A FEATURE EXTRACTING METHOD FOR TAMPER DETECTION IN PRINTED DOCUMENTS

Yoshiyasu Takahashi, Takaaki Yamada and Seiichi Susaki
Systems Development Laboratory, Hitachi ltd., Totsuka-ku Yoshida-cho 292, Yokohama, Kanagawa, Japan

Keywords: Feature Extracting Method, Dot-Pattern Code, Digital Watermarking, Electronic Delivery, e-Government.

Abstract: In this paper, we propose our feature extracting method for tamper detection in printed documents. To detect the tamper of the printed document, a feature extracting method is needed. In this paper, we describe our feature extracting method. Our feature extracting method is based on the location of the mean point of each dot. We have estimated our feature extracting method's probability of the collision, its uniformity of the distribution, its invariability during D/A and A/D transform and its invariability during the ordinal change of paper. We have found that our feature extracting method can extract desirable feature value.

1 INTRODUCTION

Since 2000, the Japanese Government has promoted the IT strategy along with the implementation of various structural reforms. (IT Strategic Headquarters, 2007).

However, several issues still remain to solve. One of the issues is the lower utilization rate of e-government.

We consider the reasons of the lower utilization rate are three; complicated preparation for online application, security worry and the lack of the means of the online delivery.

To resolve the third reason, we can envisage a system that the government delivers certificates online and the user also receives them online, and the user prints them with their home printers and use or submit them to the destination in paper format.

To bring the scenario to fruition, the two means are needed; a means to prove that the paper is original print and not copy, and a means to prove that the content of the paper is correct and not tampered.

To detect the copy, the "copy indicator" technique is widely known.

To detect the tamper of the paper, the extracting method of the paper's feature is needed. With that method, we can extract the feature of the printing document and print it within the paper. After the paper is distributed, the verifier can extract both the feature of the paper and the feature embedded, and compare them. However, the extracting method of the paper's feature is not well known.

In this paper, we describe the requirements for the feature extracting method and introduce our innovative method to extract the paper's feature. We also make estimation of our method and compare with another extracting method.

2 REQUIREMENTS FOR THE FEATURE EXTRACTING METHOD

For the feature extracting, the requirements are following five.

1) The feature values calculated from two different documents should not be equal

2) The distribution of the feature values should be uniform.

3) The document should not be calculated from its feature value.

4) The feature values should not vary during the D/A and A/D transform.

5) The feature values should not vary if the paper document is folded or got wrinkles or tainted.

For the feature extracting, the problem is more difficult than the hash values because of the requirement 4 and 5. Because the D/A and A/D transform is inevitable, the feature value should not vary during them. Moreover, because paper is used, folding, wrinkles or taint easily occur, the feature value should keep the same value during the ordinal alteration of the document. At the same time, if the significant change occurs in the document, even if

Takahashi Y., Yamada T. and Susaki S. (2008).
A FEATURE EXTRACTING METHOD FOR TAMPER DETECTION IN PRINTED DOCUMENTS.
In *Proceedings of the International Conference on e-Business*, pages 429-432
DOI: 10.5220/0001905804290432

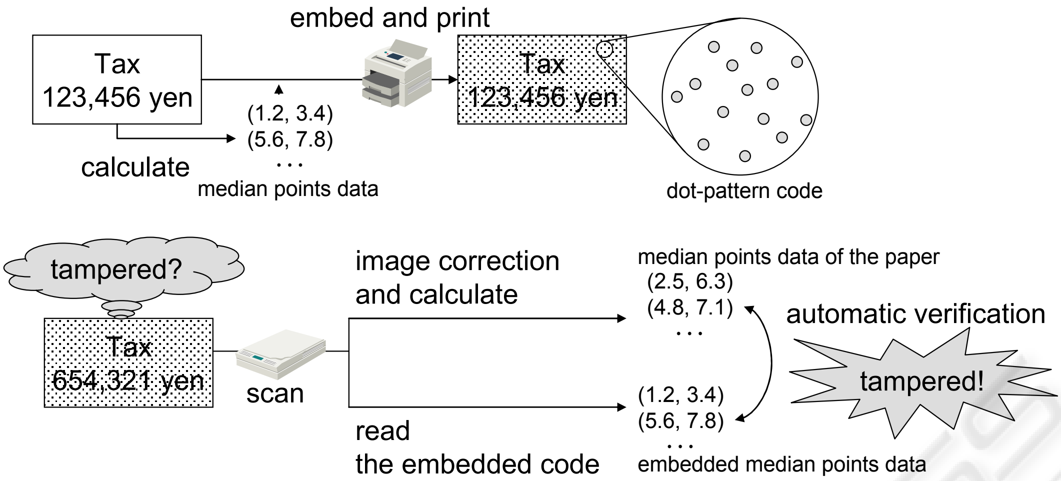

Figure 1: Overview of the verification.

the change is minute, the feature value should change the other values to represent the change of the document.

3 RELATED WORKS

The ideal feature value has not been found yet. However, several techniques that meet a part of the requirements are studied to extract the paper's feature to detect the tampering.

There are three types of the paper's feature below.

1) using the document itself
2) using the semi-fragile watermarking
3) using the feature of the paper

First, there are methods to use (a part of) the original document, such as the text data or the image data. To embed them, 2D-code is usually used. To check the document's tamper, usually a man power is needed; a verifier scans the code and extract the document's text data or image data, and compare them, watching side-by-side. Its cons are its easiness to implement and its robustness for the paper's conditions, and its pros are that it needs a verifier's help and time to verify.

Second, the use of semi-fragile watermark method has been proposed. The semi-fragile watermark is a watermark which is robust to some degradation such as compression, while it is at the same time destroyed if the embedded image is tampered with. However, the semi-fragile watermark which is robust to the paper's folding or wrinkles or tainting is not well-known. Therefore, workable proposal of this type of method has not been yet.

Third, using the feature of the paper has been proposed. Suzaki and et al. has proposed the watermarking technique for printed documents by superposing dot pattern blocks on backgrounds of the document image (Suzaki, 2003).

4 PROPOSING THE FEATURE EXTRACTING METHOD

4.1 Brief Overview

Figure 1 shows the overview of our document verification system.

To print the certificate, the following procedures are performed.

1) The issuer makes the certificate.

2) The system calculates the feature value from the certificate, and embeds it into the certificate itself. The concrete means of extracting and embedding of the feature value are described later.

3) The system prints the feature value embedded certificate with printers. In the electronic delivery, this process can be performed at the user's home.

After the issue, because the certificate circulates among many people, there is a risk of tampering the certificate. Therefore, the submitted certificate needs to be verified. With the following procedures, the verifier can verify the certificate.

1) The verifier scans the certificate and detect the embedded feature value.

2) The verifier also calculates the feature value from the submitted certificate.

3) The verifier compares the two feature values. When the significant difference is detected, the verifier judges the certificate as tampered. The concrete means of extracting and embedding the feature value are described below.

4.2 Extracting the Feature Value

In our method, we use the median point as the feature value of the document, as shown in Figure 2. In the Figure 2, the plus sign illustrates the median point of the character “4”.

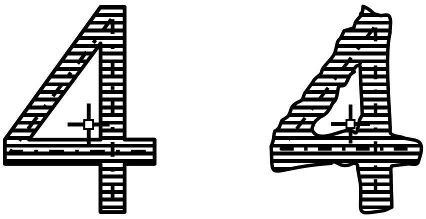

Figure 2: Sample of the feature value.

The benefits of the median point as the feature value are the robustness against the blur because of the scattering of the tonner or ink of printers and of the focus error of scanners. Figure 2 also illustrates this merit. In Figure 2 left, the character “4” is not blurred. This illustrates the character “4” to calculate the feature value in issuer’s computer. After the printing, circulating, and scanning, the character may blur. The Figure 2 right illustrates this situation.

Although the blur, the median point of the character “4” of the Figure 2 right does not vary very much. That is because the width of the line of the character is almost the same.

To embed the feature value into the Document, we use the dot-pattern code (Takahashi 2007, 2008), which the authors have developed before. The dot-pattern code is the information embedding method onto the paper with minute dots.

5 ESTIMATION

5.1 Assessment Viewpoints of the Feature Value

The viewpoints of the assessment are following five.

(A)probability of the collision
(B)uniformity of distribution
(C)probability of inverse calculation
(D)invariability during D/A and A/D transform
(E)invariability during the ordinal change of paper

5.2 Probability of the Collision and Uniformity of Distribution

To estimate the probability of the collision and the uniformity of distribution, we calculated the median points of characters. The characters are alphabet and number of Times New Roman font. We drew a character one by one on the white image and calculated its median point.

The result is shown in Figure 3. In Figure 3, the median points of each character are plotted with dots. From the Figure 3, though there is some bias, we can find that the median points are distributed almost uniformly. Moreover, there are no dots which collide against another dot.

To compare, we also calculated the area of each alphabet and number. In Figure 4, the histogram of the result is shown. The x-axis is the area and the y-axis is occurrence. From Figure 4, we can find the area is distributed not uniformly. Indeed, nine characters collided; i.e. they have the same area. Therefore we can conclude that our feature value is better than the area.

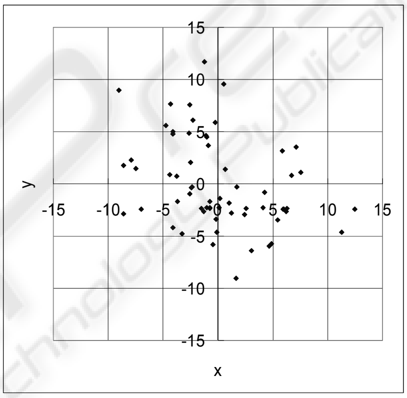

Figure 3: Median points of a character.

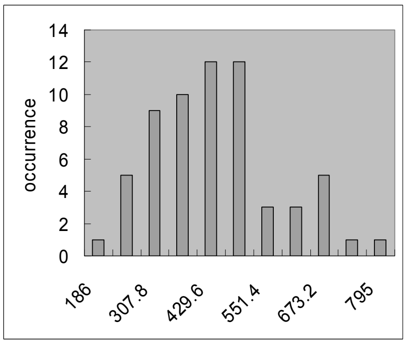

Figure 4: Area of a character.

5.3 Invariability during D/A and A/D Transform

To estimate the invariability during D/A and A/D transform, we viewed the difference of the median point’s change between before and after printing the document.

We calculate the median point and area of a character "a" in digital format. Then, we print it and scanned and calculated the median point and area again from the scanned image. Before calculating the median point and area, we converted the scanned image to binary image with various thresholds.

Table 1: Invariability during D/A and A/D transform.

thrs	median point	diff	area	diff
192	(149.12, 146.86)	2.33	260	128
128	(150.28, 148.44)	0.58	186	54
64	(150.02, 148.22)	0.76	152	20
32	(150.49, 148.31)	0.78	116	16
16	(150.41, 148.89)	0.34	108	24
8	(150.46, 148.75)	0.44	104	28
orig	(150.08, 148.98)	---	132	---

The result is shown in Table 1. From Table 1, we can find that though the threshold varies (thrs), the median point does not vary very much. On the other hand, the area varies very much and the difference of the area of the original image. There are twelve characters whose area's difference is less than 54. Therefore we cannot distinguish these twelve characters with area.

5.4 Invariability during the Ordinal Change of Paper

To estimate the invariability during the ordinal change of paper, we viewed the difference of the median points' change between before and after folding of the paper. The experiment is done as 5.3.

Table 2: Invariability during the ordinal change of paper.

thrs	median point	diff	area	diff
192	(146.03, 144.40)	6.11	296	164
128	(150.59, 148.79)	0.53	164	32
64	(150.51, 148.59)	0.57	59	73
orig	(150.08, 148.98)	---	132	---

The result is shown in Table 2. In this experiment, our median point shows near to original median point. Therefore we can conclude that our proposed method has good invariability during the folding of the paper.

6 CONCLUSIONS

In this paper, we have introduced our feature extracting method of paper document. Our feature extracting method is based on the location of the mean point of each dot, and is expected to be applicable to home printers such as inkjets.

We have estimated the probability of the collision and the uniformity of the distribution of our feature extracting method. We have found that the feature value extracted from one character is distributed uniformly and do not collide each other.

We have also checked invariability during D/A and A/D transform and we have found that almost every character can be distinguished with our feature extracting method even after D/A and A/D transform. We have also found that our feature extracting method is better than the area.

We have also checked the invariability during the ordinal change of paper and found that our method has enough and better invariability than the area.

Therefore, we can conclude our proposed method can extract desirable feature value.

REFERENCES

IT Strategic Headquarters, 2007. *Priority Policy Program 2007*. The Japan Government. Japan.

Ching-Yung Lin, Shih-Fu Chang, 2000. Semi-Fragile Watermarking for Authenticating JPEG Visual Content. In *SPIE Security and Watermarking of Multimedia Content II*, pp.140-151.

Ke DING, Chen HE, Ling-ge JIANG, Hong-xia WANG, 2005. Wavelet-Based Semi-Fragile Watermarking with Tamper Detection. In *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, Vol.E88-A, No.3*, pp.787-790.

Suzaki M., Mitsui Y., Suto M., 2003. New Alteration Detecting Technique for Printed Documents using Dot Pattern Watermarking. In *SPIE, Volume 5020*, pp. 665-676.

Takahashi Y., Yamada T., Ebisawa R., Fujii Y., Tezuka S., 2007. Research and Development of Dot Pattern Technology for Commodity Printers. In Proc. of *FIT 2007, the Forum on Information Technology*, pp. 325-326.

Takahashi Y., Yamada T., Ebisawa R., Fujii Y., Tezuka S., 2008. Information Embedding Method for Home Printing of Certifications. In Proc. of *ICACT2008, the 10th International Conference on Advanced Communication Technologies*, pp.2116-2120.

Echizen I., 2007. Digital Watermarking Technique and its Application. In *IPSJ Magazine Vol.47 No.11*, pp.1243-1249.