手写数字的特征选择和提取的研究

第五组：曾蜀童，牛成麟，张顺康，石云逸

手写数字识别在很多领域具有广泛的应用，特征的选择和提取决定着识别的精度。针对手写数字多样性、不规则性，高点阵扫描数据量大的特点，通过对手写数字的特征进行量化，使用熵权法评价体系，综合特征信息识别手写数字。特征提取的目的是把数字的结构特征提取出来，减少数字的位移、大小、字形的改变对识别的干扰，提供反映数字特征的关键信息。

比如对于数字 7，我们抓住上方横线长度和空列比、下端点的宽度三个局部特征，用四个重要位点对特征进行描述，计算数字集 7 的 5000 个样本重要位点的量化值，并加以分析。选取不为 7 的数字集中的样本，计算重要位点的量化值，判断是否满足数字 7 的特征，求得该模型的识别准确率，其中有 7 个数据集的识别准确率在 90%以上，表明正确判断率较高。

对任意两个不同类别的手写数字集所构成的集合进行研究，并获取特征提取的方法。分析集合中样本的形状、位置等拓扑结构，通过寻找两个数字间的不同点，分析有明显差异的量化值，通过观察 7 组较有特点的数字集合，比较特征的量化值，提取出 7 个结构特征，分别是欧拉数、纵横比、旋转相似度、笔画密度、轮廓变化趋势、重心距离和凹凸性。对每一组集合，通过一个数字集的样本检验另一个数字对特征的敏感程度。计算的识别准确率大部分在 70%以上，由于只是用一个特征进行判断，所以模型是较为可靠的。

在此基础上，分析 7 个特征对 10 个数字集样本的识别能力。除了数字结构的特征提取，还对每个集合中 5000 多份样本的特征信息进行统计，使主观判断结合样本的客观数据，确定量化指标值。运用熵权法，判断第t个特征对数字集i的相关程度，计算数字集i的训练样本在第t个特征上量化后的数值，确定其置信区间。计算待判断数字相应特征的量化值，若在数字集i的置信区间内，则将待判断数字在数字集i的相似度得分加上该特征的权重，否则不加，最终比较该待判断数字在 10 个数字集的相似度总得分。选出其中得分最高的数字，作为判断结果。使用 0-9 数字集中随机样本进行检验。

符 号	说 明
a	数字 7 下端笔画的宽度
b	数字 7 的空列比
c	数字 7 上半部分横线长度
D	原空间的维数
D'	特征空间的维数
μ	量化特征的均值
A	原像素矩阵
B	旋转后像素矩阵
f	矩阵上边界的排数
g	矩阵下边界的排数
w_1	图像 1/2 处的宽度
w_2	图像 1/4 处的宽度
α	图像总重心
p_{ij}	第 i 个矩形中，对其第 j 排从左往右遇到第一个黑点的步长
Δp	凹凸性指标
X	熵权法判断矩阵
h_{it}	数字集 i 在第 t 个特征上的基础得分
k_{it}	数字集 i 在第 t 个特征上的加权后得分

A 数据预处理

对扁平化后的高点阵扫描的矩阵数据，存储为0-255范围的灰度值，因此需要将原样本图像还原，对灰度进行处理。

1）还原样本：由于$784=28\times28$,所以将784维向量逆扁平化转化为28×28像素矩阵的图像，使用MATLAB的imshow命令画图，绘出手写样本图像。

2）灰度处理：为避免较淡的笔画边缘影响，将灰度值在100以下的点删去，即将灰度视为0。只保留笔画较重的点，使数字轮廓更加清晰。

由于样本自身噪声较少，不再进行降噪处理。经过预处理后，样本更加清晰易读，便于进一步处理。图1为数字集0中的部分样本经预处理后的结果。

图1 部分手写数字图像样本

B 重要位点的选取及量化

1.观察

观察手写样本，发现数字7有两个较明显的特征：1)数字7的下半部分为一条直线段；2)上半部分有一段近似平行的笔画。选取数字集合7的前5000个训练样本，使用MATLAB程序对重要位点反映的特征进行量化，统计量化结果，分别求出特征位点的平均值、标准差，计算置信区间。特征位点量化的具体过程如下：

1）下半部分直线段的量化：设样本笔画宽度为a，其值为该行最右边的像素位点与最左边的像素位点横坐标之差。在28×28像素的图像中从下向上找到第一个有像素的点，其行数记为i，考虑到书写时笔锋会使底部单个像素突出，故考虑倒数第$i-1$行，将左右两边有灰度值的位点作为位点1和位点2，计算第$i-1$行的笔画宽度a_{i-1}，如图2中1-1所示。对数字集合7的前5000个样本进行计算，其值在3左右。则带判断数下半部分直线段的值越接近3，判断为数字7的可能性越大。

2）上半部分横向笔画的量化：从横线的长度和连续程度两方面进行考虑。设数字上半部分横线长度为c，将带灰度的左上角和右上角的像素点作为位点3,和位点4，计

算两点的横向距离，如图 2 中 1-1 所示，即为横向笔画的长度。横线长度越长，越接近数字 7。

引入空列比b的概念。分别向上下、左右方向延伸至相交，形成矩形。对矩形的每一列进行检验，称没有像素的列为空列，空列总数比上矩形的列总数，即为空列比。空列比越小，说明位于图像上方的越接近直线，图 2 中 1-2 和 1-3 表示了数字 7 和 4 矩形内的空列比。

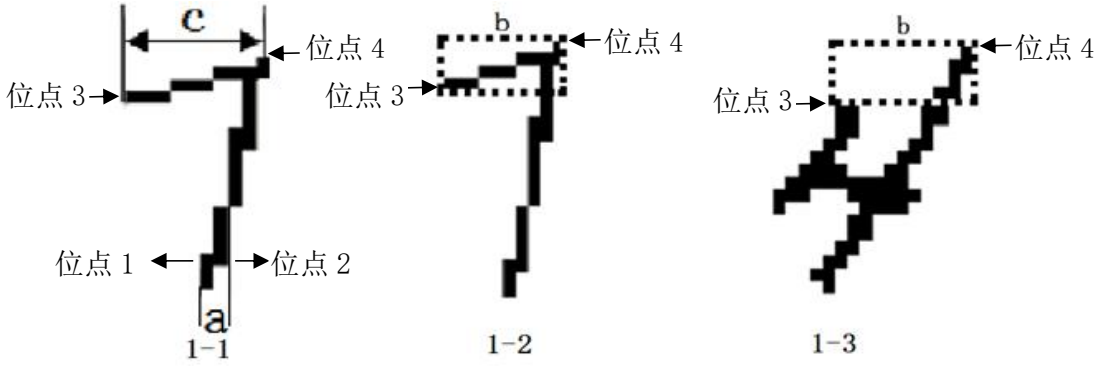

图 2 四个重要位点的定义示意图

2 识别准确率的检验

用 MATLAB 程序对数字 7 的重要位点的量化结果进行正态分布检验，可得三个位点指标均服从正态分布，置信水平与置信区间取值如表 1 所示。

表 1 数字 7 重要位点量化的置信区间

重要位点	下端点的宽度a	上方横线长度c	空列比b
置信水平	95.8%	98.3%	95.18%
置信区间	0.354-4.005	8．956-19.537	0-0.429

使用 MATLAB 程序，对数字集 7 进行识别准确率检验，取其余 9 个数字集中的 1000 个样本，对待判断数字的特征位点进行计算，若量化值均落在置信区间内，则判断该数字为 7，否则不是。计算结果如表 2 所示。

表 2 数字集 7 的识别准确率

数字集	0	1	2	3	4	5	6	8	9	平均
准确率(%	95.2	97.6	98.7	93.8	90.7	86.4	96.6	75.5	63.6	88.7

结果表明，数字集 7 重要位点识别准确率整体表现较好。对数字 0、1、2、6 的识别的正确率较高，超过了 95%，对数字 3、4、5 的正确率超过了 85%，只是对数字 8 和 9 的正确率较低。由于位点 1 和位点 2 限制了数字下端点宽度，排除了数字集 2、3、5 中的绝大多数检验样本，保证了数字 2、3、5 的准确率。位点 3 和位点 4 对上半部分的横线做了限制。横线长度c排除了数字 0、1、6 误判为 7 的可能性，保证了数字 0、1、6 判断的准确性。空列比最直接的排除了 4，由于 4 与 7 的空列比差距最大，所以保证了

数字 4 的准确率。但由于数字集 8、9 中一部分样本的书写形状满足了位点的限制，数字 8 和 9 在上半部分的圆弧由于像素较低，被误判为直线，满足数字 7 的长度和空列比。由于数字 8 书写的不规范，数字 9 下半部分特征与数字 7 相同，因此下端点的长度也没能很好的将数字 8 和 9 判断出来，从而导致准确率不是很高。

3 分析与总结

任务一选取了数字集 7 作为研究对象，通过四个重要位点反映了下端点的宽度、空列比、上方横线长度三个特征信息，对数字集 7 中的 5000 个训练样本的三个特征进行量化，统计均值标准差及其分布情况，计算判断区间。用剩下的 9 个数字集合中的样本对此判断方法进行检验，当待判断数字通过位点所得特征的值均在可接受的区间内时，则判断它为数字 7，否则判断不是数字 7。通过识别准确率这一指标，反映位点选取和特征选取的好坏。结果表明，数字 7 的识别准确率较高，平均为 88.7%，除个别数字外，识别准确率均在 90%以上。

C 任意两个数字的比较

选取两个不同类别的手写数字集所构成的集合，分析该集合的特征，比较两个手写数字的异同点，获取它们的特征。特征提取的概念是通过映射或变换的方法，把原空间的高维特征变成特征空间的低维特征，即用原始特征映射得到的较少新特征描述样本。一般情况下，特征空间的维数 d 远小于原空间的维数 D。特征提取的实质相当于，在一定条件下的一种变换 T，实现原空间 E_R 到特征空间 E_D 的映射，即 $T: E_R \to E_D$。特征提取实现了维数压缩，但应尽量不改变样本的属性。并且特征提取应更具代表性，更能反映判断手写数字的本质特点。

1 两个数字集合的研究和特征获取

10 个数字两两组合共有 45 组情况，经观察发现，其中 7 组数字组合在结构上有较大的差异，主要体现在整体结构，笔画的位置、粗细等特征。对观察得到有较大差异的数据集合进一步分析，通过对特征的量化，确定差异大小。判断其他组中此特征是否相似，计算相似组中特征的量化值，与差异较大组中的值进行比较，分析原因。其具体过程如下：

1)数字集 $\{0,8\}$ 的差异：观察发现数字 8 有两个封闭的环，定义其空洞数为 2，而数字 0 只有一个封闭的环，其空洞数为 1。数字 0 和 8 均没有未封闭的笔画。而观察数字集 $\{6,9\}$：它们均有一个中空的环，环的上面或下面连着一条弯曲的笔画，即它们空洞

的数量为 1，未封闭的笔画数为 1。故易得出空洞数与碎片数的差取决于数字的形状。查阅资料，可用欧拉数对此特征进行描述，其值定义为：欧拉数＝空洞数－碎片数。数据集的欧拉数如表 3 所示。

表 3 集合样本的欧拉数

数字集	0	6	8	9
均值	0.966	0.768	1.669	0.852
标准差	0.360	0.591	0.652	0.479

2)数字集{0,1}的差异：观察数字 0 的宽度明显大于数字 1 的宽度，而其高度略小于数字 1，则两个数据集高度和宽度的比例有较大差异。分析数字集{5,6}的共同点，它们的外形结构比较相似，且高度和宽度很相近。高度和宽度的比例包含了字符重要的几何特征信息，将纵横比总结为第二个特征，其值定义为：纵横比＝高度/宽度。数据集的纵横比如表 4 所示。

表 4 集合样本的纵横比

数字集	0	1	5	6
均值	2.207	8.898	3.670	3.665
标准差	0.424	3.003	1.017	1.027

3)数字集{1,7}的差异：对比两个数字的总体轮廓，易得数字 1 具有良好的上下对称性，经过 180° 的旋转，图形基本不变，而数字 7 经过上下翻转后变化很大，不具有旋转不变性。而分析数字集{0,1}的共同点，两个数字具有高度的对称性，上下旋转后图形基本不变，因此将旋转相似性作为第三个特征，其值定义为：旋转相似度＝旋转后不变的像素点数/总像素点数。数据集的旋转相似度如表 5 所示。

表 5 集合样本的旋转相似度

数字集	0	1	7	8
均值	0.523	0.523	0.248	0.427
标准差	0.241	0.241	0.109	0.144

4)数字集{0,7}的差异：不同的数字是由不同的笔画构成，数字 0 由一条封闭的曲线围成，而数字 7 是由一条向下开口的折线构成。两者在大小相同的情况下，数字 7 的笔画长度更小，即在最小外接矩形中，数字 0 的相对笔画密度比 7 大。观察数字集{4,5}的笔画，在书写时，4 和 5 的笔画较多，它们需要的像素点也较多，在外接矩形中它们相对的笔画密度差异不大。因此笔画密度能够反映数字笔画的总体趋势，其值定义为：笔画密度＝有像素的点数/外接矩阵的总点数。其量化如表 6 所示。

表 6 两个集合的手写数字训练样本笔画密度统计

数字集	0	4	5	7
均值	0.422	0.325	0.329	0.313
标准差	0.081	0.065	0.079	0.066

5)数字集{0,7}的差异，每个数字的轮廓有较明显的变化规律，如数字 0 从上到下外轮廓宽度的变化先由小到大，再从大到小（如图 3 所示）。而数字 7 从上到下的外轮廓宽度的变化趋势从大突然变小，此后基本不变。分析数字集{0,4}的共同点，虽然数字 0 外形比较光滑，线条弯曲程度平缓，而数字 4 外形尖锐，线条较直，但它们从上到下的外轮廓宽度的变化趋势相同，均先从小到大，再从大到小。因此概括轮廓变化趋势作为一个特征，定义其值为：轮廓变化趋势=中部宽度/上部宽度。其中上部取最小外接矩形的四分之一处，中部取二分之一处。其量化如表 7 所示。

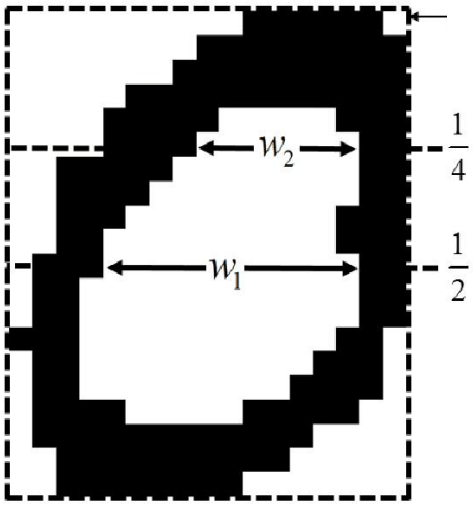

图 3 轮廓变化趋势示意图

表 7 两个集合的手写数字训练样本轮廓变化统计

数字集	0	4	7
均值	1.3143	1.3337	0.6061
标准差	0.2296	0.7268	0.5672

6)数字{7,8}的差异，将样本的最小外接矩形均分成四个区域，因为数字 8 具有明显的对称性和稳定性，所以它的总重心和四个区域重心的平均位置几乎重合。观察图 4，数字 7 的总重心α和分区重心$\alpha_1,\alpha_2,\alpha_3,\alpha_4$的平均位置明显不同。观察数字集{0,1}的重心，由于两个数字集的样本均具有轴对称性，因此具有总重心与四个区域重心的中点重合的特点。由于不同数字的笔画不同，像素点的分布不同，造成不同字符的重心位置不同，因此重心位置的偏移很好的反映了不同数字的特征，将其作为识别数字的一个重要特征。其量化值如表 8 所示。

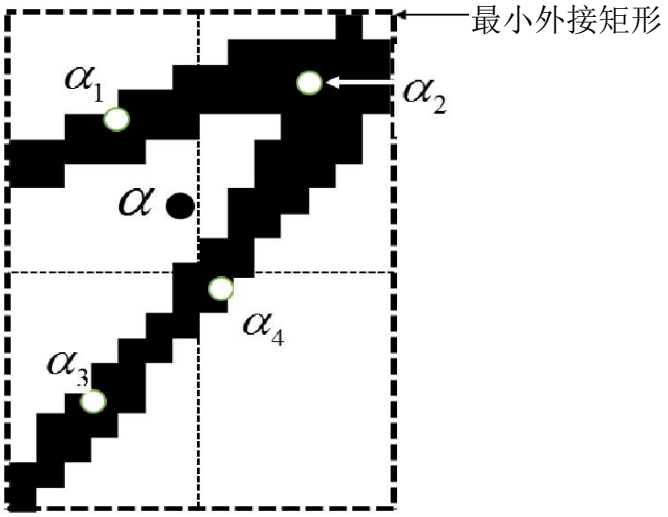

图 4 重心距离示意图

表 8 两个集合的手写数字训练样本重心距离统计

数字集	0	1	7	8
平均值	25.824	16.831	23.7868	22.4418
标准差	2.4992	2.5399	2.8963	1.7618

7)数字集{3,7}的差异，凹凸性是刻画图像的重要指标，观察数字集{3,7}，尝试寻找凹凸性的特点。易得两个数字集的上半部分是向右凸出的，但下半部分有明显差异，数字 3 的下半部分向右凸出，而数字 7 的下半部分略向左凸出。再分析数字集{3,6}的凹凸性，如图 5，将最小外接矩形分为按上下左右分成四个部分 $\Delta p_1, \Delta p_2, \Delta p_3, \Delta p_4$，每部分均分成三行，令 p_1, p_2, p_3 分别为上中下三部分首次出现像素点的横坐标，计算中间首次出现像素点的横坐标与上下两部分首次出现像素，即 $\Delta p_i = \max p_2 - \min\{p_1, p_3\}$ i=1,2,3,4，Δp_i 越大，则说明数字的凹凸性越强。观察数字 6 易得，其右下部分向右凸出，与 3 的凹凸性相像。由此将凹凸性作为一个特征信息。其量化值如表 9 所示。

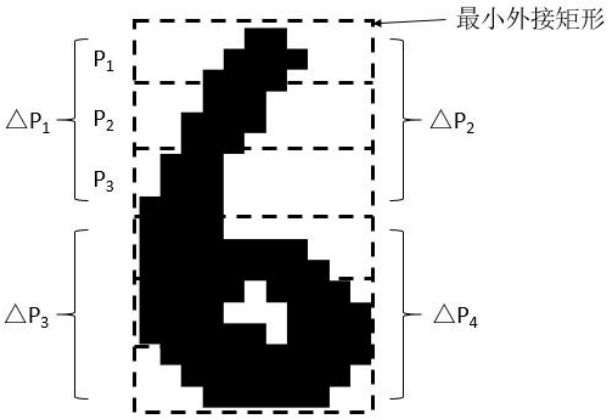

图 5 凹凸性示意图

表 9 两个集合的手写数字训练样本凹凸性统计

数字集	统计	左上	右上	左下	右下	凹凸性
3	平均值	0.7296	0.2776	0.8536	0.6806	0.634
	标准差	0.4442	0.4479	0.3535	0.4663	0.22714
5	平均值	0.4430	0.7578	0.7544	0.6388	0.6534
	标准差	0.4968	0.4285	0.4305	0.4804	0.23216
7	平均值	0.8286	0.2452	0.1578	0.0624	0.3194
	标准差	0.3769	0.4302	0.3646	0.2419	0.15478

2 特征指标的量化

对每组两个手写数字组成的集合，要对总结出的特征识别方法进行准确率验证，则需要先对该特征的评判标准进行量化。再以量化结果作为判断未知数字的依据。

1)对集合{0,8}，随机选取数字 0 和数字 8 样本中的各 1000 个作为训练样本，求出该集合的 2000 个欧拉数值，经 MATLAB 程序验证符合正太分布。计由“小概率事件”和假设检验的基本思想，使不符合该特征的数字 0 或数字 8 比例小于 5%，根据正态分布的“3σ 原则”取判断的区间$(\mu-3\sigma,\ \mu+3\sigma)$。算得数字 0 的欧拉数均值$\mu=0.966$，数字 8 的欧拉数均值$\mu=1.669$。则数字 0 的欧拉数置信区间为$(0.966+0.36，0.966+0.36)$，数字 8 的欧拉数置信区间为$(1.669-0.652，1.669+0.652)$。

同理，对下列各集合中的两个数字所对应的特征进行量化，并取合适的置信区间。

2)对集合{0,1}，数字 0 的纵横比置信区间为$(2.207-0.8,2.207+0.8)$，数字 1 的纵横比置信区间为$(8.898-5,8.898+5)$。

3)对集合{8,7}，数字 8 的旋转相似度置信区间为$(0.248-0.15,0.248+0.15)$，数字 7 的旋转相似度置信区间为$(0.427-0.11,0.427+0.11)$。

4)对集合{0,7}，数字 0 的笔画密度置信区间为$(0.422-0.07,0.422+0.07)$，数字 7 的笔画密度置信区间为$(0.313-0.07,0.313+0.07)$。

5)对集合{0,7}，数字 0 的轮廓变化趋势置信区间为(1.314-0.27,1.314+0.27)，数字 7 的轮廓变化趋势置信区间为(0.606-0.56,0.606+0.56)。

6)对集合{8,1}，数字 8 的重心距离置信区间为(16.831-3.5,16.831+3.5)，数字 1 的重心距离置信区间为(22.44-2.5,22.44+2.5)。

7)对集合{3,7}，数字 3 的凹凸性置信区间为(0.634-0.207,0.634+0.207)，数字 7 的凹凸性置信区间为(0.3194-0.18,0.3194+0.18)。

3 识别准确率的计算

由于以上每个集合分别对应一个特征，将第一个数字的 1000 个测试样本代入第二个数字在该特征上的置信区间，若落在置信区间内，则判断其为第二个数字，否则判断为第一个数字。反之，将第二个数字的 1000 个测试样本代入第一个数字在该特征上的置信区间，若落在置信区间内，则判断其为第一个数字，否则判断为第二个数字。

1)对集合{0,8}，将数字 0 的 1000 个测试样本代入数字 8 在该特征上的置信区间，若落在置信区间内，则将该数字判断其为数字 8，否则判断为数字 0，经计算，得出有 23%的数字 0 样本被判断为数字 8，即误判率为 23%；反之，将数字 8 的 1000 个测试样本代入数字 0 在该特征上的置信区间，若落在置信区间内，则判断其为数字 0，否则判断为数字 8，经计算，得出有 2%的数字 8 样本被判断为数字 0，即误判率为 2%。

同理，对集合 2）到集合 7）中的两个数字互相进行识别准确率的验证，总结果如表 10。

表 10 各数字集合互相判断的准确率检验结果

特征	欧拉数		纵横比		旋转相似		笔画密度		轮廓变化		重心距离		凹凸性	
测试样本	0	8	0	1	8	7	0	7	0	7	8	1	3	7
判断区间	8	0	1	0	7	8	7	0	7	0	1	8	7	3
准确率(%)	77	98	100	98	72	71	70	81	84	73	93	91	74	63

4 检验

由于特征提取的主要目的识别数字，故用各数字自身的 1000 个样本进行类似以上的检验，例如对集合{0,8}，其对应特征为欧拉数，将数字 0 的 1000 个测试样本代入数字 0 本身在欧拉数上的置信区间，若落在置信区间内，则判断其为数字 0，否则判断不为数字 0，统计判断正确的概率为 86%。同理，对集合 2）到集合 7）中的每数字进行自身识别准确率的验证，总结果如表 11。

表 11 各数字集合自身判断的准确率检验结果

特征	欧拉数		纵横比		旋转相似		笔画密度		轮廓变化		重心距离		凹凸性	
测试样本	0	8	0	1	8	7	0	7	0	7	8	1	3	7
判断区间	0	8	0	1	8	7	0	7	0	7	8	1	3	7
准确率(%)	86	91	98	98	70	78	60	66	80	93	93	91	85	93

5 分析与总结

对选取的 7 个集合、两个不同类别的手写数字集所构成的集合研究不同点，获取特征提取的方法。分析集合中样本的形状、位置等拓扑结构，通过寻找两个数字间的不同点，分析有明显差异的量化值，通过观察 7 组较有特点的数字集合，比较特征的量化值，提取出 7 个结构特征，分别是欧拉数、纵横比、旋转相似度、笔画密度、轮廓变化趋势、重心距离和凹凸性。对每一组集合，通过一个数字集的样本检验另一个数字对特征的敏感程度。得出的识别准确率大部分在 70%以上，由于只是用一个特征进行判断，所以模型是较为可靠的。

D 0-9 手写数字集的特征选择和提取

特征提取的目的就是把数字的某些结构特征提取出来，使数字的位移、大小变化、字形畸变等干扰相对减小，而把那些反映数字特征的关键信息提供给模型，间接地增加模型的容错能力，降低误识率和拒识率，可见为了有效地进行数字识别，特征提取是必要的。

1 基于结构模式的特征提取

首先抓住手写数字本质不变的整体拓扑结构，类似于人类通过视觉对图像进行处理。将图像逐级分解成部件、笔划或笔段，由像素得到笔划，由笔划结合成部件，提取其中的特征，根据各位点或属性的关系进行判断，这样避免了数字形状随人书写风格而变化。基于任务二中通过两两对比数字集合的共同特征，在此基础上总结出 0-9 手写数字集的特征选择和提取的基本方法。

1)欧拉数：在拓扑学中，欧拉数表示空间完整性，欧拉数＝空洞数－碎片数。用 MATLAB 软件自带程序计算样本欧拉数，统计结果见表 12。

表 12 各手写数字训练样本的欧拉数统计

数字集	0	1	2	3	4	5	6	7	8	9
均值	0.966	-0.055	0.355	0.063	0.076	0.051	0.768	0.080	1.669	0.852
标准差	0.360	0.265	0.554	0.370	0.368	0.376	0.591	0.373	0.652	0.479

2)纵横比：扫面数字图形最靠上、下、左、右的四个像素点，定义上下两点的竖直距离为图形的纵长d_1，左右两点横向距离为图形的横长d_2，纵横比$\frac{d_1}{d_2}$，统计结果见表 13。

表 13 各手写数字训练样本的纵横比统计

数字集	0	1	2	3	4	5	6	7	8	9
均值	2.207	8.898	3.244	3.567	3.584	3.670	3.665	4.590	3.144	3.978
标准差	0.424	3.003	0.831	1.120	0.937	1.017	1.027	1.321	0.715	1.016

3)旋转后相似度：先将图形做类似于二值化的处理：用 MATLAB 程序将灰度低于 150 的像素点视为 0，高于 150 的像素点视为 1。设原矩阵为 A，将数字顺时针旋转 180°，得到新的矩阵 B，从 1 到 28 遍历 i, j，依次判断 A、B 矩阵的 ij 元是否相同，用相同的数量比上矩阵的阶数，得到旋转后的相似度值，统计结果见表 14。

表 14 各手写数字训练样本的旋转后相似度统计

数字集	0	1	2	3	4	5	6	7	8	9
均值	0.523	0.523	0.374	0.375	0.384	0.329	0.340	0.248	0.427	0.377
标准差	0.241	0.241	0.152	0.135	0.142	0.152	0.115	0.109	0.144	0.115

4)笔画密度：先用最小外接矩形将数字框起来：用 MATLAB 程序扫面数字图形最靠上、下、左、右的四个像素点，定义上下两点的竖直位置为图形的上下边界，左右两点的横向位置为图形的左右边界，四条直线相交得到一个矩形，称为最小外接矩形。再同上面的方法将矩形内的点阵做同上类似于二值化的处理。然后统计最小外接矩形内值为 1 的点数占整个最小内接矩形的点数的比例，得到笔画密度，统计结果见表 15。

表 15 各手写数字训练样本的笔画密度统计

数字集	0	1	2	3	4	5	6	7	8	9
均值	0.422	0.397	0.364	0.375	0.325	0.329	0.390	0.313	0.417	0.374
标准差	0.081	0.146	0.073	0.078	0.065	0.079	0.074	0.066	0.085	0.069

5)轮廓变化趋势：首先同上，用最小外接矩形将数字框起来。设该矩形上下边界分别为第 f 排和第 g 排，找第 $\frac{f+g}{2}$ 排最左边有像素的点到最右边有像素的点的距离，即为 $\frac{1}{2}$ 处的宽度 w_1，找第 $\frac{3f+g}{2}$ 排最左边有像素的点到最右边有像素的点的距离，即为 $\frac{1}{4}$ 处的宽度 w_2，定义轮廓变化趋势为 $\frac{w_1}{w_2}$，统计结果见表 16。

表 16 各手写数字训练样本的轮廓变化趋势统计

数字集	0	1	2	3	4	5	6	7	8	9
均值	1.3143	1.0483	0.6623	1.2108	1.3337	1.333	2.1116	0.6061	0.7274	0.8409
标准差	0.2296	0.2766	0.5288	0.7168	0.7268	1.107	1.5283	0.5672	0.3997	0.3206

6)重心距离：首先定义重心的位置$(\bar{m},\bar{n})$, $m=1,2,\ldots,M-1,\ n=1,2,\ldots,N-1$。

$$\bar{m}=\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}f_{m,n}\frac{m}{\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}f_{m,n}},\bar{n}=\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}f_{m,n}\frac{n}{\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}f_{m,n}} \quad (1)$$

其中

同上，用最小外接矩形将数字框起来，求出重心位置α。然后将该最小外接矩形平均分为四个小的矩形，分别求它们的重心α_1，α_2，α_3，α_4，计算α到α_i（1≤i≤4）的欧式距离，求和得到重心距离，统计结果见表17。

表17 各手写数字训练样本的重心距离统计

数字集	0	1	2	3	4	5	6	7	8	9
平均值	25.82	16.83	25.68	24.84	21.00	24.07	22.02	23.78	22.44	20.55
标准差	2.499	2.539	2.667	2.140	2.508	2.86	2.367	2.896	1.761	2.042

7)凹凸性：首先同上，用最小外接矩形将数字框起来。然后将该最小外接矩形横向平均分为六个小的矩形（若像素的排数不是6的倍数，则四舍五入取近似值）。第i个矩形中，对其第j排从左往右扫描，遇到第一个黑点时行走的步长记为p_{ij}。比较上面三个矩形：取p_{1j},p_{2j},p_{3j}的平均值分别记为p_1,p_2,p_3，步长差$\Delta p_1=(p_2-p_1)+(p_2-p_3)$,再用相同方法从右往左扫描，得出步长差$\Delta p_2$，同理，比较下面三个矩形，得到从左往右的步长差$\Delta p_3$，从右往左的步长差$\Delta p_4$。用$1\times4$的矩阵表示数字的凹性：$(\Delta p_1,\Delta p_2,\Delta p_3,\Delta p_4)$，取平均值得到凹凸性指标$\Delta p$，统计结果见表18。

表18 各手写数字训练样本的凹凸性统计

数字集	统计	左上	右上	左下	右下	凹凸性
0	平均值	0.0938	0.1452	0.1994	0.4246	0.2144
	标准差	0.2916	0.3523	0.3996	0.4943	0.21891
1	平均值	0.0304	0.0118	0.0272	0.0218	0.0235
	标准差	0.1717	0.1080	0.1627	0.1460	0.08792
2	平均值	0.8030	0.1732	0.7164	0.7046	0.5978
	标准差	0.3978	0.3785	0.4508	0.4563	0.21143
3	平均值	0.7296	0.2776	0.8536	0.6806	0.634
	标准差	0.4442	0.4479	0.3535	0.4663	0.22714

4	平均值	0.4236	0.2436	0.8346	0.1500	0.4101
	标准差	0.4942	0.4293	0.3716	0.3571	0.18335
5	平均值	0.4430	0.7578	0.7544	0.6388	0.6534
	标准差	0.4968	0.4285	0.4305	0.4804	0.23216
6	平均值	0.0204	0.4140	0.2174	0.8144	0.3687
	标准差	0.1414	0.4926	0.4125	0.3888	0.17797
7	平均值	0.8286	0.2452	0.1578	0.0624	0.3194
	标准差	0.3769	0.4302	0.3646	0.2419	0.15478
8	平均值	0.4944	0.5120	0.2628	0.3134	0.3978
	标准差	0.5000	0.4999	0.4402	0.4639	0.23879
9	平均值	0.4188	0.2278	0.6672	0.025	0.3324
	标准差	0.4934	0.4195	0.4713	0.1561	0.1787

2 基于统计模式的特征提取

确定了 8 个特征后，每一个手写数字在这 8 个特征上的反映不同，量化结果也不同，需要确定第i个数字在第 t 个特征上的权重，权重越大，说明第t个特征越能贴切地勾勒出第i个数字的特征和属性。熵权法是一种客观赋权方法。根据信息论基本原理，信息是系统有序程度的度量；而熵则是系统无序程度的度量。因此，可用系统熵来反映其提供给决策者的信息量大小，系统熵可通过熵权法得到。计算步骤如下：

(1)构建 10 个手写数字 8 个特征的判断矩阵$X=\left(x_{ij}\right)_{n\times m}$，建立评价指标体系。

$$\text{即 } X_{ij}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad i=1,2,\ldots,m,\ j=1,2,\ldots,n \tag{2}$$

(2)异质指标同质化处理

由于各项特征指标的计量单位并不统一,因此在用它们计算综合指标前,我们先要对它们进行标准化处理,即把指标的绝对值转化为相对值,并令 xij=|xij|，从而解决各项不同质指标值的同质化问题。而且，由于正向指标和负向指标数值代表的含义不同(正向指标数值越高越好，负向指标数值越低越好)，因此,对于高低指标我们用不同的算法进行数据标准化处理。其具体方法如下:

越大越优型（正向指标）：$$X_{ij}=\frac{x'_{ij}-\min(x'_{1j},x'_{2j},\cdots,x'_{nj})}{\min(x'_{1j},x'_{2j},\cdots,x'_{nj})-\max(x'_{1j},x'_{2j},\cdots,x'_{nj})} \tag{3}$$

越小越优型（负向指标）：$$X_{ij}=\frac{\max(x'_{1j},x'_{2j},\cdots,x'_{nj})-x'_{ij}}{\max(x'_{1j},x'_{2j},\cdots,x'_{nj})-\min(x'_{1j},x'_{2j},\cdots,x'_{nj})} \tag{4}$$

则X_{ij}为第i个手写数字的第j个特征指标的数值，$i=1,2\ldots,n$,$j=1,2,\ldots,m$ 。

(3)计算第i个手写数字的第j个手写数字的比重

$$P_{ij}=\frac{x_{ij}}{\sum_{j=1}^{m}x_{ij}} \tag{5}$$

(4)根据熵的定义，根据各项特征指标，可以确定各特征指标的熵，若取$k=\frac{1}{\ln m}$，则$0\le e_i\le 1$。计算第i个特征指标的熵值

$$e_i=-k\sum_{j=1}^{m}(P_{ij}\cdot\ln P_{ij}) \tag{6}$$

(5)计算第i个特征指标的差异系数：

对于给定的e_i越大，因素评价值的差异性越小，则因素在综合评价中所起的作用越小。定义差异系数$g_i=1-e_i$。则当因素g_i越大时，因素越重要。

(6)定义熵权。定义了第n个特征指标的熵后，可得到第n个特征指标的熵权，熵权法最终确定的权重W_{ij}算法为

$$W_{ij}=\frac{g_{ij}}{\sum_{j=1}^{m}g_{ij}}$$

(7)

由MATLAB程序得到，各手写数字在各特征上的权重值，如表19。

表19 各手写数字在各特征上的权重体系

判断权重	欧拉数	纵横比	旋转相似度	笔画密度	轮廓变化趋势	重心距离	凹凸性	总和
0	0.0145	0.0185	0.1031	0.0223	0.0127	0.0056	0.8229	1
1	0.0080	0.0140	0.0389	0.0199	0.0088	0.0032	0.9069	1
2	0.1687	0.0581	0.1395	0.0324	0.4388	0.0109	0.1513	1
3	0.1417	0.0990	0.1457	0.0536	0.3583	0.0073	0.1942	1
4	0.1092	0.0725	0.2115	0.0487	0.2782	0.0161	0.2634	1
5	0.1251	0.0532	0.2179	0.0554	0.3839	0.0152	0.1489	1
6	0.1406	0.0606	0.1039	0.0316	0.4114	0.0091	0.2425	1
7	0.1039	0.0669	0.1646	0.0338	0.4238	0.0119	0.1947	1
8	0.0763	0.0525	0.1246	0.0440	0.1039	0.0064	0.5919	1
9	0.0965	0.0787	0.1140	0.0434	0.1525	0.0127	0.5019	1

总和	0.9850	0.5746	1.3641	0.3854	2.5727	0.0988	4.0190	10

易得重心距离这项指标在每个数字上的权重都小于 0.1，说明重心距离不能很好地反映各数字的特征属性，对于判断的影响可以忽略，故在最终的判断矩阵中删去重心距离指标，变成 10×6 的判断矩阵。

$$X = \begin{bmatrix} 0.0145 & 0.0185 & 0.1031 & 0.0223 & 0.0127 & 0.8299 \\ 0.0080 & 0.0140 & 0.0199 & 0.0088 & 0.0088 & 0.9069 \\ 0.1687 & 0.0581 & 0.1395 & 0.0324 & 0.4388 & 0.1513 \\ 0.1417 & 0.0990 & 0.1457 & 0.0536 & 0.3583 & 0.1942 \\ 0.1092 & 0.0725 & 0.2115 & 0.0487 & 0.2782 & 0.2634 \\ 0.1251 & 0.0532 & 0.2179 & 0.0554 & 0.3839 & 0.1489 \\ 0.1406 & 0.0606 & 0.1039 & 0.0316 & 0.4114 & 0.2425 \\ 0.1039 & 0.0669 & 0.1646 & 0.0338 & 0.4238 & 0.1947 \\ 0.0763 & 0.0525 & 0.1246 & 0.0440 & 0.1039 & 0.5919 \\ 0.0965 & 0.0787 & 0.1140 & 0.0434 & 0.1525 & 0.5019 \end{bmatrix}$$

3 各特征指标置信区间的确定

为了得出判断未知数字相应特征的量化值，对于数字集 i 的第 t 个特征，计算数字集 i 的训练样本在第 t 个特征上量化后的数值，并用 MATLAB 程序对这组数值取 95%以上的置信区间，如表 20。

表 20 各手写数字在各特征上的置信区间

	欧拉数	凹凸性	轮廓变化趋势	纵横比	笔画密度	旋转相似度
0	1	[0,0.5]	[0.914,1.714]	[1.507,2.907]	[0.272,0.572]	[0.173,0.873]
1	0	[0,0.25]	[0.438,1.658]	[3.898,13.898]	[0.117,0.677]	[0.093,0.953]
2	0,1	[0.25,1]	[0,1.462]	[1.494,4.994]	[0.214,0.514]	[0.115,0.635]
3	0,1	[0.25,1]	[0.160,2.261]	[1.667,5.467]	[0.225,0.525]	[0.115,0.635]
4	0,1	[0.25,0.75]	[0.286,2.387]	[2.083,5.084]	[0.195,0.455]	[0.124,0.644]
5	0,1	[0.25,1]	[0,3.033]	[2.010,5.330]	[0.179,0.478]	[0.049,0.609]
6	0,1,2	[0.25,0.75]	[0,4.511]	[2.005,5.325]	[0.259,0.520]	[0.110,0.570]
7	0,1	[0,0.5]	[0,2.006]	[2.441,6.739]	[0.192,0.433]	[0.018,0.478]
8	0,1,2,3	[0,0.75]	[0.127,1.327]	[1.921,4.367]	[0.261,0.572]	[0.177,0.677]
9	0,1,2	[0,0.5]	[0.280,1.400]	[2.367,5.589]	[0.248,0.499]	[0.167,0.587]

判断未知数字的第 t 个特征上量化后数值是否落在数字集 i 的第 t 个特征的置信区间内。设未知数字在数字集 i 的第 t 个特征上的基础得分为 h，若落在其第 t 个特征的置信区间内，则 h=1，否则 h=0。设未知数字在数字集 i 在第 t 个特征上的加权得分为 k，

则$k_{it} = h \times x_{it}$，未知数字在数字集 i 上的总得分为$\sum_{t=1}^{7} k_{it}$。最终比较该待判断数字在 10 个数字集的相似度总得分。选出其中得分最高的数字，作为判断结果。

4 检验

用 MATLAB 程序每次在 0-9 数字集中分别随机选取 100 个测试样本，对未知手写数字进行判断，统计判断正确百分比，多次运行程序，检验最终识别准确率。其中数字 1 的识别率高达 93%，数字 5 和数字 7 的识别率也分别达到了 76%和 83%，其余数字的准确率也超过了 60%，表明模型有较好的识别效果。

5 分析与总结

我们分析了 7 个特征对 10 个数字集样本的识别能力。除了数字结构的特征提取，还对每个集合中 5000 多份样本的特征信息进行统计，确定量化指标值。运用熵权法，判断第t个特征对数字集i的相关程度，计算数字集i的训练样本在第t个特征上量化后的数值，确定其置信区间。检验时，计算待判断数字相应特征的量化值，若在数字集i的置信区间内，则将待判断数字在数字集i的相似度得分加上该特征的权重，否则不加，最终比较该待判断数字在 10 个数字集的相似度总得分。选出其中得分最高的数字，作为判断结果。使用 0-9 数字集中随机样本进行检验，得出识别准确率在 60%-100%，结果表明模型有较好的识别效果。

E 总结

选择四个重要位点判断数字 7 的方案有待改进。虽然题目只要求用不同于数字 7 的其它集合判断，计算识别准确率进行检验，识别准确率的结果也较高，但在代入数字 7 的 1000 个测试样本时，发现对于数字 7 自身的判断正确率为 67%，说明评判标准过于严苛。结果表明，该模型对手写数字的判断仍有提升空间。大部分判断准确率在 60%以上，但准确率很难达到 90%的水平。因此模型的特征及其权重可以继续优化，提高判断准确率。

对手写数字的特征提取结合了结构特征和大量样本统计的两种方法。将不同数字的特征由局部到总体逐一分析，总结出欧拉数、纵横比、旋转相似度、笔画密度、轮廓变化趋势、重心距离和凹凸性 7 个特征，将各特征的数量、相互关系以及反映的方面作为判断识别的依据，最大限度地避免了数字识别受人书写风格变化的影响。运用人类视觉对图像进行处理、识别的原理，抓取数字本质不变的整体拓扑结构进行判别。同时，结合统计特征提取方法，将手写数字的识别看成一个模式分类问题。通过统计样本的数据，赋予不同特征因素对不同数字集影响的权重，将样本的信息最大化，使模型更加可靠。

通过对手写数字结构的特征提取，减少因数字位置、大小和字形变化对识别的影响。对反映数字特征的关键信息进行处理，间接地增加了模型的容错能力，降低误识率；而且通过特征提取，简化了程序的复杂度，减少了数据处理量和运算时间，使模型具有较强的适用性。