



硕士毕业论文

(毕业设计)

题目：？？？

姓名：林丽

学院：信息科学与技术学院

系： 计算机科学系

专业：计算机科学与技术专业

年级：2014级

学号：23020141153152

指导教师：曲延云 职称：副教授

2017年4月10日

基于汉字组合特征的自然场景中文文本检测

**[摘要]** 自。

**[关键词]**自然场景文本检测

Chinese Scene Text Detection Based on Chinese Character

Combination Features

**[Abstract]** Scene text detection and recognition is a hot topic in the fields of computer vision. It has been widely used in intelligent navigation, image retrieval, vision assistant system, etc. Although there exist a lot of research activities on scene text detection, Chinese text detection is still remained as a challenging problem. In this paper, we proposed a text detection method based Chinese character combination features. Our contributions are as follows: 1) We proposed simple Chinese character combination features, based on which a Chinese components merging algorithm, a word analysis algorithm, and a two-layer filtering algorithm are implemented afterwards. 2) We proposed an integrated algorithm for Chinese text detection, and the experimental results demonstrate the availability of our approach. 3) We built a dataset for Chinese scene text detection and recognition to evaluate the performance of our algorithm. The recall of our approach is 81.068% and the precision is 85.978%.

**[Key Words]**Scene text detection, Chinese character combination features, context

# 绪论

## 研究背景

### 1.1.1 研究意义

# 目标检测方法

目标检测（Object detection），作为计算机视觉领域的一个重要分支，也是ILSVRC（ImageNet Large Scale Visual Recognition Challenge）比赛的任务之一。

**目标检测**指的是给定输入图像和类别（一种或多种类别），找出图像中所有属于给定类别的目标的位置和所属类别。

根据目标检测的定义，只有目标的位置和类别同时符合要求才是正确的检测。如图2.1的上图（A1-A4）为给定类别为{“Steel drum”}的目标检测标准（A1）和检测结果（A2-A4）。图A1为GroundTruth，即要检测的目标类别为“Steel drum”,且目标位置为两个绿色框所在的矩形。图A3的绿色框类别判断正确但框太大，位置错误。图A4的橙色框找到的位置是正确的，但是类别判断错误。因此，只有图A2是位置和类别都正确的。【插图2.1】



图2.1 单类目标检测与多类目标检测

目标检测按照要求类别的多少，可以分为单类目标检测（也称为“**定位”**）和多类目标检测。如图2.1，上图为给定类别为{“Steel drum”}的单类目标检测标准和检测结果，下图为给定类别为{“Steel drum”, “Person”，“Microphone”，“Folding chair”}的四类检测标准和检测结果。单类检测只需找到要求的那类的所有目标，多类检测需要找到所有要求的类别的所有目标。

目标检测按照目标的不同，又可分为行人检测，车辆检测，人脸检测，交通标志检测，文本检测等等【插图2.2】。虽然不同子领域因为目标本身属性的特殊性，有各自的不同算法。例如，基于Viola and Jones的人脸检测算法，基于HOG-SVM的行人检测算法。然而，这些检测方法同属于目标检测的领域，殊途同归，许多算法的框架和本质都是相通的。比如，HOG-SVM除了可以用在行人，用在交通标志检测中也能取到很好的效果。【引用】本文的文本检测方法正是从通用的目标检测方法中得到启发，并将其迁移至文本检测领域。因此，本章着重探讨通用的目标检测方法——从传统方法到近几年基于深度的检测方法，通过目标检测方法的发展引出本文选择SSD作为核心算法的相关背景和原因。

## 传统的目标检测方法

传统的目标检测方法一般分为两大类，基于滑动窗口和基于区域的方法。

基于滑动窗口的方法就是用不同大小的窗口在原图上进行滑动，并用分类模型判断每一个窗口是否包含目标，最后把分类模型判断为正样本的检测结果进行一定的后处理，比如非极大值抑制等。这种方法需要用不同尺度，不同长宽比的窗口对图像的所有位置进行遍历，虽然尽可能降低了目标漏检的可能性，但时间复杂度高，产生的候选区域过多，给后续的分类模型过滤噪声步骤带来很大的计算压力。

基于区域的方法首先根据低级特征（比如，光强，颜色，梯度等）把图像的像素分成不同的区域，再用分类模型对这些候选区域进行判断，过滤其中的噪声区域。该类方法得到的候选区域比滑动窗口法得到的窗口少得多，因此在时间和实用性上比基于滑动窗口的方法更具有优势。另一方面，候选样本的减少以及使用的低级特征的鲁棒性差也极大限制了该类算法的性能。

两类方法各有优劣，其共同的部分在于特征表示与模型分类，主要不同在于基于区域的方法需要额外的候选区域提取算法来选择候选样本。近年来，候选区域提取，特征表示，模型分类这三个关键技术受到了越来越多的关注，发展迅速。

### 候选区域提取

文章【引用What makes for effective detection proposals?】对比了目标检测领域近几年有代表性的十二种候选区域提取算法，包括Bing， CPMC， Edge Boxes，Endres，Geodesic， MCG， Objectness， Rahtu， RandomizedPrim’s**，** Rantalankila**，**Rigor**，**Selective Search。【引用】其中，Selective Search和Edge Boxes相较于其他算法，其性能和速度的综合表现更好，使用也更加广泛。Selective Search是使用图像分割技术得到初始的小区域，再根据临近区域的颜色，纹理，尺寸，填充等相似性用贪心算法进行合并。Edge Boxes是用一个框内包含的完整的轮廓个数来判断框内是否包含目标，其依据是如果一个框内包含目标，则框内应该包含足够多的完整的轮廓，而不会与框的边界有很多相交的边缘。

### 特征表示

图像特征对图像分类，目标检测，语义分割等计算机视觉技术非常重要，好的图像特征可以极大的提高算法的性能。在没有引入深度学习之前，几乎所有好的目标检测系统都是基于HOG（Hitogram of Gradient）, SIFT（Scale-Invariant Feature Transform）这两大经典特征提取算法。【引用】HOG统计了图像的局部方向梯度的直方图，其核心思想是用光强梯度（或边缘方向）来描述物体的局部轮廓。SIFT通过在尺度空间中寻找极值点作为特征点，并用与位置，尺度，方向相关的描述子来表征目标。

### 模型分类

常用的分类模型有很多，比如支持向量机（Support Vector Machine，简称SVM）, 随机森林（Random Forest）， AdaBoost，K近邻（K-Nearest Neighbors，简称KNN），朴素贝叶斯（Naïve Bayes）, 神经网络（Neural Network，简称NN）等等。基于结构风险最小化的支持向量机，尤其是采用了核方法的SVM，由于其鲁棒性强而被广泛应用于各种检测算法中。

## 基于R-CNN框架的目标检测方法

近年来深度学习受到了计算机视觉领域的极大关注，短短几年内在各种不同的视觉任务中取得了巨大的突破【引用】。相比于传统方法使用大量人工设计的特征，深度神经网络可以通过层次化的多层网络自动学习出具有更强鲁棒性的高级特征与语义信息。在众多深度网络结构中，CNN（Convolutional Neural Network）是发展最早、最快的结构之一。2012年，【引用】在ImageNet的图像分类比赛中大放异彩，拉开了CNN快速发展的帷幕。2014年，Ross B.Girshick【引用】使用了Region Proposal与CNN结合的方法（R-CNN）代替了传统目标检测中滑动窗口与手工设计特征的方法，提出了R-CNN框架，使得目标检测取得了巨大突破，并开启了基于深度学习目标检测的热潮。

### R-CNN

1. **R-CNN的主要思想**

R-CNN的主要思想是利用候选区域算法从图像中提取出一定量的候选区域，并将其分别输入到CNN网络中，进行特征提取和分类判别。

1. **R-CNN的算法流程**

图【插图】是R-CNN的算法流程图。其检测过程主要分为3步：

第一， 用Selective search提取2000个候选区域；

第二， 用7层CNN网络模型（5个卷积，2个全连接层）提取卷积特征；

第三， 用多类线性SVM对每一个区域进行分类；

第四， 用softmax分类器对每一个区域进行回归得到更精确的位置。

1. **R-CNN的实验效果**

R-CNN算法通过把Region proposal与CNN结合起来，在Pascal VOC 2012的检测结果比当时业内最高的记录高出了30%【插入表格或图】。紧随其后的SPP, Fast R-CNN， Faster R-CNN，MultiBox正是基于R-CNN的这种框架进行的改进，这些方法被称为是R-CNN系列的目标检测法。

### SPP

* 1. **R-CNN算法的问题**

**问题一**：需要对Selective Search提取到的每一个候选框（proposal）都用前向CNN进行特征提取，大量重叠区域的卷积重复计算，影响算法的实时性；

**问题二**：由于存在全连接层，CNN的输入图像要求统一大小。因此，每个proposal需要统一裁切（crop）或者缩放（warp）到相同尺度大小，如图【插图】，这在一定程度上造成了图像畸变，影响算法的性能。

* 1. **SPP解决R-CNN问题的思路**

针对问题一，SPP采用整张图像作为CNN的卷积模块的输入，得到全图的卷积特征图。再依据每个候选框在原始图像中的位置，在卷积特征图上取出对应的卷积特征。采用这种策略，整张图只需要一次卷积，避免了重叠区域的重复卷积计算。由于卷积计算是整个CNN中最耗时的部分，因此，采用共享卷积计算的方式可以大大提高计算速度。

针对问题二，SPP在卷积和全连接层之间引入了一个称为Spatial Pyramid Pooling的池化层。该池化层可以将任意维度的卷积特征转换成固定维度的特征。该方法不但可以避免图像畸变，而且满足全连接层输入大小固定的要求。

* 1. **SPP的关键技术——Spatial Pyramid Pooling**

Spatial Pyramid Pooling是一种特殊的池化层，类似于图像金字塔。它的特点是不管输入图像大小如何，输出是固定维度的（一维）特征向量。

如图【插图】，Spatial Pyramid Pooling的输入是任意大小的图像（图中是卷积特征图），设为w\*h\*c，w和h分别表示每张图像的宽度和高度，c表示图像的通道数（图中c = 256）。设Spatial Pyramid Pooling的层数为k （图中k = 3），将每层的每个通道的图像分成nk\*nk个bin（图中，n1 = 1，n2 = 2，n3 = 4），在图像的每个bin中进行池化（图中采用最大池化，也可以用平均池化），并把c个通道的nk\*nk个bin的池化结果串联成一个nk\*nk\*c的一维向量（如图中的256 = 1\*1\*256， 4\*256 = 2\*2\*256， 16\*256 = 4\*4\*256分别为第1层，第2层，第3层的特征向量）。最后把k层的特征向量依次串联作为Spatial Pyramid Pooling的输出结果（输入到后续的全连接层）。

* 1. **SPP和R-CNN的网络结构对比**

采用了Spatial Pyramid Pooling后的SPP与传统的R-CNN网络结构不同，如图【插图】。传统R-CNN中，需要对每个候选区域进行裁切或压缩后再分别输入到卷积层，全连接层得到每个候选区域的分类结果。而SPP则首先对整张图进行全卷积，再对每个候选区域提取卷积特征，并分别输入到Spatial Pyramid Pooling和全连接层，得到对应的分类结果。

* 1. **SPP的实验效果**

SPP利用了共享卷积计算和Spatial Pyramid Pooling改进R-CNN算法，不但大幅提高了算法速度，而且还提高了检测精度。【引用】中提到，SPP与R-CNN相比，训练速度提高了3倍，测试速度提高了10~100倍，同时在精度上也有所提高。【引用Fast RCNN】

### Fast R-CNN

1. **R-CNN和SPP算法的问题**

**问题一**：R-CNN和SPP都是多步训练。R-CNN和SPP都需要分三步训练，先微调CNN网络（softmax分类器）用来提取特征，再训练SVM进行分类，最后训练回归器（regressor）对bounding-box进行精调。

**问题二：**R-CNN和SPP的训练都非常耗时，需要大量存储空间。对于R-CNN，每张图的每个候选区域的特征都要单独计算，而且需要将结果写入磁盘。SPP虽然共享了卷积计算，一张图像只需一次卷积，消除了R-CNN的卷积计算冗余问题，但是仍然需要将特征写入磁盘，这不但降低了训练速度，而且针对大训练集，需要大量存储空间。

**问题三：**R-CNN的检测速度很慢，SPP虽然提高了检测速度，但是CNN网络的卷积层参数无法更新，影响模型整体性能。R-CNN对每个候选区域单独使用CNN提特征，因此速度很慢。SPP共享卷积进行加速，同时为了解决不裁切和压缩的情况下，可以在CNN中输入任意大小的图像，加入了Spatial Pyramid Pooling，但无法解决Spatial Pyramid Pooling之前的卷积层的参数更新问题，导致SPP只能更新全连接层的参数，限制了模型的整体性能。

1. **Fast R-CNN的解决思路**

针对问题一和问题二, Fast R-CNN采用的策略是把特征提取，分类，bounding box回归整合成一个多任务（multi-task）的问题。对于问题一， Fast R-CNN摒弃了之前先用softmax训练CNN来提取特征，再用提好的特征来训练分类器和回归器这种分阶段训练的思路。Fast R-CNN不再用SVM，而是直接用CNN网络的softmax作为最后的分类器，甚至直接用另一个softmax作为回归器，把两个softmax同时接入损失层作为一个多任务问题来学习。这样直接把特征提取，分类，位置回归集成到一个网络中，使得训练只需一步，非常简洁的端到端模式。对于问题二，因为R-CNN和SPP的特征提取和SVM分类是分开的，所以必须先提取好所有样本的特征，这个数据量一般情况下比较大需要放入磁盘中。当要训练SVM时，需要把特征从磁盘中读入再训练。而Fast R-CNN将特征提取、分类、位置回归三者整合到一个多任务问题后，训练整个CNN模型就是同时学习了特征提取、分类、位置回归的参数，一步到位，不存在用磁盘作为中间媒介存储特征的过程。

针对问题三，Fast R-CNN采用了ROI Pooling来处理任意大小候选区域的输入问题，同时引入了分层采样策略（Hierarchical sampling，也称为Mini-batch Sampling）解决卷积层的参数更新问题。SPP的Spatial Pyramid Pooling层采用的是多层金字塔结构，Fast R-CNN将它进行简化为单层，称为ROI Pooling层，使得网络结构和训练更加容易。SPP无法更新卷积层的根本原因是每批训练样本（ROI）来自于不同的图像，而每个ROI往往感受野很大，甚至可能是整张图像大小。也就是说大量的训练样本（ROI）无论在前向传播还是后向传播都要处理几乎整张图像大小的输入数据，因此SPP如果要更新整个网络的参数（尤其卷积部分），将非常低效。Fast R-CNN的分层采样正是解决了这个问题。分层采样指的是用分层的方法采集训练样本——每次用来做批量随机梯度下降更新参数的样本（【角标，说明梯度下降法】mini-batch Stochastic Gradient Descent，本文用的随机梯度下降都是这个方法，以下用SGD简单表示）。假设每批需要R（R = 128）个ROI样本，首先采样N张图像，然后每张图像采样R/N个ROI。由于N取得小（N = 2），因此每张图像采样的ROI区域很多（R/N = 128/2 = 64），而这些来自于同一张图像的ROI在前向传播和反向传播中都是计算共享的（卷积）。实际上，R-CNN和SPP的采样策略是N = R, R/N = 1的特殊情况，每个ROI都来源于不同的图像，共享计算的区域越少，重复计算就越多，算法效率越低。因此，Fast R-CNN的采样策略比R-CNN和SPP的快了64倍。分层采样的策略大大提升了模型的训练速度，使得整个网络包括卷积部分的参数更新成为可能。

1. **Fast R-CNN的关键改进**

除了多任务，ROI Pooling和层次化采样，Fast R-CNN还采用了Truncated SVD来压缩模型和加速训练。下面简要介绍下这四个关键的改进技术。

* **多任务损失**

多任务损失包括了每个训练样本（ROI）的分类损失和bounding box的回归损失（特征提取任务作为分类和位置回归的前置步骤，必须接分类或回归来指导训练，没有独立的损失）。分类损失使用了softmax损失（log损失），回归损失使用了L1-smooth损失，而不是普通的L2损失，因为回归问题是无边界的，为了避免梯度爆炸，L2损失会对学习损率产生微妙影响，而鲁棒性的L1损失会消除这种敏感性【插图，L1损失，L2损失，L1-smooth损失】。

【公式】如公式，Lcls是分类损失项， Lloc是回归损失项。P = （p0, …，pK）表示在样本（某一个ROI）在K+1类的分类概率。K为总的类别数目，K+1表示除了总分类数，还要加上一个背景类。U是类别的索引，背景类索引为0， 其他类索为1， 2，…K。

表示样本的bounding box的回归, k表示属于第k类（每个样本对每一类都有一个回归）的回归，(x, y)表示对bounding box的质心的位置回归，w, h分别表示对bounding box的宽度和高度的回归。V= 表示和样本匹配的groundTruth的bounding box。超参数用来控制分类损失和回归损失的权重，在【论文】中， = 1。

[u>=1] = 1, u>=1, else 0（用公式表示）



* **ROI Pooling层**

ROI Pooling是2.2.2中介绍的Spatial Pyramid Pooling的特例，是一个k = 1的单层金字塔结构的Spatial Pyramid Pooling。二者的对比如下图【插图】所示，此处不再赘述其原理（详细请参考2.2.2）。

**【插图，多个尺度和单个尺度的不同】**

* **层次化采样**

层次化采样的原理在本节的第（2）部分“Fast R-CNN的解决思路”中已经介绍过了，这里不再赘述，只简要说明层次化采样策略可能潜在的问题。层次化采样中每批样本有一半以上来自于同一张图像，这些样本可能是相关的，所以有可能降低训练的收敛速率。然而，【引用】中的实验却表明这个问题并不存在，Fast RCNN收敛的迭代次数比R-CNN更少。

* **Truncated SVD加速**

如图【插图】，【引用】通过实验表明，对于检测任务，由于ROI数目很多，因此算法有一半的时间是花在了全连接层上。【引用】利用了Truncated SVD分解把一个大的全连接层换成了两个小的全连接层，该方法不但减少网络参数，压缩了网络模型，而且使得检测算法更快（训练和测试）。

1. **Fast R-CNN的算法流程**

Fast R-CNN算法的检测步骤主要有两步：

第一，用候选区域提取算法（Selective search）得到候选区域集（ROI）；

第二，将图像和ROI输入到Fast R-CNN中，得到每个ROI的分类概率和bounding box的回归值，用阈值过滤后得到检测结果。

如图，【插图】是Fast R-CNN的网络框架，可以分为四步：

第一， 用ConvNet得到整张图像的卷积特征图；

第二， 用ROI在第一步得到的卷积特征图上进行映射，得到每个ROI对应的特征；

第三，将每个ROI的卷积特征用ROI pooling层池化得到固定长度的特征，并输入到后续的两个全连接层；

第四，将全连接层得到的特征分别输入到用来分类和位置回归的两个softmax层，得到分类的概率分布和bounding box的回归值。



1. **Fast R-CNN的实验结果**

Fast R-CNN使用了多任务损失，分层采样，ROI pooling，Truncated SVD等方法改进了R-CNN和SPP算法。和R-CNN相比，Fast R-CNN的训练速度快了9倍，测试速度快了213倍。和SPP相比，Fast R-CNN的训练速度快了3倍，测试速度快了10倍。且Fast R-CNN的检测精度比R-CNN和Fast R-CNN都更高。

### Faster R-CNN

1. **Fast R-CNN的问题**

在Fast R-CNN的检测流程中，候选区域生成算法（Selective Search）成为了整个算法的瓶颈——体现在提取算法本身是用CPU实现，速度慢，且得到的候选区域过多，增加了后续步骤的计算压力。而且，候选区域生成算法的存在使得整个检测流程分成了级联的两个部分，不但影响检测速度，而且不便于训练和优化。

1. **Faster R-CNN的解决思路**

【引用】提出了一个称为Region Proposal Network（RPN）的深度全卷积网络代替了Fast R-CNN的候选区域生成算法（Selective Search），用来生成候选的区域。并且，把RPN和Fast R-CNN整合成一个检测网络，该网络称为Faster R-CNN。RPN网络可以在GPU中完成计算，且提取的候选区域比Selective Search等候选区域生成算法更好（更少但能保证足够的召回），更重要的是，和Fast R-CNN能够联合训练，成为一个完整的端到端检测框架。

1. **Faster R-CNN的网络结构**

* **Faster R-CNN的网络结构和检测流程**

图【插图】是Faster R-CNN的网络结构图，主要是由RPN网络和Fast R-CNN网络组成。RPN网络和Fast R-CNN网络共享了前面的卷积部分。因此，整个检测流程可以分为三步：

第一，用共享的卷积网络部分得到卷积特征图；

第二，将卷积特征图输入到RPN网络，得到候选的区域集；

第三，将卷积特征图和候选区域集同时输入到Fast R-CNN中，得到检测结果。

* **共享的卷积网络**

RPN网络和Fast R-CNN网络共享卷积的网络结构从Zeiler and Fergus model（ZF）【引用】和Simonyan and Zisserman model（VGG-16）【引用】继承来的。ZF模型和VGG-16模型的网络结构如图【插图】所示。

1. **Faster R-CNN的RPN**

* **锚点（Anchor）**

锚点（Anchor）是RPN的核心概念。它受滑动窗口机制和图像金字塔的启发。**Anchor**指的是用滑动窗口搜索**特征图**的每一个位置，将与每一个位置相关联的具有一定尺度、长宽比的区域映射在原图像上的矩形区域。

Anchor的主要思想是一旦网络结构定义好（参数固定），输出的候选区域个数必须是固定的（不管判断是背景还是目标，都作为候选区域输出，只是由置信得分大小来判断是背景还是目标）。既然是固定个数的，那么这些候选区域初始化时分布在整张图的每一个位置，并且拥有多种尺度，多种长宽比无疑是最合适的。Anchor就是这样的初始值，也可以看作是固定的“参考对象”。不论图像上有没有目标，目标在哪里，这些“参考对象”的个数，位置，大小都始终不变。有了这些“参考对象”，那么网络的学习过程实际上就是以下两步；

第一，把groundTruth和“参考对象”进行匹配，得到匹配误差；

第二，反复调整网络参数，使预测结果和“参考对象”的匹配误差接近第一步得到的匹配误差，二者越接近，预测结果越接近groundTruth，模型精度就越高。

如图，【插图】利用256个3\*3的滤波器对卷积特征图进行卷积（3\*3的滑动窗口），每个位置卷积后得到256维的特征，将特征分别输入到分类层（cls）和回归层（reg）得到分类结果与bounding box的回归结果。因为每个位置与k（k = 9）个anchor相关联（3种尺度，3种长宽比），每个anchor都对应了类别信息和位置信息。【引用】中为了简化，类别信息只包含了两类信息，是背景或者不是背景，每个anchor的每个类别都有对应的得分概率，故每个位置的类别信息共2k维。位置信息包括了bounding box的中心点坐标（水平和竖直），宽度和高度这4维信息，故每个位置的位置信息共4k维。

* **RPN的损失函数**

和Fast R-CNN一样， RPN的损失函数由两部分组成，分类损失和bounding box的回归损失。如公式【公式】所示，设一个mini-batch中和第i个anchor（设为anchor\_i）相关联的训练样本为ai，当ai是正样本时，记与ai匹配的groundTruth为ai\*。pi表示用网络预测ai是正样本（目标）的分类概率，pi\*表示ai实际是正样本（目标）或负样本（背景），当ai是正样本时，pi\*=1, 否则pi\* = 0。ti是ai与anchor\_i的bounding box的差值，Ti\*是ai\*与anchor\_i的bounding box的差值。Ti和ti\*都是四维的向量，分别对应bounding box的中心点的水平坐标x,竖直坐标y，bounding box的宽度w和高度h。ti和ti\*的具体计算如公式【公式】所示。在公式【公式】中，第一项是分类损失，第二个是bounding box的回归损失，前者使用Smooth L1损失，后者使用二类的Log损失。Ncls和Nreg分别是正则参数，表示对应的样本个数。用来平衡两项损失，【引用】中取10，使得二者相近。





RPN和Fast R-CNN的损失不同在于，Fast R-CNN回归器的输入特征是由任意大小的ROI得到的，也就是说所有尺度的ROI都共享回归器的参数，只有一组回归器参数。而RPN的回归器有k组，每一组回归器负责一种尺度和一种长宽比的anchor，k组回归器没有参数共享，各自独立。因此，虽然每个anchor的输入特征维度是固定的（每张feature map上的每个位置都是同一个用3\*3的卷积核进行卷积），但是仍然可以预测多尺度的bounding box。

1. **Faster R-CNN的训练**

由于Faster R-CNN的RPN和Fast R-CNN共享卷积部分的特征，因此这两个网络的训练都会更新卷积部分的参数，这需要有特别的联合训练方式而不是独立训练两个网络。【引用】中提出了三种训练方式，交替训练，近似联合训练，非近似联合训练。这里我们介绍【引用】真正使用的四步交替训练流程：

第一，使用ImageNet上预训练好的模型训练RPN；

第二， 以第一步得到的RPN的候选区域为ROI输入，使用ImageNet上预训练好的模型，训练Fast R-CNN；

第三，固定RPN和Fast R-CNN共享的卷积部分，训练RPN；

第四，以第三步得到的RPN的候选区域为ROI输入，固定RPN和Fast R-CNN共享的卷积部分，训练Fast R-CNN；

1. **Faster R-CNN的实验结果**

Faster R-CNN将RPN和Fast R-CNN通过卷积共享整合成一个完整的端到端检测网络，使用GPU实现整个检测系统，大幅度提升了检测速率，并进一步提高了检测精度。

### R-CNN系列方法的总结

如图【插图】， 从R-CNN到Faster R-CNN，目标检测的四大基本步骤——候选区域生成，特征提取，分类，bounding box的回归（位置精修），逐步被统一到一个深度网络框架中。从检测的训练、测试速度来看，每一次改进都是一次跳跃，直到最后完全在GPU中实现，没有了重复计算。检测精度也是如此，稳步上升。R-CNN发展到Faster R-CNN，最初的传统方法已经不可与之比拟了。可以说，R-CNN系列的方法将整个目标检测领域带入了一个新的阶段。尽管Faster R-CNN不是最好的检测方法，但后来发展起来的YOLO，SSD，R-FCN等Single stage的方法都是基于R-CNN系列进行的改进或者是受到的启发，足见R-CNN系列方法对目标检测领域的重要性和影响力之大。

## Single stage的目标检测方法

Faster R-CNN虽然是一个end-to-end的完整的检测框架，但是仍然是一个two-stage的思路，即先检测出候选区域，再对候选区域进一步分类和精调，这在无形中限制了训练和检测的速率。随Faster R-CNN之后发展起来的YOLO（You Only Look Once）, SSD（Single Shot MultiBox Detector）都是基于R-CNN系列的改进，且都属于Single stage的方法。由于本文重点在SSD，本节关于YOLO将做简要介绍，详细请参考文献【引用】。

### YOLO

1. **YOLO的主要思想**

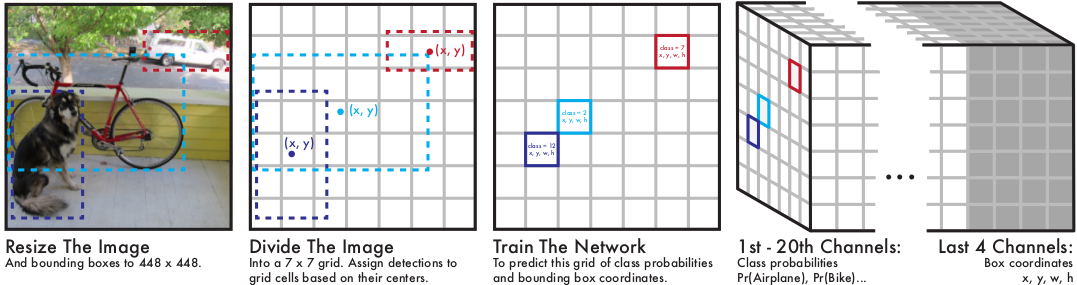
R-CNN系列的方法虽然经过了一系列的加速，但仍然无法达到实时的25fps的要求。影响速度的原因有很多，主要有以下几点：

第一，图像至少都要过两遍，第一遍找出候选区域，第二遍确定检测结果；

第二，为了保证性能，对输入图像大小有一定要求（例如，Faster R-CNN默认不小于600，虽然实际可以任意大小输入，但图像太小影响性能）；

第三，候选区域的数量偏多（例如，Faster R-CNN的默认是300）。

YOLO的主要亮点和目标就是快速，实时。YOLO延续了Faster R-CNN的端到端训练的特点，并把目标检测看作是一个回归问题，而且只需过一遍图像，这从它的名字You Only Look Once就可以看出来。它的设计思想是：把图像分割成多个网格，假如某个目标的中心落在其中某一个网格中，则该网格负责检测该目标的位置和类别。如图【插图】所示，图中的网格为7\*7大小，狗，自行车，小车的中心分别是紫色点，蓝色点，红色点。这三个点所在的网格负责输出紫色，蓝色，红色对应的矩形框，以及各自的分类信息（狗，自行车，小车）。



1. **YOLO的检测流程**

如图【插图】，YOLO的检测流程非常简单。只需三步：

第一，把原始图像缩放成固定大小；

第二，运行单个卷积网络；

第三，使用非极大值抑制后处理得到物体的位置和类别。



1. **YOLO的模型**

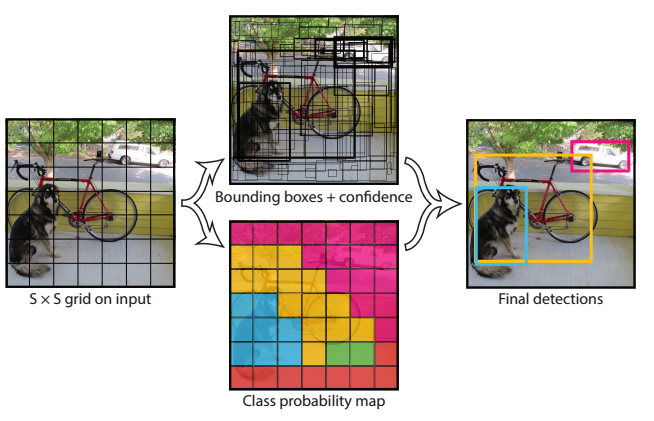
YOLO的模型如图【插图】所示。由于YOLO把检测看做是回归问题，因此，模型的主要任务就是回归目标的类别和位置。模型的逻辑流程如下：

第一，将缩放后的图像平均分割成S\*S的单元格；

第二，每个单元格负责预测B个候选框，每个框包括三类信息：位置信息，二分类信息，多分类信息。位置信息包括了bounding box的（x,y, w,h）共4个维度。（x,y）表示中心点坐标，w,h分别是宽度和高度。二分类信息用来判断框内是否包含了目标（不考虑具体是哪一类），只有1个维度。多分类信息提供了框内目标的所属类别，假设共有C类，则该信息为C个维度。由于YOLO假定以每个单元格为中心的目标只有一个，因此，在一个单元格上的B个候选框的多分类信息是共享的。也就是说，对于每个单元格，输出个数应为(4+1)\*B + C = 5B+C。对于测试，每个候选框的置信概率预测为公式【公式】，分类的总概率由二分类概率与多分类概率的乘积得到。

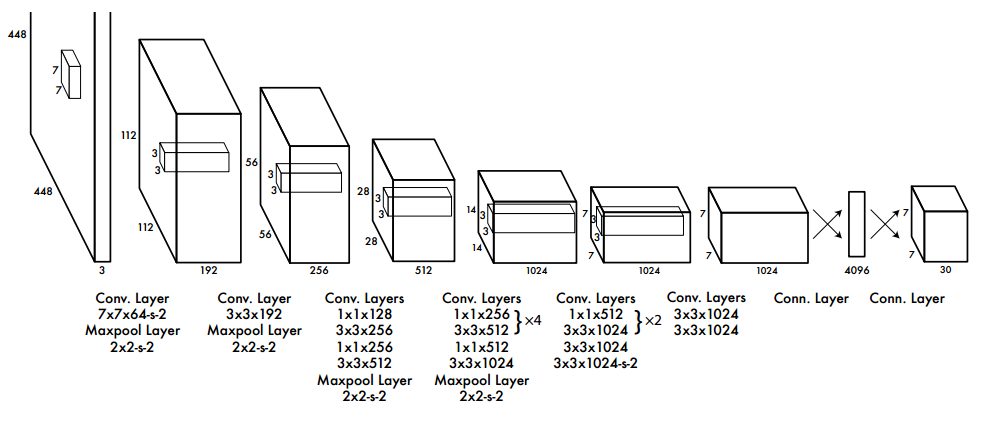


第三，最后的总输出层应有S\*S\*(5B+C)个单元。【引用】中，S = 7, B = 2, C = 20，输出单元为7\*7\*（4\*2+20）= 7\*7\*30 = 1470。



1. **YOLO的网络结构**

YOLO的网络结构参照了GoogleNet【引用】的设计。如图【插图】是YOLO的网络。可以看出，YOLO使用了24个级联的卷积层和2个全连接层。图中的多个3\*3卷积前接入1\*1的卷积是为了降维，这是参考了GoogleNet的经典Inception结构，如图【插图】。



1. **YOLO的实验结果**

YOLO的输入图像大小从Faster R-CNN的1000\*600减少到了YOLO的446\*446，候选区域的数量从300减少到了98，检测速度得到了巨大的提升（45fps），完全达到了实时的要求，但是精度比Fast R-CNN低一些。

1. **YOLO的优缺点**

* **优点**

第一，快速，已经达到实时的效果。把检测看作是一个回归问题，流程非常简洁；

第二，有全局的上下文信息。在网络结构中有对整张图像的全连接层操作，因此利用到了全局的上下文；

第三，泛化能力强。【引用】中用自然图像训练模型，并在艺术作品中测试，效果比其他的检测方法更好。

* **缺点**

第一，精度不够高，尤其对于小目标检测效果不好；

第二，由于网络结构中使用了全连接层，所以初始图像只能缩放到固定大小。另外，由于对初始图像缩放，可能造成了一定的形变，使得无法检测极端缩放比的目标；

第三，每一个单元格只预测一个类别，当多个物体中心落入同一个单元格时，无法被同时检测出。

### SSD: Single Shot MultiBox Detector

在保证精度的同时兼顾速度，一直以来都是目标检测的方向。在速度方面，YOLO一直遥遥领先，达到并超过了实时的基本要求，但是其精度差强人意，与state of the art有很大差距。另一方面，从精度来说，最为业内所追捧的当属R-CNN系列的Faster R-CNN，同样Faster R-CNN也因为被人所诟病的速度问题难以被投入实际应用。而在YOLO和Faster R-CNN之中取到完美平衡的正是SSD，它同时吸收了二者的精髓，Faster R-CNN的anchor box的思想，以及YOLO的回归和single stage的策略，并提出了多层特征融合的关键技术，很好的解决了YOLO的精度问题和Faster R-CNN的速度问题，创造了目标检测领域的新纪录。

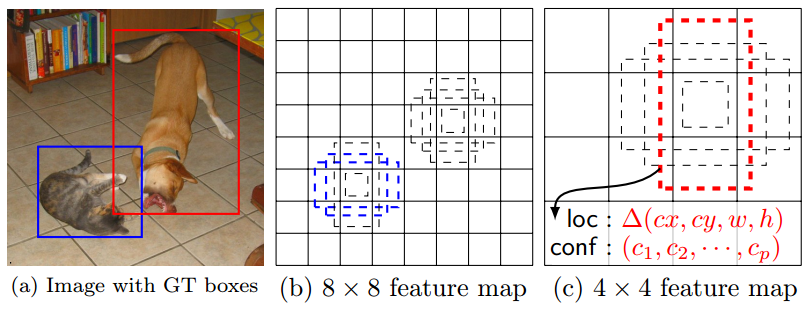
1. **SSD的核心思想**
2. **SSD的检测流程**

SSD的检测比YOLO更简单，无需将图像缩放成固定大小，因为SSD网络是全卷积网络，可以接受任意大小的输入图像。



1. **SSD的框架**

SSD和YOLO一样，把目标检测看成是一个回归问题。与YOLO不同，SSD不是在原图上画上网格，而是在特征图上，用滑动窗口提取每一个位置的特征，然后用特征回归得到目标的位置信息和类别信息。此外，SSD参考了Faster R-CNN的anchor思想，让每一个滑窗的位置对应了多个不同尺度，不同长宽比的anchor——在SSD中被称为default box，且default box不只应用于最后一个卷积层上，而是应用到多个不同大小的的卷积层上，采用了多尺度的卷积特征融合来共同回归目标。



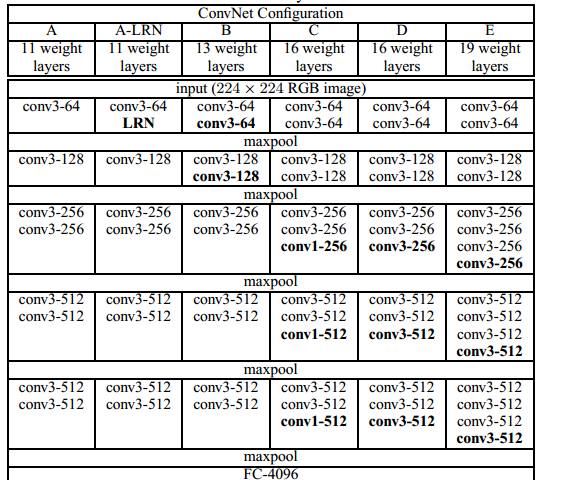
如图【插图】所示，SSD的原图和groundTruth对应的bounding box如图（a）所示，SSD只需提供这两个信息即可完成训练。图（b）和图（c）是两个大小分别为8\*8和4\*4的示例特征图，图上的矩形框分别表示default box在对应特征图上的映射（实际default box的大小符合原图尺寸，这里方便显示对default box进行了比例缩放）。从图中可以看出，不同大小的特征图，其defaut box的尺度也不同（8\*8的矩形框比4\*4的矩形框小）。不同大小的目标可能在不同尺度的特征图上回归得到（猫比较小，出现在8\*8的特征图上的蓝色框里。狗比较大，出现在4\*4的特征图上的红色框里）。由于SSD的多尺度特征图融合方法，使得不同大小的目标都可以检测到。

1. **SSD的网络结构**

SSD的网络结构如图【插图】所示，可以分解为4个部分：

第一， 使用了VGG-16的conv1~comv5\_3





SSD和YOLO都是single stage的方法，他们的网络有许多相似性。



1. **SSD的关键技术**
2. **SSD的损失**
3. **SSD的训练**
4. **SSD的实验效果**
5. **SSD与YOLO的对比**
6. **SSD与Faster R-CNN的对比**

## 对比总结与扩展延伸

### 对比与总结

### 扩展延伸

YOLO2, R-FCN， MultiBox，ION，OHEM，PVANet, CRAFT

## 本章总结

# 基于SSD的水平文本检测



## 基于SSD的水平文本检测

* + 1. **文字检测的不同粒度**
    2. **Default box的改进**
    3. **基于投票的多尺度输入融合**

## 实验与分析

* + 1. **Default box尺度大小的影响**
    2. **Default box长宽比的影响**
    3. **数据扩增的影响**
    4. **训练尺度的影响**
    5. **多尺度输出融合的影响**
    6. **多尺度输入融合的影响**
    7. **学习率的影响**
    8. **性能可视化**
    9. **样本大小的敏感性**

## 本章总结

# 基于SSD的倾斜文本检测

## 汉字组合特征

# 实验结果分析

## 数据库

## 评价标准

## 实验环境

## 字符级，单词级，文本线级的不同库上的对比结果（表格，pr曲线，时间）

# 总结与展望

## 总结

自然场景文本检测是计算机视觉领域的研究热点之一。自然场景中的文字蕴含着丰富的信息，在很多领域有着重要的应用。本课题旨在对自然场景中的中文文本进行检测和定位。本文主要利用基于多通道的最大稳定极值区域融合技术提取候选的文本区域。然后对文本区域进行了基于组合特征的文本检测，包括汉字部件合并，词组分析，和非文本滤波三个步骤。最后，采用机器学习的思想，利用改进的方向梯度直方图（HOG）和基于卷积神经网络（CNN）的无监督特征分别训练支持向量机（SVM），对候选区域进一步滤波。本文的主要贡献如下：

1.搭建了适用于中文文本检测识别的自然场景中文文本识别的数据库；

2.提出了汉字组合特征，并基于该特征，设计了相关的汉字部件合并算法、词组分析算法和两层非文本过滤算法；

3. 提出了对候选区域进行分类时应该结合其周围相似区域的性质来判断的思想（上下文信息），并在实验中成功运用该思想提升了实验结果，

4. 提出了面向自然场景中文文本检测的完整算法，并通过实验验证了算法的有效性。

## 展望

本文的研究工作虽然取得了一定的研究成果，但是仍然存在许多需要改进的地方，未来的工作将在以下方面开展：

1. 扩建自然场景中文文本识别数据库CCSV；

2. 结合其他基于连通分量的方法或者基于区域的方法。本文采用的是连通分量分析方法，这种方法在候选区域生成时一般都存在漏检的问题，因此可以将几种连通分量方法结合起来（比如与ER结合等）；

3.改进基于组合特征的相关算法。基于组合特征设计的算法参数均由人工调整，没有很强的自适应性。可以结合无监督学习的思想，将这些算法改进成可以自动学习的模型；

4. 卷积神经网络结构的改进。本文考虑卷积神经网络的复杂性，只利用到了卷积神经网络的特征提取方法。因此未来工作可以对卷积神经网络的训练结构进行简化后使用；

5. 设计端到端的中文文本检测识别系统。将中文文本的检测与识别结合起来，设计完整的端到端中文文本检测识别系统；

6. 将中文的文本检测算法推广到英文上，使算法具备更广泛的使用价值。

# 参考文献

1. A. Coates, B. C., C. Case, et al.“Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning”[C].in ICDAR. 2011. pp.440-445.
2. Asif Shahab, Faisal Shafait, Andreas Dengel. “Reading Text in Scene Images” [C]. in ICDAR.2011. pp.1491-1496.
3. Boris Epshtein, Eyal Ofek, Yonatan Wexler. “Detecting Text in Natural Scenes with Stroke Width Transform” [C]. in CVPR.2010.pp.2963-2970.
4. Dan Cires¸an, Ueli Meier, Jonathan Masci, et al.“A Committee of Neural Networks for Traffic Sign Classification” [C]. in IJCNN.2011.pp.1918-1921.
5. J. Matas, O. Chum,M.Urban, et al.“Robust Wide Baseline Stereo fromMaximally Stable Extremal Regions” [C]. in BMVC.2002.pp.384-393.
6. Jung-Jin Lee, Jung-Jin Lee, Pyoung-HeanLee, et al. “AdaBoost for Text Detection in Natural Scene” [C]. in ICDAR.2011.pp.429-434.
7. K. C. Kim, H. R. Byun, Y. J. Song, et al. “Scene Text Extraction in Natural Scene Images using Hierarchical Feature Combining and Verification” [C]. in ICPR. vol.2. 2004. pp.679-682.
8. Kai Wang, B. Babenko, S. Belongie.“End-to-End Scene Text Recognition” [C]. in ICCV. 2011. pp.1457-1464.
9. Hyung Il Koo,Duck Hoon Kim. “Scene Text Detection via Connected Component Clustering and Nontext Filtering” [C]. in TIP. 2013. pp.2296-2305.
10. S. M. Lucas, A. Panaretos, L. Sosa, et al. “ICDAR 2003 Robust Reading Competitions” [C]. in ICDAR. 2003. pp.682-687.
11. Lukas Neumann, J. Matas. “Text Localization in Real-world Images using Efficiently Pruned Exhaustive Search” [C]. in ICDAR. 2011. pp.687-691.
12. L. Neumann, J. Matas. “Real-Time Scene Text Localization and Recognition” [C]. in CVPR. 2012. pp.3538-3545.
13. Michael R. Lyu, Jiqiang Song, Min Cai. “A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction” [C]. in TCSVT. 2004. pp.243-255.
14. Song Gao, Chunheng Wang, Baihua Xiao, et al. “Adaptive Scene Text Detection based on Transferring Adaboost” [C]. in ICDAR. 2013. pp.388-392.
15. Tao Wang, David J. Wu, Adam Coates, et al. “End-to-End Text Recognition with Convolutional Neural Networks” [C]. in ICPR. 2012. pp.3304-3308.
16. YajuanS.,YanxiangHe,QingquanLi. "ReadingText in Street Views Using Adaboost Towardsa System for Searching Target Places".[C].in IVS. 2009. pp.227-232.
17. Yi-Feng Pan, Xinwen Hou, Cheng-Lin Liu. “A Robust System to Detect and Localize Texts in Natural Scene Images” [C]. in DAS. 2008. pp.35-42.
18. 张煜东. 降维技术与方法综述.[A].2010.
19. 吕岳. 基于汉字结构特征的自由格式手写体汉字切分.[A].2000.
20. <http://baike.baidu.com/view/564370.htm?fr=aladdin>
21. <http://blog.sina.com.cn/s/blog_68b72f57010133pz.html>
22. <http://blog.sina.com.cn/s/blog_60e6e3d50101bkpn.html>
23. <http://blog.csdn.net/v_july_v/article/details/7624837>
24. <http://baike.baidu.com/view/598089.htm?fr=aladdin>
25. <http://blog.csdn.net/zouxy09/article/details/8781543>
26. <http://blog.csdn.net/onlyzkg/article/details/11538203>

# 

# 科研成果

**参与的科研活动：**

1.2013年国家级大学生创新创业训练计划项目: 基于场景文本检测识别的地图智能定位研究（项目负责人)。

**发表论文:**

1. **Li Lin**, Yanyun Qu, Weimin Liao. "Structure Context Clues for Chinese Text Detection". in International Conference on Internet Multimedia Computing and Service(ICIMCS). 2014.(EI检索,已收录)

2. YAN-YUN QU, MENG-JIE LIAO, YAN-WEN ZHOU,TIAN-ZHU FANG, **LI LIN**, and HAI- YING ZHANG. "IMAGE SUPER-RESOLUTION BASED ON DATA-DRIVEN GAUSSIAN PROCESS REGRESSION". in International Conference on Intelligence Science and Big Data Engineering(IScIDE). 2013. pp.513-520 (EI检索).

3. Weiwei Wu, Yanyun Qu, Xiaoqing Yang, and **Li Lin**. "Traffic Sign Detection based on Camera Imaging Apriority". in Pacific-Rim Conference on Multimedia(PCM). 2013. pp.720-729 (EI检索).

# 致谢语

在本篇论文完成的同时，我首先要感谢我的导师曲延云老师。在曲老师的悉心指导下，我的论文才能如此顺利的完成。曲老师的负责、认真、严谨的作风对我的影响很深，以至于在我整个论文的撰写过程中都是带着这种精益求精的态度去做。在此我由衷感谢曲老师给我的帮助和指导，我相信这些习惯、思想会陪伴我一辈子。

另外，我想感谢研究生学长廖威敏同学。他从我接触这个领域开始就一直帮助我，带着我一步步往前走。对于我的问题，学长总会悉心解答，他不但在具体思路和编程中给予我指导，而且在生活中也像一个大哥哥一样帮助我。在这里，我衷心感谢学长。

我还想感谢实验室的所有同学，李建敏、廖梦杰、宋书阳、周燕稳、吴伟伟、卢畅、杨思颖。每当我想起大家都在实验室里一起奋斗的时候就很感动。其中，尤其感谢李建敏学长和吴伟伟学长，他们在我的论文和大学生创新性实验项目中提供了很多帮助。感谢你们！

我要感谢我们的创新性实验小队的所有成员，柳冰莹、刘勇顺、罗道文、王冰、车玉鑫。没有大家，就没有这个项目，感谢大家一直坚持我们的团队。

感谢我的舍友，感谢你们一路陪伴我，支持我！

最后，感谢我的家人，你们是我坚持的动力，谢谢你们给予的支持！