基于锚点的字符级甲骨图像自动标注算法研究

史先进[1,2],曹 爽[1],张重生[1],陶月锋[1],吕灵灵[3],沈夏炯[1]
(1. 河南大学计算机与信息工程学院,河南大学黄河文化遗产实验室,河南开封 475004; 2. 河南省电化教育馆,河南郑州 450004;
3. 华北水利水电大学电力学院,河南郑州 450045)

摘 要: 甲骨文是中国最早的系统文字,是目前能见到的最早的成熟汉字.甲骨文的研究对历史探究和文化传承具有重要的意义.但是要实现字符级别的甲骨字符图像标注,在现有技术环境下,只能通过资深甲骨学专家进行人工标注,不仅耗费人力资源,而且效率低下.针对这一问题,在前期工作中的甲骨字符图像识别模型的基础上,本文提出了一种甲骨字符图像自动标注算法.该算法通过先分列后切割的思想,先将甲骨拓片上的每一个字符图像归结到某一个特定列,再以锚点甲骨字为参考点,根据空间近邻关系找到甲骨原文中的字所对应的甲骨字符图像,从而实现了甲骨字符图像的自动标注.同时,将标注好的甲骨字符图像添加到样本数据集,并利用增广后的数据集(增加6~10倍)重新训练甲骨字符图像识别模型,有利于提高基于深度学习的甲骨文识别算法的识别准确度;以较小的成本大幅增加样本数量,也可以节约专家大量的时间和人力.

关键词: 甲骨文; 图像标注; 数据增广; 锚点; 空间近邻; 模式识别
中图分类号: TP311.5;TP391.1 **文献标识码:** A **文章编号:** 0372-2112(2021)10-2020-12
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20201191

Research on Automatic Annotation Algorithm for Character-level Oracle-Bone Images Based on Anchor Points

SHI Xian-jin[1,2], CAO Shuang[1], ZHANG Chong-sheng[1], TAO Yue-feng[1], LÜ Ling-ling[3], SHEN Xia-jiong[1]
(1. *School of Computer and Information Engineering, Laboratory of the Yellow River Cultural Heritage, Henan University, Kaifeng, Henan* 475004, *China*;
2. *Henan Electrochemical Education Center, Zhengzhou, Henan* 450004, *China*;
3. *School of Electric Power, North China University of Water Conservancy and Hydropower, Zhengzhou, Henan* 450045, *China*)

Abstract: Oracle-Bone inscriptions are the earliest systematic and mature Chinese characters presently discovered. The study of Oracle-Bone inscriptions is of great significance to historical exploration and cultural inheritance. However, in order to realize character-level Oracle-Bone image annotation, in the existing technical environment, only experienced experts in Oracle-Bone inscriptions can carry out manual annotation, which not only consumes human resources, but also is inefficient. Aiming at this problem, based on the Oracle-Bone image recognition model in the previous work, this paper proposes an automatic annotation algorithm for Oracle-Bone character images. In this algorithm, each character image on the Oracle-Bone rubbings is first reduced to a specific column. Then, the Oracle-Bone character images corresponding to the characters in the original text are found by taking the anchor point as the reference point and according to the nearest neighbor relation of space, so as to realize the automatic labeling of the Oracle-Bone character images.At the same time, the labeled Oracle-Bone images are added to the sample data set, and the original Oracle-Bone character image recognition model is retrained by using the augmented data set(6-10 times increase), which is conducive to improve the recognition accuracy of the Oracle-Bone character recognition algorithm based on deep learning. In this way, the number of samples can be greatly increased at a small cost, and a lot of time and manpower of experts can be saved.

Key words: Oracle-Bone inscriptions; image annotation; data augmentation; anchor point; spatial neighbor; pattern recognition

收稿日期:2020-10-26;修回日期:2021-09-29;责任编辑:孙瑶

1 引言

甲骨文是迄今为止我国发现的年代最早的成熟文字系统,是汉字的源头和中华优秀传统文化的根脉,有重要的史料价值[1]. 习近平在致甲骨文发现和研究120周年的贺信中指出,“殷墟甲骨文的重大发现在中华文明乃至人类文明发展史上具有划时代的意义”. 近百年来几代学者经过共同努力,编撰出版了一批工具书,在甲骨文的研究、传播方面起了很大的作用[2]. 但是由于甲骨文字年代久远、甲骨残缺和甲骨图像不清晰等,目前可识字仅有3000余个,仍存在大量的不可识字. 如何借助先进的计算机技术对其进行数字化展示、有效保护和便捷使用,具有重要的现实意义[3].

当前,借助计算机技术进行甲骨文字形研究已经成为甲骨学热门的研究方向. 文献[1]介绍了用于计算机处理的甲骨文字库、句法分析和综合智能知识库的建立方法以及计算机甲骨文辅助辨识分析的工作原理,采用区位码和拼音输入,对与现代汉字有对应关系的一千多个甲骨文字在现代汉字、音、意、词性、属性等方面作出详尽的标注解释,用VC++实现了它们之间的互查功能. 文献[3]利用分形几何的原理,通过计算字形以及各个象限的分形维数,将甲骨文字形形式化为一组分形描述码,再通过与甲骨文字形的分形特征库进行配准,从而识别甲骨文字形. 文献[4]提出一种基于图论的方法来识别甲骨文的理论和技术,它的核心思想是把甲骨文当作无向图来处理,提取它的图特征,并以此为识别依据. 文献[5]利用复杂网络对甲骨文进行了抽象和理解,并对未识甲骨文字的场景语义进行了预测.

现代文本检测技术特别是自然场景文本检测技术对甲骨字符检测极具借鉴意义. 文献[6]总结和分析了深度学习在场景文本检测与识别中的新的见解、最新的技术、重大进展和未来的发展趋势. 文献[7]提出了一种准确、鲁棒的自然场景图像文本检测方法,设计了快速有效的剪枝算法,利用最小化正则化变化的策略提取最大稳定极值区域(MSERs)作为特征候选. 文献[8]提出了一种新的文本检测器TextField,用于检测不规则场景文本,通过完全卷积神经网络学习,用二维矢量图像表示该方向场,将学习一个方向字段指向远离最近文本边界的每个文本点. 文献[9]充分利用边界-中心信息,提出了一种新的场景文本检测方法TextMountain,预测了文本的中心边界概率(TCBP)和文本中心方向(TCD). 该文提出的标注规则不会因角度变换而导致歧义问题,对多方向文本具有鲁棒性,并能很好地处理曲线文本. 然而,在甲骨文图像数据集内,由于每一句的甲骨文字相对分散,对拓片中甲骨字图像进行“成行”的识别存在很大的困难. 同时,甲骨图像本身还存在大量纹路和噪声,进一步加剧了图像检测与识别的困难. 因此,现有的甲骨标注中,仅仅对每个甲骨文拓片图像提供了篇幅级别的标注,即该甲骨文拓片图像中有哪几句话、每句话中有哪些文字,但没有提供每幅图像中每句话以及每句话中的每个甲骨字在甲骨文拓片图像中的具体坐标位置,也无法实现字符级别的甲骨图像标注.

近年来,卷积神经网络(Convolutional Neural Network, CNNs)在计算机视觉、自然语言处理及语音识别等领域得到了突飞猛进的发展,其强大的特征学习能力引起了国内外专家学者广泛的关注. 文献[10]介绍了卷积神经网络结构优化技术的发展历史、研究现状以及典型方法,对当前研究的热点与难点作了分析和总结,并对网络结构优化领域未来的发展方向和应用前景进行了展望. 文献[11]设计了一种基于卷积神经网络的高铁轨道周边路牌数字识别的智能系统,能使用目标识别、语义分割等深度学习算法自动定位并识别路牌内的数字,解决了之前人工处理的繁琐和低效率问题. 针对深度学习需要大量人工标注的训练数据耗费大量的人力成本这一最大的弊端,Palatucci等人于2009年提出了零样本学习(Zero-Shot Learning). 文献[12]分类和阐述了零样本学习的多种模型,指明了零样本学习进一步研究中需要解决的问题以及未来可能的发展方向. 随着深度学习算法的飞速发展,使用图像识别算法解决甲骨文领域中尚未解决的问题正逐渐成为未来的研究方向. 文献[13]对甲骨文字形的构形系统进行了研究,提出了一种基于深度学习架构的新型甲骨文字形识别方法,利用甲骨文的构形知识提升准确度,并且通过实验验证了其有效性. 针对甲骨文识别任务中类别样本分布不平衡的问题,文献[14]提出一种基于循环式生成对抗模型的甲骨字符数据扩充算法,增广样本量较少的字符类训练样本,解决长尾效应对甲骨字符识别性能的影响. 在这些方法中,数据起着核心作用,因为深度神经网络的性能在很大程度上取决于训练数据的数量和质量. Facebook[15]和谷歌[16]的出色工作已证明了大规模数据集在获得高质量训练模型上的有效性,并揭示了深度学习十分依赖大而复杂的训练集,在此基础上才能在无约束环境下很好地推广. 这种数据与模型有效性的密切关系在文献[17]中得到了进一步的验证. 数据量不足或数据分布不平衡会导致过拟合和过参数化问题,导致学习结果的有效性明显下降.

目前,在利用深度学习方法对甲骨文识别的研究中,没有公开的数据集可以利用. 为此本文在《甲骨文合集》中挑选了7824张图像进行人工字符级标注,并由甲骨文专家最终校对,从而构建了甲骨字符数据集. 使用此数据集,基于模板匹配的思想,利用孪生神经网络原理,本文设计了甲骨字识别模型(Oracle-Bone Recog-

nition, OBR),根据得到的实验结果看,样本库中出现次数多的甲骨字符图像样本,其识别率才比较高. 因此,训练OBR识别模型的样本数量决定了甲骨字符图像识别模型的准确度. 如果想要提高OBR识别模型的识别精度,就必须增加样本数据集的样本量. 可以先利用OBR模型进行识别,然后对识别结果进行自动标注,再用标注的结果扩充数据集,并重新训练OBR识别模型,进一步提高OBR模型的识别精度. 在此过程中,如何对OBR模型识别的结果进行自动标注是一个关键的问题.

为解决此问题,本文按照"按骥索图"的总体思想,考虑到甲骨文是人手工刻写,主要用于记录事项内容,其书写顺序遵循特定的规律性,总体上是按从上到下、从左至右或是从右至左的顺序进行刻写,为了便于计算机进行有序识别,本文设计了δ-分列算法和切割分列算法对甲骨文拓片中的字符图像进行分列处理,并在此基础上设计一种基于锚点空间近邻关系的甲骨字符图像的自动标注方法,并扩充带标签的样本数据集. 具体地,本文构建了二级、三级、四级推断算法,分别对锚点、近邻字、近邻字的下一级近邻字、近邻字的下二级近邻字进行推断,结合原文对甲骨字符图像进行回溯,利用甲骨拓片图像原文的序列子句确定甲骨字符图像与原文的对应关系,并将标注出的新字符图像识别结果添加到甲骨文样本数据集中,从而实现样本数据集的自动增广.

据知,本研究是甲骨文识别领域研究数据自动标注的首项工作,能够以极低的时间和人力成本,大幅扩充有标注的甲骨字符图像数据集的规模,对基于深度学习的甲骨文识别算法的准确度的提升有重要的帮助. 本文的主要创新贡献如下所述.

(1)本文率先提出甲骨字符图像的δ-同列概念,构造了甲骨字符图像的δ-分列算法,可以把拓片上所有的甲骨字符图像都归于某一列,并证明了这种算法将甲骨字符图像分列的唯一性.

(2)由于δ-分列算法是在垂直方向上对甲骨拓片进行分列的,可能存在将不同含义的几个片段分在同一列的情形. 本文构造了甲骨字符图像的切割分裂算法,弥补了δ-分列算法可能存在的不足,避免了甲骨拓片内容的割裂,完善了甲骨字符图像的分列算法.

(3)首次提出了锚点甲骨字和锚点甲骨字符图像的定义,并构造了锚点甲骨字符图像的识别算法,形成两者的一一对应关系,进一步给出了锚点甲骨字符图像的近邻字符图像的识别算法及改进算法.

(4)利用两个甲骨图像数据集,对本文算法进行了多次实验. 实验结果表明,本文提出的算法在正确推断甲骨字符图像数量、召回率和精度方面具有整体的优势.

2 基于分列的甲骨字符图像的空间近邻搜索方法

甲骨文拓片上的字符图像的排列虽有一定的规律,但与甲骨文原文中的字难以形成一一对应关系,需要找到一种算法,使得计算机能够实现甲骨字符图像的自动标注. 为此,本节首先将甲骨字符图像进行分列处理,为后续甲骨字符图像的识别做一个前期准备.

为了便于理解,本文将甲骨文拓片、甲骨字符图像、甲骨文原文、甲骨字符图像标注信息以及三元组等多种概念及其之间的联系以图形予以表示,如图1所示.

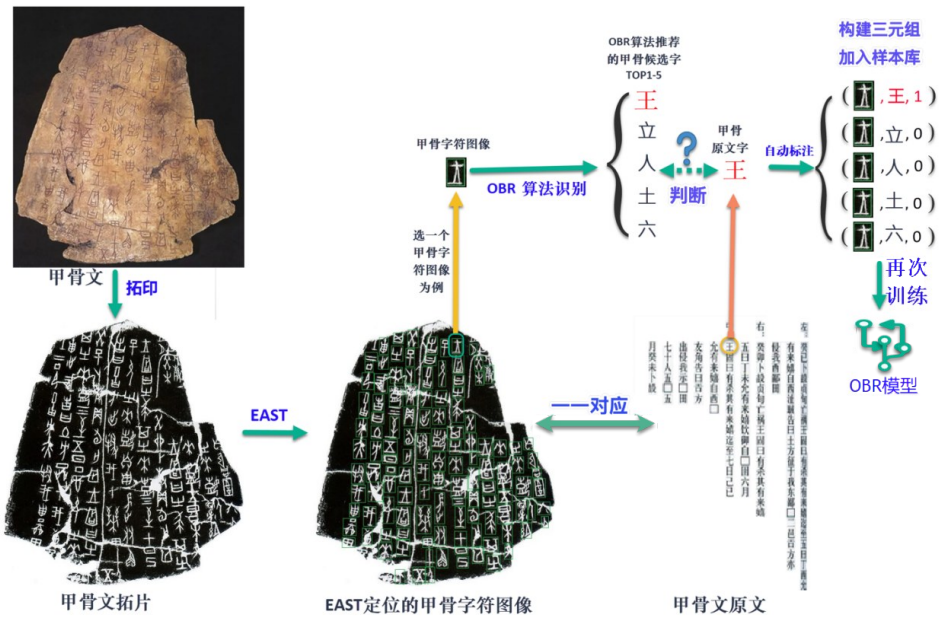


图1 甲骨文拓片、甲骨字符图像、甲骨文原文、甲骨字符图像标注以及三元组关系示意图

符号说明:记$\boldsymbol{Z}$为甲骨文字库中所有字的集合,$Z\in\boldsymbol{Z}$为甲骨文字库中的任意一个字;用$\boldsymbol{T}$表示当前拓片上所有甲骨字符图像的集合,$T\in\boldsymbol{T}$代表甲骨文拓片上的任意一个甲骨字符图像;记$\boldsymbol{Y}$为当前甲骨文拓片对应的甲骨文原文中所有字的集合,$Y\in\boldsymbol{Y}$表示甲骨文原文中的任意一个字;$\boldsymbol{M}$表示锚点甲骨字原文的集合,是$\boldsymbol{Y}$的一个子集;$\tilde{\boldsymbol{M}}$表示待定锚点甲骨字原文的集合,是$\boldsymbol{Y}$的一个子集;$\boldsymbol{T}^M$表示所有锚点甲骨字符图像的集合,T^M表示所有任意一个锚点甲骨字符图像;最后,令$f(Y)$为甲骨文原文中的字Y在现有甲骨文数据库中出现的频率,$P(T,Z)$代表甲骨文拓片上的字符图像T被识别为甲骨文字库中的字Z的概率.

不妨假设根据字符定位EAST算法[18]给出的每个甲骨字符图像T_i的坐标为$(x_{i1},x_{i2},y_{i1},y_{i2})$,其中$x_{i1}<x_{i2},y_{i1}<y_{i2}$,如图2所示. 为了便于论述,给出下面的定义.

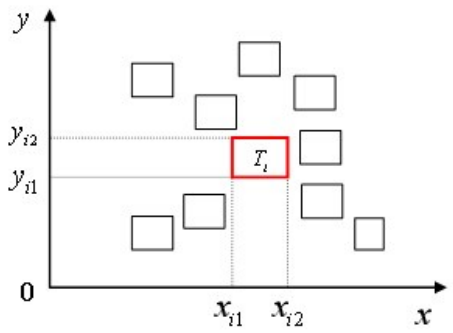


图2　甲骨字符图像的坐标示意图

定义1　给定当前拓片上所有甲骨字符图像的集合$\boldsymbol{T}$和正数$\delta>0$,对任意两个字符图像$T_i,T_j\in\boldsymbol{T}$,若有$\left|\frac{x_{i1}+x_{i2}}{2}-\frac{x_{j1}+x_{j2}}{2}\right|\leqslant\delta$,则称甲骨字符图像$T_i$和$T_j$为$\delta$-同列的,进一步称$T_i$为$T_j$的$\delta$-同列元素.

根据定义1,可以得到下述简单结论.

命题1　对任意正数$\delta>0$,任意$T\in\boldsymbol{T}$,T均为自身的δ-同列元素.

利用δ-同列元素的概念,可以对任意一幅甲骨文拓片进行分列处理,形成甲骨字符图像的分列算法,具体步骤见算法1. 算法1的基本思想是针对当前甲骨文拓片,首先遴选最右侧的一列中最上方的甲骨字符图像作为列首元素,将遴选出来的列首元素T^*放入集合中;然后找出该列首元素的所有δ-同列元素,并从该拓片的甲骨字符图像集合中划掉;接下来从剩余的甲骨字符图像集合中重复上述操作,直到把拓片上所有的甲骨字符图像全部划掉为止.

显然,根据算法1,可以得到如下结论.

算法1　甲骨字符图像的分列算法

输入: 给定甲骨文拓片的所有甲骨字符图像的集合$\boldsymbol{T}$

输出: 该甲骨文拓片的分列字符集合$\boldsymbol{T}^c$

1　**BEGIN**

2　构造集合$\boldsymbol{\Omega}$和$\boldsymbol{T}^c$,使得$\boldsymbol{\Omega}=\boldsymbol{T}$,$\boldsymbol{T}^c=\varnothing$,其中,$\boldsymbol{T}$为给定拓片的所有甲骨字符图像的集合,$\varnothing$为空集,设定列宽参数$\delta>0$;

3　**REPEAT**

4　在拓片上寻找最右列首甲骨字符图像$T^*\in\boldsymbol{\Omega}$,即对任意$T\in\boldsymbol{\Omega}$,$T$的坐标$(x_1,x_2,y_1,y_2)$和$T^*$的坐标$(x_1^*,x_2^*,y_1^*,y_2^*)$之间满足$\frac{x_1+x_2}{2}\leqslant\frac{x_1^*+x_2^*}{2}$,进一步,若$\frac{x_1+x_2}{2}=\frac{x_1^*+x_2^*}{2}$,则需要满足$\frac{y_1+y_2}{2}<\frac{y_1^*+y_2^*}{2}$;

5　$\boldsymbol{T}^c=\boldsymbol{T}^c\cup\{T^*\}$//将找到的元素$T^*$添加到集合$\boldsymbol{T}^c$中;

6　$\boldsymbol{\Omega}_{T^*}=\varnothing$//在甲骨字符图像集合中$\boldsymbol{\Omega}$寻找$T^*$的所有$\delta$-同列元素,组成一个新的集合,记为$\boldsymbol{\Omega}_{T^*}$;

7　**FOR** $T\in\boldsymbol{\Omega}\backslash\{T^*\}$ **DO**

8　令T的坐标(x_1,x_2,y_1,y_2),T^*的坐标$(x_1^*,x_2^*,y_1^*,y_2^*)$;

9　IF $\left|\frac{x_1+x_2}{2}-\frac{x_1^*+x_2^*}{2}\right|\leqslant\delta$ THEN

10　$\boldsymbol{\Omega}_{T^*}=\boldsymbol{\Omega}_{T^*}\cup\{T\}$;

11　**END IF**

12　**END FOR**

13　$\boldsymbol{\Omega}=\boldsymbol{\Omega}\backslash\boldsymbol{\Omega}_{T^*}$//从集合$\boldsymbol{\Omega}$中去除所有属于$\boldsymbol{\Omega}_{T^*}$的元素;

14　**UNTIL**　$\boldsymbol{\Omega}=\varnothing$

15　**END**

引理1　设$\boldsymbol{\Omega}_{T^*}$是由算法1中某次循环所产生的元素$T^*$的$\delta$-同列元素的集合,$\boldsymbol{\Omega}_{T^{**}}$是另外一次循环所产生的元素$T^{**}$的$\delta$-同列元素的集合,则$\boldsymbol{\Omega}_{T^*}\cap\boldsymbol{\Omega}_{T^{**}}=\varnothing$且$\boldsymbol{T}=\bigcup_{T^*\in\boldsymbol{T}^c}\boldsymbol{\Omega}_{T^*}$.

证明　若$\boldsymbol{\Omega}_{T^*}$是由当前循环产生的,则由算法的第13行可知,$\boldsymbol{\Omega}_{T^*}$中的元素都要从$\boldsymbol{\Omega}$中去除掉,而后续的循环中所产生的$\boldsymbol{\Omega}_{T^{**}}$是由$\boldsymbol{\Omega}$中的元素组成的,因此$\boldsymbol{\Omega}_{T^*}\cap\boldsymbol{\Omega}_{T^{**}}=\varnothing$.

根据算法1的第2行可知,集合$\boldsymbol{\Omega}$的初始值为拓片上所有甲骨字符图像的集合$\boldsymbol{T}$,算法每循环一次,生成一个$\boldsymbol{\Omega}_{T^*}$,然后从$\boldsymbol{\Omega}$中除掉,重复操作直到$\boldsymbol{\Omega}$为空集截止. 因此有$\boldsymbol{T}=\bigcup_{T^*\in\boldsymbol{T}^c}\boldsymbol{\Omega}_{T^*}$,其中$\boldsymbol{\Omega}_{T^*}$是每一次循环中遴选出来的列首甲骨字符图像$T^*$的$\delta$-同列元素的集合.

定理1　对于给定的$\delta>0$,算法1给出的分列是唯一的,即对任意的$T\in\boldsymbol{T}$,其属于且仅仅属于某一个列.

证明　首先证明$\boldsymbol{T}^c$的元素互为非δ-同列元素.

这里采用反证法. 假设存在 $T^*, T^{**} \in \boldsymbol{T}^c$ 且 T^* 为 T^{**}的δ-同列元素. 根据算法1第5~10行, $T^{**} \in \boldsymbol{\Omega}_{T^*}$. 由第13行, $T^{**} \notin \boldsymbol{\Omega}$. 进一步根据算法1第4行可知,所有的 $\boldsymbol{T}^c$ 中的元素均来自于集合 $\boldsymbol{\Omega}$,因为 $T^{**} \in \boldsymbol{T}^c$,所以有 $T^{**} \in \boldsymbol{\Omega}$. 这与 $T^{**} \notin \boldsymbol{\Omega}$ 是一个矛盾,故假设不成立,也即是说 $\boldsymbol{T}^c$ 的元素互为非δ-同列元素.

其次证明任意一个甲骨字符图像 T 必定是集合 $\boldsymbol{T}^c$ 的某一个元素的 δ-同列元素,且这种隶属关系是唯一的.

根据引理1可知,每一次循环中所产生的集合 $\boldsymbol{\Omega}_{T^*}$ 的并集恰好为集合 $\boldsymbol{T}$. 由于 $T \in \boldsymbol{T}$,则必然存在某一次循环中所产生的集合 $\boldsymbol{\Omega}_{T^*}$,使得 $T \in \boldsymbol{\Omega}_{T^*}$,根据5~10行可知,存在某个 $T^* \in \boldsymbol{T}^c$,使得 $\boldsymbol{\Omega}_{T^*}$ 为 T^* 的所有 δ-同列元素的集合,因此 T 为 T^* 的 δ-同列元素.

下面证明隶属关系的唯一性. 这里仍然采用反证法. 假设同时存在 $T^*, T^{**} \in \boldsymbol{T}^c$,使得 T 既是 T^* 的 δ-同列元素,也是 T^{**} 的 δ-同列元素,也即 $T \in \boldsymbol{\Omega}_{T^*}$ 且 $T \in \boldsymbol{\Omega}_{T^{**}}$,故 $T \in \boldsymbol{\Omega}_{T^*} \cap \boldsymbol{\Omega}_{T^{**}}$. 这与引理1中 $\boldsymbol{\Omega}_{T^*} \cap \boldsymbol{\Omega}_{T^{**}} = \varnothing$ 相矛盾,故假设不成立,从而说明 T 仅是集合 $\boldsymbol{T}^c$ 的某一个元素的 δ-同列元素.

定理得证.

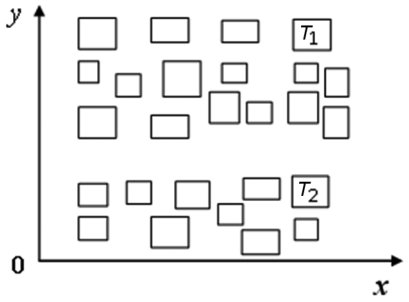
根据算法1,可以把拓片上所有的甲骨字符图像都归于某一列. 但是,在实践中,甲骨文拓片中字符图像的分布可能会出现图3所示的情况,其中,图3(a)为甲骨文拓片上的字符图像分布图,图3(b)为甲骨字符图像分布的示意图.

在图3中,根据人们的常识,拓片的内容应该是被分为上下两个部分的. 但是根据算法1,字符 T_2 很可能被归于 T_1 所在列,这是不符合常理的. 因此,需要对算法1进行矫正,将 T_1 所在列分为两个列. 为了处理类似情形,给出了下述切割分列算法2.

算法2是在算法1分好列的基础上,对每一列中的甲骨字符图像,自上而下计算相邻甲骨字符图像的中心点竖直间距,如果超出设定阈值,则让其分为两列,新增一个列首元素到集合 $\boldsymbol{T}^c$ 中,并更新该列首元素和新增列首元素的δ-同列元素.

注释1 在算法2中,更新列首元素 T^* 和新增列首元素的δ-同列元素可以采用下述方法. 假设存在 $i_1, i_2 \in \{1,2,\cdots,N-1\}$,且 $i_1 < i_2$,满足切割条件,即在本次循环中列首元素新增了 T^{*,i_1+1} 和 T^{*,i_2+1},则令 $\boldsymbol{\Omega}_{T^*} = \{T^{*1}, T^{*2}, \cdots, T^{*,i_1}\}$, $\boldsymbol{\Omega}_{T^{*,i+1}} = \{T^{*,i_1+1}, T^{*,i_1+2}, \cdots, T^{*,i_2}\}$, $\boldsymbol{\Omega}_{T^{*,i+2}} = \{T^{*,i_1+1}, T^{*,i_1+2}, \cdots, T^{*N}\}$. 以此类推.

(a)



(b)

图3 甲骨字符图像分布示意图

算法2 甲骨字符图像的切割分列算法

输入: $\boldsymbol{T}^c$ 和切割阈值 $\alpha > 0$,其中,$\boldsymbol{T}^c$ 为算法1所产生的列首元素的集合

输出: 更新后的 $\boldsymbol{T}^c$

BEGIN
  构造集合 $\boldsymbol{\Omega} = \boldsymbol{T}^c, \boldsymbol{T}^{c^*} = \varnothing$;
  REPEAT
    任取一个列首元素 $T^* \in \boldsymbol{\Omega}$,对其$\delta$-同列元素的集合 $\boldsymbol{\Omega}_{T^*}$ 中所有元素按照纵向中心点从大到小的次序进行排序,记为 $\boldsymbol{\Omega}_{T^*} = \{T^{*1}, T^{*2}, \cdots, T^{*N}\}$, 显然有 $T^{*1} = T^*$, 且 $\frac{y_1^{*1} + y_2^{*1}}{2} > \frac{y_1^{*2} + y_2^{*2}}{2} > \cdots > \frac{y_1^{*N} + y_2^{*N}}{2}$;
    FOR $i = 1, 2, \cdots, N-1$ **DO**
      IF $\frac{y_1^{*i} + y_2^{*i}}{2} - \frac{y_1^{*,i+1} + y_2^{*,i+1}}{2} \geqslant \alpha$ **THEN**
        $\boldsymbol{T}^{c^*} = \boldsymbol{T}^{c^*} \cup \{T^{*,i+1}\}$;
      END IF
    END FOR
    依次更新 $T^* \in \boldsymbol{\Omega}$ 新增列首元素的δ-同列元素的集合 $\boldsymbol{T}^{c^*}$;
    令 $\boldsymbol{\Omega} = \boldsymbol{\Omega} \backslash \{T^*\}$,即从集合 $\boldsymbol{\Omega}$ 中去除元素 T^*;
  UNTIL $\boldsymbol{\Omega} = \varnothing$
  令 $\boldsymbol{T}^c = \boldsymbol{T}^c \cup \boldsymbol{T}^{c^*}$;
END

3　甲骨字符图像的自动标注与数据增广

通过前面的描述,已经明确了对甲骨文拓片上的字符图像进行识别,即找到甲骨字符图像和甲骨文原文中的甲骨字的一一对应关系. 在众多的拓片中,一些原文字出现频率较高,可称之为高频原文字,另一部分出现频率较低,称之为低频原文字. 本文的总体思路是,对于某一张拓片,先将它上面的高频原文字与甲骨字符图像的对应关系找出来,再以高频原文字作为锚点(拓片上与它对应的甲骨字符图像也是锚点),根据甲骨文原文上下文子句,寻找它们近邻的低频原文字与拓片中甲骨字符图像的对应关系.

3.1　锚点字符图像的邻近甲骨字符图像的识别

为了便于阐述,给出如下定义.

定义 2　令$f(Y)$为当前甲骨文拓片原文中的字Y在现有甲骨文数据集中出现的频率,$\boldsymbol{Y}$为当前甲骨文拓片对应的甲骨文原文中所有字的集合,给定数$0\leqslant q\leqslant 1$,对于$Y\in\boldsymbol{Y}$(甲骨文原文字集合),若有$f(Y)\geqslant q$,则称Y为待定锚点甲骨字,记所有待定锚点甲骨字的集合为$\tilde{\boldsymbol{M}}$.

显然,待定锚点甲骨字是原文中的字,即$\tilde{\boldsymbol{M}}\subseteq\boldsymbol{Y}$,且其在甲骨文数据集中出现的频率比较高.

定义 3　对$\forall T\in\boldsymbol{T}$,记其在字库中的前$K$个识别结果为$\{Z_1,Z_2,\cdots,Z_K\}$,且满足$P(T,Z_1)\geqslant P(T,Z_2)\geqslant\cdots\geqslant P(T,Z_K)$. 记集合$\boldsymbol{T}$中所有甲骨字符的前$K$个识别结果的总集合为$\boldsymbol{R}$. 若$\exists Y\in\boldsymbol{Y}$,使得$Y\in\tilde{\boldsymbol{M}}\cap\boldsymbol{R}$,则称$Y$为锚点甲骨字,称相应的甲骨字符图像$T$为锚点甲骨字符图像. 进一步,记所有锚点甲骨字的集合为$\boldsymbol{M}$,所有锚点甲骨字符图像集合为$\boldsymbol{T}^M$.

可见,锚点甲骨字有两个特征:一是本身在样本库中出现频率高,二是与某个甲骨字符图像相似度高. 实际上,上述定义给出了寻找锚点甲骨字和锚点甲骨字符图像对应关系的办法,即先在原文中找到那些出现在甲骨文数据库中频率比较高的字,再从甲骨拓片的字符图像的识别结果中找到相同的,一一对应起来,这样锚点甲骨字符图像就被识别为相应的锚点甲骨字. 识别锚点甲骨字的步骤见算法 3.

上一小节已经对拓片上的甲骨字符图像进行了分列,把每一个甲骨字符图像都分配到某个列首元素T^*的δ-同列集合$\boldsymbol{\Omega}_{T^*}$中. 接下来,将要在锚点甲骨字符图像已经标注完毕和精确分列的基础上,讨论锚点甲骨字符图像的邻近甲骨字符图像的标注问题.

如前所述,原文中锚点甲骨字的集合$\boldsymbol{M}$所对应

算法 3　锚点甲骨字符图像的识别

输入:给定甲骨文拓片上所有甲骨字符图像集合$\boldsymbol{T}$,甲骨文原文字集合$\boldsymbol{Y}$,频率$0<q<1$和正整数K

输出:锚点甲骨字符图像集合$\boldsymbol{T}^M$和锚点甲骨字集合$\boldsymbol{M}$

```
BEGIN
  构造集合 M̃:=∅, R:=∅, T^M:=∅, M:=∅;
  REPEAT
    任取一个元素 Y∈Y,统计其在甲骨文数据集中出现的频率,记为f(Y);
    IF f(Y)≥q THEN
      M̃:=M̃∪{Y};
    END IF
    Y:=Y\Y;
  UNTIL Y=∅
  REPEAT
    任取一个甲骨字符图像 T∈T,利用甲骨识别算法,输出其在字库中前K个识别结果,记为集合 R_K;
      IF Y∈Y,使得Y∈M̃∩R_K THEN
    T^M=T^M∪{T};
    END IF
    R:=R∪R_K;
    T:=T\{T};
  UNTIL T=∅
  M:=M̃∩R;
END
```

的锚点甲骨字符图像的集合为$\boldsymbol{T}^{\boldsymbol{M}}$,也即对于任意$Y^M\in\boldsymbol{M}$,都存在一个元素$T^M\in\boldsymbol{T}^{\boldsymbol{M}}$与其对应. 这时,将甲骨字符图像$T^M$标注为原文中的甲骨字$Y^M$. 通过这种办法,可以将全部锚点甲骨字符图像标注完毕. 接下来,要以当前锚点甲骨字Y^M为着眼点,寻找下一个要标注的甲骨字符图像. 为此,构造了算法 4.

注释 2　若算法第一步中锚点甲骨字T^M位于列首或者列尾,则只有下一个或者上一个甲骨字符图像,不影响算法后续步骤的实施.

通过执行算法 4,可以标注出锚点甲骨字符图像之外其近邻的新的一批甲骨字符图像,这个过程称为二级推断,以二级推断新标注出的甲骨字符图像为着眼点,将其看成锚点甲骨字符图像,再次执行算法 4,又可以标注出新的甲骨字符图像,称为三级推断. 接着执行四级推断,就可以分别对锚点、近邻字、近邻字的下一级近邻字、近邻字的再下一级近邻字进行自动标注. 该过程的示意图如图 4 所示.

图 4 中的图标A代表锚点字(一级)的候选字,图标B代表锚点字二级近邻字的候选字,图标C代表锚点字

算法4 锚点甲骨字符图像的邻近字符图像的识别

输入：锚点甲骨字集合 $\boldsymbol{M}$，锚点字符图像 T^M

输出：新增标注数据集

```
BEGIN
  REPEAT
    任取一个锚点甲骨字 Y^M ∈ M，获取其对应的锚点字符图像 T^M 的坐标 (x_1^M, x_2^M, y_1^M, y_2^M)，并记 Y^M 的上一个和下一个甲骨字分别为 Y_1 和 Y_2
    设定正整数 P，在锚点字符图像 T^M 所在列内寻找使得 |(y_1+y_2)/2 − (y_1^M+y_2^M)/2| 最小的前 P 个甲骨字符图像 T_1, T_2, ⋯, T_P；
    FOR ∀T_i ∈ {T_1, T_2, ⋯, T_P} DO
      利用甲骨文识别算法对其进行识别，输出前 K 个识别结果，记为 R_K；
      IF Y_1 ∈ R_K（或 Y_2 ∈ R_K）THEN
        将字符图像 T_i 标注为 Y_1（或 Y_2）；
      END IF
    END FOR
  UNTIL 直到所有的锚点甲骨字符图像的邻近字符图像被标注完毕；
END
```

三级近邻字的候选字，图标 D 代表锚点字相应的四级近邻字的候选字. 推断过程：首先，对给定甲骨文拓片上的每一个甲骨字符图像用EAST算法定位，用训练好的OBR孪生网络进行识别，可以得到每个甲骨字符图像 i 个候选字；把与原文对应的甲骨字符图像在样本库出现较多次数的甲骨字符图像选作为锚点（其甲骨原文候选字分别表示为 $A_1, A_2, \cdots, A_i$），根据近邻字算法对其近邻的 j 个甲骨字符图像（其甲骨原文候选字分别表示为 $B_{11}, B_{12}, \cdots, B_{1i}, \cdots, B_{j1}, B_{j2}, \cdots, B_{ji}$）找出二级推断所对应的甲骨字符图像；分别以这 j 个甲骨字符图像为新的锚点，根据近邻字算法将每个新的锚点其 k 个近邻甲骨字符图像（其甲骨原文候选字分别表示为 $C_{111}, C_{112}, \cdots, C_{11i}, \cdots, C_{jk1}, C_{jk2}, \cdots, C_{jki}$）找出三级推断所对应的甲骨字符图像；分别以这 $j \times k$ 个甲骨字符图像为新的锚点，再根据近邻字算法对每个新的锚点其 l 个近邻甲骨字符图像（其甲骨原文候选字分别表示为 $D_{1111}, D_{1112}, \cdots, D_{111i}, \cdots, D_{jkl1}, D_{jkl2}, \cdots, D_{jkli}$）找出四级推断所对应的 $j \times k \times l$ 个甲骨字符图像及其对应的 $j \times k \times l \times i$ 个甲骨字符图像候选字.

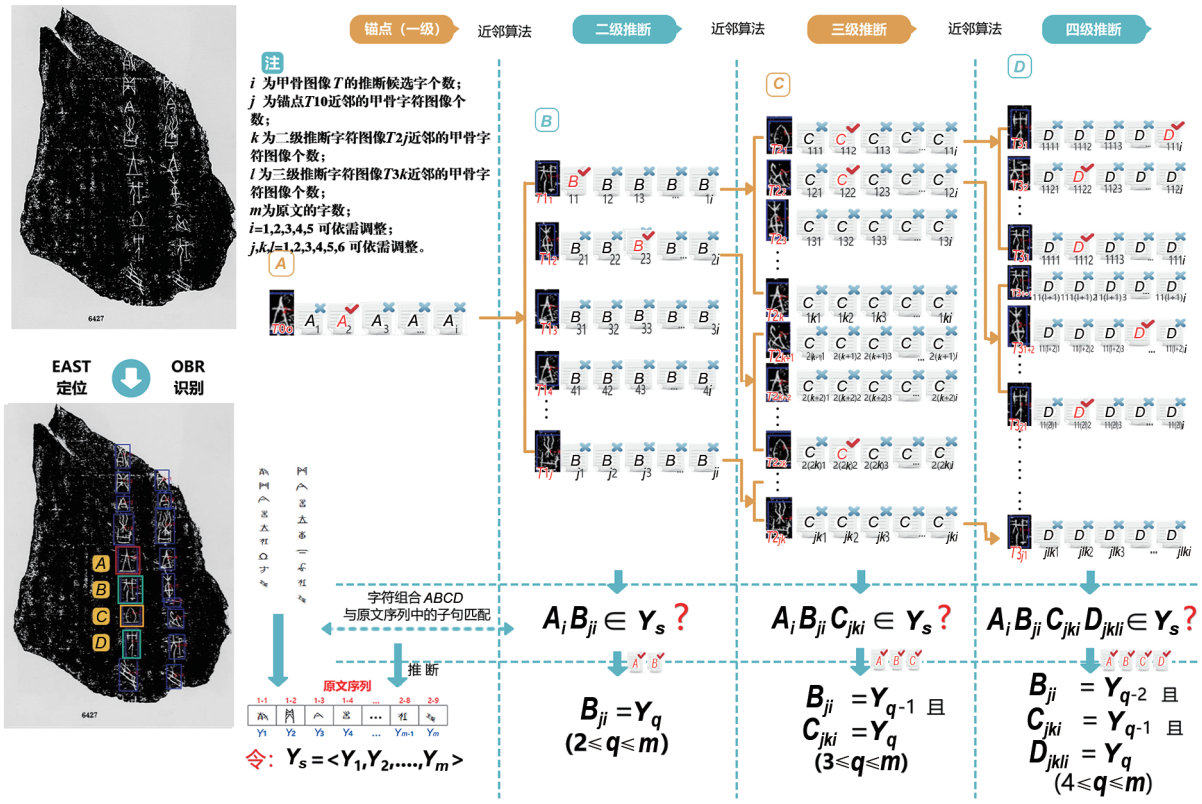


图4 利用原文序列子句推断甲骨字符图像候选字示意图

推定甲骨字符图像对应原文字过程如下. 首先将甲骨拓片对应的原文按从左至右、从上至下的顺序，将其 m 个甲骨拓片原文字进行序列化，得到甲骨拓片原文序列化的集合 $\boldsymbol{Y}_s$，这是分级推断的基础. 进行二级推

断时,如果存在子句 $A_iB_{ji}\in \boldsymbol{Y}_s$,就可以推定 $A_i=Y_{q-1}$,$B_{ji}=Y_q$(其中 $2\leqslant q\leqslant m$),一级甲骨字符图像(锚点本身)识别为 A_i,其对应的甲骨原文 Y_{q-1}(其中 $2\leqslant q\leqslant m$);锚点近邻的二级甲骨字符图像是锚点字符图像的下一个甲骨字符图像,且被识别为 B_{ji},其对应的甲骨原文为 Y_q,B_{ji} 就是 A_i 的下一个甲骨字符图像. 同理,如果存在 $B_{ji}A_i\in \boldsymbol{Y}_s$,就可以推定 $A_i=Y_{q-1}$,$B_{ji}=Y_q$(其中 $2\leqslant q\leqslant m$),B_{ji} 就是 A_i 的上一个甲骨字符图像. 在进行三级推断时,如果存在 $A_iB_{ji}C_{jik}\in \boldsymbol{Y}_s$,就可以推定 $A_i=Y_{q-2}$,$B_{ji}=Y_{q-1}$,$C_{jik}=Y_q$(其中 $3\leqslant q\leqslant m$),B_{ji} 就是 A_i 的下一个甲骨字符图像,C_{jik} 就是 B_{ji} 的下一个甲骨字符图像;如果存在 $C_{jik}B_{ji}A_i\in \boldsymbol{Y}_s$,就可以推定 $A_i=Y_q$,$B_{ji}=Y_{q-1}$,$C_{jik}=Y_{q-2}$(其中 $3\leqslant q\leqslant m$),B_{ji} 就是 A_i 的上一个甲骨字符图像,C_{jik} 就是 B_{ji} 的上一个甲骨字符图像. 同理,可以进行四级推断,得到 A,B,C,D 的顺序和分别对应的原文字.

3.2　甲骨字符图像识别的流程和优化

在前述甲骨图像自动标注的过程中,关键步骤之一是锚点甲骨字符图像的邻近字符图像的识别、推断与相互印证. 即已知锚点三元组 $\left(T^M,Y^M,1\right)$,且锚点甲骨字 Y^M 的上一个和下一个甲骨字分别为 Y_1 和 Y_2. 在锚点甲骨字符图像在 Y^M 所在列中,选择距其最近的 P 个甲骨字符图像 $T_1,T_2,\cdots,T_P$,利用甲骨文识别算法对这 P 个甲骨字符图像逐个进行识别,分别输出前 K 个识别结果,如果这 $P\times K$ 个识别结果中包含 Y_1(或 Y_2),且其对应的字符图像为 T_i,则将字符图像 T_i 标注为 Y_1(或 Y_2),并将 $\left(T_i,Y_1,1\right)$ 或 $\left(T_i,Y_2,1\right)$ 添加到样本数据集中. 通过使用上述算法,可以实现对整个甲骨文拓片的自动标注. 但是该过程没有考虑到如下问题:①拓片中可能存在重复的字符图像或者字形相似度很高的字符图像,导致算法可能会将字符图像 T_m 和 T_n 同时标注为 Y_1,即同时出现 $\left(T_i,Y_1,1\right)$ 和 $\left(T_j,Y_1,1\right)$ 的情形,这显然是不合理的;②虽然原文中甲骨字 Y^M、Y_1(或 Y_2)在位置上是严格的上下关系,但是识别结果 T_i 与锚点字符图像 T^M 并不一定在位置上是严格的上下关系. 因此,仅仅按照相似度高低进行识别,可能会导致识别错误.

为了避免上述问题,本小节提出了一种改进的甲骨字符图像识别流程和优化方法. 核心思想:利用两个或多个甲骨字符图像的空间近邻位置关系,将甲骨字符图像识别算法(OBR)对相应的甲骨字符图像的识别结果与甲骨学者提供的语句级(篇幅级)的甲骨原文序列子句文本中的甲骨字进行对应、推断和相互印证.

具体过程:当 T_i 和 T_j 都被识别为 Y_1 时,分别以 T_i 和 T_j 为新的锚点,再次使用近邻字识别算法,若 T_i 的某个近邻字可识别为 Y_1',则将 T_i 识别为 Y_1,若 T_j 的某个近邻字可识别为 Y_1',则将 T_j 识别为 Y_1,并把在此过程中所形成的三元组添加到样本库. 改进算法的核心思想描述如下.

(1)初始化,给出锚点甲骨字符图像 T^M 和其对应的锚点甲骨字 Y^M,也即锚点三元组 $\left(T^M,Y^M,1\right)$,记锚点甲骨字 Y^M 的上一个和下一个甲骨字分别为 Y_1 和 Y_2(若锚点甲骨字位于列首的话,仅有下方甲骨字 Y_2;若锚点甲骨字位于列尾的话,仅有上方甲骨字 Y_1),为了叙述方便,后续仅考虑锚点甲骨字的上方甲骨字. 进一步,记原文中甲骨字 Y_1 上方的甲骨字为 Y_1'.

(2)在锚点字符图像 T^M 所在列中选择距其最近的 P 个甲骨字符图像 $T_1,T_2,\cdots,T_P$,利用甲骨文识别算法对这 P 个甲骨字符图像逐个进行识别,各自输出前 K 个识别结果,共计给出 $P\times K$ 个识别结果,如表 1 所示,其中,$Y_j^{T_i}(i=1,2,\cdots,P;j=1,2,\cdots,K)$,是识别出的候选甲骨字,它代表甲骨字符图像 T_i 的第 j 个识别结果.

表 1　锚点字符图像 T^M 邻近字符图像的识别结果

T_i	j			
	1	2	…	K
T_1	$Y_1^{T_1}$	$Y_2^{T_1}$	…	$Y_K^{T_1}$
T_2	$Y_1^{T_2}$	$Y_2^{T_2}$	…	$Y_K^{T_2}$
⋮	⋮	⋮	⋱	⋮
T_P	$Y_1^{T_P}$	$Y_2^{T_P}$	…	$Y_K^{T_P}$

(3)判断 Y_1 是否出现在表 1 中,如果仅出现在第 m 行,则将字符图像 T_m 标注为 Y_1,生成三元组 $\left(T_m,Y_1,1\right)$ 和 $\left(T_i,Y_1,0\right),i=1,\cdots,P,i\neq m$,并将这些三元组添加到语料库中. 如果 Y_1 同时出现在表 1 的第 m 行和第 n 行,则生成三元组 $\left(T_i,Y_1,0\right),i=1,\cdots,P,i\neq m,i\neq n$,将其加入语料库,并进入下一步.

(4)分别以甲骨字符图像 T_m 和 T_n 为参考点,在其所在列中选择距其最近的 P 个甲骨字符图像 $T_1^m,T_2^m,\cdots,T_P^m$ 和 $T_1^n,T_2^n,\cdots,T_P^n$,利用甲骨文识别算法对这些甲骨字符图像逐个进行识别,各自输出前 K 个识别结果,如表 2 和表 3 所示. 其中 $Y_j^{T_i^m}(i=1,2,\cdots,P;j=1,2,\cdots,K)$ 是识别出的候选甲骨字,它代表甲骨字符图像 T_i^m 的第 j 个识别结果;$Y_j^{T_i^n}(i=1,2,\cdots,P;j=1,2,\cdots,K)$ 是识别出的候选甲骨字,它代表甲骨字符图像 T_i^n 的第 j 个识别结果.

(5)判断 Y_1' 出现在表 2 中还是表 3 中,如果出现在表 2 中,且出现在第 l 行,则说明将甲骨字符图像 T_m 识别为 Y_1 是合理的,将甲骨字符图像 T_n 识别为 Y_1 是不合理的. 这时,生成三元组 $\left(T_m,Y_1,1\right)$,$\left(T_n,Y_1,0\right)$,$\left(T_l^m,Y_1',1\right)$,$\left(T_i^m,Y_1',0\right),i=1,2,\cdots,P,i\neq l$,以及 $\left(T_i^n,Y_1',0\right),i=1,2,\cdots,P$,并将其加入样本数据集中.

表2 字符图像 T^m 邻近字符图像的识别结果

T_m	j			
	1	2	…	K
T_1^m	$Y_1^{T_1^m}$	$Y_2^{T_1^m}$	…	$Y_K^{T_1^m}$
T_2^m	$Y_1^{T_2^m}$	$Y_2^{T_2^m}$	…	$Y_K^{T_2^m}$
⋮	⋮	⋮	⋱	⋮
T_P^m	$Y_1^{T_P^m}$	$Y_2^{T_P^m}$	…	$Y_K^{T_P^m}$

表3 字符图像 T^n 邻近字符图像的识别结果

T_n	j			
	1	2	…	K
T_1^n	$Y_1^{T_1^n}$	$Y_2^{T_1^n}$	…	$Y_K^{T_1^n}$
T_2^n	$Y_1^{T_2^n}$	$Y_2^{T_2^n}$	…	$Y_K^{T_2^n}$
⋮	⋮	⋮	⋱	⋮
T_P^n	$Y_1^{T_P^n}$	$Y_2^{T_P^n}$	…	$Y_K^{T_P^n}$

本部分在流程上优化了甲骨字符图像识别/推断的流程，优化后的甲骨字符推断和识别流程，除了推断、识别锚点甲骨字的邻近甲骨字，还以上一层推断、识别出的邻近甲骨字为类似新锚点的参考点，进一步推断、识别这些类似新锚点甲骨字符图像的邻近甲骨字符图像. 需要说明的是，本文算法并没有改进甲骨字符图像识别算法(OBR)本身，而是从流程上优化了甲骨字符图像识别的流程. 其核心思想可归纳为，利用两个或多个甲骨字符图像的空间近邻位置关系，将甲骨字符图像识别算法(OBR)对相应的甲骨字符图像的识别结果，与甲骨学者提供的语句级(篇幅级)的甲骨原文序列子句文本中的甲骨字，进行对应、推断和相互印证，使得甲骨字符图像的推断和识别更加精准.

在下面的实验部分，基于字符级标注的甲骨文图像集2(由甲骨学者手工标注)，将本文算法预测(推断)出来的每个甲骨字符图像对应的甲骨字与真实标注的甲骨字(Ground Truth)进行逐一比对，以准确评估本节中的算法的性能/精度. 下一节中汇总了相关的实验结果，其中包括不同级别(深度)的推断对应的实验结果对比分析.

4 仿真实验与分析

在本文的实验中，用到了两个甲骨图像数据集，称为图像集1和图像集2. 其中，图像集1包含7824张图像. 每个字符图像都是按字标注，但是文字之间各自独立，没有保存字符的位置和句序关系. 图像集2包含955张图像，每张图片都按照句序关系进行了标注，知道每句话中的每个字在该句子中的顺序关系，含字、字序编号及坐标位置. 并且这两个图像集之间没有重复的图像，互相独立. 本文利用图像集1训练得到了字形级别的文字识别模型，称为OBR文字识别模型. 利用该模型，以图像集2中的995张图像为输入，运行本文设计的算法，设计了一个实验，验证召回率和识别精度是否达到预期要求. 以甲骨合集中39329号拓片为例，推导过程如图5所示. 注：为了便于读者理解算法的思想，本图中使用汉字替代甲骨字(原文中的甲骨字除外).

在图5所示的39329号拓片多级的推导过程中，首先对甲骨文拓片上的每一个甲骨字符图像用EAST算法定位，用训练好的OBR孪生网络进行识别，可以得到这10个甲骨字符图像分别各自对应的5个候选字. 根据样本库中原文对应的甲骨字符图像在样本库出现次数的多少，将甲骨字符图像 T_0 选作为锚点(其甲骨原文候选字分别表示为 $A_1, A_2, \cdots, A_i$)，根据近邻字算法将其近邻的 T_{12} 等 j 个甲骨字符图像(其甲骨原文候选字分别表示为 $(B_{11}, B_{12}, \cdots, B_{1i}; \cdots; B_{j1}, B_{j2}, \cdots, B_{ji})$ 找出二级推断所对应的甲骨字符图像；再分别以这 j 个甲骨字符图像为新的锚点，根据近邻字算法将每个新的锚点其甲骨字符图像 T_{21} 等 k 个近邻甲骨字符图像(其甲骨原文候选字分别表示为 $C_{111}, C_{112}, \cdots, C_{11i}, \cdots, C_{jk1}, C_{jk2}, \cdots, C_{jki}$)找出三级推断所对应的甲骨字符图像；再分别以这 $j \times k$ 个甲骨字符图像为新的锚点，再根据近邻字算法将每个新的锚点其甲骨字符图像 T_{32} 等 l 个近邻甲骨字符图像(其甲骨原文候选字分别表示为 $D_{1111}, D_{1112}, \cdots, D_{111i}; \cdots; D_{jkl1}, D_{jkl2}, \cdots, D_{jkli}$)，找出四级推断所对应的甲骨字符图像.

图5推定甲骨字符图像对应原文字对应过程如下. 首先将甲骨拓片对应的原文字按从左至右、从上至下的顺序，将其10个原文字进行序列化，得到甲骨拓片原文字序列化的集合 $\boldsymbol{Y}_s$，这是二级推断、三级推断、四级推断的基础. 如图中所示：进行二级推断时，存在 $A_1 B_{13} \in Y_6 Y_5$，就可以推定 $A_1 = Y_6, B_{13} = Y_5, B_{13}$ 就是 A_1 的上一个甲骨字符图像；存在 $B_{21} A_1 \in Y_6 Y_7$，就可以推定 $A_1 = Y_6, B_{21} = Y_7, B_{21}$ 就是 A_1 的下一个甲骨字符图像. 进行三级推断时，以甲骨字符图像 T_{21} 为例，存在 $A_1 B_{21} C_{112} \in Y_6 Y_7 Y_8$，就可以推定 $A_1 = Y_6, B_{21} = Y_7, C_{112} = Y_8, C_{112}$ 就是 B_{21} 的下一个甲骨字符图像，同理可推定 B_{12} 的上一个甲骨字符图像. 同理，可以进行四级推断，得到甲骨字符图像 T_{32} 是甲骨字符图像 T_{21} 的下一个甲骨字符图像，其对应的原文为 Y_9. 依次完成二级、三级、四级的推断，得到甲骨拓片上图像的顺序与原文中的一一对应关系.

在执行算法4的过程中，确定锚点甲骨字符图像之后，寻找下一个待识别甲骨字的过程，其在甲骨合集第5、10106拓片上寻找的情况如图6所示，其中图6(a)是在甲骨合集第5拓片上寻找的情况，图6(b)是在甲骨合

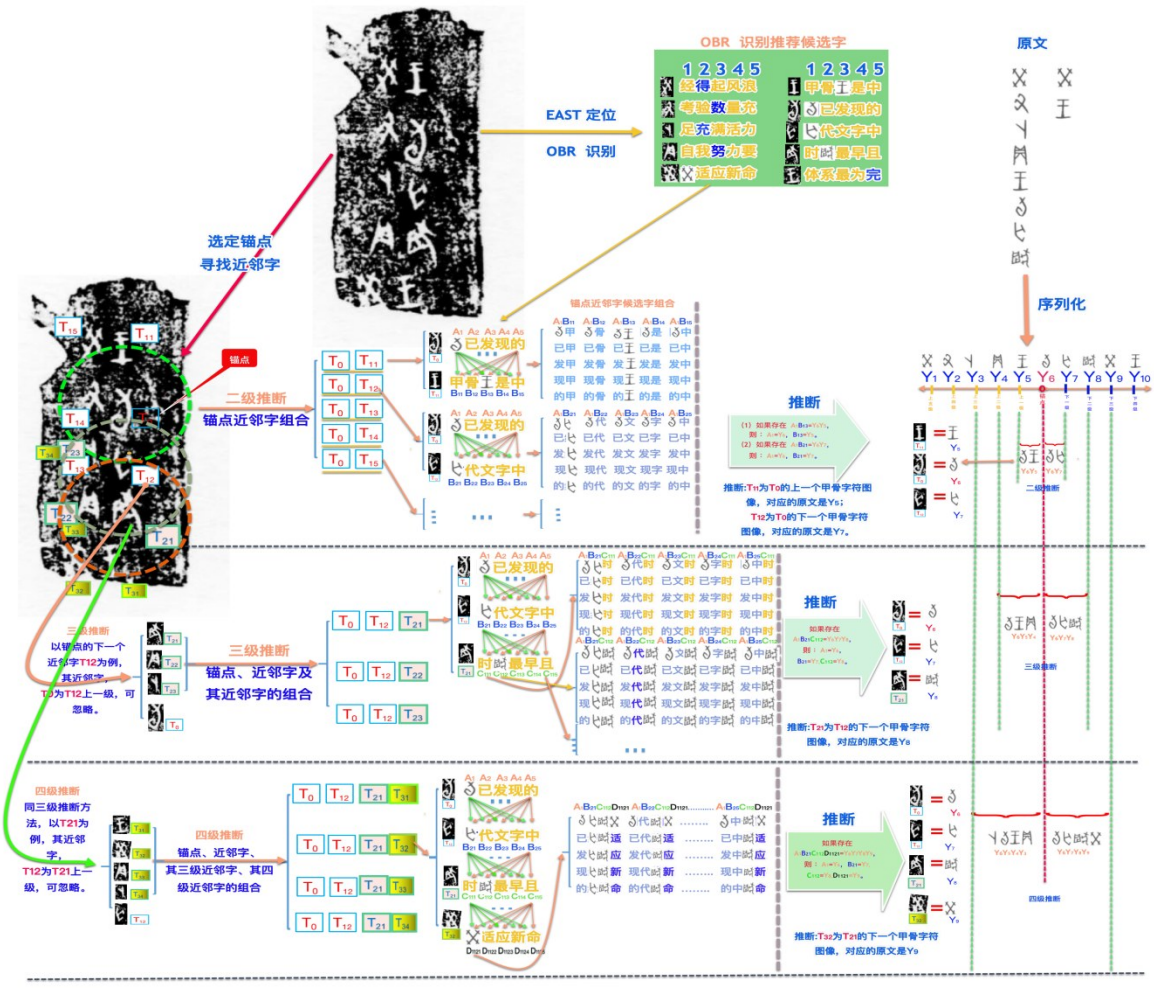
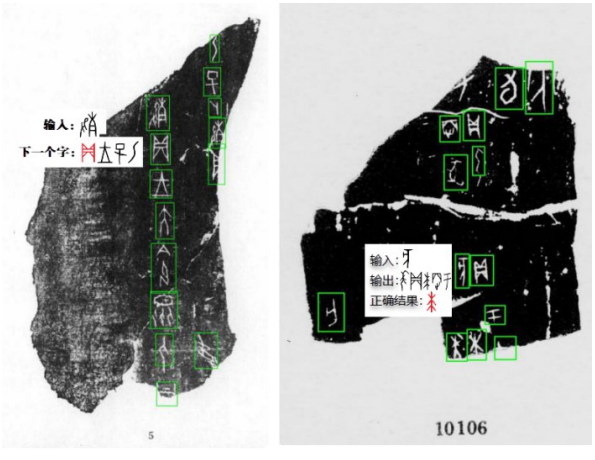


图5　39329号拓片多级推导过程示意

集第10106拓片上寻找的情况.

利用本文提出的甲骨字符图像自动标注算法,对图像集2上的995张图像执行甲骨字符图像的四级推断过程,结果如表4所示.



(a)　　(b)

图6　寻找下一个待识别甲骨字符图像的示意图

表4　算法4在图像集2上的推断结果

推断级数			二级	三级	四级
P2	有推断结果的甲骨拓片数量		882	658	452
P3	P2拓片上的甲骨字符总数		9984	8312	6298
P4	可标注的甲骨字符图像个数		6313	4846	3668
P5	正确标注的甲骨字符图像个数		4931	3828	2894
P6	P2占图像总数的比例		0.88643	0.66131	0.45427
P7	召回率	Recall1	0.49389	0.46053	0.45951
		Recall2	0.53157	0.53725	0.56511
P8	精度	Precision1	0.78109	0.78993	0.78898
		Precision2	0.85045	0.85678	0.85101

在表4中,召回率分为总召回率Recall1和平均召回率Recall2. 其中,总召回率Recall1代表推断正确的字符与有效甲骨文拓片上所有甲骨字符的比值,即Recall1=P5/P3,以二级推断为例,Recall1=4931/9984=0.49389;平均召回率Recall2是由每个有效甲骨文拓片上的Recall1值取平均值得来的. 在表4中,精度也分为

总精度和平均精度两个指标. 其中,总精度Precision1代表推断正确的字符与有效甲骨文拓片上可标注甲骨字符的比值,即Precision1=P5/P4,以二级推断为例,Precision1=4931/6313=0. 78109;平均精度Precision2是由每幅有效甲骨文拓片上的Precision1值取平均值得来的.

在实验的执行过程中,还存在下面的情况:两头的甲骨字符图像已经标注出来,中间的不确定是否匹配,但是利用标注中对应句序中的字,可以直接推断中间缺失或不确定的图像对应的字. 本实验针对三缺一(3m1)和四缺一(4m1)的情形,分别给出了召回率和识别精度的计算结果. 其中,3m1代表三个字符中,两头的已经匹配,中间的不确定是否匹配,如a*c;4m1代表四个字符中,两头的已经匹配,中间的两个中只有一个不确定是否匹配,如ab*d,a*cd. 这种情形下的实验结果如表5所示.

表5 中间缺失或不确定的图像的推断结果

推断级数			3m1	4m1
P2	有推断结果的甲骨拓片数量		840	709
P3	P2拓片上的甲骨字符总数		9736	8892
P4	可标注的甲骨字符图像个数		16599	14229
P5	正确标注的甲骨字符图像个数		6119	5523
P6	P2占图像总数的比例		0. 84422	0. 71256
P7	召回率	Recall1	0. 62849	0. 62112
		Recall2	0. 66683	0. 68566
P8	精度	Precision1	0. 36864	0. 38815
		Precision2	0. 54865	0. 55955

从上述实验结果来看,在二级推断和三级推断中,本文提出的算法在正确推断的甲骨字符图像数量、召回率和精度方面具有整体的优势. 由于推断精度影响到了人工分拣和确认的甲骨文拓片图像的数量,为了减少人工的劳动和时间成本,在高召回率的同时,精度应该尽可能地高,从这个意义上来看,本文算法取得了良好的效果. 从表5来看,对于中间缺失或不确定的甲骨文拓片图像的推断,其优点是,正确推断的甲骨字符数量最多,能够在一定程度上增加之前出现频度较少(小样本类)的样本数量,但是返回的不正确的甲骨字符图像也最多,导致召回率较高(最高),但是精度最低,只有不到40%的精度. 这给后期的甲骨学专家确认增加了大量的工作负荷. 例如3m1的方法,正确找到的甲骨字符图像的数量最多,但是返回的不相关、不正确的甲骨字符图像的推断结果较多,导致精度只有36. 864%,需要大量的人工比对、确认和筛检的时间. 时间效率方面,4m1方法耗时最长. 而二级推断方法所需的时间最少,仅需要12s. 综上所述,二级推断的"按骥索图"方法,整体上具有最好的时间效率和较好的召回率、精度,正确找到的甲骨字符图像数量最多,是整体上较优的选择.

5 结论和展望

如何让计算机自动识别任一甲骨文拓片上任一甲骨字符图像所对应的甲骨原文字,也即甲骨字符图像的自动标注问题,是本文要解决的主要问题. 本文在按字标注的甲骨图像及训练得到的OBR(甲骨文自动识别算法)文字识别模型的基础上,提出了δ-分列算法和切割分列算法,将甲骨文拓片中的所有甲骨字符图像进行合理分列处理,并利用锚点空间近邻关系给出甲骨字符图像的自动标注算法. 该算法不仅能够实现新的甲骨拓片图像中部分文字的自动标注,而且可以为基于深度学习的OBR算法增加更多的训练样本(增加6~10倍),这对进一步提高甲骨文识别准确率具有重要意义. 同时,以较小的成本大幅增加样本数量,可以节约专家大量的时间和人力,有利于甲骨文相关应用的落地,可为古文字的考释和语言理解提供重要支撑.

未来,将在以下方面开展研究:①通过提出的自动标注样本数据增广算法,利用篇幅(拓片)级的甲骨文原文与拓片甲骨字符图像间的对应关系,对新拓片进行自动标注,经甲骨学专家人工审核后,完成甲骨数据集的增广工作,以用于训练和测试甲骨识别模型,进而提升识别的准确率和识别精度;②由于小样本甲骨字在甲骨拓片中出现的次数少,但是占的比例却相当大,下一步将考虑通过基于对抗生成网络的方法进行小样本的研究,提高模型生成图像的丰富程度,扩充与增广甲骨文数据集;③目前基于图像增强学习方法的研究与应用在各个领域快速发展,下一步,拟将图像增强学习算法用于甲骨识别模型研究,提升性能和效率,使甲骨文的研究手段和识别效果得到进一步提升.

致谢 本课题的研究得到河南省黄河实验室(河南大学)的支持

参考文献

[1] 江铭虎, 邓北星, 廖盼盼, 等. 甲骨文字库与智能知识库的建立[J]. 计算机工程与应用, 2004, 40(4): 45 – 47, 60.
Jiang M H, Deng B X, Liao P P, et al. Construction on word-base of oracle-bone inscriptions and its intelligent repository[J]. Computer Engineering and Applications, 2004, 40(4): 45 – 47, 60. (in Chinese)

[2] 顾绍通. 甲骨文数字化处理研究述评[J]. 西华大学学报(自然科学版), 2010, 29(5): 38 – 42, 48.
Gu S T. Review on digitization processing of jiaguwen[J]. Journal of Xihua University (Natural Science Edition), 2010, 29(5): 38–42, 48. (in Chinese)

[3] 顾绍通. 基于分形几何的甲骨文字形识别方法[J]. 中文信息学报, 2018, 32(10): 138 – 142.

Gu S T. Identification of oracle-bone script fonts based on fractal geometry[J]. Journal of Chinese Information Processing, 2018, 32(10): 138 – 142. (in Chinese)

[4] 李锋, 周新伦. 甲骨文自动识别的图论方法[J]. 电子科学学刊, 1996, 18(S1): 41 – 47.
Li F, Zhou X L. Recohnition of Jia Gu Wen based on graph theory[J]. Journal of Electronics, 1996, 18(S1): 41 – 47. (in Chinese)

[5] 焦清局, 刘永革, 仇利萍, 等. 网络驱动的未识甲骨字特性及场景语义预测[J]. 浙江大学学报(理学版), 2020, 47(2): 142 – 150.
Jiao Q J, Liu Y G, Qiu L P, et al. Network-driven prediction of unknown oracle character's features and scene semantics[J]. Journal of Zhejiang University (Science Edition), 2020, 47(2): 142 – 150. (in Chinese)

[6] Long S B, He X, Yao C. Scene text detection and recognition: The deep learning era[J]. International Journal of Computer Vision, 2021, 129(1): 161 – 184.

[7] Yin X C, Yin X W, Huang K Z, et al. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970 – 983.

[8] Xu Y C, Wang Y K, Zhou W, et al. TextField: learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5566 – 5579.

[9] Zhu Y X, Du J. TextMountain: Accurate scene text detection via instance segmentation[J]. Pattern Recognition, 2021, 110: 107336.

[10] 林景栋, 吴欣怡, 柴毅, 等. 卷积神经网络结构优化综述[J]. 自动化学报, 2020, 46(1): 24 – 37.
Lin J D, Wu X Y, Chai Y, et al. Structure optimization of convolutional neural networks: A survey[J]. Acta Automatica Sinica, 2020, 46(1): 24 – 37. (in Chinese)

[11] 孟琭, 孙霄宇, 赵滨, 等. 基于卷积神经网络的铁轨路牌识别方法[J]. 自动化学报, 2020, 46(3): 518 – 530.
Meng L, Sun X Y, Zhao B, et al. An identification method of high-speed railway sign based on convolutional neural network[J]. Acta Automatica Sinica, 2020, 46(3): 518 – 530. (in Chinese)

[12] 张鲁宁, 左信, 刘建伟. 零样本学习研究进展[J]. 自动化学报, 2020, 46(1): 1 – 23.
Zhang L N, Zuo X, Liu J W. Research and development on zero-shot learning[J]. Acta Automatica Sinica, 2020, 46(1): 1 – 23. (in Chinese)

[13] 鲁绪正, 蔡恒进, 林莉. 基于Capsule网络的甲骨文构件识别方法[J]. 智能系统学报, 2020, 15(2): 243 – 254.
Lu X Z, Cai H J, Lin L. Recognition of Oracle Radical based on the Capsule network[J]. CAAI Transactions on Intelligent Systems, 2020, 15(2): 243 – 254. (in Chinese)

[14] Huang S P, Zhong Z Y, Jin L W, et al. DropRegion training of inception font network for high-performance Chinese font recognition[J]. Pattern Recognition, 2018, 77: 395 – 411.

[15] Taigman Y, Yang M, Ranzato M, et al. DeepFace: Closing the gap to human-level performance in face verification[A]. 2014 IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus, OH, USA: IEEE, 2014. 1701 – 1708.

[16] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering[A]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Boston, MA, USA: IEEE, 2015. 815 – 823.

[17] Hestness J, Narang S R, Ardalani N, et al. Deep learning scaling is predictable, empirically[EB/OL]. https://www. researchgate. net/publication/321487863_Deep_Learning_Scaling_is_Predictable_Empirically, 2017.

[18] Zhou X Y, Yao C, Wen H, et al. EAST: An efficient and accurate scene text detector[A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. Honolulu, HI, USA: IEEE, 2017. 2642 – 2651.

作者简介

史先进　男,1973年12月生,河南商水人.现为河南大学博士研究生.高级工程师.主要研究领域为计算甲骨学、教育大数据分析.
E-mail:shixj@henu.edu.cn

曹　爽　女,1993年2月生,河南商丘人.2021年硕士毕业于河南大学.主要研究领域为生成对抗网络、不均衡学习、计算甲骨学.
E-mail:scao@henu.edu.cn

张重生(通信作者)　男,1982年9月生,河南南阳人.现为河南大学教授、博士生导师.主要研究领域为大数据分析、深度学习.
E-mail:chongsheng.zhang@yahoo.com