手写体数字识别系统中一种新的特征提取方案*)

宋曰聪[1,2] 胡 伟[1,2]

(绵阳师范学院计算机科学与工程系[1] 绵阳师范学院程控交换与通讯网重点实验室[2] 绵阳 621000)

摘 要 本文对手写体数字识别系统中的数字特征提取方法进行了深入的研究,分析了当前用得比较多的三种特征提取方法,在此基础上提出了一种新的特征提取方案。该方案从每个字符中提取关键的13个点作为特征点,其主要特点是特征提取简单有效,节省了特征提取时间,提高了识别系统的运行速度。进行仿真时,用同一种网络对特征提取后的结果进行训练和识别,仿真结果表明,13点特征提取法用于手写体数字的识别有着极好的适应性,在运行速度和识别率上比起其它算法都有很大的提高,从而极大地证实了新算法的有效性及实用性。

关键词 手写体数字识别,特征提取,模式识别

A New Feature Extraction Method on Handwritten Digits Recognition System

SONG Yue-Cong[1,2] HU Wei[1,2]

(Department of Computer Science, Mianyang Normal University, Mianyang 621000)[1]

(Key Lab. of Program Control and Communication Network, Mianyang Normal University, Mianyang 621000)[2]

Abstract This paper discusses the handwritten digits feature extraction, analyzes current three kind of feature extraction method, points out the advantage and weakness, and in this foundation proposed one kind of new feature extraction method, namely 13 points feature extraction method, this method withdrew the key 13 spots as characteristic point from each digits character, it is simply and effective, and saves the feature extraction time, and enhances the recognition system running rate, the experiment indicats, 13 points feature extraction method to use in the Handwritten digits recognition have the extremely good compatibility, it has the very big enhancement in the running rate and recognition, tesify the new algorithms efficient and practicable.

Keywords Handwritten digits recognition, Feature extraction, Pattern recognition

1 概述

特征提取是手写体数字识别中的一个重要的环节,是模式识别的核心之一。经过预处理后,满足识别要求的模式要根据识别方法的要求抽取特征,作为识别的依据。模式识别的任务就是根据选择好的特征在 n 维空间对模式进行分类。很明显,这些特征的提取和选择对识别过程是至关重要的。如果模式选择得好,对不同类的模式就能表现出很大的差别,就能比较容易地设计出性能较高的分类器。因此特征选择会直接影响到分类器的构造和识别的效果。

国内外对于识别手写体数字的研究已经有很长时间,也发展了许多针对手写体数字的特征提取方法。本文在在对各种手写体数字进行识别时,总结和试用其他特征提取方法的基础之上,提出了一种新的特征提取方法,即13点特征提取法。经过多次尝试表明,13点特征提取法有着极好的适应性,对各种手写体数字的识别有着非常好的效果。

2 手写体数字特征提取的原则

直接把预处理后的数据作为神经网络的输入,数据量大;同时由于手写字体的多样化以及图像本身和预处理过程中附带的某些干扰的影响,为了提高识别能力,就要求神经网络有较强的容错能力,也就是要加大网络的规模。这样一来,不仅网络训练时间长,而且由于图像数据随机分布,训练出的网络也不够强壮。而特征提取的目的就是从分析数字的拓扑结构入手,把它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神经网络,这样就等于间接地增加了网络的容错能力,而且经过特征提取,数据量也大大减少,相应地,网络规模也减小了。可见,为了有效地进行数字识别,特征提取是必要的。

在模式识别中,特征的选择是一个关键问题。针对某一具体应用,所选择的特征往往直接影响到最终的识别率。由于在很多实际问题中常常不容易找到那些最重要的特征,或受条件限制不能对它们进行测量,这就使特征选择和提取的任务复杂化,因此特征的选择成为构造模式识别系统最困难的任务之一。

特征选择和提取的基本任务是如何从许多特征中找出那些最重要的特征。由于用很多特征进行分类器设计,无论从计算的复杂程度还是分类器性能来看都不是适宜的,因此研究如何把高维特征空间压缩到低维特征空间以便有效地设计分类器就成为一个重要课题。任何识别过程的第一步,无论用计算机还是由人去识别,都要首先分析各种特征的有效性并选出最有代表性的特征。尽管人们本身很精于字符的识别,但由于对大脑的识别机制了解得太少,还不很清楚当人们

*)四川省重点科技项目,项目号(02GG006-036)。宋曰聪 副教授,主要从事决策分析、群决策、优化算法及教育科研方面的研究工作。胡 伟 讲师,主要从事智能系统、网络安全及教育科研方面的研究工作。

识别一个字符时到底注重的是哪些特征，通常人们是凭经验来选择特征，由实验来验证。无论在传统模式识别还是在神经网络模式识别中，预处理的地位都是不容忽视的，而预处理的各个步骤中，特征提取显得尤为重要，所选特征合适与否直接决定着最终识别结果的好坏。

手写体字符识别的特征提取极大程度地影响着分类器的设计和性能，以及识别的效果和效率。为了保证所要求的分类识别的正确率和节省资源，希望依据最少的特征达到所要求的分类识别的正确率。因此，通常在得到实际对象的若干具体特征之后，再由这些原始特征产生出对分类识别最有效、数目最少的特征，这就是特征提取的任务。

特征提取实际上就是研究如何从众多的特征中求出那些对分类识别最有效的特征，从而实现特征空间维数的压缩。在进行手写体数字识别的过程中，特征提取应遵循的原则是：

①特征应能尽量包含字符的有用信息。

②特征的提取方法应简单而且提取快速。

③各个特征之间的相关性应尽可能小。

④特征数量尽可能少。

⑤特征应有较好的抗干扰能力。

3 手写体数字特征提取方法

经过图像的灰度化、二值化、平滑、分割、归一化等一系列的预处理操作之后，原来大小不同、分布不规律的各个字符变成了一个个大小相同、排列整齐的字符，接下来就要从被分割归一处理完毕的字符中，提取最能体现这个字符特点的特征向量。将提取出训练样本中的特征向量代入到网络之中就可以对网络进行训练了，然后提取出待识别的样本中的特征相量代入到训练好的网络中，就可以对字符进行识别了。特征向量的提取方法多种多样，根据具体情况的不同可以来选择不同的方法。对于手写体数字的特征提取方法而言，通常用得比较多的有逐像素特征提取方法，骨架特征提取方法，垂直方向数据统计特征提取法，梯度统计法，弧度统计法，角点提取等方法。

3.1 逐像素特征提取法

逐像素特征提取方法是一种最简单的特征提取方法，它是对图像进行逐行逐列的扫描，当遇到黑色像素时取其特征值为1，遇到白色像时取其特征值为0，这样当扫描结束以后就形成了一个维数与图像中像素点的个数相同的特征向量矩阵。

逐像素特征提取方法的特点是算法简单，运算速度快，可以使网络很快地收敛，训练效果好，但是这种算法的适应性不强。

3.2 骨架特征提取法

骨架特征提取法是一种利用细化的方法来提取骨架的方法。两幅图像由于它们的线条粗细不同，使得两幅图像差别很大，但是将它们的线条进行细化以后，统一到相同的宽度，如一个像素宽时，这时两幅图像的差距就不那么明显，利用图形的骨架作为特征来进行数字识别，就使得识别有了一定的适应性。

骨架特征提取的方法对于线条粗细不同的数字有一定的适应性，但是图像一旦出现偏移就难以识别。

3.3 垂直方向数据统计特征提取法

垂直方向数据统计特征提取法就是自左向右对图像进行逐列的扫描，统计每列黑色像素的个数，然后自上而下逐行扫描，统计每行的黑色像素的个数，将统计结果作为字符的特征向量。实验表明，这种方法的效果不是很理想，适应性不强。

4 一种新的手写体数字特征提取改进方案

上述的特征提取方法都存在有适应性不强的特点，当字符存在倾斜和偏移时都会对识别产生误差，另外，还有梯度统计、弧度统计、角点提取等方法，在实验中本人发现它们都存在着这样或那样的缺点，为此，本文提出了一种新的特征提取方法，即13点特征提取法。经过多次尝试表明，13点特征提取法有着极好的适应性，识别手写体数字的效果和效率良好。

4.1 13点特征提取方法的描述

13特征点提取方法的总体思路是：首先把字符平均分成8份，统计每一份黑色像素点的个数作为8个特征。分别统计这8个区域中的黑像素的数目，可以得到8个特征。然后统计水平方向中间两列和竖直方向中间两列的黑色像素点的个数作为4个特征，最后统计所有黑色像素点的个数作为第13个特征。也就是说，画4道线，统计线穿过的黑像素的数目。特征示意图如图1、图2、图3所示。

图1 八个区域

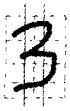

图2 竖直方向

图3 水平方向

4.2 13点特征提取方法的程序实现

从训练时间和识别率上加以对比，13特征点提取方法比其它几种方法效率都要高。13特征点提取方法有着极好的适应性，其完整的各序代码如下所示，其中函数名称为TZTQ，参数HDIBhDIB表示待提取特征的位图的句柄，num为字符的数目，dim为提取特征的维数。

```
double * * TZTQ(HDIB hDIB,int num,int dim)
{
  int i,j,k,m;
  //分配一个内存空间并得到二维指针
  double * * tezheng=alloc_2d_dbl(num,dim);
  //锁定图像句柄并获取其指针
  BYTE * lpDIB=(BYTE*);;GlobalLock ((HGLOBAL)hDIB);
  //取得图像像素数据区的起始地址
  BYTE * lpDIBBits=(BYTE*);;FindDIBBits((char*)lpDIB);
  BYTE *lpSrc;
  //获取图像高度
  LONG lHeight=;;DIBHeight ((char*)lpDIB);
  //获取图像宽度
  LONG lWidth=;;DIBWidth((char*)lpDIB);
  LONG width=lWidth/num;
  //每行的字节数
  LONG lLineBytes=WIDTHBYTES(lWidth *8);
  int b;
  //存储临时的特征
  double *tz=new double[dim];
  for (k=0;k<num;k++)
  {
    for (i=0;k<num;k++)
    //提取前8个特征
```

```
    for(m=0;m<8;m++)
    {
    for (i=int(m/2) * 8;i<(int(m/2)+1) * 8;i--+)
      for (j=m%2 * 8+k * width;j<(m%2+1) * 8+k * width;j++)
        {
          lpSrc=(unsigned char * )lpDIBBits + lLineBytes * i +j;
          b=( * lpSrc==255)? 0:1;
        }
    }
//提取第 9 个特征,总像素值
  for (i=0;i<lHeight;i++)
    for (j=k * width;j<(k+1) * width;j++)
    {
      lpSrc=(unsigned char * )lpDIBBits + lLineBytes * i+j;
      b=( * lpSrc==255)? 0:1;
      tz[8]+b;
    }
  //提取第 10、11 个特征,水平扫描切割
    i=int(lHeight * 1/3);
  for(j=k * width;j<(k+1) * width;j++)
  {
    lpSrc=(unsigned char * )lpDIBBits + lLineBytes * i +j;
    b=( * lpSrc==255)? 0:1;
    tz[9]+b;
    }
    i=int(lHeight * 2/3);
    for (j=k * width;j<(k+1) * width;j++)
    {
      lpSrc=unsigned char * ) lpDIBBits + lLineBytes * i +j;
      b=( * lpSrc==255)? 0:1;
      tz[10]+=b;
    }
    //提取第 12、13 个特征,垂直扫描切割
      j=int(k * width+width * 1/3);
      for (i=0;i<lHeight;i++)
      {
        lpSrc=(unsigned char * )lpDIBBits + lLineBytes * i +j;
        b=( * lpSrc==255)? 0:1;
        tz[11]+b;
      }
      j=int (k * width +width * 2/3);
      for(I=0;I<lHeight;I++)
      {
        lpSrc=(unsigned char * ) lpDIBBits +lLineBytes * I +j;
        b=( * lpSrc==255)? 0:1;
        tz[12]+=b;
      }
      //存储特征
      for (i=0;i<dim;i++)
        tezheng[k][i]=tz[i];
    }
    ::GlobalUnlock ((HGLOBAL)hDIB);
    //返回特征向量矩阵的指针
return tezheng;
}
```

5 仿真结果对比研究

5.1 识别系统的描述

对于整个手写体数字识别系统,程序的实现分为图像预处理和神经网络识别两大模块。在图像预处理的过程当中,采用了许多图像处理技术最后把数字的特征提取出来。这些技术包括图像数据的读取、图像的灰度化、二值化、平滑、字符的切分、归一化等图像处理技术,最后是数字字符特征的提取。其结果再利用神经网络进行字符识别,利用神经网络进行字符识别的过程主要包括网络的训练、数据的读取、字符的判定、结果的输出。系统的主界面如图 4 所示。

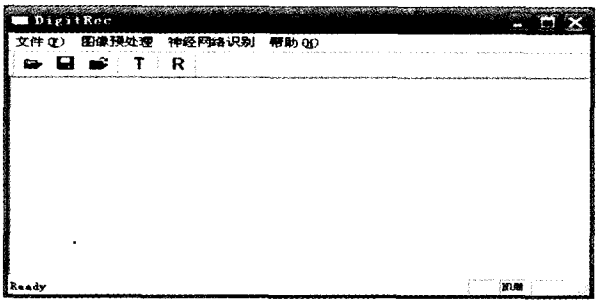

图 4 系统主界面

5.2 仿真结果分析

用本文所提出的 13 点特征提取方法对预处理后的数字进行特征提取后,再利用神经网络进行识别,图 5 为神经网络识别手写体数字的测试结果。

	识为 0	识为 1	识为 2	识为 3	识为 4	识为 5	识为 6	识为 7	识为 8	识为 9	正确	正确率
数字 0	190	0	0	0	0	0	0	0	4	6	190	95 %
数字 1	0	180	13	4	2	1	0	0	0	0	180	90 %
数字 2	0	1	181	12	2	3	1	0	0	0	181	90.5 %
数字 3	0	2	7	185	0	5	1	0	0	0	185	92.5 %
数字 4	0	1	3	2	179	2	1	2	0	10	179	89.5 %
数字 5	0	0	2	2	1	188	0	5	2	0	188	94 %
数字 6	0	0	0	3	6	8	182	0	1	0	182	91 %
数字 7	0	0	0	0	0	4	5	184	7	0	184	92 %
数字 8	2	0	0	0	1	2	4	4	180	7	180	90 %
数字 9	4	0	0	0	7	0	2	3	4	180	180	90 %
	196	184	206	208	198	213	196	198	198	203	1829	91.5 %

图 5 测试结果

对训练好后的神经网络分类器所进行的识别测试分为两种情况:

1. 当用训练集中的训练样本作为待识别样本时,其识别率可达 100%。

2. 当用所有训练样本和识别样本作为测试集时,神经网络分类器的识别率有所下降,其识别率可达 90%以上。

实验结果表明,本文所提出的 13 特征点提取法有着极好的适应性,对手写体数字的识别具有较好的识别效果,在运行速度和识别率上比起其它算法都有很大的提高,从而极大地证实了新算法的有效性及实用性。

结束语 本文系统地总结分析了手写体数字识别中的特征提取的问题。在分析当前用得最多的特征提取方法的基础之上提出了一种新的特征提取方案,即 13 点特征提取方法。该方法简单有效,用于对手写体数字进行识别时,极大地节省了特征提取的时间,提高了识别系统的运行速度,从而有效地提高了识别率。当然,这种特征提取方案只适合于单纯对数字进行的识别,比如邮政系统中信件的邮编识别,即只适合于事先知道要识别的对象是数字的识别,否则,这种特征提取方案是无法进行正常识别的。最后应当指出,手写体数字识别的研究不仅有很大的应用价值,而且有重要的理论价值,深入

进一步研究新的、效果更好的手写体数字的特征提取方法具有重大的意义。

参 考 文 献

1 Chien-cheng , Tang Yun-ching. To improve the training time of BP neural networks. Info-tech and Info-net, 2001 International Conferences on, 2001, 3: 473～479
2 Hornik K. Approximation capabilities of multilayer feedforward networks. IEEE Transactions on Neural Networks, 1991, 4: 251～257
3 Yu X H, Chen G A. Efficient estimation of dynamically optimal learning. In: Proc. of IEEE ICNN-95, 1995. 385～388
4 Sakaue S, Kohda T, Yamamoto H, Maruno S, Shimeki Y. Reduction of required precision bits for Back-Propagation applied to pattern recognition. IEEE Transactions on Neural Networks, 1993, 4 (2): 270～275
5 Kwon T M, Chen Hui. Contrast enhancement for backpropagation. IEEE Transactions on Neural Networks, 1996, 7(2): 515～524
6 Chen J Q, Jiang J P. New method to train a BP network and their application. International Joint Conference on Neural Networks, 1999, 3
7 朱学芳. 手写数字识别实验系统的研究[J]. 南京大学学报, 1996. 1
8 谢光毅, 钟义信. 神经网络用于手写体数字识别[J]. 模式识别与人工智能, 1994, 12(4)
9 刘滨. 基于神经网络的车牌字符识别研究[D]: [武汉大学硕士学位论文]. 2004. 8
10 胡小锋, 赵辉. Visual C++/MATLAB 图像处理与识别实用案例精选[M]. 北京: 人民邮电出版社, 2004. 9
11 朱小波. 基于神经网络的手写体数字识别分析与研究[D]: [武汉科技大学硕士学位论文]
12 张捷. 手写数字识别的研究与应用[D]: [西安建筑科技大学硕士学位论文]

(上接第 190 页)

个 pfd_j，将其对应的属性 a_j 剔除，同时将 pfd_j 作为下一步的 wfd，在属性集 $A-\{a_j\}$ 中继续上述的步骤，直到剩余属性数目与 S 的分形维数相同。

我们获得后向剔除属性约简算法 FDR 如下。

算法：FDR

输入：决策表 $S=(U,A,V,f)$，$A=C\cup D$，$C\cap D=\Phi$，C 为条件属性集合，D 为决策属性集合

输出：最小约简 Redu

步骤：

第 1 步：算出决策表 S 的计盒维数 wfd（全分形维）；$wfd_0 \leftarrow wfd$；

第 2 步：For $1\leqslant i\leqslant |A|$ ($a_i\in A$) 计算 U 在 $A-\{a_i\}$ 属性集上的部分分形维 pfd_i；

第 3 步：从 $pfd_1,\cdots,pfd_{|A|}$ 中选择最接近于 wfd 的一个 pfd，记为；计算 $A\leftarrow A-\{a_j\}$；$wfd\leftarrow pfd_j$；

第 4 步：如果 $|A|>wfd_0+1$，转第 2 步；

第 5 步：输出 A。

在 FDR 中，计算有 N 个元组的决策表的分形维的时间复杂度为 $O(N^2)$；从 $|A|$ 个属性中选择一个属性并剔除需要扫描 $|A|$ 次对象集 U，如果剔除 K（$K<|A|$）个属性，则需要扫描 $(K\times(2|A|-K+1))/2$ 次，所以整个算法的时间复杂度为 $O(|A|\times K\times N^2)$。

4 算法比较分析

为比较算法 BDMF 与算法 FDR 的效率，我们在 Intel 2.8GHz CPU×2（内存 1024M）上对具有 6 个属性（a_1,a_2,a_3,a_4,a_5,a_6）5500 条记录的合成数据集 FHMIN Test 6 进行了实验对比。其中，属性 a_1,a_2,a_3 对应的值由随机函数生成，属性 a_4,a_5,a_6 的值分别计算为：

$a_4=\sin(x1+x2+x3)+3$，$a_5=\sin(x1*x2+x3)$，$a_6=|x2+x3*\sin(x1+x3)|$

实验的结果如表 1。

表 1 BDMF 与 FDR 对比实验

Record	BDMF		FDR	
	Running Time (s)	Reduction	Running Time(s)	Reduction
2000	25	[x1,y1,y2,x2,]	4	[x1,x2,x3,]
3000	53	[x1,x2,x3,]	11	[x1,x2,x3,]
4000	96	[x1,x2,x3,]	16	[x1,x2,x3,]
5500	297	[x1,x2,x3,]	32	[x1,x2,x3,]

实验结果分析：由于 BDMF 算法是基于可辨识矩阵的，故对于 m 个属性、n 个元组的决策表来说，它的可辨识矩阵就是一个 $\max(m,n)\times\max(m,n)$ 的方阵，现实情况中一般属性的个数 m 都不太大，但是元组的数量 n 却是大得惊人，这时不但运算的时间很长，而且由于要占有大量的内存，使得 BDMF 算法的适用范围大大降低，实验中发现当元组数量在增加时，花费的时间急剧增加；而分形维的计算由于不需要生成可辨识矩阵，受元组数量的影响要小得多，都很快地得到了满意的结果。当 Record∈[3000, 5500]时，BDMF 得到的约简为最小约简，其运行时间从 53 秒增加到 297 秒，而 FDR 不但在数据集的记录数在各个阶段都能获得正确的结果，且花费的时间也是可以容忍的。

结论 一个数据集的分形维体现了数据集的固有特征。算出数据集的分形维数就可以确定该数据集最小约简的大小，基于这种思想，本文利用后向剔除属性的约简算法 FDR 来进行属性约简并用基于可辨识矩阵的属性约简算法 BDMF 来进行对比分析。算法复杂性分析的结果和实验的结果均说明 FDR 在时间和空间两方面都优于 BDMF。

参 考 文 献

1 Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data [J]. Kluwer Academic Publishers, 1991. 6～42
2 Miao Duoqian, Wang jue. An information-based algorithm for reduction of knowledge [J]. IEEE ICIPS'97. 1991. 1155～1158
3 Wong S, Ziarko W. On optimal decision rules in decision tables [J]. Bulletin of Polish Academy of Science. 1985, 33: 693～696
4 Miao Duoqian, Wang jue. Analysis on feature reduction strategies of rough set[J]. Jo-urnal of Computer Science and Technology, 1998, 13(2): 189～192
5 Scherf M, Brauer M. Improving RBF networks by the feature selection approach EUBAF-ES[J]. In: W. Gerstner. ed. Proc. 7th Intl Conf on Artificial Neural Networks. Lausanne, Switzerland: Springer, 1997. 391～396
6 Robert A, Stocker E. Classification and feature selection by a self-organizing neural networ-k[J]. In: Dorffner G, ed. Proc. of Int Conf on Artificial and Neural Networks, UK: Springer 1999. 651～660
7 Pernkopf F, O'Leary P. Feature selection for classification using genetic algorithms with a novel encoding[J]. In: W. Skarbek, ed. Proc. of Computer Analysis of Images and Patterms, Warschau, Poland, Springer, 2001. 161～168
8 Traina C, Traina A, Wu L, et al. Fast feature selection using fractal dimension[J]. In: C. Faloutsos, ed. Proc. of XV Brazilian Symposium on Databases, Paraila, Brazil: Springer, 2000. 78～90
9 Talavera L. Feature selection as a preprocessing setp for hierarchical clustering[J]. In: I. Bratko, ed. Proc. of the 16th Int Conf on Machine Learning. Bled, Slovenia: AAAI Press, 1999. 389～397
10 Grassberger P. Generalized, Dimensions of Strange Attractors [J]. Physics Letters, 1983, 97A: 227～230
11 Schroeder M. Fractals, chaos, power larws, 6ed [M]. New York: W. H. Freeman and Company, 1991
12 王国胤. Rough 集理论与知识获取[M]. 西安交通大学出版社, 2001. 32～67