

Research on Fragments Reassembly Based on Feature of Chinese Character and Template Matching

Gui-Sen Xu^{1,3}, Yuan-Biao Zhang^{2,3}, Yi Lin⁴, Yu-Jian Lin⁴ & Xin-Guang Lv^{2,3}

¹ Electrical and Information School, Jinan University, Zhuhai, China

² Packaging Engineering Institute, Jinan University, Zhuhai, China

³ Key Laboratory of Product Packaging and Logistics of Guangdong Higher Education Institutes, Jinan University, Zhuhai 519070, China

⁴ International Business School, Jinan University, Zhuhai, China

Correspondence: Yuan-Biao Zhang, Packaging Engineering Institute, Jinan University, Zhuhai 519070, China.
E-mail: zybt@jnu.edu.cn

Received: June 26, 2014

Accepted: July 7, 2014

Online Published: July 29, 2014

doi:10.5539/cis.v7n3p92

URL: <http://dx.doi.org/10.5539/cis.v7n3p92>

Abstract

The technology of fragments reassembly is widely employed in many scientific fields, such as judicial evidence recovery, restoration of historic documents, accessing to military intelligence and so on, which is based on computer vision and pattern recognition. In this paper, an efficient method for Chinese fragments reassembly is presented. The proposed reassembly method is based on the feature of line spacing and Chinese characters' feature that the same font has the same height. Considered the feature of English characters that English letters are connected components, this paper proposes a model for English fragments reassembly, which is based on template matching. Using the proposed methods on digitally scanned images of actual fragments of paper image prints has verified the robustness and reliability of two models.

Keywords: fragments reassembly, the feature of Chinese characters' line spacing, template matching, connectivity of English letters

1. Introduction

Fragments reassembly arises in many scientific fields, such as judicial evidence recovery, restoration of historic documents, accessing to military intelligence and so on. The manual execution of fragments reassembly is very difficult, as it requires great amount of time, skill and effort. Specifically, if the number of fragments is very large, reassemble manually will be impossible to be accomplished in a short time. Therefore, the automation of such a work is very important and can lead to faster, more efficient, painting reassembly and to a significant reduction in the human effort involved.

Fragments reassembly is an application technology based on computer vision and pattern recognition, which can be divided into two parts. The first part is to access to information of fragments by image preprocessing. The other part is to reassemble fragments based on accessed information. Conventional methods utilized corner feature of edge, outline feature, area feature and so on to reassemble fragments (Jia et al., 2005; Wolfson, 1990; Hori, 1999; Kong, 2001; Li, 1995; Gao, 2007). These methods are based on geometric characteristics of fragments. Therefore, they can't recover fragments with similar shape. In this field, some scholars have made related researches. Otsu N. presented Otsu algorithm based on the principle of least square method, which solved the problem of searching optimal threshold by calculating the maximum variance between grey classes (Otsu, 1979). Luo Z. solved the problem of feature extraction of Chinese document, by extracting line spacing and the height of characters (Luo, 2012). Gao et al., solved the problem of labeling 8-connected region in the model of English fragments reassembly, by combining the advantages of the mark line based method and region growing method (Gao, 2007).

Considering those methods, which are based on geometric characteristics, can't recover fragments with similar shape, this paper combined the advantages of above methods, and proposes subregional scoring method to optimize the extraction of Chinese line spacing. Because line spacing of English is not obvious, template matching method is proposed to reassemble English fragments, which is based on 8-connected region of English

letters. Applied the proposed methods on digitally scanned images of actual fragments of paper image prints has verified the robustness and reliability of the two models.

2. Model for Chinese Fragments Reassembly

2.1 Thought of Model

According to the feature that Chinese character in the same line have same height and same line spacing, this paper extracted the feature of line spacing and then set up the model for Chinese fragments reassembly. The process of model is divided into two sub-progress including the classification process and reassembly process based on the similarity of edges of binarized pixel matrixes.

In course of locating line spacing in fragments, firstly, the process of image binarization is completed by Ostu algorithm. The next step is to compare the value of the same pixel row in binary image matrix. If all numbers are 1, means all the pixels are white, this pixel row will be map into 1; otherwise, this pixel row will be map into 0. Finally, binary image matrix is transformed into a column vector with 0 and 1, as shown in Figure 1. This transformation not only greatly reduce the amount of data processing, but fully embodies the feature of the line spacing.

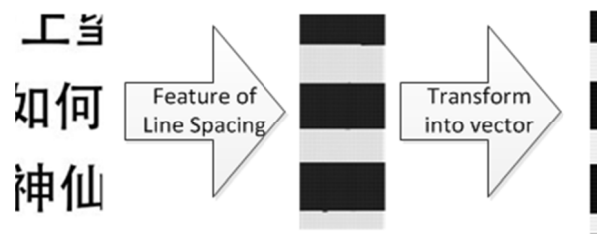


Figure 1. Transformation of binary image

If a paper is divided into two parts, the break edges must have the text line with same height and same line spacing. Therefore, according to this feature of leftmost fragments, all other fragments, which locate on the same line, can be found. However, because the size of fragments is so small, some fragments, from beginning of paragraph and the end of paragraph, contain large blank area, as shown in Figure 2. Another special case is that some fragments' line spacing feature are very similar, but not really matching, as shown in Figure 3. Those two cases caused large errors to the result of classification. In order to solve this problem, subregional scoring method is presented to optimize the extraction of line spacing. The first step is to divide the column vector into more areas. And then this paper compares the similarity between the same areas of two fragments under the appropriate accuracy. Finally, the problem, which compares similarity between two fragments, is transformed into another problem that compares similarity between those areas of two vectors. This method led to a result more accurate.

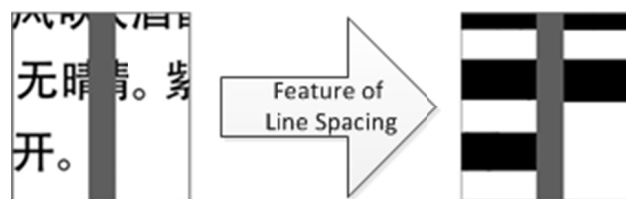


Figure 2. Fragments containing large blank area

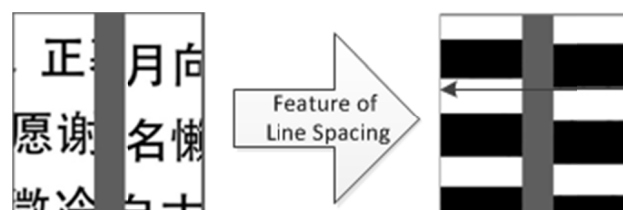


Figure 3. Fragments that very similar but not really matching



Figure 4. Schematic diagrams of subregional scoring method

After using subregional scoring method, the similarity between two fragments in Figure 3 was low, so the total score was also low. For the case in Figure 2, only the area containing a large number of blank has a low score, so the similarity of other areas is high. Therefore, subregional scoring method can ensure that two matched fragments, which containing a large number of blank, can be reassembled together. What's more, two fragments, which are very similar but not really matching, can't be reassembled by mistake. It can be said that this method have made classification result more accurate.

Finally, based on result of classification, the fragments in the same row were reassembled, according to the similarity of edges of binarized pixel matrix. The next operation is to find the correct order of each fragments row. When reassembling fragments, the case that break edge located in blank area is needed to be considered specially. Under this situation, there is no black pixel in the edge, so fragments can't be reassembled according to the similarity of edges of binarized pixel matrix. Therefore, an improved method is presented, which is based on the feature that printed documents have fixed line spacing and characters distance. For instance shown in Figure 6, the line spacing is 28px and the break edge just located in blank area. Because the line spacing above break edge is 21px, we just need to search for a fragments row with 7px line spacing under break edge. If encountering similar situation, fragments can be reassembling by the same method.

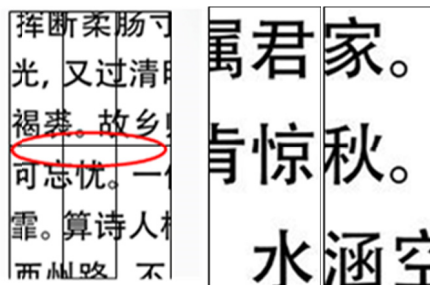


Figure 5. Fragments that the break edge located in blank area

念处。日白工恠不云。恠有旧，起伙期。这有1个归。加泪初有石一肝。
清润潘郎，又是何郎婿。记取钗头新利市。莫将分付东邻子。西塞山边白
鹭飞。散花洲外片帆微。桃花流水鳊鱼肥。主人瞋小。欲向东风先醉倒。
已属君家。且更从容等待他。愿我已无当世望，似君须向古人求。岁寒松
柏肯惊秋。

水涵空，山照市。西汉二疏乡里。新白发，旧黄金。故人恩义深。谁

Figure 6. Special case that the break edge located in blank area

2.2 Steps of Algorithm

Step 1: Comparing binary image matrixes $[A_i]$ of the first i fragments, if the first k row all is 1, then $[A_i](k)=1$; or $[A_i](k)=0$. Finally, $[A_i]$ is transformed into column vector $\{b_i\}^T$.

Step 2: Dividing $\{b_i\}^T$ into a areas, and denoting as $\{b_{i1} \ b_{i2} \ \dots \ b_{ia}\}^T$. Every area contains H/a pixels (H is the

height of fragments, unit: px).

Step 3: Calculating the similarity of the first m subregion in both $\{b_i\}^T$ and $\{b_j\}^T$:

$$Z_{score}(m) = \begin{cases} 100, & \sum_{m=1}^a \sum_{t=1}^{H/a} |\{b_{mi}\}^T(t) - \{b_{mj}\}^T(t)| \leq \delta \\ 0, & \sum_{m=1}^a \sum_{t=1}^{H/a} |\{b_{mi}\}^T(t) - \{b_{mj}\}^T(t)| > \delta \end{cases} \quad (1)$$

m represents the first m subregion in fragments. δ represents fault tolerance, which is the tolerable max numbers of different value between two areas.

Step 4: Calculating total score of two fragments

$$Z_{score} = \sum_{i=1}^a Z_{score}(i) \quad (2)$$

Within the fault tolerance δ , if $Z_{score} \geq 100(a-2)$, the first i fragment and the first j fragment are in the same row, or they are in different row.

Step 5: After classifying all fragments, reassembling fragments belong to the same row in the order, according to the similarity of edges of binarized pixel matrix. If the break edge located in blank area, fragments need to be reassembled according to the feature that printed documents have fixed characters distance.

Step 6: After reassembling each row, the next step is to extract the edge of top and bottom of each row. And then this paper takes each row as a fragment. We recover all row fragments according to the similarity of top and bottom edges. If the break edge located in blank area, fragments need to be reassembled according to the feature that printed documents have fixed line spacing. Finally, all fragments are reassembled correctly.

3. Model for English Fragments Reassembly

Considering English letters are composed of 52 uppercase and lowercase letters and English punctuations. Furthermore, the number of English punctuations is small. Therefore, it's convenient to make English characters templates. So fragments can be reassembled according to those templates.

3.1 Building of English Characters Templates Library

Found out the font type of this English document, and then found all pictures of characters. According to these pictures, the library of character templates can be set up. Finally, the process of character templates binarization was implemented, such as Figure 7.

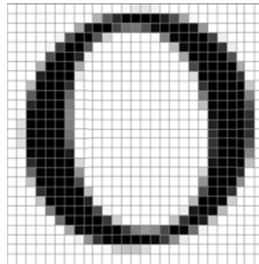


Figure 7. Binary template of letter O

3.2 Extraction of Letters

According to this feature that every English letter is an 8-connected region, each of connected components within fragments is marked in the order. The next operation is to extract the whole letter according to the boundary of connected component. Finally, all letters are extracted in the same method. Figure 8 shows the sample fragments of L and H. The following is the extraction process of letter L and H. The results are shown in Figure 9. Figure 8 and figure 9 are schematic diagrams, which is aimed to better explain the principle of 8-connected region.

Step 1: Marked connected component of L and H in the order, as shown in Figure 8.

Step 2: According to the boundary location of B and L's 8-connected region, extracted letter L and H.



Figure 8. Binary image of L and H

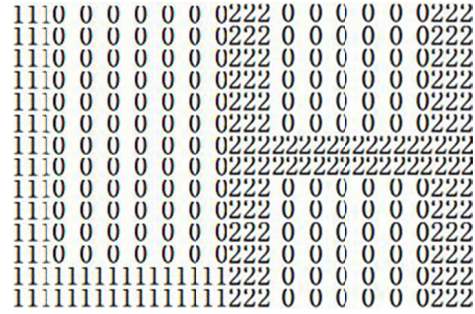


Figure 9. Marked image of L and H

3.3 Detection of Complete Letters

After extracting the letters from fragments, the first step is to transform binary images of letters into column vector $\{M\}^T$, according to the model for Chinese fragments reassembly. This process makes all the information of characters embodied in $\{M\}^T$, which is convenient to compare the characters and templates. In order to detect whether there is a same template as the fragment, the next step is to compare $\{M\}^T$ and templates with the same transformation. If there is the matched template, it shows that the character is complete. Pseudo algorithm of the method is shown as follow:

3.3.1 Initialize the Parameters

$$Sum = 10000000, \text{ number of torn pieces}(TP).$$

3.3.2 Transform those binary images of letters into $\{M\}^T$.

3.3.3 For $count = 1, 2, 3, \dots, TP$

Transform templates into column vectors;

If rows of $\{M\}^T$ is equal to rows of $\{M_{model}\}^T$

$$\{M_{XOR}\}^T = \{M\}^T XOR \{M_{model}\}^T;$$

Add each value of $\{M_{XOR}\}^T$ and assign to Sum_{XOR} .

Else

Add each value of $\{M\}^T$ and assign to Sum_{XOR} .

End if

If $Sum > Sum_{XOR}$

$$Sum = Sum_{XOR}$$

End if

If $Sum = 0$

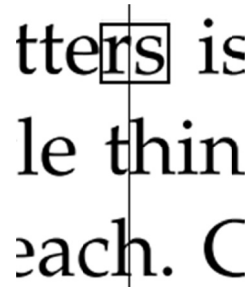
This character is complete.

End if

End for

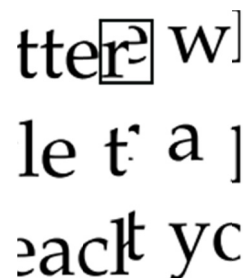
3.4 English Fragments Reassembly Based on Template Matching

Initially, to find leftmost fragments and sort those fragments, according to the feature that the left edge of the leftmost fragments contains a large number of blank areas. The next operation is to search the matching fragments from left to right according to the leftmost fragments. Then selecting a fragment from remaining fragments, and then reassembling these fragments and the left fragments. The next step is to extract the area with width of two characters around break edge in the top line, as shown in Figure 10 and Figure 11, and delete incomplete characters around the edge. Final step is to detect completeness of all characters inside the area around break edge. If there aren't incomplete characters, it means these two fragments are matched, as shown in Figure 10. Otherwise, it means these two fragments aren't matched, as shown in Figure 11. For Figure 10 and Figure 11, the area inside wireframe is the area around break edge.



tters is
le thin
each. C

Figure 10. Complete characters inside area around break edge



tters w
le t a
each yc

Figure 11. Incomplete characters inside area around break edge

In order to reassemble all fragments in the same row, the searching procedure is iteratively applied. Algorithm flowchart of the model is shown in Figure 12.

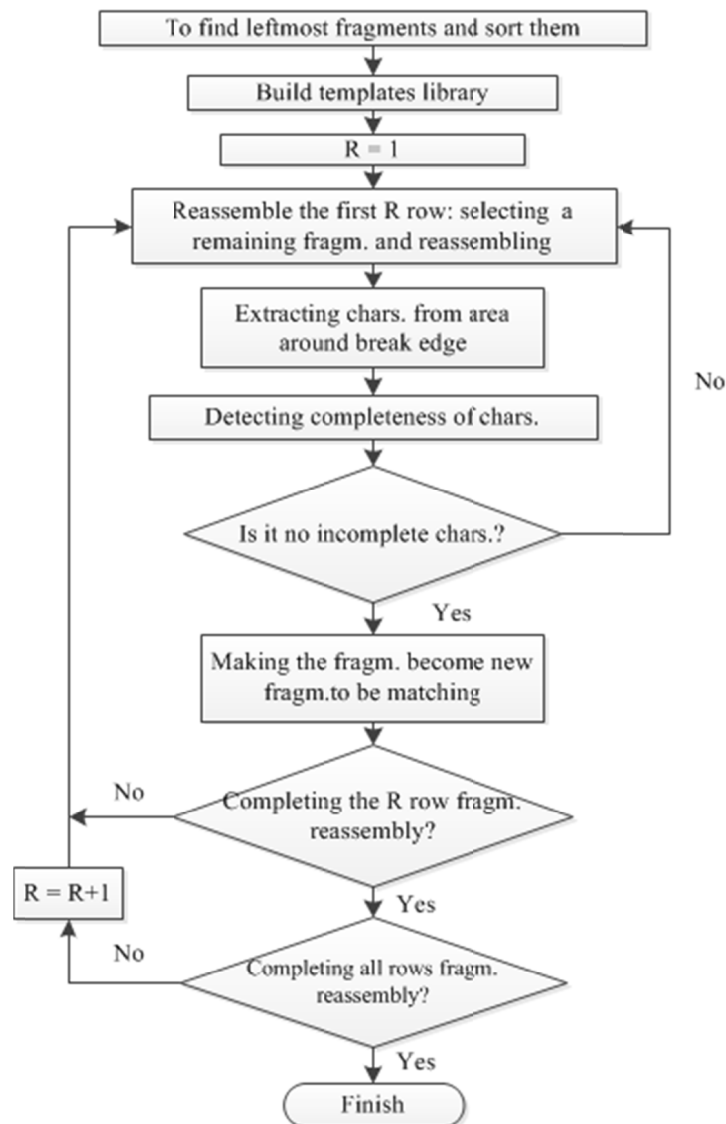


Figure 12. Algorithm flowchart of model for English fragments reassembly

4. Experiments and Results

The performance of the proposed methods was evaluated using digitally scanned images of actual fragments of paper image prints (CUMCM, 2013). There were Chinese fragments and English fragments. Furthermore, appendix 3 and appendix 4 included 209 fragment images respectively, and each fragment image had size 72×180 px (CUMCM, 2013). Model for Chinese fragments reassembly was applied to reassemble appendix 3 and model for English fragments reassembly was applied to reassemble appendix 4. The results are shown in Figure 13.

九十日春都过了，贪忙何处追游。三分春色一分愁。雨翻榆荚阵，风转柳花球。白雪清词出坐间。爱君才器两俱全。异乡风景却依然。团扇只堪题往事，新丝那解系行人。酒阑滋味似残春。

缺月向人舒窈窕，三星当户照绸缪。香生露藓见纤柔。搔首赋归欤。自觉功名懒更疏。若问使君才与术。何如。占得人间一味愚。海东头，山尽处。自古空槎来去。槎有信，赴秋期。使君行不归。别酒劝君君一醉。清润潘郎，又是何郎婿。记取钗头新利市。莫将分付东邻子。西塞山边白鹭飞。散花洲外片帆微。桃花流水鳜鱼肥。主人眼小。欲向东风先醉倒。已属君家。且更从容等待他。愿我无当世望，似君须向古人求。岁寒松柏肯惊秋。

水涵空，山照市。西汉二疏乡里。新白发，旧黄金。故人恩义深。难道东邻瘦损，凝然点漆精神。瑶林终自隔风尘。试看披鹤氅，仍是谪仙人。三过平山堂下，半生弹指声中。十年不见老仙翁。壁上龙蛇飞动。暖风不解留花住。片片著人无数。楼上望春归去。芳草迷归路。犀钱玉果。利市平分沾四坐。多谢无功。此事如何到得依。元宵似是欢游好。何况公庭民讼少。万家游赏上春台，十里神仙迷海岛。

虽抱文章，开口谁亲。且陶陶、乐尽天真。几时归去，作个闲人。对一张琴，一壶酒，一溪云。相如未老。梁苑犹能陪俊少。莫惹闲愁。且折

便邮。温香熟美。醉慢云鬟垂两耳。多谢春工。不是花红是玉红。一颗樱桃樊素口。不爱黄金，只爱人长久。学画鸦儿犹未就。眉尖已作伤春皱。清泪斑斑，挥断柔肠寸。嗔人问。背灯偷拭尽残妆粉。春事阑珊芳草歇。客里风光，又过清明节。小院黄昏人忆别。落红处处闻啼鹃。岁暮暮，须早计，要褰裘。故乡归去千里，佳处辄迟留。我醉歌时君和，醉倒须君扶我，惟酒可忘忧。一任刘玄德，相对卧高楼。记取西湖西畔，正暮山好处，空翠烟霏。算诗人相得，如我与君稀。约他年、东还海道，愿谢公、雅志莫相违。西州路，不应回首，为我沾衣。料峭春风吹酒醒。微冷。山头斜照却相迎。回首向来萧瑟处。归去。也无风雨也无晴。紫陌寻春去，红尘拂面来。无人不道看花回。惟见石榴新蕊，一枝开。

(a)

(b)

Figure 13. Results of appendix 3 and appendix 4

Table 1. Application results of the models

	Recovery rate	Comput. time
Model for Chinese Fragments	88.57%	3.03s
Model for English Fragments	96.17%	47.38s

4.1 Analysis on Result of Appendix 3

Table 1 shows that the recovery rate of model for Chinese fragments reassembly reached 88.57% and the reassembly time efficiency is high. In the process of reassembling fragments in appendix 3, two reasons affected the result of recovery. One of the important reasons is the line spacing of the second row in the last paragraph in appendix 3 was too wide, as shown in Figure 14. This reason caused serious error in the process of completing the reassembly of all rows. It can be seen that the model is more suitable for reassembling the document printing with the same line spacing. Another reason is that the line spacing of few fragments was very similar, causing error in the process of classifying.

虽抱文章，开口谁亲。且陶陶、乐尽天真。几时归去，作个闲人。对
一张琴，一壶酒，一溪云。相如未老。梁苑犹能陪俊少。莫惹闲愁。且折
便邮。温香熟美。醉慢云鬟垂两耳。多谢春工。不是花红是玉红。一颗樱
桃樊素口。不爱黄金，只爱人长久。学画鸦儿犹未就。眉尖已作伤春皱。

Figure 14. Compare of line spacing

4.2 Analysis on Result of Appendix 4

Table 1 shows that the recovery rate of model for English fragments is very high but the computation cost is higher. The reason is that this model was based on template matching, which has very high rate of letter recognition. However, we utilized method of exhaustion in the process, caused the higher computation cost.

The font in fragments of appendix 4 in this example was Zapf Calligraphic 801. In order to reassemble all fragments, Zapf Calligraphic 801's template library was set up. If we reassemble document with different font, we only need to set up a new template library. It can be seen that the English document fragments reassembly based on template matching presented in our paper can be widely employed in daily life.

In the reassembling process, occasionally there were little mismatched cases. Due to English model is based on the feature of English characters' 8-connected region. However, not all English characters are connected, such as Figure 15. These English punctuations didn't meet the requirement of 8-connected region. When this paper extracted these characters through connectivity, they were divided into two connected areas. But, all uppercase and lowercase letters meet the requirement of 8-connected region. And this paper focused on the use of the connectivity of letters in our model. Therefore, in the process of reassembling fragments by the model, the impact on accuracy, which caused by those special punctuations, could be ignored.



Figure 15. Not connected punctuations

In this example, all fragments were 180×72 px images. The method based on geometric characteristics of fragments can't solve the reassembly of these fragments. By extracting the characters according to 8-connected region, our model for English document fragments reassembly can recover the document accurately. Furthermore, most fragments in daily life have the similar shape. For example, fragments produced by shredder have the same shape. When people torn papers, they usually torn papers in half, then fold them in half and torn again. Finally, we get many fragments with same shape. Therefore, it can be seen that our model is very practical in daily life.

5. Conclusion

This paper considered the different features between Chinese characters and English characters. In order to reassemble Chinese fragments, the model based on the feature of line spacing and Chinese characters' feature that the same font has the same height is proposed. In order to reassemble English fragments, the model based on English characters' connectivity is proposed. Compared with traditional model based on geometric characteristics, the proposed models can accurately recover document fragments with the similar shapes. Therefore, the models are very practical in daily life. After a practical experiment, it can be seen that our models had high recovery rate and the robustness and reliability of models were verified.

In daily life, the handwritten documents are very common. In order to save some important fragments of handwritten documents, we need to recover those fragments. In general, the method of handwritten documents fragments reassembly is similar to the method presented in this paper. The main difference is the method of recognizing Chinese characters and English letters. In model of handwritten Chinese fragments reassembly, method of double elastic mesh is used to recognize Chinese characters (Chen, 2009). In model of handwritten English fragments reassembly, method of principal curves algorithm is used to recognize English characters (Ma, 2013).

As a future work, in order to speed up calculating, using intelligent algorithm to optimize the process of model calculation will be investigated.

Acknowledgments

The authors acknowledge the financial support of this research by the Key Laboratory of Product Packaging and Logistics of Guangdong Higher Education Institutes, the Fundamental Research Funds for the Central Universities, the project of the Natural Science Foundation of Guangdong Province (No.S2012010008773), and the projects of Zhuhai Science, Technology, Industry, Trade and Information Technology Bureau (No.2011B050102013 & 2012D0501990033).

References

- Chen, Z. H., Huang, X. H., Chen, P. F., Li, W. L., & Zhu, S. Y. (2009). Handwritten Chinese character recognition based on double elastic mesh. *Journal of Computer Applications*, 29(2), 395-397.
- CUMCM. (2013). *The problems of Contemporary Undergraduate Mathematical Contest in Modeling*. Retrieved from http://www.mcm.edu.cn/html_cn/node/d0a7128a712992392cb5cbad3da0eded.html
- Gao, H. B., & Wang, W. X. (2007). New connected component labeling algorithm for binary image. *Journal of Computer Applications*, 27(11), 2776-2777.
- Gao, X. T., Sattar, F., Quddus, A., & Venkateswarlu, R. (2007). Local natural scale based contour corner detection using wavelet transform. *Proceedings of Information, Communications and Signal Processing*, 25(6), 329-333. <http://dx.doi.org/10.1109/ICICS.2005.1689061>
- Hori, K., Imai, M., & Ogasawara, T. (1999). Joint detection for potsherds of broken earthenware. *Computer Vision and Pattern Recognition*, (2), 440-445. <http://dx.doi.org/10.1109/CVPR.1999.784718>
- Jia, H. Y., Zhu, L. J., Zhou, Z. T., & Hu, D. W. (2005). A Shape Matching Method for Automatic Reassembly of Paper Fragments. *Computer Simulation*, 23(11), 180-183. <http://dx.doi.org/10.3969/j.issn.1006-9348.2006.11.046>
- Kong, W. X., & Kimia, B. B. (2001). On solving 2D and 3D puzzles using curve matching. *Computer Vision and Pattern Recognition*, (2), 583-590. <http://dx.doi.org/10.1109/CVPR.2001.991015>
- Li, H., Manjunath, B. S., & Mitra, S. K. (1995). A contour-based approach to multisensor image registration. *Image Processing*, 4(3), 320-334. <http://dx.doi.org/10.1109/83.366480>
- Luo, Z. (2012). Semi-auto stitching of scrapped paper based on character characteristic. *Computer Engineering and Applications*, 48(5), 207-210. <http://dx.doi.org/10.3778/j.issn.1002-8331.2012.05.060>
- Ma, C., & Yu, M. (2013). Analysis and extraction of structural features of off-line handwritten characters based on principal curves algorithm. *Computer Engineering and Applications*, 49(3), 202-206.
- Otsu, N. (1979). An automatic threshold selection method based on discriminate and least squares criteria. *Denshi Tsushin Gakkai Ronbunshi*, (63), 349-356.
- Wolfson, H. J. (1990). On curve matching. *Pattern Analysis and Machine Intelligence*, 12(5), 483-489. <http://dx.doi.org/10.1109/34.55108>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).