规则碎片的拼接问题

摘要

本文针对两种切割方式以及不同的正反面特点的文件复原问题进行了建模与求解算法设置。首先对只含有纵切的情况，建立了对二值化灰度矩阵匹配度函数的TSP模型，并在求解时进一步转化为优化模型利用贪心算法求解；然后对于横切加纵切的情况，针对中英文字体特点的差异性建立了基于基线匹配的中文拼接模型以及基于四线三格匹配的英文拼接模型，完成了碎纸片复原；最后针对单双面问题，为了提高程序运行效率，我们固定两侧基准碎片减少需要遍历的情况，复原了文件且模型效果较好。

针对问题一：在只考虑纵切的情况下，首先将图像转化为灰度矩阵并进行二值化处理。首先根据页边距特点找出最左侧图片，其次利用最小二乘距离定义了任意两碎片之间的匹配度，并且将此数值赋为碎片之间的权值，将问题简化为TSP问题，最终利用贪心算法转化为优化问题求解。由结果得出该模型对中英文均适用，并且无需人工干预。

针对问题二：考虑碎纸机以横切和纵切两种方式切割得到了209张碎片，由于黑体文字的规范性且碎纸机为均匀切割，我们考虑将行列排序逐步分开处理。考虑同行文字字体高度一致，所以先将碎片按行分组。由于中英文的差异性我们建立了不同的按行聚类模型，在聚类中文时利用黑体字以及空白间距的基线信息进行分类；同时针对英文四线三格的特点利用累积灰度图确定碎片的文字位置信息进行聚类，然后利用第一问的模型进行行内拼接，最终利用行间匹配算法对行间的碎片进行最终拼接，得到了复原结果。复原过程中需要人工干预的地方出现在部分特征不明显碎片，机器单独将之归为一类，我们利用语义人工干预复原，提高了匹配的成功率。

针对问题三：针对正反面的情况，首先根据页边距确定左右两边的碎片，然后利用贪心算法进行匹配与按行分组，最终利用第二问的模型进行行内匹配和行间匹配得到最终复原结果。其中对复原错误的情况我们进行人工干预，提高了结果的可靠性和准确率。

本文针对文件复原与文字行列特点给出了针对性较强的模型，并且定义了匹配度与基线等因素进行精准复原并给出算法可自动给出碎片编号与复原图，提高了程序执行程度，降低了人工干预率。

关键词：碎片复原、二值化矩阵、TSP、匹配度、$K-Means$聚类、贪心算法

一、问题重述

破碎文件的拼接复原工作有很大应用，在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任务。随着计算机技术的发展，人们试图开发碎纸片的自动拼接技术，以提高拼接复原效率。请讨论以下问题：

1. 对于给定的来自同一页印刷文字文件的碎纸机破碎纸片（仅纵切），建立碎纸片拼接复原模型和算法，并针对附件 1、附件 2 给出的中、英文各一页文件的碎片数据进行拼接复原。如果复原过程需要人工干预，请写出干预方式及干预的时间节点。复原结果以图片形式及表格形式表达。

2. 对于碎纸机既纵切又横切的情形，请设计碎纸片拼接复原模型和算法，并针对附件 3、附件 4 给出的中、英文各一页文件的碎片数据进行拼接复原。如果复原过程需要人工干预，请写出干预方式及干预的时间节点。复原结果表达要求同上。

3. 上述所给碎片数据均为单面打印文件，从现实情形出发，还可能有双面打印文件的碎纸片拼接复原问题需要解决。请尝试设计相应的碎纸片拼接复原模型与算法，并就附件 5 的碎片数据给出拼接复原结果。

二、问题分析

2.1 问题一分析

问题一要求在只考虑纵切的情况下对印刷文件对碎纸片建立复原模型，并给出附件的复原结果。由于颜色在计算机中用$0-255$之间的数值表示，故将附件所给图片转化为矩阵来表示，由此得到信息灰度矩阵。通过研究每张图片的灰度矩阵来解决问题。图一为一维的复原问题，在只考虑纵切的情况下每张碎片的信息量较大、特征明显，匹配的成功率较高。

首先找出最左端的碎片，该纸片的特点是左侧具有一定的空白（页边距）。然后按照从左往右的顺序依次拼接。由文字特征可以知道若两张碎片相邻，则两张图对应的灰度矩阵边缘像素列灰度值相似度很高，绝大部分对应行的灰度值相等，将这个特点作为衡量任意两个碎片之间匹配度的标准，利用最小二乘法计算出两个碎片的匹配度系数，将此系数赋为这两个碎片之间的权值。此时问题转化为典型的TSP问题，碎纸片为结点，最小二乘值为边的权值，寻找最短路径。本问题中英文字体的差异对结果的影响很小。

考虑到灰度矩阵中代表颜色的数字数值过大且种类较多，在计算处理的时候会存在

误差，所以对灰度矩阵进行二值化将之化为只有0（黑色像素）和1（白色像素）2 的矩阵。常用方法有设定全局阈值以及针对特征进行处理，在此考虑到文字特征的重要性、灰色像素其实为文字与空白的边缘的特征，采取针对特征处理，将全白像素视为1，灰度以及黑全转化为0，即全白才白，有黑就黑。转化后将最小二乘距离计算出解决TSP问题提高算法效率。

2. 2 问题二的分析

问题二在问题一的基础上加入了横切，碎纸片数量增多为208张且碎片上信息量减少，直接拼接误差会很大，我们考虑在第一问模型上进行改进，先按行分组匹配之后再对各行进行拼接。考虑到中文和英文在字体特征和印刷结构的不同，需要定义不同的特征向量去描述中文和英文碎纸片上的信息。在按行分组时，文字在碎片上的位置以及行间距决定分组结果。

对于中文我们对每个碎纸片上文字的字体位置以及行间距空白信息算出基线数值，根据这个数值进行$K\text{-}M$eans 聚类得到按行分组结果。英文因为没有中文字体规整，我们考虑四线三格最中间一格的主体部分，利用累积灰度图验证找到黑色像素值突变点，再利用文字位置等得到基线信息，进行聚类。按行分组结束后利用第一问模型进行行内拼接，之后再把各行拼接起来。

2. 3 问题三的分析

对于问题三，在问题二的基础上加入了正反面信息，碎纸片的大小形状信息未改变。此时进行匹配情况非常多。需要在第二问的基础上综合考虑正反面匹配的误差，分组方式与拼接方式与第二问基本相同，最后通过语义进行部分人工干预。

综上所述，本问题可以看做对以碎纸片为结点，误差评估函数数值为边的权值的TSP问题寻找最优解。

三、模型假设

1.假设需要复原的碎片来自同一篇文章且语义通顺。

2.假设文件上文字种类、行间距、文字特征信息一致。

3.假设纸张无黑点等客观影响因素。

4.假设文章两侧空白距离明显。

四、符号说明

k	碎片文字文件编号
A	灰度矩阵
a_{ij}^{k}	第 k 个碎片文件对应矩阵的第 i 行第 j 列的灰度值
B^{k}	二值化后的灰度矩阵
$d(A_i, A_j)$	匹配度
b_{ij}	二值化后的灰度矩阵对应灰度值
δ	行高匹配度
E	碎片集合
x_{ij}	TSP 问题决策变量

五、模型的建立与求解

5.1 模型准备

5.1.1 图像数据化处理

在进行文字碎片匹配时由于计算机只能处理数字信号，我们将之转化为碎片信息矩阵，问题一中所给的碎片每一个都会被转化为 1980×72 的矩阵。我们通过 MATLAB 根据像素点的灰度值反应碎片信息，$0\text{-}255$ 代表不同颜色，“0”为黑色，“255”为白色。中间各个数值表示由黑到白的过渡色。

在这里每个碎片都可以得到各自的数字矩阵 $A_k(k=1,2,\ldots\ldots,\ n)$，$A_k$ 表示第 k 个碎片所对应的矩阵，a_{ij}^{k} 表示第 k 个碎片文件对应矩阵第 i 行第 j 列所代表的灰度值。

5.1.2 二值化

由于在定义匹配度以及误差值的时候需要用列边缘像素计算，而边缘像素信息量多且数值较大，我们采取二值化方法对灰度矩阵进行处理，提高算法效率。考虑到此题中文字特征的重要性，为了尽可能保留文字的黑色像素，我们利用如下公式二值化，建立新的灰度矩阵 B^{k}。

$$b_{ij}=\begin{cases}0 & a_{ij}\neq 255 \text{且} a_{ij}\in[0,255]\\ 1 & a_{ij}=255\end{cases} \tag{1}$$

5.1.3 匹配度的定义

定义任意两个碎片 A_i 和 A_j $(i,j=1,2,...,19且i\neq j)$ 之间的最小二乘距离称为二者的匹配度：

$$d(Ai,Aj)=\sum_{m,n=1}^{1980}(b_{m72}^{i}-b_{n72}^{j})^2 \tag{2}$$

此数值越小，匹配度越高。

5.1.4 行高中线的定义

根据碎片上文字位置信息判断碎片是否可以分为同一组时，文字行高特征是最明显的一类信息，按行分组在判断是否为同行时依据则为灰度矩阵中行高信息是否对应。若行高信息匹配则一定为同行，在此我们建立行高信息相同的标准，行高中线 h，比对碎片之间行高中线考察匹配度。

$$h_k=\frac{u_i+u_j}{2}\quad k=1,2,3\quad i、j=1,2 \tag{3}$$

其中 u_i、u_j 为像素突变处对应灰度矩阵所在行。

5.2 问题一模型的建立及求解

5.2.1 模型的建立

对于碎片的复原问题主要修复方法是像素点匹配法。此题特点性较强，解决问题时主要考虑文字特征以及文字位置。已知原图四周没有文字且假设页边距足够大，根据此特点很容易找出并固定最左侧一列碎片。

之后在对剩余碎片进行匹配时，考虑到图片特征，若两张图片被切割有两种情况：

（1）碎纸机纵切时且在文字上；

（2）正好在两列文字之间的空白部分切割。

针对情况（1）：切在文字上时，可以根据文字笔画的连续性确定碎片是否能拼接，通常情况下左、右碎片黑色像素位置是连续的，像素值相似度很高，计算两个碎片的匹配度。

针对情况（2）：当切在空白部分时，第一问仅考虑纵切，碎片信息较多误差小，针对 $1980*72$ 的矩阵匹配度数值误差影响很小。

根据匹配度的定义我们可以计算得到任意两个碎片之间的匹配度，现在问题本质则变为：已知任意两点之间的权值，寻找一个有序序列使得碎片之间的总匹配度最高，权值之和最小。该问题为典型的 TSP 求最短路径问题，我们使用贪心算法求解。

如图其中数字节点表示碎片，有向线段长度表示匹配度数值，弧尾为左边碎片，弧头为右碎片。现寻找一条回路使遍历所有节点且不重复达到匹配度最高。

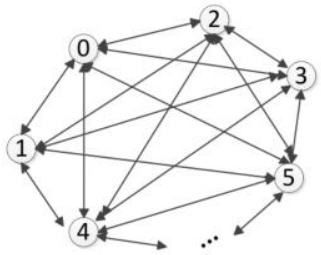

图 1 TSP 问题示意图

则该问题为以匹配度最高为目标函数建立的寻最优解的TSP问题模型，可表示为：

$$\min f = \sum_{i,j\in E} d(A_i, A_j)x_{ij}$$

$$\begin{cases} \sum_{i,j\in E} x_{ij} = 1 \\ \sum_{i,j\in E} x_{ji} = 1 \\ x_{ij} = 0,1 \end{cases} \quad (4)$$

5.2.2 模型的求解

$Step1$：将所有碎片放入指针存放；

$Step2$：根据灰度矩阵特点找出最左侧一列，并将之从存放地址中取出；

$Step3$：依次计算已经固定的最右侧一个碎片与剩余碎片的匹配度，选择匹配度最高的碎片作为当前碎片的右侧碎片，从地址中取出该碎片；

$Step4$：重复上述步骤直到碎片取完。

该算法根据匹配度求解问题最优解，信息量比较大，对中英文都适用。最终得到附件1与附件2求解结果如下：

表 1 附件 1 排序后碎片序列表

排序	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
碎片编号	8	14	12	15	3	10	2	16	1	4	5	9	13	18	11	7	17	0	6

表 2 附件 2 排序后碎片序列表

排序	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
碎片编号	3	6	2	7	15	18	11	0	5	1	9	13	10	8	12	14	17	16	4

由附录还原的文件可知两张图片都具有极高的可读性，所以我们认为复原结果正确。通过结果我们发现贪心算法可以得到全局最优解，原因是匹配度定义准确并且碎片像素信息提供较多，并且对中英文都适用。

5.3 问题二模型的建立与求解

问题二在问题一的问题上加入了横切，碎片数量增多且每个碎片具有的信息量大幅减小，所以我们需要对中英文特征的差异性进一步讨论。此问题是第一问的推广，所以我们考虑将之转化为与第一问相似的问题求解，完整化模型。

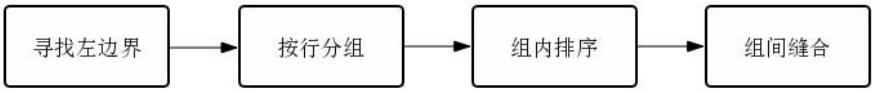

图 2 逻辑思路示意图

5.3.1 基于基线匹配的中文拼接模型

5.3.1.1 碎片的按行分组

（1）按行分组的原因

在解决第二问时我们希望整个问题可以通过完整的模型解决，但在此我们采用按行分组而非按列主要原因是此问题是加入了横切，我们应该按横切将碎片归类，且已知行与行之间的文字平行，进行同样横切的碎片文字的行高与字体位置信息是相同的，此特征会极大提高分组效率。

（2）按行分组的标准

由附件我们可以知道，每个碎片都具有三行文字的位置且大小相同，为了找出具有相同横切特征的碎片，我们对每个碎片上具有的文字位置和间距信息进行数据分析，得到每个碎片的一些文字特征信息，按照这类信息进行 $K-Means$ 聚类找出同行碎片。

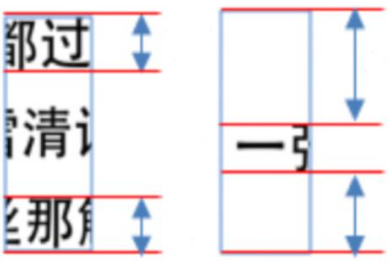

图 3 文字参数示意图

如图所示图中含有四个距离，一张碎片文字的行高信息可以由这些信息反映，我们根据这四个距离对每个碎片进行数据分析，由得出的数据进行按行聚类。在计算信息时，

为了使特征尽可能准确，我们对每个碎片的灰度矩阵数值进行分析，对矩阵每行进行扫描，遇到有黑色像素即为出现文字，一直到矩阵某一行全为白色像素认为该行文字高度结束。由此确定出每个碎片的文字距离信息，根据此数据信息进行聚类，得到行间分组结果。在扫描行高的时候考虑到系统误差，我们对出现文字的黑色像素矩阵行下方预留 $2-3$ 行的缓冲带，减小误差。

5.3.1.2 $K-M$eans聚类分析

K-均值聚类算法是一种划分聚类分割方法。划分的基本原理是：给定一个有 N 个元组或者记录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，$K \le N$。而这 K 个分组满足下列条件：

（1）每一个分组至少包含一个数据记录；
（2）每一个数据记录属于且仅属于一个分组。

对给定的 K 我们首先给出左首位的一列碎片，通过反复迭代的方法改变分组。工作原理为首先选择 K 个点，每个点初始的代表每个簇的聚类中心，然后计算剩余各个样本到聚类中心的距离，将它赋给最近的簇，接着重新计算这一簇的平均值，不断重复过程直到相邻两次没有明显变化，说明形成的簇已经收敛。具体模型为：

$$E=\sum_{j=1}^{k}\sum_{x_i\in w_j}\left\|x_i-m_j\right\|^2 \tag{5}$$

利用聚类分析进行按行分组后，会由组内碎片个数等发现明显存在错误，进行人工干预，在给出结果时我们给出人工干预的方式与时间节点。

5.3.1.3 组内碎片匹配

$Step1$：将按行分组的同组碎片以及聚类不成功未分组碎片存放；

$Step2$：根据灰度矩阵特点找出最左侧具有全白像素点的碎片，并将之从存放地址中取出；

$Step3$：依次计算已经固定的最右侧一个碎片与剩余碎片的匹配度，选择匹配度最高的碎片作为当前碎片的右侧碎片，从地址中取出该碎片；

$Step4$：重复上述步骤直到碎片取完。

在附件中给出按行聚类分组结果，可知横切的十一组分类较准确，分类不明确的碎

片被归为其余四类，在人工干预时错误也可以肉眼调整。

5.3.1.4 人工干预

由于碎纸机以横切和纵切方式切割，得到碎片过小，含带矩阵信息过少。在进行最优匹配时若两个碎片较为相似、可能出现排序错误，此时我们加以人工干预。

当行内排序拼接完成后，我们可知文字基线比较相似所以在最优匹配排序时出现乱序现象，但大部分通过语句可读性判断是正确的，针对这种部分片段我们进行人工干预。

5.3.1.5 组间缝合

在第一问纵切的模型上进行改进，利用同样的方法对只进行一种切割的碎片计算匹配度，得到任意两碎片之间的权值后，寻找最上侧存在全白像素行，确定最上侧碎片。此时问题简化为TSP问题，以碎片为结点，匹配度距离值为权值，求解最优解。由于已经进行按行分组之后的组内拼接，所以在组间缝合时碎片像素较多，信息量大，结果较为准确。

5.3.2 基于四线三格基线匹配的英文拼接模型

图 4 英文行高示意图

由文字的字形、行分布特征可以知道，中英文文字有明显差别，需要分类讨论。英文在印刷排版有严格的四线三格，根据这个特点我们通过找出黑色像素突变点确定中间一格的文字位置，抹去长笔画影响而寻找主要部分，建立了基于四线三格的英文拼接模型。思路同样为根据行分布信息进行按行聚类，之后与中文解决思路大致相同。

5.3.2.1 基于中文按行分组的模型修正

为了寻找四线三格第二格的主体部分，我们采用累积灰度图方法。由碎片二值化后的灰度矩阵画出文字的累积灰度图，找出数值突变点来确定文字位置。

由于我们在二值化时白色像素值为1，并且在此处空白行间距较为容易确定，所以我们对白色像素值进行累积，得到累积灰度图，由图片峰值对文字继续进行分析。并且进一步结合行高中线精准判断碎片中文字位置信息以及空白间距位置信息，从而进行按行分组聚类。

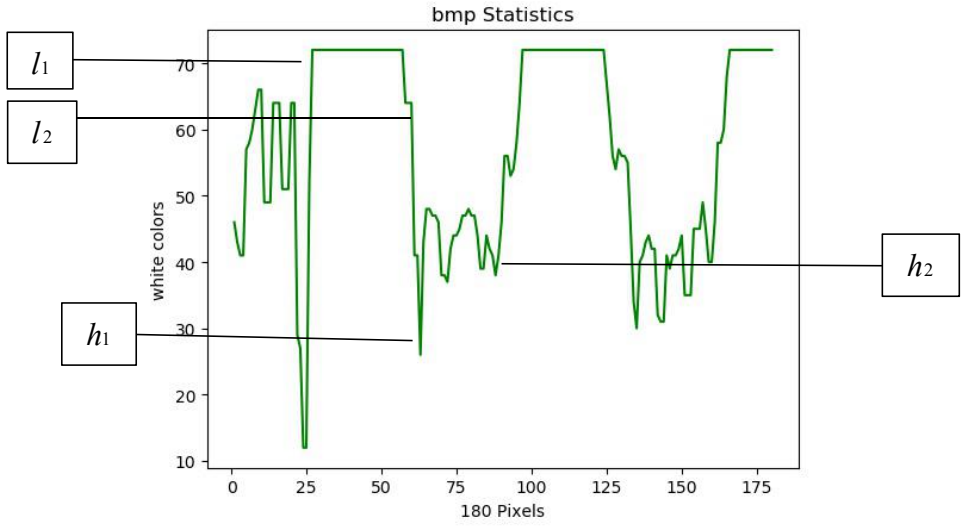

图 5 累积灰度示意图

上图为白色像素累计值的图像。按行分组的重要依据是每块碎片文字位置的基准线，因为需要确定每块碎片基准线的位置。由图片基准线是黑白像素点的分水岭，通过黑白像素点个数来确定基准线位置。方法如下：

可知当峰值持续性较高且平稳时(l_1-l_2)，当前灰度矩阵扫描行处于间隔位置，峰值较低(h_1-h_2)为黑色文字部分。通过对比任意碎片之间的累积灰度图峰值位置以及平稳程度我们进行按行聚类。为了进一步准确判断间隔的位置，我们设置阈值来判断灰度累计图是否达到空白部分。

$$\text{阈值}\delta:\begin{cases}\delta<65 & \text{处于文字部分}\\ \delta\geq 65 & \text{处于空白间隔部分}\end{cases} \tag{6}$$

5.3.3 模型求解结果

表 3 附件 3 求解结果

	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	129	28	91	188	141	11	22
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 4 附件 4 求解结果

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

红色加框表示人工干预位置，由结果可知干预次数较少。

5.4 问题三模型的建立与求解

对于问题三，在问题二的基础上加入了正反面信息，碎纸片的数据信息均未改变。解决方案基本与第二问相同，但需要在第二问基础上加入针对碎片正反面特点的改进。

5.4.1 模型的建立

（1）寻找最左侧首位碎片

同上述过程相同，我们先找出最左侧的一类碎片，根据假设这类碎片最左侧有一侧空白，我们根据灰度矩阵反应的像素数值信息索引出这类碎片。由文字信息可判断为左侧还是右侧。

（2）行内拼接，组间缝合

在加入正反面情况之后，两个碎片之间理论上存在四种拼接结果，但在固定最左侧基准碎片后，在其余碎片中选择下一个拼接碎片只存在两种情况，如图所示。

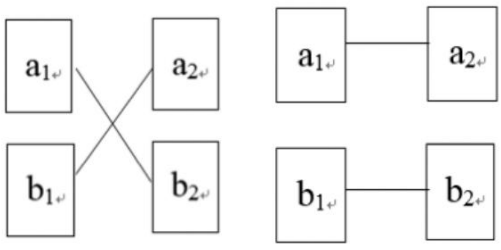

图 6 两种碎片匹配示意图

图中(a_1,b_1)为左侧基准图片的正反面。为了将正反面问题简化，我们在固定左侧右侧的基准碎片之后，用剩余的碎片根据第二问思路利用贪心算法进行匹配减少一半的计算量。

思路是利用第二问模型进行固定首位碎片的按行分组。但在此需注意匹配时正反面都需进行，即将所有碎片放入待选区域。若已经确定为正或为反面行的则直接将此碎片

对应的背面碎片归为另一面行中排序。在行内拼接完成后进行组间缝合。考虑到碎片信息过小过碎，在过程中加入人工干预。

5.4.2 模型的求解结果

根据计算机编程得到最终如下编号结果：

表 5 附件 5 一面结果

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	078b	111b	125a	140a	155a	150a	183b	174b	110a	066a	108a	018b	029a	189b	081b	164b	020a	047a	136b
2	089a	010b	036a	076b	178a	044a	025b	192a	124b	022a	120b	144a	079a	014a	059a	060b	147a	152a	005a
3	186b	153a	084b	042b	030a	038a	121a	098a	094b	061b	137b	045a	138a	056b	131b	187b	086b	200b	143b
4	199b	011b	161a	169b	194b	173b	206b	156a	034a	181b	198b	087a	132b	093a	072b	175a	097a	039b	083a
5	088b	107a	149b	180a	037b	191a	065b	115b	166b	001b	151b	170b	041a	070b	139b	002a	162b	203b	090a
6	114a	184b	179b	116b	207a	058a	158a	197a	154b	028b	012a	017b	102b	064b	208a	142a	057a	024a	013a
7	146a	171b	031a	201a	050a	190b	092b	019b	016b	177b	053b	202a	021b	130a	163a	193b	073b	159a	035a
8	165b	195a	128a	157a	168a	046a	067a	063b	075b	167a	117b	008b	068b	188a	127a	040a	182b	122a	172a
9	003b	007b	085b	148b	077a	004a	069a	032a	074b	126b	176a	185a	000b	080b	027a	135b	141a	204b	105a
10	023b	133a	048a	051b	095a	160b	119a	033b	071b	052a	062a	129b	118b	101a	015b	205a	082b	145a	009b
11	099a	043a	096b	109a	123a	006a	104a	134a	113a	026b	049b	091a	106b	100b	055b	103a	112a	196b	054b

表 6 附件 5 另一面结果

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	136a	047b	020b	164a	081a	189a	029b	018a	108b	066b	110b	174a	183a	150b	155b	140b	125b	111a	078a
2	005b	152b	147b	060a	059b	014b	079b	144b	120a	022b	124a	192b	025a	044b	178b	076a	036b	010a	089b
3	143a	200a	086a	187a	131a	056a	138b	045b	137a	061a	094a	098b	121b	038b	030b	042a	084a	153b	186a
4	083b	039a	097b	175b	072a	093b	132a	087b	198a	181a	034b	156b	206a	173a	194a	169a	161b	011a	199a
5	090b	203a	162a	002b	139a	070a	041b	170a	151a	001a	166a	115a	065a	191b	037a	180b	149a	107b	088a
6	013b	024b	057b	142b	208b	064a	102a	017a	012b	028a	154a	197b	158b	058b	207b	116a	179a	184a	114b
7	035b	159b	073a	193a	163b	130b	021a	202b	053a	177a	016a	019a	092a	190a	050b	201b	031b	171a	146b
8	172b	122b	182b	040b	127b	188b	068a	008a	117a	167b	075a	063a	067b	046b	168b	157b	128b	195b	165a
9	105b	204a	141b	135a	027b	080a	000a	185b	176b	126a	074a	032b	069b	004b	077b	148a	085a	007a	003b
10	009a	145b	082a	205b	015a	101b	118a	129a	062b	052b	071a	033a	119b	160a	095b	051a	048b	133b	023a
11	054a	196a	112b	103b	055a	100a	106a	091b	049a	026a	113b	134b	104b	006b	123b	109b	096a	043b	099b

六、模型优缺点

6.1 模型优点

（1）将一维碎纸片模型转化为排序问题用 Tsp 模型求解，利用现有模型求得全局最优解，具有拓展性；并且在进行第二问考虑横切与纵切时先以左侧空白部分为搜索特征找出左边界，之后依次进行按行分组，组内拼接，组间缝接这种方法可以很快的将相同特征碎片归类，缩小匹配的次数和数目，极大的减少了运行量提高了效率。

（2）针对中、英文不同的特点寻找各自的基线以及行内文字位置及间距信息，利用文字的这个特征去进行复原匹配。

（3）行内分组时得到了各个碎片的文字基线以及空白间距的位置，利用灰度矩阵进行 $K-Means$ 聚类分析出同行的碎片，针对碎片复原问题误差较小，可以减少人工干

预次数。

6.2 模型缺点

第二问行内拼接的时候利用两个碎片的灰度信息误差来确定匹配度，若两个碎片信息较为近似则容易出现错误顺序，误差较大导致人工干预次数增多，程序执行效率被降低。

七、模型的改进与推广

1.对于二维的碎片进行按行聚类时使用的算法是$K-Means$聚类方法，在此我们考虑改进为最大距离聚类法，此方法在聚类时对碎片的文字等信息特征进行放大，可以考虑到原来聚类未细化的数据信息从而使分类结果更精细。

2.在求解TSP问题时原来的求解算法为贪心算法，为了提高算法的适用性和高效性，可以尝试引入模拟退火算法、蚁群算法等智能算法得到全局最优解，得到好一点的结果更好

八、参考文献

[1]卓金武.李必文.魏永生.秦健.《MATLAB 在数学建模中的应用》.北京.北京航空航天大学出版社.2014.7 页

[2]司守奎.孙兆亮.《数学建模算法与应用》.北京.国防工业出版社.2017.4.7 页

[3]谭忠.碎纸片的拼接复原.http://special.univs.cn/service/jianmo/index.shtml.2019.8.7

[4]百度百科.贪心算法.https://baike.baidu.com/item/%E8%B4%A A%E5%BF%8 3%E7%A E%97%E6%B3%95/5411800?fr=aladdin.2019.8.6

附录

问题一原图

城上层楼叠巘。城下清淮古汴。举手揖吴云，人与暮天俱远。魂断。魂断。后夜松江月满。簌簌衣巾莎枣花。村里村北响缲车。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，佩向脸边浓。小郑非常强记，二南依旧能诗。更有鲈鱼堪切脍，儿辈莫教知。自古相从休务日，何妨低唱微吟。天垂云重作春阴。坐中人半醉，帘外雪将深。双鬟绿坠。娇眼横波眉黛聚。妙舞蹁跹。掌上身轻意态妍。碧霄轻笼两凤，寒烟淡拂双鸦。为谁流睇不归家。错认门前过马。

我劝髯张归去好，从来自己忘情。尘心消尽道心平。江南与塞北，何处不堪行。闲窝阻。谁念累损襄王，何曾梦云雨。旧恨前欢，心事两无据。要知欲见无由，痴心犹自，倩人道、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。临水纵横回晚辔。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，人在清凉国。江山如画，望中烟树历历。省可清言挥玉尘，真须保器全真。风流何似道家纯。不应同蜀客，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。酒阑不必看茱萸。俯仰人间今古。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。

俎豆庚桑真过矣，凭君说与南荣。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。借君拍板与门槌。我也逢场作戏、莫相疑。晕腰嫌枕印。印枕嫌腮晕。闲照晚妆残。残妆晚照闲。可恨相逢能几日，不知重会是何年。茱萸仔细更重看。午夜风翻幔，三更月到床。簟纹如水玉肌凉。何物与侬归去、有残妆。金炉犹暖麝煤残。惜香更把宝钗翻。重闻处，余熏在，这一番、气味胜从前。荫暗荷枯一夜霜。新苞绿叶照林光。竹篱茅舍出青黄。霜降水痕收。浅碧鳞鳞露远洲。酒力渐消风力软，飐飐。破帽多情却恋头。烛影摇风，一枕伤春绪。归不去。凤楼何处。芳草迷归路。汤发云腴酽白，盏浮花乳轻圆。人间谁敢更争妍。斗取红窗粉面。炙手无人傍屋头。萧萧暮雨脱梧楸。谁怜季子敝貂裘。

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 glisters 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-

问题二原图

便邮。温香熟美。醉慢云鬟垂两耳。多谢春工。不是花红是玉红。一颗樱桃樊素口。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。背灯偷揾拭尽残妆粉。春事阑珊芳草歇。客里风光，又过清明节。小院黄昏人忆别。落红处处闻啼鴂。岁云暮，须早计，要褐裘。故乡归去千里，佳处辄迟留。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、雅志莫相违。西州路，不应回首，为我沾衣。料峭春风吹酒醒。微冷。山头斜照却相迎。回首向来潇洒处。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。团扇只堪题往事，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。搔首赋归欤。自觉功名懒更疏。若问使君才与术，何如。占得人间一味愚。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。记取钗头新利市。莫将分付东邻子。西塞山边白鹭飞。散花洲外片帆微。桃花流水鳜鱼肥。主人瞋小。欲向东风先醉倒。已属君家。且更从容等待他。愿我已无当世望，似君须向古人求。岁寒松柏肯惊秋。

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，凝然点漆精神。瑶林终自隔风尘。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。犀钱玉果。利市平分沽四坐。多谢无功。此事如何到得侬。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。对一张琴，一壶酒，一溪云。相如未老。梁苑犹能陪俊少。莫惹闲愁。且折

bath day. No news is good news.

Procrastination is the thief of time. Genius is an infinite capacity for taking pains. Nothing succeeds like success. If you can't beat em, join em. After a storm comes a calm. A good beginning makes a good ending.

One hand washes the other. Talk of the Devil, and he is bound to appear. Tuesday's child is full of grace. You can't judge a book by its cover. Now drips the saliva, will become tomorrow the tear. All that glitters is not gold. Discretion is the better part of valour. Little things please little minds. Time flies. Practice what you preach. Cheats never prosper.

The early bird catches the worm. It's the early bird that catches the worm. Don't count your chickens before they are hatched. One swallow does not make a summer. Every picture tells a story. Softly, softly, catchee monkey. Thought is already is late, exactly is the earliest time. Less is more.

A picture paints a thousand words. There's a time and a place for everything. History repeats itself. The more the merrier. Fair exchange is no robbery. A woman's work is never done. Time is money.

Nobody can casually succeed, it comes from the thorough self-control and the will. Not matter of the today will drag tomorrow. They that sow the wind, shall reap the whirlwind. Rob Peter to pay Paul. Every little helps. In for a penny, in for a pound. Never put off until tomorrow what you can do today. There's many a slip twixt cup and lip. The law is an ass. If you can't stand the heat get out of the kitchen. The boy is father to the man. A nod's as good as a wink to a blind horse. Practice makes perfect. Hard work never did anyone any harm. Only has compared to the others early, diligently

问题三原图

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.

Never go to bed on an argument. The course of true love never did run smooth. When the oak is before the ash, then you will only get a splash; when the ash is before the oak, then you may expect a soak. What you lose on the swings you gain on the roundabouts.

Love thy neighbour as thyself. Worrying never did anyone any good. There's nowt so queer as folk. Don't try to walk before you can crawl. Tell the truth and shame the Devil. From the sublime to the ridiculous is only one step. Don't wash your dirty linen in public. Beware of Greeks bearing gifts. Horses for courses. Saturday's child works hard for its living.

Life begins at forty. An apple a day keeps the doctor away. Thursday's child has far to go. Take care of the pence and the pounds will take care of themselves. The husband is always the last to know. It's all grist to the mill. Let the dead bury the dead. Count your blessings. Revenge is a dish best served cold. All's for the best in the best of all possible worlds. It's the empty can that makes the most noise. Never tell tales out of school. Little pitchers have big ears. Love is blind. The price of liberty is eternal vigilance. Let the punishment fit the crime.

The more things change, the more they stay the same. The bread always falls buttered side down. Blood is thicker than water. He who fights and runs away, may live to fight another day. Eat, drink and be merry, for tomorrow we die. 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 monkeys. If you can't be good, be careful. Share and share alike. All's well that ends well. Better late than never. Fish always stink from the head down. A new broom sweeps clean. April showers bring forth May flowers. It never rains but it pours. Never let the sun go down on your anger.

Pearls of wisdom. The proof of the pudding is in the eating. Parsley seed goes nine times to the Devil. Judge not, that ye be not judged. The longest journey starts with a single step. Big fish eat little fish. Great minds think alike. The end justifies the means. Cowards may die many times before their death. You can't win them all. Do as I say, not as I do. Don't upset the apple-cart. Behind every great man there's a great woman. Pride goes before a fall.

You can lead a horse to water, but you can't make it drink. Two heads are better than one. March winds and April showers bring forth May flowers. A swarm in May is worth a load of hay; a swarm in June is worth a silver spoon; but a swarm in July is not worth a fly. Might is right. Let bygones be bygones. It takes all sorts to make a world. A change is as good as a rest. Into every life a little rain must fall. A chain is only as strong as its weakest link.

Don't look a gift horse in the mouth. Old soldiers never die, they just fade away. Seeing is believing. The opera ain't over till the fat lady sings. Silence is golden. Variety is the spice of life. Tomorrow never comes. If it ain't broke, don't fix it. Look before you leap. The road to hell is paved with good