doi:10.16055/j.issn.1672-058X.2015.0008.010

规则碎片拼接算法

李　猛[1], 刘家保[2]
(1.合肥市统计局, 合肥 230071;2.安徽新华学院 公共课教学部, 合肥 230088)

摘　要:针对通过纵向切割、纵横交错切割等方法得到的碎纸片模型进行拼接复原,以碎片图形边缘像素点的匹配程度为判断依据,提出一种有效的算法,很好地解决了纵横切割所得到的规则碎片模型的拼接问题;在误差调整方面,提出了以计算机辅助人工进行较正取代单纯的人工较正,并给出了辅助算法;对非纵横切所得规则图片的拼接也从理论上提出了相应的算法。

关键词:匹配率;碎片行分组;行匹配率;贪心法
中图分类号:0157.5　　**文献标识码**:A　　**文章编号**:1672-058X(2015)08-0043-06

如今世界上纸质碎片的拼接技术分人工和计算机自动拼接两种。由人工完成的拼接复原的正确率高,但效率非常低,当碎片数量多时,人工拼接难以在短时间里完成。随着计算机技术的发展,现代碎片拼接研究方向主要集中在计算机自动拼接[1-6],但在处理碎片拼接时,当今主流是采用形状匹配、边缘比较、数据库匹配等技术,对硬件运算能力和储存能力都有着极高的要求。而很多情况下,并不需要处理太过复杂的碎片,如一般办公室通常采用的是碎纸机碎纸,所得的碎片形状比较规则,拼接不需要考虑形状匹配等因素,因此使用主流的对高硬件运行速度和存储空间消耗的算法就比较浪费。因此,以文件碎片边缘的黑白色匹配程度为依据,结合动态规划[8]理论,以普通家用电脑为硬件基础,提出了一种简洁有效的规则碎片自动拼接算法。

1　相关设定与定义

1.1　设　定

(1) 无切割误差。碎纸机在切割文件时没有误差,切出来的碎片中文字方向平行于上下边缘,同一文件碎片大小完全一致;

(2) 无打印误差。纸质文件按同一标准打印出来,行号、字号、行间矩完全一致,中文字体完全一致,英文字体完全一致;

(3) 无物理误差。碎片图片无污迹,无毛边等干扰情况;

(4) 论文中使用的误差为根据拼接实验中碎片的数据选定,可根据实际碎片数据进行更改。

1.2　定　义

(1) 两文件碎片之间的匹配率[5]。读取每一个碎纸片图片文件为数据矩阵,选取适当阀值,将其用二值法化为 0-1 矩阵。任取碎纸片 i 和碎纸片 j 进行比较, 记:

$$\mathrm{rmatch}(i,j)=\frac{\text{第 } i \text{ 个碎纸片右边缘与第 } j \text{ 个碎纸片左边缘象素点数据相同的个数}}{\text{单个碎纸片象素点总行数}}$$

收稿日期:2014-10-01;**修回日期**:2014-11-28.

作者简介:李猛(1982-),男,安徽合肥人,硕士,从事应用数字研究.

(2) 已分组文件之间的行匹配率。在以将碎片文件分组排序的基础上,任取第 i 组和第 j 组进行比较,记

$$\mathrm{cmatch}(i,j)=\frac{\text{第 } i \text{ 组碎纸片最下边缘与第 } j \text{ 组碎纸片上边缘象素点数据相同的个数}}{\text{整张纸一行全部象素点个数}}$$

(3) 碎片文字行高估计值。碎片所在行的每行文字所占碎片文件象素行数的最大值与最小值的估计值,简称为文字行高估计值。根据一碎片包含文字行数,每一碎片的行高估计值约有 4~6 个。如图 1 中碎片 000.bmp 的行高估计值可记为 0,26,56,96,124,165。

图 1 待拼接碎片 000.bmp

2 碎片的按行分组

此时,仍可以碎片之间的匹配率为标准,利用贪心法[8]逐行拼接后再按行匹配率进行排序。但通过对中国 2013 年数学建模竞赛 B 题附件 3 提供碎片实际拼接结果发现,直接拼接误差非常大,正确率仅为 30.81%。因此,在此采用先按行分组,再进行拼接的方法来降低误差。

在假定中,碎片文件的四边平行于原纸张相应边缘,且其中文字是按同一标准逐行打印而成的。所以从理论上,同行碎片文件的同行文字的行高与行间距都应该是对应相等的(图 2)。

从图 2 中可以看出 000.bmp, 007.bmp, 045.bmp 行间矩基本相同,因此可以归成一类。且因为笔划、字形、灰度等情况可能会产生部分误差,通过图形对比,可把误差额度设置为(-5,5)。

分组算法:

第 1 步,按每一文件第一行象素全为 255,和有多行不为 255 分类,为避免特殊情况产生误差,仅有前面连续 2 行以内不全为 255 的情况单独归为一类。

第 2 步,将按顺序选择一个未分组文件单独分为一组,提取每一行汉字所占象素行数的最小值和最大值,可得到 3 对数据。

第 3 步,以上述 3 对数据为基础,设为上述碎片文件的初始文字行高估计值,以同一类中其它文件的相应数据进行比较,选取方差小于阀值 k 的文件,将其加入 001 文件所在组,并将碎片文字行高估计值改为已加入该组的所有碎片文件相应数据的平均值。k 的值根据实际需要选定,如根据允许误差范围(-5,5),则可取阀值 $k=6\times25=150$。

第 4 步,以新得到的估计值为标准,重复步骤 3,筛选出所有符合条件的文件,归为一组。

第 5 步,重复步骤 2、3、4,直至完成全部分组。

第 6 步,根据整张纸片切割的行数 m 和列数 n,对以上各组数据进行现次处理,按每组包含文件个数从大到小排列,以前 11 为基础。包含文件个数超过列数 n 的,以所在组每一碎片文件相应数据与碎片文字行

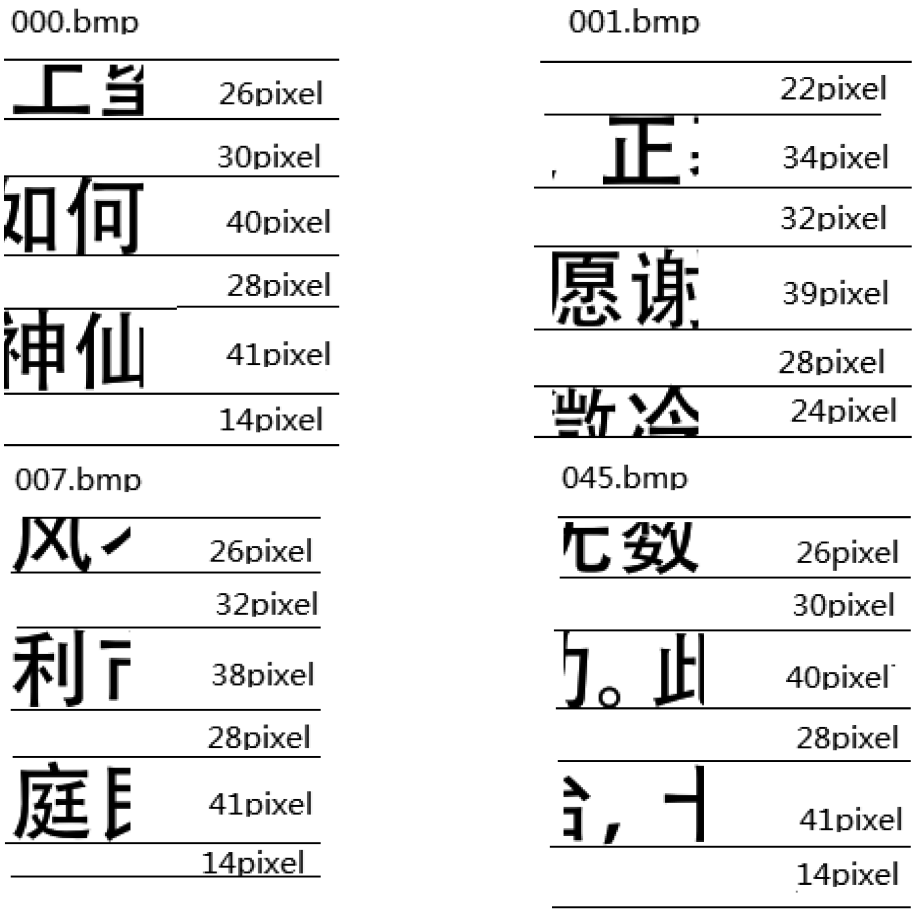

图 2　待拼接碎片示例图

高估计值比较,选出与标准差最小的 n 个文件,其余文件从组中移出。小于列数 n 的,以该组的碎片文字行高估计值为基础,将未分组文件和包含文件个数不超过 3 个的组包含文件依次比较,选出方差最小的若干个文件加入组中,补足 n 个。

以图 2 中 000.bmp, 001.bmp, 007.bmp, 045.bmp 4 个文件为例。

(1) 000.bmp、007.bmp、045.bmp 同属于多行不全为 255 的一类。

(2) 按编号顺序从 000.bmp 开始分组处理,此时分组数为 0,所以 001 不属于任何一组,将其单独分为 1 组{000.bmp}。提取碎片文件 000.bmp 每一行汉字所占象素行数的最小值和最大值,得到一组数据 1、26、57、96、125、165,将其设为组{000.bmp}对应碎片所在行的相应数据的初始估计值。经检测,001.bmp 属于第一行象素值全为 255 这一类别,此时,没有与其同一组的碎片文件,因此 001.bmp 单独分为一组{001.bmp}。

(3) 经检测 007.bmp 每一行汉字所占象素行数的最小值和最大值为 1、26、59、96、125、165,与初始估计值相比较,方差为 4 小于阀值 k, 所以将 007.bmp 并入组{000.bmp},取新的碎片文字行高估计值为 00.bmp 和 001.bmp 相应数据的平均值,1、26、58、96、125、165。

(4) 依次将 045.bmp 相应数据与之比较,最终将文件分为两组{000.bmp, 001.bmp, 045.bmp}和{001.bmp}。

每组内碎纸片的排序为:以分组数据为基础,对每一组文件,以同组文件之间的匹配率为标准,应用贪心法,将同组碎片文件排序。对已排序数据,以不同组文件之间的组匹配率为基础,应用贪心法,将各组进行排序。

3　人工较正和计算机辅助人工较正

在碎片过多的情况下,计算机拼接可能会由于打印、笔划、字型的不同等原因产生一定的误差,此时就

需要通过手工对已处理结果进行较正。但受视力、思考速度等身体条件影响,在需要较正碎片量较大的情况下,人工较正可能会产生速度较慢、有一定误差等问题。

对此,在人工较正过程中,仍可以使用计算机进行辅助,用以加快较正速度,减少较正工作量。在处理过程中,可以采用遍历法,仍以碎片之间的匹配律为基础,让计算机按顺序推荐出与错拼的文件最匹配的5个(根据需要可自由选择数目)文件,便于快速较正误差。

4 拼接实验

试验以中国2013年数学建模竞赛B题提供文件为实验对象,使用MATLAB程序作为编程工具[7]。

4.1 仅纵向切割的碎片拼接

此时所有碎片均为一行,跳过行分组阶段,直接应用贪心法进行排序,即可直接得到正确顺序。此种情况下因计算匹配率时相应象素点数量较大,因而所得匹配率数据值可信度高,拼接结果错误率极小,一般不需要人工较正。

4.2 纵横切碎片拼接的问题

第1步,把所有碎片按前面的分组算法分组。以附件3为例,得表1:

表1 分组后得到的碎片组合

组号	碎纸片序号
01	000,007,032,045,053,056,068,070,093,126,137,138,153,158,166,174,175,196,208
02	001,018,023,026,030,041,050,062,076,086,087,100,120,142,147,168,179,191,195
03	002,011,022,028,049,054,057,065,091,095,118,129,141,143,178,186,188,190,192
04	003,012,014,031,039,051,073,082,107,115,128,134,135,159,160,169,176,199,203
05	004,040,089,101,102,108,113,114,117,119,123,140,146,151,154,155,185,194,207
06	005,010,029,037,044,048,055,059,064,075,092,094,104,111,171,172,180,098,206
07	006,019,020,036,052,061,063,067,069,072,078,079,096,099,116,131,162,163,177
08	008,009,024,025,035,038,046,074,081,088,103,105,122,130,148,161,167,189,193
09	034,043,042,047,058,077,084,090,097,112,121,124,127,136,144,149,164,183,201
10	015,017,027,033,060,071,080,083,085,132,133,152,156,165,170,198,200,202,205
11	013,016,021,066,106,109,110,125,139,145,150,157,173,181,182,184,187,197,204

此处根据实际情况可适当调整阀值、误差区域。

第2步,对每一组分别排序,得表2:

表2 每一分组内部的碎片排序结果

组号	碎纸片序号																		
01	007	208	138	158	126	068	175	045	174	000	137	053	056	093	153	070	166	032	196
02	168	100	076	062	142	030	041	023	147	191	050	179	120	086	195	026	001	087	018
03	049	054	065	143	186	002	057	192	178	118	190	095	011	022	129	028	091	188	141
04	014	128	003	159	082	199	135	012	073	160	203	169	134	039	031	051	107	115	176
05	089	146	102	154	114	040	151	207	155	140	185	108	117	004	101	113	194	119	123
06	029	064	111	005	092	180	048	037	075	055	044	206	010	104	098	172	171	059	094
07	061	019	078	067	069	099	162	096	131	079	063	116	163	072	006	177	020	052	036
08	038	148	046	161	024	035	081	189	122	103	130	193	088	167	025	008	009	105	074
09	034	084	183	090	047	121	042	124	144	077	112	149	097	136	164	127	058	043	201
10	071	156	083	132	200	017	080	033	202	198	015	133	170	205	085	152	165	027	060
11	125	013	182	109	197	016	184	110	187	066	106	150	021	173	157	181	204	139	145

第 3 步,人工较正。纵横切割的碎片在计算匹配率时进行比较的象素点较少,因此出错的可能性相对仅纵切的情况较高,偶尔需要进行人工较正。

通过上述结果,观察可得,上述结果中 094 号碎片与 201 号碎片不在正确的位置上,因为只有两个错误碎片,将 094 和 201 交换所在组重新排序即可。

若错误碎片也可采用计算机辅助,利用程序依次将 094 以外的其它碎片遍历一遍,得到与 094 匹配率最高的 5 个文件 034、058、090、149、164,人工对比发现 034 号碎片与 094 号拼接最符合要求。同理可找到与 201 最适碎片 005。

第 4 步,行间排序。利用两行之间的行匹配率,应用贪心法进行排序,得到表 3 正确顺序。

表 3　附件 3 碎片附原顺序

049	054	065	143	186	002	057	192	178	118	190	095	011	022	129	028	091	188	141
061	019	078	067	069	099	162	096	131	079	063	116	163	072	006	177	020	052	036
168	100	076	062	142	030	041	023	147	191	050	179	120	086	195	026	001	087	018
038	148	046	161	024	035	081	189	122	103	130	193	088	167	025	008	009	105	074
071	156	083	132	200	017	080	033	202	198	015	133	170	205	085	152	165	027	060
014	128	003	159	082	199	135	012	073	160	203	169	134	039	031	051	107	115	176
094	034	084	183	090	047	121	042	124	144	077	112	149	097	136	164	127	058	043
125	013	182	109	197	016	184	110	187	066	106	150	021	173	157	181	204	139	145
029	064	111	201	005	092	180	048	037	075	055	044	206	010	104	098	172	171	059
007	208	138	158	126	068	175	045	174	000	137	053	056	093	153	070	166	032	196
089	146	102	154	114	040	151	207	155	140	185	108	117	004	101	113	194	119	123

5　倾斜切割的理论算法

从理论上来说,碎纸机切割纵横切割相结合,但实际上由于纸张变形、文件放置、机器精度等问题,很难做到标准的垂直和平行切割,在实际情况中难免会出现一些误差。这时仍可根据碎片文件之间的匹配率进行拼接,但如前所述,直接拼接误差太大所有仍然要先考虑分组。以下仅针对文件产生一定倾斜角的情况从理论上提出分组,其余情况可依此拓展。针对以下 3 个碎片文件(图 3):

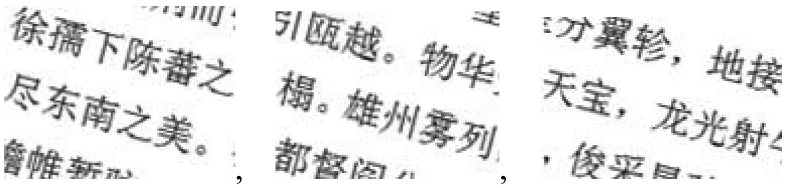

图 3　倾斜切割所得到的文件碎片

现采用两种思路进行行分组:

(1) 对碎片一分析,得到图片倾斜角大小,由此判断出与文字方向平行的空白行位置与相应行数。文字倾斜时因其上下两行文字高度始终改变,纵横切割拼接方法中按一行文字高度分组的方法就不适用了。这里采用相连接的两碎片左右文字的衔接度来分组。由于文字倾斜时一行文字所占最大值与最小值始终改变,因此采用按列抽样的方法。如对上面第一个碎片,抽取最左侧 n 列,取一行文字所占象素位置,逐列取其最大值和最小值(为避免误差,对一行文字最高处明显与前后不符的列舍去),分别取其平均值,设为碎片文

件左侧行高参数的估计值。同理,可得到碎片文件右侧参数的估计值。在允许一定的误差情况下,根据前一文件最右侧参数的估计值和最右侧参数估计值的比较进行行分组即可。

(2) 在扫描碎片文件时通过手动,将碎片中文字方向恢复为水平,如上面碎片文件可扫描为图4:

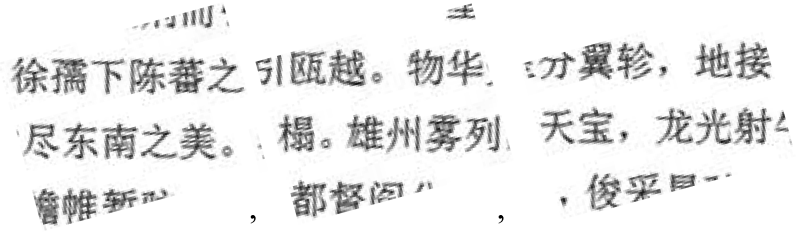

图4 将文字方向调为水平后得到的文件碎片

此时可将文件按斜行分组,且每两个碎片要以如下两个标准进行拼接:两个碎片文件之间的匹配律;前一碎片文件最右侧文字行高与后一文件最左侧文字行高相匹配。

参考文献:

[1] 罗智中. 基于线段扫描的碎纸片边界检测算法研究[J]. 仪器仪表学报,2011,32(2):289-294

[2] 王欣洁. 基于灰度矩阵的中文碎纸片的拼接复原算法[J]. 智能计算机与应用,2013,3(6):95-97

[3] 徐雅平,王运生. 碎纸片的拼接复原[J]. 上海商学院学报, 2013,4(5):79-84

[4] 李晓霞,高志鹏,张蕊倚,等. 关于中英文的碎纸片拼接复原问题研究[J]. 运城学报 2013,31(5):12-15

[5] 杨雯雯,陶佳琪,郑路通,等. 单页单面汉字纵横切碎片拼接复原算法[J]. 运城学报 2013,31(5):16-20

[6] 罗智中. 基于文字特征的文档碎纸半自动拼接[J]. 计算机工程与应用, 2012,48(5):207-210

[7] 王沫然. MATLAB6.0 与科学计算[M].北京:电子工业出版社, 2011

[8] THOMAS H C, CHARLES E L, RONALD L R.Introduction to Algorithms 算法导论[M]. 北京:机械工业出版社,2010

An Algorithm for Regular Fragments Reassembling

LI Meng[1] , LIU Jia-bao[2]

(1.Statistics Bureau of Hefei, Hefei 230088, China;
2. The department public education, Anhui Xinhua University, Hefei 230088, China)

Abstract: Only aiming at the paper fragment models based on the longitudinal cut method or the crisscrossed cut method, according to the matching degree of the pixels on the edge of the graphics, this paper proposed an efficient fragments reassembling algorithm, which solves the problem of regular fragments reassembling by crisscrossed cut. In the aspect of error regulation, a way of using the computer-aid manual work instead of the only manual adjustment and the auxiliary algorithm are proposed. Meanwhile, the corresponding algorithm for the regular fragments reassembling resulting from non-crisscrossed cut is theoretically put forward as well.

Key words: matching rate;group dividing of fragment lines;the matching rate of lines;Greedy Method