A Discussion on Automatic Reassembling of Shredded Paper via Similarity Measurement Models

Meng Fanyu[1], Qian Wen[2]

[1]Schoole of Software, ShangHai Jiao Tong University, Shanghai Municipality, China
[2]School of Electronic Information and Electrical Engineering, ShangHai Jiao Tong University, Shanghai Municipality, China
mfy0120@gmail.com, qianwen@sjtu.edu.cn

Abstract - Reassembling of shredded paper from a collection of randomly mixed fragments is a common application of computer vision technique. It plays an irreplaceable role in several applied disciplines, such as classical forensics, archaeology, and military intelligence. In this paper we address the reassembly problem using a similarity measurement model applied in data mining. We propose a general process model for automatically analyzing the similarity between the gray value matrices of the scanned images in order to judge if they were adjacent to each other in the original document. Here we take both the Euclid distance and the characteristic values into consideration when it comes to similarity measurement. The implementation results provided demonstrate the validity of the proposed technique.

Index Terms - resembling of shredded paper, image processing, similarity measurement, characteristic vector

1. Introduction

Reassembly of shredded paper from a collection of randomly mixed fragments is a problem that arises in several applied disciplines, such as classical forensics, archaeology, and military intelligence. Traditionally, the reassembly work is done by hand. By this mean, the accuracy can be highly verified, but the efficiency is so frustrating. Especially when facing a huge amount of fragments, artificial splicing has become such a daunting work. With the development of computer technology, efforts have been made to complete the work automatically in order to improve the efficiency.

To address this problem, we propose a general process model and present a specific solution to reassembling scattered documents. Assuming that the fragments are from the same text document, printed in English, black and white, as well as well cut by a shredder (that is to say all the fragments are in shape of rectangle). The proposed model has three steps:

1) Preprocessing: We need to scan the fragments and number the scanned images so that they can be conveniently processed by computer technique. Thus we have converted it into a computer vision and data mining problem. Gray information can never be avoided when it comes to a problem of image analysis. [1]So a naïve thought would be presenting each image by its gray value matrix. By now, we can solve the problem by constructing a similarity measurement model of matrices applied in data mining.

2) Automatic Reassembling: In this step of the process, we use the information provided by the gray value matrices to speculate the adjacency relationships between the fragments. [2] We developed a computer program to implement the model, which provides a set of serial number to the human analyst, implying which of the fragments can be placed adjacently and composing parts of the original document.

3) Manual intervention: When the scale of the amount of the fragments has become fairly large, the computer resembling process cannot be total automatic. Since it has recovered parts of the original document in step 2, the only job left for the human analyst is to gain the almost complete parts together according to the context environment, which is an extremely trial job compared to the original one.

In this paper we focus on the second step, that is, the automatic reassembling work done by a computer program. The rest of this paper is organized as follows: in the following section we will introduce the automatic resembling method we proposed in details. That is the mathematic model we construct and the algorithm we used. Section 3 presents an initial experimental result we obtained after applying the method to fragments shredded from an English black-and-white text document, front and back. Finally, we conclude in section 4 with a discussion on future work.

2. Reassembling Process Based on Similarity Measurement

1) Find the pictures belong to the leftmost column.

Since the document is symmetric, it will work in the same way if we start it from the rightmost.

We use a gray value matrix G_i obtained from the preprocessing step to represent each square fragment *i*, each element $g_{mn} \in [0, 255]$. It's a common sense that the left and right side of a document should be margins. Hence we can safely state that pictures with their gray value matrix containing 10 leftmost column of 255s(i.e. the pixel value is 0xff) must belong to the leftmost column of the original document.

2) Try to find the right side neighbor of a certain fragment.

For each fragment belongs to the leftmost column, we try to find the right side fragments the same line as it one by one. Here, we suppose that the more similar their marginal features are, the more likely of the two pictures to be adjacent to each other originally. The similarity measurement model is constructed on the basis of the gray value matrices by the method of Euclid distance computation and characteristic vector comparison.

Suppose the present manipulated object is fragment *i*, we first compute square of Euclid distance between the rightmost column vector of G_i with the leftmost column vector of any other gray value matrix G_j, let it be Σ_1. To describe it in a

formal way:

$$D = \min_{i\in[0,208]} \min\left(f\left(i_a + j_a\right) + f\left(i_b + j_b\right)\right), \left(f\left(i_a + j_b\right) + f\left(i_b + j_a\right)\right)$$

Here i_a and i_b represents the front and back side of fragment *i* respectively and j_a and j_b for fragment *j*. In order to simplex the algorithm and increase the efficiency, we suppose that if is larger than 1,200,000, then *j* cannot be a candidate of the optimal solution, and the number of the candidates is bounded to 30. [3]Moreover, we set a threshold to improve the reliability. Let the least Euclid distance be *R*, and the second least one be R`. If $\frac{R}{\mathrm{R`}}$<0.4 , where the threshold 0.4 is obtained from our experimental experience, we can conclude that the optimal solution we have obtained must be correct, the matching process of fragment *i* can halt. Otherwise, more similarity measurement work should be done to the candidates by other standards.

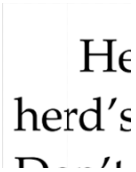

Figure 1. Joint fragment matched only by Euclid Distance

3) Further matching process according to characteristic vector.

In this step of matching, we make use of the fact that the text should be aligned to construct the characteristic vector. The element of the vector is defined as:

$$\lambda(k) = \begin{cases} 0, \text{ if all elements of the } k\text{th line of Gi is 0xff.} \\ 1, \text{ otherwise.} \end{cases}$$

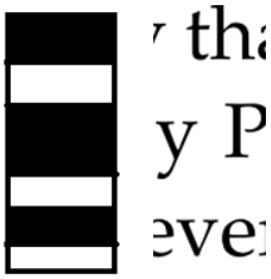

Figure 2. Characteristic vector illustration.

The construction process of the characteristic vector is shown in figure 1, where the black area implies the according element of the vector to be 1, and the blank area to be 0. [4]

Here, we propose another formula to measure the similarity of fragment i and j where i_1 and i_2 stands for the front and back side of fragment i respectively and j_1, j_2 for fragment *j*.

$$f\left(i+j\right) = \sum_{k-0}^{179} \left| \lambda_{i,k} - \lambda_{j,k} \right|$$

$$F\left(i+j\right) = f\left(i_1 + j_1\right) + f\left(i_2 + j_2\right)$$

If F > 110, we rule out fragment j from the candidate set obtained in (2), return the new candidate set.

The above matching process aims at the situations where there are some pieces of fragments with blank area on the left and right margin. Thus the margins supply such a small amount of information that the matching results of Euclid distance comparison can be incredibly inaccurate. The accuracy of the results can be improved by leaps and bounds due to characteristic vector comparison.

4) Repeat the matching process in (2) on the new candidate set. Checkout whether the present results can meet the threshold condition mentioned in (2). If thus, return the optimal solution, the fragment to the right of i in the original document. Otherwise, pause the left to right matching process of the present line. The current result is shown in figure 3:

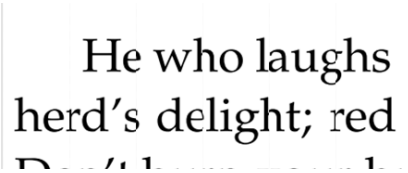

Figure 3. Joint fragment obtained from process (1)-(4).

By now, the algorithm flow chart can be shown as followed:

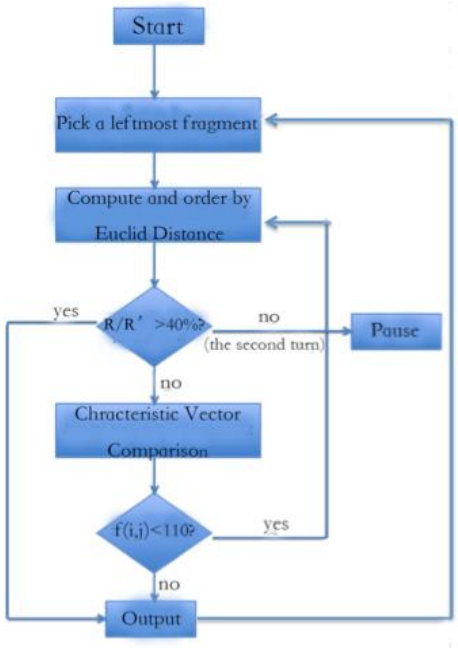

Figure 4 Algorithm flow chart.

5) Execute the matching process in (1)~(4), only from right to left this time.

6) If the reassembling process cannot go further, we can turn to the information supplied by the top and bottom margin of the fragments. At the same time, the DFS exploration can be applied to obtain more location relationships among the fragments as illustrated in figure 5.

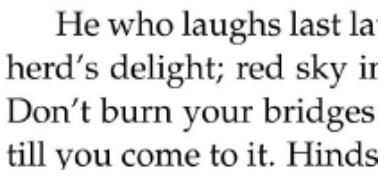

Figure 5 Joint fragment of the front obtained by DFS Search.

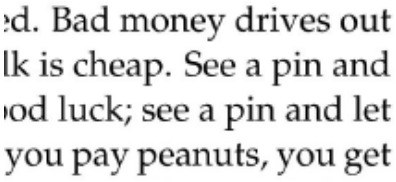

Figure 6 Joint fragment of the back obtained by DFS Search.

7) After the above process, if the reassembling of the document still cannot be completed, then some manual intervention is necessary here. But the left work must be trial now.

He who laughs last laughs longest. Red sky at night shepherd's delight; red sky in the morning, shepherd's warning. Don't burn your bridges behind you. Don't cross the bridge till you come to it. Hindsight is always twenty-twenty.

Figure 7 Front side of a complete part of the original

What can't be cured must be endured. Bad money drives out good. Hard cases make bad law. Talk is cheap. See a pin and pick it up, all the day you'll have good luck; see a pin and let it lie, bad luck you'll have all day. If you pay peanuts, you get

Figure 8 Back side of a complete document

3. Experiments & Results

We tested our computer program on an English text document, front and back. We used a paper shredder to cut it into a collection of fragments in 11 rows * 19 columns, each piece contains 180*72 pixels. Our approach has been proved to be effective with a automatic reassembling frequency of 198/209≈94.74%.

Some interesting exceptions have occurred during the experiment. For example, character 'r' has been split into two halves. However, there is hardly vertical stroke of it left in the left part. Thus the similarity measured by the Euclid distance between the two fragments would be poor, which clearly violates the reality.

4. Conclusion

The given experimental results have verified the efficiency and accuracy of the similarity measurement approach to automatically reassemble shredded paper. However, when the scale of the amount of the fragments has been large enough, the information supplied by each piece will seem to be limited, some manual intervention will be necessary. Thus, Further investigation is necessary to establish a library of Chinese characters and semantic analyzer to perfect the automatic ability of the computer program. To sum up, the supposed approach can solve the problem of reassembling shredded paper automatically to a great extent. It will favor people with facilities when it comes to the problem of automatic Reassembling of Shredded Paper in practical application areas.

References

[1] Kulesh Shanmugasundaram and Nasir Memon. Automatic Reassembly of Document Fragments via Context Based Statistical Models. IEEE Computer Security Applications Conference, 2003.

[2] Fabian Richter, Christian X. Ries, Rainer Lienhart. A GRAPH ALGORITHMIC FRAMEWORK FOR THE ASSEMBLY OF SHREDDED DOCUMENTS. Multimedia and Expo (ICME), 2011 IEEE International Conference.

[3] Florian Kleber and Robert Sablatnig. A Survey of Techniques for Document and Archaeology Artefact Reconstruction. Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference

[4] Deever, A. and Gallagher, A. Semi-automatic assembly of real cross-cut shredded documents. Image Processing (ICIP), 2012 19th IEEE International Conference.