计算机科学
COMPUTER SCIENCE

多字体印刷体维-哈-柯文关键词图像识别

沙尔旦尔·帕尔哈提，阿布都热合曼·卡的尔，阿力木江·亚森

引用本文

沙尔旦尔·帕尔哈提，阿布都热合曼·卡的尔，阿力木江·亚森.多字体印刷体维-哈-柯文关键词图像识别[J].计算机科学, 2022, 49(11A): 211100038-6.

SARDAR Parhat, ABDURAHMAN Kadir, ALIMJAN Yasin. Multi-font Printed Uyghur-Kazakh-Kirghiz Keyword Image Recognition [J]. Computer Science, 2022, 49(11A): 211100038-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100038

多字体印刷体维-哈-柯文关键词图像识别

沙尔旦尔·帕尔哈提　阿布都热合曼·卡的尔　阿力木江·亚森
新疆财经大学信息管理学院　乌鲁木齐 830012
(sardar312@163.com)

摘　要　针对印刷体维吾尔文字识别中字体单一、识别数据规模小、识别领域不区分以及哈萨克和柯尔克孜文字识别研究缺乏等问题,提出了基于卷积神经网络(CNN)的多字体印刷体维吾尔、哈萨克和柯尔克孜(以下简称维-哈-柯)文关键词识别方法。首先,针对维-哈-柯文关键词图像语料库缺乏的问题,基于图像合成技术构建包括32种字体的维-哈-柯文关键词图像数据集。然后,使用数据扩充技术对数据集的图像进行不同程度的加噪、旋转和失真操作,来进一步体现数据集的自然场景特征。最后,使用多层CNN网络在该数据集上训练图像识别模型,均得到了96.5%以上的识别准确率,并在包括3种常用字体的实际印刷体图像识别任务中得到了96%左右的准确率,该方法减少了预处理过程,并胜过了以往机器学习框架下的其他识别方法。实验结果表明,在CNN网络框架下基于合成图像和数据扩充技术的识别方法能够较好地实现多字体印刷体维-哈-柯文图像识别任务。

关键词:维-哈-柯语;OCR;图像合成;卷积神经网络;关键词图像识别
中图法分类号　TP391

Multi-font Printed Uyghur-Kazakh-Kirghiz Keyword Image Recognition

SARDAR Parhat, ABDURAHMAN Kadir and ALIMJAN Yasin
School of Information Management, Xinjiang University of Finance and Economics, Urumqi 830012, China

Abstract　Aiming at the problems of single font type, small size of recognition data, indistinguishable recognition fields and lack of research on Kazakh and Kirghiz printed character recognition, a multi-font printed Uyghur-Kazakh-Kirghiz keyword recognition method based on convolutional neural network(CNN) is proposed. Firstly, aiming at the problem of lack of Uyghur-Kazakh-Kirghiz printed image corpus, based on image synthesis technique, a Uyghur-Kazakh-Kirghiz keyword image data set including 32 font type is constructed. Secondly, using data augmentation technology to add different level of noise, rotation and distortion effects on these images to further reflect the natural scene features of the data set. Thirdly, using a multi-layer CNN network to train the image recognition model on this data set, and obtaining the recognition accuracy over 96.5%, and the accuracy of about 96% is obtained in the actual print image recognition task including 3 commonly used fonts. This method has fewer pre-processing steps and it outperforms previous recognition approaches within the classical machine learning framework. Experimental results show that the recognition method based on synthetic image data can better realize the task of multi-font printed Uyghur-Kazakh-Kirghiz image recognition.

Keywords　Uyghur-Kazakh-Kirghiz, OCR, Image synthesis, Convolutional neural network, Keyword image recognition

1　概述

印刷体关键词识别的任务是按照用户查询内容从文档图像中查找单词或短语[1-3],其在文档图像的自动分类[4-5]、多模态文档信息分析以及分类[6-7]等任务中有着很重要的作用。随着信息技术的飞速发展,大量的纸质文档被转换成电子图像文档,并被存储在数据库中,从而促进了数字图书馆的出现。图像和视频中关键词的自动识别不仅是主流语言的重要研究课题,也是多语言乃至古文字研究的重要课题。以图像形式呈现的文档是原始的,因此不便于从这些图像数据中搜索[8]、索引[9-10]或者检索[11-12]用户所需要的信息。

处理多样化的文档图像数据库时,字体和样式的多变化使得识别任务具有挑战性[13]。对于维吾尔语、哈萨克语和柯尔克孜语(以下简称维-哈-柯语)等词汇量大的派生类语言而言,在进行印刷体文字识别研究时,可以选择较小的识别单元如字符,也可以选择较大的识别单元如单词作为基本的模式识别单元。两者各有利弊,给定语言中的词语字体和书写风格多样时,选择作为识别单元的字符很容易被混淆,特别是在

基金项目:国家自然科学基金(61662073);2020年新疆维吾尔自治区天池博士计划项目;新疆财经大学校级科研基金项目(2022XGC022,2022XGC049)
This work was supported by the National Natural Science Foundation of China(61662073), 2020 Xinjiang Uyghur Autonomeus Ragion Tianchi Doctor Plan Project and Xinjiang University of Finance and Economics School Level Scientific Research Foundation Project(2022XGC022, 2022XGC049).
通信作者:阿力木江·亚森(81805794@qq.com)

手写体文字上。如果选择像单词这样较大的识别单元，则能够容易被区分，但搜索维度就会变大。然而，神经网络框架的迅速发展和深层神经网络结构的高效率使得高精度的高维分类成为了可能。

维-哈-柯文使用阿拉伯字母作为书写形式，维语中有 32 个音素，哈语中有 33 个音素，柯语中有 36 个音素。32 个维语音素用 32 种字母表示，33 个哈语音素用 33 种字母表示，而 36 个柯语音素用 36 种字母表示。维-哈-柯文字是从右到左，从左到右逐行书写的，每种字母根据其单词内的位置的不同会有 4 种不同的表面形式(首、中、尾、独)，如表 1 所列。

表 1 维-哈-柯语阿拉伯文字表面形式示例

Table 1 Examples of surface forms of Uyghur-Kazakh-Kirghiz characters in Arabic

语言	字母	首	中	尾	独
维吾尔	元音	قۇرۇق	قەلەم	باغچە	ئىرادە
	辅音	يەكتەپ	قايناق	مۇئەللىم	ئادەم
哈萨克	元音	ۇزدەۋ	قەلەم	كونە	سارە
	辅音	مىسال	الما	مۇعالىم	الەم
柯尔克孜	元音	ىراقات	سەرجىمان	كومانداۋ	مۇنىرا
	辅音	مالچى	ئىمارات	تاارىمايىم	جوقتام

从表 1 可以看出，维-哈-柯语中给定字母的首位形式会在单词的初始位置中出现，首位字母前面有标点符号或者空格，而后面与其他中或者尾位字母相连。中位字母会出现在给定单词的中间位置，且前面和后面都会与其他字母相连。尾位字母出现的位置是在给定单词的尾部，前面相连于其他的字母，而后面不会连接字母，只能有标点符号或者空格。独位字母不论前面还是后面都不会与其他字母相连，出现在给定单词时以独立的形式存在。

部分学者对印刷体维吾尔文进行了识别研究[14-18]。Chen 等[14]提出了基于区域切分及模板匹配的印刷体维吾尔文字母识别方法，并对待识别的字母与事先准备好的字母图像模板库进行匹配，来实现维吾尔文字母的识别。Bai 从印刷体维吾尔文单词中提取 Gabor 特征和梯度特征，并分别用欧氏距离和 BP 神经网络分类器，对包括 5000 个单词的印刷体维吾尔文字图像进行了识别研究[15]。Lang 用梯度特征和方向线素特征等方法提取印刷体维吾尔文字母的特征，并用欧氏距离分类器对包括 1408 个字母的印刷体维吾尔文字图像进行了识别研究[16]。Wang 提出了基于 Gabor 特征和 zernike 矩特征的印刷体维吾尔文字识别方法，用欧氏距离分类器对包括 5000 个单词的印刷体维吾尔文字图像进行了识别研究[17]。Yu 等提出了基于方向梯度直方图(Histogram of Oriented Gradient，HOG)特征结合多层感知机(Multi-Layer Perceptron，MLP)算法的印刷体维吾尔文字识别方法，先用 HOG 特征提取方法从维吾尔文字母图像中提取其 HOG 特征，然后用 MLP 算法对 1762 个印刷体维吾尔文字母图像进行了识别研究[18]。目前，在国内外学术资源上还没有公开发表的印刷体哈-柯文字识别的相关研究。

以往的印刷体维吾尔文字识别方法以单个字体样式的小规模文字图像语料库为实验对象，用传统的方法进行识别研究，其存在计算量大、人工选择特征以及对图像的分辨率变化敏感等缺点。因此，对于低资源、形态丰富的维-哈-柯等语言来说，自动、灵活的印刷体文字图像关键词识别技术是必不可少的。深层神经网络在不考虑人工选择的特征的情况下提供了方便的机器学习方法。本文用 CNN 对多字体印刷体维-哈-柯文关键词图像进行了识别研究。

2 多字体印刷体维-哈-柯文关键词识别方法

本文将维-哈-柯文字中所有常用字体的多种字体样式以图像形式准备并规范化。同时将噪声、旋转和失真效应添加到每个单词图像中，把准备好的语料库送入卷积神经网络，以监督学习方式训练卷积层和前馈层，神经网络的分类数和输出是词汇量。

2.1 基于合成技术的维-哈-柯文关键词图像数据

对于含有噪声、失真的文本图像，或者手写体粗暴的图像 OCR 而言，全词识别是较好的选择。与小单元识别方法相比，较大的上下文信息能够提供可靠的识别结果。目前，神经网络框架使得这样的结果变得更为可能。

低资源语言采集大型图像文件是不容易的，而采集到的真实图像难以覆盖大派生词汇量以及噪声和失真的图像副本。为了包括大量的词汇，有必要从文本中准备包括大量词汇的字典，然后用适当的编程语言自动生成词典词汇的图像。为了匹配真实世界的图像特性，可以以各种形式和参数向这些词典图像添加噪声和失真。这样，训练和测试的语料库是通过添加各种字体样式的词典图像副本混合的。通过分别添加噪声，旋转和失真来进一步扩展每个图像的副本。

2.2 神经网络模型的结构

卷积神经网络是一种深度学习模型，具有隐式检测因变量和自变量之间复杂的非线性关系的能力。卷积神经网络可以通过卷积核和池化层自动生成特征，避免了特征提取的不稳定性和盲目性，并且计算速度快，对输入图像的大小、旋转和位置等不敏感。因此，基于卷积神经网络的方法最近在文本图像的检测或分类方面取得了实质性的进展[19-20]。与传统特征提取方法不同，卷积神经网络通过卷积核来提取特征，每个神经元的输入连接到前一层[21]的局部感受区域，并且通过卷积核计算局部特征。通过卷积窗口的移动生成特征平面，每个特征平面共享一个相似的卷积核，从而实现权重共享，减少权重数量。卷积神经网络主要用于二维图像的识别。在卷积神经网络中，共享权值是通过监督学习获得的，可以避免人工提取特征。因此，卷积神经网络具有从训练数据中学习共享权重的优点。通常，卷积神经网络分为多个层，其中一层是卷积层，另一层是池化层，并且可以有多个卷积层和池化层。它们分别用于特征提取和特征参数处理。本文所采用的卷积神经网络框架如图 1 所示(图 1 以印刷体维文关键词识别为例)。

本文用到的神经网络底层设计窄(神经元数量少)，原因是输入图像的维度大，输入层用少量核函数或神经元可以降低模型大小和计算量。此外，识别图像所包含的基本识别单元(基元)，如线段、交、环等数量不多，输入层可以设计少量神经元，并用卷积的方式探索这些基元。

关键词图像输入卷积层之后，卷积层将输入图像分割成称为核的较小局部窗口，这些核矩阵通过最大池化层被转换成深度向量。后续而来的进一步的卷积和最大池化操作可以捕获局部表面特征并增加特征提取的深度。最后将若干个前馈神经网络连接起来，并将最后一个分类层作为输出层，输出类别的标签，输出层的大小与语料库的词汇量大小相等。

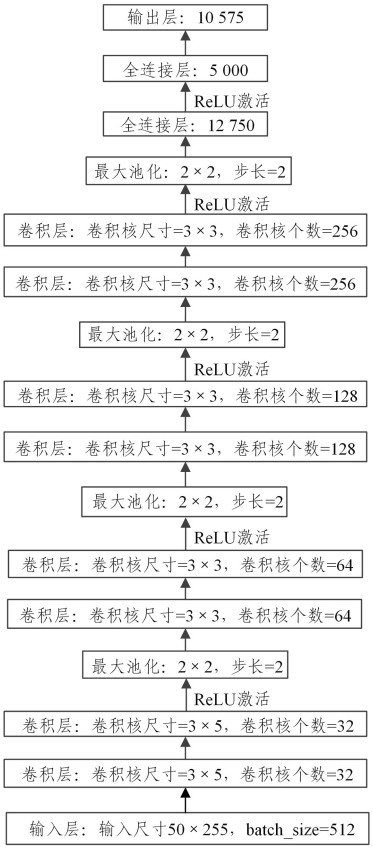


图 1　卷积神经网络框架

Fig. 1　Convolutional neural network framework

3　实验结果及分析

目前，针对印刷体维-哈-柯文字图像识别的研究还处于起步阶段，尚无公开的文字图像语料库，更没有关键词图像数据，可供我们进行特征提取和识别实验。因此，我们首先要建立印刷体维-哈-柯文关键词图像语料库，并通过添加噪声和失真来模拟真实图像。

3.1　基于合成技术的维-哈-柯文关键词图像数据

关键词是对于一篇文章或一部著作而言，能够体现其中的中心概念的词语，如：对于教育类文章而言，“学校”这个词可以是个关键词，读者在阅读给定的文章之前如果看到该文章的关键词是“学校”，那么在大体上能够判断该文章是关于教育相关的。本文开发了一个网络爬虫工具，并用此工具从人民网、天山网等权威的官方网站下载了 9 个类别的 8100 个维语文本、8 个类别的 7200 个哈语文本、6 个类别的 3219 个柯语文本。然后用 TFIDF 算法[22]从这些文本中分别选取 10575 个维语词语、9748 个哈语词语和 7246 个柯语词语，以此构建维-哈-柯语关键词数据，将这些关键词转换成具有不同字体样式的合成印刷体关键词图像。在多语言和双向操作系统中，文本到图像的转换是一项不容易的工作。阿拉伯字母在单词的不同位置以不同的表面形式出现。因此，需要用支持双向多语言的编程语言来准备具有各种字体样式的印刷体维-哈-柯文关键词图像数据。

目前没有现成的维-哈-柯等少数民族语言单词图像合成工具。C＃编程语言在 DOT. NET 平台下工作，DOT. NET 平台对维-哈-柯等字母通过链接形式构成单词的少数民族语言的支持度强，因此根据研究的需要，本文基于 C＃编程语言用 GDAL 图像处理包的 DrawString 方法，开发了一种能够将维-哈-柯语单词从文本格式转换成图像格式的工具，如图 2 所示。首先分别准备文本格式的维-哈-柯语单词，然后选择字体大小和样式之后，将其输入给该图像合成工具，该合成工具输出图像格式的单词。本文使用该合成工具从上述的关键词中分别生成包括每个词 32 种字体样式的合成印刷体维-哈-柯文关键词图像，构建了实验数据集，如图 3 的例子所示。这样，本文总共得到了 338400 个印刷体维吾尔文关键词图像，331936 个印刷体哈萨克文关键词图像和 231872 个印刷体柯尔克孜文关键词图像。该合成工具可以生成所有可用的维-哈-柯文字体和不同字体大小的单词图像。

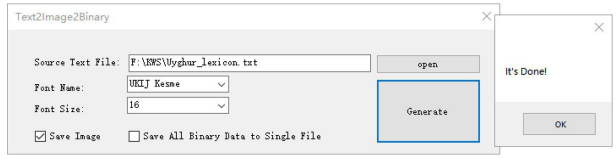
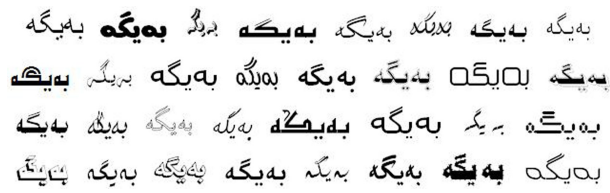
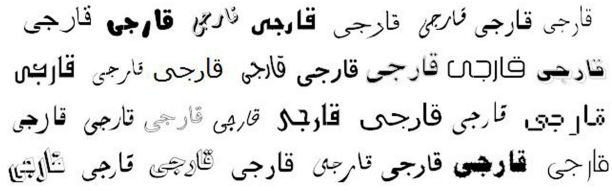


图 2　印刷体维-哈-柯文单词图像合成工具

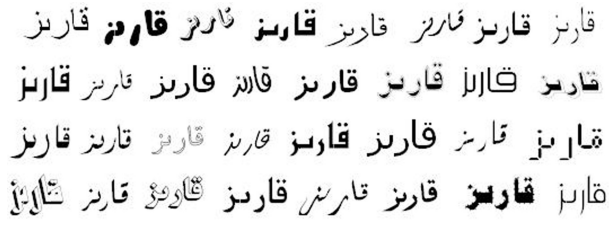
Fig. 2　Printed Uyghur-Kazakh-Kirghiz word image synthesis tool



(a)不同字体的印刷体维文单词图像示例



(b)不同字体的印刷体哈文单词图像示例



(c)不同字体的印刷体柯文单词图像示例

图 3　不同字体印刷体单词图像示例

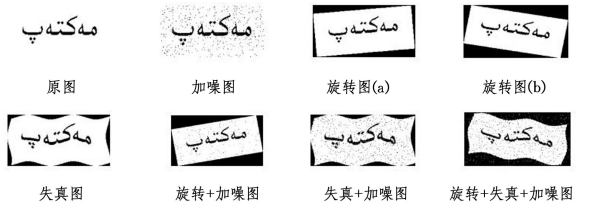
Fig. 3　Examples of printed word images with different fonts

该合成工具能够为每个词导出不同大小的矩阵图像。本文考虑到如果图像的高度是 50，则基本上可以符合存放大部分的印刷体维-哈-柯文单词，因此软件的输出高度设计为 50。至于图像的宽度，因为维-哈-柯语是派生类语言，词缀附加词干而成的单词的长度往往是不一样的，有些单词短一点，而有些单词则长很多。因此，为了使符合生成不同长度的单词图像，本文将该合成软件的输出宽度设计为 255，即选 50×255 像素的图像大小作为用于神经网络的关键词图像的尺寸。

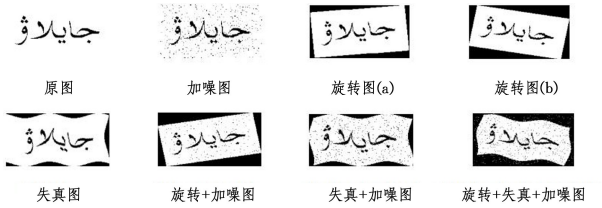
3.2　印刷体维-哈-柯文关键词图像数据扩充

由于光学扫描装置以及照相机等设备在实际应用中受到了各种噪声因素的干扰，在实际的文档图像中不可避免出现多余的黑白像素点、失真和旋转等各种情况。通常的应对策略是在模型训练期间随机转换每个输入图像，扩大训练集以提高性能[23]。为了使数据集中的合成图像能够具有实际生成的文档图像的特征，本文对所有的维-哈-柯文关键词图像

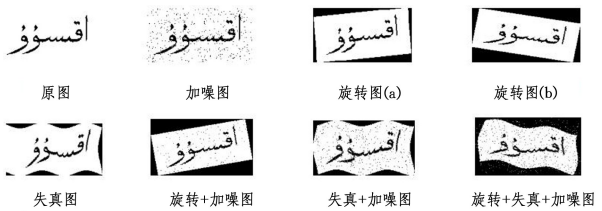
进行随机的噪声添加(本文用辣盐噪声方法,采用2种参数值,其分别为0.02和0.05)、旋转(本文用4种不同旋转角度,角度取值分别设置为10°,5°,-5°,-10°)和失真(本文用一种失真操作,其失真的幅度设置为5,周期设置为100)等操作以及这3种操作的组合,来改变轮廓的灰度分布,以此生成新的图像,如图4的例子所示(图4给出了其中比较典型的样本例子)。这样,关键词图像数据集的规模增加了30倍,并分别包括了10152000个维吾尔文关键词图像,9958080个哈萨克文关键词图像和6956160个柯尔克孜文关键词图像,通过数据扩充可以提高识别模型的泛化效果。通过使用不同的旋转角度、不同的噪声点生成方法和随机失真生成算法,可以在需要时获得更多的数据。本文将这3个数据集分成训练集(75%)、验证集(10%)和测试集(15%),使用pytorch在支持GPU的Linux CentOS操作系统上实现了CNN模型框架。



(a)加噪、旋转和失真的维文单词图像示例



(b)加噪、旋转和失真的哈文单词图像示例



(c)加噪、旋转和失真的柯文单词图像示例

图4 加噪、旋转和失真的单词图像示例

Fig. 4 Examples of noisy, rotated and distorted word images

构建维-哈-柯文关键词图像数据集后,3个数据集的每个图像都被调整为值为0和1的50×255像素的矩阵图像,以实现图像归一化。

3.3 实验参数设置

本文中输入给CNN网络的图像为50×255像素的长方形图像,因此本文在第一组卷积层上使用了尺寸为3×5的卷积核,而从第二组卷积层开始在所有的卷积层上都使用了尺寸为3×3的卷积核,卷积层上滑动窗口的步长设置为2,采用了最大池化方法,池化层上设置了尺寸为2×2的池化核,池化步长设置为2,采用了零补充方法。本文在第一组卷积层上设置了32个卷积核,在第二组卷积层上设置了64个卷积核,在第三组卷积层上设置了128个卷积核,在第四组卷积层上设置了256个卷积核,每组卷积层中的前后2个卷积层上所使用的卷积核个数是相同的。本文用了交叉熵损失函数和Adam优化函数,训练网络时,网络的学习速率设置为0.001,用dropout策略避免在训练模型时出现过拟合的问题,其dropout的值设置为0.5,网络训练采用了mini-batch训练方法,其batch_size设置为512,使用早停策略,在迭代次数的增加不能提升分类模型的准确率时停止训练。

3.4 评价指标

常用于评价分类器性能的指标有准确率、精确率、召回率等,本文使用准确率评测了所提方法的性能。对于某一个类别C_i的分类结果而言,如果正确分为该类的文本数目是a,错误划归为该类的文本数目是b,将该类文本错误划归为其他类的文本数目是c,属于其他类的文本正确分为所属类的文本数为d,则准确率的计算式如下:

$$准确率=\frac{a+d}{a+b+c+d}$$

3.5 实验结果与分析

本文实验中将大小为50×255像素的矩阵图像输入给CNN网络之后,通过尺寸为3×5的卷积核从第一组卷积层中提取其特征,并通过池化层对这些特征进行子采样,随后在第二组到第四组的卷积层上也进行特征提取以及池化等操作之后,两个全连接层和输出层被连接起来。本文用了整体识别即典型的模式识别方案,因此训练集和测试集所包含的类别数是相同的。

(1)迭代轮次对识别结果的影响

神经网络在训练过程中通过迭代计算来获得权重,经过多次迭代后得到理想的参数,本文在本次多字体印刷体维-哈-柯文关键词识别实验中共进行了100个轮次(epoch)的迭代运算,并从第一次迭代到第十次迭代运算,都生成了每次迭代的准确率和损失函数的验证结果,然后每10个轮次的迭代运算后产生一次准确率和损失函数的验证结果,如图5和图6所示。

从图5-图6可以看出,训练的迭代轮次从1到8次时,模型的识别准确率基本为0,模型的学习效果很差,迭代轮次为8~10次左右时准确率稍微提高。经过10~20次的迭代运算后,识别准确率变化显著,模型开始较好地学习图像中的特征。经过30次迭代运算之后,3个数据集上的识别率均接近80%左右。迭代轮次超过60次之后,模型在3种数据集上的准确率都超过了96%,在随后的迭代运算过程中,准确率分别达到了96.77%,96.89%和96.95%的峰值,然后开始收敛于96.7%左右。维-哈-柯文字用相同的字母,此外,构建3个语言的印刷体关键词图像数据的方法也是相同的,因此在数据集的图像上实现形状和灰度分布变化的规则是相似的,故模型在3个数据集上的识别准确率很接近。3个数据集上模型损失值随着迭代次数的增加而减少,并经过40次迭代运算后,模型在3个数据集上的损失率都降低到约为10%。显然,基于多层CNN网络的识别方法在多字体印刷体关键词识别任务中具有很好的识别效果。

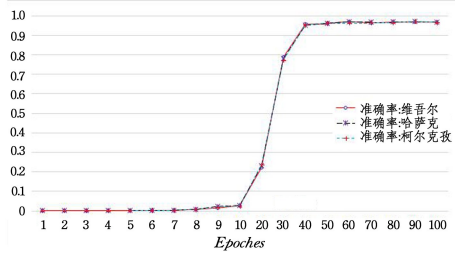


图5 迭代轮次对识别结果的影响

Fig. 5 Influence of iteration on recognition results

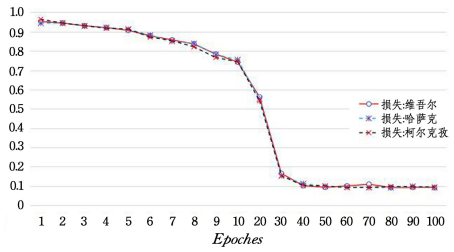


图 6　迭代轮次对识别结果的影响

Fig. 6　Influence of iteration on recognition results

(2)网络的深度(卷积层数量)对识别结果的影响

本文将设计以下几种 CNN 网络结构，并对基于这几种网络的多字体印刷体维-哈-柯文关键词识别结果与本文方法的识别结果进行了比较。

(1)CNN_Succ-2&Conv-1。该网络只包括一组前后卷积层，随后是一个池化层和全连接层。卷积层上设置了 32 个卷积核，卷积核的尺寸设置为 3×5。网络的其他参数设置都与本文方法的参数设置相同。

(2)CNN_Succ-2&Conv-2。该网络包括 2 组前后卷积层，每组卷积层后跟随一个池化层，最后是全连接层。第一组卷积层上设置了 32 个卷积核，卷积核尺寸设置为 3×5，第二组卷积层上设置了 64 个卷积核，卷积核尺寸设置为 3×3。网络的其他参数设置都与本文方法的参数设置相同。

(3)CNN_Succ-2&Conv-3。该网络包括 3 组前后卷积层，每组卷积层后跟随一个池化层，最后是全连接层。第一组卷积层上设置了 32 个卷积核，卷积核尺寸设置为 3×5，第二组卷积层上设置了 64 个卷积核，第三组卷积层上设置了 128 个卷积核，从第二组开始的两组卷积层上的卷积核尺寸都设置为 3×3。网络的其他参数设置都与本文方法的参数设置相同。

上述的所有网络中池化层选用最大池化方法，池化的步长设置为 2。以上方法都进行了 100 个轮次的迭代运算，最佳识别结果的比较如表 2 所列。

表 2　网络深度对识别结果的影响

Table 2　Influence of network depth onrecognition results

模型	数据集	准确率/%	损失
CNN_Succ-2&Conv-1	维吾尔	95.13	0.1533
	哈萨克	95.12	0.1582
	柯尔克孜	95.24	0.1557
CNN_Succ-2&Conv-2	维吾尔	95.44	0.1511
	哈萨克	95.48	0.1565
	柯尔克孜	95.69	0.1483
CNN_Succ-2&Conv-3	维吾尔	96.31	0.1071
	哈萨克	96.24	0.1024
	柯尔克孜	96.27	0.1092
CNN_Succ-2&Conv-4	维吾尔	96.77	0.0914
	哈萨克	96.89	0.0915
	柯尔克孜	96.95	0.0906

从表 2 可以看出，随着模型深度的提高，在 3 个数据集上模型的识别效果都开始增加，与基于 1 个(一组)卷积层的 CNN 模型相比，基于 4 个(四组)卷积层的 CNN 模型的分类准确率超过了 1%，而多层网络结构中所得到的损失值也是显著地小于单层网络结构中所得到的损失值。模型深度的进一步提高在一定程度上提升了模型的识别效果，这是因为模型通过更多卷积和池化层的交替计算来提取图像中更高级和更强的特征，但是模型深度的提高也会导致模型大小的增加、训练时间和模型参数数量提升等问题。

(3)数据集规模的变化对识别效果的影响

本文在测试数据规模不变的情况下，用不同规模的训练数据训练识别模型，进行了识别实验，并对结果进行了比较，模型的训练轮次设置为 100 次，如表 3 所列。

表 3　在不同数据规模下识别结果的比较

Table 3　Comparison of recognition results with different data scales

训练数据集规模	准确率/%		
	维吾尔	哈萨克	柯尔克孜
20%	93.67	93.58	93.84
50%	95.18	95.19	95.32
80%	96.64	96.72	96.89
100%	96.77	96.89	96.95

从表 3 可以看出，随着训练数据规模的增加，模型的识别效果也被提升。这是因为在提供给模型的图像数据(图像的不同副本)更多时，图像的特征更具有代表性，模型能够学习能覆盖更多场景的特征，因此模型识别率得到了进一步提高。

(4)与以往方法的识别结果比较

本文将以往的部分印刷体维吾尔文字识别有关的研究结果与本文方法的结果进行了比较，结果如表 4 所列。从表 4 可以看出，本文方法的识别结果相比以往的最高的识别结果(见文献[15])超出 0.62%。相比以往的学者发表的印刷体文字识别研究结果，本文方法有以下优点：首先，以往的学者都是在单个语言(仅维吾尔语)文字上做研究的，而本文方法在维-哈-柯 3 种语言的印刷体关键词图像上进行了识别研究；其次，在本文方法中，图像的特征提取过程是网络自动完成的，不需要像以往方法那样人工进行特征提取，因此可以节省时间并提高特征提取的效果；然后，本文实验数据规模比以往的学者所用过的数据规模 2～20 倍左右；最后，以往的识别方法只能用于单个字体的文字图像，而且文字内容不区分领域，而本文方法可以用于多个领域多种字体的印刷体文字识别任务。

表 4　与以往方法识别结果的比较

Table 4　Comparison with the recognition results of previous methods

方法	识别单元	识别方法	特征	字体	目标语言	数据集规模/(词/字)	准确率/%
文献[14]	字	模板匹配方法	结构特征	单一	维吾尔	400	94.00
文献[15]	词	欧氏距离分类器	Gabor特征	单一	维吾尔	5000	96.50
文献[16]	字	欧氏距离分类器	方向线素特征	单一	维吾尔	1408	91.26
文献[17]	词	欧氏距离分类器	zernike矩特征	单一	维吾尔	5000	70.98
文献[18]	字	MLP分类器	HOG特征	单一	维吾尔	1762	96.15
本文方法	词	CNN_Succ-2&Conv-4	自动学习特征	32种	维吾尔	10575	96.77
					哈萨克	9748	96.89
					柯尔克孜	7246	96.95

(5)实际印刷体关键词图像识别效果

本文从用于合成关键词图像的字体样式中选择维-哈-柯文印刷文档最常用的 3 种字体为实验对象，扫描生成实际关键词图像数据，并把它作为测试集进行了识别实验，验证了基于合成图像数据训练好的识别模型在实际印刷体图像上的识别效果，如表 5 所列。

表 5 模型在实际图像数据集上的识别效果

Table 5 Recognition effect of the model in actual image data set

数据集	测试字体/种	训练字体/种	测试集/个	训练集/个	准确率/%
维吾尔	3	32	31725	10152000	95.98
哈萨克			29244	9958080	96.14
柯尔克孜			21738	6956160	96.43

从表 5 可以看出，训练数据的规模比实际测试数据规模多 10 倍左右，实际的维-哈-柯印刷图像上得到的识别准确率分别为 95.98%，96.14%和 96.43%，其识别准确率比合成图像作为测试集时的识别准确率略低，其主要原因是，合成生成的训练图像数据特征不能够完全覆盖在生成实际图像数据时扫描装机以及周围环境的部分噪声因素，但是识别效果还是很理想的，通过合成技术进一步扩充训练数据来进一步提高识别效果。

结束语 多字体印刷体关键词识别是一项重要的任务，是在图像形式文档的自动归档、文档图像的分析等文档处理任务中的关键因素。维-哈-柯文书籍、杂志、报纸等各种图书资源在实际使用中常会用各种不同的字体样式且内容来自各种不同领域，不同字体的单词在形状、轮廓和灰度分布等特征上存在很大的不规则性。因此，基于单个字体及内容不区分领域的文字识别方法在多样化的字体背景和多个领域背景下识别效果不是很理想。本文讨论了一种基于图像合成技术和 CNN 网络的多字体印刷体维哈柯文关键词图像识别方法。在合成构建图像数据集上使用数据扩种技术进行加噪、旋转和失真处理，以进一步提高数据集的覆盖率，并用多层 CNN 网络构建图像识别模型，在合成图像和实际图像数据上分别得到了 96.5%以上和 96%左右的准确率。可见，基于合成图像数据的识别方法可以较好地实现多种字体背景下的维-哈-柯文关键词图像的识别效果。维-哈-柯语的派生形态结构和 OOV(Out of Vocabulary)也给派生语言的自然语言处理工作带来了困难，由于派生类特点，维-哈-柯语的词汇量是非常大的，因此模型识别语料库之外的单词图像时遇到很大的问题。本文将在未来的工作中重点研究 OOV 问题。

参 考 文 献

[1] DOERMANN D. The Indexing and Retrieval of Document Images: A Survey[J]. Computer Vision and Image Understanding, 1998, 70(3): 287-298.

[2] ALAEI F, ALAEI A, BLUMENSTEIN M, et al. A Brief Review of Document Image Retrieval Methods: Recent Advances[C]// 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016: 3500-3507.

[3] LI L. Research on document image classification and retrieval method based on convolutional neural network[D]. Huazhong: Huazhong University of Science and Technology, 2017.

[4] NOCE L, GALLO I, ZAMBERLETTI A, et al. Embedded Textual Content for Document Image Classification with Convolutional Neural Networks[C]// Acm Symposium on Document Engineering. ACM, 2016: 165-173.

[5] DAS A, ROY S, BHATTACHARYA U. Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks[J]. arXiv: 1801.09321, 2018.

[6] AUDEBERT N, HEROLD C, SLIMANI K, et al. Multimodal Deep Networks for Text and Image-based Document Classification[J]. arXiv: 1907.06370, 2019.

[7] BAGADKAR S L, MALIK L G. Review on Extraction Techniques for Images, Textlines and Keywords From Document Image[C]// IEEE International Conference on Computational Intelligence & Computing Research. IEEE, 2015: 1-3.

[8] JIANG Y X, DING S C, WU P. A Study on the Classification of Features of Multi-Modal Information Based on BiLSTM-VGG16 [J]. Information Studies: Theory & Application, 2021, 44(11): 180-186.

[9] HARLEY W A, UFKES A, DERPANIS K G. Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval[J]. arXiv: 1502.07058, 2015.

[10] AN Y H, DONG W Z. Research on Segmentation Method of Adhesive Characters based on Recognition Feedback[J]. Journal of Hebei Academy of Sciences, 2008, 25(2): 34-38.

[11] SHIN C, DOERMANN D. Structural Similarity for Document Image Classification and Retrieval[J]. Pattern Recognition Letters, 2014, 43(1): 119-126.

[12] ZHI T, HUANG W, TONG H. Detecting Text in Natural Image with Connectionist Text Proposal Network[J]. arXiv: 1609.03605, 2015.

[13] RANJAN V, HARIT G, JAWAHAR C V. Enhancing Word Image Retrieval in Presence of Font Variations[C]// Proceedings of the 2014 22nd International Conference on Pattern Recognition. IEEE Computer Society, 2014: 2709-2714.

[14] CHEN Q, YUAN B S, LI X, et al. Research on Printed Uyghur Character Recognition based on Template Matching[J]. Computer Technology and Development, 2012, 22(4): 119-122.

[15] BAI Y H. Printed Uyghur Word Recognition[D]. Xi'an: Xi'an University of Electronic Science and Technology, 2014.

[16] LANG X. Printed Uyghur Word Recognition based on Segmentation[D]. Xi'an: Xi'an University of Electronic Science and Technology, 2015.

[17] WANG X D. Research and Application of Key Technologies for Printed Uyghur Character Recognition[D]. Xi'an: Xi'an University of Electronic Science and Technology, 2017.

[18] YU L, YASIN A. Printed Uyghur Character Recognition Method based on HOG Feature and MLP Classifier[J]. Microcomputer Application, 2017, 33(6): 30-33.

[19] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Neural Information Processing Systems, 2012, 25: 1106-1114.

[20] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks [C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017: 1137-1149.

[21] HUBEL D H, WEISEL T N. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex [J]. The Journal of Physiology, 1962, 160(1): 106-154.

[22] SARDAR P, MIJIT A, ASKAR H. Research on Keyword Extraction of Uyghur-Kazakh Text based on Stem Unit[J]. Computer Engineering and Science, 2020, 42(1): 131-137.

[23] CHRIS T, TONY M. Analysis of Convolutional Neural Networks for Document Image Classifification[J]. arXiv: 1708-03273, 2017.

SARDAR Parhat, born in 1984, Ph. D. His main research interest includes text and image information retrieval.

ALIMJAN Yasin, born in 1985, Ph. D. His main research interests include programming language and formal system.