基于注意力网络集成的联机空中手写识别研究

张墨逸,邢　蕾,叶洪昶,陈海燕
(兰州理工大学 计算机与通信学院,甘肃 兰州 730050)

摘　要:针对联机空中手写识别的数据样本少、模型泛化能力不足、识别率低等问题,提出一种基于注意力网络集成的联机空中手写识别方法。该方法首先通过在形状特征中融入“联机”的时序特征,构建原始的多维数据;然后对多维融合数据降维投影到三个正交平面上,得到三组投影特征;其次,构建卷积神经网络用于提取视觉特征,同时引入字符嵌入作为图像的类标签,将类标签字符级语义特征通过注意力检测机制与三组视觉特征融合形成三组语义信息丰富的特征图,并基于特征图构建 SoftMax 分类器;最后,通过基于主学习器集成投票方法进行分类与识别。在两组空中手写数据集与哈工大(HIT-OR3C)联机数据上进行多组实验,在小样本的情况下,该方法识别率优于其他方法,分别达到 95.68%,93.02%,94.96%。实验结果表明,该方法在小样本数据的情况下,充分发掘联机空中手写数据中有效特征,提高了空中手写识别效率。

关键词:空中手写;联机手写;小样本学习;数据融合;注意力网络;集成学习;手势识别
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2024)10-0126-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0190

An Attention Network Ensemble-based Study of Online in Air-writing Recognition

ZHANG Mo-yi, XING Lei, YE Hong-chang, CHEN Hai-yan
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Aiming at the problems of small data samples, insufficient model generalization ability and low recognition rate of online In Air-Writing recognition, an ensemble-based online In Air-Writing recognition method was proposed. Firstly, the original multi-dimensional data was constructed by incorporating “online” time series features into the shape features, and the multi-dimensional fusion data was projected to three orthogonal planes to obtain three sets of projection features. Secondly, a convolutional neural network was constructed to extract the visual features, next character embedding was introduced as class labels of the image, and the class-labelled character-level semantic features were fused with the three sets of visual features through the attention detection mechanism to form three sets of semantically informative feature maps, and a SoftMax classifier was constructed based on the feature maps. Finally, the classification and recognition was performed by the main learner-based integrated voting method. Multiple sets of experiments were carried out on two sets of In Air-Writing datasets and the HIT-OR3C online dataset, and in the case of small sample recognition, the recognition rates of the proposed method were better than that of other methods, which were 95.68%, 93.02% and 94.96% respectively. The experimental result showed that the proposed method fully explored the effective features in the In Air-Writing data under the condition of small sample data, and improved the efficiency of In Air-Writing recognition.

Key words: In Air-Writing; On-Line Writing; small sample learning; data fusion; attention network; ensemble learning; gesture recognition

0　引　言

手势空中书写(简称空中手写)指用户通过手势在虚拟的输入区内书写文本,使计算机具有能够像人一样的认字能力,是真正自然的人机交互形式。空中书写技术是大数据与交互计算技术的主要内容,推动大数据驱动的人机混合智能与机器学习平台建设,从根本上提升智能交互装备的核心竞争力。

联机空中手写指书写轨迹通过定时采样即时输入

收稿日期:2024-01-04　　**修回日期**:2024-05-07
基金项目:国家自然科学基金项目(62161019)
作者简介:张墨逸(1985-),女,副教授,博士,研究方向为机器视觉、机器学习、模式识别;邢　蕾(1999-),男,硕士研究生,研究方向为模式识别与人工智能。

到计算机中。联机空中手写不同于传统书写,没有起笔和落笔的信息,字符都是一笔写成,缺少限定书写起始的分隔序列等问题,使得字符识别任务更具挑战性[1]。

传统空中手写识别的方法提取不同的特征,选取合适的分类器进行分类识别,包括动态时间规整(DTW)[2]、支持向量机(SVM)[3]、隐马尔可夫模型(HMM)[4]等。这些方法容易被人理解,也可以实现一些简单的空中手写识别,但是通用性不强,需要大量的经验和多次尝试才能选择出较好的特征,当手势字符扩展后,模型较复杂时,识别精度往往会下降。

随着深度学习技术的兴起,空中手写识别技术有非常大的进展,不同的网络结构及其集成[5]都被应用到空中手写中,如孪生网络[6]、RNN(循环神经网络)[7-8]、LSTM[9](长短时记忆网络)、Bi-LSTM[10]、CNN + LSTM[11]、RNN + LSTM[12]、TCN、1DCNN[13]、CNN+RNN[14],并取得了较好的结果。

后期,研究者针对手写特点对神经网络进行改进,常见的有加入注意力机制更好地提取特征。Ren等[15]在传统的 RNN 网络中引入注意力机制和权值方差约束对其进行改进,显著地提升了模型的识别准确率。党小超等人[16]使用基于注意力机制的双向循环神经网络模型进行训练,对空中手写数字进行识别。陈路等人[17]引入注意力机制,采用稠密卷积网络作为编码器,门控循环单元(GRU)作为解码器对手写数学公式进行识别。

使用编码器进行特征提取,是另一个改进策略。付鹏斌等人[18]将联机模式和脱机模式联合,设计了一种基于编码器-解码器框架的双模式识别模型。Wang等人[19]采用类标签的语义嵌入来生成注意力图进行特征编码,以实现更好的图像表示。

虽然上述文献都是基于图像级特征与时序的研究,但经过实验发现在小样本下进行联机空中手写识别时,各方法识别准确率较低。为了提高联机空中手写识别准确率,该文结合上述注意力机制与编码特征的优点,创新性地引进类标签作为文本特征进行编码,提出一种基于注意力网络集成的联机空中手写识别方法,最后通过实验说明其性能的优越性。

1　空中手写识别流程

空中手写识别中,整个识别流程如图 1 所示。采集到数据后,首先通过数据预处理,用所得到指尖轨迹作为获得的书写内容,然后对获得的指尖轨迹进行特征提取,最后送入分类器识别,预测出空中手写的识别结果。该文研究的范围在获得书写轨迹后,研究特征提取与分类器的设计,最终识别出书写字符。

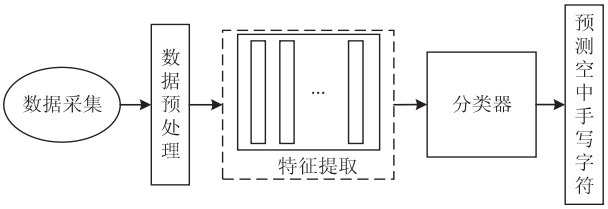

图 1　空中手写流程

如图 2 所示,该文主要从多模态数据融合、多模特征变换、注意力网络、基于主学习器集成的投票算法四个方面来研究空中手写识别。(1)对获取的联机空中手写数据进行预处理,将得到的平面位置 x,y 坐标与时序数据的 z 坐标作为融合后的 3D 数据;(2)将融合后的 3D 数据,使用特征变换的方法分别投影到 3 个正交的坐标平面上,获得 3 组不同的特征;(3)用分布式字符嵌入表示类标签,通过字符嵌入模型得到字符的语义特征;(4)将 3 组图像特征通过注意力检测器与字符特征结合,新特征在不同的学习器上学习训练得到 3 个基分类器。最终将每个基分类器的输出结果输入到基于主学习器集成投票算法中进行分类。

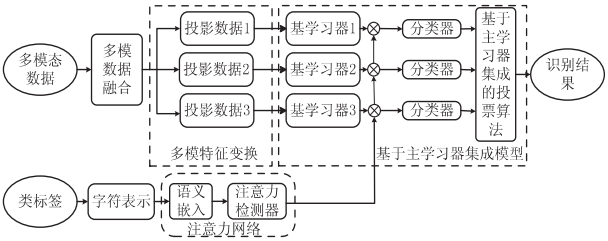

图 2　方法流程

2　数据预处理

2.1　数据采集

数据采集,首先计算动态手势序列间的差异度,通过极大值选取算法得到关键帧序列;然后使用手部骨架提取的方法对关键帧序列图像进行手指指尖轨迹追踪得到目标区域(如图 3 所示);计算目标区域的中心点作为当前的书写位置。最终采集到的数据如公式 1,其中 x_n,y_n 表示中心点位置坐标,n 表示关键帧序列。

$$X = \{(x_0,y_0),(x_1,y_1),\cdots,(x_n,y_n)\} \tag{1}$$

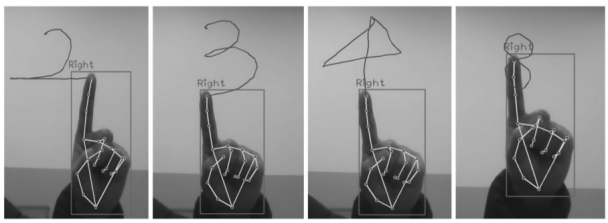

图 3　基于手部骨架方法追踪目标域手指之间轨迹序列

空中书写数据和在线手写数据都是时间序列,但它们之间存在差异。在空中书写中,字符具有一笔的特点,因此笔画中存在多余的部分(图 4 b.1 和 b.2),

而在线手写(图 4 a.1)可以用多个字符笔画完成。例如,在数字“4”和“5”写字的情况下,它们分别有两笔画,因此,在空中书写时有一个冗余部分。此外,对于相同字符,书写顺序的不同,会导致空中书写的差异。如图 4 所示,b.1 和 b.2 中的数字“5”(圆点是起点,箭头标记书写方向)。

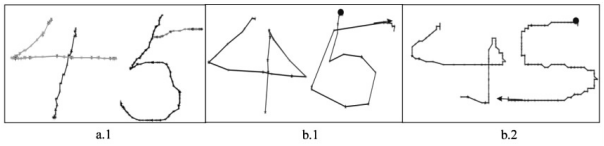

图 4 在线手写字符(a.1)和空中书写手写字符(b.1 和 b.2)

2.2 归一化

该文采用 Z-score 标准化数据预处理的方法。

基于原始数据的均值和标准差进行数据的标准化,经过处理的数据符合标准正态分布,即均值为 0,标准差为 1。其转化公式为:

$$X_n = \frac{X - \mu}{\delta} \tag{2}$$

其中,μ 为原始数据的均值,δ 为原始数据的标准差,是当前用的最多的标准化公式。

3 多模态数据融合

得到的图像数据与时序数据类型不同,为获取多模态数据,使用多模态数据融合方法将其融合为 3D 数据。方法如下:原始图像数据 $\{(x_0,y_0),(x_1,y_1),\cdots,(x_n,y_n)\}$ 是一个 2D 的时序数据,将 x,y 坐标作为融合后 3D 数据的 x,y 坐标,将时序数据作为融合后 3D 数据的 z 坐标,这样得到了融合后的 3D 数据。

$$\{[x_1,y_1,z_1],[x_2,y_2,z_2],\cdots,[x_n,y_n,z_n]\} \tag{3}$$

其中,n 是序列的长度,x_i 和 y_i 是原始图像数的 x,y 坐标,z_i 表示样本的关键帧序列。

对空中手写数字“9”进行多模态数据融合,得到融合后的 3D 数据见图 5。

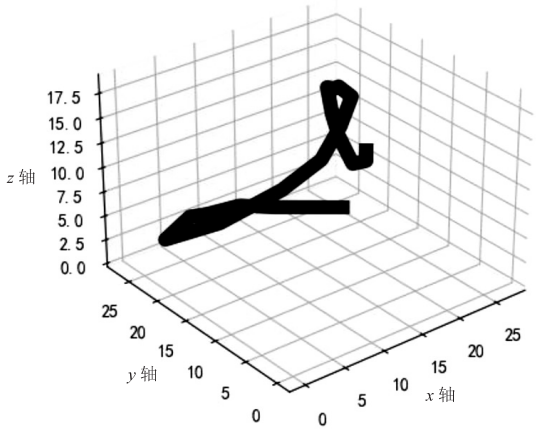

图 5 空中手写数字“9”融合后的 3D 数据

4 多模态特征变换

对于融合后的 3D 数据,使用特征投影方法,将其分别投影到 3 个正交的坐标平面上,获得 3 组不同的特征。其方法如下:融合后 3D 数据的 3 个坐标轴分别为 x,y,z 对 3D 数据降维分别取在 $x-y$ 平面,$x-z$ 平面,$y-z$ 平面的投影,得到 3 个相互正交的 2D 数据。通过这种方式进行数据降维变换,充分提取数据特征,从而提高最终的识别效率。其表达公式如下:

(1) $x-z$ 面投影变换。

其投影变换矩阵为:

$$\boldsymbol{T}_V = \boldsymbol{T}_{xoz} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

$\boldsymbol{T}_V$ 为 $x-z$ 面的投影变换矩阵。于是,由三维特征到 $x-z$ 面的投影变换矩阵表示为:

$$[x',y',z'] = [x,y,z] \cdot \boldsymbol{T}_V = [x,0,z] \tag{5}$$

(2) $x-y$ 面投影变换。

其投影变换矩阵为:

$$\boldsymbol{T}_H = \boldsymbol{T}_{xoy} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6}$$

$$[x',y',z'] = [x,y,z] \cdot \boldsymbol{T}_H = [x,y,0] \tag{7}$$

(3) $y-z$ 面投影变换。

其投影变换矩阵为:

$$\boldsymbol{T}_W = \boldsymbol{T}_{yoz} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

$$[x',y',z'] = [x,y,z] \cdot \boldsymbol{T}_w = [0,y,z] \tag{9}$$

5 注意力网络

该文利用分布式字符嵌入来表示类标签,以捕捉不同概念之间的语义关系。例如,将中文数字“九”作为图像“9”的类标签。使用字符级别的 Word2Vec 模型[20]处理类标签。这种方法将每个字符看作是一个“伪单词”,并将其视为训练样本中的一个单位。将字符序列作为输入,通过训练 Word2Vec 模型得到每个字符的向量表示。然后构建注意力图生成器,使其成为从字符嵌入和图像特征到最终构成注意力图的映射函数。

首先通过卷积神经网络提取图像视觉特征,在全连接层对输入数据进行线性变换,并通过激活函数(ReLU)对结果进行非线性映射,得到图像特征。公式如下:

$$v_i = f(W_v x_i + b_v) \tag{10}$$

其中,x_i 表示图像特征,W_v 与 b_v 表示模型参数。

然后从类标签的语义嵌入中创建一个注意力检测

器 $h = W_s c + b_s$,其中 c 表示类标签的语义嵌入, W_s 与 b_s 表示模型参数。其次将生成的注意力检测器应用于每个全局图像特征向量,以获得其初始注意力置信度得分 $a_i^{'} = h^{\mathrm{T}} v_i$,对类标签相关局部区域进行加权处理,以突出与类别标签相关的特征,并抑制与类别标签无关的特征。这样可以提高模型的判别力和准确性。其公式如下:

$$a_i = \frac{b(a_i^{'})}{\sum_i b(a_i^{'})} \tag{11}$$

其中, $b(\cdot)$ 表示通过 ReLU 函数使得注意力分数为正。

最后构建注意力图像。

$$g = \sum_{i=1}^{|D|} v_i a_i \tag{12}$$

其中, D 表示全局特征的集合。

6　基于主学习器集成的投票算法

基于主学习器集成的投票算法流程如图 6 所示。该算法是将基学习器 h_i 分别在三个正交的注意力图像上学习,其中 $i = 1,2,3$。按所识别样本的语义信息,将 $x - y$ 面投影数据融合的注意力图像作为主学习器 h_1 的识别数据,其他两个投影面融合的数据作为辅学习器 h_2 , h_3 的识别数据。将两个辅学习器的预测结果通过相对多数投票算法集成(公式 13),得到分类结果 $H(x)$ 。如果 $H(x)$ 出现的概率 $p(H(x))$ 比主学习器 $p(S(x))$ 大,则将 $H(x)$ 作为最终输出类别,否则将主学习器得到的结果作为最终的输出结果,其中 $S(x)$ 是主学习器分类结果。

相对多数投票算法,其公式如下:

$$H(x) = c_{\underset{j}{\operatorname{argmax}} \sum_{i=1}^{T} h_i^j(x)} \tag{13}$$

上述公式中,基学习器 h_i 在样本 x 上的预测输出为向量 $(h_i^1(x), h_i^2(x), \cdots, h_i^j(x))$,其中 $h_i^j(x)$ 是 h_i 在类别标记 c_j 上的输出。$H(x)$ 表示由投票法产生的分类结果,即预测为投票最高的标记。

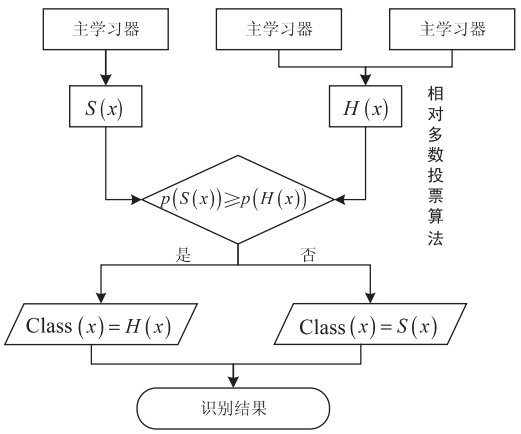

图 6　基于主学习器集成的投票算法流程

基于主学习器投票算法的具体步骤如算法 1 所示。步骤 4 中, $h_i^j(x)$ 是 h_i 对类别标记 c_j 的输出。第 5 步,如果是 $p(S(x)) \geqslant p(H(x))$,则选择 $S(x)$ 作为最终输出,否则选择识别结果 $H(x)$ 作为最终输出。

算法 1:基于主学习器集成的投票算法

输入:训练集 D_i ,学习器 $\mathcal{L}$.

流程:

1:for $i = 1,2,3$ do

2:　$h_i = \mathcal{L}(D_i)$

3:end for

4: $S(x) = h_1^j(x_1)$; $H(x) = c_{\underset{j}{\operatorname{argmax}} \sum_{i=1}^{3} h_i^j(x)}$;

5:if $p(S(x)) \geqslant p(H(x))$

6:　$\mathrm{Class}(x) = S(x)$

7:else

8:　$\mathrm{Class}(x) = H(x)$

9:end if

输出: $\mathrm{Class}(x)$

7　实验结果与分析

7.1　实验平台

硬件环境:Intel(R) Core(TM) i5-4210H CPU @ 2.90 GHz;内存:8.00 GB;系统类型:64 位操作系统,基于 x64 的处理器;软件环境:Windows10 操作系统下,每个网络使用 Tensorflow5.0 环境在 Python3.8.0 上训练 1 000 个 epoch,BatchSize 设置为 64,优化器采用 Adam,学习率为 10^{-3}。使用 LSTM 等算法训练时,每个字符在录制过程中的序列长度存在差异,为了创建相同的样本序列长度,使用了零填充。

7.2　数据集

数据集 1:实验数据集的动态手势视频库通过基于 OpenCV 提取手部骨架关键点,追踪手指指尖书写数字(0-9)轨迹信息。邀请 3 位实验者分别做这 10 种手势,每个手势重复 10 次,得到 300 个手势样本(3×10×10)。每个手势持续时间不同,约为 3 s ~ 12 s(帧率 δ =25 fps)。

数据集 2:采用文献[1]中所制作数据集。基于 LMC JavaScript 版本 2 API 设计了一个简单的交互式 web 界面,用于收集空中手写的数字(0-9)。参与者分别写了数字至少 10 次,结果有将近 1 200 个样本,采样频率为 60 Hz。

数据集 3:哈尔滨工业大学开设汉字识别语料库(HIT-OR3C)。选取联机的数字(0-9),1 220 个样本。

图 7 是对数据集 2 分别以 10% ,20% ,30% 的比例划分训练集,测试算法的识别准确率。由图可以看出,随着训练集比例的增加,识别准确率上升,且文中方法在不同比例的训练集测试中准确率均高于 CNN 模型。

为了验证文中方法在小样本情况下的优势,所有数据集的训练集按 10% 比例分割。

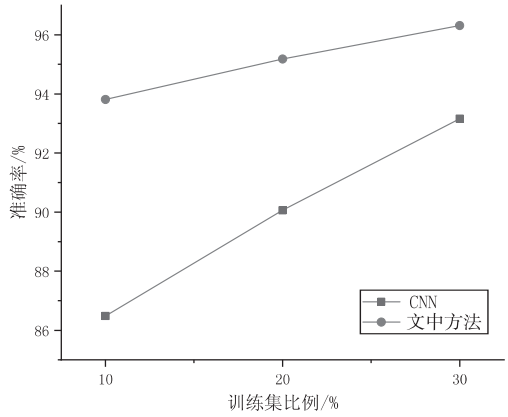

图 7 按不同比例选取训练集准确率对比

数据集 1 的划分:选取 30 个样本做训练集,30 个样本做验证集,240 个样本做测试集,数据随机分割。

数据集 2 的划分:选取 948 个样本,其中 95 个样本做训练集,758 个样本做测试集,95 个样本做验证集,数据随机分割。

数据集 3 的划分:选取 122 个样本做训练集,976 个样本做测试集,122 个样本做验证集,数据随机分割。

7.3 各学习器参数设计

在文中方法中,视觉特征提取采用三个 8 层 CNN 模型作为基学习器,8 层网络从前往后分别为输入层 32×32,16 个 5×5 的卷积层,最大池化层,36 个 5×5 的卷积层,最大池化层,平坦层,128 个神经元的隐藏层,输出层。对类标签字符嵌入表示,使用 Word2Vec 模型[20]来提取语义嵌入。其中视觉特征输出与字符语义嵌入输出设置为 300 维度。最后将融合后的注意力特征图通过 SoftMax 分类器进行分类。训练过程使用交叉熵损失函数进行优化。图 8 表示基于注意力网络集成算法模型。

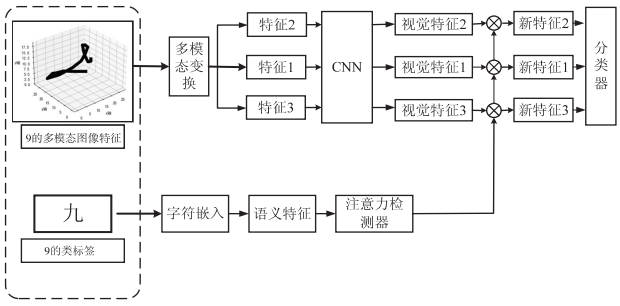

图 8 基于注意力网络集成算法模型

LSTM 使用了双层 LSTM[9] 和双向 LSTM[10],根据使用的数据集序列长短差异,每个块中隐藏单元分别为 30 个、130 个、300 个。采用 SoftMax 回归的线性全连接层将每个输入分类。

1DCNN[21-22] 为 9 层网络,4 个卷积层 100 个滤波器,内核为 10,步长为 1,2 个最大池化层,平坦层,128 个神经元的隐藏层,输出层。

1DCNN-LSTM[23] 是在一维卷积的输入之后添加了一个执行 LSTM 的额外层,120 个隐藏单元。

TCN-Dynamic[1] 通过将提取的特征映射到与输入长度相同的预测序列来处理数据序列。一共有三层,每层的膨胀因子增加 2l。使用大小为 7 的内核,每层 32 个隐藏单元和扩展列表为 1、2、4。

TCN-Static[1] 表示将图像以数组的方式展开,即每一行都附加到前一行。TCN-Static 的实际输入是 0 和 1 的序列(即图像像素的值)。

CNN-CBAM[24] 模型由 2 个卷积层、2 个池化层、3 个 BN 层、1 个 Droupout 层以及 1 个 Softmax 层组成。在第二次卷积池化层后嵌入 CBAM 注意力模块。其中,卷积核大小统一设置为 5×5,步长设置为 1,池化层的窗口设为 2×2,步长为 2。

Model[25] 模型包括 7 个堆叠的卷积层,卷积核大小固定为 3×3,在第一层有 16 个过滤器,逐渐增加到最后一层的 256 个过滤器。在卷积操作后,使用最大池化来减小特征图的大小。最后,将特征图展平并送入一个全连接层通过 Softmax 以获取最终的表示。

7.4 评价指标

为了验证文中方法的可靠性,采用四种评价指标对文中模型进行评估,分别是准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 分数(F1 Score)。准确率表示正确分类的样本占样本总数的比例,计算公式为式 14,其中 TP 表示将正样本归类为正样本,FP 表示将负样本归类为正样本,TN 表示将负样本归类为负样本,FN 表示将正样本归类为负样本。

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{14}$$

精确率计算公式如式 15 所示,召回率计算公式如式 16 所示。

$$\mathrm{Prc} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{15}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

召回率和精确度之间往往需要权衡。当模型能找到更多的正样本时,它往往会将更多的负样本归类为正样本,也就是说,当召回率高时,精度往往会降低。F1 指标旨在平衡这两个指标,其计算公式如式 17 所示。

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Prc} \cdot \mathrm{Rec}}{\mathrm{Prc} + \mathrm{Rec}} \tag{17}$$

7.5 实验结果

该文设计了消融实验,得到最优迭代周期和归一化方法。迭代周期分别选择 500、1 000 次进行迭代训

练,采用三种归一化方法在 CNN 模型中进行实验对比,对比结果如表 1 所示。

表 1　迭代次数与归一化对比结果

归一化方法	准确率/%	
	500 次迭代	1 000 次迭代
CNN-M	83.53	85.51
CNN-Z	85.38	86.96
CNN-C	84.92	85.92

表 1 中 CNN-M 表示 CNN 模型与线性归一化[1]相结合,CNN-Z 表示 CNN 模型与 Z-score 标准化相结合,CNN-C 表示 CNN 与坐标归一化[22]相结合,分别在数据集 2 中进行训练预测。由表 1 可以看出,在保证模型有效收敛的情况下,迭代 1 000 次,使用 Z-score 标准化,模型的准确率最高,相比 500 次迭代,准确率提升 1.58 百分点。

为了进一步验证字符嵌入注意力模块的有效性,在数据集 1 上进行了对比实验,实验将网络中的字符嵌入注意力模块后进行对比。由表 2 可知,添加字符嵌入注意力模块比未增加该模块的识别率高 1.22 百分点左右,识别效果明显提升,证明了添加字符嵌入注意力模块在提高识别准确率方面更加有效。

表 2　消融实验结果对比

数据	方法	准确率/%
数据集 1	无字符嵌入	94.46
	有字符嵌入	95.68

将训练得到的每个最佳模型进行保存,然后在测试集中对它们进行评估。选取数据集 2 中两个表现最好的方法的混淆矩阵图进行对比,如图 9(2DCNN 模型)和图 10(文中模型)所示。从图 9 与图 10 中可以看出,数字“0”和“2”、“1”和“2”、“4”和“1”、“4”和“7”、“8”和“3”、“8”和“9”书写轨迹具有一定的相似性,存在一定程度误判。同时从图中可以看出,文中方法与 CNN 模型相比各个类别准确率都有提升。

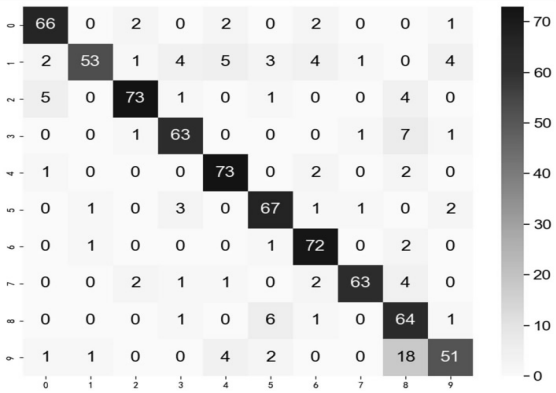

图 9　2DCNN 的混淆矩阵

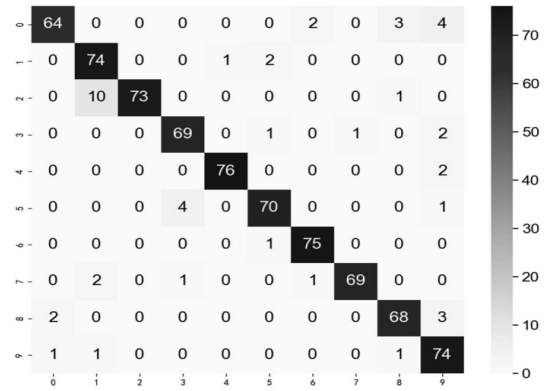

图 10　文中方法混淆矩阵

表 3 表示文中方法的平均识别时间,由表可以看出文中方法能够满足所需的实时性。

表 3　文中方法平均识别时间

数据集	平均识别时间/s
数据集 1	0.002 07
数据集 2	0.001 97
数据集 3	0.001 53

表 4 ~ 6 是常用各模型及文中方法在数据集 1 ~ 3 上分别对准确率(Acc)、精确率(Prc)、召回率(Rec)、

表 4　常用方法在数据集 1 上的各相关指标对比

模型	Acc/%	Pre/%	Rec/%	F1/%	Time/s
LSTM[9]	73.75	75.44	73.78	73.37	0.281
BLSTM[10]	82.92	85.58	83.14	82.79	0.079
TCN-Dynamic[1]	74.17	77.45	74.55	74.01	0.086
RNN[7]	70.83	75.60	71.29	70.38	0.046
1DCNN[22]	70.00	64.39	70.11	65.24	0.167
1DCNN-LSTM[23]	70.42	65.78	70.49	65.71	0.191
TCN-Static[1]	61.67	61.01	61.67	60.82	0.133
2DCNN	83.92	87.07	83.98	83.32	0.091
CNN-CBAM[24]	93.33	93.89	93.33	92.98	0.094
Model[25]	87.92	88.60	87.92	87.95	0.097
文中方法	95.68	96.39	95.68	96.03	0.095

表 5 常用方法在数据集 2 上的各相关指标对比

模型	Acc/%	Pre/%	Rec/%	F1/%	Time/s
LSTM[9]	46.11	51.67	46.18	46.35	0.053
BLSTM[10]	50.59	54.20	50.72	50.83	0.041
TCN-Dynamic[1]	45.19	51.56	45.20	45.78	0.039
RNN[7]	42.42	44.38	42.43	42.03	0.028
1DCNN[22]	52.44	55.54	52.50	52.85	0.062
1DCNN-LSTM[23]	53.23	55.46	53.30	53.80	0.076
TCN-Static[1]	57.92	56.51	57.92	56.28	0.076
2DCNN	86.96	88.24	87.06	87.07	0.049
CNN-CBAM[24]	90.12	91.00	90.17	89.92	0.052
Model[25]	90.13	90.31	90.14	90.13	0.056
文中方法	93.02	93.36	93.05	93.06	0.054

表 6 常用方法在数据集 3 上的各相关指标对比

模型	Acc/%	Pre/%	Rec/%	F1/%	Time/s
LSTM[9]	57.58	59.86	57.65	57.41	0.258
BLSTM[10]	66.50	67.74	66.45	66.50	0.143
TCN-Dynamic[1]	63.01	63.87	63.16	62.86	0.130
RNN[7]	60.55	61.31	60.67	60.65	0.061
1DCNN[22]	71.87	72.95	71.22	71.54	0.401
1DCNN-LSTM[23]	73.26	75.05	73.14	73.10	0.491
TCN-Static[1]	71.72	72.21	71.72	71.59	0.512
2DCNN	86.58	87.79	86.58	86.57	0.330
CNN-CBAM[24]	90.47	91.03	90.47	90.52	0.334
Model[25]	93.55	93.81	93.54	93.58	0.336
文中方法	94.96	95.24	95.07	95.15	0.335

F1 值及运行时间对比。

以准确率作为主要评估指标。由表 4 ~6 可知,在小样本的情况下,采用文中方法在准确率等评估指标上相比其他几种方法最优。文中方法的识别率最高,CNN 模型紧随其后,性能略低,LSTM 等模型在小样本中识别率低,特别是当时间序列长度超过 300 时(数据集 2),其各项性能指标均无法达到所需标准。在实时性上,与其他模型相比,RNN 模型的运行时间最短,这是因为其具有相对较少的训练参数。

在表 4 中,Model[25] 模型相比于传统 CNN 模型提升 4 百分点,但比模型 CNN-CBAM[24] 识别率低,在表 5、表 6 中,随着训练数据的增加逐渐上升,其识别率超过模型 CNN-CBAM,可见,该模型在小样本下,对空中手写字符识别率并不高。表 5 中可以看出,文中方法将平均识别率提升到 93.02%,比 CNN 提升了 6.06 百分点,相较于基于注意力的 CNN-CBAM 模型提升 2.9 百分点。由实验可知,在空中手写领域,文中方法简单易行,数据利用充分,在小样本上学习,可以显著地提高识别效率。

8 结束语

该文提出了一种基于注意力集成的联机空中手写识别方法。由于联机数据拥有不同模态的特征,为了能够充分利用数据的不同特性,将联机时空数据特征融合,然后对多模数据进行投影变换,将融合的多模数据投影在不同的平面,产生多组特征数据集。同时,引入字符嵌入作为类标签,将字符嵌入的语义信息与视觉图像信息通过注意力检测器相融合,最终在每个子数据集上学习得到基分类器并集成,集成时使用基于主学习器的投票算法。实验结果表明,文中方法是有效的、可靠的,能够提高小样本下联机空中手写的识别率,并且能够满足空中手写识别的实时性要求。该文进一步的工作拟结合 LDA 方法与子空间投影降维方法进行特征提取并集成,使得特征空间描述的类内方差小,类间方差大,进一步提升小样本大类别下的空中手写识别率。

参考文献:

[1] BASTAS G, KRITSIS K, KATSOUROS V. Air-writing recognition using deep convolutional and recurrent neural network architectures[C]//2020 17th international conference on frontiers in handwriting recognition. Dortmund: IEEE, 2020:7-12.

[2] TANG J, CHENG H, ZHAO Y, et al. Structured dynamic time warping for continuous hand trajectory gesture recognition[J]. Pattern Recognition, 2018, 80:21-31.

[3] 黄贻望,雷　彪. 基于 SVM 的数字识别系统设计[J]. 信息技术与信息化,2022(12):52-57.

[4] 梅家俊,王卫民,戴兴雨. 基于二阶隐马尔可夫模型的连续手语识别[J]. 计算机系统应用,2022,31(4):375-380.

[5] CHANG W D, MATSUOKA A, KIM K T, et al. Recognition of uni-stroke characters with hand movements in 3D space using convolutional neural networks[J]. Sensors, 2022, 22(16):1-15.

[6] SABRI N I A, SETUMIN S. One-shot learning for facial sketch recognition using the siamese convolutional neural network[C]//2021 11th IEEE symposium on computer applications & industrial electronics (ISCAIE). Penang: IEEE, 2021:307-312.

[7] REN H, WANG W, LU K, et al. An end-to-end recognizer for in-air handwritten Chinese characters based on a new recurrent neural networks[C]//2017 IEEE international conference on multimedia and expo (ICME). Hong Kong, China: IEEE, 2017:841-846.

[8] ZHANG X Y, YIN F, ZHANG Y M, et al. Drawing and recognizing chinese characters with recurrent neural network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4):849-862.

[9] ALAM M S, KWON K C, ALAM M A, et al. Trajectory-based air-writing recognition using deep neural network and depth sensor[J]. Sensors, 2020, 20(2):1-16.

[10] ABDULLAHI S B, CHAMNONGTHAI K. American sign language words recognition using spatio-temporal prosodic and angle features: a sequential learning approach[J]. IEEE Access, 2022, 10:15911-15923.

[11] CHOUDHURY A, SARMA K K. A CNN-LSTM based ensemble framework for in-air handwritten Assamese character recognition[J]. Multimedia Tools and Applications, 2021, 80(28):35649-35684.

[12] SIMAYI W, IBRAYIM M, HAMDULLA A. Study the preprocessing effect on RNN based online Uyghur handwritten word recognition[J]. Wireless Networks, 2021, 295:1-12.

[13] YANAY T, SHMUELI E. Air-writing recognition using smart-bands[J]. Pervasive and Mobile Computing, 2020, 66:1-16.

[14] GAN J, WANG W, LU K. A unified CNN-RNN approach for in-air handwritten English word recognition[C]//2018 IEEE international conference on multimedia and expo (ICME). San Diego: IEEE, 2018:1-6.

[15] REN H, WANG W, LIU C. Recognizing online handwritten Chinese characters using RNNs with new computing architectures[J]. Pattern Recognition, 2019, 93:179-192.

[16] 党小超,殷　杰,郝占军,等. 基于 CSI 的空中手写数字识别方法[J]. 传感器与微系统,2022(9):29-33.

[17] 陈　路,陈道喜,陆一鸣,等. 基于注意力机制编码器-解码器的手写数学公式识别模型[J]. 计算机应用,2023,43(4):1297-1302.

[18] 付鹏斌,李树军,杨惠荣. 基于双模编码器-解码器框架的联机手写数学公式识别[J]. 北京工业大学学报,2024,50(1):50-60.

[19] WANG P, LIU L, SHEN C, et al. Multi-attention network for one shot learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017:2721-2729.

[20] CHURCH K W. Word2Vec[J]. Natural Language Engineering, 2017, 23(1):155-162.

[21] YANA B, ONOYE T. Fusion networks for air-writing recognition[C]//MultiMedia modeling: 24th international conference. Bangkok: Springer, 2018:142-152.

[22] 甘　吉. 手写文字识别及相关问题算法研究[D]. 北京:中国科学院大学(中国科学院计算机科学与技术学院), 2021.

[23] MEIBL F, EIBENSTEINER F, PETZ P, et al. Online handwriting recognition using LSTM on microcontroller and IMU sensors[C]//2022 21st IEEE international conference on machine learning and applications. Nassau: IEEE, 2022:999-1004.

[24] 李波燕,张　勇,袁德荣,等. 基于注意力机制的手写体数字识别[J]. 计算机科学,2022,49(S2):626-630.

[25] AO X, ZHANG X Y, LIU C L. Cross-modal prototype learning for zero-shot handwritten character recognition[J]. Pattern Recognition, 2022, 131:1-13.