pindrop®

Improving Voice-Bio Using Free Data: Audio-Visual Diarization of Youtube Videos

Mid-Summer Presentation

Desmond Caulley

Georgia Institute of Technology
Research Intern, Pindrop

About Me

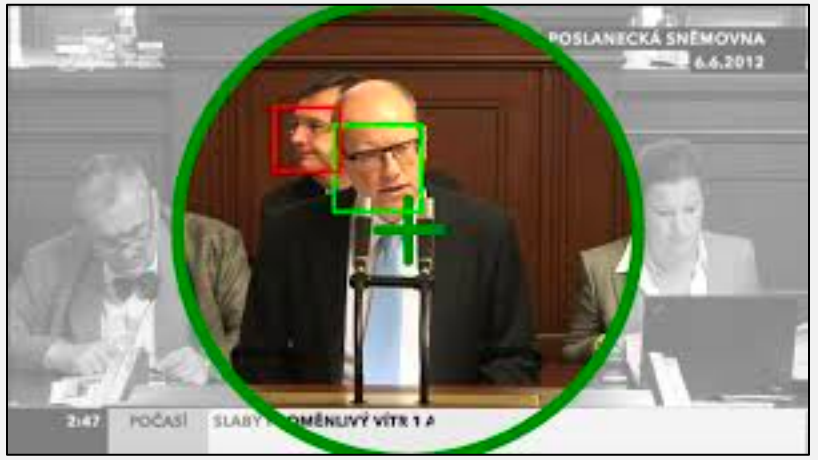
- Born in Ghana, West Africa
- Grew up 20 miles east of Atlanta

- B.S at Cornell University (2016)
- M.S at Georgia Tech (2017)

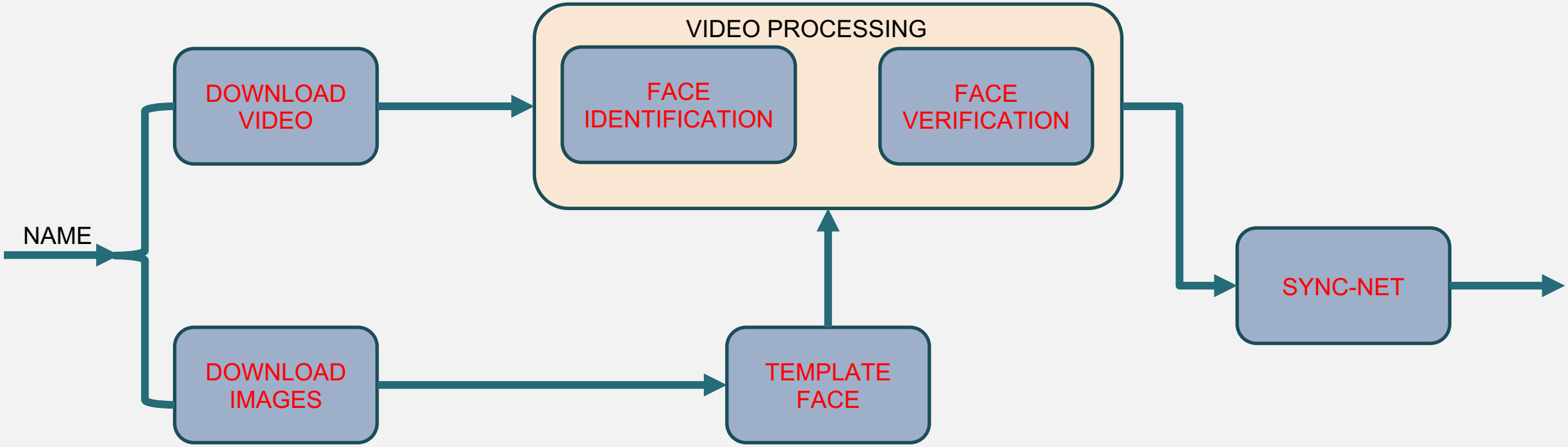
- Ph.D Student at Georgia Tech

Summer Project Overview

- Free additional data to improve voice bio
- Audio-Visual Diarization
- Identify a person of interest in a video
- Verify that the person is talking



SYSTEM OVERVIEW



Step 1: Compiling List

- Celebrity List: Korean Celebrities
- Compiled list of 130 Celebrities
  - Actors, musicians, politicians

Step 2: Template Faces

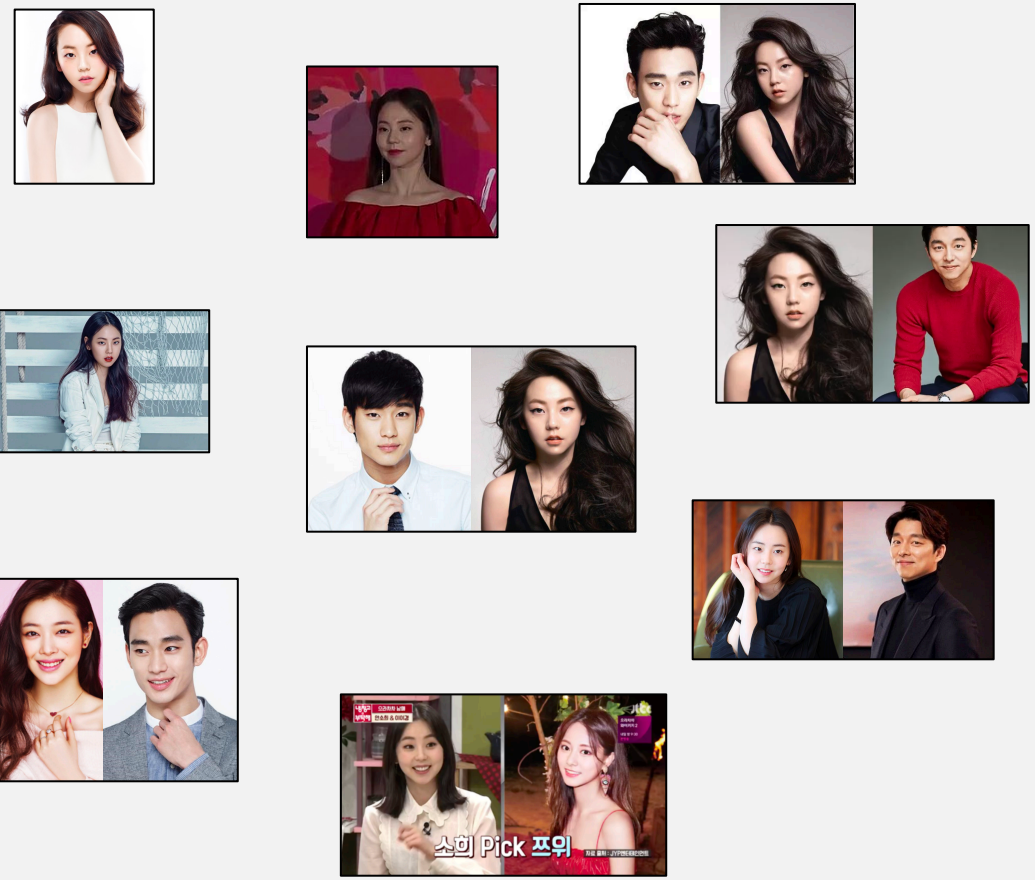
Objective: We need a face with which to compare the video faces to verify person of interest

- Search celebrity face using google:
  - Download top 25 faces returned
  - Ex: 안소희 얼굴

celeb name | face

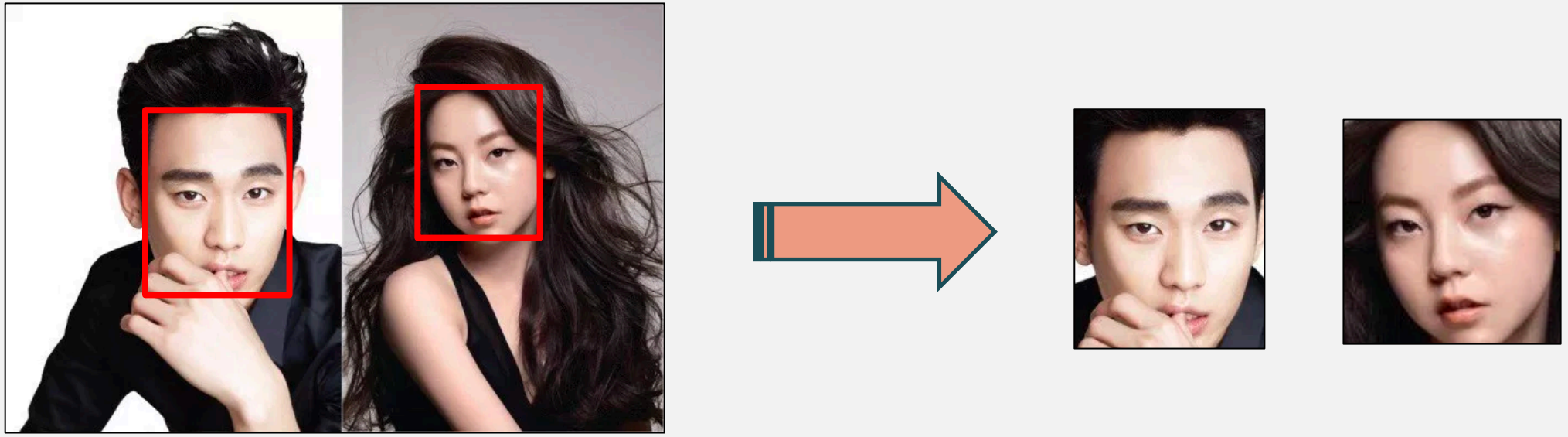
- Search using Korean names
  - More accurate

- Issues:
  - Multiple faces in one picture
  - Non person-of-interest



Creating Template Face

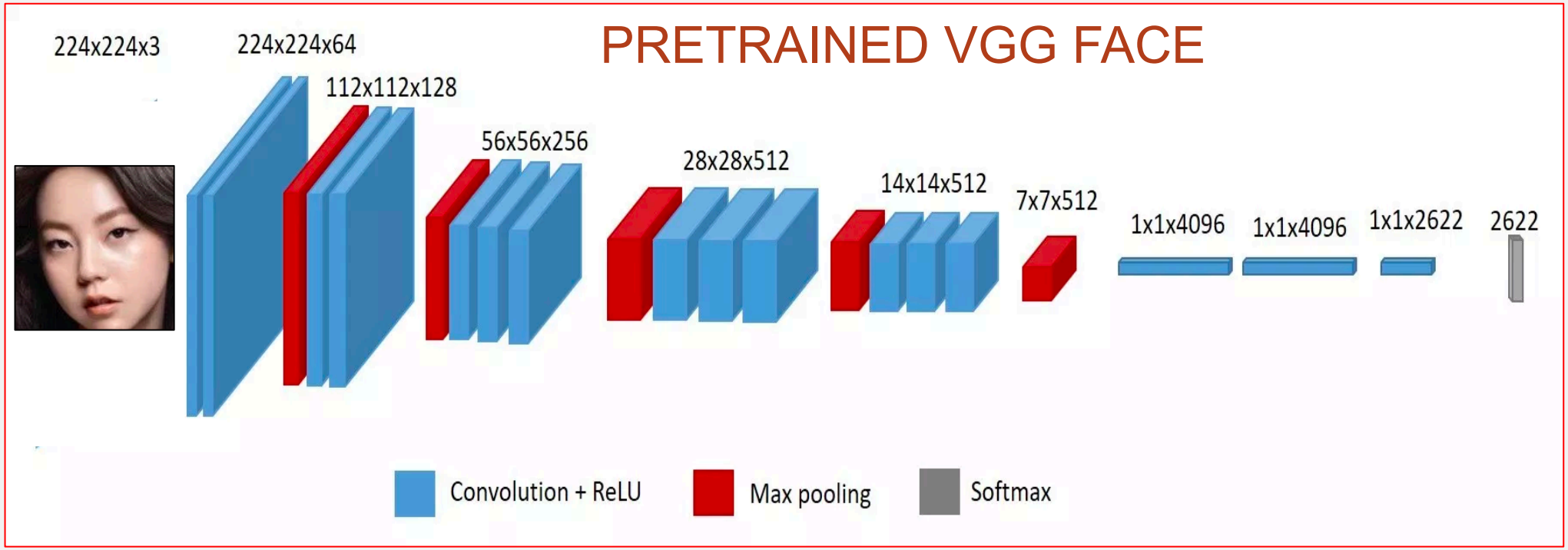
Task 1: Face Detection



We ran this on all faces in all pictures

Creating Template Face

Task 2: Extracting VGG Face Features



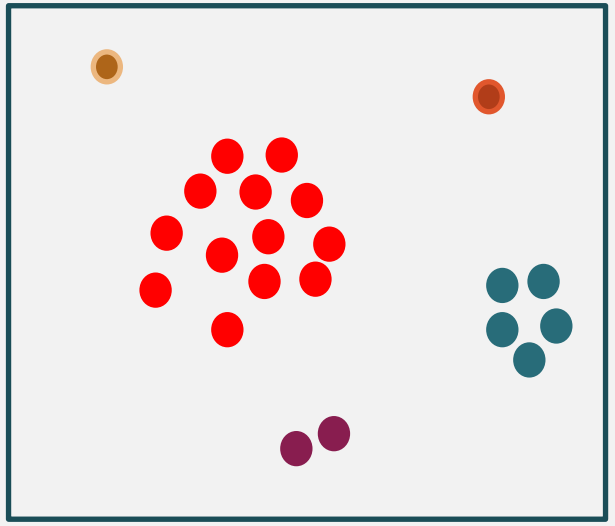
Extract over 2000-d feature vector per image

Creating Template Face

Task 2: DBSCAN Clustering

- Cosine scoring between the 2622-dimensional embedding
  - Threshold is set 0.4
  - Means if distance between two embeddings < 0.4, then they can't be immediately be put together
- DBSCAN finds clusters with high density and builds from those

VGG Embedding Clustering



TAKE MODE TO GET LARGEST CLUSTER (Assuming it's the person of interest)

Store embedding for comparison later

Output: 2 2 1 0 3 2 2 2 -1 1 2 2 2 2 1 0 1 2 2 2 -1 2

Video Search and Processing

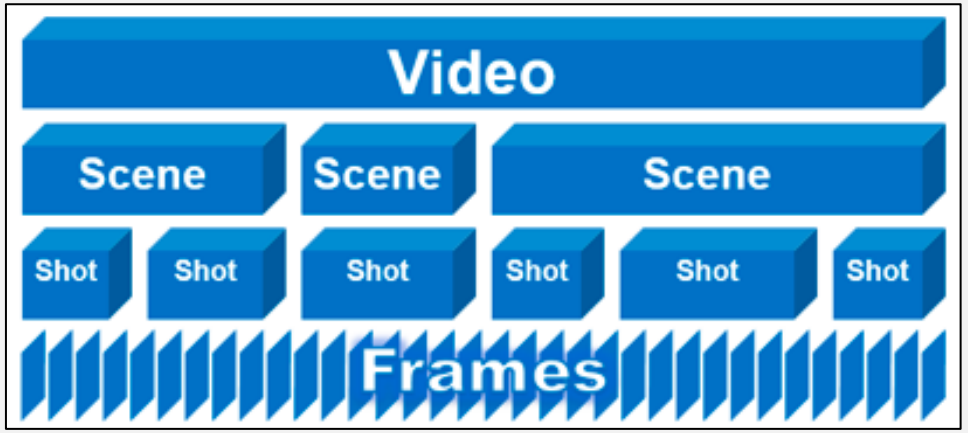
- Search youtube and download top search for celebrity
  - Youtube-dl
  - Download top 30 videos

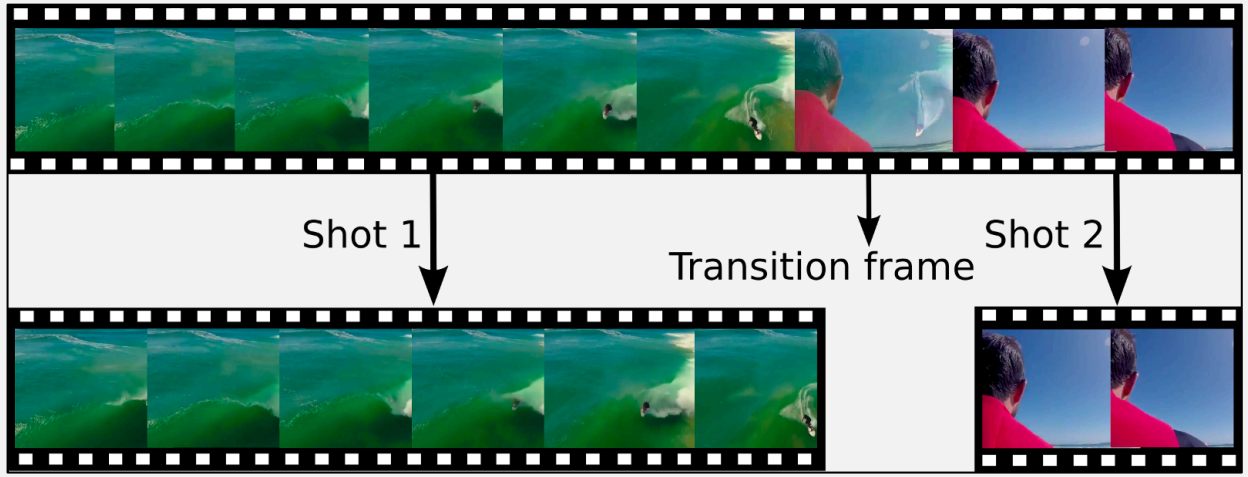
- Search using Korean names
  - Append word interview
  - Ex: 안소희 인터뷰

Step 3: Scene Detection

Objective: Identify scene changes in a video. Might signal to do face detection since people or objects present changes in a scene

- Python pkg – pyscenedetect
- color histogram-based scene detection algorithm in the HSV/HSL color-space
- Store frames where scenes transitioned

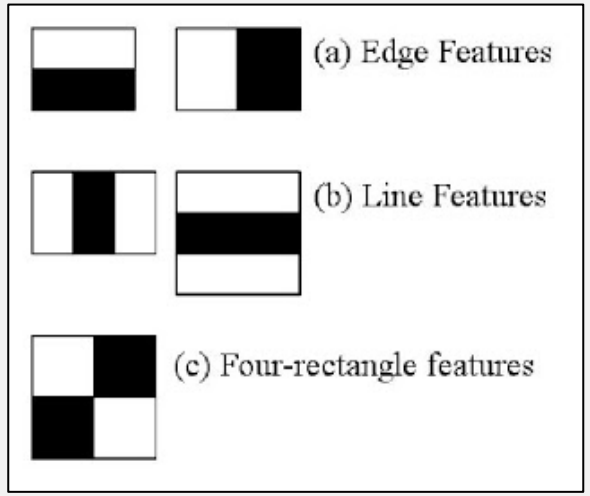


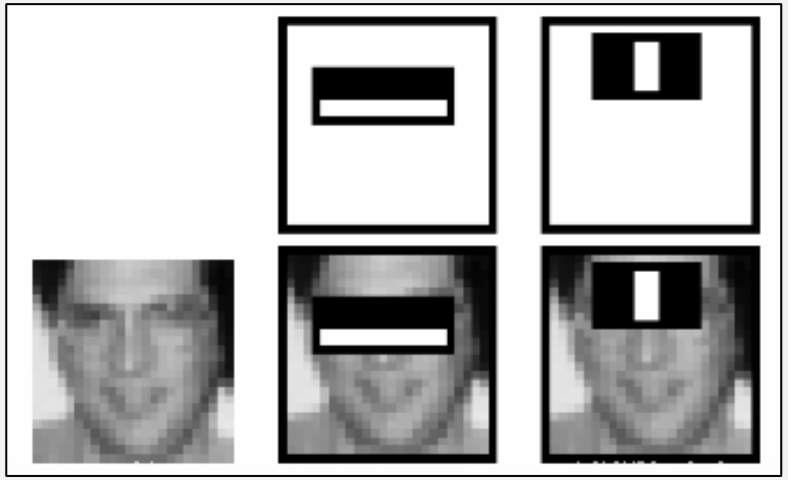


Step 4: Face Detection (HAAR BASED)

Objective: Identify faces in a particular frame before we can compare and check if it's person of interest

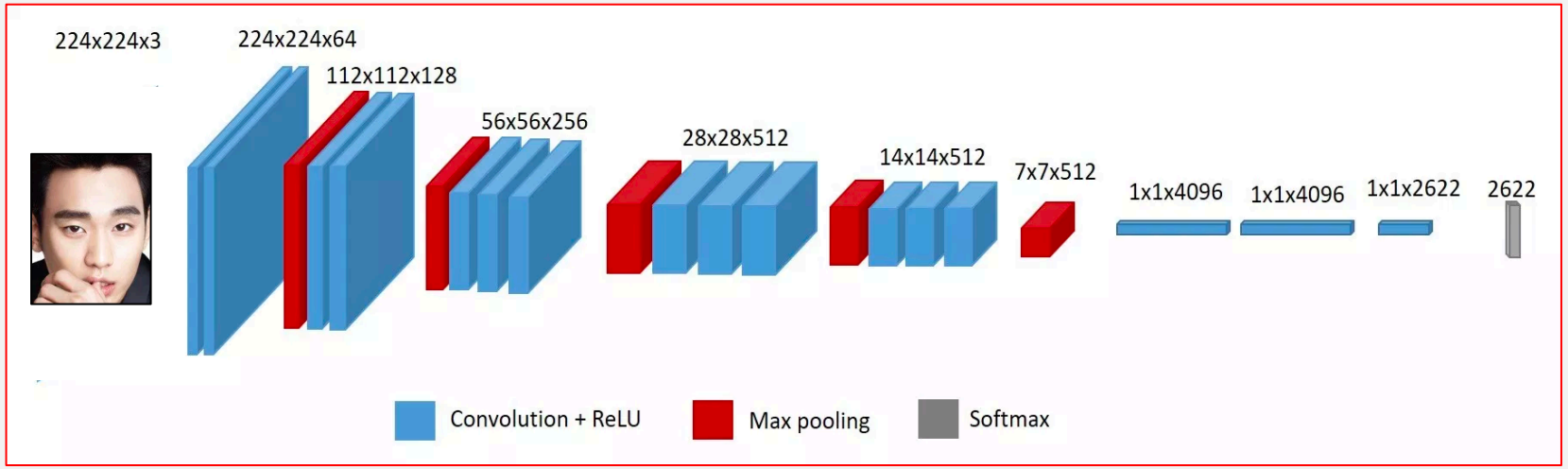
- Change image from RGB to gray scale.
  - Train on positive and negative images
- HAAR FEATURES – 16000+
  - 1st feature: region of eye darker than region of cheeks
  - 2nd feature: eye is darker than bridge of nose
- Weak Classifier (Adaboost)





Step 4: Face Verification

Objective: Compare detected face in video to template face learned from earlier step



- Extract features for detected face → detected_face
- Compare feature to template face → template_face
  - val = cosine(detected_face, template_face)
- If val >= 0.6 → There is a match!!

WHAT NEXT?

Attempt 1: Detect and Verify all Faces

Find face in every frame and then do verification on that!!

ISSUE:

- Extremely slow since we have to process entire frame
- Not smooth since verification might fail if used on every frame (1:50)

Attempt 2: Detect and Track Faces

Find a verify a particular face. Track that face for the duration of scene.

ISSUE:

- If scene detect misses some scenes, tracking continues past subject of interest.
- Fails if current shot zooms in or out slowly
- (3:30, 8:12, 8:20, 2:57 (ISSUE))

Attempt 3: Redetect Face within Local Region

pindrop

Find a verify a particular face. Track that face by redetecting that face in a local region

ISSUE:

- Tracking box jitters a little more than wanted
- Slight change in face orientation or object on face leads to misdetection
- (0:57, little jittery… 1:04)

Attempt 4: Combine Tracking and Detection

Best Solution

Find a verify a particular face. Track that face for 5 frames. Then redetect face in the local region

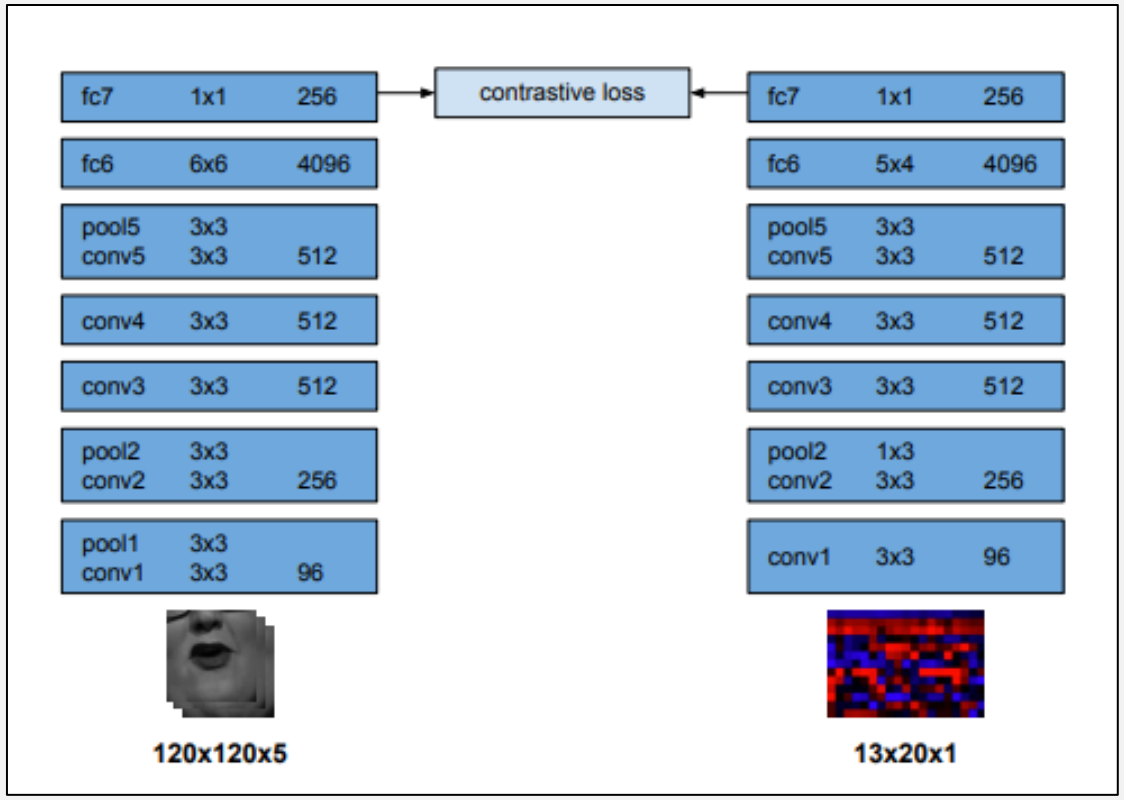
(0:48)

pindrop

Step 5 : SyncNet

- Synchronize audio to lip movement
- Take Mel Frequency Cepstral Features from audio
  - Short term power spectrum on mel scale
- Take Images from video as input
- Train two models simultaneously.
  - One model for audio
  - One model for images on video frame

Syncnet



Models Trained Simultaneously

pindrop

MORE DEMO

pindrop

THANKS!!
QUESTIONS?