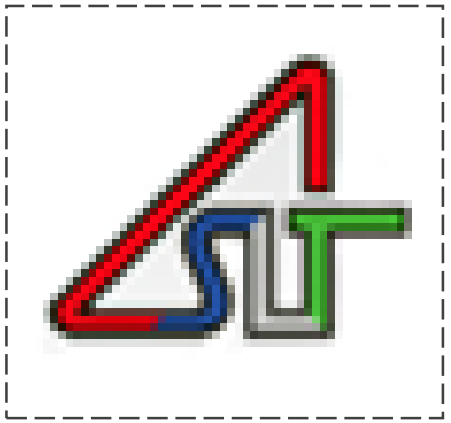


CSLT 项目阶段性报告

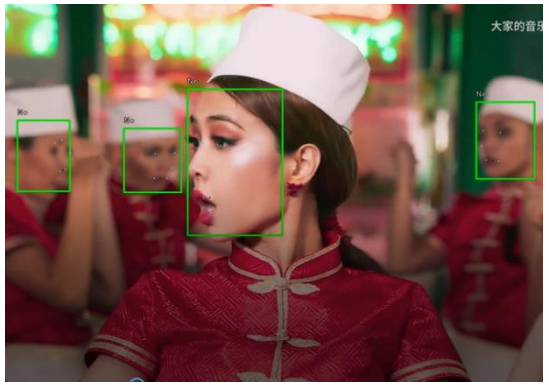
Reporter: Sitong Cheng, Pengyuan Zhang
2019.8.19

目 录

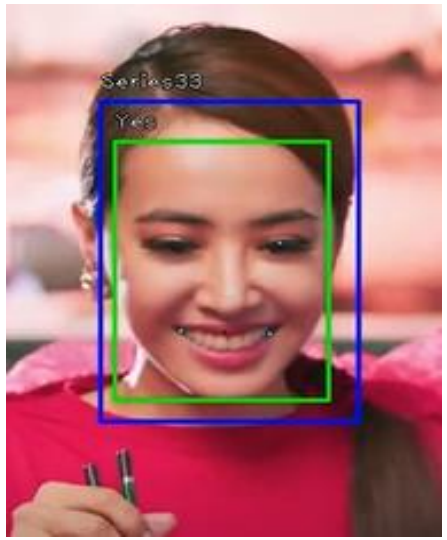
01 原始模型简介

A Brief Introduction to Original Model

原始模型简介



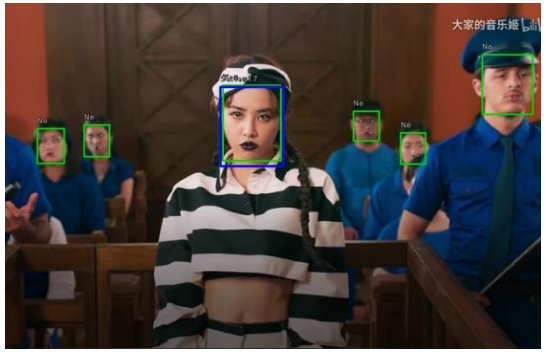
RetinaFace：Face Detection
https://arxiv.org/pdf/1905.00641.pdf



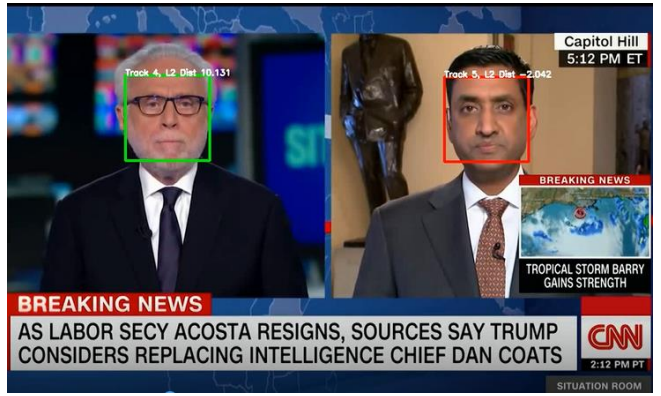
Face Tracker

- opencv的追踪算法
- 判断视频两帧之间的相关性
- 检测范围为一帧的整张图像
- 用于存储目标POI出现的视频序列，作为syncnet的输入

- 根据Tracker存储的唇动视频序列输入，输出唇动和语音之间的匹配程度confidence
- 原模型的应用场景为同时出现多个人，confidence较大的判断为说话者
- 在本项目中的作用：判断单一POI的唇动和语音是否匹配



ArcFace：Face Recognition
https://arxiv.org/pdf/1801.07698.pdf



SyncNet：Speaker Validation
http://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/chung16a.pdf

原始模型简介

种类	视频数量	反正例率FPR
interview	20	18.4%
entertain	7	39%
tvs	2	41%
vlog	2	39%
song	2	41%
All	**33**	**26.8%**

其他问题

- 旁白出声时，可能视频在播放主人公的影视作品，主人公嘴的动作会被识别为匹配成功，导致截取了旁白的片段;
- 得到片段过短（大部分短于一秒），结果太碎;
- 各种类别FPR效果相差太大。

02

模 型 优 化 改 善

Optimization and Improvement of Our Model

模 型 优 化 改 善

A SyncNet设定多组阈值

- 原始模型阈值：start confidence = 4, end confidence = 3. 不同类型视频结果相差较大
- 经过多次实验，确定三种最佳阈值，分别适用于场景简单、一般和复杂的视频

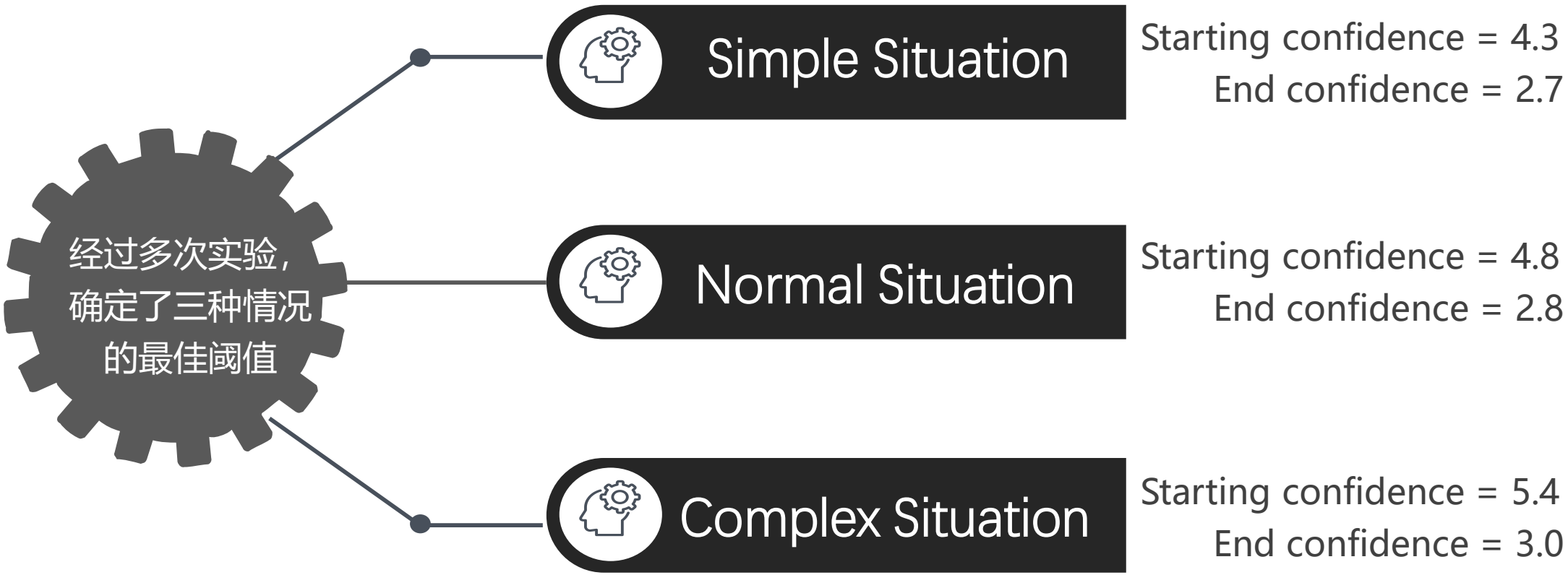
B 数据清洗

- 合并相距时间较短的片段
- 合并因tracker逻辑问题而出现的重叠片段

C 加入声纹模型

- 声纹模型能较好地分隔不同人说话的片段，但无法识别POI
- 将声纹结果与SyncNet结果做后处理

A.SyncNet设定多组阈值



A.SyncNet设定多组阈值

种类	视频数量	反正例率FPR	FPR变化	召回率Recall
interview	17	17.25%	-1.15%	59.80%
entertain	7	32.69%	-6.31%	28.59%
tvs	2	35.98%	-5.02%	14.41%
vlog	2	34.58%	-4.42%	22.22%
song	3	27.12%	-13.88%	21.86%
All	**31**	**22.44%**	**-4.36%**	**43.73%**

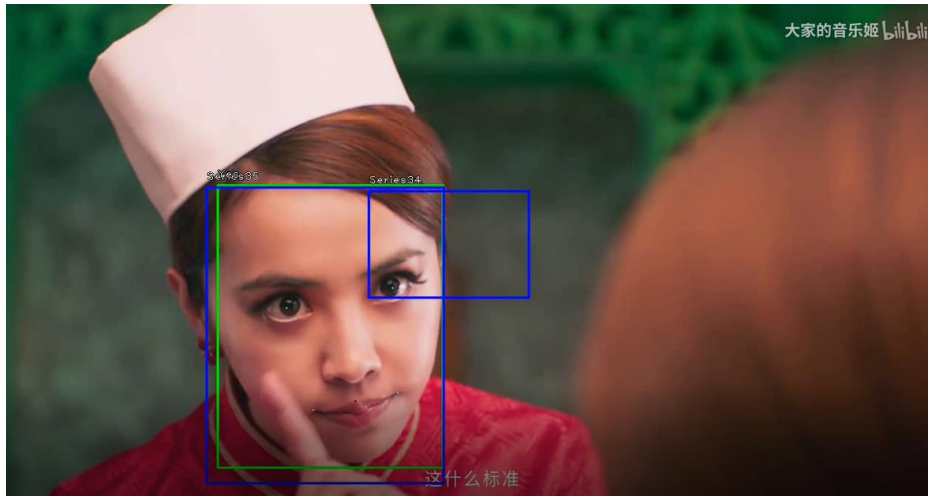
结果:

- FPR有所改善，但不同类别效果依然存在较大差异
- 旁白问题没有解决
- 片段长度得到改善，但仍未达到预期（依然有很多短于一秒的片段）
- 片段之间有重叠的部分

B.数据清洗

计算两个片段间距，小于某个值就会将它们拼接在一起

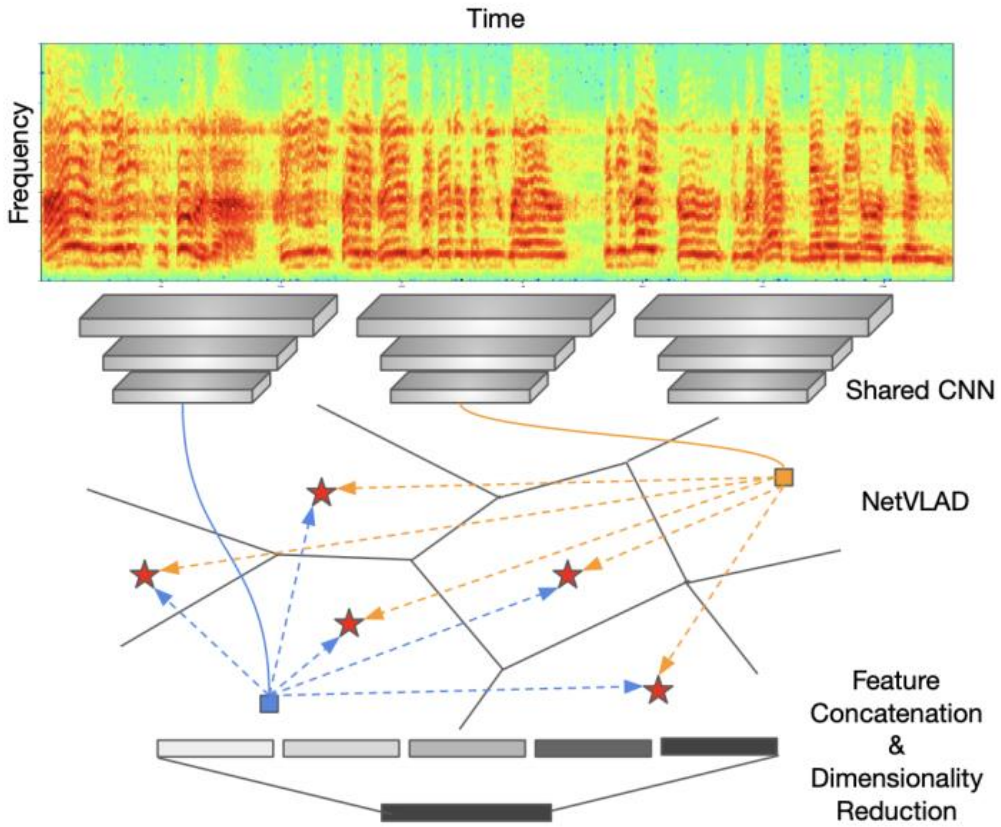
- 间距小于10帧，其他人不足以说话
- 消除片段的重叠，片段重叠的原因是同时出现两个tracker，图中两个紫框代表两个tracker，是由主程序逻辑产生的



结果：FPR没有明显升高，Recall小幅度提高。得到片段长度有明显改善

C.声纹模型——模型介绍

VGG-Speaker-Recognition(2019.5.17)
https://arxiv.org/pdf/1902.10107.pdf

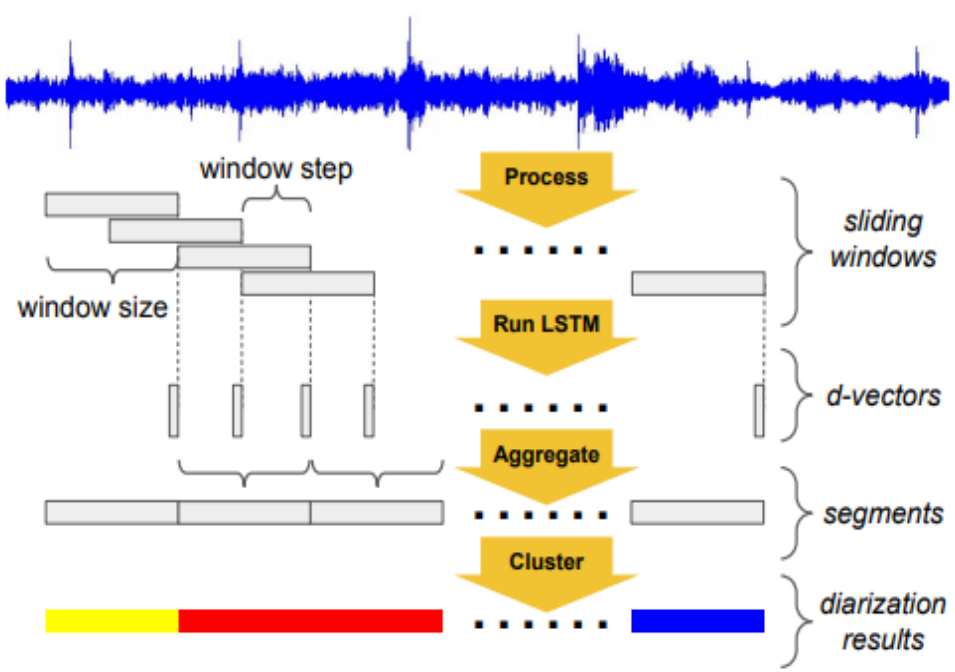


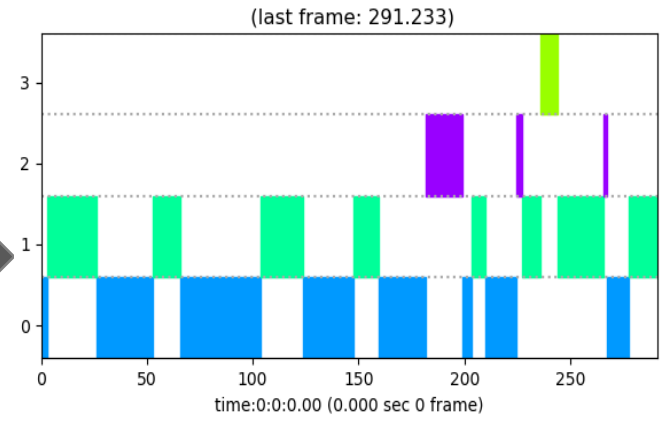
Module	Input Spectrogram ($257 \times T \times 1$)	Output Size
Thin ResNet	conv2d, 7×7, 64	$257 \times T \times 64$
	max pool, 2×2, stride $(2, 2)$	$128 \times T/2 \times 64$
	$\begin{bmatrix} \text{conv}, 1 \times 1, 48 \\ \text{conv}, 3 \times 3, 48 \\ \text{conv}, 1 \times 1, 96 \end{bmatrix} \times 2$	$128 \times T/2 \times 96$
	$\begin{bmatrix} \text{conv}, 1 \times 1, 96 \\ \text{conv}, 3 \times 3, 96 \\ \text{conv}, 1 \times 1, 128 \end{bmatrix} \times 3$	$64 \times T/4 \times 128$
	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 256 \end{bmatrix} \times 3$	$32 \times T/8 \times 256$
	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 512 \end{bmatrix} \times 3$	$16 \times T/16 \times 512$
	max pool, 3×1, stride $(2, 2)$	$7 \times T/32 \times 512$
	conv2d, 7×1, 512	$1 \times T/32 \times 512$

$$L_i = -log\frac{e^{s(\cos\theta_{y_i}-m)}}{e^{s(\cos\theta_{y_i}-m)} + \sum_{j\neq y_i} e^{s\cos(\theta_j)}}$$

C.声纹模型——模型介绍

Google UIS-RNN(2018.10.10)
https://arxiv.org/pdf/1810.04719v1.pdf






```
00:02:44:21 00:02:46:19
00:02:47:22 00:02:50:5
00:02:54:9  00:02:56:4
00:02:56:22 00:02:59:0
00:02:59:22 00:03:02:20
00:03:08:13 00:03:12:15
00:03:20:12 00:03:21:8
00:03:21:23 00:03:23:17
00:03:34:14 00:03:38:10
00:03:43:10 00:03:45:18
00:03:58:4  00:04:06:20
00:04:12:11 00:04:15:7
00:04:32:7  00:04:36:5
00:04:52:19 00:05:06:21
00:05:46:22 00:05:48:19
00:06:02:24 00:06:04:21
00:06:07:7  00:06:11:7
00:06:38:5  00:06:39:23
00:07:13:6  00:07:15:23
00:07:16:13 00:07:19:6
00:07:19:21 00:07:24:1
00:07:25:11 00:07:26:22
00:07:39:22 00:07:41:15
00:07:47:3  00:07:47:19
00:07:51:21 00:07:52:22
============ 1 ============
00:00:41:4  00:00:55:12
00:00:57:14 00:00:58:18
00:01:55:20 00:02:00:6
00:02:06:22 00:02:08:20
00:02:46:19 00:02:47:22
00:03:33:2  00:03:34:14
00:06:15:23 00:06:17:6
00:06:49:6  00:06:50:11
00:06:52:13 00:06:55:2
00:07:03:4  00:07:03:19
00:07:10:21 00:07:11:22
00:07:36:22 00:07:37:18
00:09:08:14 00:09:27:22
00:09:32:15 00:09:54:23
00:10:19:23 00:10:44:22
00:10:48:14 00:10:50:4
00:11:22:6  00:11:47:2
00:11:49:1  00:12:04:21
00:12:50:15 00:13:19:19
00:13:48:8  00:13:51:3
============ 2 ============
00:00:55:12 00:00:57:14
00:00:58:18 00:01:13:16
00:01:16:9  00:01:24:24
```

03

最 终 程 序 逻 辑

The final program logistic

程序逻辑介绍

1. 合并、清洗SyncNet得到的结果片段
2. 将声纹模型得出的**每个人**的结果与SyncNet结果取交集，根据交集占该人物总时长的比例确定出POI
3. 对确定的POI的声纹片段与SyncNet结果取交集，根据交集占该片段时长的比例确定片段是否正确
4. 根据第三步得到的片段来判断视频质量，淘汰掉效果差的视频
5. 修剪、合并第三步得到的片段，得到最终的结果

```
start to clean data..
id: 0  length: 2383 交: 104 proportion: 0.043642467477968946
id: 1  length: 1565 交: 36 proportion: 0.023003194888178913
id: 2  length: 1211 交: 60 proportion: 0.0495458298926507
id: 3  length: 138 交: 50 proportion: 0.36231884057971014
id: 4  length: 5146 交: 318 proportion: 0.061795569374271275
id: 5  length: 8078 交: 359 proportion: 0.04444169348848725
id: 6  length: 1746 交: 299 proportion: 0.17124856815578465
id: 7  length: 3137 交: 1148 proportion: 0.36595473382212307
id: 8  length: 4442 交: 2230 proportion: 0.5020261143628996
id: 9  length: 1974 交: 117 proportion: 0.05927051671732523
id: 10  length: 256 交: 5 proportion: 0.01953125
id: 11  length: 3353 交: 1463 proportion: 0.4363256784968685
id: 12  length: 8963 交: 4887 proportion: 0.5452415485886422
id: 13  length: 835 交: 407 proportion: 0.4874251497005988
id: 14  length: 7916 交: 4288 proportion: 0.5416877210712481
id: 15  length: 1729 交: 732 proportion: 0.4233661075766339
id: 16  length: 2951 交: 1679 proportion: 0.568959674686547
id: 17  length: 612 交: 230 proportion: 0.3758169934640523
id: 18  length: 326 交: 36 proportion: 0.11042944785276074
id: 19  length: 666 交: 287 proportion: 0.43093093093093093
id: 20  length: 1117 交: 591 proportion: 0.5290957923008057
id: 21  length: 81 交: 0 proportion: 0.0
id: 22  length: 78 交: 30 proportion: 0.38461538461538464
id: 23  length: 191 交: 101 proportion: 0.5287958115183246
id: 24  length: 110 交: 0 proportion: 0.0
wav poi: 12 :  4887 frame
wav poi: 14 :  4288 frame
```

```
file writing..
total valid frames:  33400
predict correct frames:  8940
total missed frames:  188
FPR:  0.020595968448729185
Recall:  0.26766467065868266
Proportion:  0.15437694493302664
f1 micro:  0.4204288939051919
```

结果验证

A.只使用SyncNet的结果

种类	视频数量	反正例率FPR	召回率Recall
entertain	18	28.00%	29.40%
interview	40	16.59%	48.95%
act	2	12.45%	67.10%
tvs	4	41.13%	15.25%
vlog	4	16.83%	43.95%
recite	1	1.10%	25.00%
advertise	2	0.00%	24.86%
song	3	28.13%	20.67%
movie	2	52.20%	13.50%
All	76	21.24%	39.76%

B.使用声纹（eps1.5）和数据清洗，但不淘汰得到的结果统计（与只使用SyncNet比较）

种类	视频数量	反正例率FPR	FPR变化	召回率Recall	Recall变化
entertain	18->17	20.01%	-7.99%	21.80%	-7.60%
interview	40	18.54%	+1.95%	43.93%	-5.02%
act	2	20.55%	+8.10%	60.70%	-6.40%
tvs	4	20.02%	-21.11%	10.18%	-5.07%
vlog	4	14.03%	-2.80%	47.45%	+3.50%
recite	1	0.00%	-1.10%	11.90%	-13.1%
advertise	2	0.00%	0.00%	22.60%	-2.26%
song	3	22.43%	-5.70%	10.50%	-10.17%
movie	2	35.35%	-16.85%	6.90%	-6.60%
All	76->75	18.63%	-2.61%	34.43%	-5.33%

C.使用声纹（eps1.35）和数据清洗，但不淘汰得到的结果统计（与只使用SyncNet比较）

种类	视频数量	反正例率FPR	FPR变化	召回率Recall	Recall变化
entertain	18	14.40%	-13.60%	20.50%	-8.90%
interview	40	18.37%	+1.78%	47.14%	-1.81%
act	2	21.45%	+9.00%	67.30%	+0.20%
tvs	4	20.02%	-21.11%	10.20%	-5.05%
vlog	4	7.15%	-9.68%	50.50%	+6.55%
recite	1	0.00%	-1.10%	13.10%	-11.90%
advertise	2	3.25%	+3.25%	41.80%	+16.94%
song	3	30.03%	+1.90%	6.60%	-14.07%
movie	2	36.70%	-15.5%	8.60%	-4.90%
All	76	17.31%	-3.93%	36.39%	-3.37%

D.使用声纹（eps1.35）和数据清洗，并淘汰质量不好视频得到的结果统计（与只使用SyncNet比较）

种类	视频数量	反正例率FPR	FPR变化	召回率Recall	Recall变化
entertain	18->14	10.79%	-17.21%	20.12%	-9.28%
interview	40->34	19.84%	+3.25%	50.65%	+1.7%
act	2	21.45%	+9.00%	67.30%	+0.2%
tvs	4->2	0.00%	-41.13%	4.00%	-11.25%
vlog	4	7.15%	-9.68%	50.50%	+6.55%
recite	1	0.00%	-1.10%	11.40%	-1.36%
advertise	2->1	6.50%	+6.50%	76.80%	+51.94%
song	3->2	24.25%	-3.88%	4.50%	-16.17%
movie	2->1	5.20%	-47.00%	16.40%	+2.9%
All	76->61	15.69%	-5.55%	40.36%	+0.6%

04

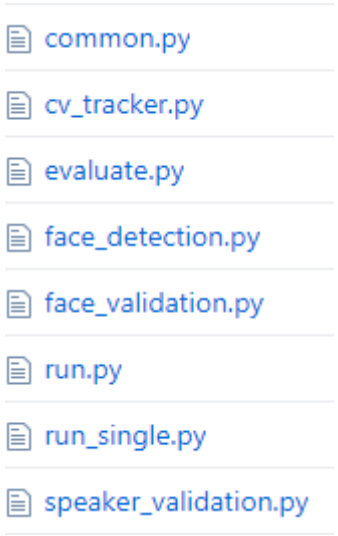
完整项目流程

The complete process of program

项目流程介绍

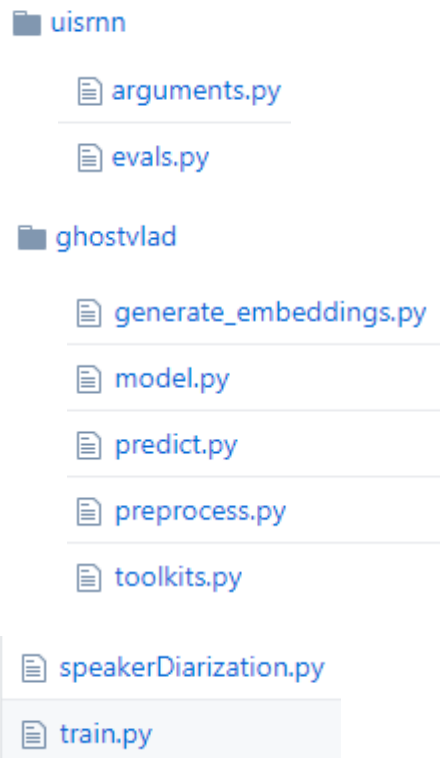
1. 选定要收集的人物，填入excel文档
2. 获取主人公的各类视频
3. （可选）对视频进行转帧、统一命名 [自动化]
4. 获得主人公的面部图片 [自动化]
5. 获得主人公的声纹模型结果 [自动化]
6. 修改好路径，运行主程序 [自动化]
7. 使用结果切割原视频 [自动化]
8. 人工校验

主程序介绍



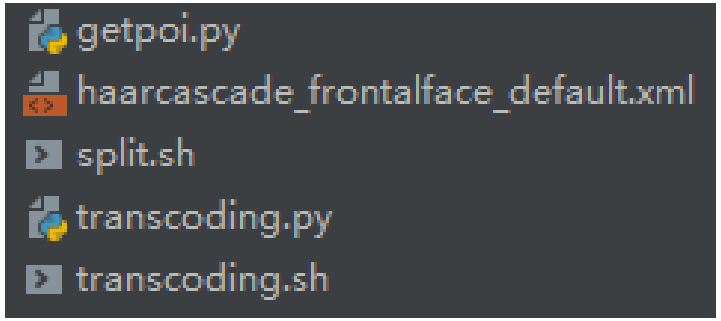
- **common**: 配置类。储存超参数，包括路径、阈值、模型参数等等。
- **cv_tracker**: 追踪类。对目标进行追踪，同时保存追踪过程中的人物嘴部图片，是最主要的程序之一。
- **evaluate**: 结果评估、数据处理。将机器得到的结果与验证集进行比较。还包括了数据清洗、与声纹模型合并以及结果写入。
- **face_detection**: 人脸检测。检测出每帧图像中人脸五点坐标。
- **face_validation**: 人脸识别。确定图像中主人公位置，判断追踪是否正常。
- **run**: 主程序。
- **run_single**: 处理单个视频。一般用于debug。
- **speaker_validation**: 使用SyncNet判断嘴唇动作与音频是否一致，输出初步的结果。

声纹模型程序介绍



- **uisrnn**: uis-rnn网络的主要实现
  - **arguments**: uis-rnn网络参数。影响预测结果和速度的参数主要有beam_size, look_ahead, test_interation
  - **evals**: 用uis网络进行预测
- **ghostvlad**: speaker recognition模型和VLAD层的主要实现
  - **generate_embeddings**: 产生embeddings作为输入
  - **model**: speaker_recognition的模型
  - **predict**: 用speaker_recognition模型进行预测
  - **preprocess**: 数据预处理
  - **toolkits**: 初始化GPU、读入数据等非模型的操作
- **speakerDiarization**: 主程序。按文件路径批量生成speaker diarization的时间标签txt。

工具程序介绍



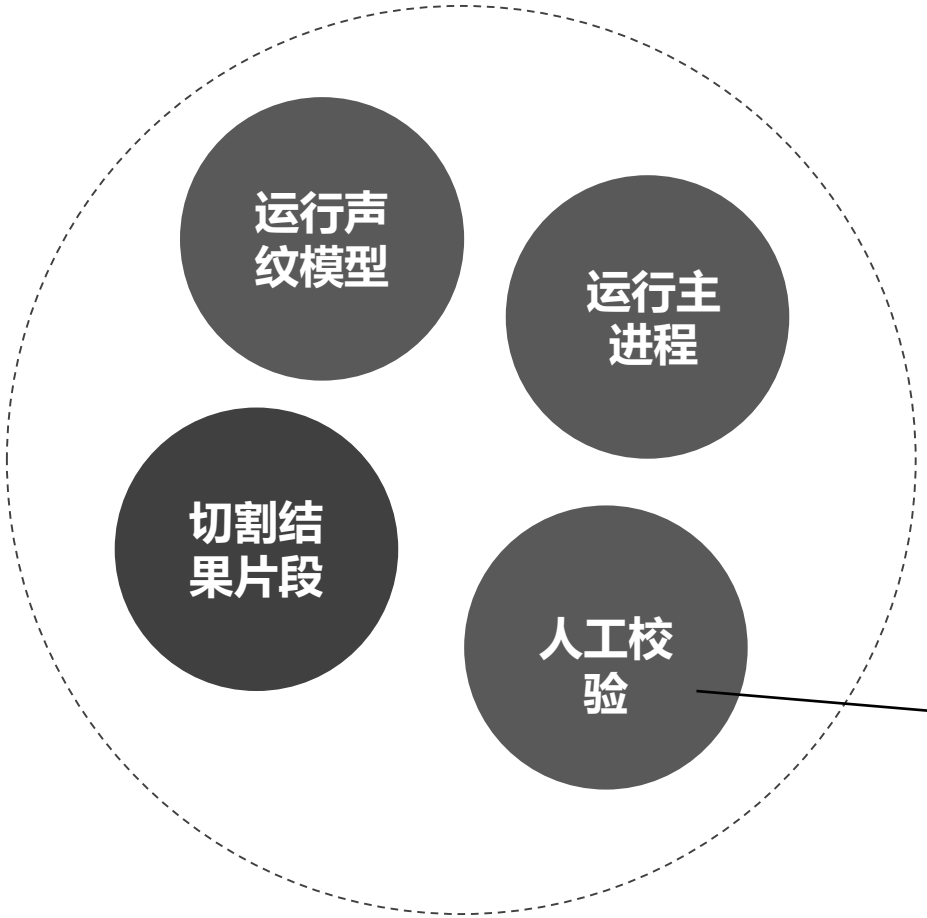
- getpoi: 自动获得主人公人脸图片。包括图片爬虫和人脸识别。
- split: 根据记录了片段开始时间和持续时间的txt文件对原视频进行切割，得到易于人工校验的视频片段。
- transcoding: 对视频进行向25帧率的转换

05

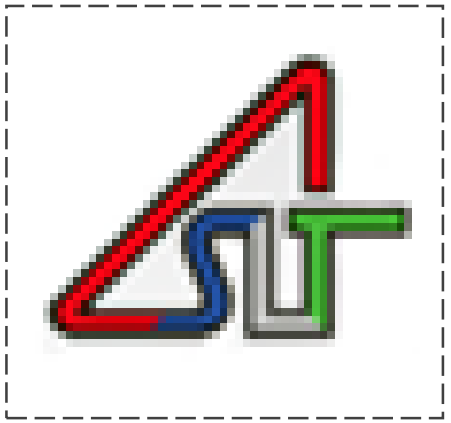
后续工作目标

Future plan

后　续　工　作　目　标



- 舍弃错误片段
- 使用PR重新剪辑



Thanks for listening!

Reporter: Sitong Cheng, Pengyuan Zhang
2019.8.19