Summary

In this paper, 3 mathematical models of computer-aided paper fragments have been set up based on the multivariate statistical analysis, where system clustering method, the Fuzzy C Means clustering method and some other methods are applied. And we have the given fragments automatic reassembly very well.

Question 1, using the method of system clustering to classify gray vectors of image bounds for length cutting paper fragments, we have set up the 1st mathematical model of computer-aided paper fragments. And based on the model, a full-automatic reassembly algorithm of fragments is designed. Firstly, we use MATLAB to deal with fragments in attachment 1 and 2 by digital image processing. Therefore, each image can be represented by a data matrix. We only extract the left and right edge vectors of each matrix and save the 38 sample data. Then we use the system clustering method to classify the samples into 19 classes, of which the left sample can match to the proper right sample exactly. At last, we can match the fragments completely without any manual intervention. The fragments serial number of the complete Chinese image is as follows: 8, 14, 12, 15, 16, 1, 3, 10, 2, 4, 5, 7, 9, 13, 18, 11, 17, 0, 6; while the fragments serial number of the complete English image is as follows: 3, 6, 2, 7, 15, 18, 11, 1, 0, 5, 9, 13, 10, 8, 12, 14, 17, 16, 4. The two complete image figures are in appendix 1.1 and 1.2, respectively.

Question 2, on the basis of the question one, we improved the model and algorithm to establish the 2nd mathematical model of computer-aided paper fragments for the transverse and longitudinal cutting paper fragments. And based on the model, a semi-automatic reassembly algorithm of fragments is designed. At first, blacked the transverse text of every fragment, using the Fuzzy C Means clustering method to classify the blackened images, and we can obtain 11 image classes, in which one class includes about 19 paper fragment images. Next, using the method of system clustering to classify the paper fragment images in each above class, similar to Question 1. Then there are over than 11 classes in all, but not very much. At second, also using the method of system clustering to classify the incomplete recovery paper fragment images. In the last process, the manual intervention will be applied to complete the whole several big fragments with only less than15 minutes. The fragments serial number of the complete Chinese image and its figure in attachment 3 are in appendix 1.3, while the fragments serial number of the complete Chinese image and its figure in attachment 4 are in appendix 1.4.

Question 3, in the fundamental of 1st and 2nd model and their algorithms, as well as the English printing characteristics, we create the 3rd mathematical model of computer-aided paper fragments for the duplex printing paper fragments. And based on the model, a semi-automatic reassembly algorithm of fragments is designed. This algorithm is almost the same as the 2nd algorithm, but it need more manual intervention. At the beginning, we make xxxa file in attachment 5 separated from xxxb file, and create a positive identification algorithm program. After the identification of the pros and cons of each file, we can accord to the 2nd algorithm to recover the files.

Key words: system clustering method Fuzzy C - Means MATLAB
full-automatic reassembly semi-automatic reassembly

Mathematical Model of Computer-aided Paper Fragments Based on the Multivariate Statistical Analysis

1 Restatement of the Problem

Broken file recovery have important applications from judicial evidence recovery, the repair of historical documents and acquisition of military intelligence etc. Traditionally, stitching must be completed by the artificial restoration work, the accuracy is quite high, but the efficiency is very low. It is difficult by artificial stitching to complete the task in a short period of time, when a huge number of fragments need repair. With the development of computer technology, people are trying to develop an automatic stitching technique for paper fragments to improve the stitching recovery efficiency. Please discuss the following questions:

1. You are expected to establish a stitching recovery model and algorithm of the paper fragments, where the paper fragments comes from the same page of a printing text file (longitudinal cutting only). And making the given paper fragments complete, one is a page of Chinese fragments in attachment 1, while the other one is a page of English fragments in attachment 2. If the recovery process requires manual intervention, please write the means of intervention and the time node. Results should be in the form of image and table together.

2. If the shredder is slitting and crosscutting, please design a stitching recovery model and algorithm, and making the given paper fragments complete, one is a page of Chinese fragments in attachment 3, while the other one is a page of English fragments in attachment4. If the recovery process requires manual intervention, please write the means of intervention and the time node. Results expression is like the above.

3. The given paper fragments in the above are all printed documents with one side. Embarking from the reality situation, there are also paper fragments from duplex printing file to be recovered. Please try to design the corresponding paper fragments recovery model and algorithm, and complete the given paper fragments in attachment 5. Results expression is like the above.

2 Assumptions and Symbol Definition

2.1 Assumptions

1) The given data is true and correct

2) The paper fragments are all regular geometric figure.

2.1 Symbol Definition

Notation	Definition	Units
U	membership matrix	
V	clustering center matrix	
G	the class by System clustering method	
X	Observation data matrix or Gray level matrix	
D_{KL}	The distance between G_K and G_L	
$J(\mathrm{U},\mathrm{V})$	the weighted square distance between gray vector in all kinds of classes and clustering center	
d_{ij}	The distance between x_i and x_j	
x_i	The i_{th} gray vector samples($i=1,2,\cdots,n$)	
m	The weight of gray vector samples x_k belong to the class i	
c	The class number of n gray vector samples	

3 Analysis of the Problem

3.1 Analysis of Question 1

The paper fragments in attachment 1 and attachment 2 are longitudinal cutting by shredder of one page in a printing text file respectively, of which the shapes are regular. To recover the paper fragments is to match them together in correct written rules. Firstly, using digital image processing deals with the BMP format images in attachments. That is, importing the paper fragment images to MATLAB cell array, and every image can be represent by a grayscale data matrix. Considering that noise may be introduced into the image in image collection, transfer and conversion processes, which may make the image fuzzy, anamorphic, and image grey value cannot mutate, here we use the median filter to protect the edge information and smooth noise. Secondly, saving column vector of every boundary, 38 in all. Thirdly, using the system clustering method classifies the 38 samples. In this way, the 1st

mathematical model of computer-aided paper fragments by length cutting have developed. After getting the classification result, there will be 19 classes, and it is easy to design a short program to match them together automatically and show us a complete figure.

Specific steps are as follows:

1) Digital image processing: after sampling and quantifying the fragment images given in attachments, we should import the image data into the MATLAB workspace, thus each image can be represent by a 1980 x 72 grayscale data matrix, 19 in all. Then storing the 19 matrix to a 1 x 19 cell array.
2) Median filtering processing
3) Extracting the sample data: after Median filtering processing, storing the column vectors of every matrix boundary one by one into a 2 x 19 cell array. The first row of the cell is the boundary on the left hand side of each image, namely the first column of every grayscale data matrix, while the second row of the cell is the boundary on the right hand side of each image, namely the 72th column of every grayscale data matrix. There are a total of 38 sample data.
4) System clustering method: using system clustering method classifies the 38 sample data, and we can get the result.
5) Recovery and saving: the sample data will be result in 19 classes. Every left boundary has its corresponding right boundary. So we can design a program to match them together automatically into a complete figure and save it.

3.1 Analysis of Question 2

As for Question 2, in which the shredder is slitting and crosscutting, we are improving the model and algorithm on the basis of Question 1. In this problem, we will not only consider to match the corresponding left and right boundary, but also the up and down boundary. And the information of the fragment's four sides are few. If we apply the 1st model to solve the problem directly, we may fail to get satisfactory result within the effective time. However, Chinese characters give us some kind of inspiration. Since the characters of the printing text file is arranged in a row, which is equivalent to a series of equal squares arranging in a row one by one. But there are no such a rule between columns. So primarily, we are going to blacken the rows with characters as black tapes, then using the Fuzzy C Means clustering method to classify those black tapes into several subsets according to a certain standard. The characters in one row of the printing text file are close together, and the space between rows is almost the same. It is easy for us to design a program to classify the 11 x 19 pieces into 11 classes in the above standard, for the information is enough to get quite correct classification. Then we can use the system clustering method in model 1 to match the fragments in one class together and recover every line. Subsequently, we will match the 11 lines in its right order. But it cannot help to use the 1st model again, since the first model is specialized for grayscale vector of boundaries, and there is likely to cut a lot of blank space. However, as we know, the space between rows is almost the same. So we can also design a program to match the space together, setting the gap width of the border in a numerical range between the right sequences. Finally, the

paper fragments can be complete.

3.1 Analysis of Question 3

The 3rd mathematical model will include the above two model, for the 2nd model nests the 1st model. At start, we'd like to separate xxxa files in attachment 5 from xxxb file, and import them into MATLAB workspace respectively. Then, it is necessary for us to design a program to identify whether the pictures in xxxa file from the same side of one page. If it is, we can recover a and b sides of the file in accordance with the 2nd model. Otherwise, we'd like to design another program to find out the pros and cons of paper documents, and then use the 2nd model to recover the paper fragments.

4 Preparation of the Problem

4.1 data processing

The given data in this topic is all BMP image data. Firstly, we'd like to convert the data into the data available to the program, and use digital image processing to deal with these image data. Then importing the paper fragment images to MATLAB cell array, and every image can be represent by a grayscale data matrix. There are 19 such matrix in attachment 1 and 2 respectively, where every matrix size is 1980 x 72. Finally, store the 19 matrixes into a 1 x 19 cell array. Data structure transformation is shown in figure 1. Image data of other accessories is to do the same conversion.

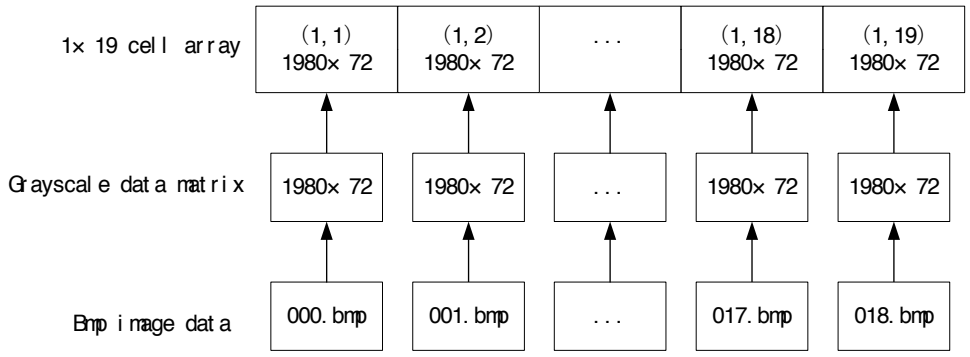

Figure 1 data structure transition diagram in attachment 1