**场景文本标注手册**

本文件属CASIA·PAL组所有，

未经书面许可，不得以任何形式传播。

****

**文件建立/修改记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 版本 | 建立/修改者 | 日期 | 修改内容 |
| 1 | 1.0 | 何文浩 | 2016.12.20 | 建立文档 |
| 2 | 1.1 | 冯媛 | 2016.12.20 | 修改文档内容 |
| 3 | 1.2 | 冯媛 | 2016.12.25 | 修改文档内容 |
| 4 | 1.3 | 冯媛 | 2016.12.29 | 修改文档内容 |
| 5 | 1.4 | 何文浩 | 2017.01.06 | 修改文档内容 |
| 5 | 1.5 | 何文浩 | 2017.04.06 | 增加一些样例 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. **标注目的**

将图像中的所有文本区域用四边形框的形式标注出来，并且标注文本区域的文本内容。

1. **标注工具**

AnnotationTool（可标任意四边形，按shift键调整顶点）.exe

使用方法：直接画框可添加，双击框可修改，通过框角点可拖拉、旋转，按ctrl可移动框，按shift可调整角点。

1. **标注范围**

简体中文、英文、阿拉伯数字、常见符号（包括逗号、句号、空格等）。

**注：英文文本标注时区分大小写。**

1. **标注要求**
2. **每个文本框用任意凸四边形表示，四边形尽量要刻画出文本行的透视变化信息，并且贴合文本的边界。**

具体的工具使用方法不做介绍，如图 4\_1所示：



图4\_1

1. **每个文本框都需要记录相应的文本内容。**
2. **一个文本框只能包含一行文本。**

如图4\_2红色框是错误标记，绿色框是正确标记：

图4\_2 图4\_3

1. **所有能认定出来的文本区域都需要用四边形标注出来。**

如上图4\_2中的比较模糊的文字，如图4\_3中未被遮挡的文字，都需要标注出来。

1. **有明显语义关联的字，不管相隔多远，都要作为一个文本行整体进行标记。**

如图4\_4所示：



图4\_4

1. **如果有明显的语义断开，就断开标记；**

**如果断开处有标点连接，就连成一行标记；**

**如果中文短句之间有空格，分两块标记。**

如图4\_5(a)中，红色方框为有明显语义断开，而且无标点连接，绿色框为有标点连接。

如图4\_5(b)中，紫色方框中有文字被遮挡，为有明显语义断开，则断开标记为两个文本行。

如图4\_5(c)中，背景图案将字明显分割，中间纹理太过复杂，则断开标记；

如图4\_5(d)中，中文短句之间有空格，分两块标记；

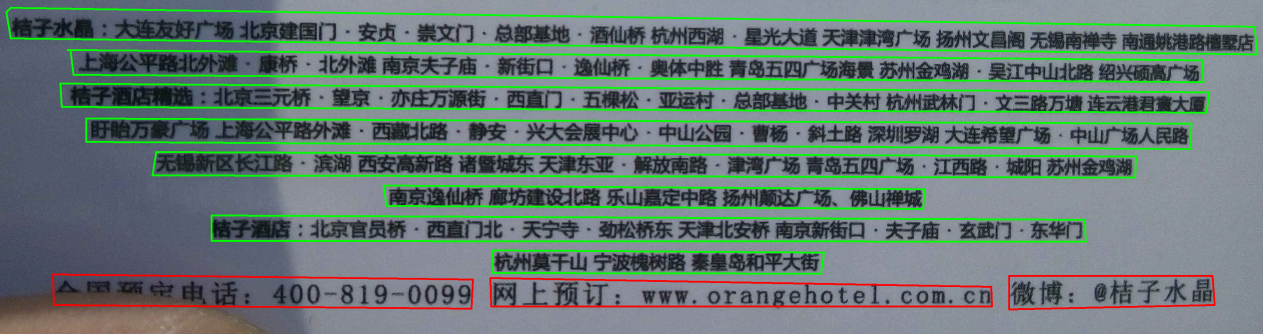


图4\_5(a)

C:\Users\sophie\Documents\Tencent Files\447517238\Image\Group\PJ3${(MJKCW57TL~WH[5WT7.jpg

图4\_5(b)



图4\_5(c)

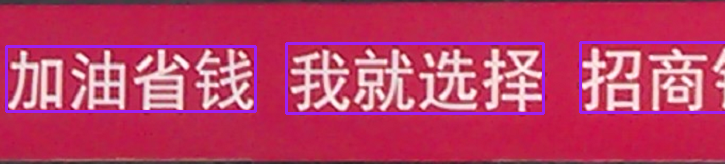


图4\_5(d)

1. **如果字体大小变化较大，且字段之间有空格，分成两个文本行进行标记；如果没有空格，如果有语义连接，那么就连为一行，不然标为两行**

如图4\_6(a) 所示，字段之间有空格，分为两行标记

如图4\_6(b) 所示，Vista为一个单词，连在一起标记

如图4\_6(c) 所示，字段之间没有语义关联性，分开标记



图4\_6 (a)



如图4\_6 (b)



如图4\_6 (c)

1. **遇到常见符号，需要在记录文本内容时，一并记录，包括包括逗号、句号、空格等。**
2. **文本行出现重叠时，重叠的文本行均需标记。**

如图4\_7所示。



图4\_7

1. **文本行出现竖向书写时，按一列进行需标记。**

如图4\_8所示。

图4\_8

1. **文本行出现中英混合情况时，假如中英文之间有空格，分开标记。**

如图4\_9所示。

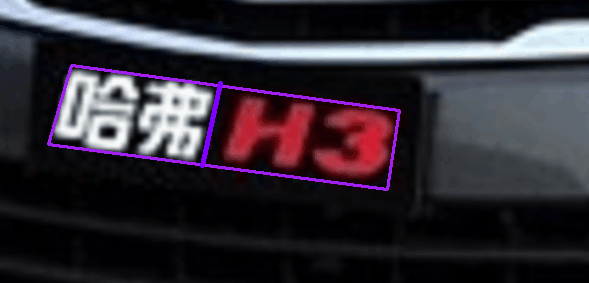
****

图4\_9

1. **标注纯英文还有数字时，将英文单词视为中文的汉字，所以英文还有纯数字按行标注。如图4\_10, 4\_11所示**

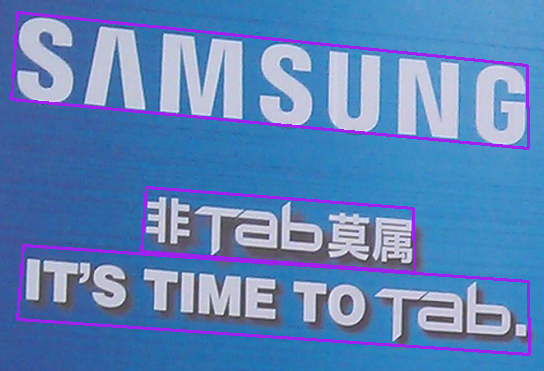


图4\_10

****

图4\_11

1. **只将有语义关联性的字与词框为一起，不然单独标注，如图4\_12**



图4\_12

1. **Ignore情况（只标注文本框，不填写文本内容）**
2. **文本模糊，标注者的肉眼难以辨认，只标注文本框，不填写文本内容。**

如图5\_1所示:



图5\_1

1. **字符被全部/大部分遮挡，标注者的肉眼难以辨认，只标注文本框，不填写文本内容。**

如图5\_2所示:

****

图5\_2

1. **字符上下颠倒，只标注文本框，不填写文本内容。**

如图5\_3所示:

****

图5\_3

1. **如果文字区域的字太小，可以标记一片区域为ignore**

如图5\_4所示:

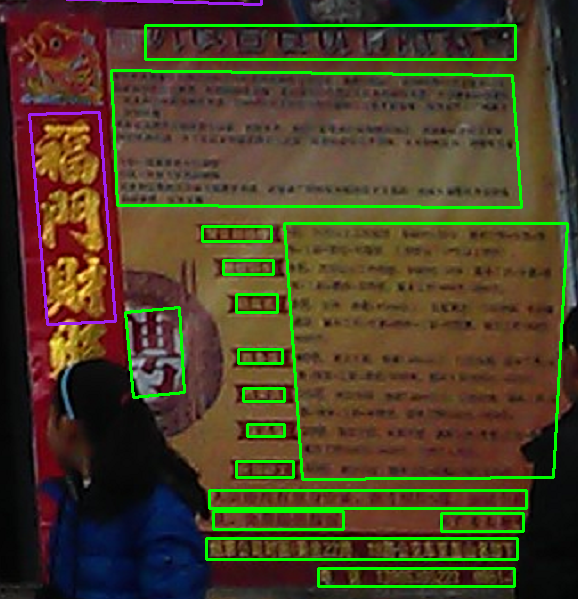


图5\_4

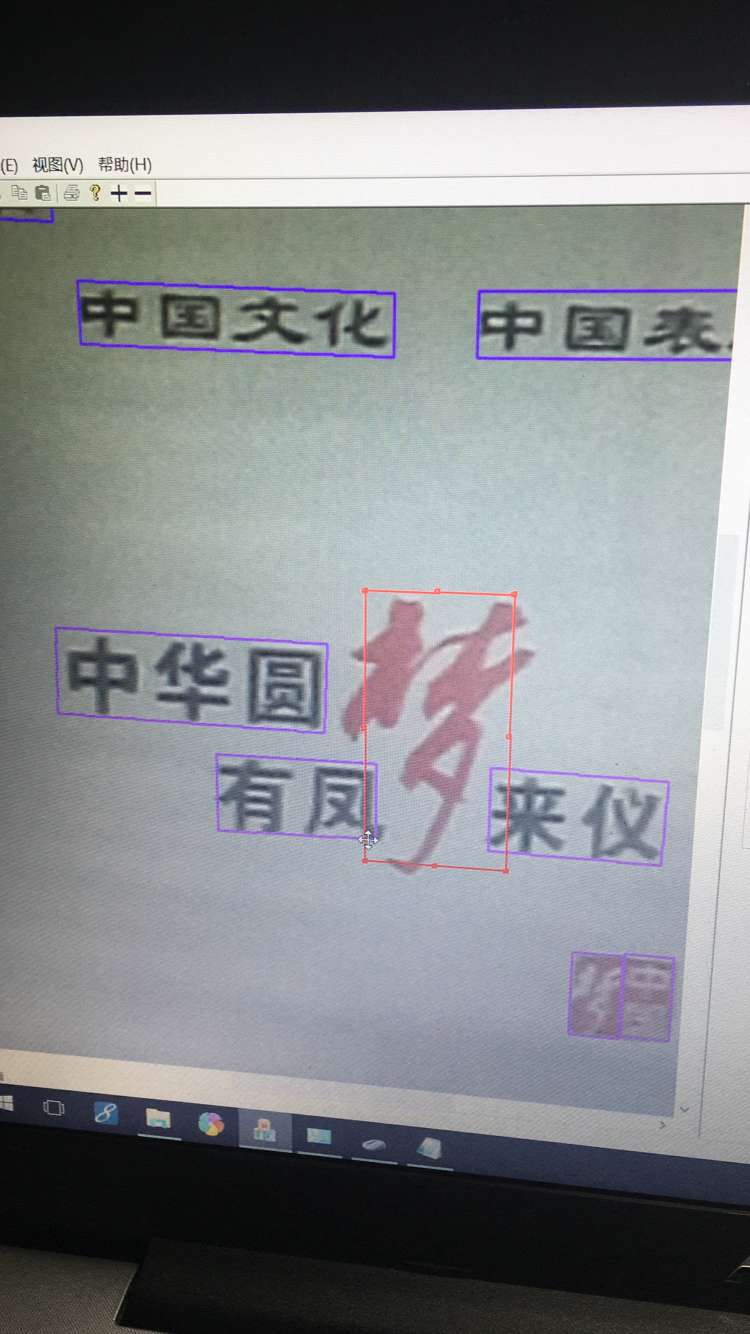
1. **不标注情况**
2. **不属于标注范围提到的文本或者字符。**
3. **标注中遇到的疑问汇总**

该章节罗列标注过程中遇到的一些问题，前六章节相当于讲课，该章节相当于习题

错误的标注：MILK BAR应该标为一行



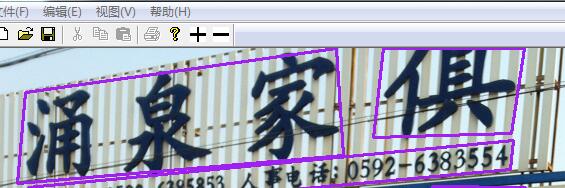
正确的标注：因为“梦”与“中华圆”大小相差太多。当然如果“中华圆梦”标为一个整体也是可以的。这种属于皆可的标注情况。



错误的标注：“招商银行”应该标为一个整体，而且整体应该标记为“Ignore”，下图“NA MERCH”应该标为整体，“S BANK”也应该标为整体



错误的标注：原因同上，“俱”应该与“涌泉家”标为一行



复杂公式整体标注为Ignore

