哈尔滨华德学院毕业设计（论文）评语

姓名：刘雨　　班号：1801112　　专业：计算机科学与技术

毕业设计（论文）题目：基于机器视觉的高精度文字识别方法研究与应用

工作起止日期：2021 年 9 月 13 日起　2021 年 11 月 18 日止

指导教师对毕业设计（论文）进行情况，完成质量及评分意见：

指导教师签字：　　指导教师职称：

评阅人评阅意见：

评阅教师签字：　　评阅教师职称：

答辩委员会评语：

_____

_____

_____

_____

_____

_____

根据毕业设计（论文）的材料和学生的答辩情况，答辩委员会作出如下评定：

学生 _____ 毕业设计（论文）答辩成绩评定为：_____

对毕业设计（论文）的特殊评语：

_____

_____

_____

_____

_____

_____

_____

_____

答辩委员会主任（签字）：_____职称：_____

答辩委员会副主任（签字）：_____ _____

答辩委员会委员（签字）：_____ _____ _____ _____

_____ _____ _____ _____ _____ _____

年 月 日

哈尔滨华德学院毕业设计（论文）任务书

姓　　名：刘雨　　学　　院：数据科学与人工智能学院

专　　业：计算机科学与技术　　班　　号：1801112

任务起至日期：　　2021 年 9 月 13 日　至　2021 年 11 月 18 日

毕业设计（论文）题目：基于机器视觉的高精度文字识别方法研究与应用

立题的目的和意义：

如今文字识别在电子笔记、图书报刊数字化、邮政编码及自动分拣、表单名片识别等都产生了非常广泛的应用的同时，传统 OCR 技术不能有效应对自然场景下复杂背景的文字识别任务，

所以本次毕设旨在研究、找寻一种能够以较高精度处理自然场景下文字识别任务的识别方法，并将其成果实际应用到现实中的场景，实现理论技术与现实生产力的转换。

课题研究主要内容：

1.文字图片采集与预处理方法研究。

2.文本范围智能检测与段落处理研究。

3.提高文字识别精度与响应时间的方法研究。

4.将处理妥善的图像素材导入建立完毕的方法模型中进行文字识别流程，获得精度响应时间等性能参数。

5.性能统计与对比，并得出实验结论。

进度安排：

序号	名称	周数	起止时间	备注
1	可行性研究	1 周	2021.09.13~2021.09.19	
2	需求分析	2 周	2021.09.20~2021.10.03	
3	系统设计	3 周	2021.10.04 ~2021.10.24	
4	系统实现	3 周	2021.10.25~2021.11.14	
5	系统调试	1 周	2021.11.14~2021.11.21	
6	撰写论文	3 周	2021.11.22~2021.12.12	

同组设计者及分工：

独立完成

指导教师签字_____

年　　月　　日

系主任意见：

系主任签字_____

年　　月　　日

摘　要

随着近年来计算机技术的发展，技术研究者们在各个领域都取得了显著的成果，并且已广泛的应用在我们日常的工作学习中，其中也包括文字识别相关领域。

目前，传统的也是主流的识别技术是 OCR（Optical Character Recognition，光学字符识别），尽管 OCR 技术已经较为成熟，但是在处理自然环境下的文本，往往由于环境背景复杂、文本不定长、文本内容混杂以及文本扭曲畸变等相关因素影响，OCR 并不能得到较为理想的识别效果，还需要寻找其他方法来解决诸如上述的问题。所以本次课题旨在研究一种能够以较高精度够处理识别自然场景下的文本方法。

基于机器视觉的高精度文字识别方法研究与应用，本片论文主要包含了以下工作：

（1）为了解决能够处理自然环境下内容混杂、不定长以及可能由于文本范围不规范导致发生畸变的文本，提出了基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法。该方法可仅需较小的感受野便可精准的在字符级别处理复杂背景下的文字检测与识别任务。

（2）搭建传统的 OCR 识别流程作为基线模型。通过常规的预处理，轮廓检测，轮廓遍历，预处理后将图像素材导入 OCR 进行识别。并通过控制变量，对比 OCR 与本文提出的基于卷积循环神经网络与一种弱监督方式训练的字符级检测网络构成的识别方法进行性能对照，验证该方法相较于传统的 OCR 技术是否具有优越性。

通过上述实验流程并综合实验结果，很明显的发现本课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法相比传统 OCR 技术在处理如文本格式字体不规范、场景环境复杂等文字识别检测任务时表现出了更好的效果。

关键词：机器视觉；循环神经网络；检测网络；文字识别

Abstract

With the development of computer technology in recent years, technical researchers have achieved remarkable results in various fields, and they have been widely used in our daily work and study, including text recognition related fields.

At present, the traditional and mainstream recognition technology is OCR (Optical Character Recognition, Optical Character Recognition). Although OCR technology is relatively mature, when processing text in natural environments, it is often due to complex environmental background, variable length of text, and mixed text content. As well as the influence of related factors such as text distortion and distortion, OCR cannot get a more ideal recognition effect, and it is necessary to find other methods to solve the problems such as the above. Therefore, the purpose of this project is to study a text method that can process and recognize natural scenes with high accuracy.

Research and application of high-precision text recognition methods based on machine vision. This paper mainly includes the following work:

(1) In order to solve the problem of dealing with mixed content, variable length, and text that may be distorted due to irregular text range in the natural environment, a character-level detection network based on convolutional recurrent neural network and a weakly supervised training method is proposed. CRAFT and CRNN text recognition method under weakly supervised learning (CCRW). This method can accurately process text detection and recognition tasks under complex backgrounds at the character level with only a small receptive field.

(2) Build a traditional OCR recognition process as a baseline model. Through conventional preprocessing, contour detection, contour traversal, the image material is imported into OCR for recognition after preprocessing. And through the control variables, compare the performance of OCR with the recognition method based on the convolutional recurrent neural network and a character-level detection network trained in a weakly supervised manner to verify the performance of the method compared with the traditional OCR technology.

sex.

Through the above experimental process and the comprehensive experimental results, it is obvious that the text recognition method based on the convolutional recurrent neural network and a character-level detection network trained in a weakly supervised manner is compared with the traditional OCR technology in processing such as text. It shows better results in text recognition and detection tasks such as irregular fonts and complex scene environments.

Keywords： Machine vision Recurrent Neural Network Detection network Text recognition

目　录

第 1 章　绪论

1.1　课题背景及意义

通过可靠的研究与报导，可以得知视觉是我们人类获取信息的与感知的重要渠道。在人们平时的生活工作中，会接触到各式各样的、海量的信息，这其中最多的是以文本的形式存在。文本中包含的文字是人类思考、交流、沟通所必不可少的一种工具，在文明的延续与传承中有着举足轻重的作用。通过对文本所展现的信息进行深度挖掘，我们往往可以得到更多、更完善的信息。为了完成这一任务，OCR（光学字符识别，Optical Character Recognition）技术顺应时代而生，其底层原理为使用计算机技术与光学技术，对存在于目标图像、物体之上的文本进行检测，并转换、识别、输出成机器与人都可阅读的形式，为后续的文字录入、处理工作提供便利。现如今，OCR 技术已经相当成熟，对于处理环境简单的文本效果非常理想。但处理如文本字体不规范，背景色彩元素过多等文字识别检测任务时就需要针对环境带来的影响在某些复杂的场景内结合特定技术实现场景文本的识别（Scene Text Recognition,STR）。而现实情况中却大都是一些复杂场景，这就需要做场景文本识别(STR)，我们需要从并不单一的场景中识别、提取出文本信息。导致这一结果的原因很多，且都非常普遍，如拍照时器材成像质量不一、牌匾广告等文本排版不规则、海报涂鸦文字形式多样等，这就导致识别的难度上升，识别的精度下降。而 OCR 技术应付起这些场景就显得乏力，这时我们便需要设计新的方案来解决这一问题。场景文本识别技术的识别精度还有待提高，尚不能达到理想效果，而文本中的信息也不仅仅是提取出来就可以了，我们还应该对其使用自然语言处理技术，对整个文本的信息、内容、语义、语法进行理解分析，并通过分析结果对识别结果进行校正，从而避免一些识别错误，尽一切可能的去提高精度。

现如今，文字识别技术已经渗透进我们的生活，变得无处不在，如社交软件的截屏转换文字、办公软件的证件识别、邮寄快递的单号扫描等。本课题主要为找寻一种较为合理的文字识别方案，试图解决目前 OCR 技术所无法解决的一些问题，并可以应用于课件文本转换等一些实际的应用场景，所以研究一种实用性与良好的性能兼备的文字识别技术是有意义的。

1.2 国内外研究现状

在计算机领域甚至是在整个科学界,人工智能都是非常热门的研究话题，在过去两年，与人工智能和深度学习、机器学习相关的文章已经多次登上 Nature 与 Science 并被作为封面文章。而著名的期刊《科学美国人（Scientific American）》去年亦发表文字表示“人工智能的春天“来到了。

在国内，人工智能的春天也正在悄悄来临，从国家科研机构、国家科技部门以及政府等，都在积极的进行着人工智能与机器学习方面的学术研究与其在相关产业中的应用推广。其中视觉感知是人工智能所要解决的重要的问题之一，而文字识别又是非常重要的计算机视觉技术。因为符号、文字是人类感受、认知世界的最主要手段，无论是与人交流还是自身学习，都离不开文字的参与。在日常生活当中，这种包含着重要信息的“符号”更是无处不在，若是不借助文字的帮助我们将很难准确的理解社会和世界。而文字所包含信息的重要性同时也表现在多个层面，它是人类文明的印记，是传递表达心意途径，积累学习知识的重要工具，是记录过往、经验以及描绘未来的载体。人类感知、解读社会于世界的信息中有超过 91%来自于视觉感知，所以想要赋予机器感知世界的能力，视觉感知能力是基础也是热点，而文字识别又是视觉感知中的基础与重点。

随着机器视觉技术的不断发展，解决问题的手段逐渐丰富。二十一世纪到来后，伴随着智能移动设备的普及与全球互联网用户的激增，基于设备图像采集和光学字符识别技术（OCR），以及手写文字特征检测与识别技术等到了人们的关注。

传统的文字识别的一般框架包括预处理、特征提取、分类器设计三个主要模块，各个模块都需要缜密的分析、设计。主流的文本识别方法大都是基于滑窗的方法或是基于连通域分析的方法。这些方法虽然具有速度快、无需庞大的数据来进行模型训练的优点，同时也面临识别性能不够高的问题。

近年来随着技术人员对深度学习技术的研究，相继提出了很多基于深度学习的方案，诸如 Faster R-CNN/YOLO/SSD/R-FCN 等基于深度学习的方案为解决此类问题提供了新的方向。尽管文字检测也可以粗略的认定为物体检测的一种，但简单地把深度学习中的物体检测框架做检测是达不到理想效果的。虽然可以在整个流程中加入深度学习技术，来为这些问题提供优化与解决方案。而简单、暴力的利用深度学习解决这些比较难的问题，效果并不是很理想。所以我们要找寻更合适但我方法，或是针对不同语言设计方法，或

是针对不同字体设计方法，来使深度学习与传统方法结合，从而更好的服务与不同的应用场景。

1.2.1 主流的的识别模型

STR 包含有文本的检测与识别这两个阶段。检测顾名思义便是在传入的图像中检测文字所在的范围，找出文本范围后用矩形框框选出该范围并生成框坐标。识别则是对检测到的文本范围内的的信息进行检测与提取。其中因为识别是在检测的基础上进行的，对于一个文字识别方案来说，文本范围检测的是否准确对后续的识别精度有着很大的影响。所以，为了达到更好的效果，也可以用识别的结果去监督检测的结果。

初期的文本检测技术大都使用人工设计特征的这一传统方法，所谓人工设计特征便是根据数据中特定的目标做出连通域与纹理的特征设计，并分析底部的特征从而去除背景区域来确定出字符范围。但是随着人工智能技术的飞速更新与发展，基于机器学习与神经网络的文字识别技术也在被不断的提出与研究。神经网络与深度学习在特征提取上有着极大的优势，使得文字识别的精度也得到极大的提高。常见的模型有上文提到的 Faster R-CNN/YOLO/SSD/R-FCN 等，其中 Faster R-CNN 这一模型又衍生出了 CTPN（Connection Text Proposal Network）模型。CTPN 模型是具有重大意义的，其首次引入了循环神经网络，用循环神经网络来预测、判断、学习上下文的语义，从未提高了检测网络的精度。而 YOLO 与 SSD 这类方法则与 Faster R-CNN 不同，它们不需要生成备选框。

1.2.2 自然场景文字识别的困难

正如上文所述，OCR 虽然是一种相当成熟的文字识别技术，但是往往在处理自然场景时得不到一个理想的效果，它的应用场景仅仅局限于环境单一的文档之类的扫描，而自然场景在实际应用中出现的频率显然要多余单一场景，这也是文字识别的一个具有挑战性的应用，究其原因在于自然场景的背景环境总是糅杂的，甚至时高曝光的、极端的，这就导致被检测图像的纹理或轮廓被破坏，导致图像整体的颜色混乱或部分线条扭曲，最终非文本区域也有可能被误检测为文本区域，最终干扰检测的精准度。

除了上述情况，现实中往往还存在大量不规则的文本，如图 1-3，这类文本由于本身是扭曲的、不规则的，这也对范围检测形成了不小的考验，上文中曾说过检测环节对后期的文本内容提取与识别有着至关重要的作用，所

以如何处理不规则文本来达到一个较为理想的状态也是一个挑战。

因为不规则的文本一般都有着不规则的排版，这就导致字与字之间的边界变得模糊，间距不可确定，甚至出现相互覆盖，在这种情况下检测网络很容易出现混乱以至于丢失检测目标，并且在丢失目标的同时还容易伴随统一性的丢失。

在面临这种情况的时候，我们常用的手段是使用一个矫正网络，采用几何技术去校正，使其尽量规整。

1.3 主要研究内容

为解决传统 OCR 技术在处理文字不规范、字体多样以及背景复杂等自然场景下的文字识别任务效果不理想的问题，并经过上述分析与课题相关内容的国内外研究现状，本次课题将对现有的文字识别技术进行研究，在整体过程中对部分流程做出合理的修改，进行适当的算法优化，以此来获得一种能够以理想效果处理自然场景下复杂情况的文字识别任务的文字识别方法，并借助自主训练的数据集在实验中验证该方法的可行性本次课题的整体流程图如下图 1-4 所示。

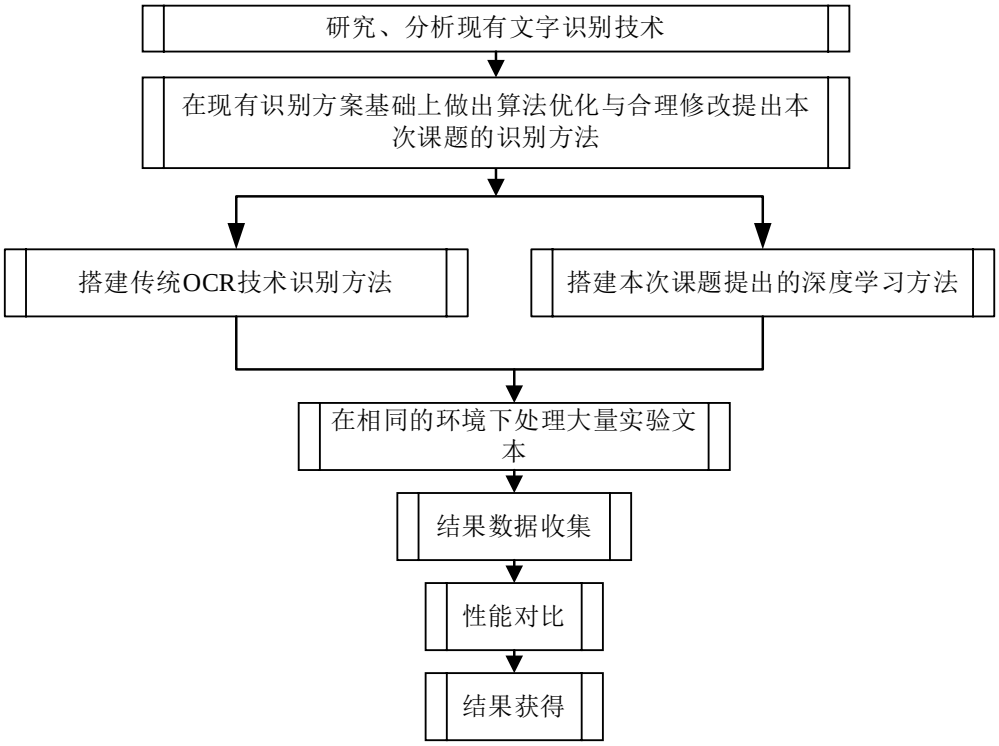

图 1-4 本次课题研究的整体流程

本次课题主要研究内容如下。

1.方法提出与可行性验证

在本次课题中提出一种基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法(CRAFT and CRNN text recognition method under weakly supervised learning ,CCRW)。为了解决传统 OCR 技术在处理复杂背景环境文本乏力的情况,在提出的新方法中优化了多重网络细节，采用了字符级别的检测网络，并在多种复杂的情况下通过实验，对比验证其可行性。

2.算法调优与模型建立

因本次课题旨在解决自然环境下的文本检测，所以要对文中提出的自然环境文本检测所面临的问题做出应对。在本次课题中引入 CRNN 以替代 N-Gram 语言模型，在神经网络之后采用 seq2seq 模型对语义语法做出判断，并引入字符级的检测网络 CRAFT 进行辅助，进一步提高整个方法的文本识别准确度，其中为了训练完善该方法，在条件允许的情况下尽可能的采集了大量的文本信息建立了自己的数据模型,对整个方法做出了有针对性的训练。

3.性能对比

将本次课题提出的一种基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法与传统 OCR 技术建立对照实验，通过多种不同场景的检测实验对比各项性能参数，经过对比验证该方法是否可行，相较于传统 OCR 技术是否具有响应时间、识别准确度等性能上的优越性。

1.4 论文结构

本次课题为基于机器视觉的高精度文字识别方法研究与应用，本文在论文结构层面，一共分为六个章节，各章节任务描述如下。

第一章：该章节为绪论部分。从研究背景，研究意义上介绍了自然场景文本检测技术并探讨了其存在的必要性，并根据当前文字识别领域所处的国内外研究状况和一些实际存在的问题阐述了应对方案并讨论了其可行性，基于上述内容确立了本次课题的主要研究内容，且粗略的介绍了研究流程。

第二章：该章节主要为相关概念与技术介绍。在本章中首先按介绍了与课题相关的一些技术与基础知识，然后介绍了编码器与解码器的概念和部分模型，并针对课题提出的方法中涉及的部分流程做出介绍。

第三章：该章节主要为基于卷积循环神经网络与一种以弱监督方式训练

的字符级检测网络构成的文字识别方法的研究。在此章节里首先按介绍了该方法提出的背景及应用场景和不同于传统 OCR 技术的新特点，然后详尽的论述了整个方法各个环节。

第四章：该章节主要为本次课题提出的一种以弱监督方式训练的字符级检测网络构成的文字识别方法在不同场景中的实现，验证各个环节的稳定性、论述其模型算法、实现过程，并与传统的 OCR 技术做多组对比实验，对比其响应时间与识别精度等性能指标。

第五章：该章节主要为课题总结。本章对整个课题的成果进行一个概括，并对不满意之处提出未来可行的改进方向。

第 2 章　相关知识与技术介绍

该章节主要介绍本次课题中涉及的相关概念与技术，从实现原理、模型特性与相关背景方面做出较为合理详尽的介绍。

2.1　卷积循环神经网络(CRNN)

卷积循环神经网络(CRNN)根据定义可表面的看作是一种端到端的的网络，因其并在识别的过程中并没有一个切割的过程，所以严格意义上不能算是端到端的网络，而是一种识别网络。其网络结构如图 2-1 所示。

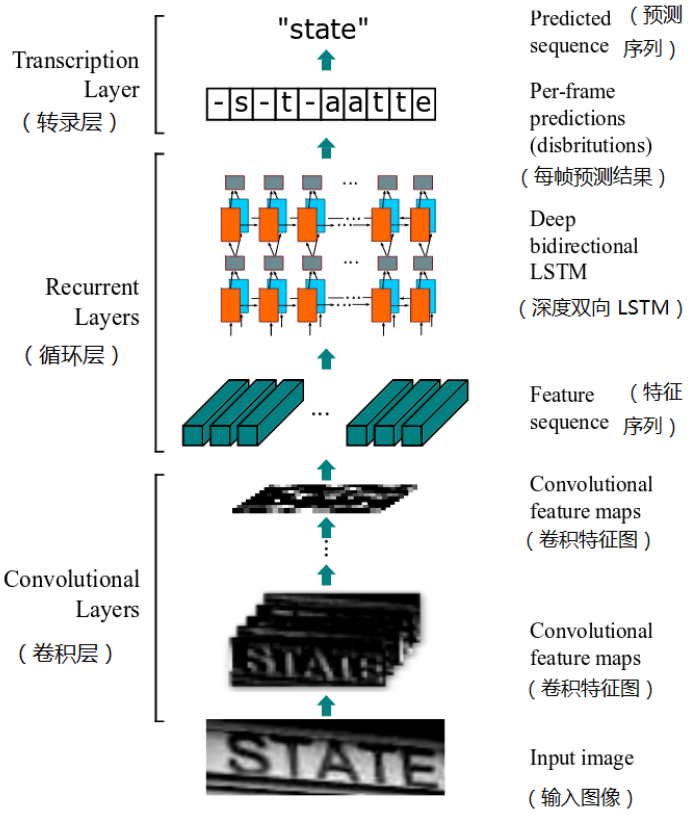

图 2-1 CRNN 网络结构

CRNN 可以理解为是一个卷积神经网络(CNN)+循环神经网络(RNN)的网络，底层的 CNN 进行特征提取，是由图像变为序列的过程。网络为了提高效率，防止参数过多造成错误等问题，将全连接层做了删除。中间为 RNN，

对特征做出预测，是一个把序列转变成结果的过程。最上层为联结时序分类层(Connectionist Temporal Classification,CTC)，主要进行翻译，将单帧预测结果翻译为字母序列。

在卷积循环神经网络的卷积神经网络部分采用了金字塔网络的结构，并且在其网络结构上做出修改与调整，具体如下。

1.在卷积神经网络层面为了可以有效的将提取到的特征输入，输入进循环神经网络中，在卷积循环神经网络中做出修改，把第三和第四个 maxpooling（最大值池化）的核的尺度从 2 乘 2 缩小为 1 乘 2。

2.从优化网络的训练效率层面考虑，卷积循环神经网络在 5、6 卷积层后增加了 BN（批量归一化）层。

从上述的第一点可以看出，为什么将缩小 maxpooling 的核尺寸是为了方便的将卷积神经网络提取到的特征作为循环神经网络的的输入。需要注意网络的输入规格为 W×32（W 代表任意宽度），意味着卷积循环神经网络对输入图片的宽没有特殊的要求，但是输入的高度尺寸都必须修改到 32。

CRNN 中循神经环网络部分是对于卷积神经网络输出的特征序列 $x=x_1$，….，x_t，对每一个输入的特征序列 x 都存在相应的输出 y。为了避免训练时出现的梯度的消失，采用了 LSTM 神经单元作为 RNN 的单元。

CRNN 中 CTC 层可以看作将 RNN 的输出转化为一个字符串，而转化的输入与输出长度不对应而且输入可以是不同长度的序列。也可以看作一个 loss 计算概率到实际输出的概率。

其的损失函数为最小化负对数似然函数，如公式 2-1 所示。

$$\sum_{(x,y)\in\mathcal{D}} -\log p(Y \mid X) \qquad \text{公式（2-1）}$$

2.1.1 卷积神经网络（CNN）

由科学家乐村提出的一种改良神经网络，采用次采样步骤由多层卷积层组成的神经网络，常被用于语言处理，是前馈网络的代表之一。并且因其具有位移不变性，也被称为位移不变网络。

卷积神经网络不同于其他网络，在卷积神经网络中它的各层神经元是一个多维排列的：height、width 和 depth。在这三个参数中 height 和 width 是最容易理解的，毕竟卷积自身便是一个二维存在，但是 dept 在整个网络中指的是激活数据体，它并不是卷积神经网络的深度的深度，网络整体的深度是网

络的层数。而 CNN 在结构的最后将图像概括为包含有评分的向量，这个向量的方向就是沿着 dept 方向的。其结构模型如下图 2-3 所示。

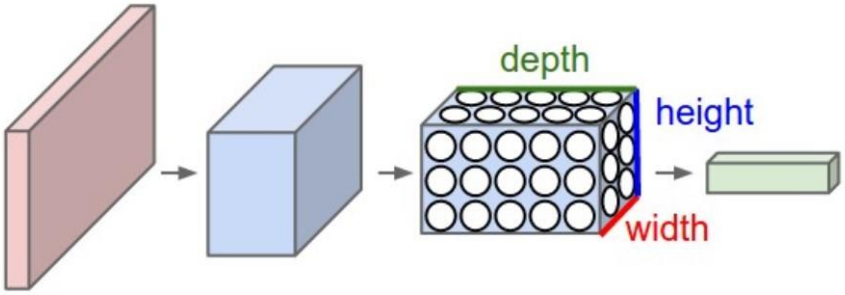

图 2-3 CNN 网络结构模型

卷积神经网络的主体由输入层、卷积层、ReLU 层、池化层和全连接层（全连接层与传统神经网络相同）组成。 通过叠加这些层，构建一个完整的卷积神经网络。 在处理图像等高维输入时，让每个神经元都与上一层的所有神经元完全连接是不切实际的。 相反，只将每个神经元连接到输入数据的局部区域。 连接的空间大小称为神经元的感受野，它的大小是一个超参数（实际上是过滤器的空间大小）。 在深度方向，这个连接的大小总是等于输入深度。

卷积神经网络的优点在于所需参数少，因其并非全连接模式，每个神经元不用与上一层的全部神经元相连，只将每个神经元连接到输入数据的局部区域且一组连接可以共享同一个权重，而不是每个连接都拥有一个不同的权重，这样就缩减了很多参数降低了网络复杂度。在特征提取层使用图像局部相关性的原理，对图像做子抽样，能够有效缩减需处理的数据量并使得有用信息得到保留。通过去掉不重要的样本，进一步减少了参数数量。相比全连接网络会出现的参数过多、层数限制等问题卷积神经网络都可有效避免。因此，在文字识别、图像处理方面，卷积神经网络相比全连接网络具有无可匹敌的优势。

2.1.2 循环神经网络（RNN）

循环神经网络是一种循环式的结构，其将前一步得到的结果进行保留用作下一步的输入，从而形成一种闭合循环结构。

RNN 具有一定的“记忆能力”，对于一个计算结果，它会记下与计算结果有关的信息，且对所有输入使用同一个参数，这源于它对所有输入或隐藏层执行相同的任务来产生输出，这与其他网络大不相同，且降低了参数在复

杂性方面的数值。RNN 基本模型如图 2-4 所示。

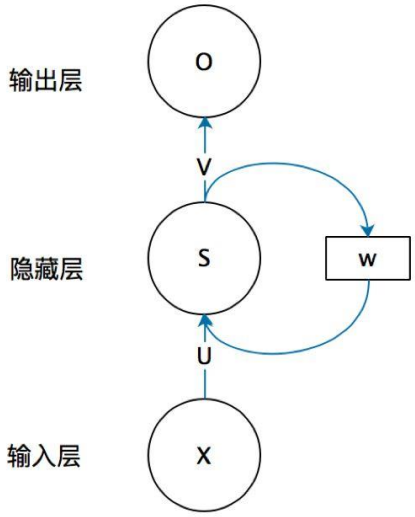

图 2-4 RNN 单层模型

RNN 的思想是利用序列信息。在常规的网络结构中，所有的输入输出可以认为是互不干扰的、独立的。而对于大多数的场景，这种模式是不佳的。如果你需要预测一个句子中的下一个字，知道前边的字会是很有帮助的。在理论上循环神经网络对任何长度的信息做出利用。

然而，当输出的文本序列长度过长，当前节点的位置距离关联信息点较远时，普通的 RNN 或许会出现梯度爆炸或消失的情况。为了处理这类现象，部分网络模型在 RNN 的基础上加入了门控机制，如 LSTM，其内部包含输入门、遗忘门、输出门共三个门控单元和一个记忆单元。在有重要信息输入时，输入门的值将趋近于 1，并被保留在记忆中，当之前的信息变得不在重要，遗忘门的值趋近于 0，网络中将旧记忆替换为新信息。若输入无用信息，输入门的值趋于 0，遗忘门趋于 1，网络保留旧信息，相当于拥有了长期记忆能力。

在实际的应用中，RNN 相比传统的 N-Gram 模型相比具有体量小、处理效果好的优势，在 N-Gram 模型中想要处理一段文本，需要考虑当前处理位置受之前 N 个字词的影响权重，所以为了提高处理效果，往往需要增大 N，而这也会导致，模型体量成指数增长。而 RNN 可以单向的往前或往后扫描任意长度的字词，这就解决了 N-Gram 所会出现的一些主要问题使得 RNN 成为了文字信息处理的理想网络。

2.2 CRAFT 检测网络

CRAFT 检测网络是一种基于字符感知的文本检测方法。CRAFT 检测网络的基本原理是以理想的精准度扫描确定文本中的每个字符，然后把扫描定位到的字符做出连接，使其成为一个整体的文本。因为在 CRAFT 检测网络中是对每个字符做出检测与定位，只用考虑字符间的距离，并没有把关注点放在整个文本上。所以 CRAFT 检测网络不需要大的感受野，而对不规则、畸变的文本也有较好的效果。

在网络结构方面，CRAFT 检测网络使用的是 VGG16-BN 以及类似于 U-net 的结构， CRAFT 网络结构如图 2-5 所示。

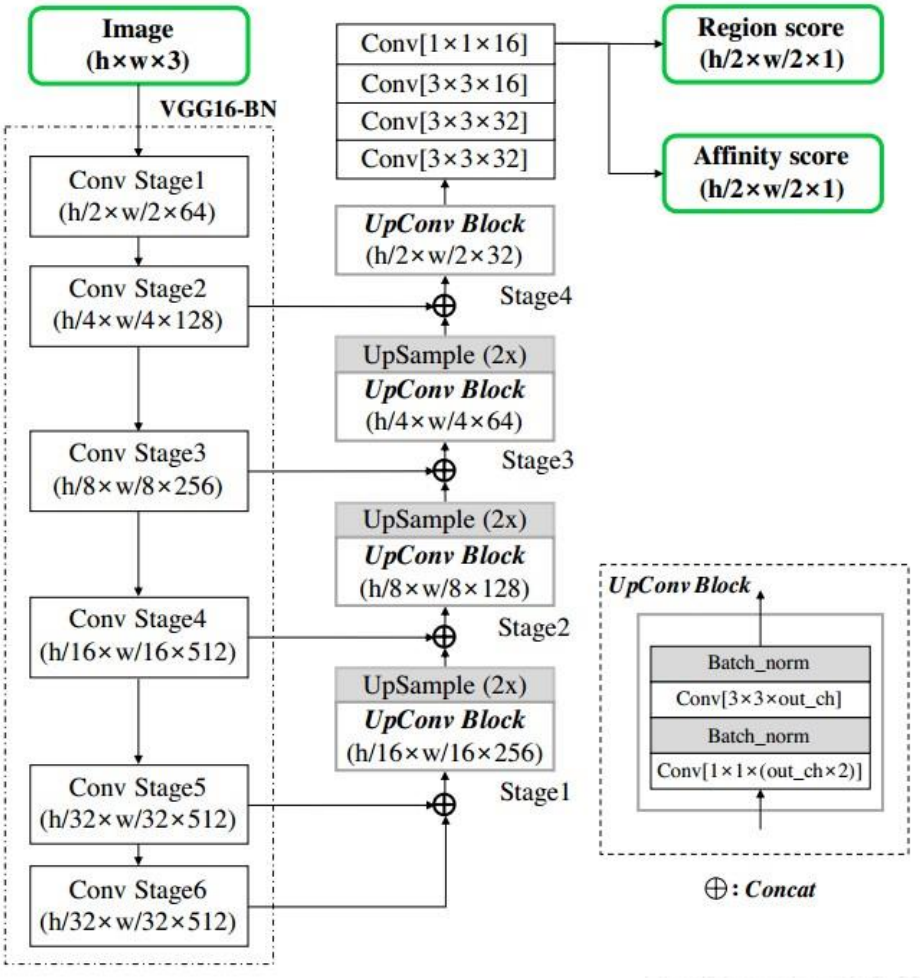

图 2-5 CRAFT 网络结构

CRAFT 检测网络的思想概括如下。

1.图像分割的思想，采用 u-net 结构，先下采样再上采样。

2.并不是常见的像素级别方法，CRAFT 把 character 看作一个目标，而不是一个单词，也就是说 CRAFT 并不执着于整个文本。这与 CTPN 很像，优先检测基础的单位，在根据检测到的单位与其之间的距离做出连接。这种做法的优点在于检测网络只需要对基本的字符做出关注，可以用很小的感受野处理不定长的信息。 而 CTPN 检也是如此，对长度较长的处理对象处理结果相对较好. 和这种思想相对应的方法是 EAST，遭到感受野的限制，对较长文本的处理结果往往表现出两端的效果不理想。

3.对于公开数据集大部分都是文本框级别的标注，而非字符级别的标注，提出了一种弱监督学习思路，先利用合成样本进行预训练，再将预训练模型对真实数据集进行检测，得到预测结果，经过处理后得到高斯热度图作为真实数据集的字符级标签.

若想训练 CRAFT 检测网络，第一步需要生成 region score 和 affinity score 的 ground truth。在 ground truth 的生成过程中没有用到离散化的二进制图，而是使用高斯热图对字符中心的概率进行编码（高斯热图在处理非严格限制的 ground truth 区域时具有高度灵活性），来表示 ground truth。

给定字符框（Character Boxes），region score 和 affinity score 的 ground truth 的生成流程如下。

1.使用字符框生成亲和框（Affinity Boxes）

2.准备一个在二维层面面向各向的高斯映射

3.对每个亲和框与高斯图之间的变化做出计算

4.通过将计算完的高斯图映射到亲和框得到 region score GT（affinity score GT）

亲和框定义：对每个检出出的字符进行绘框，在绘制出的字符框中将对角线做连接形成两个对称三角形，然后两个三角形的中点与相字符框中的三角形中点连接便是一个亲和框。见图 2-6。

绘制亲和框可以有效的确定文本范围内单个字符的权重值分布，进一步细化了识别网络在不规范的或是畸变的文本识别任务中所能提供给后续识别网络的信息，为后续的整体流程提供了帮助。

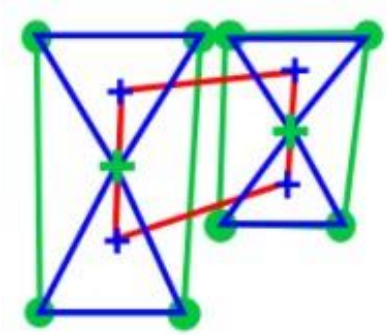

图 2-6 亲和框

通过弱监督训练模型，得到的字符框与实际的字符框存在差距，因而需增加该部分的损失来估量产生的伪 GTs（pseudo-GTs）。其中每个字符注释的置信度图应与检测到的字符数对 ground truth 字符做除法的结果成比例。

弱监督这部分的损失 L 定义如公式 2-2 所示

$$L = \sum_p S_c(p) \cdot \left(\|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2 \right) \quad \text{公式（2-2）}$$

弱监督学习字符分割流程如下。

1.从原始图像裁剪词（行）图像；

2.通过训练的模型中得到 region score；

3.用分水岭算法对字符存在区域做分割，使区域被字符的边界覆盖；

4.对剪切流程做逆变换还原初始坐标。

具体流程如图 2-7 所示。

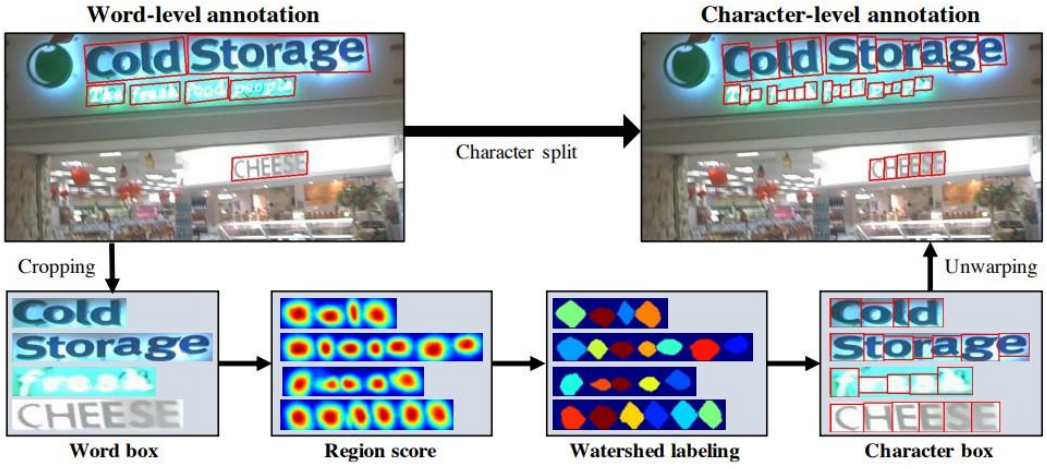

图 2-7 弱监督学习分割字符流程

2.3 序列到序列模型（seq2seq 模型）

序列到序列模型（seq2seq 模型）最先由 Google 引入，seq2seq 属于 encoder-decoder 结构的一种，主要思想为使用两个循环神经网络，分别将其作为编码器与解码器。编码器的功能是将得到的输入序列压缩成指定长度的向量，该向量就可以视作是这个序列的语义，而这个流程即为编码流程，最简单的得到语言向量的方法是把最后输入的隐含状态当作语言向量。或是通过将最后一个隐含状态变换获得。

解码器通过语义向量产生指定的序列，这个流程即为解码流程，最容实现的方式是将编码器获得的语义变量视为初始状态输入到解码器的循环神经网络中，就可以获取输出序列。这时前一刻的输出就会变成这一刻的输入，而且其中语义向量 C 只作为初始状态参与运算，后面的运算都与语义向量 C 无关。seq2seq 网络模型结构见图 2-8。

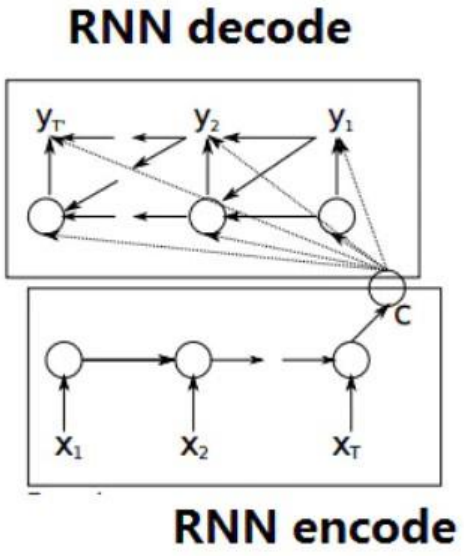

图 2-8 seq2seq 模型

Encoder-Decoder 框架可以说是识别网络中的常客，因其是处理自然环境下的文本，需要使用的一个通用的，重要的框架便是编码器-解码器（Encoder-Decoder）。其中 Encoder 承担接受输入的工作，在接收到输入后生成固定向量，Decoder 负责对 Encoder 生成的向量进行解码并输出对应的结果。由于其对自然场景处理具有着举足轻重的实际功能意义，所以在各种识别网络中常有出现。

2.4 光学字符识别技术（OCR）

OCR （Optical Character Recognition，光学字符识别）是指用仪器（例如文本录入器）扫描纸质载体上存在的字符，经过明、暗处理得到其轮廓，然后通过一种文字识别算法把检测到的轮廓信息转换成文字的流程，也就是对于纸质载体上的字符应用光学处理手段把文本处理为灰度矩阵的形式，并通过识别工具将图像中的文字转换为信息化的文本，以便利后续文字录入工等作再继续进行。怎样校正错误识别与提高准确度，成为 OCR 的主要研究课题，ICR（Intelligent Character Recognition）这个名词也是因此才得以诞生。评判一个光学字符识别算法的优秀与否主要从响应时间、准确度等性能指标出发。

目前主流的 OCR 识别流程如下。

1.检测出图像中文本存在的区域

2.对文字区域矩形分割，将其拆解成各个字符

3.对拆解出的字符做出分类

4.识别出文字（最终识别出整个字符）

5.后期的处理校正，主要对识别出的字符与近似字符做出校正。

本章小结

本章主要对卷积循环神经网络、卷积神经网络、循环神经网络等一系列与课题相关的技术与相关背景以及编码器与解码器的概念和部分模型做出介绍，其中重点介绍了卷积循环神经网络(CRNN)的网络结构与相关构成；介绍了 RNN 网络的结构特性与相关重要公式；介绍了 CRAFT 检测网络、seq2seq 模型下的 encoder-decoder 框架和 OCR 传统文字识别技术。

第 3 章　基于 CCRW 的自然场景文字识别研究

通过上文的论述，面对自然场景文本识别检测领域的现状和存在问题的研究，本章主要提出并实现了一个基于卷积循环神经网络与一种字符级检测网络的自然场景文本检测识别方法(CCRW)，具体内容将涵盖研究动机、模型设计、实验细节、实验结果等方面进行论述。

3.1　研究动机

在文字识别发面，早期的重点是通过识别器识别打印文本并对器进行数字化录入。在这种背景下，OCR 技术随之诞生，随后在长时间的改进与优化中 OCR 技术以得到了极大的完善。尽管 OCR 技术在文字识别领域已经非常成熟，可以非常理想的处理文档信息数字化等常见的文本识别任务，但在处理自然场景下背景复杂的文本识别任务时往往无法得到理想的效果。

自然场景中文本识别的难点在于以下几点。

1.环境复杂、不够单一 自然场景中文字信息所处的环境往往是不单一的，甚至是高度紊乱的。在这种情况下极可能破坏信息文本的纹理与形状，有些背景元素也有可能表现出与信息文本相似的特征，不便区分，且大都包含丰富的色彩与修饰等干扰检测识别的元素，最终导致关键信息遭到干扰，无法准确的检测识别文本。

2.信息文本范围不规范、字体多样 在文档识别任务中出现的文本字体大都为打印体等规范字体，字体特征明显，颜色单一规范。而在科技与生活文化蓬勃发展的当下，自然场景中的文本往往为了起到吸引眼球或其他美化目的，往往在文本形状、字体上是不规范的、风格化的甚至是丧失统一性的。这些文本与传统文档字体大不相同，由于它们自身包含太多特殊信息，常常在识别过程中无法准确定位范围与边界。

3.光影影响 自然环境中的文本，如灯红酒绿的步行街，高楼林立的城市街景等，都存在着大量的光影干扰，这些不稳定的光线对文本存在范围的照射会使得文本的边界模糊、颜色信息失真。

不同环境文本对比如图 3-1 所示。

而本次课题与其相关工作正是以找寻一种在任何场景下都可以有较高精度的文字识别效果的文字识别方法为目的所展开的。

图 3-1 不同环境下的文本对比

3.2 CCRW 方法原理简述

本次课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法(CRAFT and CRNN text recognition method under weakly supervised learning ,CCRW)流程为首先在检测网络部分通过字符级检测网络对包含文本信息的图像做处理，确定文本信息在图像中的具体位置并将信息传入识别网络。然后在识别网络部分通过 CRNN 提取文本的特征信息，同时得益于 RNN 的特性，该方法不需要对文本的长短做出判断，可以直接处理。在通过 seq2seq 模型进一步提取得到的特征信息的序列特征。

由于 CCRW 在检测网络部分采用了字符级的检测网络，是以较高精度定位每个单一字符，后再将每个单一字符做连接得到整个文本。在整个将测网络中不需要考虑文本的整体范围，只用考虑单一字符之间的距离，故而不需要很大的感受野，优化了性能，在处理弯曲、变形或者极长的文本都可以表现出良好的效果。在识别网络部分使用了 CRNN+seq2seq 模型，可以有效的根据上下文语义做出预测，能够对识别结构与检测结果做出有效监督与矫正，并且由于网络为端到端的形式，不需要对输入做出过多修改便可直接处理。综上，整个 CCRW 方法表现出网络体量小、配置要求低、响应时间短的特性。

3.3 模型建立

基于卷积循环神经网络与一种字符级检测网络的文字识别方法，针对自然场景文字识别的数据特点，融合了 seq2seq 模型设计了整体流程。从整体

上看，该方法可分为六个部分：图像前期预处理、导入检测网络检测文本范围、卷积循环神经网络提取图像特征、seq2seq 模型进一步提取卷积特征中的序列特征、根据得到的特征进行反向监督与文本校正、内容输出。具体流程见图 3-2。

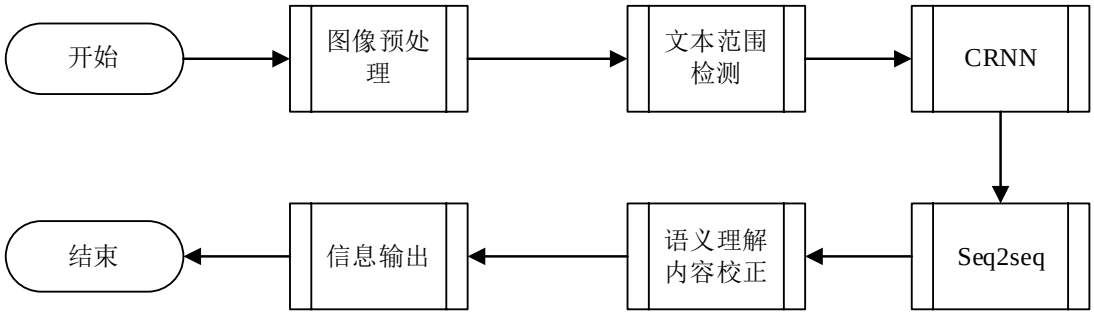

图 3-2 总体识别流程

seq2seq 循环神经网络中的一个重要模型，也称为 Encoder-Decoder 模型，可以看作是一个 N 乘 M 的模型。seq2seq 模型由两部分构成：Encoder 处理编码序列的相关信息，把不定长的信息处理到一个向量中。而 Decoder 作为解码器，在获取到解码器处理的信息向量后对其进行解码，然后以序列的形式输出。在 seq2seq 模型中，编码器采用堆叠的卷积神经网络进行静态特征提取，得到图图像的特征序列，然后实验 GLU 将得到的特征序列进行编码，将编码结果传入到 Decoder。Decoder 部分用 BiLSTM 和注意力机制构成，在对传入信息解码后，输出识别出的信息结果。模型框架见图 3-3。

本次课题在 CCRW 方法中引入 seq2seq 模型，主要目的为提高 CCRW 方法的整体识别精度，因为 seq2seq 模型可以根据文本的前置信息来对后续可能出现的信息做出预测，而预测结果既可以用来获得结果补全对话，亦可以用来反向监督 CCRW 流程前期的检测网络的输出结果。前者适用于文本缺失或是极度污损的情况，理论上可以在一定程度上根据预测结果对文本缺失内容做出修补。而后者可以通过预测结果多次反向监督，在极大程度上收束 CCRW 的检测网络所用作约束的权值，以达到提高识别精度的结果。

从 CCRW 的整体流程中看，seq2seq 模型发挥着至关重要的功能，是该方法不可缺失的一部分。

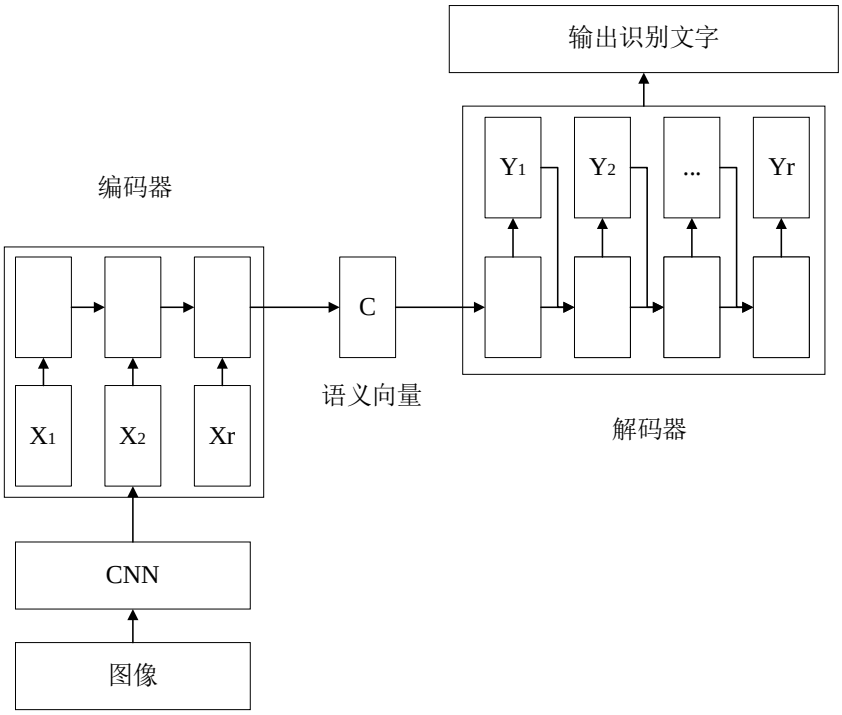

图 3-3 seq2seq 模型结构

3.3.1 数据预处理

在整个结构化分析中，采用自顶向下的分析方式，通过确定外部实体关系，归纳处理过程，寻找数据流向，使功能逐次分解，

图像的预处理首要的内容是同一图像尺寸、转换为灰度图并进行归一化。详细的说，就是对采集到的包含文本信息的自然场景下的图像进行预处理操作，通过对图像的 resize 和填充等操作流程将图像调整到适用于卷积神经网络的输出尺寸大小，做出统一的图像使得数据训练环节更加容易。影响图像识别的主要因素是梯度因素与特征因素，可以用此提高特征因素与梯度因素的提取精度，使得识别流程的处理时间极大的缩减。

综上所述，图像的预处理将分为灰度变换、灰度归一化、图像 resize 三个流程，经处理得到规格等尺寸的灰度图，整个流程如图 3-4 所示。

图 3-4 图像预处理流程

1.转换原始图像为灰度图 在绝大多数场景中，图像为了凸显出自身的信

息，吸引眼球，大都以彩色的形式出现，可以说色彩也是图像所包含的一种信息，但计算机在做文字识别的流程中并不会使用到图像的色彩信息。且在输入原始图像进行灰度图转换的过程中，图像是通过丢弃色彩信息并用灰度表示图像对应区域的亮度来决定图像的灰度，借此来减少无关参数的影响，提升模型的运行速度。在预处理的开始阶段，首先要将初始图像转换为灰度图，这一过程是为了使颜色分量 R，G，B 相等，采用加权平均值发进行计算。

2.归一化操作 灰度处理后，原始图像由彩色图转变为了灰度图，由于自然环境的复杂，图像在取景时得到的背景往往也是复杂的，其取景过程容易收到天气、光照量等多种因素影响，很有可能导致图像的灰度集中在一个或多个区段，会产生由彩色图转换为灰度图后依旧模糊。这时可以采用灰度拉伸的方式将其拓展为 256 灰度级。归一化的运算如公式 3-1 所示。

$$g(i,j) = \frac{f(i,j) - \min}{\max - \min} * 255 \qquad \text{公式（3-1）}$$

3.图像 resize 在统一图像尺寸的环节，可以采用双线性内插算法通过归一化处理后的灰度图四周的像素值计算缩放处理后的像素值。

图像灰度转换与双线性内插处理的效果如图 3-5 所示。

a(灰度转换)　　b(双线性内插)

图 3-5 灰度转换与双线性内插处理的效果

3.3.2 文本范围检测

文本范围检测阶就是通过检测网络，判断识别出图像中包含的文本所在的范围，并生成范围坐标告知后续网络，以作识别等操作。在本次课题中使用的检测网络是一种字符级别的检测网络-CRAFT 检测网络。

网络的核心思想如下。

1.图像分割的思想，采用 u-net 结构，先下采样再上采样。

2.并不是常见的像素级别方法, CRAFT 把 character 看作一个目标，而不是一个单词，也就是说 CRAFT 并不执着于整个文本。这与 CTPN 很像，优先检测基础的单位，在根据检测到的单位与其之间的距离做出连接。这种做法的优点在于检测网络只需要对基本的字符做出关注，可以用很小的感受野处理不定长的信息。而 CTPN 检也是如此，对长度较长的处理对象处理结果相对较好. 和这种思想相对应的方法是 EAST, 遭到感受野的限制，对较长文本的处理结果往往表现出两端的效果不理想。

3.对于公开数据集大部分都是文本框级别的标注，而非字符级别的标注，为此提出了一种弱监督学习思路，先利用合成样本进行预训练，再将预训练模型对真实数据集进行检测，得到预测结果，经过处理后得到高斯热度图作为真实数据集的字符级标签。

网络采用 VGG-16 的 backbone 做向下采样，该流程一共将进行 5 次，故而再将图像传入网络前的预处理阶段,将会对输入图片的长和宽进行处理为最近似为 32 的倍数的整数尺寸。例如输入图片原始尺寸为 700x600，则会再预处理阶段处理到 704x608，经过这样的预处理后可有效避免分割时会发生的像素漂移。利用 Unet 的思想，对下采样的特征图再进行上采样和特征图 concat 操作，最终获得尺寸为原图大小二分之一的两个通道特征图：region score map 和 affinity score map，分别为单字符中心区域的概率和相邻字符区域中心的概率。网络结构图见上文图 2-5。

CRAFT 要进行训练首先需要生成 region score 和 affinity score 的 ground truth。生成 ground truth 使用的不是离散化的二进制图，而是使用高斯热图对字符中心的概率进行编码（高斯热图在处理非严格限制的 ground truth 区域时具有高度灵活性），来表示 ground truth。

给定字符框（Character Boxes），region score 和 affinity score 的 ground truth 的 生成流程如下所述。

1.使用字符框生成亲和框（Affinity Boxes）

2.准备一个在二维层面面向各向的高斯映射

3.对每个亲和框与高斯图之间的变化做出计算

4.通过将计算完的高斯图映射到亲和框得到 region score GT(affinity score GT)

生成了 Ground Truth，还无法进行训练，生成 Ground Truth 的前提条件是要有字符框，但是在真实的数据集中一般都是词（行）注释，没有字注释，所以需要把词（行）注释转换成字注释。于是提出弱监督学习框架使用词（行）

注释来估计字注释，再生成 ground truth。

通过弱监督训练模型，得到的字符框与实际的字符框存在差距，所以需要在训练时增加这一部分损失来衡量生成的伪 GTs(pseudo-GTs)。每个词（行）注释的置信度图与检测到的字符数除以 ground truth 字符的数量成比例。

对于训练数据的词（行）注释样本 w，令 R(w) 代表边框区域 l(w)代表单词的长度。获得估计的字符长度之后，样本 w 的置信分数计算如公式 3-2 所示。

$$s_{\text{conf}}(w) = \frac{l(w) - \min(l(w), |l(w) - l^c(w)|)}{l(w)} \quad \text{公式（3-2）}$$

一张图像的像素级置信图计算如公式 3-3 所示。

$$S_c(p) = \begin{cases} s_{\text{conf}}(w) & p \in R(w) \\ 1 & \text{otherwise} \end{cases} \quad \text{公式（3-3）}$$

其中 p 代表在区域 R（w）中的像素。则弱监督的损失 L 定义如公式 3-4 所示。

$$L = \sum_p S_c(p) \cdot \left(\|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2 \right) \quad \text{公式（3-4）}$$

弱监督学习的字符分割流程如下。

1.从原始图像裁剪词（行）图像；

2.将裁切好的图像输入网络；

3.通过训练的模型中得到 region score；

4.用分水岭算法对字符存在区域做分割，使区域被字符的边界覆盖；

5.对剪切流程做逆变换还原初始坐标。

在字符分割结束后，检测网络通过 Ground Truth 的生成流程来生成伪 GTs。

如果置信分数低于 0.5，预测出的字符边界框应该忽视，边界的置信分数过低时会对模型的训练产生负面影响。在这种情况下，可以把单个字符的宽度处理为常数，并通过使用字符区域 R(w)除以字符数 l(w)的方法计算字符集预测。然后设置为置信分数低于 0.5 的为不学习的忽略文本。

具体流程如图 3-6 所示。

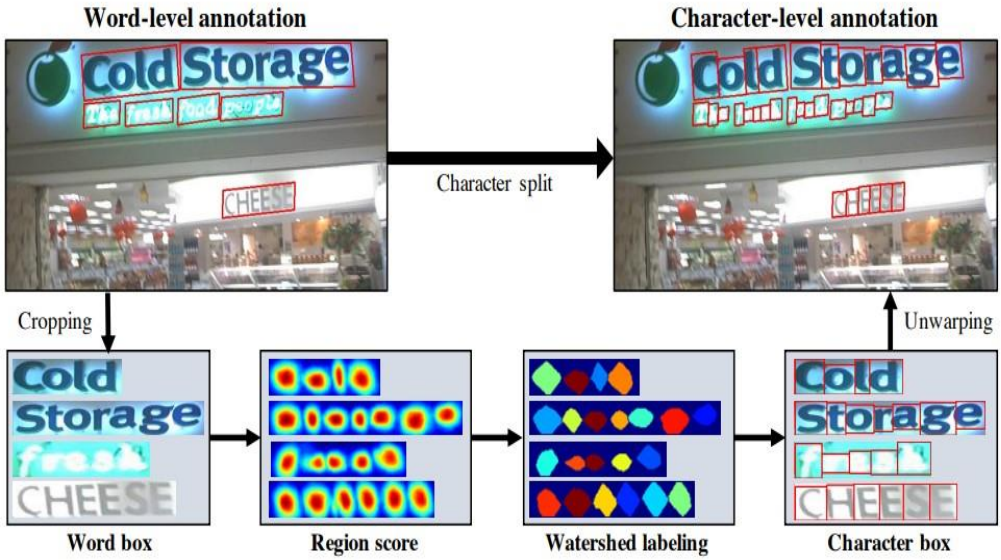

图 3-6 弱监督字符分割流程

在后处理阶段，需生成多边形框来处理不规则的曲线文本，首先，沿扫描方向找到字符区域的局部最长线（图 3-7 蓝线），局部最长线的长度都设置为它们中的最大长度。然后，连接全部局部最长线的中点生成中心线（图 3-7 黄色），并且旋转局部最长线以垂直于中心线（反应字符的倾斜角度（图 3-7 红线））。局部最长线的端点是文本多边形的候选点。最后，为了完全覆盖文本区域，沿着中心线（黄线）向外移动两个最外边的局部最长线（红线），形成最终的控制点。连接全部的控制点，组成多边形。

通过示意图可以直观的看出本次课题使用的检测网络具有高度的灵活性与泛用性，可以用理想的效果来处理文本畸变以及文本不规范的自然场景识别任务。而在确定了不规范网络的文本范围后，也可以准确的更具绘框范围继续进行亲和框的绘制，对范围内的每个字符做出权值分布定位，这样就进一步细化了识别网络在不规范的或是畸变的文本识别任务中所能提供给后续识别网络的信息，对整个 CCRW 方法在处理这类问题时的识别精度做出了补充。

而这也突显出了 CCRW 在处理曲线文本等一系列复杂问题时具有一定的应对性，并不是将其与其他场景一概而论的处理。

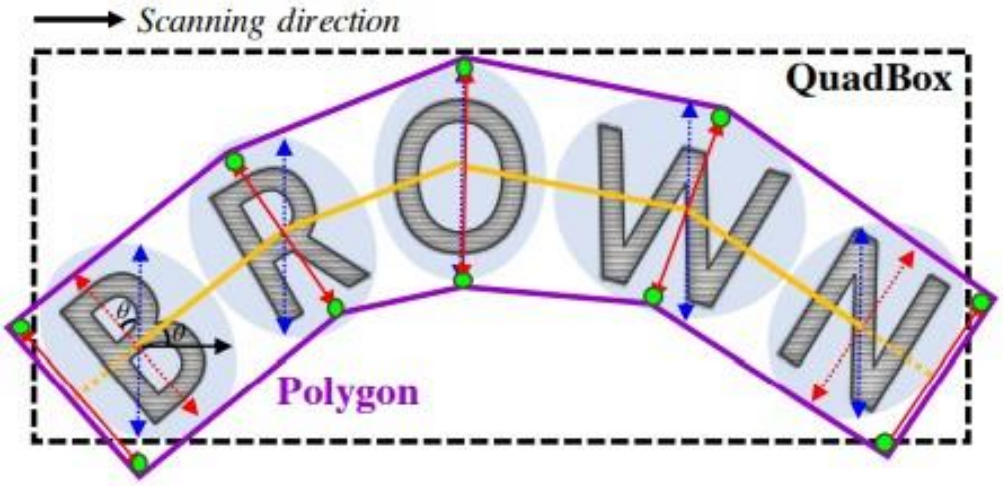

图 3-7 处理曲线文本

实际运行效果如图 3-8 所示。

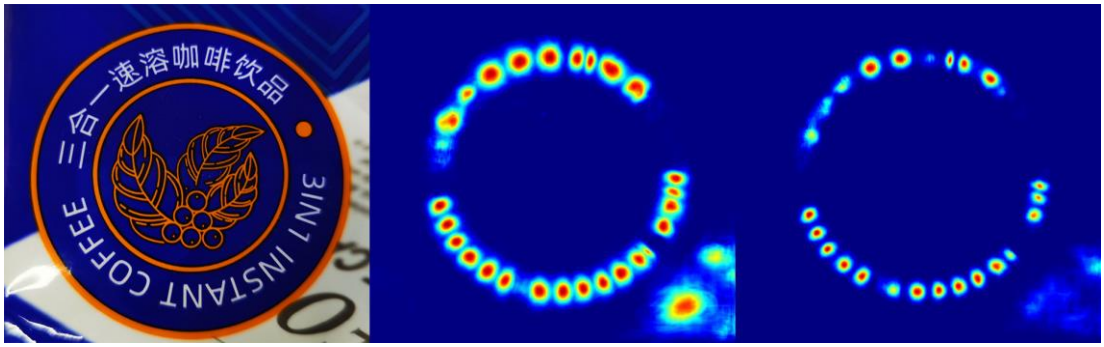

图 3-8 实际运行效果（热力图）

在 CRAFT 检测网络中为了在上一步得到的检测结果上进一步获得单独的绘框，可以增加一个网络对连接做出细化处理，也就是 LinkRefiner。

在训练过程中，仅使用 CTW1500 数据集训练 LinkRefiner。该网络输入是 region score，affinity score 和 原始 CRAFT 模型 Stage4 的 feature map。采用 ASPP 来确保可以把同一行的字和词组合到一个大的感受野中。该网络输出是 link score，用来代替 affinity score。然后使用上述的生成流程生字符绘框。CRAFT 模型定位单个字符，LinkRefiner 模型将字符以及有空格分隔的单词组合在一起，用来做 CTW1500 评估。

LinkRefiner 结构如图 3-9 所示。

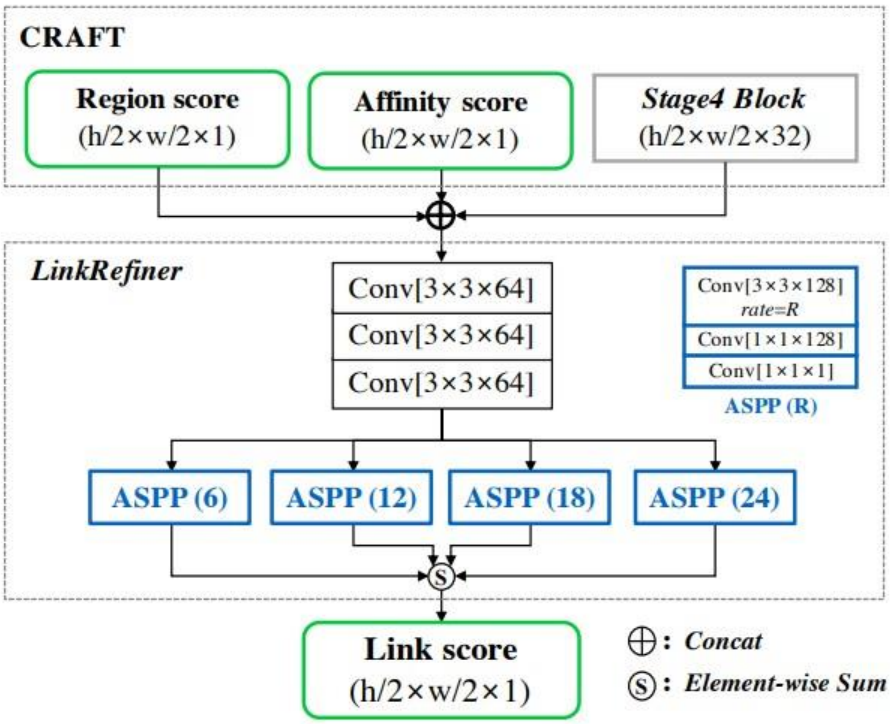

图 3-9 LinkRefiner 结构

本章小结

该章节首先介绍了本次课题的研究动机，并在研究动机的基础上提出了 CCRW 识别方法。在后续的篇幅中主要介绍可基于卷积循环神经网络所提出的 CCRW 方法。在该方法中首先确认了以一种可以在字符级别检测文本范围的检测网络作为 CCRW 的检测网络并确认了以 CRNN 与 seq2seq 模型组合而成的识别网络。并在理论上对 CCRW 网络的可行性提出了详尽的分析与支持。

第 4 章　实验验证与分析

本章节主要为本次课题提出的一种以弱监督方式训练的字符级检测网络构成的文字识别方法在不同场景中的实现，验证各个环节的稳定性、论述其模型算法、实现过程，并与传统的 OCR 技术做多组对比实验，对比其响应时间与识别精度等性能指标。

4.1　实验目的

本次实验的目的为验证课题在找寻一种基于机器视觉的能够在自然环境下用较高精度处理文本识别任务的文本识别方法这一诉求下提出的 CCRW（基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法）的可行性并通过建立对照实验对比验证 CCRW 与机器视觉领域传统的 OCR 技术所搭建的基线方法在识别效果、方法稳定性、泛用性上是否存在优势、是否能够胜任自然场景下的负载背景文本识别任务。

4.2　实验过程

本次实验将采用对照实验的模式进行，的具体流程如图 4-1 所示。

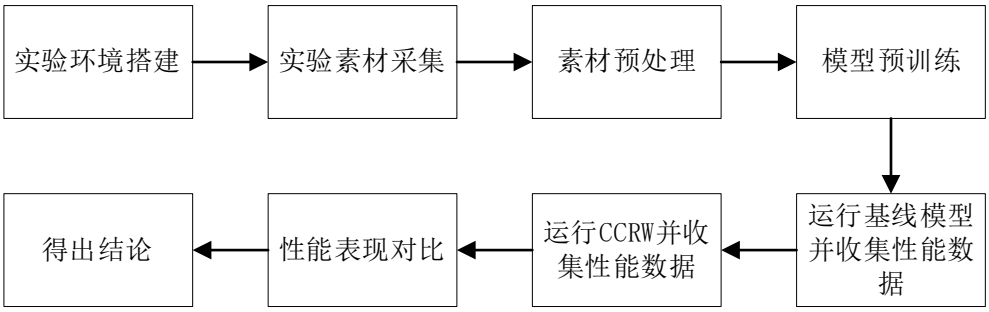

图 4-1 本次实验流程图

4.2.1 实验环境搭建

本章节旨在通过建立对照实验对比本次课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法与传统 OCR 技术进行对比，验证该方法是否可行并在特定场景下具有优越性。

实验过程中使用的 OCR 流程为：使用 OpenCV 开源库进行图像的预处理操作，在对原始图像处理后对其进行 OCR 识别，该方法下数据集的训练

采用开源训练工具进行训练。

而课题提出的方法的数据集训练采用深度学习模型进行训练，因为数据集文本量庞大，在训练过程中需要使用 GPU 加速，故模型的测试选择在 Torch 框架下配合英伟达公司的 CUDA11 技术加速 GPU 完成。具体实验环境如表 4-1 所示。

表 4-1 实验环境

	实验环境配置
内存	16G
CPU	Intel(R) Core(TM) i7-8550U @ 2.00 GHz
GPU	NVIDIA MX150
开发语言	Python
图像预处理	OpenCV
深度学习框架	TensorFlow、Torch

4.2.2 素材采集

本次实验的素材主要采集自互联网、实际文本扫描以及实际摄影设备对显示街景的拍摄。

为了保证最终实验结果的可靠性与可信度，素材采集阶段在不影响实验结果与走向的前提下对素材的采集做出了严格的筛选，采集到用来做训练集的素材全都字迹清晰且多样， 可以有效的训练出特征多样的数据集，用作效果验证的素材在大部分符合课题所要解决的自然场景下复杂背景识别任务的要求的同时也准备了单一、简单背景的文本图像作为对照组来验证 CCRW 在处理自然场景之外的其他识别任务的效果，以此来验证 CCRW 是否在特定场景之外的任务中具有泛用性。

在符合上述要求的基础上，借助网络爬虫、实际摄影设备拍摄等技术手段共采集素材图像 62 万张，并在自主采集的 62 万张素材库中选取 5000 张图像作为训练集供模型训练，然后随机抽取 50 张图片进行测试，为了保证模型的泛化性，设有备选测试集，保证实验结果的客观性，并将两种方法的实验结果建立对照来论证设想。

采集到的素材图像部分如图 4-2 所示。

图 4-2 实验素材部分

4.2.3 素材预处理

在完成素材的采集与筛选后，我们还需要对素材做出预处理。

图像的预处理首要的内容是同一图像尺寸、转换为灰度图并进行归一化。通过对图像的 resize 和填充等操作流程将图像调整到适用于卷积神经网络的输出尺寸大小，做出统一的图像使得数据训练环节更加容易。影响图像识别的主要因素是梯度因素与特征因素，可以用此提高特征因素与梯度因素的提取精度，使得识别流程的处理时间极大的缩减。

灰度处理后，原始图像由彩色图转变为了灰度图，由于自然环境的复杂，图像在取景时得到的背景往往也是复杂的，其取景过程容易收到天气、光照量等多种因素影响，很有可能导致图像的灰度集中在一个或多个区段，会产生由彩色图转换为灰度图后依旧模糊。这时可以采用灰度拉伸的方式将其拓展为 256 灰度级。归一化的运算如公式 4-1 所示。

$$g(i,j) = \frac{f(i,j) - \min}{\max - \min} * 255$$

公式（4-1）

鉴于编码阶段对本课题所编写的 CCRW 方法的网络训练方式以及后续数据集建立的便利性考虑，本次实验的预处理相对于传统的计算机视觉领域的预处理流程多出一个步骤。

我们需要将预处理后的图像文本信息提取出来生成相应的信息文本并与

处理后的图像采用一一对应的形式存放，存放形式如图 4-3 所示。

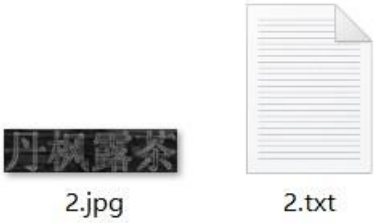

图 4-3 存放形式

4.2.4 模型预训练

采集到的初始图像在经过预处理转换为灰度图并归一化后，在计算机的视觉内会变得清晰，对比度更加明显，除此之外还可以减少阐述计算量，方便后续的数据处理。

为了确定模型各参数值，首先在数据集中选取标签样本进行预训练，减少同时训练的参数并降低模型的计算量，起到避免模型出现收敛速度过慢的情况，并在训练中对导入参数不断优化，以尽可能达到理想的识别效果，同时提高模型的训练速度。

在训练本次课题提出的识别方法时需要手动建立训练集表，训练集表为.txt 格式文本文件，内部信息格式为目标图像路径+目标图像包含信息中间用空格分开，具体如图 4-4，4-5 所示。

图 4-4 原始素材

data/train/line-081/1001.jpg 道又不干我事与我无干" 武松道

图 4-5 内部格式

在训练时可以采用指数衰减法提高收敛速度，通过将学习率初始值选定为 0.01，然后通过数据迭代不断更新学习率，这样可以在算法多次迭代后确定一个最优数值。模型在训练完成后可使用之前备用的数据集验证其泛化性。

在模型训练完成后，将其导入需要建立对比实验的两种文字识别方法中，并测试多组相同的测试图像，得到响应时间、预测率等性能参数。

需要注意的是在训练模型时需考虑不同的卷积神经网络中不同 loss 的影

响，不同 loss 在 MATLAB 的展现如图 4-6 所示。

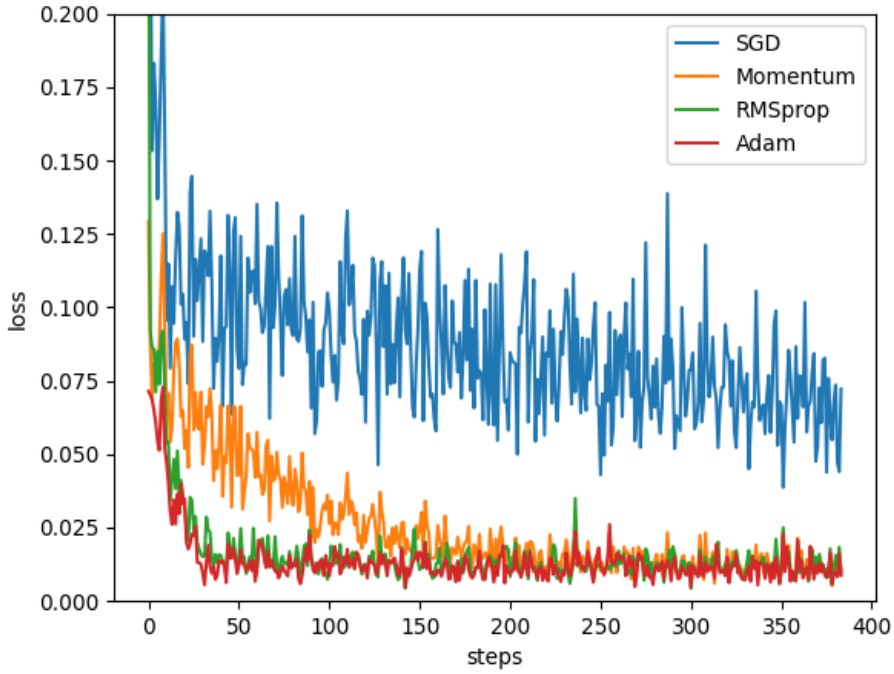

图 4-6 不同 loss 在 MATLAB 的展现

4.2.5 运行 CCRW 与基线模型(OCR)并收集数据

根据对照实验的准则，在严格符合实验要求的情况下将本课题提出的 CCRW 与根据传统 OCR 技术搭建的基线模型导入搭建好的实验环境中并使导入自制测试集进行实验并收集两种方法的性能数据。

经过实际实验，可以明显发现在处理如文档扫描等背景信息单一、干扰度低的文本识别任务时 CCRW 与基线模型都表现出了良好的效果，且基线模型凭借着 OCR 技术的高度成熟性在平均响应时间长来到了接近 200ms 的水平，优于 CCRW 的 240ms 的平均响应时间。在识别准确度上稳定在 93%到 94%之间。而 CCRW 的准确度稳定在 92%到 93%之间，两者没有太过明显的差距。

用作例举的单一背景文本如图 4-7 所示。

CCRW 处理背景单一的文本时的实际运行效果如图 4-8 所示

基线模型(OCR)处理单一背景文本时的实际运行效果如图 4-9 所示。

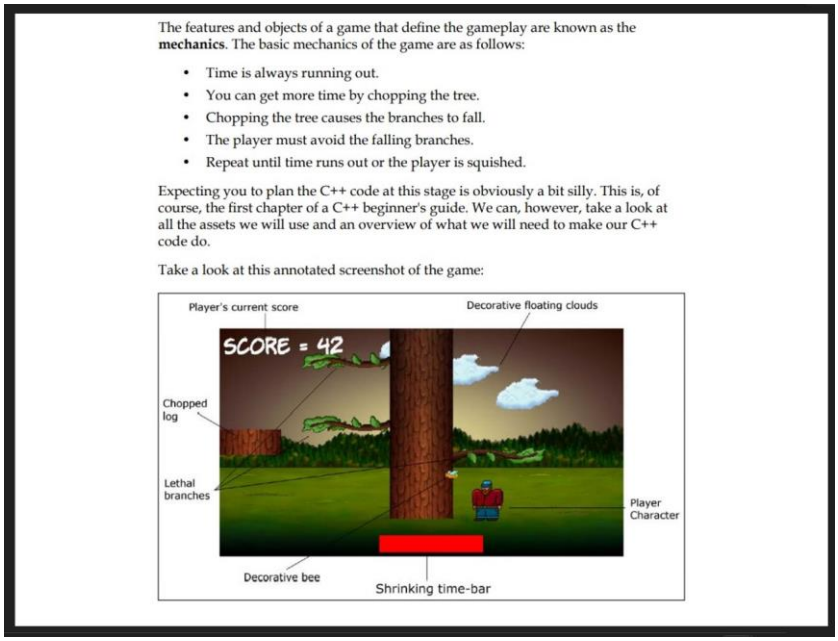

图 4-7 用作例举的的单一背景文本

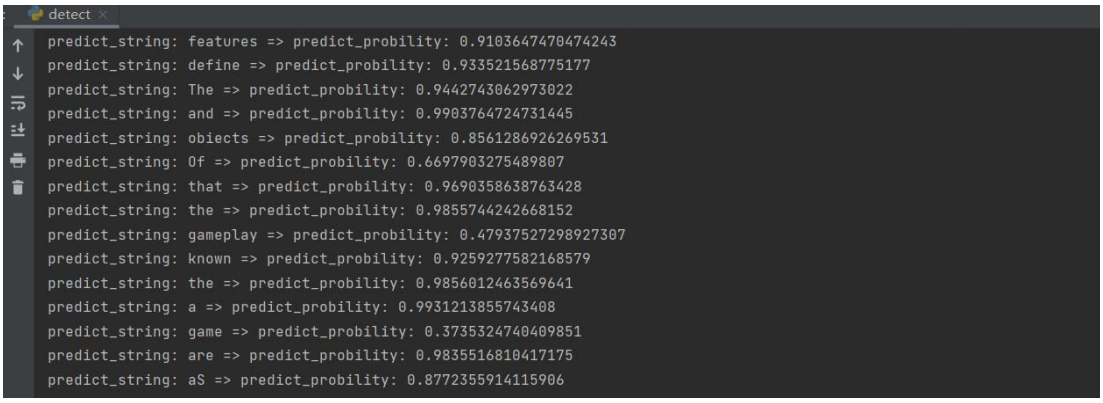

图 4-8 CCRW 处理效果

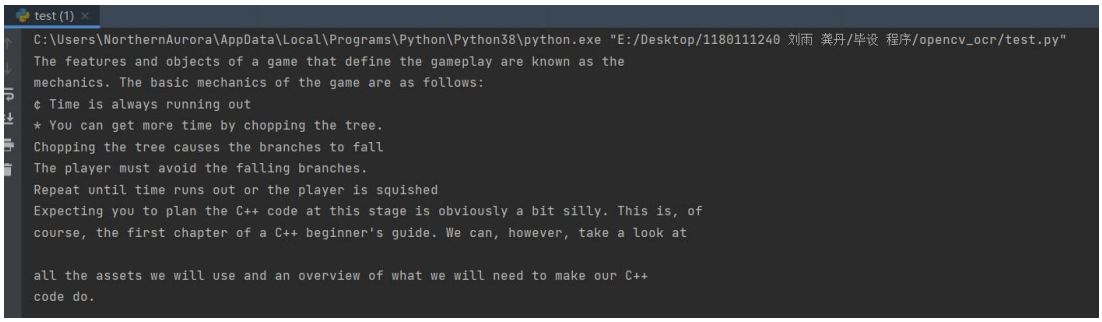

图 4-9 基线模型处理效果

为了直观展现两者的处理效果，制作了表 4-2

表 4-2 两者处理效果

	原文	处理效果
CCRW	• Time is always running out.	• Time is always running out.
基线模型	• Time is always running out.	$Time is always running out.

经过实际实验，在处理自然场景下的文本识别任务时由于文本环境呈现出复杂多样的特征，执行处理时间在理论上应该得到增长，但是在实际实验的 CCRW 与基线模型的执行情况反馈来看仅有 CCRW 的执行时间有了明显的增长而基线模型(OCR)的执行时长虽也有所增加，但是相对于 CCRW 的增长数值来说基线模型的耗时增长并不明显。CCRW 执行耗时如图 4-10 所示，基线模型(OCR)执行耗时如图 4-11 所示。

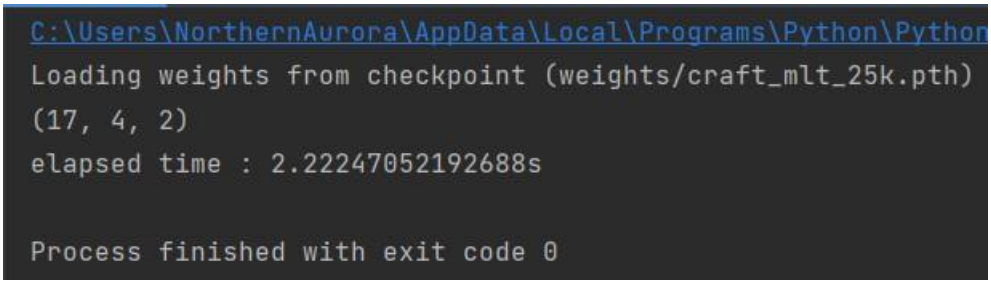

图 4-10 CCRW 执行耗时

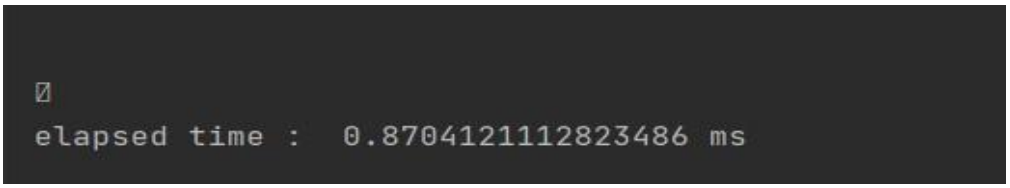

图 4-11 基线模型执行耗时

于是对 CCRW 与基线模型的执行结果做出检验，发现基线模型受限于 OCR 技术遇到无法检测的信息即会将其丢弃这一特性对实验素材图像中无法识别的信息做出了丢弃，而因为实验素材背景过于复杂，基线模型(OCR)几乎对所有信息都做出了丢弃，所以基线模型的执行耗时没有明显增长。但是由于丢弃了绝大多数的文本信息导致了基线模型的最终执行效果几乎不具备任何信息于意义，就文本识别或信息提取者方面的目的来说，其执行结果可以认为是毫无意义的。

而 CCRW 虽然执行耗时有了明显增加，但是仍然可以理想的检测并识别出场景中的文本信息，且程序运行稳定，可以认为其执行结果具有文本识别、信息提取方面的意义。

用作例举的自然环境下复杂背景文本如图 4-12 所示。

基线模型处理自然环境下图像的执行效果如图 4-13 所示。

CCRW 检测自然环境下图像的执行效果如图 4-14 所示。

CCRW 处理自然环境下图像的执行效果如图 4-15 所示。

图 4-12 用作例举的自然环境下复杂背景文本

图 4-13 基线模型处理自然环境下图像的执行效果

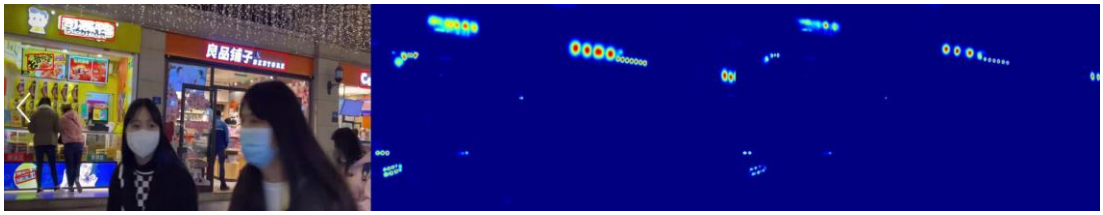

图 4-14 CCRW 检测自然环境下图像的执行效果

```
predict_string: 周黑鸭 => predict_probility: 0.94765836062579181
predict_string: 良品铺子 => predict_probility: 0.99263272734234723
predict_string: 去骨 => predict_probility: 0.96723742384238428
predict_string: 鸭掌 => predict_probility: 0.98932832834289342
```

图 4-15 CCRW 处理自然环境下图像的执行效果

为了直观展现两者的处理效果，制作了表 4-3

表 4-3 两者处理效果

	原文	处理效果
CCRW	周黑鸭|良品铺子|去骨鸭掌	周黑鸭|良品铺子|去骨鸭掌
基线模型	周黑鸭|良品铺子|去骨鸭掌	欢恤|黑莺|薪政|僵 0 政 .溶_

4.2.6 实验结果分析

通过实际实验，对多组实验素材的实际运行，可以发现限于设备算力与在校用电时间等客观因素影响，训练的数据集训练量无法达到最优，在采集的 62 万文本中仅挑选了 5 千文本参与训练,使得在环境背景单一的文本识别测试中本课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法对比传统 OCR 技术并没有表现出明显的优势，具体性能参数如表 4-4 所示所示，测试集样张如图 4-16 所示。

表 4-4 处理单一、复杂场景测试集

	测试集识别平均精度	平均耗时
本课题提出的方法 (单一场景)	92.1%	232ms
传统 OCR 技术（单一场景）	93.5%	214ms
本课题提出的方法（复杂场景）	97.5%	1013ms
传统 OCR 技术（复杂场景）	15.2%	238ms

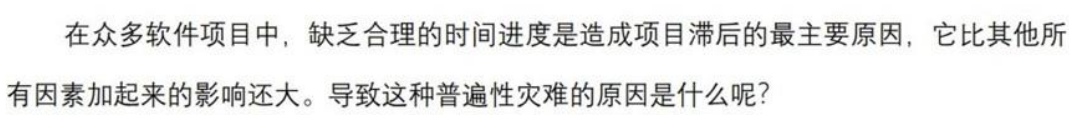

图 4-16 单一背景测试图像样张

而在测试多组自然环境下采集的背景复杂的数据集时，本次课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法在检测处理时间上有明显增加，而传统 OCR 技术的执行时间几乎没有变化，经排查后发现，在处理部分复杂背景的图像时传统 OCR 技术已经完全丧失处理能力，在文本检测阶段未能正确检测到图像包含的文本范围，故无法进行下一步的识别流程，所有在整体的执行时间上没有明显波动。通过实际的检测效果，我们可以惊喜的发现本次课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法在预测理想状态下可以达到 99%以上的预测率。测试集样张如图 4-17，4-18 所示。

图 4-17 复杂背景测试图像样张

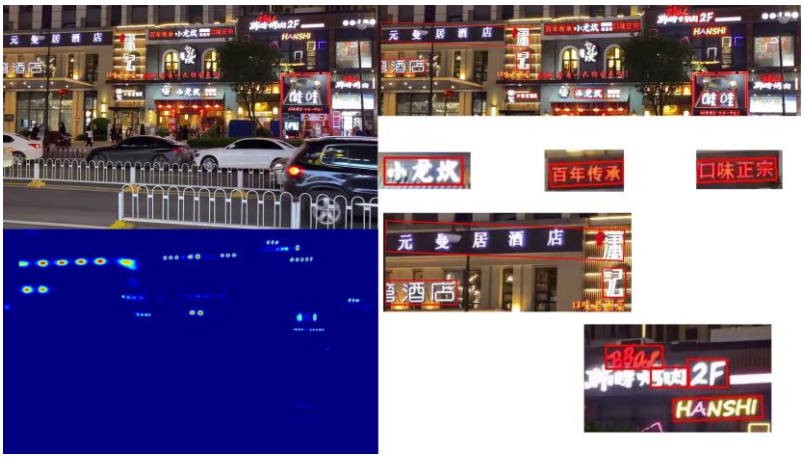

图 4-18 CCRW 在复杂环境下的表现

通过上述实验流程以及实际实验所得出的结果，我们可以有效的证明本次在以找寻一种基于机器视觉的高精度文字识别方法这一目标下提出的

CCRW（基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法）在实际的文本识别任务中具有可行性。

同时在本次实验中通过基线模型(OCR)证明了OCR技术在处理文档扫描类任务方面有着良好的表现，在处理自然环境下复杂背景文本识别任务时表现出不理想的执行效果。而 CCRW 以此做对比，在训练规模仅有 5000 文本张素材的前提条件下表现出了良好的识别效果，可以在预测率稳定在 94%极好情况下达到 99%的高精度表现处理自然场景下复杂背景的文本识别任务，同时该模型在多组测试集中表现稳定，具有泛化性。由此证明了 CCRW 相对于 OCR 技术在自然场景文本识别任务中的优越性。

4.3 实验总结

通过上述实验，我们可以直观的得到，尽管受限于部分客观因素未能使本次课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法表现出更好的性能，但也在取得了在处理单一文本与传统 OCR 技术近乎相同的效果，在处理自然场景下背景复杂、文本畸变、局部对比度过高等一些列复杂文本中相较于传统 OCR 技术展现出极大优势。

可以认为本次实验较成功的验证了本课题提出的基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法的可行性与在处理复杂情况文本下对于传统 OCR 技术的优越性，整体达到了较为理想的状态。

本章小结

本章节通过将本次课题提出的一种以弱监督方式训练的字符级检测网络构成的文字识别方法与传统 OCR 技术建立对照实验，并论证了各个环节的稳定性、论述其模型算法、实现过程，并在实验中验证了该方法的可行性与相较于传统 OCR 技术在复杂场景文本识别任务下的优越性。总体流程明了顺利。

第 5 章　研究成果应用

技术研究因以实际应用为本，将理论成果转换为实际应用带来显示意义是极为重要的。本章节将讨论 CCRW 技术在现实中的应用场景与可行性，在该流程中将使用 Web 前端技术结合 CCRW 粗略搭建一个网页文字识别系统来辅助论证。

5.1　CCRW 可应用的场景

在全面数字化的当今社会，信息数字化的推进浪潮不可避免，信息记录、传递已不局限于纸质载体，信息的数字化已成为科技生活的基础，如何高效率、高准确性的将信息数字化也成为了科技领域与大众生活中的重要课题。而本次课题提出的 CCRW 方法可以有效的参与到信息数字化的过程中。

在大众的日常学习工作中常见的信息数字化场景为证件信息提取、票务信息提取等生活场景，而 CCRW 表现出的网络体量小、配置要求低、响应时间短、识别准确度高等特性都可以很好的服务于上述应用场景。

为此，本次课题将实验 Web 前端技术结合 CCRW 搭建一个简单的网页文本识别系统来验证 CCRW 具有实际应用能力。

5.2　设计与实现

5.2.1 功能设计

为了体现 CCRW 的实际可应用性，系统应具备一个文字识别系统如用户文件选择、用户选择图像 URL、用户自主选择识别等功能。

上述基础功能全部由 Web 前端技术结合本次课题提出的 CCRW 识别方法合作完成。

5.2.2 功能实现

相关功能展示如下。

1.选择文件 用户进入网页选择、上传需要处理的图像文件。如图 5-1 所示。

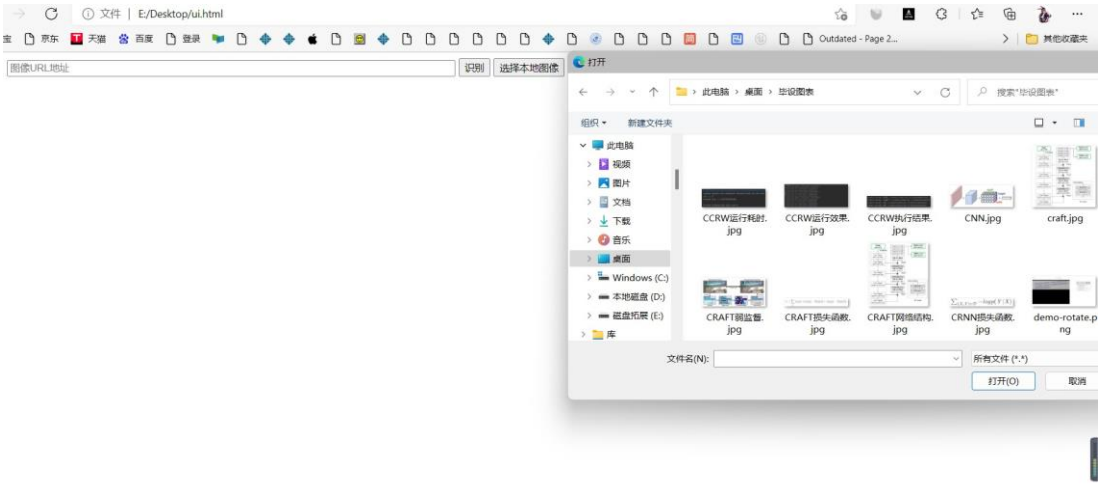

图 5-1 图像选择、上传界面

2.文本信息识别 文本信息识别功能，用户在网页端完成所需处理图像的选择与上传后可以单击识别按钮进行文字的识别。识别功能如图 5-2 所示。

序号	值	文本检测置信度
0	既然是自己选的,就不要抱怨,有的时候一条路开始了就不能回	0.99
1	头,也不要回头。不为了别的,只因为既然在开始时你有勇气选	0.98
2	择,就要有本事自己承担起后果。	0.97

图 5-2 文本信息识别

5.2.3 识别效果展示

经过实际测试，识别网页工作正常，可以正确调用 CCRW 识别用户传入网页的图像文本信息，识别效果与置信度理想。测试素材如图 5-3 所示。识别结果如图 5-4 所示。

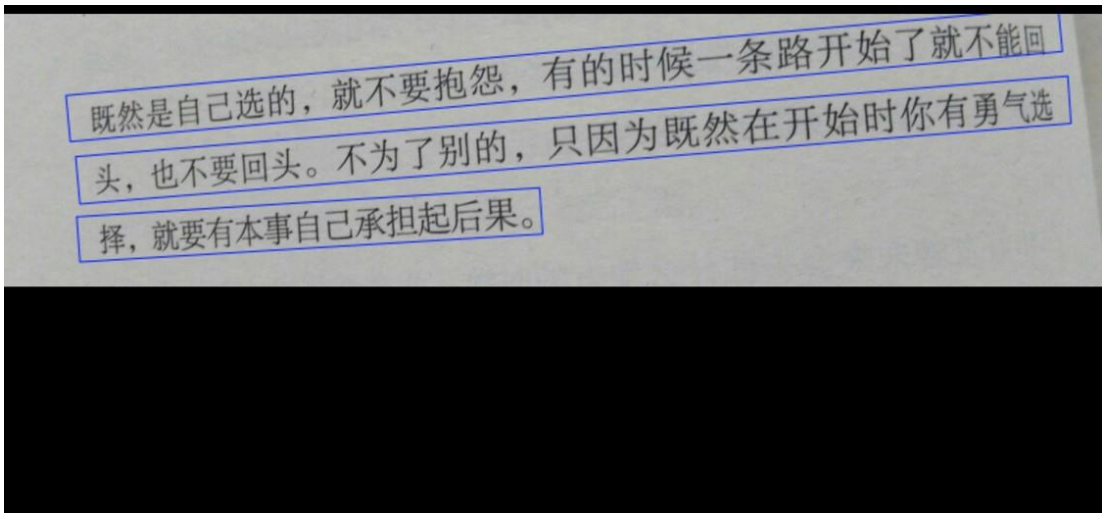

图 5-3 网页测试素材

序号	值	文本检测置信度
0	既然是自己选的,就不要抱怨,有的时候一条路开始了就不能回	0.99
1	头,也不要回头。不为了别的,只因为既然在开始时你有勇气选	0.98
2	择,就要有本事自己承担起后果。	0.97

图 5-4 系统识别效果

5.3 CCRW 方法的可应用性总结

通过使用 Web 前端技术并结合本次课题提出的 CCRW 方法搭建的网页版文本信息识别系统经实际检测具有响应快、配置要求低、处理精度高、场景适应性强的特性。可以胜任大众日常生活与学习工作中出现的绝大部分文本信息识别场景。

在具体的使用过程中，识别网页整体体现出了逻辑学习成本低、响应时间短、体量小、稳定性好的特性，可以极大的便利文本提取、文本录入等工作任务，为大众日常的学习工作提高了效率。

通过上述流程，展现了 CCRW 方法在现实中的应用，论证了本次课题在基于机器视觉的高精度文字识别方法研究与应用这一目标下提出的 CCRW 文本识别方法具有实际的可应用性。

体现出了 CCRW 方法在文本信息处理以及参与文本信息数字化流程的巨大潜力。

本章小结

本章主要目的为验证本次课题提出的 CCRW（基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法）在现实场景中是否具有可应用性，是否能够完成从理论到现实生产力的转变。通过使用 Web 前端技术结合 CCRW 识别方法所建立的网页文本识别系统在实际现实场景的使用中表现出逻辑学习成本低、响应时间短、体量小、稳定性好的特性，可以充分证明 CCRW 方法具有实际的应用能力。

结　论

本文阐述了机器视觉、文字识别技术在在当今人工智能技术蓬勃发展的大环境下的地位与面临的问题，也确定了课题为找寻一种基于机器视觉的高精度文字识别方法的方向与目的。对提出的技术要达到的预期及以往出现的问题做了详细阐述。介绍了文字识别领域的现状。其次列举了目前文字识别方面主流的 OCR 技术在自然场景文字检测的应用任务中所不能解决的问题。然后通过对本次课题提出的 CCRW（基于卷积循环神经网络与一种以弱监督方式训练的字符级检测网络构成的文字识别方法）的出发点、理想性能、方案的整体设计以及最终搭建做了详细的说明。并构建基于 OCR 技术的基线模型与本次课题提出的 CCRW 方法进行对比实验，通过符合流程的严格实验，证明了 CCRW 方法相较于 OCR 技术在处理自然场景下文本识别任务的优越性以及网络体量小、泛化性好等特性。

本次课题亦没有止步于理论分析与研究，通过在课题的后半程通过实验 Web 前端技术结合本次课题提出的 CCRW 识别方法搭建的网页文本识别系统，对 CCRW 方法进行了从理论实际生产力的转换，经过实际的使用体验与测试，辅佐论证了 CCRW 方法在实际应用中的可行性。也展现出了该方法在文本信息录入、文本信息数字化等工作领域的巨大潜力。

而本次课题也并非毫无遗憾，受限于一些客观客观原因，在初期采集到的 62 万张素材图像中仅训练了 5000 张作为数据模型以及被迫采用弱监督的方式训练识别网络，这都在很大层面上限制了 CCRW 方法的性能发挥与表现，希望以后能有更好的条件对其做出进一步的完善。

致　谢

时光荏苒，日月如梭，不知不觉间已是即将毕业的年纪，回首在学院求学的这几年，付出了许多，也收获了许多，在大学生活即将结束之际，向四年来指导过我的老师、帮助过我的同学表示感谢。

首先，感谢 HDACM-icpc 项目组导师，也是我本次课题的指导老师龚老师在大学的学习生涯中提供了项目组这个跳板让我有机会见识更多其他高校的技术精英并与之同台竞技，开阔了我的阅历、拓宽了我的技术视野，且在本次课题的整个流程中对我的指导与鞭策，从课题伊始的报告撰写阶段便不停的引导、督促，为课题方向提供了明确的思路。

其次，感谢学院以及四年来其他教导过我的老师，让我在学院度过了充实的大学生活。同时也感谢 HDACM-icpc 小组给了我良好的学习氛围并切实提升了我的编码能力。

还要感谢和我一起学习、生活四年的同学们，为大学生活在学习之余添加了一抹欢乐。

最后感谢批阅、评审该论文的老师在百忙之中审阅验收我的课题论文。

参考文献

[1] 章炜，机器视觉技术发展及其工业应用[J]红外,2005 ,27 (2) :11 - 17.

[2] Jan Koutnik, Klaus Greff, Faustino Gomez, Juergen Schmidhuber. A Clockwork RNN[J]. Proceedings of The 31st International Conference on Machine Learning, pp. 1863–1871, 2014.

[3] Haoxiang Li, Zhe Lin, XiaohuiShen, Jonathan Brandt, Gang Hua. “A convolutional neural network cascade for face detection”, CVPR.2015.

[4] Lichao Huang, Yi Yang, Yafeng Deng, Yinan Yu.“DenseBox: Unifying Landmark Localization with End to End Object Detection” CVPR 2015.

[5] 段峰，王耀南，雷晓峰,等. 机器视觉技术及其应用综述[J] . 自动化博览, 2002 (3) : 59 - 62.

[6] 刘焕军,王耀南. 机器视觉中的图像采集技术[J] . 电脑与信息技术,2003 (1) :18 - 21.

[7] Jaeger H, Haas H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication[J]. Science, 2004, 304(5667): 78-80.

[8] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Eprint Arxiv, 2014.

[9] 张中良 .基于机器视觉的图像目标识别方法综述[J]. 科技与创新, 2016(14):32-33.

[10] 叙勇,石绘机器视觉自动检测技术[M]. 北京:化学工业出版社, 2013,10:6-7.

[11] 蒋树强,闵巍庆,王树徽. 面向智能交互的图像识别技术综述与展望[J]. 过算机研究与发展, 2016,53(1):113-122.

[12] 秦亚航,苏建欢,余荣川机器视觉技术的发 展及其应用[J]. 科技视界, 2016(25).

[13] Zhang Z. A flexible new technique for camera calibration[J]. IEEE Transactionson Pattern Analysis & Machine Intelligence, 2000, 22(11):1330-1334.

[14] 哮辉.基于 OpenCV 的双目测距系统研究[D].武汉:武汉纺织大学, 2011.

[15] 周鹏. 基基 OpenCV 的双目立体视觉系统定标与三文重构技术研究[D].银川:宁夏大学.

[16] 紜峰 视频流中图标检测与识别的方法研究及其应用[D].上海:上海交通大学,2009.

[17] 王卓.基于 RGB 三基色原理的颜色检测仪的设计[D].天津:天津大学,2006.

[18] 王浩,许志闻,谢坤,等.基基 OpenCV 的双目测距系统[J].吉林大学学报信息科学版，2014,32(2): 188-194.

[19] 黄山园. 基于颜色特征的图像检索技术研究[D].太原:山西大学, 2013.

附录 1　译文

深度学习技术

深度学习(DL, Deep Learning)是机器学习(ML, Machine Learning)领域中一个新的研究方向，它被引入机器学习使其更接近于最初的目标——人工智能(AI, Artificial Intelligence)。

深度学习是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。 深度学习是一个复杂的机器学习算法，在语音和图像识别方面取得的效果，远远超过先前相关技术。

深度学习在搜索技术，数据挖掘，机器学习，机器翻译，自然语言处理，多媒体学习，语音，推荐和个性化技术，以及其他相关领域都取得了很多成果。深度学习使机器模仿视听和思考等人类的活动，解决了很多复杂的模式识别难题，使得人工智能相关技术取得了很大进步。

深度学习是一类模式分析方法的统称，就具体研究内容而言，主要涉及三类方法：

(1)基于卷积运算的神经网络系统，即卷积神经网络(CNN)。

(2)基于多层神经元的自编码神经网络，包括自编码(Auto encoder)以及近年来受到广泛关注的稀疏编码两类(Sparse Coding)。

(3)以多层自编码神经网络的方式进行预训练，进而结合鉴别信息进一步优化神经网络权值的深度置信网络(DBN)。

通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征表示后，用“简单模型”即可完成复杂的分类等学习任务。由此可将深度学习理解为进行“特征学习”（feature learning）或“表示学习”（representation learning）。

以往在机器学习用于现实任务时，描述样本的特征通常需由人类专家来设计，这成为“特征工程”（feature engineering）。众所周知，特征的好坏对泛化性能有至关重要的影响，人类专家设计出好特征也并非易事；特征学习（表征学习）则通过机器学习技术自身来产生好特征，这使机器学习向“全自动数据分析”又前进了一步。

近年来，研究人员也逐渐将这几类方法结合起来，如对原本是以有监督学习为基础的卷积神经网络结合自编码神经网络进行无监督的预训练，进而利用鉴别信息微调网络参数形成的卷积深度置信网络。与传统的学习方法相比，深度学习方法预设了更多的模型参数，因此模型训练难度更大，根据统计学习的一般规律知道，模型参数越多，需要参与训练的数据量也越大。

20 世纪八九十年代由于计算机计算能力有限和相关技术的限制，可用于分析的数据量太小，深度学习在模式分析中并没有表现出优异的识别性能。自从 2006 年， Hinton 等提出快速计算受限玻耳兹曼机(RBM)网络权值及偏差的 CD-K 算法以后，RBM 就成了增加神经网络深度的有力工具，导致后面使用广泛的 DBN(由 Hinton 等开发并已被微软等公司用于语音识别中)等深度网络的出现。与此同时，稀疏编码等由于能自动从数据中提取特征也被应用于深度学习中。基于局部数据区域的卷积神经网络方法今年来也被大量研究。

典型的深度学习模型有卷积神经网络(convolutional neural network)、DBN 和堆栈自编码网络(stacked auto-encoder network)模型等，下面对这些模型进行描述。

卷积神经网络模型

在无监督预训练出现之前，训练深度神经网络通常非常困难，而其中一个特例是卷积神经网络。卷积神经网络受视觉系统的结构启发而产生。第一个卷积神经网络计算模型是在 Fukushima(D 的神经认知机中提出的，基于神经元之间的局部连接和分层组织图像转换，将有相同参数的神经元应用于前一层神经网络的不同位置，得到一种平移不变神经网络结构形式。后来，Le Cun 等人在该思想的基础上，用误差梯度设计并训练卷积神经网络，在一些模式识别任务上得到优越的性能。至今，基于卷积神经网络的模式识别系统是最好的实现系统之一，尤其在手写体字符识别任务上表现出非凡的性能。

深度信任网络模型

DBN 可以解释为贝叶斯概率生成模型，由多层随机隐变量组成，上面的两层具有无向对称连接，下面的层得到来自上一层的自顶向下的有向连接，最底层单元的状态为可见输入数据向量。DBN 由若 2F 结构单元堆栈组成，结构单元通常为 RBM（RestIlcted Boltzmann Machine，受限玻尔兹曼机）。堆栈中每个 RBM 单元的可视层神经元数量等于前一 RBM 单元的隐层神经元数量。根据深度学习机制，采用输入样例训练第一层 RBM 单元，并利用其输出训练第二层 RBM 模型，将 RBM 模型进行堆栈通过增加层来改善模型性

能。在无监督预训练过程中，DBN 编码输入到顶层 RBM 后，解码顶层的状态到最底层的单元，实现输入的重构。RBM 作为 DBN 的结构单元，与每一层 DBN 共享参数。

堆栈自编码网络模型

堆栈自编码网络的结构与 DBN 类似，由若干结构单元堆栈组成，不同之处在于其结构单元为自编码模型(auto-en-coder)而不是 RBM。自编码模型是一个两层的神经网络，第一层称为编码层，第二层称为解码层。

附录 2　英文参考资料

Deep Learning Technology

Deep learning (DL, Deep Learning) is a new research direction in the field of machine learning (ML, Machine Learning). It is introduced into machine learning to make it closer to the original goal-artificial intelligence (AI, Artificial Intelligence).

Deep learning is to learn the inherent laws and representation levels of sample data. The information obtained in the learning process is of great help to the interpretation of data such as text, images and sounds. Its ultimate goal is to allow machines to have the ability to analyze and learn like humans, and to recognize data such as text, images, and sounds. Deep learning is a complex machine learning algorithm that has achieved results in speech and image recognition far surpassing previous related technologies.

Deep learning has achieved many results in search technology, data mining, machine learning, machine translation, natural language processing, multimedia learning, speech, recommendation and personalization technology, and other related fields. Deep learning enables machines to imitate human activities such as audiovisual and thinking, and solves many complex pattern recognition problems, which has made great progress in artificial intelligence-related technologies.

Deep learning is a general term for a type of pattern analysis method. In terms of specific research content, it mainly involves three types of methods:

(1) Neural network system based on convolution operation, namely Convolutional Neural Network (CNN).

(2) The self-encoding neural network based on multi-layer neurons includes two types: Auto encoder and Sparse Coding, which has received widespread attention in recent years.

(3) Pre-training with a multi-layer self-encoding neural network, and then combining the identification information to further optimize the deep belief network (DBN) of the neural network weights.

Through multi-layer processing, after the initial "low-level" feature

representation is gradually transformed into the "high-level" feature representation, the "simple model" can be used to complete complex classification and other learning tasks. Therefore, deep learning can be understood as "feature learning" or "representation learning".

In the past, when machine learning was used for real-world tasks, the characteristics of the description samples usually needed to be designed by human experts, which is called "feature engineering". As we all know, the quality of features has a crucial impact on generalization performance, and it is not easy for human experts to design good features; feature learning (representation learning) uses machine learning technology to generate good features, which makes machine learning "Fully automated data analysis" is another step forward.

In recent years, researchers have gradually combined these types of methods, such as unsupervised pre-training of convolutional neural networks based on supervised learning and auto-encoding neural networks, and then fine-tuning network parameters using discriminative information. Convolutional deep belief network. Compared with traditional learning methods, deep learning methods preset more model parameters, so model training is more difficult. According to the general law of statistical learning, we know that the more model parameters, the greater the amount of data that needs to be trained.

In the 1980s and 1990s, due to the limited computing power of computers and the limitations of related technologies, the amount of data available for analysis was too small, and deep learning did not show excellent recognition performance in pattern analysis. Since 2006, Hinton et al. proposed the CD-K algorithm to quickly calculate the weights and deviations of restricted Boltzmann machine (RBM) networks. RBM has become a powerful tool to increase the depth of neural networks, leading to the widespread use of DBN(Developed by Hinton and others and has been used in speech recognition by companies such as Microsoft) and other deep networks appeared. At the same time, sparse coding is also used in deep learning because it can automatically extract features from data. Convolutional neural network methods based on local data regions have also been extensively studied this year.

Typical deep learning models include convolutional neural network, DBN, and stacked auto-encoder network models. These models are described below.

Convolutional Neural Network Model

Before the advent of unsupervised pre-training, training deep neural networks was usually very difficult, and one of the special cases was convolutional neural networks. Convolutional neural networks are inspired by the structure of the visual system. The first convolutional neural network calculation model was proposed in Fukushima (D's neurocognitive machine). Based on the local connections between neurons and the layered organization image conversion, the neurons with the same parameters are applied to the previous layer. Different positions of the neural network resulted in a translation-invariant neural network structure. Later, based on this idea, Le Cun et al. used error gradients to design and train a convolutional neural network, and obtained superior performance in some pattern recognition tasks. Performance. So far, the pattern recognition system based on convolutional neural network is one of the best implementation systems, especially for handwritten character recognition tasks.

Deep Trust Network Model

DBN can be interpreted as a Bayesian probability generation model, which is composed of multiple layers of random latent variables. The upper two layers have undirected symmetrical connections, and the lower layer gets top-down directed connections from the upper layer. The bottom unit The state of is the visible input data vector. DBN is composed of a stack of 2F structural units. The structural unit is usually RBM (RestIlcted Boltzmann Machine). The number of neurons in the visible layer of each RBM unit in the stack is equal to the number of neurons in the hidden layer of the previous RBM unit. According to the deep learning mechanism, the input samples are used to train the first-layer RBM units, and their output is used to train the second-layer RBM models, and the RBM models are stacked to improve the model performance by adding layers. In the unsupervised pre-training process, after the DBN code is input to the top RBM, the state of the top layer is decoded to the unit of the bottom layer to realize the reconstruction of the input. As the structural unit of DBN, RBM shares parameters with each layer of DBN.

Stacked auto-encoding network model

The structure of the stacked auto-encoding network is similar to that of the

DBN, consisting of a stack of several structural units. The difference is that the structural unit is an auto-en-coder instead of RBM. The self-encoding model is a two-layer neural network, the first layer is called the coding layer, and the second layer is called the decoding layer.