambling the sync signal. Inexpensive devices can restore the sync signal, but most viewers prefer to obtain the signal legally.

Countermeasures to reduce the threat of removing marks include:

— making the distortion needed to conceal marks intolerable,
— making it more difficult to remove marks,
— making it more expensive to redistribute a bootleg copy of the document than the original, and
— making it difficult to forge valid marks.

Marks can be destroyed by dithering the spaces that contain data by amounts greater than the marks. The distortion introduced by this attack is much greater than the distortion introduced by marking the document, particularly when there are many more places to hide marks than there are marks. The change in the quality of the document makes the illegal document less valuable.

Marks are made more difficult to remove by controlling the version of the document that a user has access to, and the structure of the document. In order to distribute the processing a publisher may distribute the text and marks separately, but must not make this representation easily accessible. A formatted document (postscript, pdf, etc) with the data encoded, is an economical means of storing and printing marked documents. However, a program can be written to automatically remove the marks.

As an alternative, a user can be limited to a bit map or a paper copy of a marked document. With a bit map, the user must perform character recognition on a very clean document in order to reset the document and remove the marks. If the document is straight text, in a single font, character recognition can be completely automated. If, however, the document has several different fonts - *for headers, emphasize, etc.* - has text in different places - *in figure captions, figures, tables, etc.* - or has unusual patterns of text - *as in equations* - then it likely that human assistance will be required. If the attacker is given a paper version then the document must first be scanned into a computer. The bit map will be noisy, particularly when the paper copy that has been distorted by copying or faxing, and the amount of computer work or human intervention increases.

Our final countermeasure against an attacker is to make the mark difficult to forge. As character recognition techniques improve it will be possible to automatically reset any document. While we can not prevent an attacker from removing a mark, we can make it almost impossible to replace a mark with another valid mark. A document without a valid mark is as an illegal document.

The ability to recognize illegal copies is not as strong a deterrent as the ability to identify the illegal distributor, but in many applications it is sufficient. Most major corporations or universities do not encourage or tolerate employees using illegal documents. Most individuals would rather not lose sleep over an illegal copy of an inexpensive newspaper or magazine article.

Public key cryptography[15,16] provides a standard means of including difficult to forge marks in a document. This type of a mark must contain information specific to the document, to prevent an attacker from taking a valid mark from one document and placing it on another. In addition, short encrypted messages are not secure. As an alternative we can make marks difficult to forge by using a randomly selected subset of the possible marks. For instance, assume that there are 1,000 copies of a document and that 30 bits can be encoded in the document. If the 1,000 patterns are selected at random from the $2^{30} \approx 10^9$ possible patterns, then the probability of an attacker selecting a valid mark is about 1 in a 1,000,000.

Although marking originated as a method to protect electronic documents, its ability to also protect paper documents is an increasingly important application. There are many sensitive paper documents that must be restricted. In addition, as scanning and character recognition technologies improve. paper documents will become as easy to copy and redistribute as electronic documents, Marking techniques should make electronically generated paper documents more secure than, and preferable to, mass printed documents.

Complete strategies for applying document marking to electronic publishing are described in references 1 and 17. A variant on these strategies is shown in figure 1. In this figure, the publisher multicasts the same encrypted document to every recipient. The publisher unicasts each user a decryption program that contains a unique identification number. The program decrypts the document, inserts the marks corresponding to the recipient's identification number, and converts the document to a bit map. If the recipient trys to redistribute the document, not only is it marked, but the recipient must transmit more bits than the publisher.

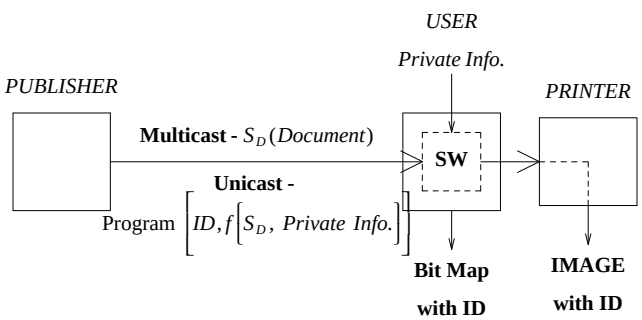

Figure 1. A Document Marking distribution System

A recipient may give away the program from the publisher instead of the document. An illegal recipient decrypts the multicast transmission rather than receiving the document from another recipient. This tactic is discouraged by requiring personal information about the original recipient in order to use the program. For instance, the secret key for the decryption program may be included in a function with the recipient's credit card number. It is unlikely that a legal recipient will give his credit card number to anyone who steals documents.

This system makes it possible to scale to large numbers of receivers by distributing the processing and sharing the transmission bandwidth. At the time of the JSAC experiment trusted, client side software would have been difficult to create and distribute. Since that experiment, JAVA has evolved as a common client side platform, and is more trusted than general programming languages.

3: Marking techniques

3.1: Encoding

A text document consists of objects of different sizes, such as paragraphs, lines, words, characters, figures, and captions. The basic idea is to encode information by moving these objects by small amounts. For instance, a text line can be moved up to encode a '1' or down to encode a '0'. The movements may be as small as a pixel, or 1/300th inch at 300 dot-per-inch (dpi) resolution. The motivation for encoding data in this manner is that moving an entire object is less perceptible than distorting the object. Encoding techniques that distort the object include dithering and modifying the transform components. We have been able to hide enough 'signal strength' in object movements to achieve accurate detection without the marks being perceptible.

For text marking we can move a paragraph vertically (or horizontally), move a text line vertically, move a block of words or a single word horizontally, or move a character horizontally. The movements can be nested or combined to encode more information. For instance a line can be moved slightly up and down while words within that line can be moved slightly left or right.

One of the major distortions suffered by documents when they are printed, photocopied or scanned, is translation and expansion or shrinkage. To compensate for this effect, we always surround a text object that is marked (moved) by two control objects that are not marked. Hence a marked text line always has two neighboring control lines that are not moved, a block of words that has been marked always has two neighboring control blocks, etc. The relative change in spacing between an object and its two control objects is used for detection, rather than the absolute movement of the object.

There is a tradeoff between imperceptibility and detection accuracy. The fractional change in a space is more perceptible than the absolute value. For instance, the space between two paragraphs is typically larger than that between two text lines, which is in turn larger than that between words or characters. Hence encoding can use a larger change in paragraph spacing than in line spacing, and larger change in line spacing than in word or character spacing, and still remain imperceptible. The text objects are different communication channels and the change in their spacing is the signal power. Increasing the signal power decreases the probability of detection error and increases the channel capacity. Imperceptibility imposes an upper bound on the signal power that can be transmitted on these channels. Hence the paragraph spacing channel has the highest allowable signal power, followed, in order, by the line spacing, word spacing and character spacing channels.

We cannot conclude, based upon the signal energy, that paragraph spacing is the best channel to hide text marks. The different channels also have different noise powers and different numbers of parallel channels (there are less paragraphs than lines, less lines than words or characters). The overall capacity for each type of channel is a function of the allowable signal power, the noise power, and the number of parallel channels. For instance paragraph spacing has the highest allowable signal power, but also has the smallest number of parallel channels per document page, and, with one of our decoding techniques, this channel has the largest noise power. Character spacing has the largest number of parallel channels but has the smallest allowable signal power and, with the same decoding technique, has a relatively high noise power.

3.2: Decoding

The first step in decoding is to remove as much of the distortion and noise introduced faxing, photocopying and scanning as possible. When the correlated components of noise are removed, the remaining noise is "whiter." Common distortions include speckle noise, skewing, translation, expansion or shrinkage of the text, blurring, random shifting, and uniform or spatially varying changes in intensity. Most speckles and skewing are removed using standard document analysis techniques.[18,19] The effects of translation are reduced by including stationary control objects in the image. The effects of expansion and shrinkage are reduced by differential encoding that uses relative, rather than absolute, values of movement. The effects of blurring and fading are reduced by using the center of mass of an object, rather than the edges.

Our decoding methods are based on the profiles of a document. The bit map of a page is represented as

$$f(x,y) \in [0,1], \quad x=0,1,\cdots,W, \; y=0,1,\cdots,L$$

where W and L are the width and length of the page in pixels, and $f(x,y)$ is the grayscale value of a pixel.

The *horizontal profile*,

$$h(y) = \sum_{x=0}^{W} f(x,y) \ ,$$

is the intensity of a horizontal scan-line. The value is high where there are lines of text and low in between lines. The *vertical profile*,

$$v(x) = \sum_{y=t}^{b} f(x,y) \ ,$$

is the intensity of the scan-lines that form a single line of text. This profile is high where there are characters and low in between characters. Figure 2 shows a typical horizontal profile of three text lines and a typical vertical profile of six words.

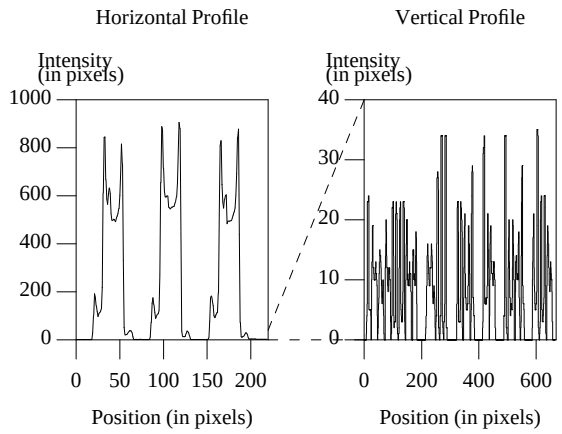

Figure 2. Horizontal and vertical profile (resolution = 300 dots-per-inch)

We have used two approaches for detection. The first approach uses the correlation between the received profile and the uncorrupted, marked profile[20] and the second approach uses the change in the 'distance' of a marked block relative to the two control blocks. With the second approach, position has been estimated using both centroids and boundaries. The boundary for a line of text is the baseline[4] and that of a word is the bounding block.[21] We have found empirically that detection based on centroids is less susceptible to noise than that based on boundaries.[4] A performance comparison of correlation and centroid detection is presented in reference [22].

4: Conclusion

Document marking is a feasible method for discouraging the wholesale copying and redistribution of text documents. Marked documents are registered with the original recipient. We can trace copies back to the original recipient or detect that the mark has been altered. The system can be scaled to large numbers of receivers.

There are several methods for decoding information that has been placed in text. Some of the techniques only require general knowledge about the structure of the document while others also use the characteristics of the specific document. In general, a decoder can more accurately extract a signal in the presence of noise when it has more information about the signal.

For instance, assume that a signal is encoded by moving lines of text up or down by a small amount relative to one another. The movement can be decoded by measuring the distance between the baselines of the text, where the baseline is the invisible line that the characters are printed on. Alternatively, the decoder can operate by determining the location of the center of mass of the line. The baselines in an unencoded text document are equally spaced. Therefore, the decoder can determine

which baselines have been shifted without receiving additional information about the document. On the other hand, the center of mass of a line is determined by the specific characters on the line and the centers of mass are not equally spaced. In order to know if a centroid is shifted up or down we must know its original location. We have found that decoders that use the center of mass are more immune to the types of noise that are encountered in paper copying and faxing than decoders that use baselines.

In general, enforcement agencies should take advantage of document specific information. These agencies operate on the behalf of the publisher and have access to the information. In addition, an enforcement agency may have to decode information from paper copies of documents that have been distorted, either intentionally or through the copying processes.

A document recipient, on the other hand, should use decoding techniques that only depend on the known characteristics of a document. Additional characteristics of specific documents have to be transmitted separately. In addition, the recipient operates on electronic copies of documents that are not distorted.

An interesting observation is that it may be useful to apply both decoding techniques to the same encoded signal. Consider information that is encoded in line displacements. Recipients may use baseline decoding to extract the signal from an undistorted, electronic copy of the document. The encoded information can provide useful information to store or operate on the document or may verify the document's authenticity. If a distorted, paper copy of the document is recovered by an enforcement agency, the agency may apply centroid decoding to the same information to extract the signal in the presence of noise.

REFERENCES

[1] J. T. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying," IEEE Infocom '94, June 14-16, 1994, Toronto, Canada, pg. 1278-87.

[2] J. T. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, "Marking Text Features of Document Images to Deter Illicit Dissemination", Int. Conf. on Pattern Recognition Israel, Oct., 1994 (in press).

[3] N. Maxemchuk, S. Low, J. T. Brassil, L. O'Gorman, "Document Marking and Identification using Both Line and Word Shifting" Infocom '95, Boston, Mass. Apr 4-6, 1995, pg 853-860.

[4] J. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying," IEEE Journal on Sel. Areas in Commun., Oct. 1995, vol. 13, no. 8, pp 1495-1504.

[5] S. H. Low, N. F. Maxemchuk, A. M. Lapone, "Document Identification for Copyright Protection Using Centroid Detection," Submitted for Publication, IEEE Trans on Commun.

[6] J. Crowcroft, D. Estrin, H. Schulzrinne, M. Schwartz, "Special Issue: The Global Internet," IEEE JSAC, Oct. 1995, vol. 13, no. 8.

[7] J. T. Brassil, A. K. Choudhury, D. M. Kristol, A. M. Lapone, S. Low, N. F. Maxemchuk, L. O'Gorman, "SEPTEMBER: Secure Electronic Publishing Trial," IEEE Communications Magazine, vol. 34, no. 5, May 1996, pp. 48-55.

[8] Ross Anderson, editor, **Proc. of First International Workshop on Information Hiding**, Cambridge, U.K., May/June 1996, Vol. 1174 of Lecture Notes in Computer Science, Springer Verlag

[9] Ross Anderson, I. Cox, S. Low, N. Maxemchuk, "Copyright and Privacy Protection," IEEE JSAC, Scheduled for publication in May 1998.

[10] N. R. Wagner, "Fingerprinting," Proceeding of the 1983 Symposium on Security and Privacy, April 25-27, 1983, Oakland, CA, pp. 18-22.

[11] D. Wasserman, "The Encoding and Automated Recognition of Audio," National Association of Broadcasters Convention, April 21, 1993, Las Vegas, NV.

[12] K. Matsui, K. Tanaka, "Video-Steganography - How to secretly embed a signature in picture," Dept. of Computer Science, The National Defense Academy.

[13] F. M. Boland, J. J. K. O Ruanaidh, C. Dautzenberg, "Watermarking Digital Images for Copyright Protection", Trinity College, Dublin, Ireland Rheinisch-Westfaelische Technische Hochschule, Aachen, Germany

[14] I. Cox, J. Kilian, T. Leighton, T. Shamoon, "Secure Spread Spectrum Watermarking For Multimedia," Proc. First Workshop on Information Hiding, Newton Institute, Cambridge, UK, May 1996.

[15] M. Willett, "A Tutorial on Public Key Cryptography" Computers & Security I (1982) pp. 72-79, North-Holland Publishing Company.

[16] W. Diffie, "The First Ten Years of Public-Key Cryptography," Proceedings of the IEEE, Vol. 76, No. 5, May 1988, pp. 560-577.

[17] A. K. Choudhury, N. F. Maxemchuk, S. Paul, H. Schulzrinne, "Copyright Protection for Electronic Publishing over Computer Networks," IEEE Network Mag., May/June 1995, vol. 9, no. 3, pg. 12-21.

[18] 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.

[19] L. O'Gorman, "The Document Spectrum for Structural Page Layout Analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(11), November 1993.

[20] S. H. Low, A. M. Lapone, N. F. Maxemchuk, "Document Identification to Discourage Illicit Copying ," IEEE GlobeCom 95, Nov. 13-17 1995, Singapore.

[21] J. T. Brassil, L. O'Gorman, "Watermarking Document Images With Bounding Box Expansion," Proc. First Workshop on Information Hiding, Newton Institute, Cambridge, UK, May 1996.

[22] S. H. Low, N. F. Maxemchuk, A. M. Lapone, "Document Marking and Identification Techniques and their Comparison," Technical report, University of Melbourne, Department of Electrical & Electronic Engineering, 1996. Submitted for publication.