

基于智能算法的破碎文件拼接复原技术的研究

霍敏霞 薛博桓

(重庆邮电大学移通学院计算机科学系 重庆 401520)

摘 要 破碎文件的拼接在司法物证复原、历史文献修复以及军事情报获取等领域都有着重要的应用。随着科技的发展,破碎文件的自动拼接技术以及拼接复原效率是目前研究的热点问题。对于单一方向的破碎文件,文中采用 0-1 匹配、Pearson 相关系数的灰度匹配,针对破碎文件的长短、多少,以及中英文等情况建立了两种模型;对于横切和纵切方向的破碎文件,采用聚类分析和灰度匹配建立模型并进行相应处理,从而实现对这些破碎文件的全自动或半自动拼接复原。

关键词 破碎文件,0-1 匹配,灰度匹配,聚类分析,拼接复原

中图分类号 TP391 文献标识码 A

Research on Splicing Recovery of Broken Files Based on Intelligent Algorithms

HUO Min-xia XUE Bo-huan

(Department of Computer Science, College of Mobile Telecommunication, Chongqing University
of Posts and Telecommunications, Chongqing 401520, China)

Abstract The technique of the splicing of broken files in the areas of recovering judicial evidence and historical documents and acquiring military intelligence has an important application. With the development of science and technology, the technique of automatic splicing for broken files is now a hot researching spot. As for unidirectional broken files, the technique in this paper was to build two kind of models based on the length and size of broken files and the condition of the usage of English and Chinese, with matching 0-1 and Pearson related coefficient gray matching. As for broken files in transverse direction or longitudinal direction, the technique in this paper was to build models and perform relevant processing, with the way of cluster analysis and gray matching to realize the full-automatic and semi-automatic recovery of those broken files.

Keywords Broken files, 0-1 matching, Gray matching, Clustering analysis, Splicing recovery

1 引言

破碎文件的拼接在司法物证复原、历史文献修复以及军事情报获取等领域都有着重要应用^[1]。传统上,拼接复原工作需由人工完成,准确率较高,但效率很低,特别是当碎片数量巨大、人工拼接很难在短时间内完成时。因此,我们需要找到提高拼接速度和准确率较高的方法来对破碎的文件进行拼接,在提高拼接复原效率的同时提高其精度。

本文研究的主要内容有两方面:1)仅考虑单一碎纸方向(比如纵切),对给定的来自同一页印刷文字文件的碎纸机破碎纸片建立有效的碎纸片拼接复原模型和算法,并对碎片数据进行拼接复原实例验证;2)对于碎纸机既纵切又横切的情形,设计了基于聚类分析和灰度匹配相结合的算法进行拼接复原,同时给出了基于中文文本复原的实例验证。

2 单一碎纸方向的拼接复原研究

本节给出了两种算法模型,从不同角度对单一碎纸方向进行拼接复原研究。给定的破碎纸片来自同一页印刷文件,假设纵切的中文碎片为大小相同的矩形,其边缘分为规整的

上、下、左、右,位于碎纸片左右边缘的文字有些保全而有些则残缺不全,通过对碎纸片的左右边缘文字的分析归类,将图形进行灰度化处理^[2],即根据整体图像呈现出的黑白效果将图像中点的灰度值设为 0~255 的数值。结合其左右边缘的灰度值特征把边缘文字完全或残缺的情况用数值表示出来,得到每张碎纸片特定的矩阵,再对各矩阵进行比对配对。同理,对纵切的英文文本进行拼接复原。

2.1 基于 0-1 匹配模型的碎纸片拼接复原机制

本文首先从一组包含 19 张碎纸片的图库中将第 i 张碎片还原成灰度值,得到 72×1980 的矩阵。

$$X_i^1 = \begin{bmatrix} a_{11}^{i1} \\ \vdots \\ a_{m1}^{i1} \end{bmatrix} = \begin{bmatrix} a_{11}^{i1} & \cdots & a_{1n}^{i1} \\ \vdots & \ddots & \vdots \\ a_{m1}^{i1} & \cdots & a_{mn}^{i1} \end{bmatrix}$$

其中, $n=1,2,\dots,72, m=1,2,\dots,1980, a_{mn}^{i1}=0,1,2,\dots,255$ 。

用函数将各点值划分为两类:

$$u_{mn}^{i1} = \begin{cases} 1, & a_{mn}^{i1} = 0 \\ 0, & a_{mn}^{i1} = 1, 2, \dots, 255 \end{cases}$$

将矩阵 X_i^1 的首列与末列提取出来,得到:

$$X_i' = \begin{bmatrix} u_{11}^{i1} & u_{172}^{i1} \\ \vdots & \vdots \\ u_{m1}^{i1} & u_{m72}^{i1} \end{bmatrix}, i=1,2,\dots,19, m=1,2,\dots,1980$$

然后将矩阵 X_i' 分成相应的块,得到矩阵 X_i'' ,使矩阵的每一行能够反映每一行文字的情况:

$$X_i'' = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ \vdots & \vdots \\ A_{281} & A_{282} \end{bmatrix}, A_{kj} = \begin{bmatrix} u_{ij}^k \\ \vdots \\ u_{nj}^k \end{bmatrix}, k=1,2,\dots,28, j=1,2$$

其中, $u_{ij}^k - u_{mj}^k$ 表示一行汉字加上一个行间距的距离。 A_{kj} 的取值遵循如下原则:若存在 $u_{ij}^k = 1, m \leq s \leq n$, 则 $A_{jk} = 1$; 若存在 $u_{ij}^k = 0, m \leq s \leq n$, 则 $A_{jk} = 0$ 。

根据所得 A_{kj} 的取值 ($A_{kj} = 1$ 代表由 A_{kj} 表示的文字是完整文字, $A_{kj} = 0$ 代表 A_{kj} 表示的文字是残缺字), 确定 X_i' 的矩阵, 从而得到由 0-1 代表的残缺字和完整文字构成的矩阵。

本文规定相似度计算方式为:任取两个矩阵 $X_i'' =$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ \vdots & \vdots \\ A_{281} & A_{282} \end{bmatrix} \text{ 和 } X_j'' = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \\ \vdots & \vdots \\ B_{281} & B_{282} \end{bmatrix}, \text{ 其中 } A_{ij} = 0 \text{ 或 } 1,$$

$B_{ij} = 0$ 或 1, 规定相似度公式 $M = \sum_{i=1}^{28} |A_{ij} - B_{jk}|, j, k=1,2$ 且 $j \neq k$ 。这里 M 越小, 相似度越高, 两张碎片匹配度越高; 反之, M 越大, 相似度越低, 匹配度越低。

本文通过编程^[3]将纵切的中文碎片数据中的 19 张纸碎片分别进行上述处理, 得到相应的 19 个含有 0-1 的数量矩阵, 选取其中一个矩阵, 将剩下的 18 个矩阵分别与之左右进行匹配, 使相似度最高的矩阵与之匹配, 依此类推, 求解各矩阵间的 M 值, 搜索最大值, 将所得矩阵进行自动拼接得到初步拼接图像如图 1、图 2 所示。由于拼配过程中存在少许误差, 因此在对图片进行初步拼接之后, 通过适当的人工干预进行检验及修正, 进而得到完整的图像, 如图 3 和图 4 所示。

上屋楼叠嶂。城下清淮古汴。举手揖吴云。人与暮天俱远。魂断。城
下松江月满。被蓑衣市桥卖花。村里村北响鹌鹑。牛衣古柳卖黄
瓜。海棠珠缀一重重。清晓近帘栊。胭脂谁与匀淡。偏向脸边浓。小
瓶非
常记。二南依旧能诗。更有鲈鱼堪切脍。儿菜莫教知。自古相从休
论。何妨低唱微吟。天垂云重作春阴。坐中人半醉。窗外雪将深。双
翠绿
阴。娇眼横波眉黛翠。妙舞飘飏。掌上身轻姿媚。妆穿轻罗两凤。香
烟淡
拂双蛾。为谁流睡不归家。错认门前过马。
我动归思归去好。从来自己忘情。坐心消尽道心平。江南与塞北。何
处不堪行。雨阁阴。谁念素娥无主。何曾梦云雨。旧恨前欢。心事两无
凭。要知欲
见无由。痴心犹自。情人道。一声传语。风卷珠帘自上钩。萧萧乱
叶报新
秋。独携纤手上高楼。临水纵横回顾。归来转觉情怀动。梅笛烟
中闻几弄。秋阴重。西山雪淡云凝冻。凭高眺远。见长空万里。云无留
迹。桂
魄飞来光射处。冷浸一天秋碧。玉宇琼楼。乘鸾来去。人在清涼国。江
山如画。望中楼阁历历。省可清言挥玉尘。真须保霜全真。风流何似
道家
纯。不应当窗客。惟爱卓文君。自惜风流云雨散。关山有限情无限。待
君重
见寻芳伴。为说相思。目断西楼燕。莫恨黄花未吐。且教红粉相扶。酒
醒见
寻芳。看朱颜。俯仰人问今古。玉骨那愁瘴雾。冰姿自有仙风。海仙时
遣
探芳丛。倒挂绿毛么凤。
烟丝雨柳。凭君说与南窗。愿闻吴越报丰登。君王如有问。结
袜赖王
生。师唱谁家曲。宗风嗣阿谁。惜君拍板与门槌。我也逢场作戏。
莫相疑。
翠尊堪惜。印枕犹记。闲照晚妆残。残夜晓风。可恨相违。莫相疑。
能几日
。不知重会是何年。莫要细更重看。午夜风翻枕。三更月到床。
紫纹如
水。肌凉。何物与依归去。有残妆。金炉犹记断香残。情香更把宝
钗斜。更
闻处。余熏在。这一番。气味胜从前。碧梧枝上。新色绿
叶照林
光。竹篱茅舍出青黄。落梅水痕收。浅碧轻黄。浅洲。酒力渐消风
力软。晚
随。破帽多情却恋头。短影摇风。一枕伤春。归不去。风楼何
处。芳
尘归路。清发云鬓半白。露浮花乳轻圆。人间准拟更争妍。斗
处。芳
面。炙手无人傍屋头。萧萧晚雨。晚梧树。谁怜季子微裘。红粉

图 1 计算机处理后所得的图像复原图(中文)

fair of face.

The customer is always right. East, west, home's best. Life's not all beer and skittles. The devil looks after his own. Manners maketh man. Many a mickle makes a muckle. A man who is his own lawyer has a fool for his client.

You can't make a silk purse from a sow's ear. As thick as thieves. Clothes make the man. All that glitters is not gold. The pen is mightier than sword. Is fair and wise and good and gay. Make love not war. Devil take the hindmost. The female of the species is more deadly than the male. A place for everything and everything in its place. Hell hath no fury like a woman scorned. When in Rome, do as the Romans do. To err is human; to forgive divine. Enough is as good as a feast. People who live in glass houses shouldn't throw stones. Nature abhors a vacuum. Moderation in all things.

Everything comes to him who waits. Tomorrow is another day. Better to light a candle than to curse the darkness.

Two is company, but three's a crowd. It's the squeaky wheel that gets the grease. Please enjoy the pain which is unable to avoid. Don't teach your Grandma to suck eggs. He who lives by the sword shall die by the sword. Don't meet troubles half-way. Oil and water don't mix. All work and no play makes Jack a dull boy.

The best things in life are free. Finders keepers, losers weepers. There's no place like home. Speak softly and carry a big stick. Music has charms to soothe the savage breast. Ne'er cast a clout till May be out. There's no such thing as a free lunch. Nothing venture, nothing gain. He who can does, he who cannot, teaches. A stitch in time saves nine. The child is the father of the man. And a child that's born on the Sab-

图 2 计算机处理后所得的图像复原图(英文)

城上屋楼叠嶂。城下清淮古汴。举手揖吴云。人与暮天俱远。魂断。城
下松江月满。被蓑衣市桥卖花。村里村北响鹌鹑。牛衣古柳卖黄
瓜。海棠珠缀一重重。清晓近帘栊。胭脂谁与匀淡。偏向脸边浓。小
瓶非
常记。二南依旧能诗。更有鲈鱼堪切脍。儿菜莫教知。自古相从休
论。何妨低唱微吟。天垂云重作春阴。坐中人半醉。窗外雪将深。双
翠绿
阴。娇眼横波眉黛翠。妙舞飘飏。掌上身轻姿媚。妆穿轻罗两凤。香
烟淡
拂双蛾。为谁流睡不归家。错认门前过马。
我动归思归去好。从来自己忘情。坐心消尽道心平。江南与塞北。何
处不堪行。雨阁阴。谁念素娥无主。何曾梦云雨。旧恨前欢。心事两无
凭。要知欲
见无由。痴心犹自。情人道。一声传语。风卷珠帘自上钩。萧萧乱
叶报新
秋。独携纤手上高楼。临水纵横回顾。归来转觉情怀动。梅笛烟
中闻几弄。秋阴重。西山雪淡云凝冻。凭高眺远。见长空万里。云无留
迹。桂
魄飞来光射处。冷浸一天秋碧。玉宇琼楼。乘鸾来去。人在清涼国。江
山如画。望中楼阁历历。省可清言挥玉尘。真须保霜全真。风流何似
道家
纯。不应当窗客。惟爱卓文君。自惜风流云雨散。关山有限情无限。待
君重
见寻芳伴。为说相思。目断西楼燕。莫恨黄花未吐。且教红粉相扶。酒
醒见
寻芳。看朱颜。俯仰人问今古。玉骨那愁瘴雾。冰姿自有仙风。海仙时
遣
探芳丛。倒挂绿毛么凤。
烟丝雨柳。凭君说与南窗。愿闻吴越报丰登。君王如有问。结
袜赖王
生。师唱谁家曲。宗风嗣阿谁。惜君拍板与门槌。我也逢场作戏。
莫相疑。
翠尊堪惜。印枕犹记。闲照晚妆残。残夜晓风。可恨相违。莫相疑。
能几日
。不知重会是何年。莫要细更重看。午夜风翻枕。三更月到床。
紫纹如
水。肌凉。何物与依归去。有残妆。金炉犹记断香残。情香更把宝
钗斜。更
闻处。余熏在。这一番。气味胜从前。碧梧枝上。新色绿
叶照林
光。竹篱茅舍出青黄。落梅水痕收。浅碧轻黄。浅洲。酒力渐消风
力软。晚
随。破帽多情却恋头。短影摇风。一枕伤春。归不去。风楼何
处。芳
尘归路。清发云鬓半白。露浮花乳轻圆。人间准拟更争妍。斗
处。芳
面。炙手无人傍屋头。萧萧晚雨。晚梧树。谁怜季子微裘。红粉

烟丝雨柳。凭君说与南窗。愿闻吴越报丰登。君王如有问。结
袜赖王
生。师唱谁家曲。宗风嗣阿谁。惜君拍板与门槌。我也逢场作戏。
莫相疑。
翠尊堪惜。印枕犹记。闲照晚妆残。残夜晓风。可恨相违。莫相疑。
能几日
。不知重会是何年。莫要细更重看。午夜风翻枕。三更月到床。
紫纹如
水。肌凉。何物与依归去。有残妆。金炉犹记断香残。情香更把宝
钗斜。更
闻处。余熏在。这一番。气味胜从前。碧梧枝上。新色绿
叶照林
光。竹篱茅舍出青黄。落梅水痕收。浅碧轻黄。浅洲。酒力渐消风
力软。晚
随。破帽多情却恋头。短影摇风。一枕伤春。归不去。风楼何
处。芳
尘归路。清发云鬓半白。露浮花乳轻圆。人间准拟更争妍。斗
处。芳
面。炙手无人傍屋头。萧萧晚雨。晚梧树。谁怜季子微裘。红粉

烟丝雨柳。凭君说与南窗。愿闻吴越报丰登。君王如有问。结
袜赖王
生。师唱谁家曲。宗风嗣阿谁。惜君拍板与门槌。我也逢场作戏。
莫相疑。
翠尊堪惜。印枕犹记。闲照晚妆残。残夜晓风。可恨相违。莫相疑。
能几日
。不知重会是何年。莫要细更重看。午夜风翻枕。三更月到床。
紫纹如
水。肌凉。何物与依归去。有残妆。金炉犹记断香残。情香更把宝
钗斜。更
闻处。余熏在。这一番。气味胜从前。碧梧枝上。新色绿
叶照林
光。竹篱茅舍出青黄。落梅水痕收。浅碧轻黄。浅洲。酒力渐消风
力软。晚
随。破帽多情却恋头。短影摇风。一枕伤春。归不去。风楼何
处。芳
尘归路。清发云鬓半白。露浮花乳轻圆。人间准拟更争妍。斗
处。芳
面。炙手无人傍屋头。萧萧晚雨。晚梧树。谁怜季子微裘。红粉

图 3 经人工干预后得到的图像复原图(中文)

fair of face.

The customer is always right. East, west, home's best. Life's not all beer and skittles. The devil looks after his own. Manners maketh man. Many a mickle makes a muckle. A man who is his own lawyer has a fool for his client.

You can't make a silk purse from a sow's ear. As thick as thieves. Clothes make the man. All that glitters is not gold. The pen is mightier than sword. Is fair and wise and good and gay. Make love not war. Devil take the hindmost. The female of the species is more deadly than the male. A place for everything and everything in its place. Hell hath no fury like a woman scorned. When in Rome, do as the Romans do. To err is human; to forgive divine. Enough is as good as a feast. People who live in glass houses shouldn't throw stones. Nature abhors a vacuum. Moderation in all things.

Everything comes to him who waits. Tomorrow is another day. Better to light a candle than to curse the darkness.

Two is company, but three's a crowd. It's the squeaky wheel that gets the grease. Please enjoy the pain which is unable to avoid. Don't teach your Grandma to suck eggs. He who lives by the sword shall die by the sword. Don't meet troubles half-way. Oil and water don't mix. All work and no play makes Jack a dull boy.

The best things in life are free. Finders keepers, losers weepers. There's no place like home. Speak softly and carry a big stick. Music has charms to soothe the savage breast. Ne'er cast a clout till May be out. There's no such thing as a free lunch. Nothing venture, nothing gain. He who can does, he who cannot, teaches. A stitch in time saves nine. The child is the father of the man. And a child that's born on the Sab-

图 4 经人工干预后得到的图像复原图(英文)

表1—表4分别列出了不同语言文本格式下图像碎片经计算机处理和人工干预的复原顺序。可以看出,0-1模型能够较为精准地将图像复原,虽然存在少许误差,但是由于首尾

不契合的偏差很容易被人工发现并修正,因此这样的误差能够在人为条件下避免。同时,该算法对英文文本具有一定的鲁棒性。

表1 经计算机处理后得到的图像复原碎片顺序(中文)

复原编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
原始编号	6	8	14	12	15	3	10	2	16	1	4	5	9	13	18	11	7	17	0

表2 经人工干预后得到的图像复原碎片顺序(中文)

复原编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
原始编号	6	8	14	12	15	3	10	2	16	1	4	5	9	13	18	11	7	17	0

表3 经计算机处理后得到的图像复原碎片顺序(英文)

复原编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
原始编号	3	6	2	7	15	18	11	0	5	1	9	13	10	8	12	14	17	16	4

表4 经人工干预后得到的图像复原碎片顺序(英文)

复原编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
原始编号	3	6	2	7	15	18	11	0	5	1	9	13	10	8	12	14	17	16	4

2.2 基于 Pearson 相关系数的灰度匹配模型

由于0-1匹配模型中存在0~255中选定的一个值,在划定0-1界限时无标准可循,使得模型的计算存在一定的偏差,因此提出灰度匹配模型对已有的0-1模型进行改进,同时对结果做进一步验证。

基于灰度的算法的基本思想是,假设两幅图像对应点的小邻域内具有相似的灰度分布,以统计的观点将图像看作一维信号,采用相关的方法利用图像上的局部相关性来确定两个对应点间的匹配。基于灰度的匹配方法是把一幅图像中某一像素的灰度邻域作为模板,在另一幅图像中搜索具有相同(或相似)灰度值分布的对应点邻域,从而实现两幅图像的匹配。在搜索过程中,本文以相关系数作为两个搜索邻域间的相似性测度。当搜索区域中的元素使相似性准则最大化时,认为元素是匹配的^[4]。Pearson相关系数^[5]反映了变量之间线性相关的密切程度,是评判变量是否具有线性关系的重要依据。

Pearson相关系数常用 r 表示,其计算公式为:

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

相关系数 $r \in [-1, 1]$ 。当 $r > 0$ 时,称为正线性相关;当 $r = 0$ 时,称为线性无关;当 $r < 0$ 时,称为负线性相关。通过计算两张碎片相邻边的相关系数来确定图像的相对位置。

具体的算法流程如下。

首先,将19张图片转换成灰度值,得到如下矩阵:

$$X^i = \begin{bmatrix} a_{11}^i \\ \vdots \\ a_{m1}^i \end{bmatrix} = \begin{bmatrix} a_{11}^i & \cdots & a_{1n}^i \\ \vdots & \ddots & \vdots \\ a_{m1}^i & \cdots & a_{mn}^i \end{bmatrix}$$

其中, $i=1,2,\dots,19,n=1,2,\dots,72,m=1,2,\dots,1980,a_{mm}^i=0,1,2,\dots,255$ 。

其次,将上述矩阵的首列和尾列提取出来得到19个新的 72×1980 矩阵,即

$$X_p^i = \begin{bmatrix} a_{11}^i & a_{172}^i \\ \vdots & \vdots \\ a_{m1}^i & a_{m72}^i \end{bmatrix}, i=1,2,\dots,19, m=1,2,\dots,1980$$

计算每一个矩阵的奇数项与剩余18个矩阵的偶数项的相关系数,即为每一个矩阵左边与剩余矩阵右边的相关系数,以此判断两矩阵的相关程度,得到一个 38×38 的 R 矩阵。同一偶数行表示一个碎片矩阵 X_p^i 左边与其他各项右边列的相关系数,即为某一碎片首列与其余碎片末列的相关程度;同一奇数列表示一个碎片矩阵 X_p^i 右边与其他各项左边列的

相关系数, R 矩阵即 $R = \begin{bmatrix} r_{11} & \cdots & r_{138} \\ \vdots & \ddots & \vdots \\ r_{381} & \cdots & r_{3838} \end{bmatrix}$,其中 $r =$

$$\frac{n \sum a_{i1}^i a_{j2}^j - \sum a_{i1}^i \sum a_{j2}^j}{\sqrt{[n \sum (a_{i1}^i)^2 - (\sum a_{i1}^i)^2][n \sum (a_{j2}^j)^2 - (\sum a_{j2}^j)^2]}}, i, j = 1, 2, \dots, 19。$$

最后,根据矩阵进行判断,首先选取 R 中全为NaN值的一行,由于分母中的方差为:

$$Var = [n \sum (a_{jk}^j)^2 - (\sum a_{jk}^j)^2], k=1,2$$

若某列全为相等数,则其方差为0,若某行 r 全为NaN值,定义其为碎片的起始碎片,再通过找出 R 矩阵中该行与其他各列的相关系数的最大值,来依次找出与之契合得最好的其他列,依此类推即可得到完整的图片。

以中文为例,计算首尾列合成矩阵的相关系数,进而得到 R 矩阵,根据其特征可以看出第17行所有值均为NaN,代表第17个矩阵的左边行全为相同值,则第17个碎片被定义为起始碎片,再搜索与第17个矩阵右边列所求得相关系数最高的矩阵,即找到第17个矩阵右边列与其余列左边列的线性相关性最高的矩阵,得到与起始碎片匹配的碎片,从而复原出整个图形。经灰度匹配模型处理后,图像复原碎片顺序如表5、表6所列。

表5 经灰度匹配模型处理后的图像复原碎片顺序(中文)

复原编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
原始编号	6	8	14	12	15	3	10	2	16	1	4	5	9	13	18	11	7	17	0

表6 经灰度匹配模型处理后的图像复原碎片顺序(英文)

复原编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
原始编号	3	6	2	7	15	18	11	0	5	1	9	13	10	8	12	14	17	16	4

由此可以看出,相较 0-1 匹配模型,灰度匹配模型有更高的精准度。但总体而言,两个模型的正确率都较高,均能够达到很好的图像复原效果。

3 横切和纵切碎纸方向的拼接复原研究

针对既有横切又有纵切的情况,根据中文印刷字体几乎同在一个水平线的特点,本文通过扫描得到每一张碎片中每前 3 行字的 3 条标记线,来得到每一张碎片的位置矩阵;对各标记矩阵进行系统聚类^[6],得到一个 11 类的分类,该分类反映了各个标记线的类别特征,同一类的碎片代表该类碎片处于同一水平面上;对分类后的碎片进行人工干预^[7]精确分类,再应用灰度匹配模型对同一类中的碎片进行拼接,对复原后的横向图像进行竖向拼接得到复原全图。

3.1 预处理

建立模型之前,对得到的灰度值进行扫描从而得到标记线的位置矩阵,横向扫描每行中的灰度值,记下灰度值从 0 到全部为 0 的位置,即为所求标记值,每张碎纸片可得到如下矩阵:

$$B = \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix}$$

3.2 基于系统聚类的碎纸文件拼接复原研究

系统聚类开始时,每个样本自成一类,类与类之间的距离与样本之间的距离是相同的;然后,在所有的类中选择距离最小的两个类合并成一个新类,并计算所得新类与其他各类的距离;接着再将距离最近的两类合并,这样每次合并两类,直至将所有的样本都合并成一类为止。

系统聚类法的聚类原则决定于样本间的距离(或相似系数)及类间距离的定义,类间距离的不同定义产生了不同的系统聚类分析方法。通过比较多种聚类方法,我们认为离差平方和法(WALD 方法)所得的答案最佳,因此选用离差平方和法对位置矩阵进行聚类分析。基本步骤如下:用 d_{ij} 表示矩阵 $X_{(i)}$ 和 $X_{(j)}$ 之间的距离,矩阵间的亲疏关系采用相似系数 C_{ij} 表示,令 $d_{ij} = 1 - |C_{ij}|$,用 D_{ij} 表示类 G_i 和 G_j 间的距离。

将 n 个矩阵分成 k 类,记为 $G_1, G_2, \dots, G_k, n_i$ 表示 G_i 类的矩阵个数, $\bar{X}^{(i)}$ 表示 G_i 的重心, $X_{(i)}^{(t)}$ 表示 G_i 中的第 i 个矩阵 ($i=1, \dots, n_i$),则 G_i 中矩阵的离差平方和为:

$$W_i = \sum_{t=1}^{n_i} (X_{(i)}^{(t)} - \bar{X}^{(i)})' (X_{(i)}^{(t)} - \bar{X}^{(i)})$$

其中, $X_{(i)}^{(t)}$ 和 $\bar{X}^{(i)}$ 为 m 维向量, W_i 为一数值 ($t=1, 2, \dots, k$)。

k 个类的总离差平方和为:

$$W = \sum_{i=1}^k W_i = \sum_{i=1}^k \sum_{t=1}^{n_i} (X_{(i)}^{(t)} - \bar{X}^{(i)})' (X_{(i)}^{(t)} - \bar{X}^{(i)})$$

当 k 固定时,选择使 W 达到极小的分类。

用 WALD 方法把某两类合并后增加的离差平方和看作类间的平方距离,即令 $D_{pq}^2 = W_r - (W_p + W_q)$ 表示类 G_p 和 G_q 的平方距离,其中 $G_r = \{G_p, G_q\}$, W_r, W_p, W_q 分别为 G_p, G_q, G_r 类中样品的离差平方和。利用 W_r 的定义可得:

$$\begin{aligned} W_r &= \sum_{t=1}^{n_r} (X_{(r)}^{(t)} - \bar{X}^{(r)})' (X_{(r)}^{(t)} - \bar{X}^{(r)}) \\ &= \sum_{i=1}^{n_p} (X_{(i)}^{(p)} - \bar{X}^{(r)})' (X_{(i)}^{(p)} - \bar{X}^{(r)}) + \sum_{i=1}^{n_q} (X_{(i)}^{(q)} - \bar{X}^{(r)})' (X_{(i)}^{(q)} - \bar{X}^{(r)}) \end{aligned}$$

其中, $\bar{X}^{(r)} = \frac{1}{n_r} [n_p \bar{X}^{(p)} + n_q \bar{X}^{(q)}]$ 。经整理可得:

$$D_{pq}^2 = \frac{n_p n_q}{n_r} (\bar{X}^{(p)} + \bar{X}^{(q)})' (\bar{X}^{(p)} + \bar{X}^{(q)})$$

当矩阵间的距离采用欧氏距离时,上式可表示为:

$$D_{pq}^2 = \frac{n_p n_q}{n_r} d_{pq}^2$$

其中, d_{pq}^2 表示 G_p 和 G_q 的重心 $\bar{X}^{(p)}$ 与 $\bar{X}^{(q)}$ 的平方距离,即 $d_{pq}^2 = d^2(\bar{X}^{(p)}, \bar{X}^{(q)})$ 。当 G_p 和 G_q 合并为 G_r 后, G_r 与其他类 G_k 的距离有如下递推公式:

$$D_{rk}^2 = \frac{n_k + n_p}{n_r + n_k} D_{pk}^2 + \frac{n_k + n_q}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$$

由于系统聚类模型所得的关联度包含了位置矩阵 3 个标记线的位置,能够比较全面地反映碎片所处的水平位置,因此聚类分析从统计的角度对各位置矩阵进行划分以能够客观地反映该行文字的不同水平。

本文采用 R 语言^[8]编程通过系统聚类将其分为 11 类,得到如表 7 所列的分类表格。

通过系统聚类,将 209 张碎片分成了 11 个水平上的类别,但可以看到,虽然系统聚类有效的位置矩阵间的距离可以通过运算得到,但其并未平均地将其分为 11 类,如通过系统聚类后第 7 类含有 31 个位置矩阵,也就是说把 31 个碎纸片的标记线分为同一水平,显然比预想的要多。这说明可能存在细微差别的类别,因此可多次通过系统聚类,将所含位置矩阵较多的项进行进一步细分,得到更加准确的聚类。此处将第 7 类、第 8 类、第 10 类进行二次系统聚类,将类别拓展为较为精确的 14 类。通过人工干预,对已分类图像进行简单筛选,剔除明显差异的图片,并对某些类别进行合并,得到比较精确的分类,如表 8 所列。

表 7 初步的系统聚类分析结果(中文)

类别	所属类别碎片号
第 1 类	002,011,022,028,049,054,057,065,091,095,118,129,141,143,178,186,188,190,192,071
第 2 类	004,013,040,062,101,102,113,114,117,119,123,140,146,155,182,185,194,207
第 3 类	003,012,029,031,039,051,073,082,107,115,128,134,135,159,160,169,176,199,203
第 4 类	001,016,018,023,026,030,041,050,076,086,087,100,120,142,147,168,179,191,195
第 5 类	021,066,106,109,110,139,145,150,157,173,181,184,187,197,204
第 6 类	014,089
第 7 类	000,007,032,045,052,053,056,068,070,085,093,127,137,138,153,158,165,166,170,174,175,196,208
第 8 类	006,015,017,019,020,027,033,060,063,067,069,072,078,079,080,083,096,099,116,131,132,133,152,156,162,163,177,198,200,202,205
第 9 类	008,009,024,025,034,035,038,042,043,046,074,098,104,111,171,172,180,201,206
第 10 类	005,010,037,044,048,055,059,075,098,104,111,171,172,180,201,206
第 11 类	036,047,058,061,077,092,103,105,151,154

表 8 经人工干涉后的聚类分析结果(中文)

类别	所属类别碎片号
第 1 类	002,011,022,028,049,054,057,065,091,095,118,129,141,143,178,186,188,190,192
第 2 类	004,040,089,101,102,108,113,114,117,119,123,140,146,151,154,155,185,194,207
第 3 类	003,012,029,031,039,051,073,082,107,115,128,134,135,159,160,169,176,199,203
第 4 类	001,018,023,026,030,041,050,062,076,086,087,100,120,142,147,168,179,191,195
第 5 类	013,016,021,066,106,109,110,125,139,145,150,157,173,181,182,184,187,197,204
第 6 类	000,007,032,045,053,056,068,070,093,126,137,138,153,158,166,174,175,196,208
第 7 类	015,017,027,033,060,071,080,083,085,132,133,152,156,165,170,198,200,202,205
第 8 类	034,042,043,047,058,077,084,090,094,097,112,121,124,127,136,144,149,164,183
第 9 类	008,009,024,025,035,038,046,074,077,081,088,105,122,130,148,161,167,189,193
第 10 类	006,019,020,036,052,061,063,067,069,072,078,079,096,099,116,131,162,163,177
第 11 类	005,010,014,037,044,048,055,059,075,092,098,104,111,171,172,180,201,206

将经过聚类后的碎纸片置于同一水平线,运用与第 2 节类似的灰度匹配模型进行类别内的匹配,在此不再赘述。聚类分析后经灰度匹配的图像复原碎片顺序如表 9 所列。复原结果如图 5 所示。

表 9 聚类分析后经灰度匹配的图像原碎片顺序(部分)

复原序号	1	2	3	4	5	6	7
原始编号	49	54	65	143	186	2	57
复原序号	8	9	10	11	12	13	14
原始编号	192	178	118	190	95	11	22
复原序号	...	...	...	...	...	...	...
原始编号	...	...	...	...	...	...	...
复原序号	204	205	206	207	208	209	
原始编号	4	101	113	194	119	123	

便邮。温香熟类。船慢云鬓垂两耳。多谢春工。不是花红是玉红。一鞭楼
 桃堡素口。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。背灯偷拭残妆粉。春事闹幽芳草
 歇。客里风光。又过清明节。小院黄昏人忆别。落红处处闻啼鸟。岁云暮。
 须早计。要梅装。故乡归去千里。佳处便迟留。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。莲
 志莫相违。西园路。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。
 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。应须只
 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。香生雾縠见纤柔。倦宵眠耿耿。
 自觉功名便更疏。若问使君才与水。何如。占得人间一昧愚。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。记取钗头新样。夏将分付东邻子。西家山边白
 鹭飞。散花洲外片帆微。桃花流水鳜鱼肥。主人谪小。欲向东风先醉倒。
 已闻君家。且更从容等待他。愿我已无当世望。似君须向古人求。岁寒松
 柏青堪秋。
 水涵空。山照市。西汉二疏乡里。新白发。旧黄金。故人恩义深。谁
 道东阳都瘦损。偶然点漆精神。瑶林终自随风尘。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。雕钱玉果。
 利市平分沾四座。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
 一壑琴。一壶酒。一溪云。相如未老。梁苑犹能陪俊少。莫愁闲愁。且折

图 5 中文文本复原图

由图 5 可以看出,由此顺序生成的复原图完整准确,因此所提模型得到的图像复原图较为精确。

结束语 本文首先针对单一方向的纵切问题进行研究。运用 0-1 匹配模型,根据碎纸片左右两边的文字是否残缺,将

碎纸片转化为特定矩阵,将所得矩阵通过相似度原理进行匹配,拼接出图像。这种方法将复杂问题简单化,易于理解,但需要进行部分人工干预。因此,本文针对这一不足采用灰度匹配思想加以改进,以统计的观点将图像看作二维信号,采用统计相关的方法寻找信号间的相关匹配。同时采用 Pearson 相关系数作为两信号的相关函数,评价它们的相似性以确定同名点,结果证明此法可行且不需要对结果进行人工干预,对于中文文本碎片和英文文本碎片的拼接复原效果都很好。最后,本文针对既有纵切又有横切的中文文本碎片拼接复原问题采用聚类分析方法进行研究,有效地解决了科学研究中多因素、多指标的分类问题。本文采用离差平方和法计算类与类之间的距离,具有一定的优越性。未来我们的研究方向还可以进一步扩展,如在既有纵切又有横切的英文文本碎片拼接复原方面依据其字符的特殊性进行精确的操作以及在更为复杂的双面情况下对碎片进行拼接复原。

参 考 文 献

- [1] 贾海燕. 碎纸自动拼接关键技术研究[D]. 长沙:国防科学技术大学,2005.
- [2] 张桂华,冯艳波,陆卫东. 图像处理的灰度化及特征区域的获取[J]. 齐齐哈尔大学学报(自然科学版),2007,23(4):49-52.
- [3] SVOBODA T, KYBIC J, HLAVAC V. Image Processing, Analysis & Machine Vision- A MATLAB Companion[M]. Thomson Learning, 2007.
- [4] 乔杰忠,吴中雷. 基于灰度匹配算法在图像处理中的研究[J]. 装备制造技术,2009(10):42-43.
- [5] YANG N, JI D G, LI S Q. The Application of Pearson Correlation Analysis Method in Air Quality Analysis of Beijing-Tianjin-Hebei Region[J]. Agricultural Science & Technology, 2015(3):590-592.
- [6] 胡雷芳. 五种常用系统聚类分析方法及其比较[J]. 统计科学与实践,2007(4):11-13.
- [7] 卢健. 少量人工干预的碎纸片拼接[J]. 自动化与仪器仪表,2015(10):51-53.
- [8] 贾里德 P·兰德. R 语言[M]. 北京:机械工业出版社,2015.