

# A 题 电商平台图片中文字的识别

## 一、背景：

由于电子商务环境下用户无法接触商品实物，电商网站提供的商品信息是用户做出购买决定的重要依据。不同的电商平台展现商品信息的风格虽然各有不同，总的来说商品信息除了一部分以文本、表格形式给出外，越来越多地以图片的形式呈现。图 1 是京东商城上某电水壶商品页面的示例。



图 1、京东商城的某商品页面

京东商城商品页面中的“规格参数”部分以文本表格的形式给出了商品的常规属性，对特定的商品类目这部分往往有比较固定的内容格式，如图 2 所示。

|                                                                                |                                                                                               |
|--------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| 料理机<br>电压力锅<br>豆浆机<br>微波炉<br>电磁炉<br>煮蛋器<br>电饭锅<br>电饼铛<br>电烤箱<br>其它厨房电器<br>电热饭盒 | 榨汁机<br>电压力锅<br>咖啡机<br>电烤箱<br>面包机<br>酸奶机<br>电水壶/热水壶<br>多用锅<br>电饼铛<br>电烤箱<br>其它厨房电器<br>养生壶/电热水壶 |
| 同类其他品牌                                                                         | 美的<br>苏泊尔<br>科立泰<br>北鼎<br>SKG<br>欧博<br>MION                                                   |
| 九阳 (Joyoung)                                                                   | 九阳 (Joyoung)                                                                                  |
| 九阳 (Joyoung) JYK-17S08电热水壶                                                     |                                                                                               |

|        |              |      |              |      |      |                    |       |
|--------|--------------|------|--------------|------|------|--------------------|-------|
| 商品介绍   | 规格参数         | 包装清单 | 商品评价(189533) | 售后服务 | 京东服务 | 客户服务热线: 4006789799 | 加入购物车 |
| 品牌     | 苏泊尔          |      |              |      |      |                    |       |
| 型号     | SWF17K2-180  |      |              |      |      |                    |       |
| 颜色     | 不锈钢色         |      |              |      |      |                    |       |
| 类型     | 普通电水壶        |      |              |      |      |                    |       |
| 材质     | 拉丝不锈钢        |      |              |      |      |                    |       |
| 容量     | 1.7升         |      |              |      |      |                    |       |
| 温控器类型  | 其它           |      |              |      |      |                    |       |
| 加热方式   | 底盘加热         |      |              |      |      |                    |       |
| 360度底座 | 支持           |      |              |      |      |                    |       |
| 电动出水   | 不支持          |      |              |      |      |                    |       |
| 滤网材质   | 304不锈钢       |      |              |      |      |                    |       |
| 自动断电   | 支持           |      |              |      |      |                    |       |
| 保温功能   | 不支持          |      |              |      |      |                    |       |
| 适用人数   | 2-5人         |      |              |      |      |                    |       |
| 技术参数   |              |      |              |      |      |                    |       |
| 额定电压V  | 220v         |      |              |      |      |                    |       |
| 额定频率Hz | 50Hz         |      |              |      |      |                    |       |
| 额定功率W  | 1800W        |      |              |      |      |                    |       |
| 产品净重kg | 1.1          |      |              |      |      |                    |       |
| 产品毛重kg | 1.3kg        |      |              |      |      |                    |       |
| 包装尺寸mm | 205*174*227  |      |              |      |      |                    |       |
| 电源线长m  | 不小于80cm(含插头) |      |              |      |      |                    |       |
| 产地     | 中国 浙江 绍兴     |      |              |      |      |                    |       |
| 特性     | 健康生活 新鲜水开始   |      |              |      |      |                    |       |

图 2、某品牌电水壶的规格参数信息

京东商城的“商品介绍”部分主要以图片列表的形式提供有关商品的更多信息，包括商品**独特的设计特点、商品的优势、适用场景**等。商品信息图片提供了很多“规格参数”部分未包含的商品信息，是顾客了解和选择商品的重要参考，如图 3 所示。

图片是一种更为生动和形象的视觉表达方式，但也为电子商务网站的管理者带来了技术上的挑战。电商平台对商家发布的商品信息有严格的管理规范，如《京东开放平台禁发商品及信息管理规范》：

<http://help.jd.com/Vender/viewQuestion-852-2011.html> 和《京东开放平台商品标题及广告语发布规范》

<http://help.jd.com/Vender/viewQuestion-852-2687.html>。在这里，我们主要考虑商品信息图片中的文本信息的管理问题：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。

最常见的问题是，商品信息图片中的内容和商品规格参数中的内容不一致。例如某电热水壶产品规格参数中标明材质为“拉丝不锈钢”，而商品信息图片中介绍为 304 不锈钢。这将造成用户理解上的困扰：到底哪一个是准确的呢？实际上“拉丝”是不锈钢的一种表面处理工艺，“拉丝不锈钢”并没有说明所采用不锈钢材料的成分所符合的国家标准。

商品信息图片可能出现的另一个问题是，采用图片的方式来规避对敏感或违禁词的检测。我国的《广告法》和《反不正当竞争法》分别对虚假广告和利用广告作引人误解的虚假宣传作了禁止性规定。文本形式的商品介绍信息较易进行检测，而使用图片将大大增加电商网站监管的难度和工作量。除了电商网站，微信等社交软件中也存在大量以图像格式存在的文本信息，所以这一问题具有普遍性。

要解决以上的这些问题，需要设计一个系统来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。

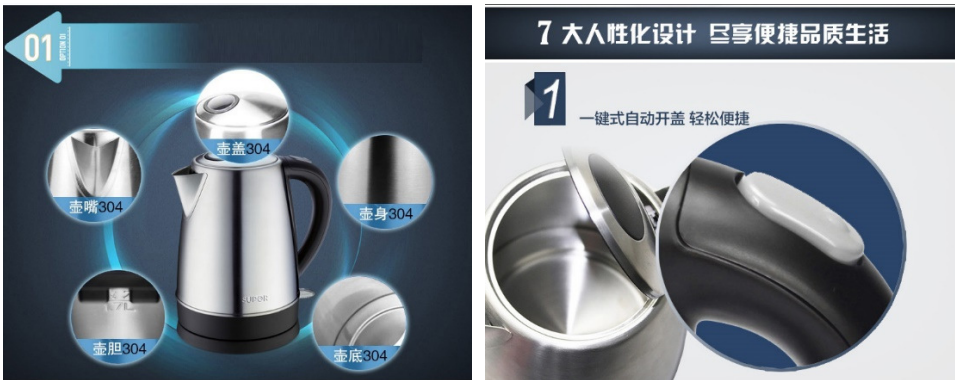
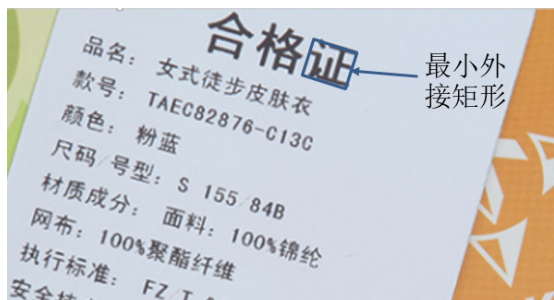


图 3、某电水壶的商品信息图片（部分）

**二、根据附件提供的数据，设计相应的算法，解决以下任务，并做出评测（F-score）。**

**1、图片中的字符检测**

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（字符的最小外接矩形，如图 4 所示）。以图片左下角为原点，输出字符的最小外接矩形的坐标 4 元组：（左下角 x 坐标，左下角 y 坐标，右上角 x 坐标，右上角 y 坐标），如图 5(b)中的后四列所示。



最小外接矩形 → 洗涤保养 MAINTENANCE

#### 鞋类保养注意事项：

1. 保持干燥、存放于通风阴凉处。
2. 切忌暴晒、火烤、水洗、锐物撞击和接触化学物质。
3. 磨砂皮、绒面皮和布类严禁擦油或用水擦洗、清洁鞋面可用软刷顺方

图 4 字符的最小外接矩形

## 2、字符识别

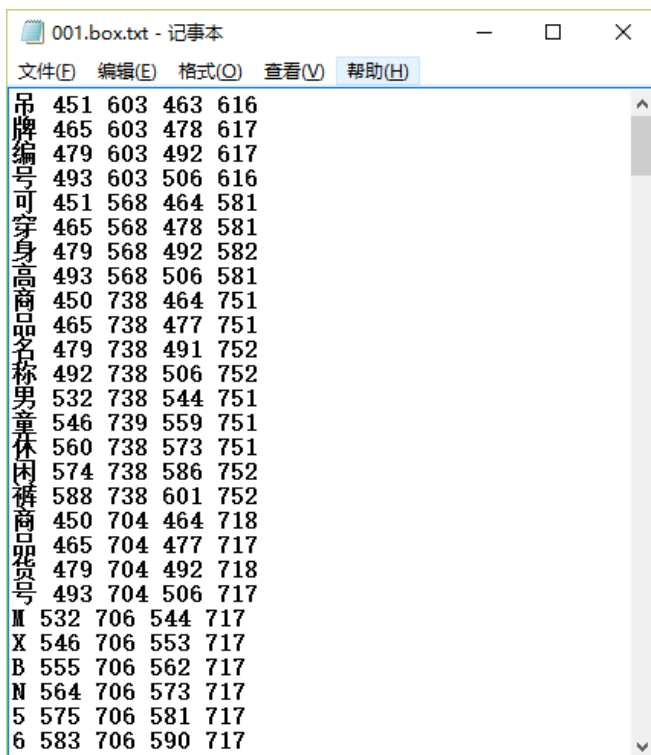
对所提取字符进行自动识别，识别结果列在字符位置四元组之前，如图 5(b)所示。

## 三、数据说明

1、附件 1 为训练样本，包含了 1000 张商品图片及对应的 box 文件。图 1(a)是某品牌服装商品的图片。图 1(b)是图 1 所对应的 box 文件。



(a)



(b)

图 5. 附件 1 数据 (a)商品信息图片(b)对应的 box 文件

2、附件 2 为检测样本，包含了 500 张商品图片，将于 5 月 14 日 8:00 在赛题发布页面(<http://www.tipdm.org/qk/729.jhtml>)中公布。参赛队需在最后提交作品期限之内，以 box 文件方式完成识别结果。