目录

数字字符识别——特征设计

1 问题描述

在第一次作业的基础上，尝试设计不同的特征并用于模板匹配，比较不同特征情况下对模板匹配性能的影响。

2 解决方案分析

本次实验的前期处理模块（主要是二值化处理）和对识别结果的正确率判定模块均与第一次实验相同。本次实验中，我尝试了两种特征，其中一种效果比较一般（正确率只有 60%-70%），另一种稍微好一点（正确率 80%-90%）。下面分别介绍。

2.1 方案一：基于分布的特征

该方案的主要思路是：对于每个数字模板（二值化后），统计其每列的黑点个数，将其组合一起，便得到了一个数字模板沿着列方向的黑色像素个数的分布图。这样处理后，每个数字模板对应的分布如图 1 所示（点数统一为 50）。

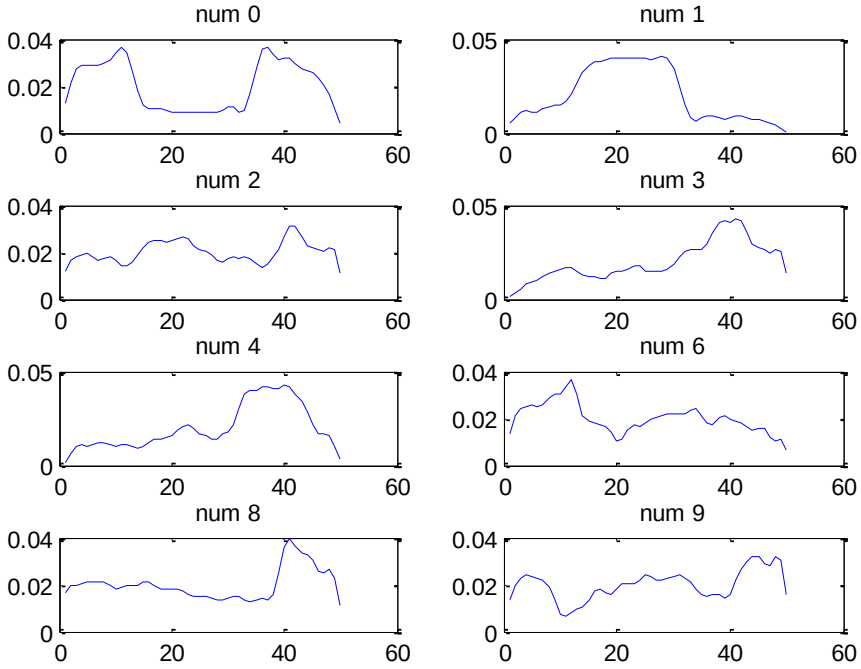

图 1 各数字模板沿列方向的黑像素个数分布

这样之后，我们可以将测试图像的若干连续行全部取出，按照同样的方法统计其沿列方向的黑色像素个数的分布。如果取出的连续行恰好包含分离的数字的话，我们可以得到如图 2 为例的分布图。

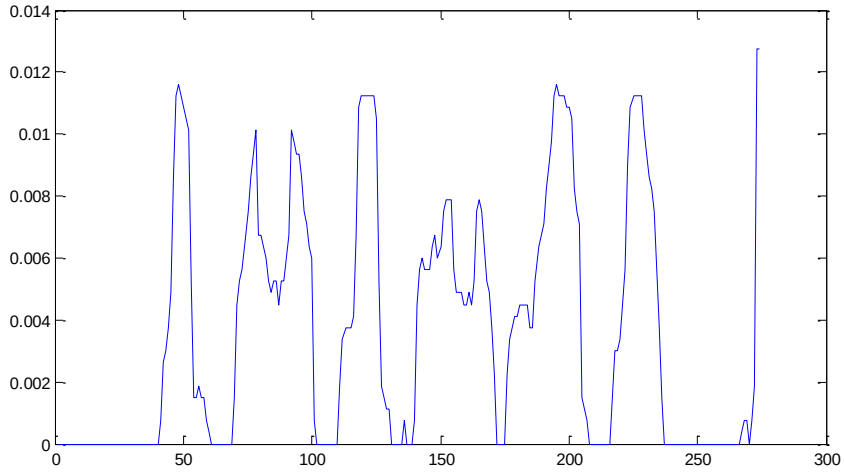

图 2 某若干连续行图像块沿列方向的黑像素个数分布

从图 2 可以看出，图中存在明显的几个相互分离的分布，将其中的每个分布单独取出（需要排除一些明显不可能是数字的分布，如峰值过大或过小等），依次计算与各个数字模板的分布相似性（需要先统一分布的点数），如果与某个数字模板的分布相似性高于某阈值，我们可以认定该区域被识别出数字。如图 2 中的第一个分布，与数字“1”的分布极为相似，表明该区域很可能是数字“1”。

至于如果衡量两个分布的相似性，显然不能用两个分布的相关，因为某个分布和全 1 分布（纯黑色区域）的相关总是最大的。通过查阅资料，KL 散度是衡量两个分布距离的一个度量。假设两个分布分别是 p 和 q，则其 KL 距离定义为

$$D_{KL}(p,q)=\sum_{i=1}^{n}p(i)\log\frac{p(i)}{q(i)} \tag{1}$$

但由于 KL 距离不满足对称性，实际上不能算作距离的定义，因此，可以将其对称化，由此导出的两个分布 p 和 q 的距离为

$$D_{sym}(p,q)=\frac{D_{KL}(p,q)+D_{KL}(q,p)}{2} \tag{2}$$

基于此，便可以根据分布的相似度对测试图像中的数字进行识别。

2.2 方案二：基于图像块的强度和重心

实验中方案一的效果并不是很好，所以我又结合作业要求的提示，用图像块的强度和重心作为特征，对该方案进行了实现。

与方案一相反，该方案是对图像块逐行统计其相关特征。特征提取方式如下。

1) 强度

将模板或扫描窗的每行分为左右两部分。对于每部分，统计其黑色像素点的个数，作为该图像块的强度特征。将这些强度值放在一起，即可得到数字模板或扫描窗对应的强度向量

$$s=\begin{pmatrix}s_{11} & s_{12} & s_{21} & s_{22} & \cdots & s_{h1} & s_{h2}\end{pmatrix}^{T} \tag{3}$$

2) 重心

将模板或扫描窗的每行分为左右两部分。对于每部分，寻找其所有黑色像素的坐标，这些坐标的平均值即可认为是其黑色像素的中心，如果该部分没有黑色像素，则认为其中心在图像边缘外侧。将这些重心值放在一起，即可得到数字模板或扫描窗对应的重心向量

$$c=\begin{pmatrix}c_{11} & c_{12} & c_{21} & c_{22} & \cdots & c_{h1} & c_{h2}\end{pmatrix}^T \tag{4}$$

获取这些特征向量之后，对于每个扫描窗，我们将其强度向量 s 与重心向量 c 分别与每个数字模板的强度向量 $s_{pattern}$ 和重心向量 $c_{pattern}$ 对比，求其距离。为了协调不同特征之间的影响，我们可以将两个距离进行组合，得到最终的距离度量，即

$$d=\alpha\left\|s-s_{pattern}\right\|+(1-\alpha)\left\|c-c_{pattern}\right\|,\ \alpha\in(0,1) \tag{5}$$

公式(5)中，特征向量均已长度归一化。获得距离度量后，便可结合距离阈值，对测试图像进行识别。

3 实验结果与分析

3.1 方案一实验结果

经过对参数阈值的调整、测试，识别结果如表 1 所示，识别结果中的“□”表示该位置未识别出数字。

测试图片	识别结果	识别错误/未识别个数
1.bmp	60130163 161241	5
2.bmp	2612□122 161861	7
3.bmp	20130123 161261	4
4.bmp	28140122 161261	6
5.bmp	□□130162 161641	5
6.bmp	20140168 161641	4
划痕.bmp	□□□6□□□9 □□□□□□	13 （未对划痕进行处理）
噪声.bmp	601□□1□3 141□□1	8
补充 1.bmp	2812□163 161261	8
补充 2.bmp	20140163 161641	4

表 1 方案一识别结果

3.2 方案二实验结果

经过对参数阈值的调整、测试，识别结果如表 2 所示，识别结果中的“□”表示该位置未识别出数字。

测试图片	识别结果	识别错误/未识别个数
1.bmp	281381□9 181641	3
2.bmp	201381□□	3

	181641	
3.bmp	2813812□ 181641	3
4.bmp	28138129 181641	2
5.bmp	88138129 181641	3
6.bmp	28138129 181241	3
划痕.bmp	20□□01□9 18□64□	5 （未对划痕进行处理）
噪声.bmp	2813812□ 181□11	5

表 2 方案二的识别结果

划痕、噪声、大小不一致的图片处理方法均与上次实验相同。

3.3 结果分析

从表 1 和表 2 的结果来看，方案一的识别效果明显不如方案二。我对这些结果的分析如下。

3.3.1 对方案一的分析

在方案一中，由于统计的特征的整个数字模板的黑色像素分布，因此没有进行分块处理。没有分块是否会对识别结果有严重影响，我暂时无法确定。

通过分析结果，我发现一个现象：4 被误判为 6 的情况非常多。实际上，4 和 6 都表现出左边黑色像素少而右边黑色像素多的趋势，从图 1 也可以看出，4 和 6 的分布特征是比较接近的。由于我们的样本数量很少（每个数字只有一个），很难根据大量样本的平均得到稳定的模板分布，我想这可能是结果不太理想的一个原因。

但方案一也有着其优点，如对图像的大小特征不敏感，只要能获取到一个可能是数字的分布，我们便可以根据分布差异的度量进行判定，因为所有的分布被我们统一到相同的长度。

如果我们能够获取较多的训练样本，进一步精确数字模板的分布，同时再结合不同方向的特征，方案一识别的准确率应当有比较明显的提升。如，我们再加入沿行方向的黑色像素分布，便可将 4 与 6 明显地区分开来，因为 4 是上面像素多，而 6 是下面像素多。

3.3.2 对方案二的分析

较方案一，方案二对图像进行了分块，并且结合了强度和重心两个特征，因此识别效果相对好一些。由于添加了不同的特征，识别过程中需要的参数设定也比较多，从实验过程中也能感受到该方案表现出的参数敏感性。如果再根据测试图像的特征对参数进行调整测试，应该能得到更好的识别效果。时间原因，我没有就参数对识别效果的影响进行进一步探究。

最后，综合对比方案一、二，我有如下想法：

除去其他因素不谈，**在训练样本较少的情况下，基于统计特性的特征应该要优于基于分布特性的特征**。就像一个随机变量一样，获取其概率分布总是不容易的，而其一阶矩、二阶矩等统计特性是比较容易估计的。当训练样本较少时，我们直接基于分布特征去匹配，这对训练集和测试集样本的一致性要求过高了，因而很容易受到一些随机噪声的影响（可以对比划痕图片的识别结果，在不对划痕做任何处理时，方案一的效果远不如方案二）。强度、重心这些特征是基于统计特性的，这些特征本身带有平均的意义，因此在一定程度上能增加对可能的模式的识别率。