ASAC

Basic Face recog is not great for videos

- https://arxiv.org/pdf/2111.14448.pdf
- https://arxiv.org/pdf/2401.12039.pdf
- https://arxiv.org/pdf/2008.04237.pdf
- https://cs.emis.de/LNI/Proceedings/Proceedings245/39.pdf
- Assuming diarized audio is provided initially (basically what this was trying to do in stage 1)
  - Stage 1 technique can technically be used to diarize the audio
  - LWTNet's diarization can also be used
  - Both are not great, not suitable for films
  - Best scenario, cleaned clips for each speaker is provided - recorded separately - for films
- Need to associate each audio clip with speaker/face on each frame of the video, and get x,y,z of the speaker for each frame, so can map audio to that spatial location
- Running face recognition is unreliable since even a small change in appearance will throw the model off (would already have a picture of speaker along with audio clip, maybe also use k means clustering on faces recognised from the video itself)
  - This would only work if there is very less changes in facial features/costume or there are very few people on the scene - even though similarity is low there is enough distinction between faces
  - We need something that performs well with side angles of faces, rotation invariant and also occlusion (partial or full)
  - Some possible models
    - dlib
    - deepface
  - Use AVRNet face tracking and association maybe, handles not clearly seen speakers
- Or run face recognition on whole video, fill in the frames where it could not recognise with object tracking algorithms - less compute than optical flow
  - IOU with hungarian
  - Kalman + IOU with hungarian
  - SORT
  - DeepSORT
  - OCSort
  - **Player Tracking in Sports Videos**
    https://www.researchgate.net/publication/338938041_Player_Tracking_in_Sports_Videos
  - https://viso.ai/deep-learning/object-tracking/
  - https://learnopencv.com/object-tracking-and-reidentification-with-fairmot/
  - https://learnopencv.com/understanding-multiple-object-tracking-using-deepsort/
  - https://broutonlab.com/blog/opencv-object-tracking/
  - https://www.thinkautonomous.ai/blog/computer-vision-for-tracking/

  > https://github.com/google-research/human-scene-transformer

- Or we could run optical flow using various local descriptors, this would solve above problems

- Optical flow - statistical - various features, descriptors - https://nanonets.com/blog/optical-flow/
  - Doesnt work when intensity of image changes, focal length of camera changes, lots of camera movement along with object movement
- Deep learning based optical flow methods - http://sintel.is.tue.mpg.de/results
- Descriptors could be attention maps from LWTNets, basically their optical flow pipeline, it's pretty damn good, fix the memory issue — its not great at all
  - The model uses excessive amounts of CPU memory for long clips, around 30GB of RAM for a 10 sec video clip, very inefficient
  - With multiple speakers moving around, it falters big time, especially suffers from occlusion and identity switching

Future Work

Three avenues

- Taking into account gaze of the person to direct audio in that direction
- Movie audio/foley sounds generation using scene understanding
- Spatial sound track generation
- https://proceedings.neurips.cc/paper/2020/file/227f6afd3b7f89b96c4bb91f95d50f6d-Paper.pdf
- https://xypb.github.io/CondFoleyGen/
- https://v-iashin.github.io/SpecVQGAN
- maybe extend this to object sounds
  - for object sounds we need the model to be creative in sound spatialization, generate multiple possible configurations
  - or maybe act like a copilot to the mixer, recommending things, maybe guided through natural language?

Github link: https://github.com/Raghav010/ASAC

ALL Demo vids: https://iiitaphyd-my.sharepoint.com/:f:/g/personal/raghav_donakanti_students_iiit_ac_in/EvqZFTsVfodIrZPtSZfoluMBRguBnbpNWU5Xe=rnGwa2

Presentation : https://docs.google.com/presentation/d/1Nm6lUIljJoAEm0JnqoT8RTJsSnMTMF1ZWjaIzgpfgG8/edit?usp=sharing