计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600157

基于注意力机制的复杂场景文本检测

刘 燕 温 静
山西大学计算机与信息技术学院 太原 030006
(449258197@qq.com)

摘 要 传统的文本检测方法大多采用自下而上的流程,它们通常从低级语义字符或笔画检测开始,然后进行非文本组件过滤、文本行构建和文本行验证。复杂场景中文字的造型、尺度、排版以及周围环境的剧烈变化,导致人的视觉系统是在不同的视觉粒度下完成文本检测任务的,而这些自底向上的传统方法的性能很大程度上依赖于低级特征的检测,难以鲁棒地适应不同粒度下的文本特征。近年来,深度学习方法被应用于文本检测中来保留不同分辨率下的文本特征,但已有的方法在对网络中各层特征提取的过程中没有明确重点特征信息,在各层之间的特征映射中会有信息丢失,造成一些非文本目标被误判,使得检测过程不仅耗时,而且会产生大量误检和漏检。为此,提出一种基于注意力机制的复杂场景文本检测方法,该方法的主要贡献是在VGG16中引入了视觉注意层,在细粒度下利用注意力机制增强网络内全局信息中的显著信息。实验表明,在载有GPU的Ubuntu环境下,该方法在复杂场景文本图片的检测中能保证文本区域的完整性,减少检测区域的碎片化,同时能获得高达87%的查全率和89%的查准率。

关键词:文本检测;深度学习;注意力机制
中图法分类号 TP391

Complex Scene Text Detection Based on Attention Mechanism

LIU Yan and WEN Jing
School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

Abstract Most of the traditional text detection methods are developed in the bottom-up manner, which usually start with low-level semantic character or stroke detection, followed by non-text component filtering, text line construction, and text line validation. However, the modeling, scale, typesetting and surrounding environment of the characters in the complex scene change drastically, and the task of detecting text is carried up by human under variety of visual granularities. It's difficult for these bottom-up traditional methods to maintain the text features under different resolution, due to their dependency on the low lever features. Recently, deep learning methods have been widely used in text detection in order to extract more features under different scale. However, in the existing methods, the key feature information is not emphasized during the feature extraction process of each layer, and will be lost in the layer-to-layer feature mapping process. Therefore, the missing information will also lead to a lot of false-alarm and leak detection, which causes much more time-consuming. This paper proposes a complex scene text detection method based on the attention mechanism. The main contribution of this method is to introduce a visual attention layer in VGG16, and use the attention mechanism to enhance the significant information in the global information in the network. Experiments show that in the Ubuntu environment with GPU, this method can ensure the integrity of the text area in the detection of complex scene text pictures, reduce the fragmentation of the detection area and can achieve up to 87% recall rate and 89% precision rate.

Keywords Text detection, Deep learning, Attention mechanism

1 引言

在自然场景中识别和理解文本一直是计算机视觉中的热点研究问题[1-4]。文本检测是实现文本识别的前提和关键,然而文本样式的多样化和高度杂乱的背景为精确文本定位提出了巨大的挑战。

传统的文本检测方法大多采用自下而上的流程[2-4],它们通常从低级语义字符或笔画检测开始,然后进行非文本组件过滤、文本行构建和文本行验证。这些方法主要分为基于连接组件(Connected-Components,CC)的方法和基于滑动窗口

到稿日期:2019-06-26 返修日期:2019-09-18 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金青年科学基金(61703252);山西省1331工程项目;山西省应用基础研究计划项目(201701D121053)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (61703252), 1331 Engineering Project of Shanxi Province and Shanxi Province Applied Basic Research Programs (201701D121053).
通信作者:温静(wjing@sxu.edu.cn)

的方法两类。基于 CC 的方法通过使用快速滤波器来区分文本和非文本像素，然后使用低级属性（如强度、颜色和梯度等）将文本像素分为笔划或候选字符。基于滑动窗口的方法通过在图像中密集地滑动多尺度窗口来检测候选字符，然后通过预先训练的分类器，使用手动设计的特征[3]或最近的卷积神经网络提取的特征进行区分。这些自底向上的方法的鲁棒性和可靠性较差，其性能在很大程度上依赖于低级特征（如基于 SWT[5]，MSER[2]，HoG[3]的文本检测）。如果没有上下文信息，其难以鲁棒地单独识别各个笔划或字符；如果没有检测重点，这类方法很容易对一些类似文本的纹理特征判断不清晰。这些限制在字符检测中通常会导致大量非文本组件被误检，而且在后续步骤中难以对它们进行处理。同时，传统方法的误检[3]容易在自下而上的过程中形成累积误差。

近年来，深度学习在一般物体检测[6]方面获得了巨大的成功。然而，这些通用目标检测方法很难被直接应用于自然场景文本检测。这是由于在通用目标检测中，每个目标都有一个明确的封闭边界，而在文本中可能不存在这样一个明确定义的边界，因为文本行或单词是由许多单独的字符或笔划组成的。检测文本需要在不同粒度下完成识别任务，需要覆盖文本行或字的整个区域。近来，深度学习也被应用到文本检测方向。深度学习在文本检测中的运用大致可以分为两个方向：基于目标检测和基于目标分割。例如，文献[7]是基于目标分割直接对边框进行回归，不产生目标边框；文献[8]是基于目标检测，在生成候选提议框时回归一个任意多边形。但现有的一些深度学习方法在每层的特征提取过程中没有明确重点特征信息，在各层的特征映射过程中会丢失一些有价值的信息，也会误判非文本的目标，使得检测过程不仅耗时，而且会产生大量误检和漏检。

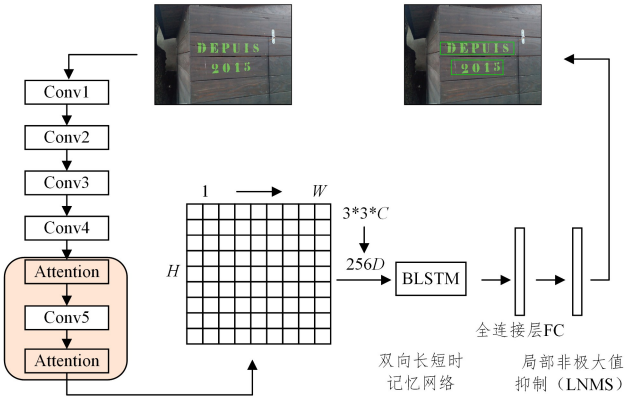

图 1 注意力卷积神经网络的结构

Fig. 1 Structure of attention convolutional neural network

为此，本文将重要的深度特征信息直接在卷积映射中体现，在 VGG16[9]的最后两层加入视觉注意层，利用视觉注意层来加强局部细节的关注点和提取；之后使用一些小的文本提议框，允许区域候选网络[6]（Region Proposal Network，RPN）使用单尺度窗口检测多尺度目标；接着使用可以关联前后序列的循环神经网络[10]（Recurrent Neural Networks，RNN），利用前后文的信息来测定文本位置；最后运用局部感知非极大值抑制（Locality-Aware Non-Maximum Suppression，LNMS）来更加准确地确定文本框的位置。

本文提出了图 1 所示的网络结构注意力卷积神经网络（Attention Convolutional Neural Network，ACNN），在特征提取的网络层的后期加入新的视觉注意层，并且在非极大值抑制（Non-Maximum Suppression，NMS）合并阶段运用了局部感知非极大值抑制。

2 基于注意力机制的网络

注意力在人类感知中起着重要作用。人类视觉系统的一个重要特性是人们不会一次尝试处理整个场景。相反，人类利用一系列局部瞥见并选择性地聚焦于显著部分来更好地捕捉视觉结构。

注意力机制最初应用于目标检测时获得了很好的效果，根据文献[11]的分析，将注意力引入 VGG16 后，目标检测的均值平均精度（mAP）值由原来的 78.9% 提升到了 79.3%。在进行目标检测时，通常需要对网络的每一层都加入注意力模块。而有别于一般的目标检测问题，文本检测更侧重细粒度上的检测任务。文本有自己的纹理特征，选择何种特性的注意力模块与以何种顺序设置和组合模块同样重要。

我们根据 VGG16 的网络结构和每个卷积层对不同特征的敏感度，经过实验对比，选择性地在网络层之间加入注意力模块。图 1 中矩形框框住部分即为本文的视觉注意层，即在 VGG16 的基础上加入注意力模块[11]，这是一种用于前馈卷积神经网络的简单而有效的注意力模块。给定一个中间特征图谱，模块沿着两个独立的维度、通道和空间依次推断注意力映射，然后将注意力映射图谱与输入特征图谱相乘以进行自适应特征细化。

2.1 视觉注意层

在考虑增加网络基数[12]的前提下，本文提出的视觉注意层主要通过使用注意力机制来关注重要特征并抑制不必要的特征。卷积运算是通过将多个通道信息和空间信息混合在一起来提取信息特征，本文采用注意力层就是为了加强这两个主要维度中有意义的特征：通道和空间轴。为了实现该目标，依次应用通道和空间注意模块（如图 2 所示），这样每个分支都可以分别在通道和空间轴上学习“什么”和“在哪里”。因此，注意力模块通过了解要强调或抑制的信息，有效地帮助网络中的信息传递。

视觉注意层的整体模块设计如图 2 所示，在特征处理阶段分为两个层面，一个是通道上的注意力，另一个是空间上的注意力。我们给定中间特征映射 $F \in R^{C*H*W}$ 作为输入，模型依次推断出一个一维通道注意力图谱 $M_C \in R^{C*1*1}$ 和一个二维的空间注意力图谱 $M_S \in R^{1*H*W}$。

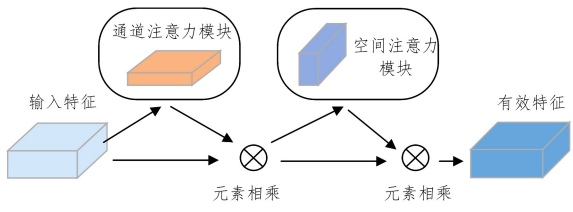

图 2 注意力模块

Fig. 2 Attention module

整个注意过程可以被概括为：

$$F'=M_C(F)\otimes F \tag{1}$$

$$F''=M_s(F')\otimes F' \tag{2}$$

其中，$\otimes$表示逐元素乘法。

在乘法过程中，注意力值沿空间维度通道被广播，F''是最终的输出。下面分别介绍每个注意力图谱的计算过程以及每个模块的细节。

(1)通道注意模块。通过利用特征的通道间关系来产生通道注意力图谱。当场景文字特征图谱的每个通道都被视为特征检测器时[13]，对于给定的输入图像，通道注意力集中在“什么”上是有意义的。我们的目标是检测场景中的文本，所以显著性的文本是注意力的重点。为了有效地计算通道注意力，首先将场景图片的特征图在空间维度上进行压缩，得到一个一维矢量后再进行操作。为了聚集空间信息，对输入特征图进行空间维度压缩时，不仅考虑了平均池化层，还引入了最大池化层作为补充。全局平均池化层对特征图上的每一个像素点都有反馈，而全局最大池化层在进行梯度反向传播计算时，把梯度直接传给前一层的某个特定像素，而其他像素不接受梯度，也就是为 0。通道注意力模块的结构如图 3 所示。

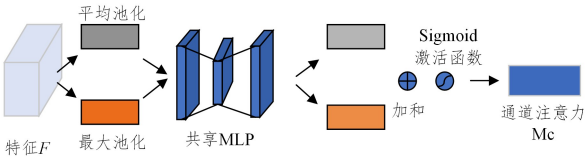

图 3　通道注意力模块

Fig. 3　Channel attention module

通过使用平均池化和最大池化操作来聚集特征图谱的空间信息，从而生成两个不同的空间上下文描述符 F^c_{avg}和 F^c_{max}，其分别表示平均池化特征和最大池化特征。然后将两个描述符转发到共享网络，以产生通道注意力图谱 $M_C\in R^{C*1*1}$。共享网络由带有一个隐藏层的多层感知机(Multilayer Perceptron，MLP)组成。为了减少参数开销，将隐藏的激活值设置为 $R^{\frac{C}{r*1*1}}$，其中 r 是压缩率。在将共享网络应用于每个描述符后，使用逐元素求和来合并输出特征向量。简而言之，通道注意力的计算如下：

$$\begin{aligned}M_C(F)&=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))\\&=\sigma(W_1(W_0(F^c_{avg}))+W_1(W_0(F^c_{max})))\end{aligned} \tag{3}$$

其中，σ 表示 Sigmoid 函数，$W_0\in R^{C/r*C}$，$W_1\in R^{C*C/r}$。W_0 和 W_1 是多层感知机的权重，共享输入 W_0 的激活函数。

(2)空间注意模块。利用特征的空间关系生成空间注意力图谱。与通道注意力不同的是，空间注意力集中在“哪里”，它与通道注意力相辅相成。除了在通道上生成注意力模型，在空间层面上也需要网络明白特征中哪些部分应该有更高的响应优先级。首先，使用平均池化和最大池化对输入特征图进行压缩操作，但这里的压缩变成了通道层面上的压缩，对输入特征分别在通道维度上做平均和最大操作。沿着通道轴应用池化操作是有效的，有效的空间特征会突显在信息区域中。然后得到两个二维的特征图：$F^s_{avg}\in R^{1*H*W}$ 和 $F^s_{max}\in R^{1*H*W}$。将这两个二维特征图按通道维度拼接在一起，得到一个通道数为 2 的特征图。在级联特征描述符上，使用一个包含单个卷积核的隐藏层对得到的特征图进行卷积操作，应用卷积层来生成空间注意力图谱 $M_S(F)\in R^{H*W}$，表示其强调或抑制的位置。空间注意力模块如图 4 所示。

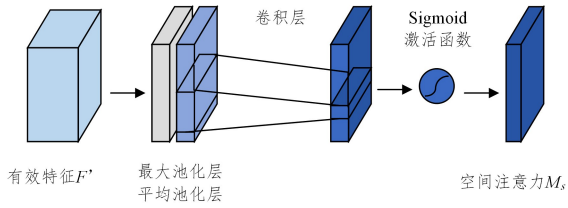

图 4　空间注意力模块

Fig. 4　Space attention module

简而言之，空间注意力的计算如下：

$$\begin{aligned}M_S(F)&=\sigma(f^{7*7}([AvgPool(F);MaxPool(F)]))\\&=\sigma(f^{7*7}([F^s_{avg};F^s_{max}]))\end{aligned} \tag{4}$$

其中，σ 表示 Sigmoid 函数；f^{7*7}表示卷积操作，卷积核大小为 $7*7$。

2.2　Fine-granularity(细粒度)文本检测

细粒度图像分析一直是计算机视觉中的热门研究领域。细粒度检测，即模型需要识别非常精细的子类别。而在本文的任务中，细粒度检测的含义是检测到场景文本中比较显著的文本区域，减少背景和类似文本的目标干扰，降低漏检率和误检率，从而加快检测速度。我们在特征提取模型中插入新的视觉注意力模块，目的是加大文本区域的注意力。其中，模块的放置位置至关重要。由于前期的特征提取都属于局部信息，而视觉注意具有全局性，因此我们在 VGG16 中表征全局信息的后两层分别加入了 6 个注意力模块。经多次实验证明，这一设计可以使得模型的效果达到最好。VGG16 的具体模型如图 5 所示。

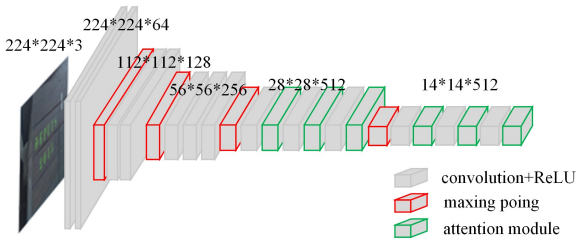

图 5　VGG16 注意力模型

Fig. 5　VGG16 attention model

网络的前几层大尺度特征图(较靠前的特征图)可以用来检测小目标，而小尺度特征图(较靠后的特征图)用来检测大目标。由于我们的需求是减少漏检和误检，因此注意力应放在显著的文本区域。网络在提取特征后，运用了区域提取网络和双向长短时记忆网络，允许输入任意大小的图像。通过在卷积特征映射中密集地滑动小窗口来检测文本行，从而输出一系列细粒度的文本提议。这里使用一个小的空间窗口 $3*3$ 来滑动最后的注意力层特征。通过前几层网络对小目标的特征检测，我们可以把文本区域和类似文本特征的部分全部检测出来；而在网络的后两层加入注意力机制，把关注点放在场景文本显著的区域，可以丢弃一些含有弱文本信息的模糊区域。其次，后面所连接的双向长短时记忆网络(Bi-di-

rectional Long Short-Term Memory Recurrent Neural Network, Bi-LSTM), 可以有效地联系上下文, 使得文本的检测区域比较完整, 降低漏检率。由于我们加入了注意力机制, 网络检测的部分比较明确, 从而减少了与文本特征相似的目标的误检, 使得网络更关注于我们所注意的目标区域。

给定输入图像, 通过最后一个注意力层输出的特征图大小为 $W * H * C$, 其中 C 是特征映射或通道的数目, $W * H$ 是空间范围。当检测器通过特征图密集地滑动 $3 * 3$ 窗口时, 每个滑动窗口通过 $3 * 3 * C$ 的卷积特征来产生预测。对于每个预测, 水平位置(x 轴坐标)和 $k=10$ 个文本提议框的位置是固定的(宽度为 16 个像素, 高度在输入图像中从 11 个像素变化到 273 个像素, 高度每次除以 0.7。检测器在每个窗口位置输出 k 个文本提议框的文本/非文本分数和预测的 y 轴坐标(v)。文本具有强大的序列特征, 序列的上下文信息对于做出可靠决策至关重要。使用递归神经网络(Recursive Neural Network, RNN)对用于文本识别的上下文信息进行编码, 它可以丢弃一些含有弱文本信息的模糊区域。因此, 本文在最后的注意力层上利用长短时记忆(Long Short-Term Memory, LSTM)架构作为 RNN 层。对每个 LSTM 使用一个 128 维的隐藏层, 从而产生 256 维的 RNN 隐藏层。隐藏层内部状态被映射到后面的 FC(全连接层), 并且其输出层用于计算第 t 个提议的预测。

2.3 局部感知非极大值抑制

为了形成最终的结果, 需要用非极大值抑制(Non Maximum Suppression, NMS)对经过阈值处理后的几何形状进行合并。传统的 NMS 过程是去除冗余的重叠 Boxes, 对全部的 Boxes 进行迭代-遍历-消除。首先将所有框的得分排序, 选中最高分及其对应的框; 然后遍历其余的框, 如果其与当前最高分框的重叠面积 IOU 大于一定的阈值, 则将该框删除; 接着从未处理的框中继续选一个得分最高的, 重复上述过程。传统 NMS 算法的时间复杂度是 $O(n^2)$, 其中 n 是候选几何的数量。假设邻近像素的几何形状高度相关, 那么可以逐行合并几何图形, 在对同一行几何图形的得分进行加权平均合并的同时, 将当前遇到的几何图形与上一次合并的几何图形进行迭代合并。改进后算法的时间复杂度是 $O(n)$。改进算法不仅加快了运行速度, 而且减少了计算量。

3 实验结果与分析

数据集 ICDAR2017-MLT 包含了英文、中文、韩文、日文、阿拉伯文等 9 种文字。数据集中文字的设计感强, 同一幅图片中的文字具有不同的尺度、颜色和字体(如图 6(a)中第 5 行第 1 幅图中"pillow"); 此外, 拍摄时摄像头与文字平面的角度变化剧烈(如图 6(a)中第 4 行第 1 幅图), 文字周围的环境纹理复杂, 易产生误检和文字区域过分割。ICDAR2015 数据集是复杂的街拍图, 大多为商场里的随拍, 图中目标多且文本区域过小(如图 7 中第 4 行), 还存在各种角度和模糊拍摄的客观影响因素, 检测难度较大。ICDAR2013Standard 数据集是普通生活区域的外景街拍图片, 文字变体比较多, 一幅图片中存在多种尺寸和多种字体(如图 7 中第 2 行), 对细节辨识度的要求很高。

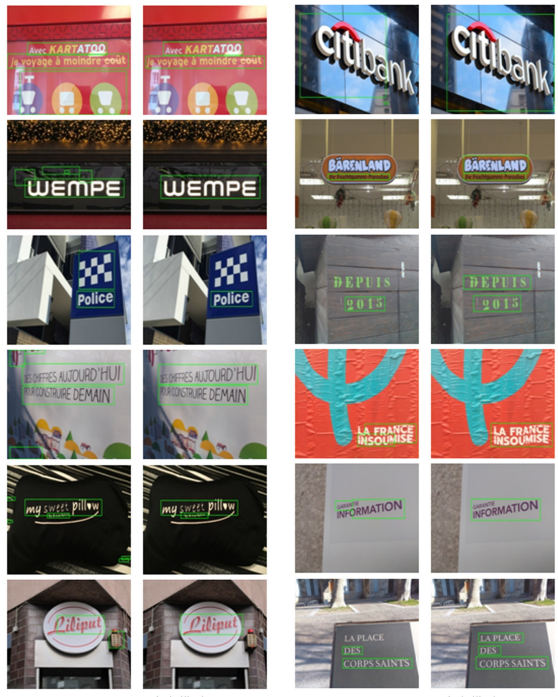

图 6 ICDAR2017-MLT 数据集上的实验对比

Fig. 6 Experiment comparison on ICDAR2017-MLT dataset

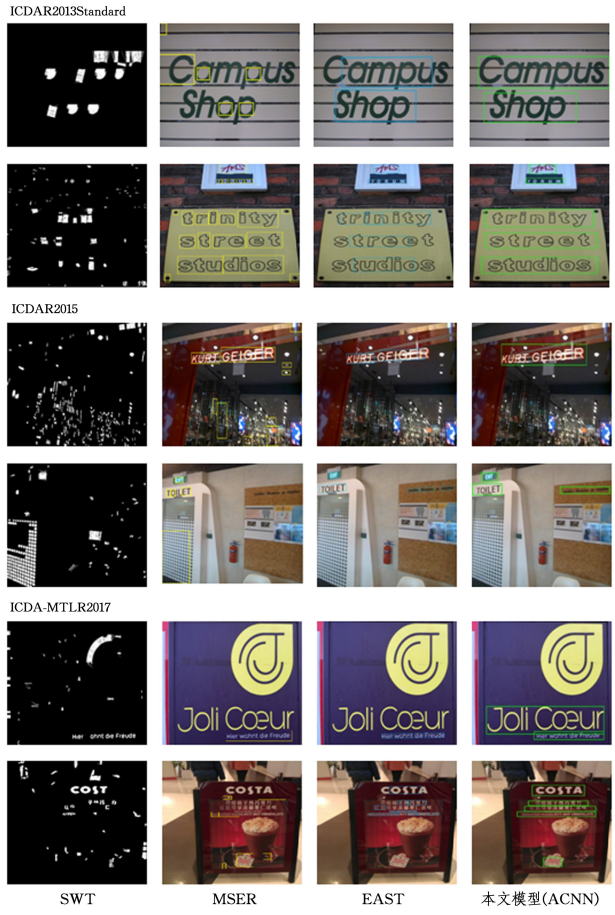

图 7 3 个数据集上的实验结果对比

Fig. 7 Experiment comparison on three datasets

3.1　实验对比图

实验 1　把注意力机制加入本文模型的基础特征层中与 CTPN[14] 进行对比，实验结果如图 6 所示。从图 6(a)中可以明显看出：加入了注意力机制的模型有效地减少了误检，对于接近文本纹理的部分，CTPN 易将其作为文本部分做出响应，而本文的方法则减少了这种失误。从图 6(b)可以看出：CTPN 存在明显的漏检问题，而本文模型由于引入了视觉注意机制，能够最大程度地分辨文本区域，因此能有效地避免漏检的发生。

实验 2　把局部感知非极大值抑制和注意力机制都加入基础网络层中，同时将其与传统方法 SWT[5]，MSER[2] 以及具有代表性的深度学习方法 EAST[15] 进行对比，实验结果如图 7 所示。可以看出：SWT 方法所获取的文本检测虽然响应很多，但无法保留大部分的文本区域，是 4 种方法中误检最多的；MSER 方法较 SWT 方法能够获得更多的文本部分，但依然不能有效地减少误检，同时不能检测出全部的文本区域；而 EAST 能够区别出文本和背景，同时获得大部分文本区域，但该方法将文本区域过度分割，产生了很多不必要的碎片以及重叠区域，使得图 7 第 2 行的图片中同一行的文字被分割成了两个区域，此外第 5 行中明显的文字部分由于存在艺术化设计而被漏检；本文方法明显优于对比的方法，能够最大程度地保留文本区域的完整性，在图 7 最后一行的文字检测中，本文方法可以同时检测出英文和中文字体，保证了检测的结果接近人眼获取信息的特性和人们的阅读习惯。

实验 3　在随机街拍的图片集上进行实验，图片中包含中文、日语和印度语，实验结果如图 8 所示。可以看出，本文方法可以同时适用于多种语言检测，并获得了令人满意的性能和效果。

图 8　多语言场景文字的检测

Fig. 8　Multi-language scene text detection

3.2　实验评估

通过对比实验可以发现，本文提出的基于视觉注意机制的深度学习网络能够在文本检测中获得令人满意的效果。为了评测本文方法的性能，我们在 ICDAR13Standard，ICDAR2017-MLT 和 ICDAR2015 数据集上对其进行了定量分析，计算出了其对文字检测的查全率和查准率。由于文字区域分散在图片中的各个部分，我们在计算文字区域时按照人们的阅读理解习惯，例如：同一行的文字算作同一个区域。实验结果如表 1 所列，可以看出，在 ICDAR13-Standard 数据集上，本文方法能够获得 87%的查全率和 89%的查准率；如表 2 所列，在 ICDAR2015 数据集上本文方法也获得了 58%的查全率和 80%的查准率；如表 3 所列，在 ICDAR2017-MLT 数据集上本文方法获得了 86%的查全率和 93%的查准率。

表 1　ICDAR13Standard 数据集上的实验结果

Table 1　Experiment results on ICDAR13Standard dataset

方法	*Recall*	*Precision*	*F*
SSD[16]	0.60	0.80	0.68
I2R NUS FAR[17]	0.69	0.75	0.72
I2R_NUS[17]	0.66	0.73	0.69
CASIA NLPR[17]	0.68	0.69	0.73
Bai et al[18]	0.68	0.69	0.73
USTB TexStar[19]	0.66	0.88	0.76
Ours	0.87	0.89	0.88

表 2　ICDAR15 数据集上的实验结果

Table 2　Experiment results on ICDAR15 dataset

方法	*Recall*	*Precision*	*F*
CTPN[14]	0.52	0.74	0.61
MCLAB_FCN[20]	0.43	0.71	0.54
Yao et. al.[21]	0.57	0.72	0.64
Ours	0.58	0.80	0.67

表 3　ICDAR17-MLT 数据集上的实验结果

Table 3　Experiment results on ICDAR17-MLT dataset

方法	*Recall*	*Precision*	*F*
FOTS[22]	0.58	0.81	0.67
Pixel-anchor[23]	0.60	0.80	0.68
CRAFT[24]	0.68	0.81	0.74
Ours	0.86	0.93	0.89

结束语　本文提出了一种基于视觉注意机制的复杂场景文本检测，通过在不同卷积层引入视觉通道，增强了细粒度下的特征。复杂场景中，在文本尺度变化剧烈、字体设计感重，以及文本与背景接近等多种情况下，所提方法都能获得较优的结果。然而，在实验中也发现由于不能保留融合各个层的特征，因此在卷积操作时容易丢失特征信息，未来可以考虑加入新的模块来保留各个分辨率下的特征。

参 考 文 献

[1]　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.

[2]　HUANG W, QIAO Y, TANG X. Robust scene text detection with convolutional neural networks induced mser trees[C]// European Conference on Computer Vision (ECCV). 2014: 3.

[3]　TIAN S, PAN Y, HUANG C, et al. Text flow: A unified text detection system in natural scene images[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 4651-4659.

[4]　YIN X C, PEI W Y, ZHANG J, et al. Multi-orientation scene text detection with adaptive clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1930-1937.

[5] EPSHTEIN B, OFEK E, WEXLER Y. Detecting text in natural scenes with stroke width transform[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2963-2970.

[6] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015: 91-99.

[7] HE W, ZHANG X Y, YIN F, et al. Deep direct regression for multi-oriented scene text detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 745-753.

[8] LIU Y, JIN L. Deep matching prior network: Toward tighter multi-oriented text detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1962-1969.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.

[10] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.

[11] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[12] HE T, HUANG W, QIAO Y, et al. Text-attentional convolutional neural network for scene text detection[J]. IEEE Transactions on Image Processing, 2016, 25(6): 2529-2541.

[13] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision. Cham: Springer. 2014: 818-833.

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

[15] 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.

[16] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[17] ICDAR 2013 robust reading competition challenge 2 results [OL]. http://dag. cvc. uab. es/icdar2013.

[18] BAI B, YIN F, LIU C L. Scene text localization using gradient local correlation[C]// 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013: 1380-1384.

[19] YIN X C, YIN X, HUANG K, et al. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970-983.

[20] ZHANG Z, ZHANG C, SHEN W, et al. Multi-oriented Text Detection with Fully Convolutional Networks[C]// Computer Vision and Pattern Recognition. 2016: 4159-4167.

[21] YAO C, BAI X, SANG N, et al. Scene Text Detection via Holistic, Multi-Channel Prediction[J]. arXiv: 1606. 09002, 2016.

[22] LIU X, LIANG D, YAN S, et al. FOTS: Fast Oriented Text Spotting with a Unified Network[C]// Computer Vision and Pattern Recognition. 2018: 5676-5685.

[23] LI Y, YU Y, LI Z, et al. Pixel-Anchor: A Fast Oriented Scene Text Detector with Combined Networks[J]. arXiv: 1811. 07432, 2018.

[24] BAEK Y, LEE B, HAN D, et al. Character Region Awareness for Text Detection[C]// Computer Vision and Pattern Recognition. 2019: 9365-9374.

LIU Yan, born in 1990, master. Her main research interests include computer vision and so on.

WEN Jing, born in 1982, Ph.D, associate professor, master tutor, is a member of China Computer Federation. Her main research interests include computer vision, image processing and pattern recognition.