第四届“泰迪杯”
全国大学生数据挖掘竞赛

优
秀
作
品

“泰迪杯”优秀作品

作品名称：基于 MSER-CNN 的商品图片字符识别

作品单位：国防科学技术大学

作品成员：吴冰冰　邓志鹏　傅瑞罡

指导教师：指导组

基于 MSER-CNN 的商品信息图片字符检测与识别

摘　要：在互联网+时代，越来越多的人们选择在电商网站购物，商家提供的商品信息是用户做出购买决定的重要依据。图片这种生动、形象的视觉表达方式在给消费者带来便利的同时，也给电子商务网站的管理者带来了技术上的挑战：图片中的文本以光学字符的形式表达，不能使用计算机直接检索和处理。在大数据背景下，自动地从商品信息图片上提取其中的文本信息，这将有助于电子商务企业更好地做出商品推荐、售后服务和信息监管。本文将基于字符的检测与识别技术对商品信息图片进行文本信息提取与挖掘。

在本次数据挖掘过程中，我们首先对获取得到的图片和标注数据进行预处理，过滤掉少量错误的标注信息，以提高标注信息的正确性。

接着，对图片文本信息进行检测，主要分为离线和在线两个处理过程。在离线阶段，我们根据标注样本提取字符区域和非字符区域，正则化处理后得到字符样本集和非字符样本集，采用三种分类方法：基于 HOG 和 SVM 的分类方法、基于 LeNet 的分类方法和基于 Fast-RCNN 的分类方法。在线处理阶段，我们首先采用 MSER 算法对图片的 8 个通道进行字符检测，接着根据先验知识，对候选区域的面积、长、宽进行粗筛选，然后根据候选区域的行间距把左右相邻的字符区域进行联通，再对这些行区域进行形态学处理和垂直投影，得到单个字符区域。再把这些字符区域输入三种分类器进行背景区域的去除，得到最终的检测结果。

然后，对检测出的字符区域进行基于 CNN 网络的识别。识别之前统一对字符图像进行灰度化、“字亮底暗”的预处理，以缩小样本空间，提高网络识别率。识别分为离线过程和在线过程。在离线阶段，自主设计了 CNN 模型，训练后，这种单网 CNN 在测试集中得到了 93.07%的正确率。然而，由于给定的训练样本在种类上分布极不均匀，训练得到的 CNN 网络可能存在过拟合。我们尝试四种不同方法改进原网络：CNN+HOG、集成 CNN、双网、迁移 CNN。实验表明，在没有从根本上改变训练样本种类分布的情况下，提高识别率比较困难；此外，虽然本文的迁移 CNN 没有得到理想效果，但仍然是解决少样本，零样本问题，最有潜力的方法；最后，本文选择单网 CNN 作为识别模型。在在线阶段，我们把提取出的字符区域输入识别模型，得到预测的字符标签。

在实验过程中，我们分别对比了 Fast-RCNN、Faster-RCNN 等检测算法，同时对比分析了基于卷积神经网络和 SVM 分类器的优劣，以及不同的卷积神经网络模型，并对我们的检测识别方法的适用性以及参数设置进行了详细的分析，在给出的测试集中，检测率 F-scroe 为 0.524，识别正确率为 70.5%,最终平均 F2score 为 0.2676。验证了本文方法的有效性。同时本论文也提供了方便交互使用的软件界面，可以为网络信息监管工作提供有力的技术支持。

关键词：字符检测识别；MSER；SVM；CNN；迁移学习；

The Character detection and recognition of product information based on MSER-CNN

Abstract: In Internet plus era, people choose to shop in the website more. Goods information is an important basis for users to make a purchase decision on businesses. Vivid picture of visual expression bring convenience to consumers. But also brought technical challenges to the electronic commerce website managers: the text of the picture is expressed in the form of optical character and can't use computer retrieval and process it directly. In the context of large data, extract the text information from the image of the product information automatically will help e-commerce companies to make better product recommendation, after-sales service and information regulation. This paper will carry out the extraction and mining of text information based on the character of the detection and recognition technology.

In order to improve the accuracy of tagging information, we first preprocessed the image and the labeled data, filter out a small amount of error tagging information.

Then, the text information of image is detected, which is divided into two processes: off-line and on-line. When off-line, according to the label sample extraction of regional character and non-character regions, regularization treatment is used to get character and the non-character sample set. Using three kinds of classification methods: Based on the hog and SVM classification; on the LeNet and on the Fast-RCNN. When on-line. We use MSER on picture of eight channels to detect character. According to priori knowledge, the area, length, width of candidate regions are coarse screened. Then Unicoming according to the candidate region spaced around the adjacent character region, morphological processing and the vertical projection. Then single character region is got. These characters are input into three kinds of regions to remove the background, and the final detection results are obtained.

Then, based on the recognition of CNN network, the detected character areas are identified. Before recognition, the reunification of image character is grayed, "word all dark", in order to reduce the sample space and improve the rate of recognition network. Identification is divided into off-line and on-line. When off-line, the CNN model was designed, and the accuracy rate of the test set was 93.07% after training. However, due to the different distribution of training samples, the training of the CNN network may be over fitting. We tried four different methods to improve the original network: CNN+HOG, CNN, CNN, integrated dual migration. Experimental results show that in the absence of fundamentally changing the distribution of the different kinds of training samples improving the recognition rate is difficult. In addition, although the migration CNN did not get the ideal result, but still solve the small and zero sample problem, the most potential method. Finally, we choose single CNN as recognition model. When on-line, we put the extracted character region into the recognition model, get the prediction of the character tag.

In the experiment, we compared Fast-RCNN, Faster-RCNN detection algorithm, and comparative analysis of the based on the pros and cons of the convolutional neural network and SVM classifier, and different convolutional neural network model and of our method for detection and recognition of applicability and parameter settings were detailed analysis, focused on the given test, detection rate

F-scroe..... , the correct rate of recognition is 78%. The validity of the method is verified. At the same time, this paper also provides a convenient and interactive interface, which can provide strong technical support for network information supervision.

Key words: character recognition; MSER; SVM; CNN; transfer learning.

“泰迪杯”优秀作品

目 录

“泰迪杯”优秀作品

1. 挖掘目标

（1）挖掘意义

当下，消费者对于商品购买需求的多样化、便捷化促使网络购物这一消费渠道快速发展。各大电商平台例如京东、淘宝、一号店等都迎来了高速生长期。从各大交易平台来看，商品种类繁多，信息量大，消费者获取商品信息最直观的途径是查阅商品信息图片。图片作为商家提供的商品信息的载体，是消费者做出购买决定的重要依据。因此，各商家争相把商品信息图片做的丰富、生动、形象，以期吸引更多的消费者购买。更有甚者，无良商家把某些敏感信息、虚假信息、违禁词汇放入图片中，以期蒙骗消费者，躲过电商平台的监管。由于图片中的文本以光学字符的形式表达，不能使用计算机直接检索和处理。面对海量的图片信息，人工的去审查过滤这些图片很不现实。因此，自动地从商品信息图片上提取其中的文本信息，对于电子商务企业更好地做出商品推荐、售后服务、信息监管以及维护良好的网购环境，保障消费者权益，有着重要的研究意义和实用价值。

此外，除了电商网站，论文查重、广告图片、宣传海报、微信微博等社交软件中也存在大量以图像格式存在的文本信息，对这些图片里的文本信息进行提取，对于防止抄袭、舆情监管、图片审查等有着重要的实用价值，因此图片中的文本信息提取这一问题具有普遍性。

（2）挖掘目标

本次建模针对电商平台中的大量商品信息图片进行文本信息的提取，主要分为字符检测和字符识别两个过程，首先通过 MSER 算法提取大量的候选字符区域，然后根据字符的特征进行筛选得到待识别的字符区域，输入预训练的卷积神经网络进行字符的识别，以期正确输出每幅图片所包含的文本信息（字符的类别标签及其在图片中的位置），以实现全自动的图片文本信息检测识别过程，并将算法转化为软件成果，为网络信息监管工作提供有力的技术支持。

2. 分析方法与过程

2.1. 数据分析

本文建模使用的实验数据分为训练样本集和测试样本集。其中，训练样本集包含了 1000 张商品图片及其对应的标注文件（.box 格式存储）。标注文件提供了相应商品图像中字符内容及其最小外界矩形的位置坐标。测试样本集有 25 张商品图片及其标注数据。在去除掉错误的标注文件之后，我们对每一幅图片和对应的标注文件进行了显示操作。通过对训练样本集和测试样本集的观察分析，本文总结样本数据特征及可能存在的技术难点如下：

1. **待检测字符类别繁多。**

任务要求对图像中中文字符、英文字符、数字及标点符号均需检测出。其中：

1） 英文字符包含 52（A-Z，a-z）种类别。

2） 数字包含 10（0-9）种类别。

3） 常用标点符号 12 类。包括：

，	逗号	。	句号（中）	.	句号（英）	！	感叹号	？	问号
：	冒号	—	破折号	（）	括号	/	反斜杠	%	百分号
℃	摄氏度	、	顿号						

4） 中文字符种类繁多，约 3000 多类。

2. **部分字符间区分度较小。**如：

英文字符中 C/c，O/o，S/s，V/v，W/w，Z/z 大小写区分性很小；

数字 0 同英文字符 o 之间容易混淆。

3. **字符背景分为简单干净背景和复杂背景。**在复杂背景下，由于字符较小，信息量较少，在检测识别过程中受到复杂背景信息的干扰，影响检测识别结果。

(a) (b)

图 1. 图 1(a)左上角为简单背景，其余为复杂背景样例；图 1(b)吊牌中存在反光、字符形变等复杂情况。

4. **字符存在旋转角度（0-360°）。**在各类旋转角度下，字符均要能够被检测识别出来，定义旋转角度下字符的最小外接矩形不发生旋转。

(a) (b)

图 2. 图 2(a) 存在各种旋转角度的字符；图 2(b) 水印文字 logo 背景复杂，检测难度较大。

5. **水印文字 logo 需要被检测出来。**如图 2（b）所示。

6. **存在大量零样本、少样本、样本不均的字符。**

1） 存在非常规字体的字符。其中包括了通常意义下的艺术字体和经过特殊处理的字符，如字符有断裂样式、被划斜线等。

2） 少量样本的繁体字。

3） 边界不完整字体，如遮挡、图片边界位置等。

图 3. 图 3(a) 数据集中多样式艺术字；图 3(b) 字符存在断裂、划线情况；图 3(c) 少量繁体字样本；图 3(d) 边界不完整字体。

7. **可拆分字符检测识别歧义。**如中文字符“品”，在检测过程中可被分割检测为三个“口”字，左右结构和上下结构的汉字较多，容易被分别检测到，如何对检测结果优化，添加辅助判决信息，获取正确检测结果有待解决。

图 4. 可拆分字符检测识别歧义。“品”检测为三个“口”，“加”检测为“力”和“口”

8. **图片格式为 JPG 格式。**这种格式是一种压缩格式，图片放大后存在失真和模糊，对字符的检测和识别难度很大。

综上所述，网络商品图片的文本信息提取存在着诸多挑战，但是这种背景下的字符检测与识别也有较其他应用背景下（如自然场景的文本检测、手写体识别等）特有的**优势：**

①商品信息图片都是人工设计出来的，整体图像质量较好，不存在太多的噪声、光度等影响；

②商品信息图片中的文本都比较规整，都是已有的字符库生成的，没有太多的形变因素。

③商品信息图片中的文本为了便于阅读，基本还是按行或列的方式呈现，整体来说比较规律，不会出现随意摆放的情况，这个可以作为检测的先验知识。

2.2. 总体流程

题目要求根据商品信息图片的特点，设计算法完成字符检测与识别任务，并输出字符的位置和类别信息，再结合上述数据的分析，本文将挖掘任务主要分为字符检测与字符识别两个部分。采用的技术流程如下图所示：

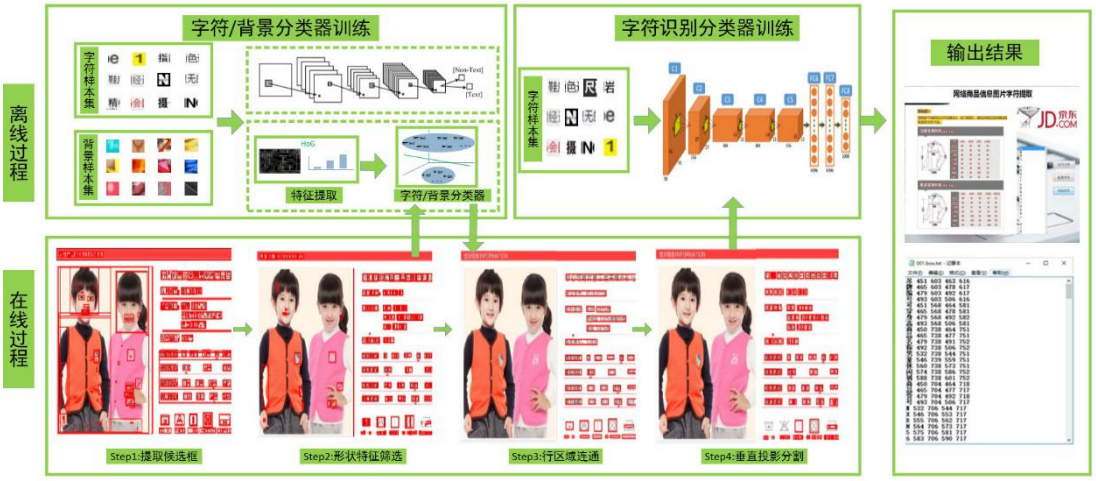


图 5 商品信息图片文本信息提取流程图

本论文的分析流程可大致分为以下三步：

第一步：预处理：对题目给出的标注文件进行预处理和显示，去除错误的标注文件，如空文件和坐标值为负数的文件等；

第二步：字符检测：

①根据标注样本提取字符区域和非字符区域，正则化处理后得到字符样本集和非字符样本集，通过提取梯度直方图特征训练得到字符 SVM 分类器；

②采用 MSER 算法对图片的 8 个通道进行字符检测，接着根据先验知识，对候选区域的面积、长、宽进行了粗筛选，然后根据候选区域的行间距把左右相邻的字符区域进行联通，再对这些行区域进行形态学处理和垂直投影，得到单个字符区域。再对这些字符区域提取梯度直方图特征，输入字符 SVM 分类器，采用非极大值抑制策略对背景区域和重叠区域进行筛选。最后根据标注样本对预测的候选区域进行回归，得到更加准确的字符区域；

第三步：字符识别：

①对常用的 3500 个汉字及所有英文字符（包括数字标点），通过系统字体变换，生成了最终字符数为 210 万的初始字符训练集，并利用该初始训练集，采用卷积神经网络的方法，训练得到一个初始的分类模型；

②固定初始卷积神经网络模型的底层参数，加入题目提供的训练样本集对分类模型进行调整，以实现模型的迁移。最终输入待检测的字符区域对字符的类别标签进行预测。

2.3. 具体步骤

2.3.1. 数据预处理

针对题目给出的图片和标记样本，我们首先利用 MATLAB 对给出的样本标记文件进行解析，并对图片和标记样本框进行显示。在显示过程中，发现了以下几个问题：

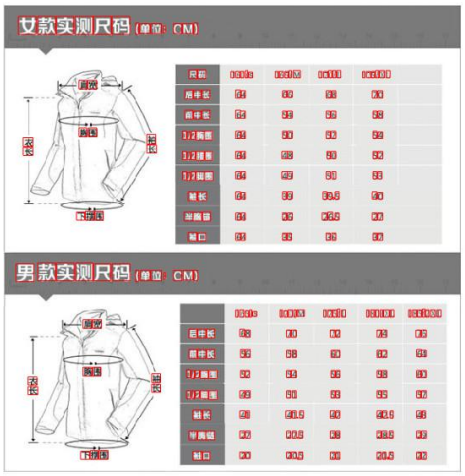
①有 8 个 box 文件是空文件，本文对其图片和标注文件进行了剔除；

②标准的标记文件格式为【字符，左下角 x,左下角 y,右上角 x，右上角 y】，如图 6（b）所示，但是部分标记文本出现了五个坐标值，本文对其末尾的数值删除后能正确的显示出标记框的位置，因此本文对这种情况下的末尾数值进行了批量删除。

③部分标记文件的坐标为负值，由于数量较少，本文对这些坐标行进行了删除。

经过标记文件的预处理后，所有的图片及其对应的标记框的位置均能正确的显示。如图 6（a）所示。

对于题目给出的图片文件，由于都是商家制作的，整体图像质量较好，只存在由于 JPG 格式压缩造成的放大后失真现象，因此本文对图片文件不做进一步的预处理。



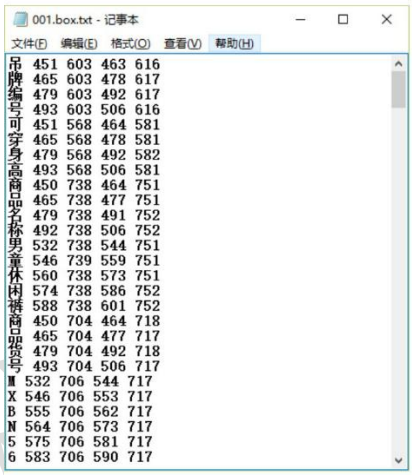


图 6 (a) 图片和标记框的显示　　(b) box 文件的内容

2.3.2. 字符检测

2.3.2.1. 字符检测的关键问题

题目要求根据商品信息图片中字符和背景的特点，设计算法从图片中检测字符（包括中文字符、英文字符和数字、标点）。检测结果保存在以 .box 为后缀的文本文件中，输出字符在原图中的位置范围。

具体到本文要解决的商品信息图片中的字符检测，需要解决好以下几个问题：

①漏检率低是首要任务，尽量不要出现检测不出的字符，这样会漏掉一些关键字符，给网络监管带来漏洞；

②在保证漏检率低的情况下尽量检测正确率要高，即允许存在少量的背景区域被检测出；

③字符检测的结构要完整，尤其是汉字的左右结构、上下结构等，要能够尽量完整的检测出汉字的区域；

④字符的位置要尽量准确，题目要求给出字符的最小外截矩形区域；

2.3.2.2. 当前字符检测算法

目标检测是图像处理领域的经典课题，主要的检测思想大致分为两类：一种是采用滑窗的策略对整幅图像进行遍历，另一种是首先提出大量可能的候选区域，然后再进行筛选。无论采用哪种策

略，其核心思想都是为了区分目标和背景，进而更加准确的找到目标所在的位置。具体到字符检测领域，我们对字符的特点总结如下：

①符的颜色与之周围的背景颜色有较大的差别。

②图像中的字符行往往是水平方向或者是垂直方向的，因为人们的阅读习惯主要就是水平方向从左至右进行的。

③提取的字符区域往往在距离上是很相近的。不管是英文文本还是中文文本，单个字符存在的概率是很小的，因为单个字符携带的信息很少。

④字符区域外接矩形的宽高比、空洞比、梯度方向直方图（Histogram of Oriented Gradient，HOG）等特征区分性比较强，可以用来区分背景和字符。

传统的字符检测算法主要集中在扫描文档、发票等简单背景下的字符检测任务，目前的研究热点在于自然场景下的字符检测任务，这种检测识别技术可以被广泛应用到自动驾驶、车拍识别、手机拍照翻译等领域，而针对商品信息图片的字符检测研究相对较少，属于大数据时代下的新问题。

常规的字符检测算法主要是基于区域特征的方法，代表性的方法有两个，一个是最大稳定极值区域(Maximally Stable Extremal Regions，MSER)[1]，这种方法认为字符区域是图像中的稳定区域，通过设定一系列阈值来找到最稳定的区域作为字符的候选区域，这种方法也是目前的主流方法，但其不足之处在与对光照影响不具有鲁棒性。另外一种方法是笔画宽度变换(Stroke Width Transform，SWT)[2]算法，这种算法认为字符的笔画是固定宽度的，因此可以有效区分出字符区域和背景区域，这种方法对于非水平方向的文本行检测有着较好的效果，但是总体检测效果不如 MSER 优越。

此外，基于卷积神经网络(Convolutional Neural Network，CNN)[3]的字符检测算法也成为了近两年的研究热点，该类方法能够克服传统的手工设计特征普适性不强的弱点，从大量的数据样本中学习出最佳的字符特征表达方式，最终进行字符的检测与识别，取得了较好的效果，但是该方法的缺点需要大量的标记样本和较长的训练时间。

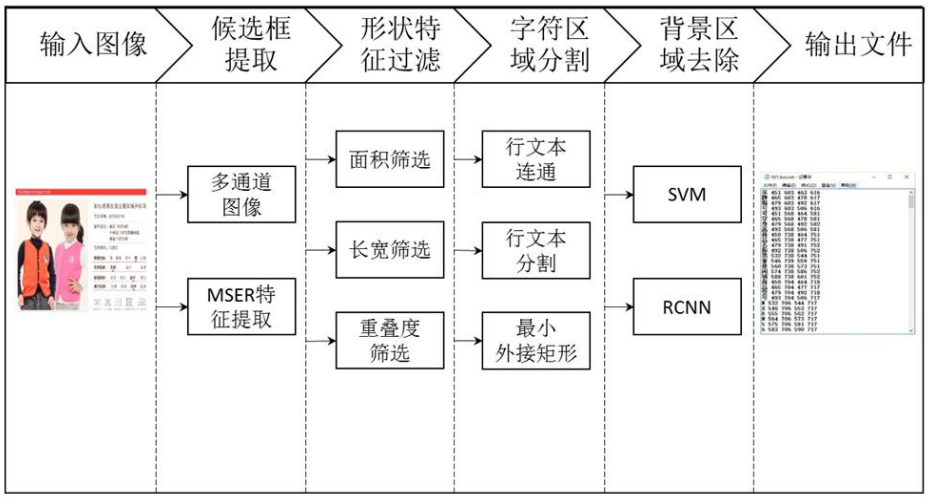


图 7 字符检测流程图

2.3.2.3. 本文的字符检测算法

由于商品信息图片中字符数量较多，所以本文采用先提取大量候选字符区域，然后再进行有效

筛选的策略来检测字符。字符筛选过程中重点关注两个问题：一是字符区域与背景区域的分类，二是字符结构的准确分割。本文检测算法主要采用多通道的 MSER 算法，筛选策略主要参考字符的长宽、面积、纹理、梯度等特征与背景进行有效区分，最后采用形态学处理方法以及字符行方向的先验信息对字符区域进行投影分割，得到最终的字符区域。

主要步骤如图 7 所示：产生候选(candidate),字符过滤，字符结构分割、检测结果的回归。

（1）基于 MSER 算法的多通道候选字符区域提取

MSER 算法是目前公认的目前性能最好的仿射不变区域。它是基于宽基线得到的显著性区域，具有显著性、仿射不变性、稳定性等特征，能在发生几何形变的不同图像中提取相同的目标区域。MSER 的基本思想是对一幅灰度图像(灰度值为 0-255)取阈值进行二值化处理，阈值从 0 到 255 依次递增。在得到的所有二值图像中，图像中的某些连通区域变化很小，甚至没有变化，则该区域就被称为最大稳定极值区域。这类似于当水面持续上升的时候，有些被水淹没的地方的面积没有变化。用数学公式解释如下：

对于图像 $p(x)$，$x \subset Q$，Q 是一个包含像素元素的实函数有限集，Φ 是一个拓扑结构参数。换句话说，我们定义 $Q=[1,2,\ldots,N]$，Φ 可以是四邻域或者八邻域，并且不限制 $n=2$，设图像 $p(x)$ 一个水平集 $S(x)$，$x \subset Q$，为灰度小于或等于 $p(x)$ 的集合：

$$S(x)=\{y \in Q \mid p(y) \le p(x)\} \quad (1)$$

序列 $(x_1,x_2,...,x_n)$ 是一个包含许多像素的连通序列(比如 x_i 和 x_{i+1} 是四邻域或者八邻域，$i=1,2,\ldots,n-1$)，Q 的一个连通分量 R 是 Q 的子集，$R \subset Q$，每对像素 $(x_1,x_2) \in R^2$ 都通过 R 中的一条路径相连。如果任何包含 R 的连通分量 R' 都等于 R，则称 R 为最大连通分量。极值区域 C 就定义为水平集 $S(x)$ 的最大连通分量。于是用 $C(i)$ 表示图像 P 所有的极值区域的集合。在 $C(i)$ 所有的极值区域中，此算法只对满足一定平稳标准的区域感兴趣。它的数学定义为：

$$q(i)=|Q_{i+\Delta}-Q_{i-\Delta}|/Q_i \quad (2)$$

其中，Q_i 表示阈值为 i 时的某一连通区域，△为灰度阈值的微小变化量，$q(i)$ 为阈值是 i 时的区域 Q_i 的变化率。当 $q(i)$ 为局部极小值时，则 Q_i 为最大稳定极值区域。

由 MSER 的数学定义可以看出，△的取值对最后的最大稳定极值区域的选取是有很大影响的。△取值小，则检测出来区域多，但不稳定；△取值大，则检测出来区域较少，但相对稳定。另外，△的取值大小跟 MSER 的提取时间消耗也有关系，△值越大，MSER 提取耗时越短，△值越小，MSER 提取耗时越长。目前还没有找到一种自适应调节参数△的方法来对应各种图像。

一般情况下，极值区域的提取是从原始图像的灰度图中提取的。对于商品信息图片中的文本区域，往往在笔画中包含了丰富的颜色分量信息，因此进行灰度图中的文本区域提取，会浪费掉图像中颜色的信息。如果能从图像的各种颜色模型中选择有效的颜色通道，对每一通道的二维图像进行极值区域提取，会大大增加商品信息图像中字符区域的检测率。因此本文采用了 R，G，B，I 和梯度强度通道 ∇，这里的 I 通道即 HSI 颜色空间的 Intensity 通道，强度梯度等级通道图像中，每个像素值是 HSI 空间的 I 通道图像中对应该位置像素值与其相邻像素值的强度差中的最大值，如图 8 所

示。

（a）原图 （b）I通道 （c）梯度强度图

图 8 多通道对比图

为了最大可能提取有价值的极值区域，分别增加了 R，G，B，I 这四个通道的取反通道，即用 255 减去这四个通道图像的各像素点的值而得到的图像。这样得到理论上的 9 个通道的图像。对这 9 个通道分别进行极值区域提取后，可以提取出大量包含字符区域的候选框。

本文采用 vlfeat 工具包实现 MSER 特征的提取，对提取的稳定极值区域求取最小外截矩形可以得到字符的矩形区域。△经验性的取 10 能取得较好的检测效果，对其中一个样本提取效果如图 9 所示。

（a）step1候选框 （b）面积较大候选框 （c）面积较小候选框

图 9 多通道候选框的检测

从图 9（a）可以看出，多通道提取的候选区域，基本完全覆盖了所有的字符区域，但是这些区域数量庞大，且存在着大量的重叠区域以及错误区域，需要接下来对这些候选区域进一步处理。

（2）基于字符区域形状特征的粗筛选

面积筛选：由于设定的△阈值较小，因此所提取的字符区域大小不一，形态各异，其中图 9（b）和（c）分别显示了面积大于 4000 和面积小于 50 的候选区域，这些区域的存在会干扰后续的字符特征统计，因此本文对其进行了剔除，剩余的候选框如图 10（a）所示，可以看出筛选后的候选框面积适中，检测效果较好。

字符垂直投影：但是存在的一个问题在于表格的方框等较宽的区域都得到了保留，很多字符是以两两相连的形式被检测出，而本文的主要任务是检测出单个字符的区域，因此本文在这里对这些候选区域进行了第一次的垂直方向投影，同时设定投影直方图的信息熵阈值为 0.25，对那些信息熵小于阈值的图像块进行去除，具体投影算法在第二次垂直投影时会详细介绍，以期得到单个字符区

域，投影变换后的字符检测如图 10（b）所示，可以看出垂直投影可以消除部分的背景区域，同时去除了表格区域。

字符宽度筛选：由于单个字符的长宽比满足一定的形态约束，本文计算投影分割后的字符区域的长宽比，去除掉那些较宽的字符候选区域，筛选效果如图 10（c）所示，对比图 10（b）可以看出长宽比的筛选有利于去除较宽的背景区域。

（a）面积筛选　　（b）投影筛选　　（ c）宽度筛选

图 10 字符区域的形状特征筛选

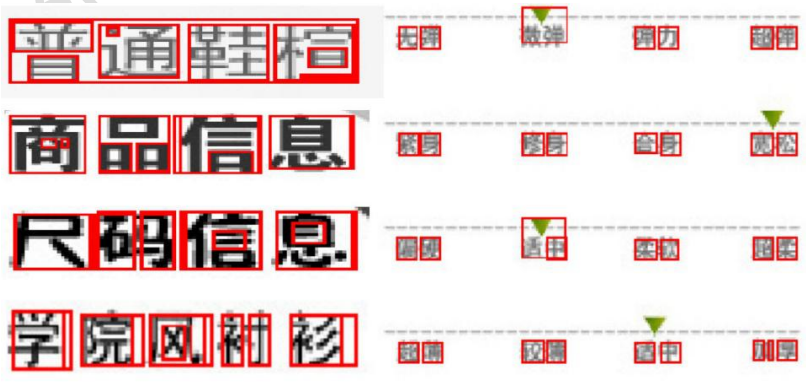
（3）字符行区域的连通

内部结构筛选：在上述处理结果中，我们可以发现由于汉字结构通常由简单的偏旁或者部首组合而来的，在投影过程中容易把这些结构拆分开来，如图 11（a）所示，“普”字被拆分为上下结构，“品”字被拆分为三个“口”字，“码”被拆分为左右结构，以及还有“风”被拆分为包含结构。对于这种情况，本文分两种情况来处理，对于包含结构的汉字，我们希望能去除被包含的那个字符，对此我们首先计算候选框之间的重叠度，候选区域 a 和候选区域 b 的重叠度计算公式如下：

$$overlapRatio(a,b) = intersectAB/min(aere(a), aere(b)) \quad (3)$$

其中，intersectAB 为区域 a 和区域 b 的重叠面积，对于重叠度为 1 的区域，我们保留面积较大的候选区域作为字符区域，去除内框后的效果如图 12（a）所示，可以看出包含在字里的候选框都得到了有效的剔除。

对于左右结构的汉字，将在后续的垂直投影中采用形态学的方法予以合并。



（a）汉字结构拆分　　（b）字符行排列规律

图 11 汉字错误检测的情况

行区域连通：字符的检测容易被周围的信息干扰，造成错误检测，如图 11（b）所示。但是通过观察字符排列的规律，可以发现，大多数的字符排列通常按照从左到右的按行排列，或者从上到下的按列排列，此外，每一行文本，字符与字符之间的间距较小，这是为了便于消费者阅读，这两个特点可以作为文本检测的重要先验知识。

常规的扫描文档的字符检测识别算法往往采用按行或者列的投影方式来进行字符分割，但是由于商品信息图片背景较为复杂，无法准确的找到字符行所在的位置。而通过观察图 11（b）可以发现，本文初步筛选后的字符区域，在行坐标上的分布具有局部一致性，这样可以统计这些候选字符区域的行坐标，通过局部聚类，可以大致找到每一个文本行所在的行坐标，再通过判断字符与字符之间的间距，对每一个文本行进行连通，这样就可以得到连通后文本行区域。

我们设定文本行之间的聚类阈值为 3 个像素（即行坐标相差 3 个像素以内的候选框我们认为是同一行的文本），字符间距阈值为 2.5 倍的平均字符宽度（即字符候选框之间水平距离大于 2.5 倍平均字符宽度的文本行需要进行拆分）,文本行连通的效果图如图 12（b）所示，其行坐标统计如图 12（c）所示，其中纵坐标轴为行坐标值，横坐标轴为候选字符框的个数统计。通过观察可以发现文本行都得到了有效的连通，此外，采用行连通的方法也能有效的去除了大量的背景区域，这是由于背景区域的行坐标没有字符区域这种局部聚类的排列规律。

（a）去除内框　　（b）字符行串联　　（c）字符行水平投影

图 12 局部字符候选区域行连通

（4）字符行垂直投影

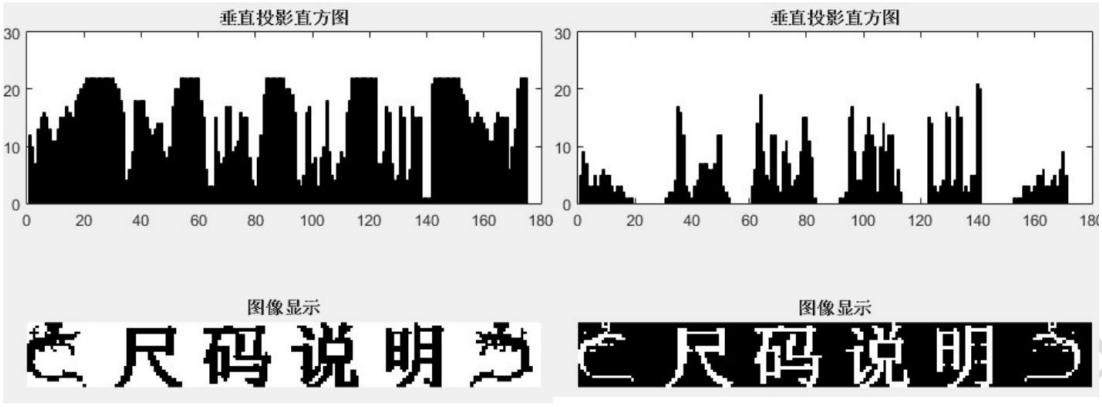
在得到文本行的基础之上，我们希望能够在垂直方向上把字符分割开，且不能出现左右结构拆分的情况，如果情况允许，还需要单独把标点符号分割出来。在这里，我们首先截取候选字符区域，把三通道彩色图像变换为灰度图像，然后对其做二值化处理，其阈值采用如下的计算公式：

$$level = \left(fmax - \left(fmax - fmin \right) / 3 \right) / 255 \qquad (4)$$

$fmax$ 和 $fmin$ 分别表示灰度图像的最大灰度值和最小灰度值，灰度阈值的选取直接决定了后续分割的效果。

二值化：在二值化处理之后，我们希望能得到黑底白字的二值化效果图，这是由于黑底白字的垂直投影直方图能更加便捷的投影出字符区域。图 13（a）所示为白底黑字的显示效果和垂直直方投影直方图，可以看出，有字符的区域位于波峰，间隔区域位于波谷，这种方式也能够对字符区域

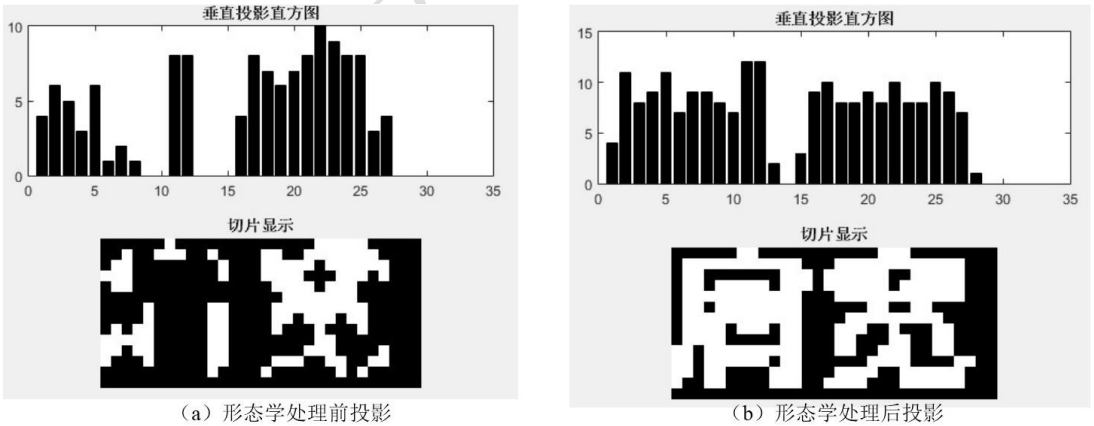
进行分割，但是分割位置并不准确，图 13（b）所示为黑底白字显示效果和垂直投影，可以看出间隔区域直方图统计为零，字符区域直方图数值较高。本文通过直方图的显示规律，来对这两种显示模式进行判断，如果直方图区域连续为正值，则判断为白底黑字，要对二值图像进行取反操作。



（a）白底黑字投影　　（b）黑底白字投影

图 13 两种二值化效果图

左右结构连通：在对二值化图像块进行垂直投影后，通过找到这些零值列和非零值列的转折点，即可判断字符的位置信息。但是由于 JPG 格式的图像高度压缩，导致局部字符放大后，单个字体模糊从而出现二值化后的字符连接中断，如图 14（a）所示，“肩”字由于横笔画较细，二值化时出现了中断，在垂直投影直方图上容易被才分为两个字符，此外，左右结构的字符也容易被拆分为两个字符。考虑到左右结构的字符以及这种中断连接的字符，内部结构的间距会比字符与字符之间的间距要小一些，因此本文采用形态学闭运算以及调小二值化阈值的处理方法，对这些图像块进行处理，由于字符较小，本文采用的处理模板为3×3，二值化阈值设定为$level\times0.8$，处理后的效果如图 14（b）所示。这种情况下就可以正确的投影出字符区域。



（a）形态学处理前投影　　（b）形态学处理后投影

图 14 汉字左右结构的形态学处理

（5）字符/背景区域的分类器训练与测试

训练集构建：字符检测的主要任务在于有效区分出字符区域和背景区域，这就需要训练出一个分类器，来对这些候选框进行分类。由于本次数据挖掘，题目给出了商品信息图片的标注信息，因此可以利用这些标注出的字符区域作为训练的正样本，背景区域作为负样本。在采集负样本的过程中，本文采用公式（3）计算候选字符区域与真实字符区域的重叠度，重叠度为 0 的候选区域为背景区域，即负样本。通过对所有图片进行遍历和随机抽取，本文共截取出了正负样本图像块个 1.5 万

个，图像块大小标准化为28×28 ，具体数据如图 15 所示。

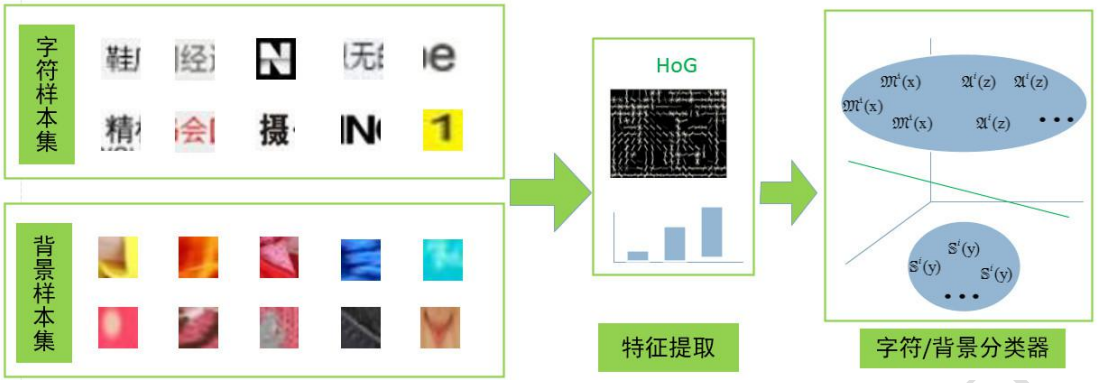


图 15 字符分类器训练

特征提取与训练：本文所采用的图像特征为梯度方向直方图 HOG，它通过计算和统计图像局部区域的梯度方向直方图来构成特征，这是一种对字符比较有效的图像特征。

方法一：基于 HOG 特征和 SVM 分类器的字符检测

梯度方向直方图通过计算和统计图像局部区域的梯度方向直方图来构成特征，这是一种对字符比较有效的图像特征。分类器采用经典的支持向量机，其通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，即通过统计如图 16 所示的支持向量，来获得一个最优的分割面。

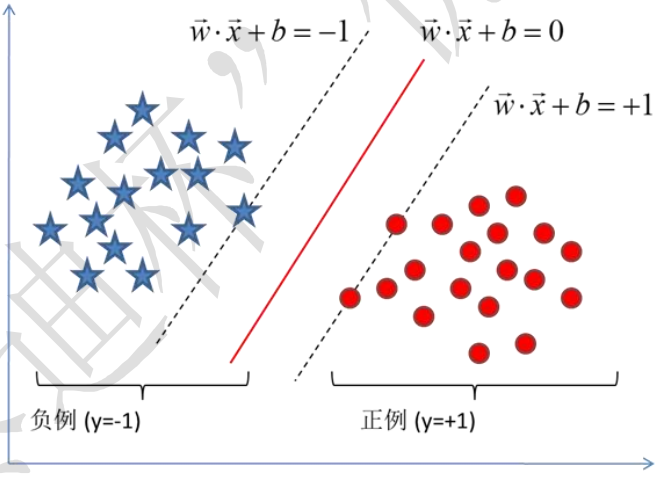


图 16 SVM 最优分类界面和支持适量示意图

给定一个训练样本$\{x_i,y_i\}$，其中$\{x_i,|i=1,2,...,1\}$为样本值，$y_i\in\{1,-1\}$表示类别标签。令将两类样本完全分开的函数为：

$$f(x)=\mathbf{w}^Tx+b \tag{5}$$

显然，如果 x 是位于超平面上的点则有$f(x)=0$。在这里我们要求对于满足$f(x)<0$的点，其对应的 y 等于-1，而$f(x)>0$ 对对应于 y 等于 1。为了寻找确定最优界面的参数以及考虑离群点的情况时，需要求解一个二次凸优化问题：

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t.\ \ y_i\left(\mathbf{w}^T x_i + b\right) \geq 1-\xi_i, i=1,2,...,n \\ &\xi_i \geq 0, i=1,2,...,n \end{aligned} \tag{6}$$

其中 $\xi_i \geq 0\,(i=1,2,...,n)$ 为松弛变量（slack variable），对应于数据点 x_i 允许偏离的函数间隔量，C 用于控制目标函数中寻找间隔最大的超平面和保证数据点偏差最小之间的权重，为 SVM 的一个寻优参数。

通过对 3 万个样本进行训练集和测试集的拆分，利用 libsvm 工具包，核函数选用 RBF 核，采用交叉验证的方式对参数 C 和 g 进行寻优，得到最优的参数 C=8，g=2，在测试正负样本集上得到 92.6%的正确率，并将模型参数进行保存。

样本预测：在测试阶段，上一步经过垂直投影后的字符检测区域如图 17（a）所示，可以看出字符字符区域都得到了有效的检测和分割，但是仍然存在部分背景区域被检测出来，本文对所有候选区域进行截取和正则化处理，提取 HOG 特征，然后输入 SVM 分类器，对于预测标签为-1 的区域，我们把它当做背景区域给予去除，最终的检测效果如图 17（b）所示，可以看出该分类器能够有效的对背景区域进行去除。

（a）文本行垂直投影效果图　　（b）SVM分类效果图

图 17 字符检测效果图

方法二：基于 LeNet 卷积神经网络的字符检测

图像特征的好坏直接影响字符和背景分类的精度，方法一采用对 HOG 特征是一种手工设计的特征，为了进行对比，本文还采取了基于卷积神经网络的分类方法，对于上述提取的 3 万个字符/背景数据集，采用 LeNet 的网络结构，如图 18 所示，这是一种经典的手写体识别网络，本文设置输出值为 2，即背景和字符两类。本文采用 Linux 系统下的 Caffe 工具包进行实现，在上述整理的字符/背景数据集上可以达到 97.1%的分类正确率。

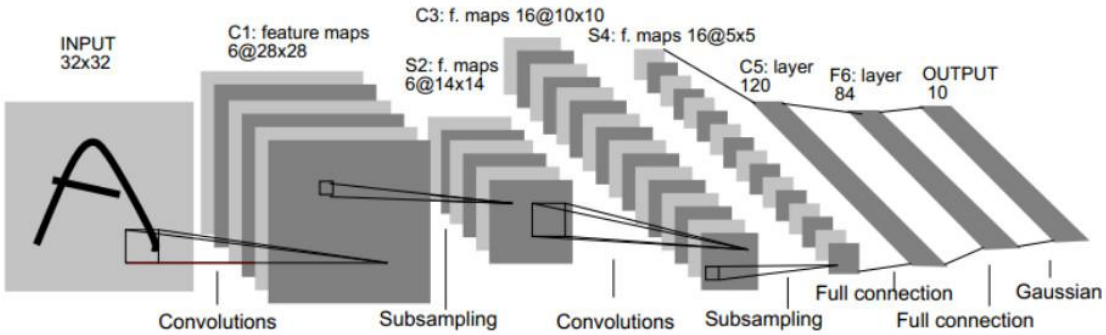


图 18 LeNet 网络结构

在测试阶段，对于每一幅图像提取出的候选框，输入 LeNet 网，输出标签为背景的候选框予以剔除。

方法三：基于 Fast-RCNN 的字符检测

在目标检测领域，R-CNN(Region CNN)、Fast-RCNN[4]、Faster-RCNN 是三种非常经典且一脉相承的方法，其中 R-CNN 是基础，其具体原理如图 19 所示，它把检测问题转化为分类问题，主要包括四个步骤——候选框检测、深度特征提取、SVM 分类器分类、检测框回归。详细过程如下：首先从图像中采用显著性方法提取出约 2000 个候选样本框，然后对这些候选框做正则化处理，输入预训练好的卷积神经网络提取深度特征，最终把这些深度特征输入 SVM 分类器，进而得到每个候选区域的类别，最后对检测框进行回归，以得到更加准确的位置。

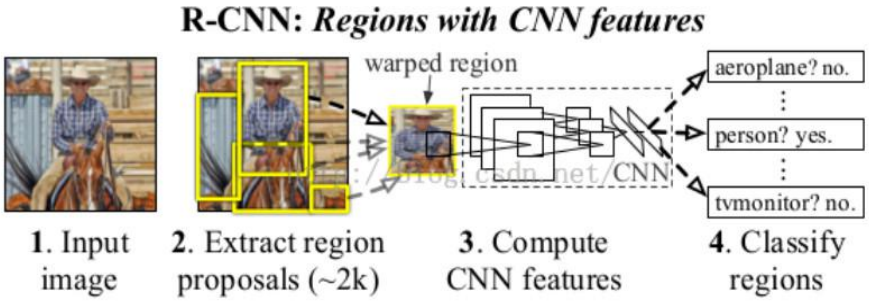


图 19 R-CNN 算法流程

由于 R-CNN 计算较慢，需要对每一个候选区域提取深度特征，并且需要把特征存储在硬盘上，会耗费大量的储存空间。因此作者又提出了改进后的 Fast-RCNN，流程图如图 20 所示，主要改进是对整幅图像提取深度特征，然后找到候选框在 Feature Map 中的位置，最后分类器采用 softmax 函数判别，且回归和分类统一为一个框架同时进行，因此大大加速了检测速度。

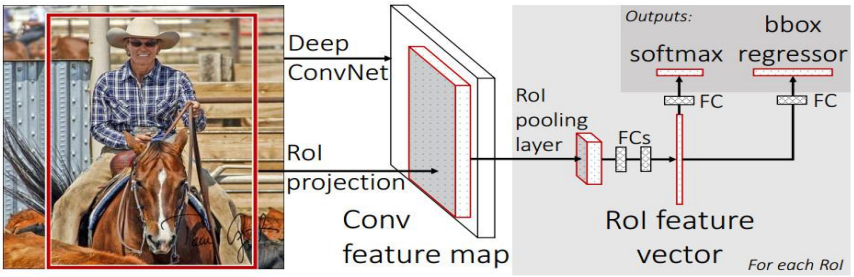


图 20 Fast RCNN 算法流程

后续的 Faster RCNN 又在 Fast-RCNN 的基础上，对提取候选框进行了改进，用深度网络取代了显著性检测，来进行候选区域的提取，速度和检测效果都得到了很大的提升。

本文的检测思路借鉴 RCNN 这一系列检测算法的思想，其优势主要有：①这种方法在常见 21 类目标检测中取得了非常好的检测效果；②可以有效的去除背景；③可以对最后的检测结果进行回归，因此检测更加准确。

在方法选取上，由于 RCNN 由于计算缓慢，且存储较大，因此不适合做大批量的字符检测，Faster RCNN 计算最快，但是其提取候选框的思想是采用神经网络的方法，提取的候选区域较少，而本文字符检测任务通常一幅图像有好几百个字符目标，因此这种方法也不适用。然而在采用 Fast-RCNN 的过程中，存在着一个很重要的问题在于，基于显著性检测的候选区域提取方法提取的候选框普遍分布在衣服、模特等区域附近，如图 21 所示，真正分布在字符区域的候选框很少，这样不利于后续的分类和回归，而本文采用 MSER 提取的候选框，大量分布于字符区域，再经过前期的先验知识筛选，可以作为 Fast-RCNN 进行字符检测的候选区域，因此本文用 MSER 算法提取候选框，然后采用 Fast-RCNN 对检测网络进行训练。输出类别为字符和背景这两类，算法采用作者公开的 Fast-RCNN 的代码，在 LINUX 系统下的 Caffe 平台进行实现。

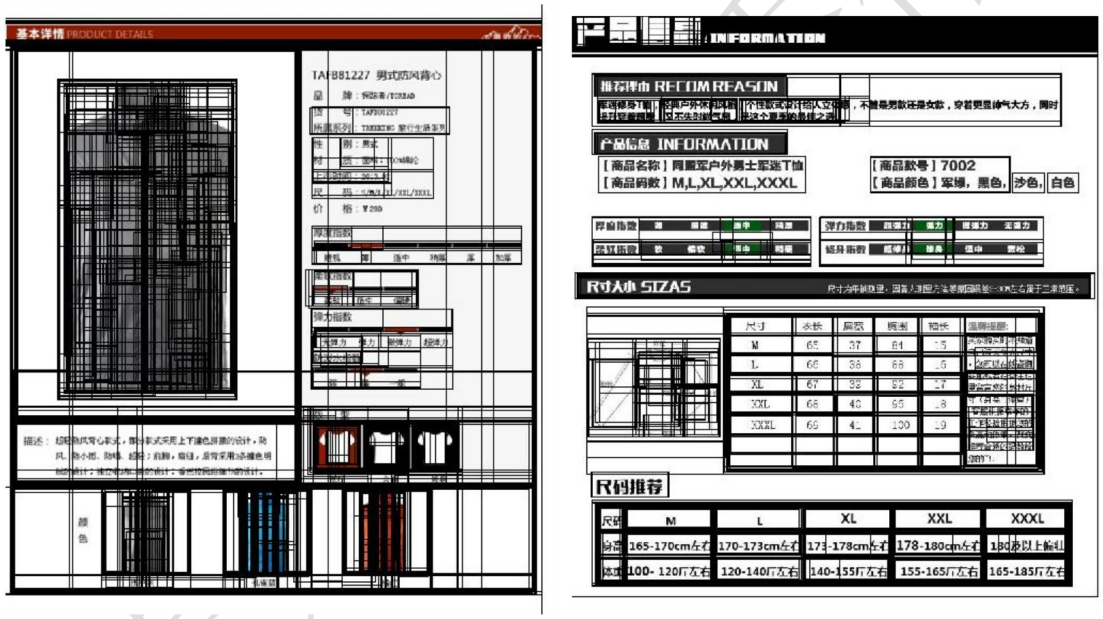


图 21 Fast-RCNN 中基于显著性提取候选框

在测试阶段，首先对待测试图像采用 MSER 算法提取候选区域，然后输入 Fast-RCNN 网络对每个候选区域进行预测分类，最终检测的结果如图 22 所示。

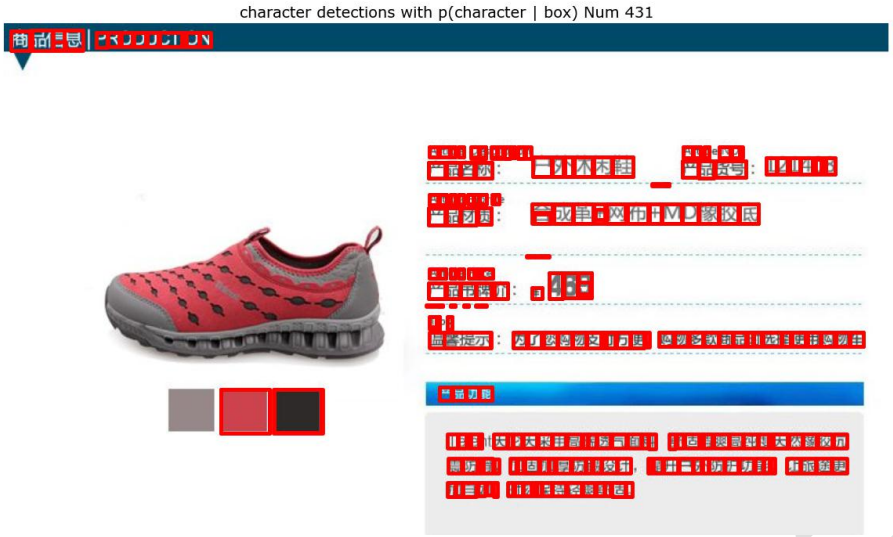


图 22 Fast-RCNN 检测结果

2.3.3. 字符识别

2.3.3.1. 字符识别关键问题

本题中给出的字符图像来源于商品图像，多为印刷体字符。字符识别即对商品图像中所提取的字符图像进行自动识别，相应的识别算法需要具备的能力包括：

① 有能力应对大规模字符集，其中大规模字符集指集合中字符种类多，字符图像多；

② 实时性好，算法必须是高效的；

③ 正确识别率高。

2.3.3.2. 当前字符识别算法

当前，印刷体字符图像的识别方法主要分为三大类[5]：

1、结构模式识别

印刷体字符是一种特殊的模式，其结构虽然比较复杂，但具有比较严格的规律性。换言之，其文字图形含有丰富的结构信息，所以可以设法提取含有这种信息的结构特征及其组字规律作为识别的依据。结构模式识别是早期印刷体文字识别研究的主要方法。其主要出发点是印刷体文字的组成结构。从构成上讲，印刷体文字是由笔划或更小的结构基元构成的。由这些结构基元及其相互关系完全可以精确地对印刷体文字加以描述，就像一篇文章由单字、词、短语和句子按语法规律所组成一样。所以这种方法也叫句法模式识别。识别时，利用上述结构信息及句法分析的方法进行识别，类似逻辑推理器。

然而，在实际应用中，该方法面临着诸多挑战。首先，与早期印刷体字符不同，现在的印刷体字符多字体，字体的不同，其结构基元也是不同的。其次，该方法抗干扰能力差，假如字符图像中存在干扰，如倾斜，扭曲，断裂，粘连，纸张上的污点，对比度差等，这些因素将直接影响到结构基元的提取，当结构基元不能准确地得到，后面的推理过程就成了无源之水。最后，结构模式识别的描述比较复杂，匹配过程的复杂度因此也较高。所以在印刷体文字识别领域中，结构模式识别的方法逐渐式微。

2、统计模式识别

统计决策论发展较早，理论也较成熟。其要点是提取待识别模式的一组统计特征，然后按照一定准则所确定的决策函数进行分类判决。印刷体文字的统计模式识别是将字符点阵看作一个整体，其所用的特征是从这个整体上经过大量的统计而得到的。代表性的方法有模板匹配、基于特征的模板匹配等。

统计特征的特点是抗干扰性强，匹配与分类的算法简单，易于实现。但不足之处在于细分能力较弱，区分相似字的能力差。另外，当数据集较大时，实时性差。

3、机器学习

在识别算法中，特征提取是重要环节。算法泛化性的好坏很大程度上受特征影响。在前两类识别方法中，字符特征均为经验性特征，在遇到背景变化的情况时，算法的识别率得不到保证。

机器学习技术可以在数据中挖掘好特征。特别的，深度学习作为当今机器学习的主流，在图像分类识别领域取得了非常大的成就。深度学习技术应用于字符识别并不少见，但对象往往是手写体数字或手写体汉字识别[6][7][8][9][10]。目前，深度学习应用于字符识别的常用模型有：卷积神经网络[8]、深度置信网络[11]等。卷积神经网络因其优异的识别性能，备受关注。

卷积神经网络（Convolution Neural Network）是神经网络的一种，区别于一般的神经网络，CNN的优点在于“权共享”，即让一组神经元使用相同的连接权，这可以节省网络训练的开销。以CNN进行手写数字识别任务为例，图23是一个简单的用于手写体数字识别的卷积神经网络。网络的输入是28×28的手写数字图像，输出是其识别结果，CNN复合多个卷积层和采样层对输入信息进行加工，然后在连接层实现与输出目标之间的映射。每个卷积层均包含多个特征映射，每个特征映射是一个由多神经元构成的“平面”，通过一种卷积核提取输入的一种特征。例如，图中的第一个卷积层由6个特征映射构成，每个特征映射是一个24×24的神经元阵列，其中每个神经元负责从5×5的区域通过卷积核提取局部特征。此外，为了使网络具有非线性的性能，卷积层的输出一般都会经过一个非线性的激活函数，传统的激活函数常常使用Sigmoid函数（近来，常常用修正线性函数ReLU代替）。采样层的作用是基于局部相关性原理进行亚采样，从而在减少数据量的同时保留有用信息。例如图中的第一个采样层有6个12×12的特征映射，其中每个神经元与上一层中对应特征映射的2×2邻域相连，并据此计算输出。通过复合卷积层和采样层，CNN将原始图像映射成120特征向量，最后通过全连接的方式与10维的输出特征向量相连，从而完成识别任务。

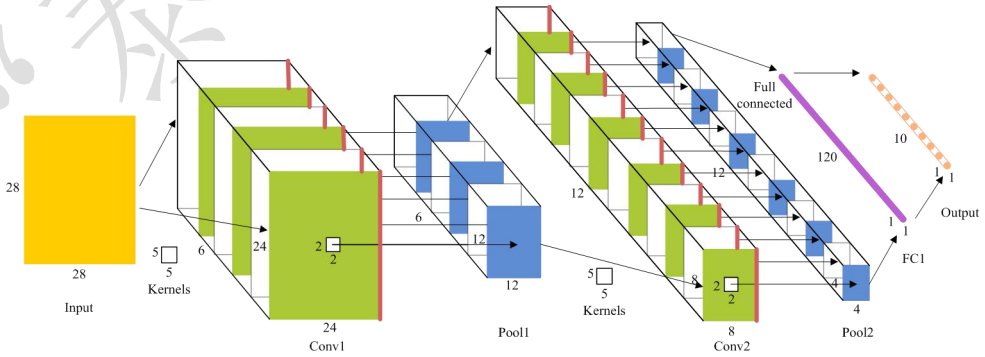


图23 一个用于Mnist识别的简单CNN模型

CNN可用BP算法进行训练，但在训练中，无论是卷积层还是采样层，其每一组神经元都使用

相同的连接权，从而可以大幅地减少需要训练的参数数目。

CNN 提取图像特征的过程嵌在网络除却全连接层的之前几层，图 23 中，指 Conv1-Pool2。可以看到，网络的特征是网络从训练数据集中学习到的，这使其在识别能力上比一般的非学习算法性能更优。且数据越多，网络学到的特征性能越好（基于数据正确、分布合理的前提），所以 CNN 适合处理大规模数据；虽然训练比较耗时，但是在得到网络参数后，识别仅是一个前向运算过程，相比于前两类识别方法而言实时性更好。

故本文选用基于 CNN 的字符识别方法。

2.3.3.3. 本文的 CNN 结构

现代 CNN 于 2012 年后发展迅速，至今，在图像分类及识别领域取得了不少成绩。伴随着向 ILSVRC 的不断挑战，公认的性能较优的 CNN 模型有：Alexnet、Zeiler/Fergus net、GoogLeNet、VGGnet、PReLUnet、ResNet 等。但是这类网络针对的是自然图像库，不能直接使用。

[12-15]介绍了一些用于自然场景中的字符识别方法，受他们的启发，我们自主设计了一个用于字符识别的卷积神经网络，如图 24 所示。网络结构为：1x28x28--20C5--MP2--50C5--MP2--500C4--2*l*N--Relu--*l*N--softmax。

其中输入图像是一张 28×28 的灰度图像；网络第一层为卷积层，共 20 个特征映射，5×5 的卷积核；第二层为最大值采样层，采样间隔为 2；第三层为卷积层，50 个特征映射，5×5 的卷积核；第四层为最大值采样层，采样间隔为 2；第五层为卷积层，500 个特征映射，4×4 的卷积核；第六层为全连接层，共 2l 个神经节点，其中 l 为样本中存在的字符种类数量；第七层为激活层，我们使用修正函数 ReLU 作为网络的激活函数；第八层为全连接+softmax 层，输出为 l 维的矢量 s，用于表达输入图像归属于各类别的得分（概率）。若 s 在第 j 处取得最大值，则认为输入图像归属于网络的第 *j* 类。

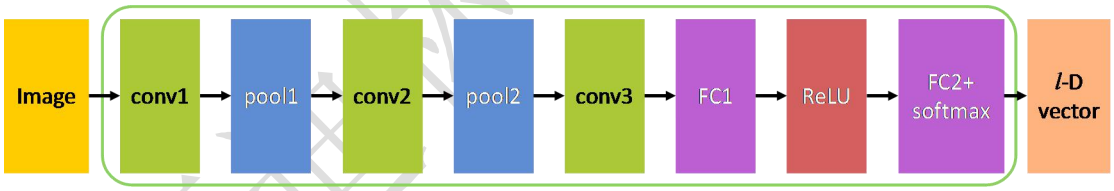


图 24 本文 CNN 结构

结构中，网络前五层在功能上为特征提取，后三层为模式识别。

具体实验中，我们从训练数据集中分出一部分作为验证数据集，通过观察训练过程中验证数据集的误差变化，可以判断网络是否产生了过拟合。

2.3.3.4. 字符图像“字亮底暗”预处理

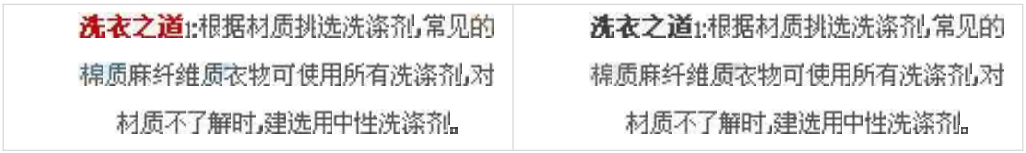
在使用网络进行识别之前，我们对输入的字符图像进行“字亮底暗”的预处理。这能有效缩小样本空间，从而提高网络的识别能力。以下篇幅中，我们将图像中的字称为前景，非字部分称为背景。具体的图像预处理操作如下：

1. 灰度化图像；

颜色是图像的重要信息，一般利于图像识别。之所以选择灰度化图像，主要有以下两点考虑：

1、为更好地利于用户了解商品，商品图像中的字符其前景与背景的色差一般都比较大，在灰度化后，图像中的字符基本能保证是清晰可认的，信息丢失可以接受；

2、若考虑颜色，网络将相应地变为三通道，这会导致网络的模型参数增多，调优困难，且训练时，对训练样本数量的需求会更大。



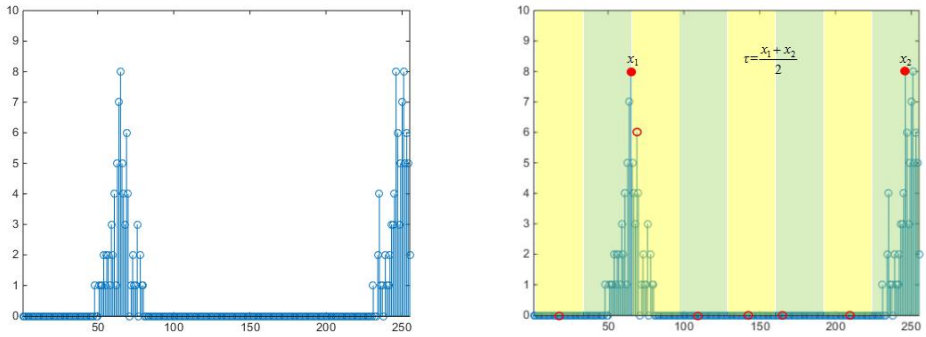
a 原图像　　b 灰度化后的图像

图 25 灰度化前后的字符图像

从灰度化前后的图像对比来看，在对字符图像进行灰度化处理后，字符信息仍是饱满的。

2. 描述图像灰度，寻找可以区分前景背景的阈值

样本中，字符图像的前景与背景之间一般具有比较大的灰度差异，图 26.b 第一张字符图像“洗”的灰度直方图如图 26.a 所示：



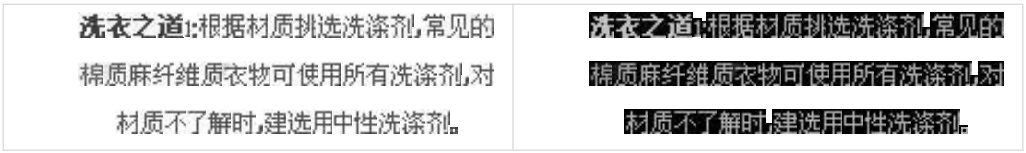
a 字符图像灰度直方图　　b 寻找阈值

图 26 对象：图 3.b 第一张字符图像“洗”

图 26.a 中，两个灰度峰清晰可见，它们对应着前景与背景,且前景背景的灰度值以区间的形式分布。阈值即可以将前景背景区分开的灰度值。我们将直方图分为 N 个区间，并分别在 N 个区间中找到其直方图的最大值，实验中，N=8。取 N 个区间中最大的两个值所对应的灰度值 x_1，x_2，可以认定为前景背景的灰度值，最后取其平均值作为阈值τ。如图 26.b 所示。

3. 确定前景背景

在上一步的过程中，我们找到了可以将图像分为前景与背景两部分的阈值，但是我们并不能就此确定前景是灰度值大的部分还是小的部分。事实上，字符图像存在两种模式：“字亮底暗”与“底亮字暗”。我们可以在图像的四周寻找灰度值小于阈值的像素点，若总数多于总像素点的一半，则该图像为“字亮底暗”模式，否则为“底亮字暗”模式，对“底亮字暗”模式我们要进行反色操作。



a. 灰度图像　　b. “字亮底暗”处理后的图像

图 27 “字亮底暗”处理前后的字符图像

实验结果显示，该方法能有效对“底亮字暗”的模式进行反转。需要注意的是，我们并没有将图像进行二值化，使其成为“黑纸白字”，其原因是因为绝大多数的字符图像并不具备可以二值化的条件，以下是根据阈值τ所做的二值化处理图像。

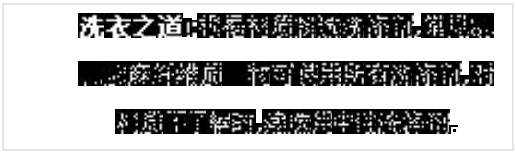


图 28 二值化处理后的字符图像

2.3.3.5. 网络的训练与测试

对训练图像进行预处理后，就可以进行网络的训练与测试。在 1000（预处理后实际采用 992 张）张图像的训练集中，应用图 23 的 CNN 结构，可以在 500（实际可用 498）张图像的测试集中得到 93.07%的正确识别率。

然而，在深入研究了训练样本集后，我们发现训练集中各类别**样本数量分布极不均匀**。所以应当对模型做进一步的改进。

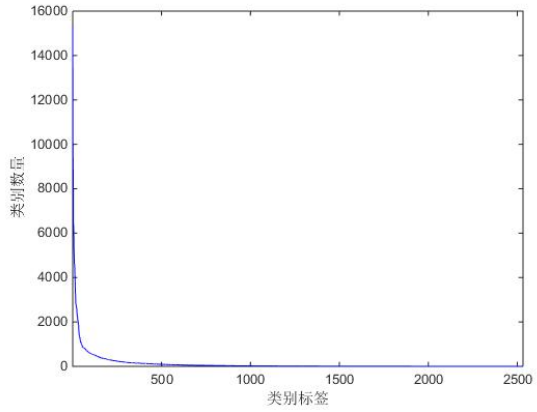


图 29 图像集各类别样本数量分布极不均匀

如图 29 所示，训练图像集中共有 2532 类字符，数量多的字符达到成千上万，数量少的字符却只是个位数。在这种训练图像集的基础上，训练出的网络有过拟合的倾向，对少样本字符的识别能力较差。

针对种类样本数量不均匀的问题，最本质的解决方案是增加少样本字符，在训练图像集确定的情况下，可以通过加噪声、旋转平移等方法增加少样本字符。但是鉴于本题训练集中，种类样本数量分布极不均匀，这种方法在具体实践时难度较大。于是，我们考虑以下 4 种思路：

1. CNN+HOG

为所有 2532 类字符配置若干 HOG 特征作为模板，在 CNN 对待识别图像的响应 s_j 小于阈值τ时，用模板匹配的方法进行识别。具体方法为：

假设某类字符图像在训练集中共 N 张，我们随机在这些图像中抽选 n 张制备其对应的 HOG 特征模板，实验中，取 n=10。若 $N<n$，则取 $n=N$。实验中，HOG 特征取 6×6×9=324 维。最后我们可以在训练集中得到 324×15831 的 HOG 特征字典，其中 324 为 HOG 特征的维度，15831 为实际提取的模板数。

阈值τ不宜过大亦不宜过小，事实上，当τ=0.0 时，就是单纯使用卷积神经网络识别的情况，当τ=1.0 时，就是单纯使用模板匹配识别的情况。暂时可取值τ=0.8。

2. Ensemble CNN（集成 CNN）

将 1000 张训练样本图平均分为 N 组，虽然分开后的样本集仍然存在样本不均匀的情况，但是相比于原始情况而言，样本不均匀的问题会有所好转。本文中，取 N=10。子样本类别分布情况如图 30 所示。

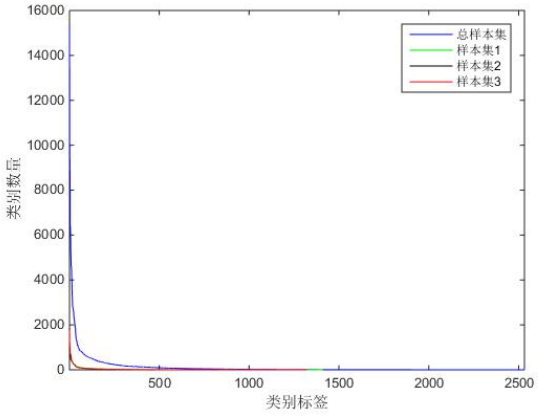


图 30 子样本集分布示意图

我们按照图 23 的卷积神经网络结构分别对这 N 个训练样本集进行训练，可以得到 N 个卷积神经网络。需要注意的是，这些网络所能表示的字符类别各不相同。

我们按如下的流程图得到最后的识别结果：

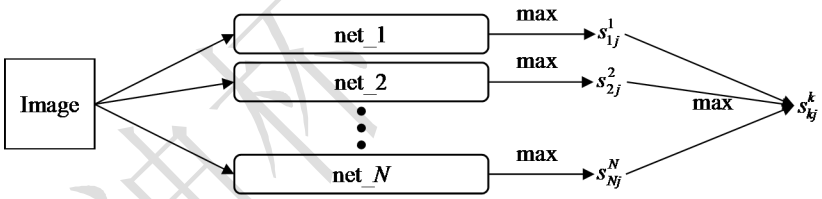


图 31 集成 CNN 工作流程图

令其中某一 CNN 标号为 net_i（i=1,...,N），该网络对输入图像的响应为 s_{ij}^i，表示 net_i 认为图像以 s_{ij}^i 的概率归属于网络 i 的第 j 类。将所有 N 个网络均应用于输入图像的识别，那么不同的网络可能将输入图像归属于不同类，概率分别为 s_{ij}^i（i=1,...,N）。我们寻找这些概率的最大值 s_{kj}^k，$s_{kj}^k = \max_{i=1,...,N}(s_{ij}^i)$，即表示输入图像最大概率地归属于网络 k 的第 j 类。

3. Double CNN（双网）

一种解决少样本问题的经典方法是在预训练的网络上进行 fine-tuning[16]。

我们将样本集合按照其样本种类数量分为相同的两堆，即将 2532 类字符分为 1266 类两份。如图 32 所示：

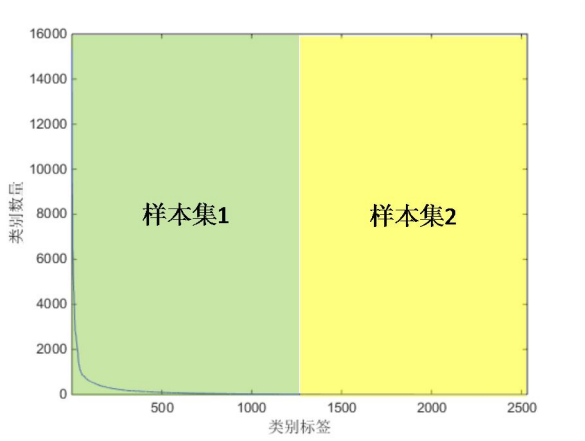


图 32 按样本数量将样本集合分为种类数相同的两堆

其中，样本集 1 中样本较多，样本集 2 样本较少。若我们分别在这 2 个样本集中训练图 23 的 CNN，那么这两个网络在结构上是完全相同的。经验地，由于样本集 1 中样本较多，所以训练得到的网络 net_1 在识别性能上会优于样本集 2 上的网络 net_1。于是我们可以先训练 net_1，并用 net_1 前五层参数（特征提取）初始化 net_2 的前五层参数。net_2 只需要在 net_1 的基础上微调即可。

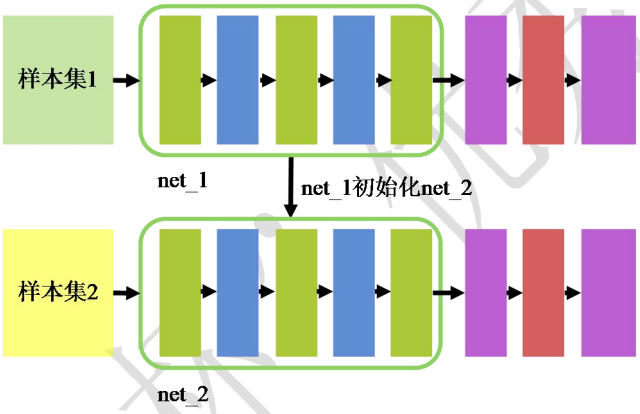


图 33 双网中，net_1 是 net_2 的预训练网络

在获得了两个 CNN 后，用与集成 CNN 中相同的方法确定输入字符图像的最终归属。

4. Transfer CNN（迁移 CNN）

训练样本集中共有 2532 类字符，包括中文、英文、标点等，但实际上，此 2532 类字符并不能涵盖所有常用字符。单就常用汉字而言就有 3500 余种。所以，实验中不仅存在少样本问题，还存在零样本问题，于是我们考虑合成一个印刷体字符的训练样本库。

我们在 windows 系统中下载了常用的字体格式，并对每一个常用字符（包括汉字、标点、英文）合成了不同字体的文字。以“阿”为例，样图如图 34 所示：

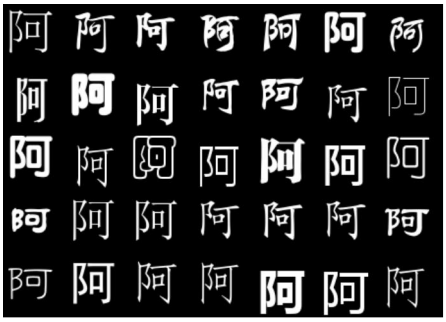


图 34 自主合成的字符训练集

此外，为了增加数据量，我们对每种字体图进行简单的位置平移变换。对每一类字符，均制备了 300~400 个样本，最终生成了字符数为 210 万的字符训练集。

沿用图 23 的网络结构，我们对该合成数据库进行训练，并将训练后的网络应用于测试集的识别。原计划包括在该网络的基础上进行实际训练样本的 fine-tuning，但由于时间仓促，我们遗憾未能将该想法付之于实验。

2.4. 结果分析

2.4.1. 字符检测结果分析

对于测试集给出的 500 幅图像，本文通过预处理的方法对错误 box 文件进行删除，得到 498 个测试样本。采用 F1-score 的评价方式来评估检测结果，具体计算公式如下：

$$f_1 = 2/\left(1/p+1/r\right) \tag{7}$$

其中 p(precision)是准确率，r(recall)是召回率，这两者的计算公式如下：

$$p=\frac{\#\text{正确检测}}{\#\text{检测样本}} \qquad r=\frac{\#\text{正确检测}}{\#\text{真值样本}} \tag{8}$$

p 和 r 只要有一个很低，$F1$ 值都会很低。

判别检测正确的定义如下：①检测框中心与标定框中心距离/最大检测框和标定框的边长<0.15，②最小检测框和标定框的边长/最大检测框和标定框的边长>0.8。

本文对其理解如下：①检测框与标定框距离要小，且小于 max（BOX_DET_w,BOX_DET_h, BOX_GT_w,BOX_GT_h）*0.15; ②检测框和标定框边长要接近，即 min(BOX_DET_w, BOX_GT_w)/max(BOX_DET_w,BOX_GT_w)>0.8,min(BOX_DET_h,BOX_GT_h)/max(BOX_DET_h,BOX_GT_h)>0.8。③这三个条件必须同时满足。其中 BOX_DET 和 BOX_GT 分别表示检测框和标定框，后缀 w 和 h 分别表示宽和高。

在这种评价准则下，本文对比了三种检测方法的结果，分别是 HOG+SVM、LeNet-CNN、Fast-RCNN。三者均采用 MSER+筛选策略进行候选框的提取，然后采用三种分类器进行背景和字符的分类。第一种方法在 Windows 系统下基于 MATLAB 进行实现，后两种方法在 LINUX 系统下基于 caffe 平台和 PYTHON 语言进行实现。台式机配置为 i7 处理器，GTX960 显卡，16G 内存。

(1)平均耗时

表 1 检测平均耗时统计

	MSER+筛选策略	HOG+SVM	LeNet-CNN（GPU）	Fast-RCNN（GPU）
时间(单位：秒)	9.34s	1.53s	1.12s	0.86s

从计算时间上来看，单幅图像耗时最多的在于 MSER 提取候选框的部分，其中经历了筛选、投影、面积计算等过程，因此比较耗时，后续的背景和字符分类器耗时较短，采用 GPU 加速可以大大提高检测时间。

（2）HOG 特征和深度特征对比

HOG 特征是一种手工设计特征，深度特征是一种通过大数据学习得到的特征，为了评估这两种特征区分字符和背景的性能，本文分别对其进行了测试和特征可视化。对整理得到的字符和背景数据集，用正负样本各 10000 个进行训练，剩余正负样本各 5000 个进行测试，用分类正确率来评价分类精度。其中 SVM 分类器采用交叉寻优的方式，得出 c=8，g=2,表 2 给出了两种特征的分类精度,可见深度特征对于字符和背景分类效果更好更佳。

图 37 为 LeNet 网络结构提出的各层深度特征，图 38 为 HOG 特征和深度特征的特征分布图，可视化工具为 t-SNE 工具包，把特征进行降维显示。由图可见，深度特征的区分性更好，这也进一步验证了深度特征在区分背景和字符方面有着更好的效果。

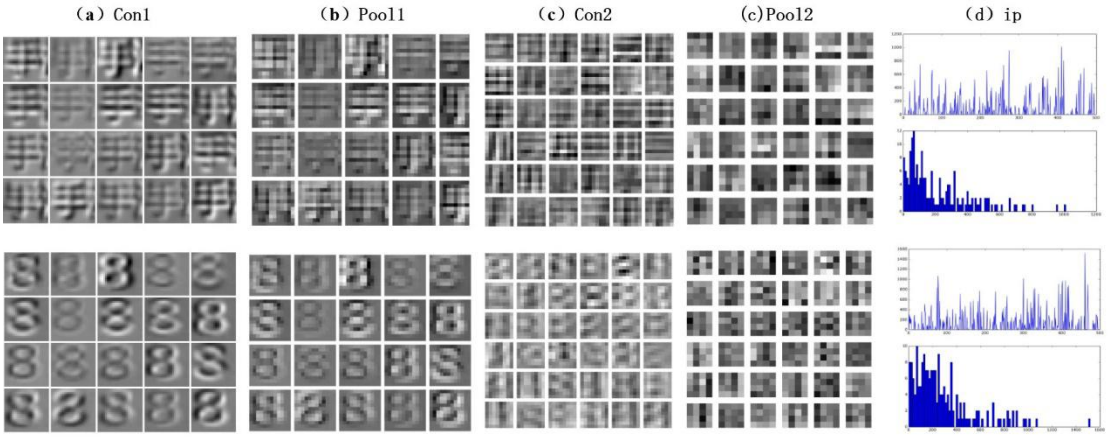


图 37 LeNet 网络提出的深度特征

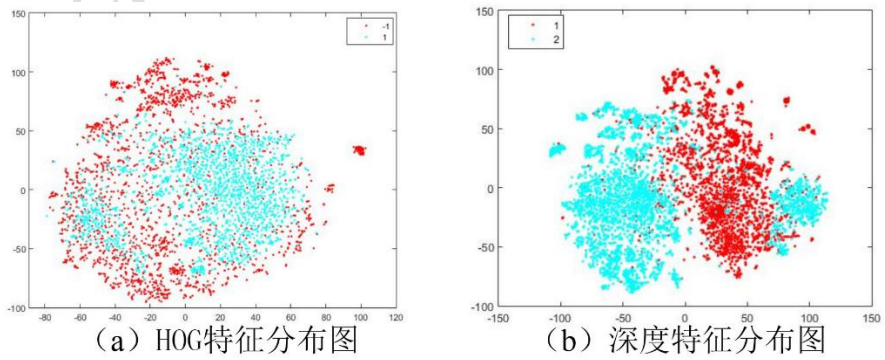


图 38 特征分布图

表 2 背景/字符分类精度

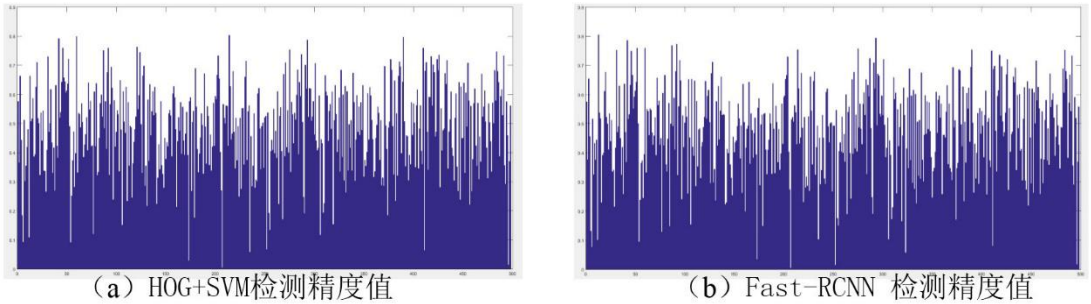
	HOG+SVM	LeNet-CNN
Acc	92.6%	97.4%

（3）检测方法精度对比

为了便于批量化的对比实验结果，本文对 HOG+SVM 和 Fast-RCNN 这两种方法进行检测评估。两种方法在不同的系统下运行，Fast-RCNN 的筛选阈值设定为 0.1，最大极值抑制阈值为 0.2，结果都保存为题目规定的.box 格式，然后调用计算程序对真值框和检测框进行比较，计算得到 F1 的值，图 39 给出了两种方法在 498 个测试样本中的 F1 值。表 1 给出了两种方法的均值，对比可以发现，基于 Fast-RCNN 的方法 F1 的检测值要高，这是因为 Fast-RCNN 对候选框进行了打分和回归，因此检测的精度更高。

表 3 检测精度评估

	HOG+SVM	Fast-RCNN
F1	0.478	0.524



（a）HOG+SVM检测精度值　　（b）Fast-RCNN 检测精度值

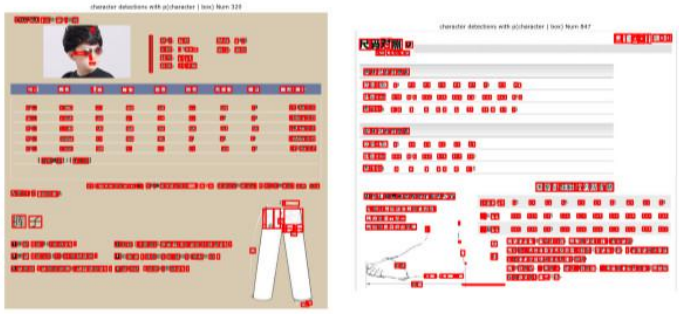
图 39 检测精度图

(4) 部分正确检测与错误检测结果分析

图 26 给出了部分正确检测结果对于背景比较简单的图片，本文的检测算法效果较好，基本能够检测出大部分区域。其中，Fast-RCNN 的方法检测的效果更佳。

图 27 给出了部分错误检测情况。主要难以检测的情况出现于汉字的结构拆分、水印难以检测出来、字符行的倾斜等。

（a）HOG+SVM检测结果



（b）Fast-RCNN检测结果

图 26 部分正确检测结果

2.4.2. 字符识别结果分析

方法一中，我们分别在τ=0.0、0.2、0.4、0.6、0.8、1.0 时做了相关的识别实验，实验结果如图 35 所示。单网 CNN（τ=0.0）在识别率上是最高的，正确识别率达到 93.07%；而单纯使用模板匹配识别的方法（τ=1.0）在识别率上是最低的，仅得到了 68.08%的正确率。

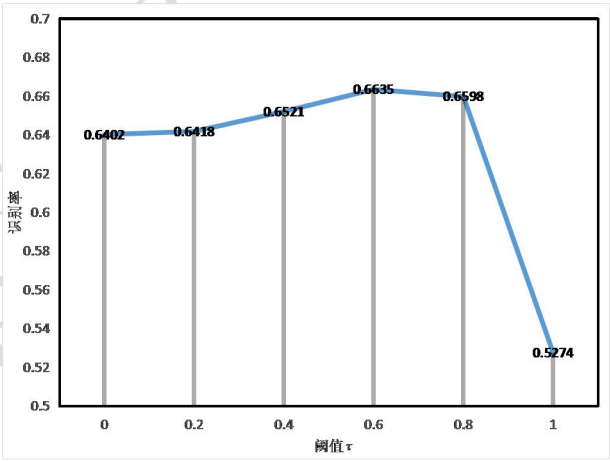


图 35 阈值τ对 CNN+HOG 方法的影响

从图 35 中可以看到，随着阈值τ的变大，网络的识别率不断减小，这说明用 HOG 特征配合 CNN 的方法并不能解决数据分布不均匀的问题，此外，识别率在τ=0.8 之前都比较稳定。分析其原因，可能的原因有以下两点：1、在大数据情况下，图像数量多、背景复杂，少量的模板不足以完成字符识别任务；2、CNN 网络的输出置信度高，模板匹配方法配合困难。

方法二的正确识别率为 88.65%。

集成学习通过构建并结合多个个体学习器来完成学习任务，得到的效果往往能好于单个个体学习器。但是该方法的识别率比单网识别率低。我们认为这是由于训练集的缩小导致的。事实上，由于样本集的缩小，实验中产生的 10 个网络对测试集的单独识别率只分布在 40%~60%之间。该方法的识别率差强人意，缺点在于对时间的要求较高。然而，随着硬件多核技术的发展，这个缺点对于并联结构的网络来说并不是致命的。

方法三的正确识别率为 86.85%

该方法的正确识别率不及单网识别。原因可能为以下三方面：1、虽然这种微调模式在一定程度上可以改进少样本问题，但是样本集 2 中的样本实在太少；2、样本集 1 本身并不是一个完美的大数据图像库，网络的泛化性不强；3、和方法二相同，双网的训练集缩小了，实验表明，大样本集 1 所对应的 net_1 应用于测试集的识别率高达 92.09%，而小样本集的 net_2 识别率仅为 0.74%，所以将两者集成在一起似乎并不是一个好的选择。

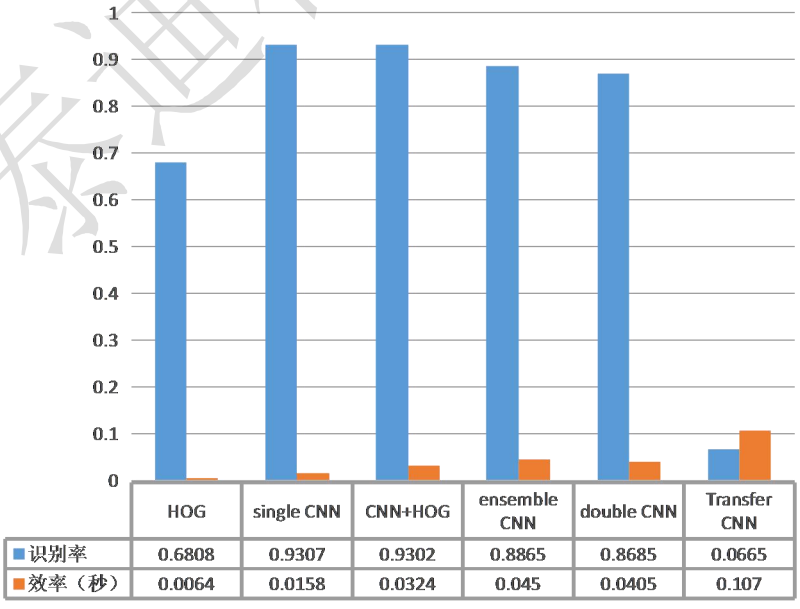
方法四的正确识别率为 6.65%。

虽然方法四的识别率很低，但是在理论上可以接受。实验反映了一些技术事实，具有很大的研究价值：

首先，虽然该网络的识别率不高，但是网络是具备字符识别能力的。网络的识别率之所以不高，其主要原因在于网络的训练数据集是合成的，网络并没有用到实际的训练数据，一方面合成图像并不是字符的最小外接矩形框、另一方面，合成图像是二值化的，与实际的字符数据差别较大。且由于标点难以统计，在数据的准备中难以避免会有所疏漏。

其次，为了提高网络的识别率，和方法三类似，我们可以在该网络的基础上进行实际训练样本的 fine-tuning，将该网络所学习到的信息迁移至本题的训练样本中，这在不少的文献中都有提及，可供参考。我们仍在努力实现该想法。

综上，上述的 1~3 方法并没有真正地解决样本分布不均匀的问题。其本质原因在于没有从根本上改变训练样本库的种类分布。方法四是最值得期待的方法！各方法的识别率、效率如下图所示：



	HOG	single CNN	CNN+HOG	ensemble CNN	double CNN	Transfer CNN
识别率	0.6808	0.9307	0.9302	0.8865	0.8685	0.0665
效率（秒）	0.0064	0.0158	0.0324	0.045	0.0405	0.107

图 36 本文各种识别方法的识别率、效率对比

就效率而言，简单的 HOG 匹配方法效率最高（建立在小字典的基础上），迁移 CNN 耗时最多，这是因为其全连接层神经元个数较多。

综合考虑各种因素后，选择单网 CNN 作为本文的字符识别算法。然而，我们还是认为在解决少样本、零样本问题，迁移 CNN 的潜力最大。

2.4.3. 检测识别效果评估

本文采用 F2-score 的评价方式来评估全部的检测识别结果，具体计算公式如下：

$$f_2 = 2/\left(1/p+1/r\right) \tag{9}$$

其中 p(precision)是准确率，r(recall)是召回率，这两者的计算公式如下：

$$p=\frac{\#\text{正确检测识别}}{\#\text{检测样本}} \qquad r=\frac{\#\text{正确检测识别}}{\#\text{真值样本}} \tag{10}$$

其中判定为正确检测识别的规则是在公式（8）的基础之上再判断本文识别的字符与标定的字符是否相等。

本文采取效果较好的 Fast-RCNN 作为检测算法，识别算法采用单网模型，首先采用检测算法提取检测框，然后调用识别算法对每一个检测区域进行识别，结果存入.box 格式文件，最后调用评估算法计算 F2-scorez 值，表 4 给出了 498 个测试样的平均统计结果，其中 Recognition rata 表示在正确检测基础上识别的正确率。较在训练集上测试的识别的正确率有所下降，这可能是因为部分像素截取位置的偏差，会影响识别的精度。图 N 给出了 498 个样本的参数值，可以看出不同的样本图像差别较大，最低的 F2 值有接近零的，最高的可以达到 0.61，由此可见，针对不同的图像，需要设定不同的阈值，选取不同的方法，接下来应该进一步在阈值自适应和图像种类分类讨论方面下功夫。

表 4 检测识别结果统计

	precision	recall	Recognition rata	F2-score
本文方法	0.3093	0.2435	70.64%	0.2676

“泰迪杯”获奖作品

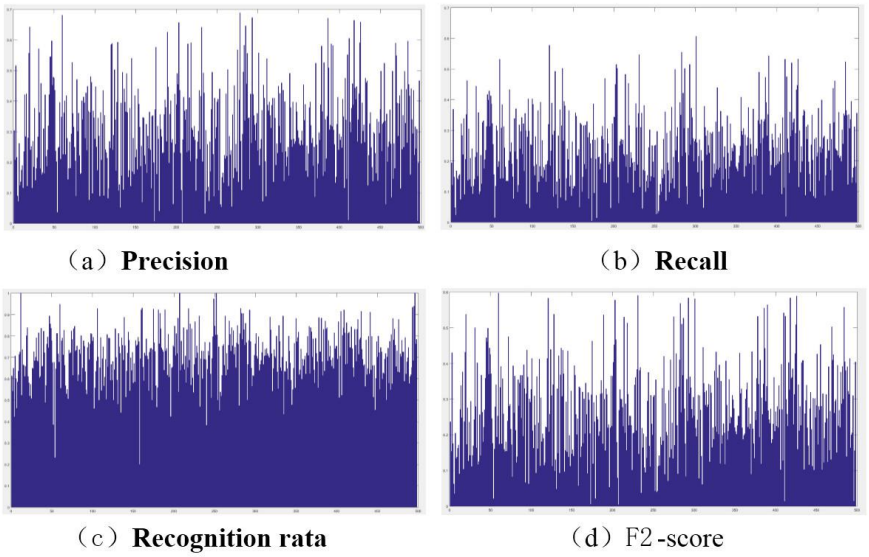


图 40 本文检测识别结果

2.4.4. 软件介绍：

为了便于直观地展现本文对商品信息图片文字检测与识别的结果，本文利用 GUI 绘制了简单的软件操作界面，可视化字符检测和字符识别的过程，软件界面如图所示：

图 40 软件界面

具体操作如下：

1. 点击“打开图片”，选择待处理的商品信息图片；

图 41 (a) 打开图片 (b) 字符检测

2. 点击“字符检测”，弹出对话框显示“正在检测……”，当检测完成后，图片上会显示检测出的红色 box 框；

图 42 (a) 检测结束 (b) 字符识别

3. 点击“字符识别”，出现进度条提示识别进度。当识别结束后，右侧会显示检测出的字符。

4. 双击 Listbox 中的字符，可以显示对应的检测字符位置。

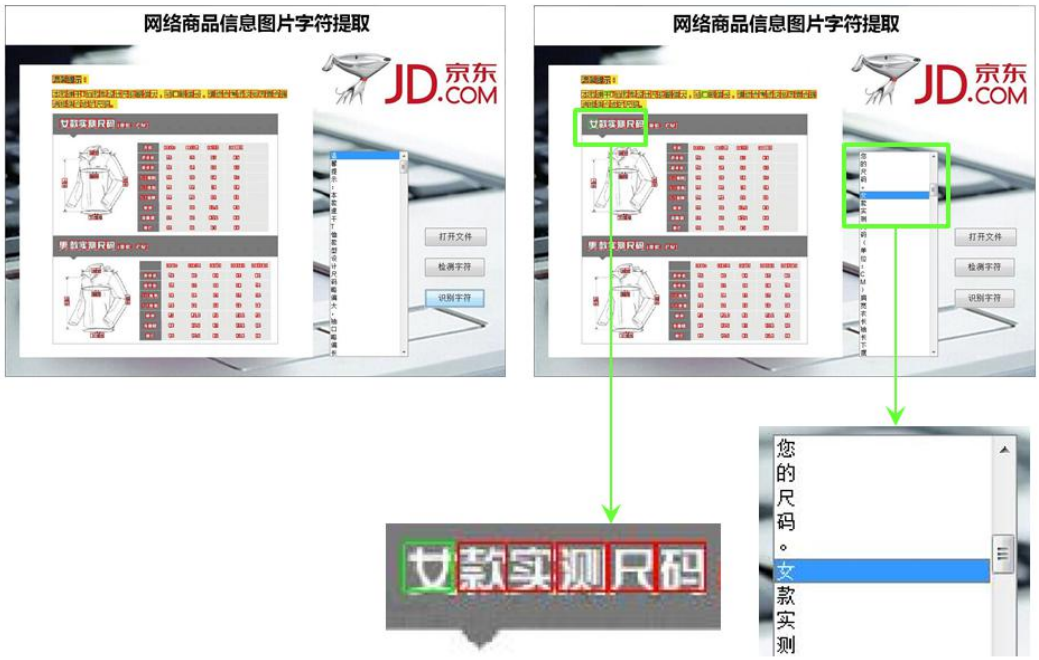


图 43 (a) 识别结束 (b) 对应字符显示

同时，本文利用“汉王 ocr”软件对商品信息图像进行了处理，通过实验可以看出，“汉王 ocr”对复杂商品信息图片中的光学字符提取也仍旧存在一定问题，对比普通白色背景下 pdf 文件字符提取效果，商品信息图片中主要存在字符检测大量漏检，检测出的字符普遍识别正确。可见，在此业务领域字符的高效准确检测仍然存在重大的技术挑战，而识别的过程中，字符库的构建十分重要，“汉王 ocr”拥有庞大的字符数据库，因此在识别过程中，其性能相对稳健。

图 本文检测方法与汉王检测对比

3. 结论

本文通过对预处理后的商品信息图片中的光学字符，利用基于 MSER 算法、SVM 分类器、卷积神经网络等方法建立了多种数据挖掘模型，得到了具有价值的挖掘结果，实现了对商品信息图片中字符的检测和识别，对网络监管提供了有力的技术支持。通过字符检测算法，构建离线和在线处理过程，实现了复杂多样字符的识别。利用大型数据训练出的识别网络，解决了传统方法上低效、样本不足等缺点，能够有效识别字符。对于本文给出的挖掘结果，可为进一步获得商品语义信息提供技术支持。

从分析结果来看，在字符检测部分，由于商品信息图片背景复杂，包含字符数量较多，字符字体颜色大小变化多样，算法中的阈值等各参数难以调整至获得某种意义上的最优。因此算法普适性仍有待提升。在字符识别部分，通过大型字符数据库训练深度学习网络，线下构建大型字符数据库需要投入较多的准备工作。并且，这样一种基于大数据驱动的数据挖掘技术对硬件要求较高，未来在突破硬件条件限制的情况下，识别能力将有更大的提升。本文方法在字符检测与识别的相关领域进行了深入的学习和研究，当前检测和识别是两个独立的挖掘过程，后续将致力于基于深度学习的字符检测方法研究以提高检测算法的普适性，进一步将检测的深度网络和识别深度网络融为一个整体。

4. 参考文献

[1] Ye Q, Doermann D. Text Detection and Recognition in Imagery: A Survey[J]. IEEE Trans. PAMI, 2015, 37(7):1480-1500.

[2] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[C]// 2013 IEEE Conference on CVPR. IEEE, 2010:2963-2970.

[3] Huang W, Qiao Y, Tang X. Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees[C]// Computer Vision–eccv. 2014:497-511.

[4] Girshick R. Fast R-CNN[C]// IEEE (ICCV). IEEE, 2015:1440-1448.

[5] 梁涌. 印刷体汉字识别系统的研究与实现[D]. 西北工业大学, 2006.

[6] Cireşan D, Meier U, Gambardella L, et al. Deep, Big, Simple Neural Nets forHandwritten Digit Recognition[J]. Neural Computation, 2010, 22(12):3207-3220.

[7] Lecun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zipcode recognition[J]. Neural Computation, 1989, 1(4):541-551.

[8] Cun, Y. Le, Boser, B, Denker, J. S, et al. Handwritten digit recognition with a back-propagation network[C]// Advances in Neural Information Processing Systems.Morgan Kaufmann Publishers Inc. 1990:465.

[9] Schmidhuber, Jurgen. Multi-column deep neural networks for image classification[C]// IEEE Conference on CVPR. 2012:3642-3649.

[10] Jarrett K, Kavukcuoglu K, Ranzato M, et al. What is the Best Multi-StageArchitecture for Object Recognition?[C]// ICCV. 2009:2146 - 2153.

[11] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets.[J].Neural Computation, 2006, 18(7):1527-54.

[12] Netzer Y, Wang T, Coates A, et al. Reading Digits in Natural Images with Unsupervised Feature Learning[J]. Nips Workshop on Deep Learning & Unsupervised Feature Learning, 2011.

[13] Sermanet P, Chintala S, Lecun Y. Convolutional neural networks applied to housenumbers digit classification[J]. 2012:3288-3291.

[14] Contes A, Carpenter B, Case C, et al. Text Detection and Character Recognition inScene Images with UnsuperVised Feature Learning[C]// International Conferenceon Document Analysis & Recognition. IEEE, 2011:440-445.

[15] T. Wang, D. J. Wu, A. Coates, et al. End-to-end text recognition with convolutionalneural networks[C]// ICPR. IEEE,2012:3304-3308.

[16] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, et al. Deep Learning for Content-Based Image Retrieval[C]// the ACM International Conference. 2014.