Information Hiding and Extracting In Printed Chinese Documents by Missing Feature Method

Hui Lin[1,a], Jie Lin[2,b], Bo Fu[2,c]

[1]Leshan Vocational and Technical College, Sichuan, 614000, China

[2] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

[a]lslinhui@126.com, [b]linjie@uestc.edu.cn, [c]bxfu@163.com

Keywords: Information hiding; information extracting; watermark

Abstract. In this paper, we proposed a new method for hiding and extracting information in printed chinese documents. This new method embeds information into chinese documents by varying the typefaces of chinese words. In addition, an information extracting method based on missing feature concepts is proposed to handle the partial occlusion, generated by printing and scanning operations. The new improved approach has been evaluated on numerous printed documents and demonstrated improved performance over other systems.

Introduction

Information hiding has recently become important in a number of application areas, for example the information security and concealing data in printed papers. The most typical technique for the digital information hiding is digital watermarking [1].

Digital watermarking techniques aim to embed or hide a data called “watermark” into host digital media (e. g: audios, images and printed papers) such that the watermarked media is not easily observed by human eyes. Because watermark is unchangeable in the new copies of that watermarked media, the watermark inside can later be extracted.

This paper focuses the research on the information hiding in the printed documents. For this research topic, there exits one vital problem needed to be resolved. The digital watermark embedded in the documents undergoes a printing and scanning process, before it can be extracted. These two processes may introduce some random noises into the images, consequently impairing the performance of extracting system. Several researchers try to solve the problem, for example [2-6]. P. Bas et al[2] proposed the technique of geometrically invariant watermarking using feature points. S. V. Voloshynovkiy et al[3] introduced a scheme for watermarking in the printing channel using Gel’fand-Pinsker construction. R. Lhawchaiyapurk et al [4] and S. Ibrahim et al [5] presented the schemes capable of removing printing and scanning distortions by assigning different regulation factors to different partitions and removing the distortion before extracting. Recently, K. Thongkor[6] proposed a pre-processing step to the scanned watermarked images in order to remove the printed information as much as possible before performing the watermark extraction process.

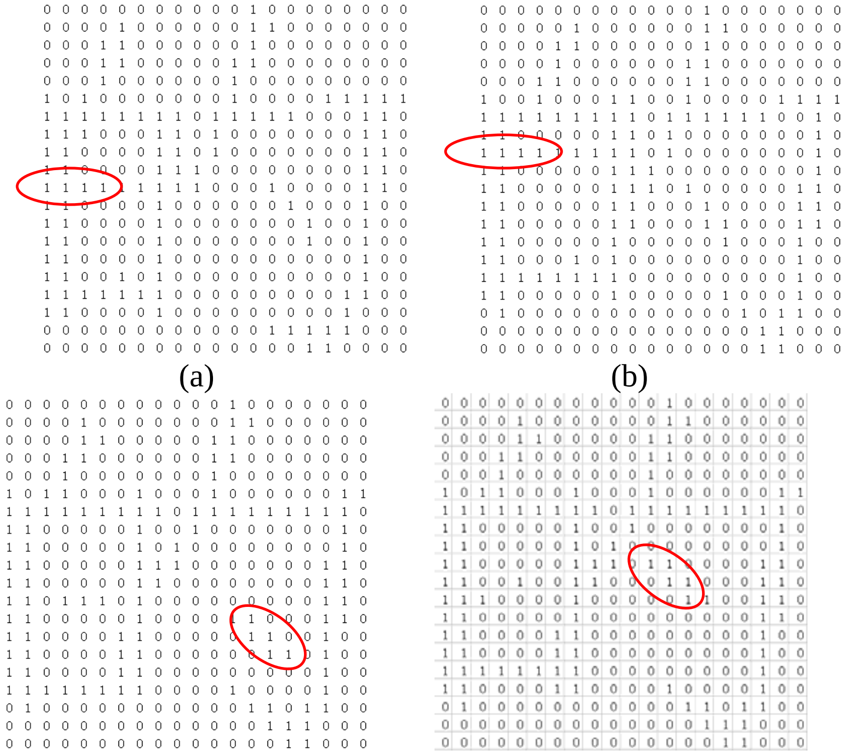
In this paper, we also focuses on the problem how to resolve the accuracy extracting information from the printed chinese documents suffering from unknown partial image occlusions and distortions. We proposed a new method to embed information into a chinese documents by varying the typefaces of many chinese words under that documents. Moreover, a new extracting method was presented for resolving the partial occlusion problem. This new method based on the basic Cosine similarity calculation, enhanced by a missing feature model proposed previous in our research for robust face recognition. This missing feature model ignores severely mismatched local features and focuses the recognition mainly on the reliable local features. It thereby improves the robustness while assuming no prior information about the corruption.

Information hiding into the printed documents

Published by Atlantis Press, Paris, France.

In this paper, we focus the research on the information hiding and extracting in printed chinese documents. Information was hided into the words over the documents, by small varying in the pixels of words. These varying of pixels can be handled by reassigning the locations of some character components or the modification in thick levels of the character components. Different pixels expressing of one word indicates different patterns and typefaces, and each typefaces may represents an information.

Let us discuss the scheme in detail through an example. Assume the information can be coded with an array of quaternary numbers. There need a four different typefaces for one word to present the "0", "1", "2", "3", respectively. Only the word commonly appearing in texts are selected to embed information rather than every word among the fonts. Since how to change the pixels of the word character components is not our focus in this paper, we randomly change the locations of one or two character components in one word to construct four typefaces for the word. Fig.1 show a binary image example of four typefaces for a chinese word((a) representing "0", (b) representing"1",(c) representing "2" and (d) representing "3". Fig.1(a) moves down horizontal, in contrast Fig.1(b), as well as Fig.1(c) and (d) move the location of the point. All the modifications are enclosed within circles over the images shown in Fig.1.



```
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 1 1
1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 0 1 1 0
1 1 1 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1 0
1 1 1 1 1 1 1 1 1 0 0 0 1 0 0 0 0 1 1 0
1 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0
1 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0
1 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0
1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0
1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0
1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0
1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0
```

(a)

```
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0
0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
1 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 1 1 1 1
1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 1 0
1 1 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 0
1 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0 1 0
1 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0
1 1 0 0 0 0 0 1 1 1 0 1 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 1 0
1 1 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0
1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0
1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0
1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0
0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0
```

(b)

```
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0
0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
1 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1
1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0
1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0
1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0
1 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 0
1 1 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 1 0
1 1 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 1 0 0
1 1 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 1 0 0
1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0
1 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 1 0 0
0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0
```

(c)

```
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0
0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
1 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1
1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0
1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0
1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0
1 1 0 0 0 0 0 1 1 1 0 1 1 0 0 0 0 1 1 0
1 1 0 0 1 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0
1 1 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0 1 1 0
1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0
1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0
1 1 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0
0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0
```

(d)

Fig.1 an example of four typefaces for a chinese word, representing "0":(a), "1":(b), "2":(c), "3":(d) in quaternary.

When the number of selected words is enough big, few pages, even one of document containing four word typefaces, will accommodate embedded data with sufficient space.

Information extracting

The information extracting process referring to our embedding scheme can be illustrated as an image pattern recognition processing, stepped as follow:

Step.1 Word cutting in an image of a document.

Step.2 Locating these selected words, which have multi-typefaces; saving the image blocks of these words, and recognizing and saving the labels of each blocks.

Step.3 Classifying image blocks one by one into one of four classes, according to the relation

Published by Atlantis Press, Paris, France.

labels of that image block.

Step.4 Combining the sorting results to encode the data embedded in the document images.

The Step.1 and Step.2 can be easily implemented with a typical word cutting and chinese character recognition system[7]. This paper focus on how to recognize word image blocks. While extracting the information from the printed documents, the recognition systems normally suffer from image distortion which is made by either printer's quality or scanner's precision. Hence images of each word involve unexpected mismatch. The recognition method for information extracting must be robust to those partial distortion and occlusion. In order to address this problem, we present a new approach in follow section 2.2 to recognize the typefaces of each word image block, subject to unknown partial distortion and occlusion. This approach ignores severely mismatched local features and focusing the recognition mainly on the reliable local features, thereby improves the robustness while assuming no prior information about the corruptions. We call this approach as missing feature method. It has previously been applied successfully in robust face recognition[8].

The word Feature extracting from printed documents

All the image blocks in our system are transformed to a binary image before recognition. The location and trick of character components distinguish the different typefaces of one word. Therefore the position information of the whiter pixels about the image blocks could be selected as the features for representing different typefaces. Note the values of $i-th$ pixel is $x_i \in \{1,0\}$. The features of a word is X = (x_1, x_2, ..., x_M). M is the number of pixels in a block image.

Recognition the typeface by Missing Feature Method

We briefly describe the missing feature method in this section. The method is a transformation of the cosine similarity. It can be robust to partial image occlusions by focusing the similarity calculation only on reliable feature subset, thereby improving the partial mismatch robustness, while assuming no prior information about the partial corruption on images. Assume that the training and testing word block images all can be divided into N -local images and represented by an N -part feature vectors. The cosine similarity for comparing a testing vector X = (x_1, x_2, ..., x_N) and a reference vector Y =(y_1, y_2, ..., y_N), each being expressed as N local vectors, can be written as follows:

$$S(X,Y)=\frac{X\cdot Y}{\|X\|\|Y\|}=\sum_{n=1}^{N}\frac{x_n\cdot y_n}{\|x_n\|\|y_n\|}\cdot\frac{\|x_n\|\|y_n\|}{\|X\|\|Y\|}=\sum_{n=1}^{N}S(x_n,y_n)\omega_n$$

(1)

Where *S(a, b) = ab/* ||*a*||||*b*|| is the inner product between vectors a and b normalized by their respective norms. Equation (1) shows that the overall cosine similarity equals the sum of the local cosine similarities $S(x_n, y_n)$ weighted by ω_n, which are the comparisons of the individual local 'energies' $\|x_n\|\|y_n\|$ to the overall 'energies' ||X||||Y||.

Assume that some of the local x_n are corrupted but knowledge about the number and identities of the corrupted x_n is not available. As ω_n is a function of the overall energy ||X||, it will be adversely affected by any local corruption within X. To remove this coupling, we can assume an equal ω_n for the entire local and simplify the cosine similarity equation (1) to [8]

$$S(X,Y)\approx\sum_{n=1}^{N}S(x_n,y_n)$$

(2)

Based on (2), we may reduce the effect of local occlusion on similarity calculation by estimating optimal overall similarity $S(X_{\hat{I}_Q}, Y_{\hat{I}_Q})$, where $\hat{I}_Q$ = ($\hat{n}_1$, $\hat{n}_2$, ..., $\hat{n}_Q$) defines the indexes of the Q optimal local similarities $S(x_n, y_n)$, without assuming prior knowledge about $\hat{I}_Q$. However searing for the optimal set of reliable features to calculate the similarity can be computationally expensive. This problem can be address by transforming the simplified cosine dissimilarity to an exponential form, which is proportional to the simplified cosine similarly, as follow:

$$G(X_{\hat{I}_Q}, Y_{\hat{I}_Q}) = M^{S(X_{\hat{I}_Q}, Y_{\hat{I}_Q})} = M^{\mathrm{S}(\mathrm{x}_{\hat{\mathrm{n}}_1}, y_{\hat{n}_1})} M^{\mathrm{S}(\mathrm{x}_{\hat{\mathrm{n}}2}, y_{\hat{n}_2})} \ldots M^{\mathrm{S}(\mathrm{x}_{\hat{\mathrm{n}}\mathrm{Q}}, y_{\hat{n}1Q})}$$

(3)

Where $M > 1$ is a positive number. Against the equation (3), now the estimation of $S(X_{\hat{I}_Q}, Y_{\hat{I}_Q})$ can be replaced with estimating $G(X_{\hat{I}_Q}, Y_{\hat{I}_Q})$, where $G(X_{\hat{I}_Q}, Y_{\hat{I}_Q})$ can be approximated by summing $G(X_{I_Q}, Y_{I_Q})$ over all Q-sized local feature subsets, assuming that the optimal $G(X_{\hat{I}_Q}, Y_{\hat{I}_Q})$ will dominate the sum [8]. So we have

$$G(X_{\hat{I}_Q}, Y_{\hat{I}_Q}) \propto \sum_{I_Q \subset \{1,2,\ldots,N\}} G(X_{I_Q}, Y_{I_Q}) = \sum_{n_1 n_2 \ldots n_Q} M^{S(x_{n_1}, y_{n_1})} M^{S(x_{n_2}, y_{n_2})} \ldots M^{S(x_{nQ}, y_{nQ})}$$

(4)

The optimal similarities for each classes ω and input face features X can thus be obtained by choosing appropriate Q for maximizing equation (5):

$$MS(\omega, X) = \sum_{Y \in \omega} \max_{1 \le Q \le N} F(X_{I_Q}, Y_{I_Q})$$

(5)

Where by definition

$$F(X_{I_Q}, Y_{I_Q}) = \frac{G(X_{I_Q}, Y_{I_Q})}{\sum_{\omega'} \sum_{Y' \in \omega'} G(X_{I_Q}, Y'_{I_Q})}$$

(6)

Equation (6) is defined in a way similar to the class posterior probability. In our task, the input unknown block image (features) X can be classified into one of classes based on the scores of $MS(\omega, X)$.

Experiments

We evaluated the performance of our method by comparing it with the traditional Cosine similarity and Euclidean distance on the same training and testing data. 100 words in chinese font ware selected to be varied for embedding data. Each word has four different typefaces, as Fig.1 representing “0”, “1”, “2”, “3” . The data to be hided into the documents, is coded to an array of quaternary numbers in advance and then embedded into the documents. The documents have three pages and contain 250 selected words, which show different typefaces in the printed pages according to the codes of data. In the experiment, we scan images of the documents with 2550×3670 pixels three times to product three experiments. And the document images captured on the PC screen also is treated as the fourth experiment data. We performed the word cutting and chinese word recognition with the method[8]. All of the 250 word block images in each experiment ware resized 100×100 pixels and covert to binary images as well as the word templates.

We compared our system with two other systems: (1) Cosine-based which use the traditional cosine similarity to calculate the similarity in recognition, (2) a Distance-based system, which classify the image to one of four classes by Euclidean-distance. We first divided each word block images into 16 non-overlapping local images, and then applied the feature extracting method described in section 2.1 on each local image. And M = 80 was used in our proposed system. Table. 1 shows the recognition accuracy rates by the various systems on the capturing testing image and three scanning images. The rates are averaged over the three experiments, each involving overall 250 words, as described above.

Table 1 indicates that all the systems achieved similar recognition accuracy on the captured images. The proposed new, however, outperformed the other systems on the scanning images. The new method achieved the improvement without having assumed any prior information about the number and identities of the distorted features. The other systems used the full set of features and their performance was thus impaired by the distorted features.

Table 1. Recognition accuracy (%) on capturing and scanning document images for the proposed new method, compared to a Cosine-based system, and Distance-based system.

Image type	system		
	Our	Cosine-based	Distance-based
Capturing	99.42	99.37	99.42
Scanning	97.53	81.75	83.56

Conclusions

In this paper, we presented an effective approach to embed and extract information in a printed chinese document. This new method improves the robustness of extracting system to the document image partial occlusion, by a missing feature idea. The comparing experiments under the same databases have shown enhance performance.

Acknowledgements

This paper is supported by the National High Technology Research and Development Program of China (No: 61202084), the Program of China Academy of Engineering Physics (No: ZYGX2010J076), (No: ZYGX2010J070).

References

[1] J. Kim, K. Kim, J. Choi, Technologies for Online Issuing Service of Documents, International Conference on Web Information System Engineering, 2004, pp. 169-180.

[2] P. Bas, J. M. Chassery, B. Macq, Geometrically invariant watermarking using feature points, IEEE Trans. Image Process., vol. 11(9), 2002, pp. 1014-1028.

[3] S. V. Voloshynovskiy, O. Koval, F. Deguillaume, T. Pun, Visual communications with side information via distributed printing channels: extended multimedia and security perspectives, International Conference on Society for Optical Engineering, 2004, pp. 428-445.

[4] R. Lhawchaiyapurk, N. Mettripun, T. Amomraksa, Scaling methods for printed and scanned document resilience, International Conference on Embedded Systems and Intelligent Technology, 2010, pp.64-67.

[5] S. Ibrahim, M. Afrakhteh, M. Salleh, Adaptive Watermarking for Printed Document Authentication, International Conference on Computer Sciences and Convergence Information Technology, 2010, pp. 611-614.

[6] K. Thongkor, T. Amornraksa, Improved watermark extraction for printed and scanned watermarked document, International Symposium on Intelligent Signal Processing and Communication Systems, 2011, pp.7-9.

[7] Microsoft Ocr soft on http: // microsoft- ocr- download. fyxm. net.

[8] J. Lin, J. Ming, D. Crookes, Robust Face Recognition with Partially Occluded, Illumination Variation and Limited Data by Optimal Feature Selection, Computer Vision IET, 2011,Vol(5), pp.23-32.

Published by Atlantis Press, Paris, France.