文章编号：2095-2163(2020)08-0157-05 中图分类号：TP391.4 文献标志码：A

基于光流预测的文本校正算法的研究

张文强，张亚博，左旺孟
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：场景文本识别旨在将自然图像中所包含的文本信息识别为计算机可处理的字符序列，其挑战性在于如何处理不规则分布形状的场景文本。目前的主流方法是将其解耦为文本校正与序列识别两个子任务，文本校正模块负责将不规则文本行特征扭曲为标准化的水平形式，然后送入后续的序列识别模块。由于缺乏必要的标注信息，目前大部分文本校正方法依赖于弱监督方式训练的空间变换网络，并且需要微妙的参数初始化策略和端到端的优化方法才能收敛。本文注意到场景文本通常满足一定的几何先验约束，提出一种在该约束下学习的光流网络，其生成的光流场可以用于文本校正，并在若干真实场景文本识别数据集上进行了相关实验。实验结果表明，基于本文方法的文本识别系统比传统基于 STN 网络的系统的准确率有所提升，这可以归因于本文所提出的基于光流变换的文本校正算法的有效性。

关键词：场景文本校正；场景文本识别；光流预测；几何先验约束

A text rectificationmethod based on optical flow field prediction

ZHANG Wenqiang, ZHANG Yabo, ZUO Wangmeng
(School of Computer Science and Technology, Harbin Institute ofTechnology, Harbin 150001, China)

【Abstract】 Scene Text Recognition (STR) aims to recognize the text information from natural images into computer-processable character sequences, whose challenge comes from handling the irregular-shaped scene text. To this end, current works usually decouple it into two subtasks: text rectification and sequence recognition, where the former warps the irregular text line features into canonical form which fed into the subsequent sequence recognition module. Due to the lack of annotation, most text rectification methods heavily rely on the Spatial Transformer Network (STN) and the weakly-supervised training, which require delicate parameter initialization and end-to-end optimization to converge. We noticed that scene text usually satisfy certain geometric prior constraints and propose an optical flow network, which can generate the optical flow field used for text rectification. The extensive experiments have been conducted on several real scene text recognition datasets, and the results indicate that our text recognition system has an improved accuracy than the traditional STN-based system, which should be attributed to the effectiveness of our text rectification method based on optical flow prediction.

【Key words】 Scene Text Recognition; optical flow prediction; geometric prior constraints

0 引 言

文字作为标志着人类文明诞生的重大发明，承载着丰富而具体的高级语义信息。如何从自然场景图像中自动提取文本信息，即场景文本识别[1]（Scene Text Recognition, STR），可以被广泛的应用到视觉任务中，如：图像搜索[2]、机器人导航[3]和工业自动化[4]等，因此近年来 STR 相关研究在计算机视觉领域中愈发重要。相比于传统的针对扫描文档的光学字符识别任务（Optical Character Recognition, OCR），STR 的主要处理对象为场景文本，具有文本风格和尺度多变、分布形状不规则、背景复杂、成像条件不良等多种特性，因此仍然是一个非常具有挑战性的课题。

自然场景中的文本行形状通常满足一定的几何关系约束，如字符关于文本中心线轴对称、字符法线方向趋于一致等，如果能够充分利用这些几何先验约束，文本校正的质量将会得以提升。因此，本文提出了一种基于光流预测的文本校正模型，利用预测得到的密集光流网格（dense grid）和双线性插值进行文本校正，其变换规则不再局限于仿射变换或 TPS[5]变换等特定形式，具有较强的灵活性，且在几何先验关系的约束下，可以生成更为合理的光流变换网格。

由于大部分的场景文本数据集并没有提供字符级别的标注信息，无法计算得到对应的几何属性。因此本文利用 SynthText[6]引擎，生成仿真数据集，并用于模型训练。随后，在真实数据集上对模型进行微调，利用预训练好的文本行几何属性预测（Text

作者简介：张文强（1996-），男，硕士研究生，主要研究方向：计算机视觉、文本检测与识别；张亚博（2000-），男，本科生，主要研究方向：图像处理、计算机视觉；左旺孟（1977-），男，博士，教授，博士生导师，主要研究方向：计算机视觉、机器学习与生物特征识别。

收稿日期：2020-06-02

Line Geometry, TLG)模块解析得到文本行的几何属性,再以此约束另一路密集光流网格的生成。为了加快模型的收敛速度,除了上述几何约束外,还引入了关键点回归损失和光流场平滑正则化项。为了衡量本文所提出的文本校正模块的效果,将其与一个基于 GRU[7] 网络的通用序列识别模块相结合,形成了一个完整的文本识别系统,并在 IIIT5K[8]、ICDAR15[9] 等真实数据集上进行测试。实验结果表明,基于本文所提出的文本校正模块,系统的识别准确率在以弯曲文本行为主的数据集上有较大的提升。

1 相关工作

自从 2012 年 AlexNet[10] 在 ILSVRC 竞赛中崭露头角之后,深度学习技术在计算机视觉领域的研究中迅猛发展,STR 模型也经历了从手工设计特征的传统模式向数据驱动的深度学习模式的技术变迁。传统的文本识别技术主要采用统计机器学习方法,如支持向量机(Support Vector Machine,SVM)、随机森林(Random Forest,RF)等模型,它们对于 OCR 任务而言已然足够,但对于 STR 任务来说效果仍然不佳。进入深度学习时代以来,针对场景文本识别的各种困难因素,各种各样精心设计的卷积神经网络被提出。例如,针对弯曲文本行的识别问题,Shi 等人提出的 ASTER[11] 模型将其解耦为文本校正与文本识别两个阶段,其中校正模块采用基于 TPS 变换的 STN[12] 网络,首先将弯曲文本行校正为水平形式,然而再利用后续的文本识别模块进行序列识别,模型的整体架构如图 1 所示。

图 1 基于 STN 网络的文本校正模块

Fig. 1 The text rectification module based on STN network

为了描述不规则文本行的弯曲形态,TextSnake[13] 模型提出了一种基于重叠圆盘(disk)的文本表征形式。其中每个圆盘的圆心位于文本行的中心线,且具有可变化的半径和分布方向,可以很好地拟合任意分布形状的文本行。每个圆盘的几何属性通过全卷积网络进行语义分割得到,不同的几何属性(如中心线、半径、分布夹角等)分别对应不同的语义通道。ScRN[14] 模型在此基础上提出一种新的文本行几何属性的形式化定义方法,本文模型的 TLG 模块即采用了该表示方法。如图 2 所示,将一行文本视作一个有序字符序列 $A = \{A_1, \cdots, A_i, \cdots, A_m\}$,其中 m 为字符总数,将每个字符 A_i 所对应的边界框记作 B_i。首先,构造中心点序列 $C = \{c_{head}, c_1, \cdots, c_i, \cdots, c_m, c_{tail}\}$,包含每一个字符边界框 B_i 的中心点 c_i,最左字符边界框 B_1 的左中心点 c_{head} 和最右字符边界框 B_m 的右中心点 c_{tail},文本行中心线(Text Center Line,TCL)通过按序连接这些中心点获得。每一个中心点 c_i 与一组局部的几何属性相关联,即 $geo_i = (c_i; s_i; \varphi_i; \theta_i)$,其中 s_i 描述字符尺度信息,取 B_i 边界框高度的一半,φ_i 是字符法线方向与水平方向的夹角,θ_i 是当前中心点 c_i 与下一中心点 c_{i+1} 连线所构成的文本行分布方向 $\vec{c_i c_{i+1}}$ 与水平方向的夹角。对于在 TCL 中但不在 C 中的点,其几何属性值由离它最近的两个中心点的属性进行线性插值得到。

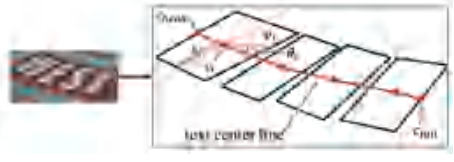

图 2 文本行几何属性定义示意图

Fig. 2 Schematic diagram of the text line geometry definition

本文提出的 TLG 模块可以预测上述定义的文本行几何属性,通过 FPN[15] 网络进行语义分割得到 TCL、scale、$\sin\theta$、$\cos\theta$、$\sin\varphi$、$\cos\varphi$ 共 6 个通道。由于 TCL 的宽度仅为单个像素,很容易造成断连,因此将 TCL 区域沿字符法线方向进行了 20%尺度的扩展,扩展区域与采样源点具有相同的几何属性。由于分割得到的中心线 TCL 为连续分布形式,为了满足后续关键点回归损失的需要,本文采用 TextSnake 模型中提出的 striding 后处理算法对其离散化操作,得到类似于序列 C 的关键点集合。

2 模型详述

2.1 几何约束损失

弯曲文本行的单个字符在理想的校正情况下具有局部线性化的特点,即字符只是按照文本行中心线弯曲的排列分布,而自身并没有被严重非线性扭曲。因此,以往工作所采用的对文本行整体进行非线性较强的 STN 变换可能会造成过度校正、训练困难的结果,而光流预测模块在 L_{geo} 约束下则可以保证变换的局部线性化。

如图 3 所示,考虑左侧红色局部区域 A',将其中心点记为 p,上下左右四条边的中点分别记为 p_0、p_1、p_2 和 p_3,假设经过光流校正后,得到右侧水平文本行中的局部区域 A,其中心点 q 位置对应的像素值应来自校正前的 p 点,其上下左右四边中点的 q_0、

q_1、q_2 和 q_3 位置的值应分别来自校正前的 p_0、p_1、p_2 和 p_3 点。除此之外，右侧沿竖直方向（即 $\overrightarrow{q_0\,q_1}$ 方向）距离中心点 q 为 d 个像素的点 q_d 所对应的源位置应为左测的 p_d 点，设其距离中心点 p 的距离为 $d^{'}$。为了保证局部变换的线性化，希望校正前后两者的比例性、共线性得以保持，即满足 $\frac{|q\,q_d|}{|q_0\,q_1|}=\frac{|p\,p_d|}{|p_0\,p_1|}$ 和 $\overrightarrow{p\,p_d}//\overrightarrow{p_0\,p_1}$，根据 TLG 模块所采用的几何属性描述方法，$|p_0\,p_1|$ 与文本尺度 scale 成正比，$\overrightarrow{p_0\,p_1}$ 的方向即为字符法线方向夹角 φ。因此，定义竖直方向的几何约束损失 L_{geo}^{y} 如式（1）所示：

$$L_{geo}^{y}=\sum_{(q,q_d)\in A}\lambda_1\left|\|\varphi_q-\varphi_{q_d}\|-d^{'}(q)\right|+\lambda_2\left|S(\varphi_q,\varphi_{q_d})-\frac{sin\,\varphi_q}{cos\,\varphi_q}\right|. \tag{1}$$

其中，点 q 为校正后图像区域 A 内的任意一点，其沿竖直方向移动 d 个像素得到点 q_d，通过累加每一组合法的点对 (q,q_d) 所对应的损失作为竖直方向上的总损失。φ_q 和 φ_{q_d} 分别表示点 q 和点 q_d 的光流，即其所对应的源位置点 p 和点 p_d。$d^{'}(q)=\frac{2d\times scale(q)}{H}$，表示按比例计算得到的 $d^{'}$，并与点 q 所在的字符的尺度相关，其中 H 为校正后的图像区域的高度（即 $\|q_0\,q_1\|$），$scale(q)$ 为 TLG 模块所预测的 scale 通道在 q 点处的值。$S(\varphi_q,\varphi_{q_d})$ 表示向量 $\overrightarrow{\varphi_q\,\varphi_{q_d}}$ 与水平方向夹角的斜率，$sin\,\varphi_q$ 和 $cos\,\varphi_q$ 分别表示 TLG 模块语义分割结果的 sin φ 通道和 cos φ 通道在 q 位置的值。λ_1 和 λ_2 为超参数，用于调节比例性和共线性的保持程度，在实验中，发现取 $\lambda_1=0.5$，$\lambda_2=1.0$ 时，效果较好。

水平方向上的几何损失 L_{geo}^{x} 与竖直方向几何损失 L_{geo}^{y} 类似，不同之处在于点 q_d 是由点 q 沿水平方向而非竖直方向移动 d 个像素得到，且使用 θ 通道而非 φ 通道对斜率 $S(\varphi_q,\varphi_{q_d})$ 进行约束，不再赘述。最终，完整的几何约束损失 L_{geo} 如式（2）所示：

$$L_{geo}=\lambda_x L_{geo}^{x}+\lambda_y L_{geo}^{y}. \tag{2}$$

其中，λ_x 和 λ_y 为超参数，实验中分别设置为0.3和1.0。

2.2　关键点回归损失

为了稳定光流预测模型的初期训练过程，本文采用了类似于控制 STN 变换的关键点约束。假设通过在 TCL 分割图上利用 striding 算法生成长度为 k 的中心点序列集合 C，对其中任一中心点 $c_i\in C$，根据式（3）计算得到其对应的上下边界点 t_i 和 b_i，分别形成长度为 k 的上边界点集合 T 和下边界点集合 B，记全体关键点集合为 $P=C\cup T\cup B$，其中任一元素为 $p_i\in P$。

$$\begin{cases}t_i=c_i+(s_i\times\cos\varphi_i,\ -s_i\times\sin\varphi_i),\\ b_i=c_i+(-s_i\times\cos\varphi_i,s_i\times\sin\varphi_i).\end{cases} \tag{3}$$

至此，共生成 $3k$ 个关键点，而它们在校正后所处的位置是先验已知的，即分别在水平文本行的上下边界和中心线均匀采样 k 个关键点的位置。如图 4 所示，若关键点 $p_i^{'}$ 所对应的源位置点为 p_i，则自然希望 $p_i^{'}$ 点处的光流 $\varphi_{p_i^{'}}$ 与 p_i 点越接近越好，这是一个简单的回归任务。

图 3　基于光流的文本校正示意图

Fig. 3　Schematic diagram of text rectification based on optical flow

图 4　关键点先验位置

Fig. 4　The prior positions of key points

因此，引入关键点回归损失 L_{lm} 如式（4）所示：

$$L_{lm}=\sum_{i=1}^{3k}\|\varphi_{p_i^{'}}-p_i\|_2. \tag{4}$$

2.3　光流平滑损失

以上的关键点回归损失只能作用在有限个独立的关键点上，为了让学习到的光流场更为平滑，本文引入了光流平滑损失作为模型训练的正则化项。给定一个二维光流向量场 φ，定义光流平滑损失 L_{TV} 如式（5）：

$$L_{TV}=\|\tilde{\mathrm{N}}_x\,\varphi_x\|^2+\|\tilde{\mathrm{N}}_y\,\varphi_x\|^2+\|\tilde{\mathrm{N}}_x\,\varphi_y\|^2+\|\tilde{\mathrm{N}}_y\,\varphi_y\|^2. \tag{5}$$

其中，$\tilde{\mathrm{N}}_x$ 或 $\tilde{\mathrm{N}}_y$ 表示 x 或 y 方向上的梯度算子；φ_x 或 φ_y 表示光流场 φ 的 x 或 y 分量通道。在具体工程实现时，梯度算子 $\tilde{\mathrm{N}}$和之前 L_{geo} 的计算结果都可以采取特征图错位相减的方法简单得到。

最终，结合几何约束损失 L_{geo}、关键点回归损失

L_{lm} 以及光流平滑正则项 L_{TV}，得到光流网络的总损失函数如式(6)：

$$L_{tot} = \lambda_{geo} L_{geo} + \lambda_{lm} L_{lm} + \lambda_{TV} L_{TV}. \quad (6)$$

3 实 验

3.1 实验设置

首先采用带有字符级别标注的 Synth90k 人工合成数据集预训练文本校正模块和识别模块，使其具有预测文本行几何属性的能力，然后分别在真实数据集 IIIT5K、IC03、IC15、CUTE80 上进行模型微调。其中识别模块总共可以识别出 95 种不同的字符，包括数字、大小写英文字母、32 个标点符号和一个特殊的结束符 EOS。

模型采用 Adam 优化器对总损失函数(6)进行优化，训练批处理大小为 128，初始学习率为 0.1，每经过 60 k 轮迭代时，衰减为原来的 0.1 倍，直到训练总损失几乎不变为止。在特定数据集上测试时，均以上述得到的参数作为初始化，在该数据集的训练集上进行微调，采用较小学习率继续训练大约 10 k 次。

3.2 实验结果

在 IIIT5K、IC15 等通用文本识别数据集上分别测试了采用 ResNet50[16] 和 FlowNet v2-SD[17] 作为主干网络的识别性能，并与基于 STN 网络的模型进行了比较，结果见表 1。

表 1 与基于 STN 网络的文本校正方法的准确率比较

Tab. 1 Comparison of accuracy with other STN-based rectification methods

主干模型	IIIT5K	IC03	IC15	CUTE80
STN-based	**93.0**	**92.8**	77.8	81.9
ResNet50	92.5	91.5	78.1	82.6
FlowNet v2-SD	92.8	92.5	**78.2**	**82.8**

可以发现，在以弯曲文本行为主体的数据集上，如 IC15 和 CUTE 等，基于光流校正的模型要优于传统的基于 STN 变换的模型。采用 FlowNet v2-SD 作为主干网络的模型对于光流的捕捉能力更强，在与基于 ResNet 模型的参数量相当的情况下，文本校正性能有了较大提升，符合预期。

对于几何约束损失 L_{geo} 起到的作用，基于 FlowNet v2-SD 作为主干光流网络，就是否施加几何约束分别进行了相关的对比实验，实验结果见表 2。不难发现，几何约束的施加，使得光流网络可以充分利用文本行几何形状的先验信息，从而提高了光流预测的准确度，进而改善了文本校正质量和识别精度。

表 2 施加几何约束与否对准确率的影响

Tab. 2 Recognition accuracies with and without geometric prior constraints

是否施加几何约束	IIIT5K	IC03	IC15	CUTE80
否	92.6	91.9	77.5	81.5
是	**92.8**	**92.5**	**78.2**	**82.8**

利用带有几何约束的光流变换进行文本图像校正的视觉效果如图 5 所示，其中左列为文本行的原始图像，右列为基于带有几何约束的光流变换的校正结果。可以看出，针对弯曲文本行，基于几何约束的光流网络的校正结果符合预期。

图 5 文本校正可视化结果

Fig. 5 The visualization results of our text rectification method

4 结束语

本文提出了一种新颖的基于光流网络的场景文本校正方法，可以在文本行几何属性的指导下，对弯曲文本行进行校正，进而提升文本识别系统的性能。该方法与传统基于 STN 网络的模型相比，既具有光流变换的灵活性，又具有几何先验约束的可控性。相关实验结果表明，通过将该文本校正模块与后续的序列识别模块相结合，可以提升模型在不规则文本行数据集上的识别准确率。

参考文献

[1] LONG S, HE X, YAO C. Scene text detection and recognition: The deep learning era [J]. International Journal of Computer Vision, 2020: 1-24.

[2] SCHROTH G, HILSENBECK S, HUITL R, et al. Exploiting text-related features for content-based image retrieval[C]//2011 IEEE International Symposium on Multimedia. IEEE, 2011: 77-84.

[3] Guilherme N DeSouza and Avinash C Kak. Vision for mobile robot navigation: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2002,24(2):237-267.

[4] MM Aftab Chowdhury and Kaushik Deb. Extracting and segmenting container name from container images[J]. International Journal of Computer Applications, 2013,74(19).

[5] BOOKSTEIN F L. Principal warps: Thin-plate splines and the decomposition of deformations [J]. IEEE Trans. Pattern Anal. Mach. Intell., 1989,11(6):567-585.

[6] 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.

(下转第 163 页)