doi:10.3969/j.issn.1672 - 4348.2014.03.014

# 基于分层聚类的仅横纵切碎中文纸片拼接分类

熊保平，祝丽华

（福建工程学院 数理系，福建 福州 350118）

**摘要**：根据仅知道碎纸机横纵切破碎中文纸片的文字之间存在统一的文字高度和行间距等文字特征，提出在匹配拼接前，把碎纸片的二维图像通过横向投影方式转变为保留文字高度、行间距等特征的一维向量，并利用它进行分层聚类，对所有碎纸片进行行分类，从而有效地减少匹配拼接的时间，提高匹配拼接正确性，实验结果表明，该方法精确，有效。

**关键词**：碎纸拼接；分层聚类；横向投影

**中图分类号**：TP371.4　　**文献标志码**：A　　**文章编号**：1672 - 4348(2014)03 - 0268 - 04

## The stitching classification of simple horizontally or vertically shredded Chinese papers based on hierarchical clustering

Xiong Baoping, Zhu Lihua

(Mathematics and Physics Department, Fujian University of Technology, Fuzhou 350118, China)

**Abstract**: Due to the Chinese characters' features of unified text height and line spacing, the 2D images of the scraps of simply horizontally or vertically shredded Chinese paper were converted into 1D vectors with the text height and line spacing via horizontal projection method. The paper shredders were differentiated in lines and classified based on hierarchical clustering to reduce the time of the stitching and to improve the accuracy of the stitching. The experimental results show that the method is accurate and effective.

**Keywords**: scraps stitching; hierarchical clustering; horitzontal projeccion

常规的破碎纸片计算机拼接算法一般都是利用破碎纸片的边缘几何特征搜索与之相匹配的相邻碎纸片进行拼接[1-9]，但这种基于边缘几何特征的拼接方法并不适用于边缘几何形状基本一样的碎纸机仅横纵切破碎中文纸片的拼接。现在许多重要中文文档的销毁都采用横纵切的碎纸机，这样的碎片基本都是形状大小一样的矩形，想利用边缘特征进行拼接是不可能的。

对这种无法使用边缘几何特征进行拼接的碎纸片，由于理论与技术的限制只能采用边缘像素信息的匹配进行拼接，但是通常单页文档经碎纸机销毁后其碎片都是大几百至上千片，计算量巨大且极易产生错误的匹配。

本文针对这一问题，以及中文文字之间存在统一的文字高度和行间距等文字特征，提出在匹配拼接前，把碎纸片的二维图像通过横向投影方式转变为保留文字高度、行间距等特征的一维向量，并利用它进行分层聚类对所有碎纸片进行行分类。

## 1 文字特征的获取

为了问题简单化，假设被切割的文档为中文且

收稿日期：2014 - 06 - 10
基金项目：福建工程学学院科研发展青年基金(GY - Z0892，GY - 13009)
第一作者简介：熊保平(1980 - )，男(汉)，江西南昌人，讲师，硕士，研究方向：生物医学仪器及智能测试，算法，图像处理。

横的切割方向正好与文字的水平方向平行，把单页文档切割为 11 行 ×19 列，如图 1 所示。

**图 1　碎纸机切割后文档**

**Fig. 1　Shredded file**

从图 1 可知，在同一行的碎纸片竖的方向具有相同的文字开始与结束，以及相同的行间距开始与结束，根据这些特点可以对破碎纸片以行分类。但是从 2 维图像中寻找具有相同的文字开始与结束或者具有相同的行间距开始与结束，不仅算法复杂，耗时大且极易出错，所以本文提出了一种简便且保留字高，行间距等特点的二维转一维的方法，即先对碎纸片图像二值化后进行横向投影，其转换方法如式(1)

$$f(x) = \begin{cases} 1\cdots\cdots\left(\sum_{y=1}^{N} I(x,y) > N-1\right) \\ 0\cdots\cdots\left(\sum_{y=1}^{N} I(x,y) \leqslant N-1\right) \end{cases} \tag{1}$$

式(1)中，$I(x,y)$ 是二值化后的碎纸片图像的像素值；$N$ 是破碎纸片图像的宽度；$f(x)$ 为转化后的二值化一维向量。式(1)的转换效果如图 2。图 2 中(a)为碎纸片的二值化图像，(b)为碎纸片转化后的二值化一维向量。其中白色为 1，黑色为 0。由图 2(a)、(b)的比较可知，(b)比较精确地表达了(a)竖向文字的开始与结束，行间距的开始与结束等文字特征。

(a)碎纸片图像　　(b)横向投影后的一维向量

**图 2　横向投影对比图**

**Fig. 2　Comparison between horizontal projections**

## 2　分层聚类

由上文可知转化后的二值化一维向量中值相等的其对应的碎纸片必然在同一行中；但是由于文字的本身特点以及横向投影算法的细小误差，导致在判定文字的开始和结束上存在细小误差，所以在行分类判定的时候，只能通过转化后的二值化一维向量之间的汉明距离来判定，而这一过程刚好与分层聚类算法相符，所以在分类的过程中采用了分层聚类方法。

分层聚类算法是一个基于相似性将给定对象集合划分为若干个聚类的过程，使得同一聚类的对象具有最高相似性，不同聚类之间具有最低的相似性。聚类的对象可以是一个测量向量，或者特征向量，也可以是空间上的一个点。相似性可以使用距离来度量，如汉明距离。所以横向投影后碎纸片的分类算法如下[10]：

1)(初始化)把每个碎纸片的二值化一维向量归为一类，计算每两个类之间的汉明距离，也就是样本与样本之间的相似度；

2)寻找各个类之间最近的两个类，把他们归为一类(这样类的总数就少了一个)；

3)重新计算新生成的这个类与各个旧类之间的相似度；

4)重复 2 和 3，直到所有样本点都归为一类，结束。

## 3 实例分析

为了验证该方法的有效性,在测试实验中把图1中的209片碎纸片进行随机编号并打乱,其完整的11行×19列的图片序号如表1所示。

首先针对表一序号所对应的碎纸片的二维图像进行图像的相似性聚类,其11类的分类结果如表2所示。

由表1与表2的数据对比分析可知,没有出现错判,但27片碎纸片无法进行程序分类。

根据本文提出的分层聚类方法,把表1中所有序号相对应的碎纸片的二维图像由式(1)进行横向投影转化为一维向量,并利用一维向量的信息进行分层聚类,通过实验发现,分类≥19后分类稳定,其结果如表3所示。

**表1 碎纸机切割后文档碎片序号**

**Tab.1 The order of shredded file**

| 行 | 列 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 1 | 49 | 54 | 65 | 143 | 186 | 2 | 57 | 192 | 178 | 118 | 190 | 95 | 11 | 22 | 129 | 28 | 91 | 188 | 141 |
| 2 | 61 | 19 | 78 | 67 | 69 | 99 | 162 | 96 | 131 | 79 | 63 | 116 | 163 | 72 | 6 | 177 | 20 | 52 | 36 |
| 3 | 168 | 100 | 76 | 62 | 142 | 30 | 41 | 23 | 147 | 191 | 50 | 179 | 120 | 86 | 195 | 26 | 1 | 87 | 18 |
| 4 | 38 | 148 | 46 | 161 | 24 | 35 | 81 | 189 | 122 | 103 | 130 | 193 | 88 | 167 | 25 | 8 | 9 | 105 | 74 |
| 5 | 71 | 156 | 83 | 132 | 200 | 17 | 80 | 33 | 202 | 198 | 15 | 133 | 170 | 205 | 85 | 152 | 165 | 27 | 60 |
| 6 | 14 | 128 | 3 | 159 | 82 | 199 | 135 | 12 | 73 | 160 | 203 | 169 | 134 | 39 | 31 | 51 | 107 | 115 | 176 |
| 7 | 94 | 34 | 84 | 183 | 90 | 47 | 121 | 42 | 124 | 144 | 77 | 112 | 149 | 97 | 136 | 164 | 127 | 58 | 43 |
| 8 | 125 | 13 | 182 | 109 | 197 | 16 | 184 | 110 | 187 | 66 | 106 | 150 | 21 | 173 | 157 | 181 | 204 | 139 | 145 |
| 9 | 29 | 64 | 111 | 201 | 5 | 92 | 180 | 48 | 37 | 75 | 55 | 44 | 206 | 10 | 104 | 98 | 172 | 171 | 59 |
| 10 | 7 | 208 | 138 | 158 | 126 | 68 | 175 | 45 | 174 | 0 | 137 | 53 | 56 | 93 | 153 | 70 | 166 | 32 | 196 |
| 11 | 89 | 146 | 102 | 154 | 114 | 40 | 151 | 207 | 155 | 140 | 185 | 108 | 117 | 4 | 101 | 113 | 194 | 119 | 123 |

**表2 图片相似性聚类的分类结果**

**Tab.2 The clustering result in accordance with images similarity**

| 类 | 图片序号 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | 22 | 28 | 49 | 54 | 57 | 65 | 91 | 95 | 118 | 129 | 141 | 143 | 178 | 186 | 188 | 190 | 192 | |
| 2 | 19 | 20 | 36 | 52 | 61 | 63 | 67 | 69 | 72 | 78 | 79 | 96 | 99 | 116 | 131 | 162 | 163 | 177 | |
| 3 | 18 | 23 | 26 | 41 | 50 | 62 | 76 | 86 | 87 | 100 | 120 | 142 | 147 | 168 | 179 | 191 | 195 | | |
| 4 | 35 | 38 | 46 | 81 | 88 | 103 | 122 | 130 | 144 | 148 | 161 | 167 | 189 | 193 | | | | | |
| 5 | 17 | 33 | 71 | 80 | 83 | 132 | 133 | 156 | 170 | 198 | 200 | 202 | | | | | | | |
| 6 | 12 | 14 | 30 | 31 | 39 | 51 | 73 | 82 | 107 | 115 | 128 | 134 | 135 | 159 | 160 | 169 | 176 | 199 | 203 |
| 7 | 42 | 43 | 47 | 58 | 77 | 84 | 90 | 94 | 97 | 112 | 121 | 124 | 127 | 136 | 144 | 149 | 164 | 183 | |
| 8 | 16 | 21 | 66 | 106 | 109 | 110 | 125 | 139 | 145 | 150 | 157 | 173 | 181 | 182 | 184 | 187 | 197 | 204 | |
| 9 | 10 | 29 | 37 | 44 | 48 | 55 | 59 | 64 | 75 | 92 | 98 | 104 | 111 | 171 | 172 | 180 | 201 | 206 | |
| 10 | 7 | 45 | 53 | 56 | 68 | 126 | 137 | 138 | 158 | 174 | 175 | 208 | | | | | | | |
| 11 | 40 | 89 | 101 | 102 | 108 | 113 | 114 | 117 | 119 | 123 | 140 | 146 | 151 | 154 | 155 | 185 | 194 | 207 | |

**表3 分层聚类的分类结果**

**Tab.3 The hierarchical clustering result**

| 类 | 图片序号 | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 125 | | | | | | | | | | | | | | | | | |
| 2 | 27 | 60 | 85 | 152 | 165 | 170 | 205 | | | | | | | | | | | |
| 3 | 3 | 12 | 31 | 39 | 51 | 73 | 82 | 107 | 115 | 128 | 134 | 135 | 159 | 160 | 169 | 176 | 199 | 203 |

续表 3

| 类 | 图片序号 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 18 | 23 | 26 | 30 | 41 | 50 | 62 | 76 | 86 | 87 | 100 | 120 | 142 | 147 | 168 | 179 | 191 | 195 |
| 5 | 4 | 40 | 101 | 102 | 108 | 113 | 114 | 117 | 119 | 123 | 140 | 146 | 151 | 154 | 155 | 185 | 194 | 207 | |
| 6 | 6 | 19 | 20 | 36 | 52 | 61 | 63 | 67 | 69 | 72 | 78 | 79 | 96 | 99 | 116 | 131 | 162 | 163 | 177 |
| 7 | 15 | 17 | 33 | 80 | 83 | 132 | 133 | 156 | 198 | 200 | 202 | | | | | | | | |
| 8 | 24 | 35 | 38 | 46 | 81 | 88 | 103 | 122 | 130 | 148 | 161 | 167 | 189 | 193 | | | | | |
| 9 | 34 | 42 | 43 | 47 | 58 | 77 | 84 | 90 | 94 | 97 | 112 | 121 | 124 | 127 | 136 | 144 | 149 | 164 | 183 |
| 10 | 2 | 11 | 22 | 28 | 49 | 54 | 57 | 65 | 91 | 95 | 118 | 129 | 141 | 143 | 178 | 186 | 188 | 190 | 192 |
| 11 | 5 | 10 | 29 | 37 | 44 | 48 | 55 | 59 | 64 | 75 | 92 | 98 | 104 | 111 | 171 | 172 | 180 | 201 | 206 |
| 12 | 07 | 45 | 53 | 68 | 126 | 137 | 138 | 158 | 174 | 175 | 208 | | | | | | | | |
| 13 | 13 | 182 | | | | | | | | | | | | | | | | | |
| 14 | 32 | 56 | 70 | 93 | 153 | 166 | 196 | | | | | | | | | | | | |
| 15 | 89 | 25 | 74 | 105 | | | | | | | | | | | | | | | |
| 16 | 14 | | | | | | | | | | | | | | | | | | |
| 17 | 89 | | | | | | | | | | | | | | | | | | |
| 18 | 16 | 21 | 66 | 106 | 109 | 110 | 139 | 145 | 150 | 157 | 173 | 181 | 184 | 187 | 197 | 204 | | | |
| 19 | 71 | | | | | | | | | | | | | | | | | | |

通过表 1 与表 3 的对比，发现分类中未出现错判的情况，而且通过分析发现单行被分多类的情况是由段结束或段开始时留下的空格而导致的匹配不一致所引起的，如表 1 中的第 4 行由于出现了段的结束所以被分为了类 8 和类 15，又如表 1 中的第 6 行由于出现了段的开始所以被分为了类 3 和类 16。

由表 2 和表 3 的数据对比可以发现，本文提出的分层聚类方法分类更加精确，误判可能性小。

## 4　结论

通过实验数据表明，上述方法可以对碎中文纸片的行进行较精确的分类，从而有效地减小了碎纸片的匹配范围，减少了拼接时间，提高了拼接的准确率。但是其缺点是仅对横切割的方向与中文文字行的方向平行的碎纸片分类比较有效。

## 参考文献：


[1] 王磊，莫玉龙. 戚飞虎. 基于 canny 理论的边缘提取改善方法[J]. 中国图像图形学报，1996，1(3)：191－195.

[2] 陶波，于志伟，郑筱祥. 图像的自动拼接[J]. 中国生物医学工报，1997，16(4)：29－35.

[3] 钟家强，王润生. 基于边缘的图像配准改进算法[J]. 计算机工程学报，2001，23(6)：25－29.

[4] 周鹏，谭勇，徐守对. 基于角点检测图像配准的一种新算法[J]. 中国科学技术大学学报，2002，32(4)：455－461.

[5] 吕科. 空间物体碎片的自动复原算法研究[J]. 计算机应用，2003(4)：756－758.

[6] 王斌君，赵兴涛，王靖亚，等. 犯罪现场碎片拟合算法研究[J]. 中国人民公安大学学报，2011，2：46－50.

[7] Freeman H, Garder L, Puzzies A J. The computer solution of a problem in pattern recognition[J]. IEEE Trans Elec Comp, 1964(13)：118－127.

[8] Woifson H, Kaivin A, Schonberg E, et al. Soiving jigsaw puzzles by computer[J]. Annales of Operations Research, 1988(12)：51－64.

[9] Grigore C B, Haim W J. Solving jigsaw puzzles by a robot[J]. IEEE Transactions on Robotics and Automation, 1989, 5(6)：752－764.

[10] 张红云，刘向东，段晓东，等. 数据挖掘中聚类算法比较研究[J]. 计算机应用与软件，2003，20(2)：5－6.


（责任编辑：陈雯）