查看: 778 | 回复: 10

TA的每日心情

开心

2019-4-4 17:54

签到天数: 2 天

[LV.1]初来乍到

10	36	1125
主题	帖子	VC币

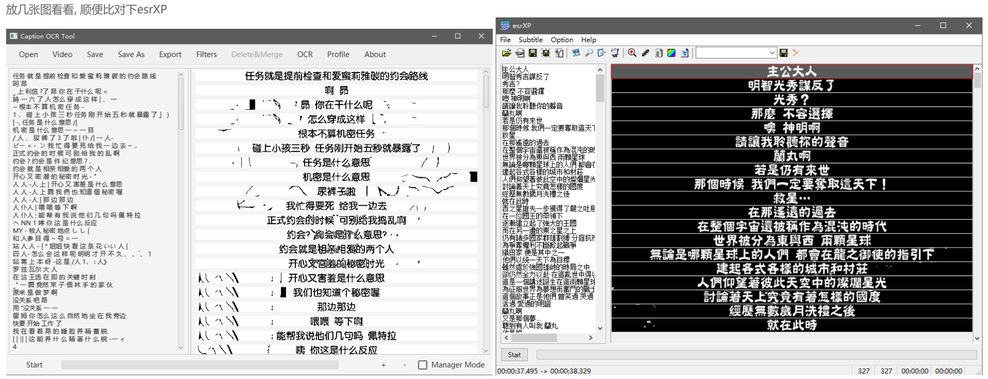
至尊会员

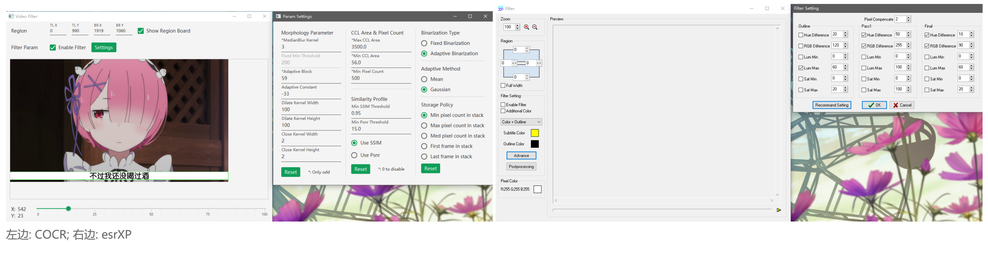
积分 290312

为了更便捷的手抄字幕, 我搞了一个新玩具

NoobNeo 发表于 2019-8-11 22:37:01 | 只看该作者

提取硬字幕, 就是对着视频把字幕一个个敲下来做成外挂字幕.
普遍用esrXP做提取工作, 这软件从05年就不更新, 兼容性不好, 其他的提取软件多数是需要idx/sub字幕, 用OCR处理成ass/srt, 与需求不符.
经过几次手抄后, 觉得esrXP不好用, 万一系统重装(bug10的日常), 又要重新整环境, 很多人就放弃了这个软件, 然后我就利用闲暇时间整了个玩具, 姑且叫"Caption OCR Tool"(Abr. COCR).
放几张图看看, 顺便比对下esrXP





左边: COCR; 右边: esrXP

附上软件试试? 现在连β版本, 呸 α版都算不上, 就不拿出来了.
稍微介绍下, 项目是java语言, 基于opencv, ffmpeg, tesseract等开源项目.
esrXP基本实现方式:
1. 通过HSL实现字幕过滤的, 整体干扰物(无用的竖线, 噪点)少, 需要考虑字体/描边颜色;
2. 去除干扰物的方法不明, 可以手动去除;
3. 时间轴应该是基于帧的, 提取间隔估计在3~5帧;
4. 重复帧的处理方式不明;
5. OCR? 提取字幕图片后导出, 再利用其它软件识别.
COCR基本实现方式:
1. 通过形态学算法实现字幕过滤的, 整体干扰物多, 文字区域少, 忽略颜色信息;
2. 利用连通区域填充去除干扰物, 不能手动去除;
3. 时间轴是基于帧的, 提取间隔可选1~5帧;
4. 重复帧是通过SSIM或PSNR算法去除的;
5. 通过Tesseract实现的, LSTM网络(RNNs的一种), 把论坛的字体包过了一遍, 简体字错误率3%, 繁体字10%, 日文10%, 我已经尽力了...心累
其实用Tensorflow + CNN组合更好, 不过调用失败, 神经网络这方面的知识不够, 只好放弃.
esrXP运行的时候占用内存极低, 才几十M; COCR根据视频大小, 经过优化还是需要1G+的内存, 最初动不动就OOM(内存溢出), 跟其他人交流后, 基本都遇到过视频过大就OOM的问题, 感叹esrXP用的什么黑科技如此省资源.

如果有好的建议, 请不要吝惜, 发给我吧, 软件我会在优化测试结束后, 发到论坛上的, 但只有我一个人, 平时也没太多时间, 咕咕咕...