Advanced Science and Technology Letters
Vol.28 (NT 2013), pp.25-33
http://dx.doi.org/10.14257/astl.2013.28.06

# Print-scan Resilient Watermarking for the Chinese Text Image

Zhihua Xia, Shufang Wang, Xingming Sun, and Jin Wang

Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044, China
School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China
xia_zhihua@163.com, wangshufang128@126.com, sunnudt@163.com and wangjin@nuist.edu.cn

**Abstract.** Digital text watermarking has been a popular way to discourage illicit reproduction of documents by embedding copyright information into them. Based on the relative heights of characters in the same text line, this study presents a robust watermarking algorithm for the Chinese text image. In the embedding process, the characters are segmented firstly by projecting the image horizontally and then vertically. And the rough segmentation is refined according to the peculiarity of Chinese characters. Then on the basis of character segmentation algorithm, watermark embedding is achieved by shifting characters up or down. In the extracting process, pre-process operations such as the binarization and image deskewing algorithm are done first to reduce the impact caused by print-scan operation. Then the messages are extracted by comparing the relative heights of characters. Experimental results show that the proposed method possesses high extraction accuracy under the tampering of print-scan operation.

## 1 Introduction

Along with the development and widespread of the Internet, text documents could be easily copied and distributed. Thus more and more attention has been paid to illicit access to text images. Digital text watermarking is a technology which prevents copyright piracy by hiding copyright information into digital texts [1-3]. Although the digital text watermarking technologies could be a good copyright protector to the digital texts, the mature techniques of printing and scanning make them invalid when the digital texts are converted to hardcopies and vice versa. Therefore, it is worth proposing print-scan resilient watermarking methods for text images.

There have been a lot of print-scan resilient watermarking researches devoted to images. Hernandez *et al.* proposed image watermarking methods in the frequency domain (Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT)) [4-6]. Although they can resist print-scan attack effectively, they cannot be adopted directly into text images. Because characters in text images have clean cuts between the foreground and background areas; a slight

ISSN: 2287-1233 ASTL

change of data in the transformation domain would lead to severe distortion of images. In order to verify this, in the experiment part, some text images are tested with the transform domain method but it can hardly succeed.

Topkara *et al.* put forward the text image watermarking method based on natural language processing which generally embeds watermark by replacing the equivalent information or voice transformation [7-9]. They are suitable for text images and own strong robustness to print-scan attack, but it is very complex to carry out natural language processing (such as participle processing, polyphone recognition, and syntax and synonyms analysis). In addition, methods that by modulating the structure of Chinese characters [10, 11] and by modifying the strokes of characters [12, 13] have better visual effect but they will lose effectiveness if the content of the document is changed.

Brassil *et al.* presented the line shift coding and word shift coding for English text images [14-16]. They embed a data by slightly shifting a block of characters vertically or horizontally. Although they have low embedding capacity and need original images when extracting messages, their idea inspires us to design the watermarking method for Chinese text images.

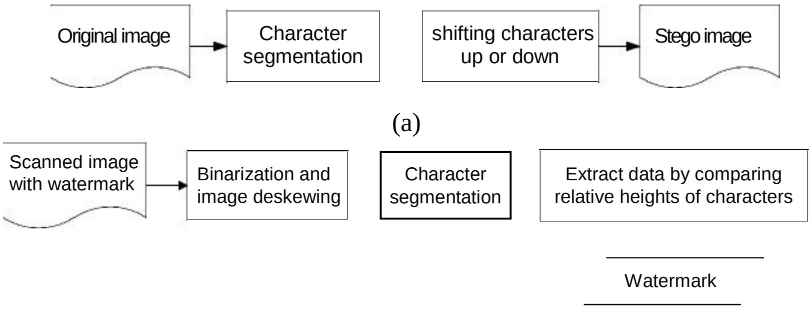
In this study, a new print-scan resilient watermarking method is proposed for Chinese text images. It achieves watermark embedding by shifting characters up or down with modulating relative heights of characters in the same line. And the extracting is the inverse process of embedding.

In section 2, we depict the framework of watermarking algorithms including the embedding and extracting process. Section 3 introduces the character segmentation algorithm in detail, which is a key step of the proposed watermark methods. The deskewing algorithm is described in section 4. The experimental results are showed in section 5. In the end, we make a conclusion in section 6.

## 2 Watermark Embedding and Extracting Process

The main thought of this algorithm is to shift characters and make their height upper or lower than the reference line. The height of a character is measured by its position in the vertical direction and its value is equal to the mean of the character’s upper and lower boundaries.

The framework of the watermarking method is illustrated in Fig. 1. In the embedding process, the Chinese characters are segmented firstly with the character segmentation algorithm which is depicted detailedly in section 3. Then messages are embedded by shifting the characters up or down. In the extracting process, in order to minimize noises and deskewing caused by print-scan operation, the scanned text image is firstly processed by binarization and deskewing algorithm. Then characters are segmented with the proposed character segmentation algorithm. At last the messages are extracted by comparing the relative heights of characters.


Original image
Character segmentation
shifting characters up or down
Stego image
Scanned image with watermark
Binarization and image deskewing
Character segmentation
Extract data by comparing relative heights of characters
Watermark

(a)

(a)

**Fig. 1.** Framework of proposed methods, (a) Watermark embedding, (b) Watermark extracting

### 2.1 Embedding Procedure

According to the result of the character segmentation, the bounding box of each character could be determined and the height of each character can be obtained. Subsequently, in each text line the reference line shown in Fig. 2 can be located and its value is the average height of all valid characters in the same line.

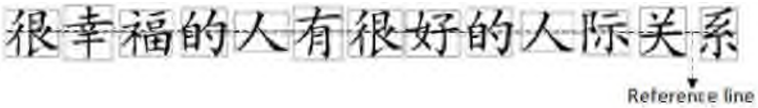

很幸福的人有很好的人际关系
Reference line

**Fig. 2.** Reference line of a text line

In order to ensure the Imperceptibility, the punctuations and the characters whose heights have big gaps with the reference line cannot be used to embed watermark. In addition, with the purpose of stronger robustness, we only embed the watermark in the text lines that have 10 valid characters at least. Finally, for embeddable text lines, we shift valid characters up or down to make their heights equal to the reference line.

With the similar idea of the first algorithm, characters in the same line are divided equally into 2 or 4 groups. If bit ‘0’ is embedded, characters in the former group are shifted with their heights higher (lower) than the reference line and the latter group remains unmoved; otherwise if ‘1’ is embedded, the latter group is shifted and the former group is kept unchanged.

According to the idea of relative heights of characters, the larger the relative height, the stronger robustness the watermark will acquire; however, if the relative height is too large, it would cause severe visual distortion. In order to balance the robustness and visual perceptibility of the watermark, the shifting pattern is made like a shape of a letter ‘n’. For example, characters in Fig. 3 are denoted as $a_i$ (i=1, 2, . . . , 5). The character $a_3$ is shifted 3 pixels up; $a_2$ and $a_4$ 2 pixels up and $a_1$ and $a_5$ 1 pixel up.

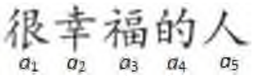


**Fig. 3.** Example of shifting pattern

### 2.2 Extracting Procedure

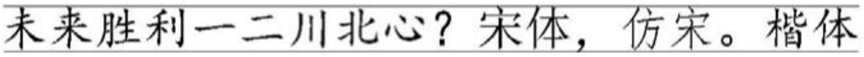
The extracting method is the inverse process of the embedding phase. On the basis of the character segmentation operation, the bounding box of each character is got. Then divide the characters into 2 or 4 groups. According to the embedding phase, we can see the sum of the heights of characters in different groups indicating their difference after watermarking. So the hiding dates could be extracted just by computing and comparing the sum of the heights in different groups.

## 3 Character Segmentation

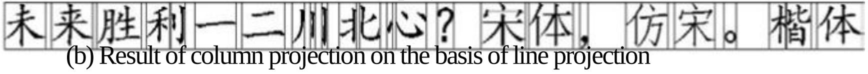
As described above, the embedding method is realized by shifting the characters up or down. So, precise character segmentation is very important to the watermarking algorithm. In this study, on the basis of connected domains, the Chinese character segmentation algorithm is mainly composed of the segmentation of connected domains and the merging of the connected domains. The particular process is in the following steps.

### 3.1 The Segmentation of Connected Domains

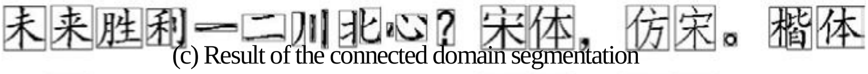
Firstly, the results of the line segmentation of a text image are got by projecting the image horizontally shown in Fig. 4 (a). On the basis of line segmentation, by projecting each text line vertically, the rough segmentation of each connected domains is obtained shown in Fig. 4 (b). According to the rough connected domain, the accurate left and right boundaries are gained. And then by projecting the each connected domain horizontally again, the minimum and maximum ordinate point can be located, which is the exact upper and lower bounds of the connected domain. So, the exact segmentation of connected domains is received shown in Fig. 4 (c).



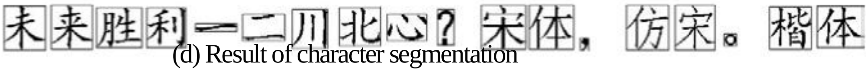
(a) Result of line projection



(b) Result of column projection on the basis of line projection



(c) Result of the connected domain segmentation



(d) Result of character segmentation

**Fig. 4.** Procedure of connected domain segmentation

### 3.2 The Merging of Different Kinds of Connected Domains

As we all know, most of the Chinese characters have a characteristic of square shape. And the aspect ratio *R* of majority characters' bounding boxes is almost equal to 1. So, the merging operation is achieved according to aspect ratio *R* of the connected domains. The concrete segmenting process is as follows:

*Step 1.* Compute $R_1$ of the selected connected domain $Q_1$;

*Step 2.* Get $Q_2$ by merging $Q_1$ and its adjacent connected domain and compute $R_2$;

*Step 3.* Compare $R_1$ and $R_2$ and the connected domain whose *R* is closer to 1 would be segmented into a separate Chinese character;

*Step 4.* Reselect the next connected domain and goto step 1.

The result of the character segmentation is shown in Fig. 4(d). During the merging there are the following conditions need to be considered.

**(1) The connected domains of L-R structure character.** Many L-R structure Chinese characters are segmented into 2 separate parts according to the domain segmentation such as '*!J' in Fig. 4(c). So in this case, the method mentioned above could be used directly.

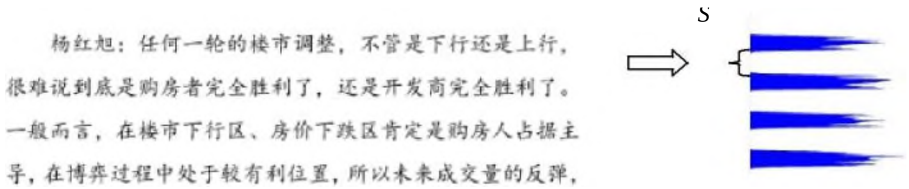
**(2) The punctuations and some characters whose *R* is not close to 1.** The aspect ratio of many punctuations such as "?", " ", characters such as "一" does not meet the feature of *R* . In this case, our solution is to distinguish them before the merging operation. Then they can be identified and segmented immediately by the feature of position coordinates.

**(3) Chinese characters whose internal spacing is bigger than common ones.** The most typical example of this case is the 'JI '. In this case, they just be identified and segmented according to an empirical threshold.

## 4 Image Deskewing Algorithm

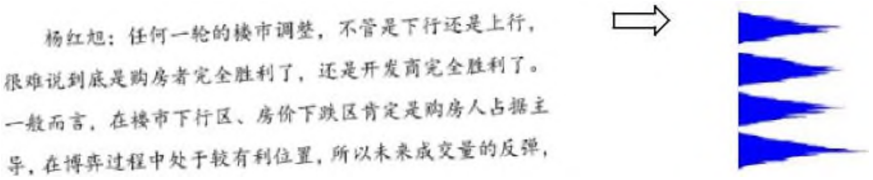
Print-scan operation would inevitably introduce image tilt and small tilt could lead to wrong detection especially for the method based on relative height of characters. So it is particularly important to execute the image deskewing algorithm before the detection operation.

When projecting the image with different inclination horizontally, the sum of the inter-line blank spacing *S* presents different results shown in Fig. 5. Experiments show that the smaller the incline angle of the image, the bigger the value of *S*. Based on this phenomenon, we propose an image deskewing algorithm with the sum of the inter-line blank spacing shown in Alg. 1.



杨红旭：任何一轮的楼市调整，不管是下行还是上行，
很难说到底是购房者完全胜利了，还是开发商完全胜利了。
一般而言，在楼市下行区、房价下跌区肯定是购房人占据主
导，在博弈过程中处于较有利位置，所以未来成交量的反弹，

(a) Horizontal projection of correct image



杨红旭：任何一轮的楼市调整，不管是下行还是上行，
很难说到底是购房者完全胜利了，还是开发商完全胜利了。
一般而言，在楼市下行区、房价下跌区肯定是购房人占据主
导，在博弈过程中处于较有利位置，所以未来成交量的反弹，

(b) Horizontal projection of tilt image

**Fig. 5.** The horizontal projection of the correct image and tilt image

**Algorithm 1** Procedure of deskewing algorithm

```
Require: I: the Chinese text image, $S_r$ ($S_l$): the S after rotating I clockwise (anti-clockwise), $S_a$ ($S_b$): 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 with angle A in correct direction; then calculate $S_a$;
      while $S_a > S_b$
        $S_b = S_a$ ;
          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
```

In order to evaluate the proposed deskewing algorithm, 10 different text images are tested with it. The images are rotated 0.4° clockwise and anticlockwise respectively and then their tilt angles are calculated by using the algorithm. The result is presented in Tab. 1 and it indicates that the scope of the error is within 0.00547°.

**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 |

## 5 Experiment Results

### 5.1 Comparisons of Different Watermarking Methods

We compared the advantages and disadvantages of different kinds of the watermarking methods in Tab. 2. Methods base on transform domain have strong robustness but they are not suitable for text images[6, 17]. Brassil *et al.* proposed three data hiding methods; they have good visual perceptibility but low embedding capacity[14]. Although method based on natural language processing possesses good imperceptibility and has strong robustness but it is very complicated to carry out the natural language process. The proposed methods are suitable for Chinese text images and the process of embedding and extracting is simple and efficient to implement. Besides, they offer robustness to print-scan attack.

**Table 2.** Comparison among different methods

| Methods | Embedding capacity | Robustness | Weakness |
|---|---|---|---|
| Transform domain | — | Print-scan | Suitable for natural images |
| Line/word shifting | 0.5bit/line | Little to print-scan | Low embedding capacity, need original image when detecting |
| Based on natural language processing | — | Print-scan | Need complex pretreatment |
| Proposed | 1.5bit/line | Print-scan | Little robustness to copy |

### 5.2 Algorithm Tests

In order to verify the performance of the proposed method, the extraction accuracy and the embedding capacity of four methods was compared. The first one embeds data by adjusting the coefficient of the 3-level DWT (Discrete Wavelet Transform)[6]. And the two of the rest are presented in our study. The last one is the combination of the two proposed methods.

We used a HP LaserJet M1536dnf MFP to print and scan the hardcopy documents with resolution of 600 dpi. A 60-page Chinese text document was used to test the methods respectively. The documents were formatted in different fonts (“Song typeface”, “Regular script”, “Imitation song”) and sizes (16pt and 15pt). The extraction result is shown in Tab. 3. The results apparently indicate that the method with transform domain is not suitable for the text image while the proposed method possess strong robustness to the print-scan attack. The result of embedding capacity is presented in Tab. 4. From the table, we can see the embedding capacity is enough to hide copyright information and it lies on the font size of the text to some extent.

**Table 3.** The extraction accuracy result by using different methods

| Method | Font (%) | | |
|---|---|---|---|
| | Song typeface | Regular script | Imitation song |
| Modulating character spaces | 99.67 | 99.61 | 100 |
| Shifting characters up or down | 99 | 100 | 100 |
| Combination of above two methods | 99.71 | 100 | 100 |
| DWT | 48.92 | 51.67 | 51.5 |

**Table 4.** Comparisons of the embedding capacity of different methods

| Watermarking method | Pages | Average document page sizes(KB) | Average embedding capacity(byte) |
|---|---|---|---|
| Modulating character spaces | 60 | 30 | 30 |
| Shift characters up or down | 60 | 26 | 30 |
| Combination of above two methods | 60 | 26 | 30 |
| DWT | 60 | 27 | — |

## 6 Conclusions and Future Works

In this paper, we propose a new watermarking algorithm which can be employed in Chinese text images. On account of the relative height of characters, it completes data hiding by shifting characters with their height higher or lower than the reference line. In the extracting process, pretreatment such as binarization, image deskewing and character segmentation were adopted firstly. Then, messages are extracted by comparing the sum of relative heights of characters. Experimental results indicate that it hold much better robustness against print-scan attack and have good visual perceptibility.

In current digital age, the images are likely to be printed and scanned for many times. So, it is necessary to put forward the watermarking method robust to multiple printing and scanning. In addition, we plan to improve the watermarking algorithm and make it robust to copying attack.

**Acknowledgements.** This work is supported by the NSFC (61232016, 61103141, 61070195, 61070196, 61173141, 61173142, 61173136, 61103215, 61373132, 61373133, and 61073191), National Basic Research Program 973 (2011CB311808), 2011GK2009, GYHY201206033, 201301030, 2013DFG12860, SBC201310569, Research Start-Up fund of NUIST (20110428), and PAPD fund.

## References

1. F. Peng, et al., "Adaptive reversible data hiding scheme based on integer transform," Signal Processing, vol. 92, no. 1, pp. 54-62, 2012.
2. K. Ntirogiannis, et al., "A combined approach for the binarization of handwritten document images," Pattern Recognition Letters, no. 2012.
3. P.-J. Chiang, et al., "Printer and scanner forensics," Signal Processing Magazine, IEEE, vol. 26, no. 2, pp. 72-83, 2009.
4. M. Cedillo-Hernandez, et al., "Robust digital image watermarking using interest points and DFT domain," in Telecommunications and Signal Processing (TSP), 2012 35th International Conference on, 2012, pp. 715-719.
5. Z. Haitao, et al., "Low luminance smooth blocks based watermarking scheme in DCT domain," in Communications, Circuits and Systems Proceedings, 2006 International Conference on, 2006, pp. 19-23.
6. W.-H. Lin, et al., "An efficient watermarking method based on significant difference of wavelet coefficient quantization," Multimedia, IEEE Transactions on, vol. 10, no. 5, pp. 746-757, 2008.
7. U. Topkara, et al., "The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions," in Proceedings of the 8th workshop on Multimedia and security, 2006, pp. 164-174.
8. H. M. Meral, et al., "Natural language watermarking via morphosyntactic alterations," Computer Speech & Language, vol. 23, no. 1, pp. 107-125, 2009.
9. M. LUO and Y.-x. ZHANG, "Watermarking Chinese Text Document Based on Structure and Semantics of Chinese Characters," Computer Knowledge and Technology, vol. 34, no. p. 058, 2011.
10. C.-C. C. Zhi-Hui Wang, Chia-Chen Lin, Ming-Chu LI, "A reversible information hiding scheme using left-right and up-down chinese character representation," The Journal of Systems and Software, vol. 82, no. 2009.
11. X. Sun, et al., "Component-based digital watermarking of Chinese texts," presented at the 3rd international conference on Information security New York, USA, 2004.
12. L. TAN, et al., "Print-Scan Resilient Text Image Watermarking Based on Stroke Direction Modulation for Chinese Document Authentication," Radioengineering, vol. 21, no. 1, p. 171, 2012.
13. X. Tang and L. Wang, "Text watermarking algorithm based on the stroke of Chinese characters," in Multimedia Technology (ICMT), 2011 International Conference on, 2011, pp. 794-796.
14. J. T. Brassil, et al., "Copyright Protection for the Electronic Distribution of Text Documents," Proceedings of the IEEE, vol. 87, no. 7, pp. 1181-1196, 1999.
15. J. T. Brassil, et al., "Electronic marking and identification techniques to discourage document copying," Selected Areas in Communications, IEEE Journal on, vol. 13, no. 8, pp. 1495-1504, 1995.
16. N. F. M. a. A. M. L. Steven H. Low, "Document identification for copyright protection using centroid detection," IEEE TRANSACTIONS ON COMMUNICATIONS, vol. 46, no. NO.3, 1998.
17. A. Al-Haj, "Combined DWT-DCT digital image watermarking," Journal of computer science, vol. 3, no. 9, pp. 740-746, 2007.