自然场景下藏文检测识别数据集与方法

侯琴[1]，胡永祥[1]，刘思宇[1]，尼玛扎西[2]，程建[2,1*]

【摘要】自然场景下藏文文本检测与识别是一项重要且有挑战性的任务。为解决藏文自然图像数据不足的问题，本文构建了一个乌金体藏文检测识别数据集，收集藏区多个城市中包含藏文的1320张自然图像，共标注图像中1979条文本的位置与内容。同时，本文提出了自然场景下藏文检测与识别方法，使用可微二值化网络实现藏文文本检测，提出基于时序卷积网络的藏文识别框架。为进一步扩充数据，本文提出自然场景藏文图像合成方法，在图像的合适位置渲染藏文文本，生成真实自然的藏文图像。本文在数据集上进行了藏文检测与识别实验，文本检测的F1分数达到86.93%，文本识别的字符识别准确率达到92.70%，实验表明本文提出的方法有较好的性能。

关 键 词 藏文数据集； 藏文检测； 图像合成； 藏文识别

中图分类号 TP391.1 **文献标识码** A

自然场景下的文本检测与识别是计算机视觉领域的一项关键任务，该任务负责将自然图像中的文字提取至计算机内，实现文本资料的数字化。文本检测与识别广泛应用于人机交互[1]、机器翻译[2]、资料检索[3]、自动驾驶[4]等领域。作为图像翻译的前置任务，文本检测与识别的精度直接影响后续机器翻译的性能。然而，自然图像中的文本形态多变，存在遮挡、光照不均、背景复杂、多语种混合等特性，导致检测与识别的准确率较低。所以，设计泛化性能好的自然图像文本检测与识别方法，是一项具有挑战性的任务。

藏文是一种历史悠久、应用广泛的文字，对藏文的智能检测与识别具有较高的研究价值与应用价值。然而相较于中文、英文的检测与识别，针对藏文的研究较少、效果较差，这主要是因为含有藏文的自然图像数量远小于中英文图像数量，导致模型训练不充分，性能较差。

为提升自然场景下藏文检测与识别效果，本文从数据集与方法两方面开展研究工作。在数据集方面，我们构建了一个自然场景下的乌金体藏文数据集，共1320张图像，包含1979条藏文文本、10552个藏文音节、22318个藏字。图像来源为网络地图街景、旅行软件、搜索引擎等**，主要内容为路牌、横幅、店铺招牌、石刻等。我们对图像中藏文文本的位置和内容进行了标注。

在藏文检测方法上，本文使用可微二值化网络(Differentiable Binarization Net, DBNet)[5]作为检测的模型。为进一步扩充训练数据，本文提出了一种图像生成方法，生成真实自然的藏文图像，缓解数据不足对模型训练的影响。在藏文识别方法上，本文提出了基于时序卷积网络的藏文识别框架，使用时序卷积网络(Temporal Convolutional Network, TCN)[6]建模序列关系来提升识别性能。

图1 数据集图像示例

Fig .1 An instance in our dataset

1 相关研究基础

随着深度学习技术的发展，基于深度学习的文本检测方法已成为该领域的主流方法，主要分为两类：一类方法是基于目标检测的，将文本检测视为文本框回归问题，如CTPN[7]、TextBoxes[8]、EAST[9]等；另一类则是基于分割的方法，通过分割预测文本的最佳边界，如PSENet[10]、DBNet[5]等。文本识别框架主要有两种：一种是基于循环卷积神经网络(Convolutional Recurrent Neural Network, CRNN)[11]的方法；另一种是基于注意力的方法，例如RARE[12]等。

藏文文本的识别工作开展较早。早期的研究主要针对背景简单、图像清晰的扫描文档。王维兰等人[13]通过提取不同方向的投影特征与字典进行匹配，实现了藏文基本字符的识别；王华等人[14]利用方向线素特征提取方法与基于置信度分析的两级分

基金项目：国家自然科学基金委员会与中国民用航空局联合资助项目（U2133211），国家自然科学基金面上项目（62071104）

*通信作者：程建，chengjian@uestc.edu.cn；

**图像的具体来源有：百度地图、高德地图、小红书、大众点评、百度图片搜索、中国西藏网

类策略实现了藏文字符的识别；梁弼等人[15]提出一种基于隐马尔可夫模型分类器的手写藏文识别方法。然而，以上方法对图像质量要求高，难以直接应用于自然场景下的藏文识别。

近年来自然场景下藏文检测与识别的工作主要基于深度学习方法。仁青东主等人[16]利用卷积循环网络与连接时域分类相结合的模型，实现了自然场景下的藏文识别。洪松等人[17]利用DBNet与CRNN实现了对自然场景下藏文的检测与识别。然而，受自然场景下藏文图像数据量的限制，目前基于深度学习的藏文文本检测与识别的方法较少，构建大规模的藏文数据集、提出高效的检测与识别方法是提升该任务性能的关键。

2 自然场景藏文数据集构建

目前，公开的自然场景文本图像数据集主要有MSRA-TD500[18]、RCTW-17[19]和COCO-Text[20]等针对中英文和数字的数据集，而针对我国少数民族文字的文本数据集较少，且没有公开的自然场景藏文数据集。因此，我们采集包含藏文文本的自然场景图像，并人工进行检测与识别标注，构建自然场景藏文数据集。

2.1 数据采集

考虑到实地采集的困难性，我们从网络地图街景、旅行软件、搜索引擎等收集了自然场景下的藏文文本图像共1320张，包含藏文文本1979条，藏文音节10552个，藏字22318个。这些图像来自西藏自治区的拉萨、日喀则、那曲、山南，四川省的甘孜藏族自治州、阿坝藏族羌族自治州等多个地区，包含路牌、店铺招牌、石刻、横幅等。这些图像在拍摄角度、光照条件、遮挡情况上差异较大，部分图像如图2所示。

图2：采集到的自然场景图像

Fig. 2 Tibetan Images in the wild

目前已知的自然场景下的藏文数据集有仁青东主等人[16]以及洪松等人[17]的研究中使用的数据集，本文数据集与其对比如下表1所示。对比之下，我们的数据集来源更加丰富，数据总量较大。

表1 现有自然场景藏文数据集对比

Tab.1 Comparison of existing Tibetan datasets in the wild

数据集	来源地	数量
洪松等人[17]	拉萨	449张图像 4321条藏文文本
仁青东主等人[16]	未说明	600张图像*
本文数据集	拉萨、日喀则、那曲、山南、甘孜、阿坝等	1320张图像 1979条藏文文本

*：仅测试集

2.2 图像标注

对于每一张自然场景图像，我们仅对藏文文本进行标注。标注的过程如图3所示。首先进行检测标注，在藏文文本周围绘制不规则四边形边界框，记录四个顶点的坐标，每个框内仅包含一条文本。然后对边界框内的文本进行识别标注。

(a)待标注图像　(b)检测标注　(c)识别标注

图3：图像标注流程

Fig. 3 Annotation process of images

针对部分文本被遮挡、不清晰或多语种同时出现的情况，本文设定如下标注规定：（a）只对藏文以及包含在藏文中的数字进行标注；（b）忽略藏文中存在的空格；（c）对于被遮挡或者不清晰的字符，用“#”进行标注，如果一条文本中出现超过两个“#”，则忽略整条文本。我们总共对1320张图像中的1987条文本进行了标注，其中有效数据量为1979条。

2.3 数据集划分

自然场景藏文数据集包含1320张图像，1979条藏文文本。对于文本检测任务，训练集包含1056张图像，测试集包含264张图像；对于文本识别任务，训练集包含1800条文本，测试集包含179条文本。

2.4 数据统计

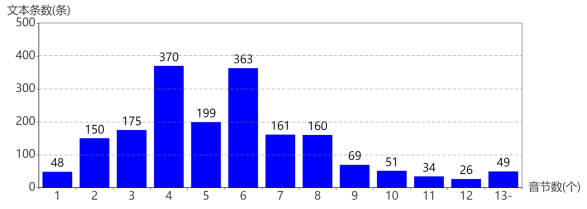
本文构建的自然场景藏文图像数据集包含1320张图像，1979条有效藏文文本，10552个藏文音节和22318个藏字。如图4所示，图像中的藏文文本特性各异，包含平面文本、凸起文本、倾斜文本、复杂背景文本、光照不均、部分遮挡、多语种混合等情

况。一张图像最多包含13条藏文文本，平均一张图像包含1.5条文本。

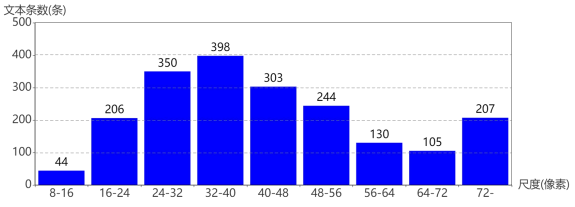
图4：多种藏文文本示例

Fig. 4 Examples of various Tibetan texts

每条文本的长度不同。以藏文中的“·”作为音节分割符，每条文本最短只包含1个音节，最长包含29个音节，平均包含5.7个音节，文本长度数据如图5(a)所示。本数据集涵盖不同尺度的藏文文本，以文本边框的短边长度来表示文本尺度，最小为11像素，最大为342像素，平均为44.5像素，大多集中在16-56像素范围内，统计结果如图5(b)所示。图6展示了数据集中最常出现的20个藏字及其频数。



(a)藏文文本长度统计



(b)藏文文本尺度统计

图5：数据集中藏文文本长度、尺度统计

Fig. 5 The length and scale statistic in our dataset

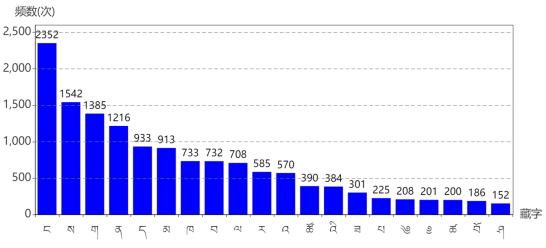


图 6：最常出现的 20 个藏字

Fig. 6 Top 20 high frequency Tibetan characters in dataset

根据以上统计数据，本数据集在文本类别、尺度和属性上具有良好的多样性和复杂性，涵盖了大多数自然场景情况，是一个具有挑战性的自然场景下藏文检测和识别数据集。

3 自然场景下藏文检测与识别方法

本文提出的自然场景下藏文的检测与识别方法由藏文文本检测和藏文文本识别两个阶段构成。在检测阶段，本文提出了一种有效的自然场景藏文图像合成方法，使用可微二值化网络作为检测模型；在识别阶段，本文提出了一种基于时序卷积网络的文本识别框架，使用时序卷积网络建模序列关系，获得了更优的识别性能。总体框架如图7所示。

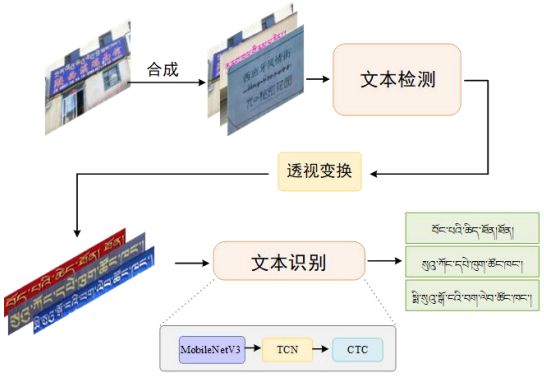


图7：藏文检测与识别总体框架

Fig. 7 General framework of Tibetan detection and recognition

3.1 藏文文本检测

文本检测的目标是确定文本在图像中出现的位置。一般情况下，基于深度学习的文本检测需要较大的数据量对模型进行训练。为解决更高的数据需求，本文在第2节所述数据集的基础上，通过图像合成方法进一步扩充模型训练数据。

在图像合成方法中，较为直接的一种方法是：首先建立藏文语料库和背景图像库，然后从藏文语料库中提取文本，通过泊松融合等方式将文本随机渲染在背景图上，该方法的合成效果如图8所示。由于该方法的文本出现位置存在随机性，因此文本出现的位置并不完全符合实际情况，例如出现在天上、地面上，这不利于模型学习文本位置的先验知识。

图8：直接合成的图像

Fig. 8 Images generated by direct method

为解决以上问题，本文提出了一种有效的合成

自然场景藏文图像的方法。为了使藏文文本渲染至合适的位置，本文首先选取自然场景中文数据集RCTW[19]的图像，随机抹除其中的文本，然后从藏文文本库中随机选择文本，根据原文本的边界框信息对藏文文本进行透视变换，然后渲染到图像的相应位置上。该合成方法流程如图9所示。

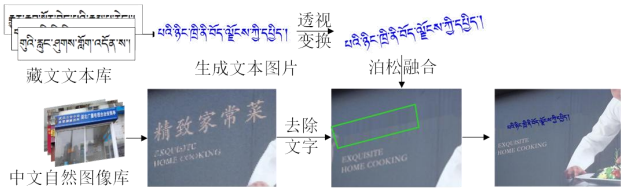


图9：本文提出的图像合成方法

Fig. 9 Image synthesis method we proposed

使用该方法生成的图像如图10所示。此方法会将藏文渲染到路牌、招牌等文字常出现的位置上，合成真实自然的图像，有利于模型学习文本位置先验信息。本文采用该方法生成了3854张合成图像作为训练集的补充。

图10：本文方法生成的图像

Fig. 10 Images generated by our method

本文使用可微二值化网络[5]作为文本检测的模型。在特征提取部分，使用MobileNetV3[21]网络提取多尺度特征图，上采样之后拼接特征，预测概率图和阈值图像，然后通过可微分的二值化算法生成近似的二值图像，最后经后处理得到文本框坐标。检测网络架构如图11所示。

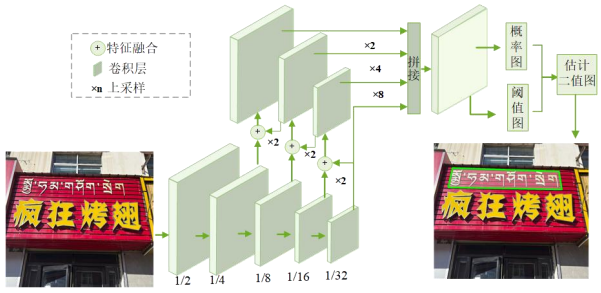


图11：检测网络架构

Fig. 11 Framework of detection network

3.2 藏文文本识别

实际采集到的1320张图像中只包含1979条可用的藏文文本，数据量较小。因此，本文首先利用藏文文本库和背景库额外生成了20000条藏文文本图像，以扩充训练集。受循环卷积网络[11]的启发，本文设计的识别模型由特征提取，序列建模和转录三个部分组成。首先按照文本检测模型的输出剪裁出包含藏文文本的图像块，通过透视变换得到正视角的文本图像，再使用卷积神经网络提取图像特征。本文使用MobileNetV3[21]提取特征，用时序卷积网络(TCN)[6]建模序列关系，完成序列预测。TCN引入了因果卷积与膨胀卷积结构，因果卷积确保了时序性，膨胀卷积增大了卷积核的感受野。最后将表征序列关系的输出特征输入转录层，得到藏文识别结果。本文提出的识别网络架构如图12所示。

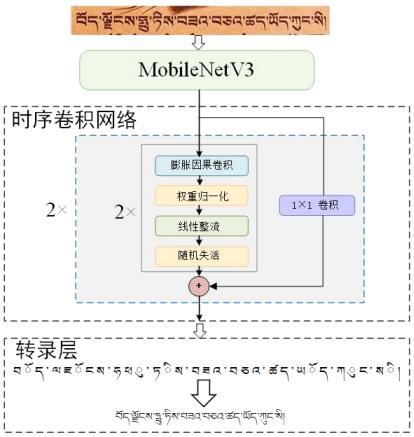


图12：藏文识别框架

Fig. 12 Framework of Tibetan recognition

4 实验

4.1 实验设置

实验设置方面，本文使用深度学习框架PyTorch 1.8实现算法，在4张GeForce RTX 2080Ti上完成实验。优化器使用Adam，批大小设置为64，初始学习率设置为0.001，权重衰减系数为0.00002。文本检测任务中，含有藏文的自然图像尺寸统一至640×640，训练800个轮次。文本识别任务中，藏文文本块的尺寸统一至32×320，训练800个轮次，在第700个轮次后将学习率降低至0.0001。

训练集与测试集按2.3节所述规定划分。3854张人工合成的图像与20000条人工合成的文本图像均划入训练集。对于检测任务，训练集共包含4910张图像，测试集共包含264张图像；对于识别任务，训练集共包含21800条文本，测试集共包含179条文本。

4.2 评价指标

在检测任务上，设置交并比阈值大小为0.5。即：对于真实框 G_i 和预测框 D_j，若满足两者的交并比 $\frac{A(G_i \cap D_j)}{A(G_i \cup D_j)} > 0.5$，则认为 G_i 与 D_j 匹配成功，记为真正例；若 G_i 或 D_j 始终匹配不成功，分别记为假反例与假正例。最后根据匹配数据计算精度(Precision)、召回率(Recall)以及F1分数作为文本检测的评价指标，具体计算公式如下：

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = \frac{2*P*R}{P+R}$$

其中 P, R, F_1 分别代表精度、召回率、F1分数，TP, FP, FN代表真正例、假正例、假反例数量。

在识别任务上，本文选用字符串识别准确率(String Recognition Accuracy, SRA)和字符识别准确率(Character Recognition Accuracy, CRA)作为评价指标。SRA即整条文本预测结果全部正确的文本比例，CRA通过计算预测文本与真实结果的编辑距离衡量二者的相似度，公式为：

$$CRA = 1 - \frac{D}{num}$$

其中，D表示预测字符串与真实结果的编辑距离，由动态规划方法计算得出，num表示文本字符数量。

4.3实验结果

在检测任务上，本文测试了数据集合成方法的效果。在真实图像、直接合成法、SynthText[22]与本文的方法合成的数据集上进行文本检测实验，结果如表2所示。仅采用真实图像进行训练，测试结果为：精度87.00%，召回率82.07%，F1分数84.46%，与真实图像数据集的实验结果相比，使用直接合成法的数据集的F1分数上升了0.45%，但精度下降了2.78%，说明采用直接合成法扩充数据在一定程度上是有效的，但也会影响模型学习先验知识。相较于真实图像数据集，使用本文提出的方法生成的数据集在基本保持精度的情况下，召回率与F1分数分别提升了4.96%与2.47%，且三项指标均高于直接合成法与SynthText。实验结果证明本文提出的自然场景藏文图像合成方法是有效的，该方法在一定程度上能弥补训练数据不足的问题，实现检测性能的提升。

表2 藏文检测实验结果

Tab.2 Tibetan detection results

数据集	训练集图像数	检测指标@mIoU=0.5(%)		
		精度	召回率	F1分数
真实图像	1056	**87.00**	82.07	84.46
直接合成法	4910	84.22	85.61	84.91
SynthText[22]	4910	80.93	86.08	83.43
本文的方法	4910	86.82	**87.03**	**86.93**

在识别任务上，相比于仅采用原始数据，本文方法在训练集扩充至21800条后取得了较大的性能提升，SRA提升了43.57%，CRA提升了13.20%。同样使用扩充后的数据集进行训练，本文的藏文识别方法的结果高于CRNN，两项指标分别提升了3.36%和0.65%。与场景文本识别的前沿算法SVTR[23]相比，本方法的SRA高于SVTR-L，CRA略低于SVTR-L。实验结果表明，本文提出的基于时序卷积网络的识别框架在藏文识别任务上是有效的。

表3 藏文识别实验结果

Tab.3 Tibetan recognition results

识别模型	训练文本条数	SRA%	CRA%
本文的方法	1800	13.41	78.85
CRNN[11]	21800	56.98	92.05
SVTR-L[23]	21800	57.54	**93.42**
本文的方法	21800	**60.34**	92.70

5 结 论

本文构建了自然场景下的藏文检测与识别数据集，包含1320张自然图像、1979条藏文文本、10552个藏文音节，22318个藏字，并标注了文本边界框与文本内容。同时，本文提出了藏文文本检测与识别方法。在检测任务中，提出了合理的图像合成方法以扩充数据集；在识别任务中，引入时序卷积网络建模序列关系。实验表明以上两种方法可提升检测与识别的性能。

在未来的工作中，我们将从更多地区收集藏文图像，增强数据集的多样性，对同一藏文文本拍摄多张图像以获得更大的数据量；另一方面，我们将探索性能更好、泛化能力更强的检测与识别方法；此外，在本文的基础上引入机器翻译模块，设计端到端的检测-识别-翻译模型；本文提出的数据扩充方法有望应用于其他数据量不足的少数民族文字上，

协助更多种文字的智能分析与理解。

参 考 文 献

[1] MISHRA A, SHEKHAR S, SINGH A K, et al. Ocr-vqa: Visual question answering by reading text in images[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 947-952.

[2] 王洪君,孙健琳,于光玉,刘珂,王小飞. 一种翻译图片中文字的方法[P]. 中国：CN105761201B,2019-03-22.

[3] WANG H, BAI X, YANG M, et al. Scene Text Retrieval via Joint Text Detection and Similarity Learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4558-4567.

[4] 倪翠竹. 基于视频的交通标志文字检测与识别算法研究[D]. 北京: 北京交通大学,2015:1-55

[5] LIAO M, WAN Z, YAO C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11474-11481.

[6] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. ArXiv preprint: 1803.01271, 2018.

[7] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C] //European Conference on Computer Vision. Springer, Cham, 2016: 56-72.

[8] LIAO M, SHI B, BAI X, et al. Textboxes: A fast text detector with a single deep neural network[C]//Thirty-first AAAI Conference on Artificial Intelligence. 2017:4161-4167.

[9] ZHOU X, YAO C, WEN H, et al. East: an efficient and accurate scene text detector[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5551-5560.

[10] WANG W, XIE E, LI X, et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9336-9345.

[11] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(11): 2298-2304.

[12] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4168-4176.

[13] 王维兰.藏文基本字符识别算法研究[J].西北民族学院学报(自然科学版),1999(03):20-23+51.

[14] 王华,丁晓青.多字体印刷藏文字符识别[J].中文信息学报,2003(06):47-52

[15] 梁硕,王维兰,钱建军.基于HMM的分类器在联机手写藏文识别中的应用[J].微电子学与计算机, 2009, 26(04): 98-101.

[16] 仁青东主,尼玛扎西.基于深度学习的自然场景藏文识别研究[J].高原科学研究,2019,3(04):96-103.

[17] 洪松,高定国,三排才让,取次.自然场景下乌金体藏文的检测与识别[J].计算机系统应用,2021,30(12):332-338.

[18] YAO C, BAI X, LIU W, et al. Detecting texts of arbitrary orientations in natural images[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 1083-1090.

[19] SHI B, YAO C, LIAO M, et al. Icdar2017 competition on reading chinese text in the wild (rctw-17)[C]//2017 14th iapr International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 1429-1434.

[20] VEIT A, MATERA T, NEUMANN L, et al. Coco-text: Dataset and benchmark for text detection and recognition in natural images[J]. ArXiv preprint: 1601.07140, 2016.

[21] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1314-1324.

[22]Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2315-2324.

[23]Du Y, Chen Z, Jia C, et al. SVTR: Scene Text Recognition with a Single Visual Model[J]. arXiv preprint: 2205.00159, 2022.

Tibetan Detection and Recognition Dataset and Methods in the Wild

HOU Qin[1], HU Yongxiang[1], LIU Siyu[1], Nyima-Tashi[2] and CHENG Jian[2,1*]

(1. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731;

2. School of Information Science and Technology, Tibet University, Lhasa, Tibet, 850000)

Abstract: Tibetan text detection and recognition in the wild is an important and challenging task. To solve the problem of insufficient Tibetan images in the wild, this paper constructs a large-scale Wujin style Tibetan text detection and recognition dataset. It contains 1320 images collected from several cities in Tibetan areas, and annotates the positions and contents of 1979 texts. In addition, this paper proposes a Tibetan text detection and recognition method in the wild. It uses a differentiable binarization network for detection and proposes a Tibetan text recognition framework based on Temporal Convolutional Network. To further expand the dataset, this paper proposes a Tibetan image synthesis method. It renders Tibetan text at the appropriate position of the image to generate more realistic Tibetan images. The experiments showed that our method had better performance. The F1-score for detection was 86.93% and the character recognition accuracy was 92.70%.

Key words: Tibetan dataset; Tibetan text detection; Image synthesis; Tibetan text recognition