**B题 碎纸片的拼接复原**

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。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（见【结果表达格式说明】）。

**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。附件5给出的是一页英文印刷文字双面打印文件的碎片数据。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，结果表达要求同上。

通过研读分析2013年B题的部分优秀论文，我们对于数学建模的流程及步骤有了较明确的认知。完整解决一个数学建模问题，有以下五个步骤：模型准备、模型假设、模型建立及求解、模型分析。

模型准备，要求我们了解问题的实际背景，明确建模目的，搜集必须的各种信息，理清对象特征。模型假设，是根据对象特征及建模目的，对问题化繁为简，凝练语言，条理假设的过程，使问题线性化、均匀化。模型构成，根据所作的假设分析对象的因果关系，利用对象的内在规律和适当的数学工具，构成各个量间的等式关系或其他数学结构。模型求解，利用数学方法和计算机语言，通过编程和数学软件实现对构建完成的模型的求解过程。模型分析，对模型解答进行数学上的分析，包括误差分析、数据稳定性分析、模型优缺点分析等。

为此，我们小组通过研究2013年B题中获一等奖的部分优秀论文进行了详细的分析和研读，通过深刻理解优秀论文中的建模思路及思考点，培养组内各个成员的建模能力。

1. **模型准备**

经过通读问题，我们了解到该题力求通过数学方式及计算机技术，解决传统人工拼接技术的低效率问题，实现高效的自动拼接技术，为司法物证复原、历史文献修复及军事情报获取等领域提供便捷。在了解问题的实际背景和建模目的后，我们对本题的对象特征进行梳理。问题一要解决单页打印纵切纸片的左右边界匹配问题，其中涉及到汉字和英文匹配的异同点；问题二要解决单页打印横、纵切的纸片复原问题；问题三要解决双面碎片的还原问题。由问题可提取出解题的对象为碎片间的边界关系，利用碎片间差异度大的特征构造特征因子，描述碎片的行列特征，用以比较、分类、匹配。

1. **模型假设**

通过组内成员讨论及研读优秀论文，我们假设本文的碎片具有以下特征。

1. 假设每一附件中的所有碎片来自同一张纸，且该纸具有完备性。
2. 假设所有碎片均为尺寸大小相等、切割边缘光滑的矩形。
3. 假设所有碎片中的文字边缘清晰，已被去除噪音。
4. 假设同一页中，文字的种类、文本颜色、行间距、段落分布、书写格式相同。
5. 假设所有碎片中的文字端正无倾斜。
6. 假设附件中编号为000的图片为第一张图片，001为第二张图片，以此类推。
7. 假设包含000a图片的页为原文件的正面。
8. **模型建立及求解**

**问题一：**

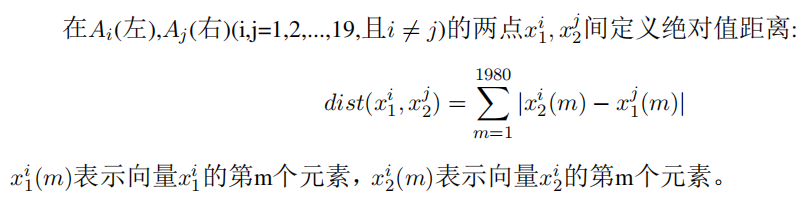
厦门大学

模型一主要解决一维纵切复原问题。首先利用灰度矩阵将碎片数值化，根据碎片左右端的特征找出文章的最左端（即最左端文字与图片左边缘距离最大的碎片），以此类推从左到右拼出全文。在判断左侧碎片的右边缘和待匹配碎片左边缘的匹配程度时，引入绝对值距离来描述，将代表碎片的灰度矩阵二值化（只有0和255），利用最大类间方差法设定一个全局阈值，当绝对值距离越小时，两个碎片的匹配度越高。

1. **模型建立**

碎片bmp000~bmp018对应的二值矩阵分别为A1~A19，特征因子xj1、xj2（j=1,2...,19）分别为Aj的左、右端列向量，可看做1980维空间内的点，共有19×2个列向量。

绝对值距离：



利用两点间的距离大小判断两碎片边缘的匹配程度，距离越大表示匹配程度越低，距离越小表示匹配程度越高。

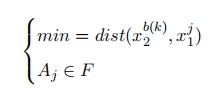
1. **模型求解**

**如论文附图**

Step1：算出19个碎片左端的留白距离，找出距离最大的碎片left，二值矩阵为Aleft。

Step2：建立两个矩阵集合E、F（E用来存放复原后的图片，F用来存放碎片），其中E=Aleft，F={A1,A2,...,A19} \Aleft；建立19列的顺序行向量b，用于存放复原后碎片对应的二值矩阵标号，b(1)=left；k为计数符号，初始值为1。

Step3：利用模型匹配第k张图右侧的图片，并将该图标记为第k+1张。



匹配完成后可得最优解Ass。

Step4：

b(k+1)=ss;

E=E∪{Ass};

F=F \ Ass;

k=k+1;

若k=19，终止；否则跳转至step3

浙江工业大学

模型一主要解决被纵切的碎片复原问题，基于碎片拼接是碎片寻找最佳排列，我们可以将碎片拼接复原问题转化为一个符合特定条件的旅行商问题。

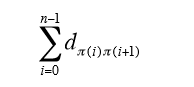
1. **模型一建立**

我们将所有碎片构成一个集合，记做｛A0, A1 … An｝。对碎片的拼接就是找到一个最佳排列，使Ai能够组成一页完整的纸张A，即找到一个排列π，使得

A = A π(0) || A π(1) || … A π(n)

其中“||”为连接符号。

我们需要对碎片进行是否能够拼接的判断，如此循环往复找到一个最佳排列。我们用dij来衡量碎片Aj拼接在Ai后的可能性，即匹配距离。匹配距离越小说明匹配度越高，即图片匹配合适度越高。匹配距离的计算根据图片边缘的情况，所有碎片之间的匹配距离确定下来后，我们利用模型求解最佳排列π，使得



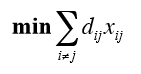
在n个碎片所有排列π集合中最小，如此找到（1）中最小和就转化为一个图论问题。

我们将所有匹配度dij构成一个完全图的邻接矩阵，碎片Ai看作图的顶点。如此，在有n条边的图中，匹配度dij表示第i个顶点到第j个顶点的权值，所以最佳排列π就是一条能够走过所有顶点（且仅被访问一次）并且使得权值和最小的一条路径，即找到一条权值和最小的哈密尔顿路径问题。

1. **求解**

在问题一中所有的碎片都是纵切的，只存在纵向差异，所以我们只需要对碎片进行一维排列。

首先，用matlab将所有碎片读入，计算碎片之间的匹配距离。根据模型一，Xij=1表示从顶点i到顶点j走过，Xij=0表示没有走过，将原模型转化为基于商旅模型的线性规划模型，如下，



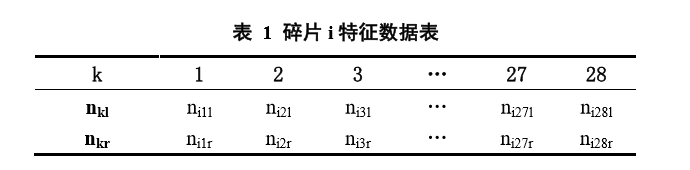
由此，用lingo语言编写程序求解。

依据上式求解出一个0-1矩阵，并不是一个真正的排列。所以，需要确定路径的源点，再根据0-1矩阵求得结果。根据日常经验，一张纸片的左右两侧的边缘像素值为背景像素值。根据这一特点将碎片分别边界碎片和中间碎片。根据算法求得左右侧碎片编号为8、6，故将8作为路径源点。

国防科学技术大学

首先，需要对图像数据进行数字化处理以及文字行特征的提取，再建立基于匹配度的模型。数字化处理主要包括五个方面：二值化处理、edge矩阵的建立、count矩阵的建立、shred with blank left的选取和碎片特征的提取。

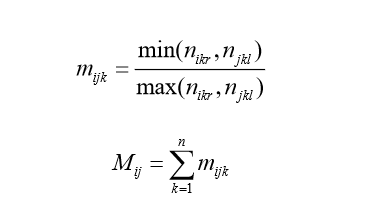
每条碎片维护者以下的特征数据表：



其中，k 为行号；nkl为第 k 行左侧边缘取值为 0 的像素点的数量，nkr为第 k 行右侧边缘取值为 0 的像素点的数量； nikl为第 i 条碎片第 k 行左侧边缘取值为 0 的像素点的数量。

1. **匹配度的定义**

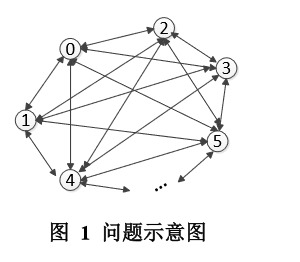
我们定义匹配度Mij作为衡量两个碎片之间匹配程度的标准，其计算公式如下：



其中，n为总行数；mijk表示碎片i和碎片j第k行之间的匹配度；Mij表示碎片i和碎片j之间的匹配度。

1. **模型建立**

通过匹配度的定义，计算得到19个碎片两两之间边缘的匹配度，即可将问题转化为：已知起始碎片和其他碎片之间的匹配度，寻找一个排列使得碎片之间的匹配度总和达到最大。如下图所示：



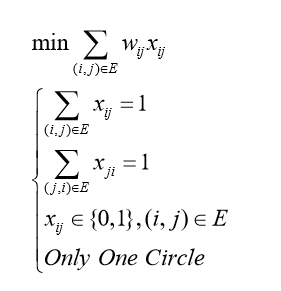
（1）为简化问题，我们引入空白碎片 0，同样可计算得到空白碎片与其他碎片的匹配度；

（2）其中节点表示碎片，有向线段长度表示费 用 wij，且 wij=1-Mij，箭头所指方向表示前一条碎片右侧边缘到后一条碎片左侧边缘。如①→②表示 1 号的碎片右侧边缘与 2 号碎片左侧边缘的费用；

（3）现在要求寻找一条回路遍历所有的节点使得费用达到最小

由此可得，这是一个典型的旅行商问题（TSP），故建立如下模型：

假设图中存在哈密尔顿回路，有 n 个节点，图中(i,j)边的权重为 wij，设决策变量为 xij=1 说明弧(i,j)进入到哈密尔顿回路中，则线性规划模型可表达为：



求解该问题就是求解该线性规划问题的最优解。

1. **求解**

基于上述模型，设计如下算法：

（1）将所有碎片放入地址池中；

（2）选出左侧全为白色像素点的碎片作为基准碎片，记为 si，并将其从碎片池中剔除；

（3）依次计算 si右侧边缘与地址池中其他碎片 sj左侧边缘的匹配度 Mij；

（4）选择匹配度最大的碎片 st作为碎片 si的右侧碎片，并将其记为 si，从碎片池中剔除该碎片；

（5）重复（3）-（4）步直至碎片池为空。

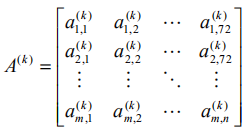
通过求解结果发现，利用贪心策略求解得到了全局最优解，原因在于匹配度的定义较好。同时由于碎片两侧提供的信息量大，很好地避免了中英文之间的差异性。若碎片规格变小，信息量减少，则需要着重考虑中英文的差异性。

西南交通大学

首先将bmp格式的碎纸片图以灰度矩阵形式存储，再将灰度矩阵转化为0-1矩阵；第二，基于0-1矩阵提取每幅图片左右边界的0-1值，建立左右边界匹配模型，确定图片序列；重复上述步骤，将附件图片拼接。

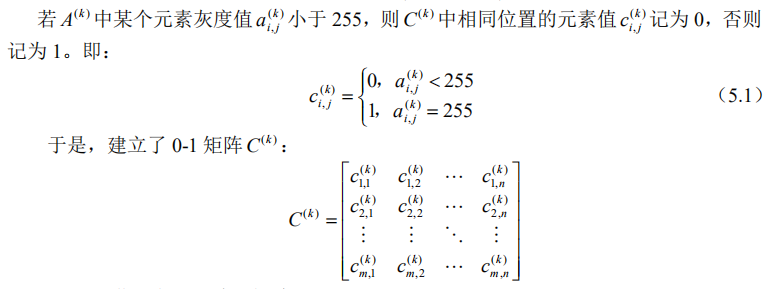
1. **模型准备**

* **灰度矩阵A(k) （k=1,2,...,19）**

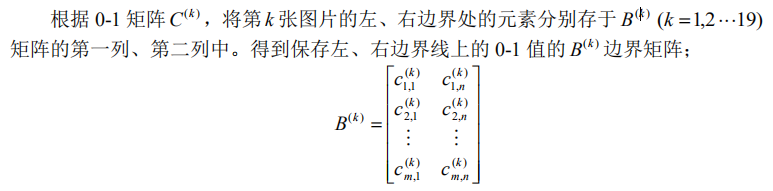


其中，第k个图片的灰度信息以0~255的灰度值存储在矩阵中，颜色越深，灰度值越大。

* **0-1矩阵C(k) （k=1,2,...,19）**



* **左右边界矩阵B(k) （k=1,2,...,19）**



1. **模型建立**

采用边界匹配法，即相邻的两图片边界差异度小。

* **右边界匹配模型**

将第k个图片的右边界与第s张(s=1,2,...,19且s≠k)图片的左边界进行匹配，即求第k个图片边界矩阵的第二列与第s张边界矩阵的第一列对应元素的差，再求差的绝对值之和：



将第k个图片的右边界依次与其余图片进行上述计算，得到n个值，通过比较取其最小值作为右边界匹配值。



* **左边界匹配模型**

将第k个图片的左边界与第s张(s=1,2,...,19且s≠k)图片的右边界进行匹配，即求第k个图片边界矩阵的第一列与第s张边界矩阵的第二列对应元素的差，再求差的绝对值之和：



将第k个图片的左边界依次与其余图片进行上述计算，得到n个值，通过比较取其最小值作为左边界匹配值。



* **最佳边界匹配模型**

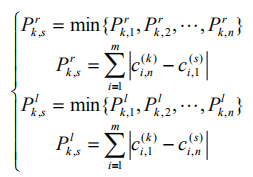
取第k个图片，与剩余的任意图片s(s=1,2,...,19且s≠k)依次进行左、右边界匹配，得到左边界匹配值和右边界匹配值，取两值中的最小值：



P\*k对应的匹配方式即为第k张图片与第s张图片的最佳匹配方式。则说明第k张图片的右边接于第s张图片的左边，记做k→s；说明第k张图片的左边与第s张图片的右边相连，记做s→k。

综上，左右边界匹配模型可总结为：





1. **模型求解**

使用matlab求解，复原结果见论文。

Step1：取附件中第i张碎片，依次与第1张碎片，第2张碎片，第i-1张碎片，...，第n张碎片进行右边界匹配，得到右边界匹配值；

Step2：取附件中第i张碎片，依次与第1张碎片，第2张碎片，第i-1张碎片，...，第n张碎片进行左边界匹配，得到左边界匹配值；

Step3：比较右边界匹配值与左边界匹配值的大小，选择两者中较小值对应的匹配方式，将两张图片按此方式匹配。

Step4：重复step1-3，直至排出原文件的图片序列。

该模型不需要人工干预。

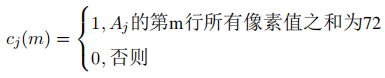
**问题二：**

厦门大学

模型二主要解决二维复原问题。增设两个特征因子空白行和向左（向右）最小边距。先找出复原行左侧的11个头碎片，再找出每行头碎片对应的碎片，构成只有横切的11个碎片，最后再将11个碎片上下拼接，完成复原。

1. **定义**

空白行特征因子：对碎片的灰度矩阵Aj（j=1,2,...,19）构造1980行的列向量cj,cj(m)代表列向量第m行元素。cj用于记录Aj的空白行位置。



向左（向右）最小边距特征因子：对灰度矩阵Ai的向左（向右）最小边距分别记为li1,li2，表示向左（向右）最小边距为碎片最左端向右（最右端向左）数的空白像素列个数。

边界误差：用于描述灰度矩阵Ai(左),Aj(右)边缘的匹配程度。



其中α为像素值的比例因子，β为像素值的偏离因子，α=1，β=0。

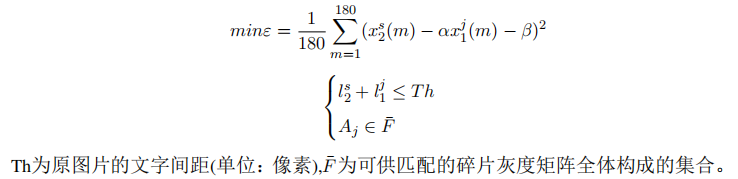
1. **模型建立**

为头碎片Aleft1寻找同行剩余碎片Aj时，需要考虑字符边缘白噪声造成的误差。此时给定一个阈值DH，若|cleft1-cj|<DH，认为Aj与Aleft1同行。

* **复原同行碎片的最优化模型**

为灰度矩阵As寻找右相邻碎片，目标为在可供匹配的所有灰度矩阵中找寻与灰度矩阵As边缘匹配程度最大者。

由于附件三、四中，部分碎片的文字不与碎片左端或右端相交，此类图片无法根据边缘匹配程度选择相邻图片，可用以下不等式排除此类图片的干扰。



* **复原同行碎片的图论模型**

问题转化为在非负有向图中寻找一条权值最小的哈密顿路径问题。

1. **模型求解**

**如论文附图**

Step1：计算每张碎片的向左最小距离，找出距离最大的11张碎片即为原图的头碎片。将图片序号分别标记为left1,left2,...,left11。

Step2：根据头碎片寻找同行剩余碎片的模型和算法分别为11个头碎片找出各自的同行碎片。

Step3：对每行碎片利用复原同行碎片的最优化模型复原，形成11个横切的行碎片。

Step4：类比问题一对11张横切行碎片进行上下拼接。

人工干预及时间节点：

在对每行碎片利用复原同行碎片的最优化模型复原时，当

1. 两图片间的ε值较其余的小，但图片实际不连续时；
2. 多组合的ε值相等且最优时；

需要进行人工干预。

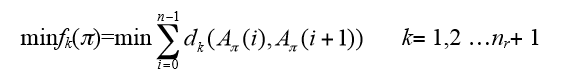
浙江工业大学

模型二主要解决即被纵切又被横切的碎片还原问题。

**1．模型二建立**

我们结合模型一，先根据文本行特征将纵切后的碎片行分组；接着利用模型一找到碎片行分组的拼接排列；再根据模型一得到碎片行的纵向最佳排列；最后根据文本行距相等这一假设条件对所得碎片还原进行自动调整，得到最终结果。

由此，可以将模型二分解为若干模型一中的旅行商问题。对一个被切成nr×nc个碎片的纸张，需要做nr次横向旅行商问题和nc次纵向旅行商问题。故，模型二目标函数如下。



其中，d(i,j)=dij表示碎片j拼接到碎片i的匹配距离。当k=1,2,…n时，fk(π)为横向旅行商问题的目标函数，其匹配距离为碎片横向匹配距离；当k=nr+1时，fk(π)为纵向旅行商问题的目标函数，其匹配距离为碎片纵向匹配距离。

* **纸张左右两侧碎片**

根据日常经验，纸张左右两侧留白比文字与文字或文字与符号之间的留白要多，所以采用逐步缩小范围的搜索方法。

* **文本行特征**

**中文**

对于一个中文字碎片Ak进行按行扫描,记δu为碎片从顶端开始编号的像素行号，记rik为碎片Ak在第δu=i行的特征值。对于任意碎片Ak，从δu=1开始递增扫描碎片的像素行，若碎片Ak在像素行δu有一个文字像素值（该行有笔画），则rik=0，否则（该行无笔画）为1。如此得到碎片Ak的行距特征向量rk=[r1k,r2k,…,rpk]T。根据统计，得到上下两个文字间的空白距离为63个像素距离。

**英文**

英文文字的书写一般在四线三格之中，只有g、j、p、q、y、Q这六个字母超过第三条线，所以可以将第三条线作为底线判别文字行。当从上往下扫描图片时，突然有90%以上的像素点为背景像素值时，就认为该行为底线，将底线上面三分之二舌吻文字像素值，底线下面的三分之一设为背景像素值。根据统计，上下两条底线之间的距离为63个像素距离。

* **同行碎片**

基于碎片行距特征向量，定义碎片Aj归为碎片Ai所在行后 由行距特征向量ri与rj作异或操作得到的模为文本行距匹配距离的dg(i,j)=|rirj|。记θ为文本行距匹配的梯度阈值，当dg(i,j)≥θ，认为碎片Aj归Ai所在行的可能性较大。

* **基于纸张碎片左分组的碎片行分组**

由于段落首行缩进或者段尾留白，在进行碎片查找时会发生未构成整行碎片数nr就停止查找。我们可以对碎片先右分组再左分组，最后由左右分组结合的方法来对行进行分组，以此减少误差。

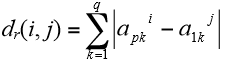
* **横向碎片拼接**

用模型一对得到的行分组进行拼接还原，得到无关联碎片行。

* **纵向碎片拼接**

通过纸张双侧向中间拼接碎片，得到碎片组。将每一行的碎片组看做图的顶点，用模型一对碎片组纵向排列。

构造碎片边界纵向匹配距离函数如下，



其中p为碎片像素矩阵Ai的行数。

表示碎纸片Ai下端边缘文字图像和碎纸片Aj上端边缘文字图像的匹配程度，匹配距离越小，则两张碎片越匹配。

* **中文字基于行距的碎片拼接**

由于我们假设纸张中的文本行距相等，所以要对纵向拼接后的结果进行一定的调整。

当一行碎片下边缘为背景色，另一行碎片上边缘为背景色时，并且纵向边缘匹配度较高，拼接在一起会导致这两行行间距变大。因此，我们将这样的碎片组拼接断开，分成几组碎片段，每个碎片段都由一些碎片行组成。

将碎片段重新排列组合，选出文本行间距最符合文本行间距相等这一条件的碎片段组合，以此作为最终碎片段还原结果。由于文字书写时行距一般相等，所以按行扫描文件，行间距会呈现一定的周期规律。故对某一碎片行的每个碎片行距特征向量做求和运算，就可以得到文本行距的变化曲线，取曲线变化规整的组合为最终结果。

* **英文字基于行距的碎片拼接**

因为英文文字排列较为稀疏，所以在横切的时候会出现多个两端空白的碎片行，所以利用碎片纵向匹配距离会出现错误。所以，我们通过之前分析得到的英文文本行特征（底线间距为63像素值）求出每个碎片行的底线位置。

引入碎片纵向距离函数de（i，j）=63-mod(l(i,j),63)，mod(l(i,j),63)为碎片Ai,与碎片Aj底线之间距离与63做取余运算。再定义新的复合距离dre=dr\*(i,j)+de\*(i,j)，以此转化为旅行商问题。

**2.求解**

**如论文附图**

其中，对于附件4，人工干预的时间节点为对行分组碎片横向拼接后。

在经过行分组碎片拼接后，会出现问题。某些碎片有一个共同特征，两侧中有一侧为白色，与之匹配的左侧或右侧碎片边缘也是白色，在这种情况下答案可能是错误的。所以在这里进行人工干预，将碎片顺序进行交换后，通过手工拼写验证，得到新的正确序列。

国防科技大学

由于中英文字存在显著差异，对算法设计有很大影响，所以要对此进行具体讨论。分析得到如下差异性特征：

（1）中文字结构复杂，笔画繁多，无法确定固定的基准线；

（2）英文字印刷格式固定，一般为“四线三格” 的形式，容易通过基准线来确定具体字母的位置信息。

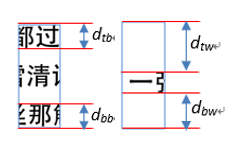
基于以上分析，我们需要针对中英文设计不同的模型进行碎纸片的拼接复原。

* **基于误差评估的汉字拼接模型**

1. **碎片的逐行分组**

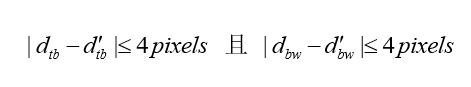
由于碎片上下边缘信息量少，分组误差较大，以及处于同一自然段开始位置的两块碎片因为首行缩进2个字符的上半部分差异导致匹配失败率高，所以不可取按列分组，应该同时考虑左侧边缘与右侧边缘的信息。

确立四个参数 dtb，dbb，dtw，dbw 其意义如下图所示。



其中 dtb，dbb代表由上下边缘至中间像素点由黑过度到白的距离；dtw，dbw代表上下边缘至中间像素点由白过度到黑的距离。

根据上述定义，对于每一幅图像，首先判断其上下边缘的过度情况，由此可以确定上下边缘的两个参数。故，我们认为如果两个碎片具有相同的参数，如同样为 dtb和 dbw，同时相同参数之间的差值不超过 4pixels，即：

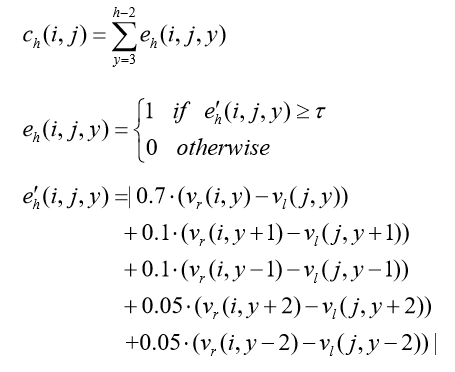


那么，我们认为两碎片在同一组内。

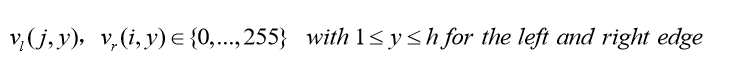
1. **组内碎片匹配算法的设计**

通过误差评估函数来评价两块碎片拼接的好坏程度。

由题意得，每页被分为了 11×19 个碎片，每个碎片具有同样的规格。那么，就有hi=hj，表示两碎片的高度相等。我们定义误差评估函数 ch(i,j)为：



其中，每个像素点以 8 位数值表示，有：



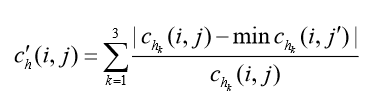
参考问题一，设计组内匹配算法如下：

（1）将组内碎片以及剩余未分组的碎片放入碎片池中；

（2）以左侧全为白色像素点的碎片为基准碎片，记为 si，并将其从碎片池中剔除；

（3）依次计算 si右侧边缘与地址池中其他碎片 sj左侧边缘的误差评估值ch1(i,j), ch2(i,j), ch3(i,j)；

（4）依次计算碎片 si与池中其他碎片 sj的相对误差:



其中，j 表示当前比较的碎片，j为碎片池中其他的碎片， k表示上下三段， 如h1,h2,h3；

（5）选取相对偏差最小的碎片，且 ch’(i,j)≤ζ时，将该碎片作为碎片si的右侧碎片，并记该碎片为 si，将其从碎片池中删除；

（6）重复（3）-（5）步 18 次直至结束。

在此过程中由于最小相对误差的碎片存在多个，计算机很难进行取舍。因此，在程序中，设计了人机交互的界面，通过人工干预的方式来选择最佳的碎片。

1. **组间片段的匹配**

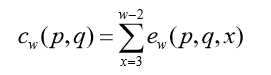
对于组间片段的匹配，主要依据以下两点原则：

（1）若两个片段在同一行，那么两个片段内碎片的数量应当互补（和为19）；

（2）以一个片段为基础，计算其与右侧片段的相对偏差，相对偏差越小，匹配程度越高。

1. **行间匹配**

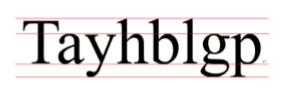
类似于组内碎片匹配的算法，通过计算行片段之间上下边缘的误差评估值 cw(p,q)，最终得到行片段的正确序列。行误差评估值的计算公式如下：



其中，p，q表示行片段号，w表示行像素点的数量。

* **基于基线的英文字拼接模型**

在英文的印刷排版中对应字体均有比较严格的基准线，基线是排成一行的西文字符主体上的一条不可见的假想线。对于英文段落上完整的一行，一定存在至少四条基准线，将整行英文分布在三个格子中，即英文的“四线三格”。本问题中给出的是衬线体（serif 字体）示意图如下：



根据上述特征，建立基于基线的碎片拼接模型。同中文字体，首先需要对 11×19 个碎片进行。

1. **按行分组**

按行分组的重要依据是每一块碎片的基准线，因此首先需要确定每一个碎片基准线的具体位置。通过观察，基准线是黑白像素点的分水岭，因此通过统计黑白渐变处黑色像素点的数量，可以确定基准线的位置。

按行分组时：

1. 选取左侧留白的碎片作为标准碎片，以基准线为依据，若碎片i和碎片j满足：

|line1i-line1j|＜2pixels。那么，认为这两个碎片处于同一行。

（2）若该碎片的基准线不和任何一行的标准碎片相吻合，那么将该碎片放入碎片池中。

最终可以得到 11 个组以及碎片池中少量的碎片。

1. **组内匹配算法设计**

由于在组内，英文碎片的边缘信息与汉字类似，因此可以同样采用误差评估函数ch (i,j)来衡量两个碎片之间的匹配程度。不同的是由于相对误差最小的碎片不一定是该碎片右侧的碎片。因此，当两碎片的相对误差ch’(i,j)＜0.5时，接受碎片j作为碎片i的右侧碎片，否则则加入人工干预的方式，帮助判断是否为右侧碎片

1. **行间匹配**

利用行碎片上下边缘的信息，计算行片段之间上下边缘的误差评估值cw(p,q)，最终可以得到行片段的正确序列。

西南交通大学

首先用问题一的方法对图像进行预处理，构建中英文文章行特征的特征向量，确认像素行的行数；建立特征匹配、左右边界匹配、上下边界匹配模型；通过matlab对模型进行求解，特征匹配模型求解后加入人工干预。

1. **模型准备**

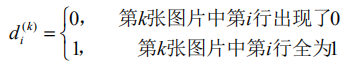
* **图像的预处理**

同问题一，建立每张图片的0-1矩阵。

* **特征灰度条向量**

（1）中文特征灰度条向量

特征灰度条是记录图片中文字的行方向信息的列向量D(k)。对于预处理后的图像，建造一个与碎片的图像行数一致的列向量D(k)，对图像中的像素行进行扫描，若此行中有像素值为0的点，则将列向量D(k)中相同行处的值设为0，否则设为1。



特征灰度条的列向量D(k)为：



（2）英文特征灰度条向量

分析英文字母的结构特性，发现所有的英文字母的主体部分都在中格，故构造特征灰度条向量时，可以只考虑每行英文字母的中格部分。对于预处理后的图像，建造一个与碎片0-1矩阵行数相同的列向量。逐次扫描图像的像素行，若某行元素之和小于M，说明该像素行中黑点较多，是英文字母的主体部分。通过多次试验，取M值为56。



特征灰度条的列向量D(k)为：



若两张碎片的灰度条相似程度达到精确要求，则他们具有相同的图像行特征，位于原文件的同一行。

1. **模型建立**

* **特征匹配模型**

将碎片k与碎片s进行特征比较(s=0,1,...,208且s≠k)，即求碎片k的特征列向量D(k)与碎片s的特征列向量D(s)对应元素的差的绝对值之和，为特征值Wk,s：

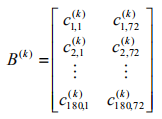


取一个合适小的置信区间[a,b]，若Wk,s∈[a,b]，则认为碎片k与碎片s来自于原文件的同一行。

* **左右边界匹配模型**

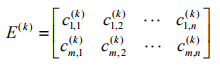
此模型构建类似于问题一的思路。

构建的左右边界矩阵B(k) 为：



* **上下边界匹配模型**

将第k张图片的上、下边界处元素分别存于E(k)(k=1,2...11)矩阵的第一行、第二行中。即上下边界匹配模型中第k行的上下边界矩阵为：



将第k行的上边界与第s（s=1,2...11且s≠k）行的下边界进行上边界匹配，即求第k行的边界矩阵的第一行与第s行的边界矩阵的第二行对应列元素的差的绝对值之和——Qu k,s。重复上述操作，得到n个值：Qu k,1 Qu k,2...Qu k,n。通过比较，取n个值中的最小值，作为上边界匹配值Qu k,s。





同理可得下边界匹配值Qd k,s：



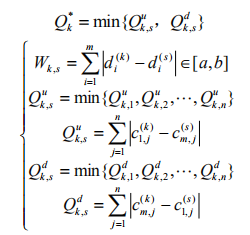


* **最佳边界匹配模型**



Q\*k对应的匹配方式即为第k行与第s行的最佳匹配方式。

综上，我们构建的模型为：



1. **模型求解**

使用matlab求解，复原结果见论文。

Step1：取附件中的第i张碎片，依次与其他碎片进行特征匹配，将匹配成功的碎片存入同一行；

Step2：重复Step1，得到具有相同行特征的行集合；

Step3：人工干预（识别碎片边缘的字迹断线、理解碎片内文字含义），将得到的类的个数降维；

Step4：取同一行中的第i张碎片，与该行中的其他碎片进行左、右边界匹配，找到最佳边界匹配方式，组合图片；

Step5：重复Step4，直至确定出每行内部图片的排列顺序。

Step6：取第i行，与其它行进行上、下边界匹配，找到最佳边界匹配方式，组合图片；

Step7：重复进行Step6，直至得到图片的原始序列。

**问题三：**

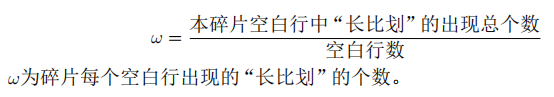
厦门大学

模型三用于解决双页横、纵切碎片的复原问题。针对英文字符的较长比划多的特点，对碎片图像进行预处理，“抹去”长比划。针对碎片差异较为明显的行特征新设特征因子组。本问对比最短距离法、最长距离法和内平方距离法三种聚类分析方法后，得到应用内平方距离法是每类所拥有的碎片个数均衡度最好的，故利用聚类分析的方法将419张碎片聚为11类，同类的大部分碎片处于同行。

1. **定义及相关方法**

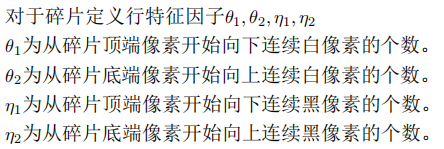
* **英文碎片图像预处理方法**

对拥有较长比划的字母做出界定，包括b、d、f、g、h、i、j、k、l、p、q、t、y以及所有大写字母。



抹去“长比划”基本思想：二值化碎片的灰度矩阵，得到只有黑色和白色的碎片图。给定阈值SH，对碎片分别计算每行的黑像素总数，若总数小于给定阈值SH，认为本像素行处于空白行中，且行中的黑像素为“长比划”的一部分，将该行的每个像素变为白像素，即抹去该长比划。

* **行特征因子组定义**



利用个碎片指标值的方差来描述各碎片此行特征的差异度，差异度越大，则利用该指标区分度越大，则指标选取较优。

观察发现，每组碎片的a、b图的行特征都相同，可知文档正反面的页边距完全相同，按照行特征分类，任意组的a、b图都将分入同一行。

根据四个指标利用聚类分析的方法将209×2张图分为11类，对聚类后的同行碎片进行人工干预，得到11类碎片，每类中有19个碎片，且每类中的19个碎片都属于同一行。故接下来只要先对同行碎片复原，再复原全文。

1. **模型建立**

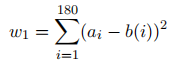
* **同行碎片复原模型与算法**

找行首碎片：观察图片知，位于行首及行尾的图片位于左端的全空白部分宽度较大，通过matlab编程调整限制条件左端全白列部分宽度及右端全白部分宽度，取左端、右端全空白图片的公共部分作为某一面的行首，得到行首编号为left1,left2,...,left11。

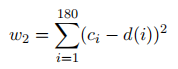
设立匹配指标函数：当图片a、b相邻时，图a正面末列与图b正面首列的匹配值比其他组合小；图a反面首列与图b反面末列的匹配值比其他组合小，故匹配指标模型如下：



其中，



ai,bi分别为前一张图正面灰度矩阵的末列的第i行数据和后一张图正面灰度矩阵的首列的第i行数据。



ci,di分别为前一张图反面灰度矩阵的首列的第i行数据和后一张图反面灰度矩阵的末列的第i行数据。



表示两矩阵特征因子θ1相似程度。分别表示前一张图和后一张图的特征因子。

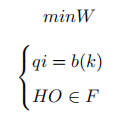
1. **模型求解**

以left1所在行例给出同行复原模型与算法。令碎片left1为新编号1，剩下碎片编号依次为2,...,19。

Step1：建立集合F用于存放未定顺序的碎片。

建立19列的顺序行向量b，并将其初始化零向量，用于存放正面复原后的碎片编号，k为计数单位。其中b(1)=left1，k=1。

Step2：为b(k)碎片的正面匹配右相邻的碎片正面图。设QI为前一张碎片，HO为后一张碎片；qi为前一张碎片编号，ho为后一张碎片编号。



得到碎片b(k)最优匹配HO\*，则：

F=F\HO\*;

K=k+1;

b(k)=ho\*;

若k=19，运算结束，否则跳转至step2。

11个行都复原后，同理问题二对11行进行排序，得到复原图。

人工干预及时间节点：

在step2中，当

（1）两图片间的W值较其余的小，但图片实际不连续时；

（2）多组合的W值相等且最优时；

需要进行人工干预。

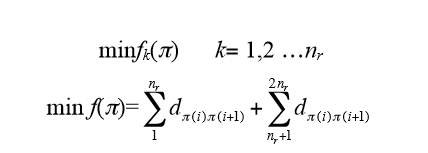
浙江工业大学

模型三主要解决双面碎片还原问题。

**1.模型三建立**

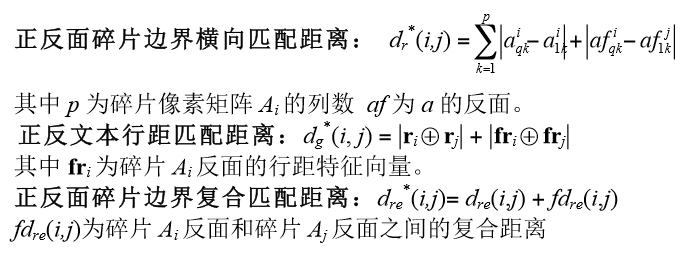
模型三类比模型二，将原问题分解为若干模型一中的旅行商问题，但是由于碎片的双面性，所以引入正反文本行距匹配距离。对碎片进行行分组和拼接等处理，方法类比模型二。

由于碎片的双面性，在对碎片进行行纵向拼接时，要用到模型一种多旅行商的情形。对一个被切成nr×nc的双面碎片纸张，需要做nr次横向旅行商问题和1次纵向多旅行商问题。故，引入模型三目标函数如下。



其中，当k=1,2,…n时，fk(π)为横向旅行商问题的目标函数，其匹配距离为正反面碎片边界横向匹配距离；当k=nr+1时，fk(π)为纵向（2个）旅行商问题的目标函数，其匹配距离为正反面碎片边界复合匹配距离。

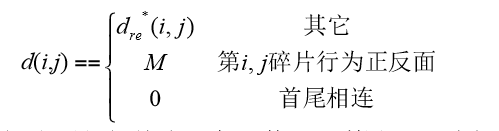
* **正反匹配距离**



* **纵向碎片拼接**

通过改变距离矩阵将2个旅行商问题简化为一个旅行商问题，我们需要把不属于同一列的首位间距定为0，相当于旅行商的再次出发。即通过边界碎片查找到该碎片行中处于原图第一行的碎片y1,y2和最后一行的s1,s2，令d(y1,s1),d(y1,s2),d(y2,s1),d(y2,s2)都为0。

以正反面碎片边界复合匹配距离来表示其两两之间的距离，因为正反面不可以相互连接，所以定义其正反距离为一个足够大的数字M，得到距离矩阵中D中第i行第j列的元素d(i,j)计算公式如下。



计算所有d(i,j)，根据模型一中的旅行商问题的线性规划模型，利用lingo求解。

**2.求解**

**如论文附图**

国防科技大学

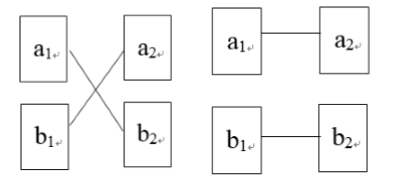
问题三与问题二的解决方案基本类似，针对正反碎片的特点，故将问题二的算法进行改进。

1. **边缘基准碎片的选择**

由于一个碎片存在正反面的信息，因此在选择边缘基准碎片时，需要考虑正反面两侧边缘留白的情况。若某一碎片正反面两侧均有留白，说明该碎片为边缘基准碎片；否则不是。这样可以选取到 22 组边缘基准碎片对（即一碎片的正反面）。

1. **分组标准的确定**

若（a1，b1）为左侧基准碎片，（a2，b2）为右侧比较碎片，那么有下图所示的两种比 较方式。



同问题二的基准线误差计算方式，可以得到两组数据： （（lt1,lb1），（lt2,lb2））。其中,(lt1,lb1) 表示在第一种情况下匹配，上下面基准线的差值；（lt2,lb2）表示在第二种情况下匹配，上下面基准线的差值。

为了便于比较，作如下简化：对于每一次匹配可以得到一组数据(l1,l2)， 其中，

l1=lt1+lb1,l2=lt2+lb2。

1. **组内匹配算法的改进**

对于同一组内的碎片，同样可以以误差评估函数ch1(i,j)来衡量两个碎片之间s的匹配程度。不同的是在匹配的过程中，我们可以计算得到上下两面的误差值分别为ch1(i,j)，ch2(i,j)，为了便于比较，同样以ch(i,j)= ch1(i,j)+ ch2(i,j)来作为衡量匹配程度的标准。ch(i,j)越小，匹配度越好，即可作为被匹配碎片的右侧碎片。

1. **行间匹配的改进**

利用问题二中行间匹配的算法，计算正反面的误差评估值的综合指标，即：

Cw(i,j)= cw1(i,j)+ cw2(i,j)

最终可以拼接复原得到原始文件。

西南交通大学

问题三的基本求解思路与问题二一致。

1. **模型准备**

* **图像的预处理**

同问题一，建立每张图片的0-1矩阵。

* **正、反面特征矩阵**

利用问题二中中英文灰度条的构建方法，得到图片k的a面特征灰度条，再根据特征灰度条得到a面的特征列向量：





同理，得到b面的特征列向量：





1. **模型建立**

* **两次特征匹配模型**

第一次特征匹配：

将碎片k的a面与碎片s(s=0,1...208且s≠k)的a面进行特征比较，即求碎片k的a面特征列向量D(k) a与碎片s的a面特征列向量D(s) a对应的元素的差的绝对值之和，得到特征值Ra,a k,s:



再将碎片k的a面与碎片s(s=0,1...208且s≠k)的b面进行特征比较，得到特征值Ra,b k,s:





取一个合适小的置信区间[c,d]，若Rl k∈[c,d]，进行第二次特征匹配：

1. 



若取一个合适小的置信区间[e,f]，若Rb,b k,s∈[e,f],则认为碎片k与碎片s的匹配方式为k的a面与s的a面处于一面的同一行，k的b面与s的b面处于另一面的同一行。

（2）



若取一个合适小的置信区间[e,f]，若Rb,a k,s∈[e,f],则认为碎片k与碎片s的匹配方式为k的b面与s的a面处于一面的同一行，k的a面与s的b面处于另一面的同一行。

* **左右边界匹配模型**

本问中此模型构建同问题二的思路。

* **上下边界匹配模型**

本问中此模型构建同问题二的思路。

1. **模型求解**

使用matlab求解，复原结果见论文。

Step1：读取图片数据，构建0-1矩阵；

Step2：任取碎片i依次与其他碎片进行二次特征匹配，确定出i与s的特定面为原文件的同一行；

Step3：重复Step2，将附件中的图片聚类，将相同特征的图片放入同一行；

Step4：加入人工干预，将类的个数降维，使得每类图片的个数相同。

Step5：取同一行中的第i张碎片，与该行中的其他碎片进行左、右边界匹配，找到最佳边界匹配方式，组合图片；

Step6：重复Step5，直至确定出每行内部图片的排列顺序；

Step7：取第i行，与其它行进行上、下边界匹配，找到最佳边界匹配方式，组合图片；

Step8：重复进行Step7，直至确定出每行的上下位置，得到图片的原始序列。

1. **优秀论文的思路总结**
2. **心得体会**

建模主要需要建立一个数学模型，可以是一组方程、一个可操作的框架或一个函数。总之，数学模型是解决数学问题的一种抽象方法。高中物理在解决物理问题上，实际上是数学建模的过程。例如，2016年全国大赛A的第一个问题可以用高中物理的正交分解法来解决，只要列出水平和垂直方向的能量平衡。以下内容会主要总结分析一下我在建模学习的过程中对于建模几个关键环节的感受与看法。

1. 关于选题

如何在不离开主题的情况下分析题目是建模的首要任务。看到一个题目，我们首先应该分析总体需求以及它属于哪个领域、你是否对这个问题感兴趣或者了解熟知。在选择一个主题时，一个团队应该选择一个大多数人认为不错的题目。下一步是找出问题中不熟悉的专业术语的含义。这也可以在百度搜索中找到。然后分析主题中的各种问题以及它们之间的关系，并大致了解可以使用什么方法来解决这些问题。我们自己做这个分析是不够的，我们必须结合参考资料来分析主题并选择主题。

1. 关于文献

很多老师都会强调论文的水平取决于参考文献的水平。仔细考虑是有道理的。如果你引用的文章非常笼统或过时，那么你的文章水平几乎一定不够高，或者长篇大论地写着其他人已经研究过的问题。如果你提到顶级文献，研究前沿问题，你参考这样文章做出来的水平也可想而知。

文献的主要来源是中国知网。首先，在搜索更多文章时使用高级搜索。以“学生成绩评价”为例，直接搜索“学生成绩评价”，搜到的结果都是含有“学生成绩的评价”整体的题目。如果在高级级别输入“学生”、“成绩”和“评价”，可能结果是“\*\*学生\*\*成绩\*\*评价\*\*”，这样可以找到更多实用而有效的内容。其次，善于改变关键词。还是以上的例子，学生可以换成学员，评价可以换成评判、分析、探讨等。关键词甚至可以适当增减，例如只搜索评价方法，可以看到许多优秀的方法，你可以拿来用在评价学生上面。

1. 关于如何挖掘创新思路

通过研读多篇优秀的建模论文我们会发现，一个创新而有效的思路常常是一篇建模文章出彩最重要的部分。除了对题目所在主题格外了解，思路水到渠成之外，还有什么方法可以挖掘到较为创新的思路呢？我觉得可能主要如下：

1. 反过来想一想，这个问题和我们熟知的问题是否有什么类似的地方？变动其中一些条件是否会造成很大的影响？
2. 对着题目里的关键字发散地想，把所有想到的东西记录下来，仔细考虑一下这些东西对问题的解决是否会存在帮助？
3. 对着想要的结果想，如果我引入一个和结果关系非常密切的变量，用初始的条件能否推出与它的关系？（如2017年b题有的优秀论文中提到的“对用户吸引力”）