http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (11): 2130-2137, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.2130.2137

Skeleton-based Chinese Text Image Watermark Algorithm Robust to Printing and Scanning

Xingming Sun, Shufang Wang, Zhihua Xia and Xinhui Wang
School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China

Abstract: The advances in digital media make the copyright protection more and more important. Digital watermarking provides a copyright protection solution to this problem. The text is the most popular medium over the internet and many researchers have proposed text watermarking methods in past years. In this study, we propose a new watermarking method based on skeleton algorithm for hiding messages in Chinese text image. In the embedding process, on the basis of character segmentation, the skeleton of each character is extracted and then the location of the bounding box of each skeleton is recorded. Secondly, in each text line, the average center line of the bounding boxes is calculated and the centers of character skeletons are shifted to a position higher or lower than the average center line. The shifting pattern constitutes the watermark. In the extracting process, we firstly conduct the binarization and deskewing operation to the printed-scanned image. Then the remaining phase is very similar to the embedding process. The experimental results prove that the proposed method can successfully resist the print and scan attack and holds much better robustness than that does not use skeleton algorithm.

Key words: Text image watermarking, skeleton extracting, print and scan attack, character segmentation, deskewing

INTRODUCTION

The advances in digital media and internet make the distribution of text documents increasingly convenient. However, these also lead to enormous copyright issues. Digital text watermarking provides a copyright protection solution to this problem by embedding copyright information imperceptibly into the digital text document in a way that is robust to various kinds of attacks (Liu and Tsai, 2007; Luo and Zhang, 2011; Peng *et al*., 2012). But the simple image interconversion between print and digital forms makes it urgent to study text image watermark which is resistant to print and scan operation.

In order to improve the robustness of watermark, many researchers embed watermark by choosing the coefficient in transform domain such as Discrete Fourier Transform (DFT) (Pereira and Pun, 2000; Cedillo Hernandez *et al*., 2012), Discrete Cosine Transform (DCT) (Chu, 2003; Briassouli *et al*., 2005), Discrete Wavelet Transform (DWT) (Nikolaidis and Pitas, 2003; Zhaoqian *et al*., 2012). Specifically, Chu described a DCT-based watermarking algorithm which employed different modifications to the DCT coefficients pertaining to different sub-images (Chu, 2003). Although the watermark embedded in transform domain has stronger robustness, it is more suitable for natural images. Different form natural images, the characters in text images are very clean cut. Therefore, it is obvious that the change of the high-frequency coefficient will just alter the edge pixels of characters and disturb the human visual effect.

Some researchers embed the watermark in text images by adjusting the document structure such as line spacing and word spacing. Brassil *et al*. (1995, 1999) proposed three methods for text documents watermarking: (a) Line shifting coding, (b) Word shifting coding, (c) Feature coding. In (a), The embedded line was shifted up or down slightly and its two neighbor lines remained unmoved. And in (b), The words in each line were divided into three blocks and the middle block was shifted left or right when embedding. The blocks adjacent to the middle one were not marked. Both of the first two methods needed control information (control lines or control blocks) when embedding. So this greatly reduced the embedding capacity. Moreover, they also required the original document during the detection phase. The third method (c) Embeds the information by altering special text features (e.g., upward or vertical endlines); the features

Corresponding Author: Xingming Sun, School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China Tel: 025-58731575 Fax: 025-58731575

are altered or not depending on the watermarking information. This method also required original document when extracting information and it was vulnerable to the noise.

The line spacing and word spacing in a text image is a relative stable feature which can resist print and scan attack. And human beings are not visually sensitive to the tiny change, so the watermark algorithm based on the text image's line spacing and word spacing have been researched a lot e (Maxemchuk, 1994; Low *et al*., 1995, 1998; Low and Maxemchuk, 2000; Huang and Yan, 2001; Yang and Kot, 2004; Zou and Shi, 2005).

This study presents a text image watermarking method based on the character skeletons. The watermarks are embedded by shifting the characters up or down. Since the character skeleton can resist print and scan attack better than the character itself, present method is expected to hold stronger robustness.

SKELETON EXTRACTION ALGORITHM

The skeleton of a region may be defined via the Medial Axis Transformation (MAT) proposed by Blum (1967). The core idea of the skeleton extraction algorithm is to delete edge points on the region iteratively. Although, the skeleton only has one-pixel width, it presents the character's geometrical and topological properties; so it is widely used in pattern recognition and image analysis (Tarabek, 2012).

The skeleton extraction algorithm is realized mainly by a series of sequential thinning algorithm using the structuring elements (Lam *et al*., 1992). And each individual thinning pass is achieved by using eight structuring elements to erode the characters repeatedly. The entire process is repeated until the character has no further changes occur. The two of the eight structuring elements are shown in Fig. 1.

Where '1' stands for the black pixel and '0' stands for the white background; '*' indicates "don't care", that means it may be 0 or 1. And the other six structuring elements are got by rotating the two with angle 90°, 180° and 270°.

Figure 2 plots the result of skeleton extraction. We can see in Fig. 2b the edge of the characters is easily to be affected by noise but there is almost no difference between Fig. 2c and d. So the skeleton is insusceptible to the effect of character edge noise and is resistant to the print and scan attack to some extent.

WATERMARK EMBEDDING AND EXTRACTING PROCEDURE

Embedding procedure: The main thought of embedding watermark is to shift the center line of the skeleton upper or lower than the average one and it performs as the shifting of characters. The detailed watermark embedding procedure is shown in Fig. 3. And Fig. 4 provides an instance of watermark embedding process.

The embedding process starts by segmenting the characters. Next, the skeletons of characters are extracted and the bounding boxes' locations of all character skeletons are obtained by using the method similar to character segmentation. Then the location of the average center line (Fig. 4d) of each text line is computed. Finally

$$\begin{bmatrix} 1 & 1 & 1 \\ * & 1 & * \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} * & 1 & * \\ 0 & 1 & 1 \\ 0 & 0 & * \end{bmatrix}$$

Fig. 1: Two of the eight structuring elements

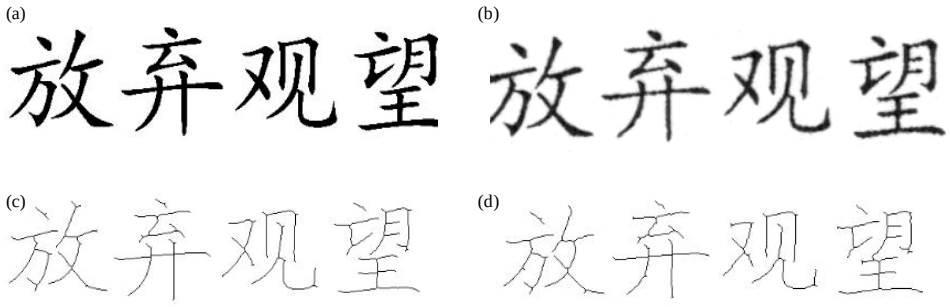

Fig. 2(a-d): Comparison of the skeleton extraction result between digital image and scanned image, (a) Digital image, (b) Scanned image, (c) Skeleton extraction result of digital image and (d) Skeleton extraction result of scanned image

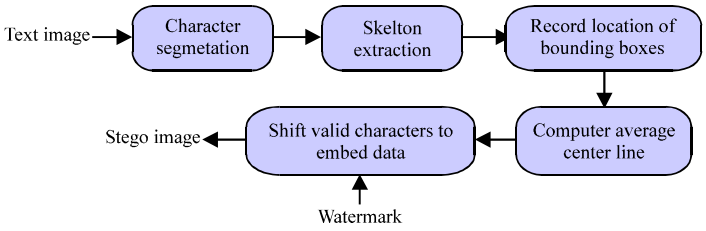

Fig. 3: Watermark embedding procedure

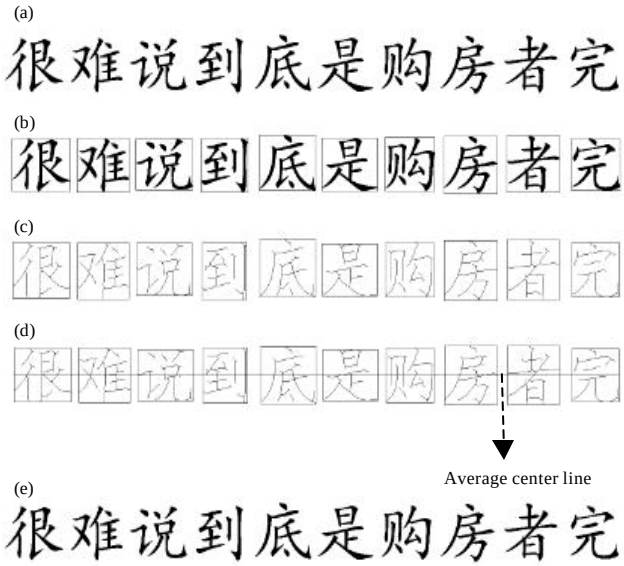

Fig. 4(a-e): Instance of watermark embedding process (a) The original text image, (b) Character segmentation, (c) Skeleton extraction and record bounding boxes' locations, (d) Shift the center lines of valid characters to the average center line and (e) Shift characters up or down to complete watermark embedding

the watermark is embedded by shifting the valid characters to make their center line is upper or lower than the average center line.

Figure 4c and d, the operation of skeleton extraction and skeleton segmentation is on the basis of character segmentation. Since the connected domain will change greatly before and after the skeleton extraction operation, it may lead to wrong segmentation of characters if we segment the skeleton directly after the skeleton extraction operation. So this operation ensures the consistency of the character segmentation before and after the skeleton extraction operation and lays the foundation for the correct extraction of the watermark.

In the same text line, the center lines of Chinese characters are basically in the same level because of their structural features. So, it is imperceptible to shift the characters up or down slightly except for some special characters like punctuation, '-' etc. Hence, we rule that in a same text line, the characters whose width, height is bigger than half of the character's average width and height in this line are the valid ones; and the invalid characters are not allowed to embed watermark.

Extracting procedure: The extraction process is described in Fig. 5. The text image we used to extract can be got by scanning a hard copy document which must contain watermark information.

First of all, several image preprocessing operations shall be used to remove the noises and distortions in the image, such as edge cropping, binarization and deskewing. Then the following operations are similar to the embedding process: character segmentation, skeleton extraction and record locations of bounding boxes. After we get the bounding box of each skeleton, we only need

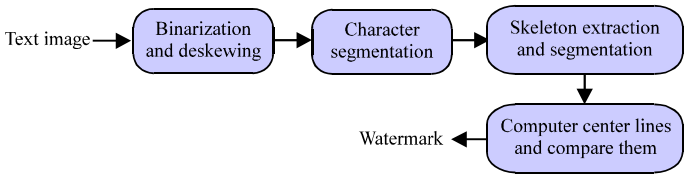

Fig. 5: The watermark extracting process

to compute center lines of valid characters' skeletons. At last, calculate the sum of center lines in different group and we can get the watermark by comparing the values in different group.

CHARACTER SEGMENTATION

Present algorithm is achieved by shifting the characters up or down, so accurate character segmentation is very important in the watermarking system. In this study all the characters are segmented into connected domains firstly by projecting the image horizontally and vertically. Then the rough segmentation is refined by merging the connected domains with given rules according to the peculiarity of Chinese characters. The specific process is as follows:

Segmentation of connected domains: By projecting the image horizontally and vertically firstly, we can get separate connected domains shown in Fig. 6a and each box indicates a connected domain.

However, the characters with conjoint strokes cannot simply adopt the above operations. Besides, if we have judged it is a conjoint strokes connected domain shown in Fig. 6b according to its height, we can segment it into separate connected domains by finding the weakest connections through vertical projection.

The merging of connected domains: Some left-right Chinese characters (很, 利 in Fig. 6a) are segmented into two connected domains in the first step, so the connected domains need to be merged to complete the character segmentation. We use aspect ratio R to merge the connected domains and its theory basis is: except for some individual characters, no matter what font, size, the R of the rest Chinese characters is close to 1.

Consequently, before merging the connected domains, we should identify and segment the special characters and not allow them to take part in the merging operation. For the punctuation shown in Fig. 7a, after segmenting the connection domains, these kinds of characters can be identified by their features such as position coordinates; For Chinese characters like '川'

Fig. 6(a-b): Different conditions of connected domains when segmentation, (a) Connected domain segmentation and (b) Conjoint strokes

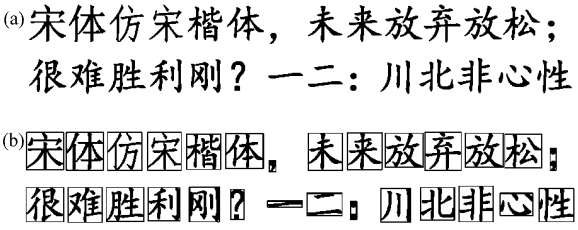

Fig. 7(a-b): Segmentation result of different types of character, (a) The original digital text image and (b) The result of character segmentation

shown in Fig. 7a, we set a threshold according to present experiment experience to identify and segment them.

Then the specific merging process is: Compute R_1 of a selected connected domain Q_1 and R_2 of the merged connected domain Q_2 got by merging Q_1 with its adjacent right one. The connected domain whose R is closer to the given threshold can be segmented into an independent character. Then use the same method as getting Q_2 to get Q_n (n = 3, 4, 5,…) and compute their R to decide whether they can be merged. If not, then repeat the operations mentioned above to complete character segmentation.

The result of character segmentation is shown in Fig. 7b.

BINARIZATION AND DESKEWING ALGORITHM

Image binarization algorithm: Binarization is of great importance in document processing such as character

segmentation and document layout analysis. Many document binarization methods have been presented in the literature (Moghaddam and Cheriet, 2011; Ntirogiannis *et al*., 2012). In this study we use an iterative method which is based on the thought of approximation shown in Alg. 1. Figure 8 shows the effect before and after the binarization.

Algorithm 1: The procedure of binarization algorithm

Require: T_{max}(T_{min}): the maximum (minimum) gray value in a Chinese text image,

Calculate T_{max} and T_{min} of a image and then set $T = T_{max}+T_{min}$;
Use T to divide the image into 2 groups and calculate the average gray value T_1, T_2 in each group;
while T_1,T_2 has changed
$T = (T_1+T_2)/2$;
Use T to divide the image into 2 groups and recalculate T_1, T_2 in each group;
end while
We can get the threshold T and then use it to binarize the image

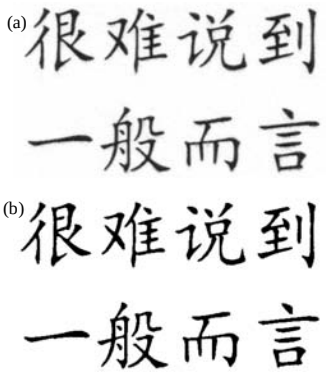

Fig. 8(a-b): Comparison of the scanned image and its binarization result (a) The scanned text image (b) The binarization result of the scanned image

Image deskewing algorithm: After being printed and scanned the image will inevitably be tilted. According to the theory of the algorithm, we can find that the detection phase is susceptible to skew correction. So before we do further operations to the image, we should carry on the deskewing operation to it.

When projecting the image horizontally, we find that, the S (the sum of inter-line blank spacing) of the correct image had a big difference from the tilt one's shown in Fig. 9. Then tests indicate that the bigger the tilt angle of the image, the smaller the sum of inter-line blank spacing. According to this feature we put forward a deskewing algorithm based on S. The detailed procedure is shown in Alg. 2.

Algorithm 2: Procedure of deskewing algorithm

Require: I: the Chinese text image, S_r (S_l): the S after rotating I clockwise (anticlockwise), S_a (S_b): the S after (before) rotating I , A: the rotation angle, P: a given precision,

Load the image I and project it horizontally and then calculate S;
Rotate I clockwise and anticlockwise and calculate its corresponding S_r and S_l;
if $S_r > S_l$
the tilt correction is upright; $S_b = S_r$;
else
the tilt correction is leftright; $S_b = S_l$;
end if
while A >= P
rotate I by angle A in correct direction; then calculate S_a;
while
$S_b = S_a$;
keep rotating I in right direction with angle A and recalculate S_a;
end while
if $S_a <= S_b$
reduce A;
end if
end while
if $A < P$
we can get the tilt angle W;
end if

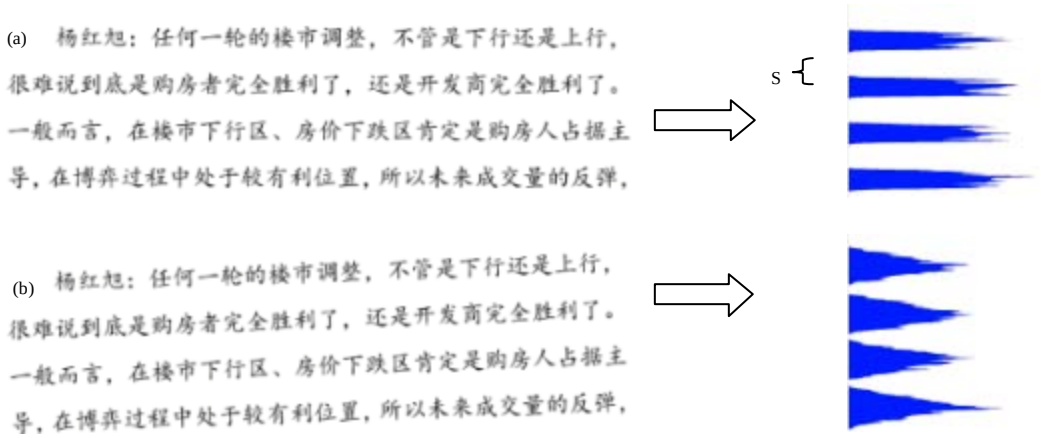

Fig. 9(a-b): Horizontal projection of the correct image and tilt image, (a) Horizontal projection of correct image (b) Horizontal projection of tilt image

In order to verify the validity of this algorithm, we test 10 correct samples. For a rotation of 0.4° clockwise and anticlockwise to the samples respectively and then use this algorithm to compute the tilt angle. The result is shown in Table 1. From the table we can see the absolute value of the computed angles and the actual angle difference is between 0° (-0.400000°) and 0.00547° (-0.394530°).

EXPERIMENT RESULTS

To test the robustness of present watermarking algorithm, the extraction accuracy of two methods were compared. One method is proposed in present study and the other method makes the average center line of characters as the conference line when shifting characters. We used them to test a 30-page Chinese text image respectively. The text images were formatted separately in three fonts (Song, Kai and Fangsong) and in four different font sizes (18, 16, 14 and 10.5 pt); thus, there were 720 images in all. Characters with 18 and 16 pt are usually used in official notice and most documents adopt characters with 14 and 10.5 pt.

The advantages and disadvantages among different methods provide in Brassil *et al.* (1995, 1999), Nikolaidis and Pitas (2003) and Zou and Shi (2005) are listed in Table 2. And we also compared the method with and without skeleton in extraction accuracy and embedding capacity; the specific process is as follows. The Parameter Settings in the Embedding and Extracting phase.

Table 1: The test result of the deskewing algorithm

Samples	-0.4°	0.4°	Samples	-0.4°	0.4°
1	-0.403238	0.397195	6	-0.401785	0.396123
2	-0.401759	0.400577	7	-0.400000	0.405286
3	-0.400330	0.401637	8	-0.394530	0.398770
4	-0.400352	0.403878	9	-0.403960	0.397799
5	-0.402768	0.396494	10	-0.400115	0.400686

During the embedding phase, binary Chinese text images with the resolution of 600 dpi were used to embed watermark. In order to ensure the robustness of watermarking, embedding one bit needed 10 valid characters at lease. In present experiment, we embedded two bits in one text line, so a valid line must have 20 valid Chinese characters at least. For example in Fig. 10, there are 24 valid Chinese characters left in the current line. So, we divided the words into 4 groups evenly and each group had 6 words. Then we embedded the first bit with the first part and the second bit with the second part. If we embed 01, then move the characters in group 2 one pixel up and move characters in group 3 one pixel up. The unmoved characters in group 1 and 4 served as reference locations in the extracting phase.

In the extracting phase, we used a HP LaserJet M1536dnf MFP to print and scan the embedded hard-copy document with resolution of 600 dpi. And the scanned image most likely had rotational distortions. Then we used watermark extracting method introduced to get the watermarking.

The extraction result of the two methods: Table 3 and 4, respectively present the extraction accuracy of the image scanning with different ways. Images used in Table 3 is got by automatically bulk scanning while in Table 4 the image is scanned through manual single placement in the

Table 2: Comparison among different methods

Methods	Suitable for Chinese text images	Embedding capacity	Robust to printing and scanning	Require the original document when extracting
Transform domain	No	-	Yes	No
Line shifting	Yes	0.5 bit/line	Yes (not good)	Yes
Word shifting	Yes	1 bit/line	A little	Yes
Inter-word space modulate	Yes	1 bit/line	Yes	No
Center line shifting (skeleton-based)	Yes	≥1 bit/line	Yes	No

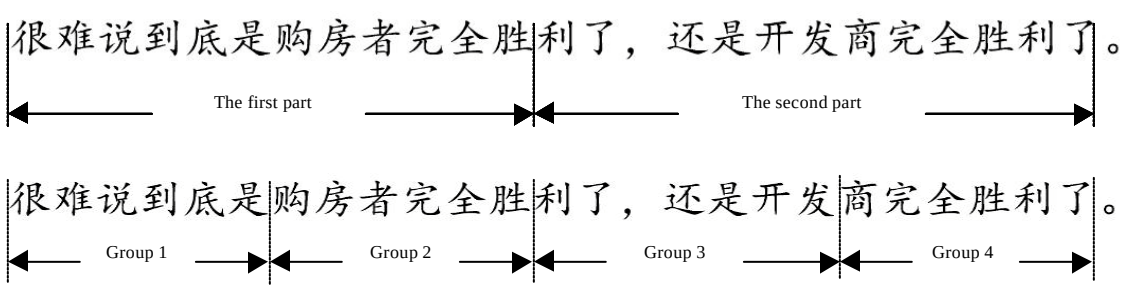

Fig. 10: An instance of grouping process

Table 3: Extraction accuracy result in automatically bulk scanning

		Font size (%)			
Font		18 pt	16 pt	14 pt	10.5 pt
Song	With skeleton	87.0	80.0	89.0	82.0
	Without skeleton	63.0	49.0	57.0	60.0
Kai	With skeleton	91.1	82.0	82.5	78.5
	Without skeleton	59.0	62.5	76.0	60.0
Fangsong	With skeleton	86.0	84.0	82.0	82.0
	Without skeleton	90.5	51.0	54.0	52.0

Table 4: Extraction accuracy result through manual single scanning

		Font size (%)			
Font		18 pt	16 pt	14 pt	10.5 pt
Song	With skeleton	95.0	93.5	94.0	98.5
	Without skeleton	85.5	84.5	94.5	89.5
Kai	With skeleton	94.0	94.5	93.5	94.5
	Without skeleton	88.5	94.0	94.0	91.5
Fangsong	With skeleton	92.0	94.0	97.0	95.0
	Without skeleton	83.5	90.5	92.5	93.5

Table 5: Embedding capacity of different font and font size

	Font size (%)			
Font	18 pt	16 pt	14 pt	10.5 pt
Song	33	39	38	74
Kai	35	40	40	78
Fangsong	35	40	40	79

scanner. From the two tables we observe that the method with skeleton performs better than the one without especially in the condition of automatically bulk scanning. Besides, in Table 4 the characters with font size in 14 pt perform better than others relatively.

The statistical result of the embedding capacity: At last, we also counted the embedding capacities of each image. According to our experience, we rule that if we want to embed one bit, we must have 10 valid characters at least and the text line which has valid characters between 10 and 20 can embed one bit, if the number is more than 30, it can embed 3 bit in one text line. Thus, the embedding capacity can be increased. So we count those conditions in. The average embedding capacity of different fonts and font sizes is shown in Table 5. From the table we can see that for different fonts, the smaller the font size, the bigger the embedding capacity.

CONCLUSION

In this study, we put forward a new watermark embedding method which is robust to print and scan attack. The method which adopts the center line of the characters skeleton as the conference line embeds data into text image by slightly shifting the characters to make its center line upper or lower than the average one. And the shifting pattern constitutes the mark, so we can extract the data by comparing the sum of skeleton's center lines in different groups. According to our experiment results, we can find the method we described performs better than the method which dose not utilize skeleton algorithm.

In the future, we plan to improve the skeleton algorithm which will be more efficient in time and much more robust to noise. We also plan to perform the method which combines the embedding algorithm of shifting the characters left or right slightly and skeleton algorithm together.

ACKNOWLEDGMENTS

This study is supported by the NSFC (61232016, 61202496, 61173141, 61173142, 61103141, 61173136, 61103215, 61070196, 61070195 and 61073191), National Basic Research Program 973 (2011CB311808), 2011GK2009, GYHY201206033, 201301030, 2013DFG12860, Research Start-Up fund of NUIST, Grant No. 20110428 and PAPD fund.

REFERENCES

Blum, H., 1967. A Transformation for Extracting New Descriptors of Shape. In: Models for the Perception of Speech and Visual Forms, Wathen-Dunn, W. (Ed.). MIT Press, Cambridge, MA., pp: 362-380.

Brassil, J.T., S. Low, N.F. Maxemchuk and L. O'Gorman, 1995. Electronic marking and identification techniques to discourage document copying. IEEE J. Selected Areas Commun., 13: 1495-1504.

Brassil, J.T., S. Low and N.F. Maxemchuk, 1999. Copyright protection for the electronic distribution of text documents. Proc. IEEE, 87: 1181-1196.

Briassouli, A., P. Tsakalides and A. Stouraitis, 2005. Hidden messages in heavy-tails: Dct-domain watermark detection using alpha-stable models. IEEE Trans. Multimedia, 7: 700-715.

Cedillo Hernandez, M., F. Garcia Ugalde, M. Nakano-Miyatake and H. Perez Meana, 2012. Robust digital image watermarking using interest points and dft domain. Proceedings of the 35th International Conference on Telecommunications and Signal Processing, July 3-4, 2012, Prague, Czech Republic, pp: 715-719.

Chu, W.C., 2003. DCT-Based image watermarking using sub sampling. IEEE Trans. Multimedia, 5: 34-38.

Huang, D. and H. Yan, 2001. Interword distance changes represented by sine waves for watermarking text images. IEEE. T. Circ. Syst. Video Technol., 11: 1237-1245.

Lam, L., S.W. Lee and C.Y. Suen, 1992. Thinning methodologies: A comprehensive survey. IEEE Trans. Pattern Anal. Machine Intel., 14: 869-885.

Liu, T.Y. and W.H. Tsai, 2007. A new steganographic method for data hiding in microsoft word documents by a change tracking technique. IEEE Trans. Information Forensics Security, 2: 24-30.

Low, S.H., N.F. Maxemchuk and A.M. Lapone, 1998. Document identification for copyright protection using centroid detection. IEEE Trans. Commun., 46: 372-383.

Low, S.H. and N.F. Maxemchuk, 2000. Capacity of text marking channel. IEEE Signal Process. Lett., 7: 345-347.

Low, S.H., N.F. Maxemchuk, J.T. Brassil and L. O'Gorman, 1995. Document marking and identification using both line and word shifting. Proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communications Societies, April 2-6, 1995, IEEE Computer Society, Washington, DC. USA., pp: 853-860.

Luo, M. and Y.X. Zhang, 2011. Watermarking Chinese text document based on structure and semantics of Chinese characters. Comput. Knowl. Technol., Vol. 34.

Maxemchuk, N.F., 1994. Electronic document distribution. ATT Tech. J., 73: 73-80.

Moghaddam, R.F. and M. Cheriet, 2011. Adotsu: An adaptive and parameterless generalization of otsu's method for document image binarization. Pattern Recognit., 45: 2419-2431.

Nikolaidis, A. and I. Pitas, 2003. Asymptotically optimal detection for additive watermarking in the dct and dwt domains. IEEE Trans. Image Process., 12: 563-571.

Ntirogiannis, K., B. Gatos and I. Pratikakis, 2012. A combined approach for the binarization of handwritten document images. Pattern Recognit. Lett. 10.1016/j.patrec.2012.09.026

Peng, F., X. Li and B. Yang, 2012. Adaptive reversible data hiding scheme based on integer transform. Signal Process., 92: 54-62.

Pereira, S. and T. Pun, 2000. Robust template matching for affine resistant image watermarks. Proc. IEEE Trans. Image Process., 9: 1123-1129.

Tarabek, P., 2012. A robust parallel thinning algorithm for pattern recognition. Proceedings of the 7th IEEE International Symposium on Applied Computational Intelligence and Informatics, May 24-26, 2012, Timisoara, pp: 75-79.

Yang, H. and A.C. Kot, 2004. Text document authentication by integrating inter character and word spaces watermarking. Proceedings of the IEEE International Conference on Multimedia and Expo, Volume 2, June 30, 2004, Taipei, China, pp: 955-958.

Zhaoqian, G., G. Fei and S. Cheng, 2012. Implementation of dwt domain-video watermarking fast algorithm in blackfin dsp. Proceedings of the International Conference on Mechanical Engineering and Technology, November 24-25, 2011, London, UK., pp: 773-778.

Zou, D.K. and Y.Q. Shi, 2005. Formatted text document data hiding robust to printing, copying and scanning. Proceedings of the IEEE International Symposium on Circuits and Systems, Volume 5, May 23-26, 2005, Japan, pp: 4971-4974.