COMP7503 Multimedia Technologies

Example Applications

Dr. Bill Luo

THE UNIVERSITY OF HONG KONG

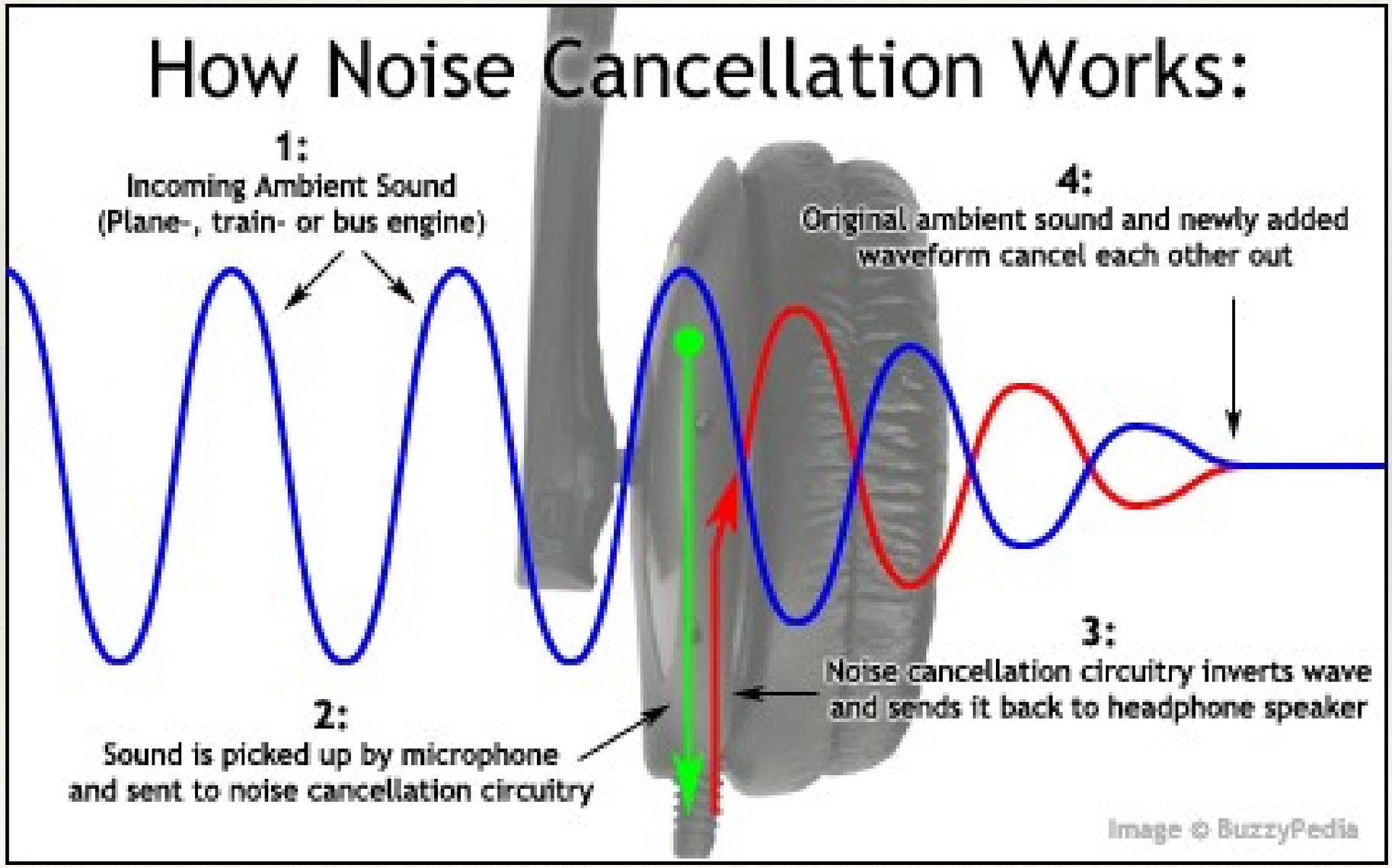
DEPARTMENT OF COMPUTER SCIENCE

Audio Processing

- Noise Cancellation
- Digital Audio Equalizer
- Speech-To-Text
- Shazam
- Audio Compression

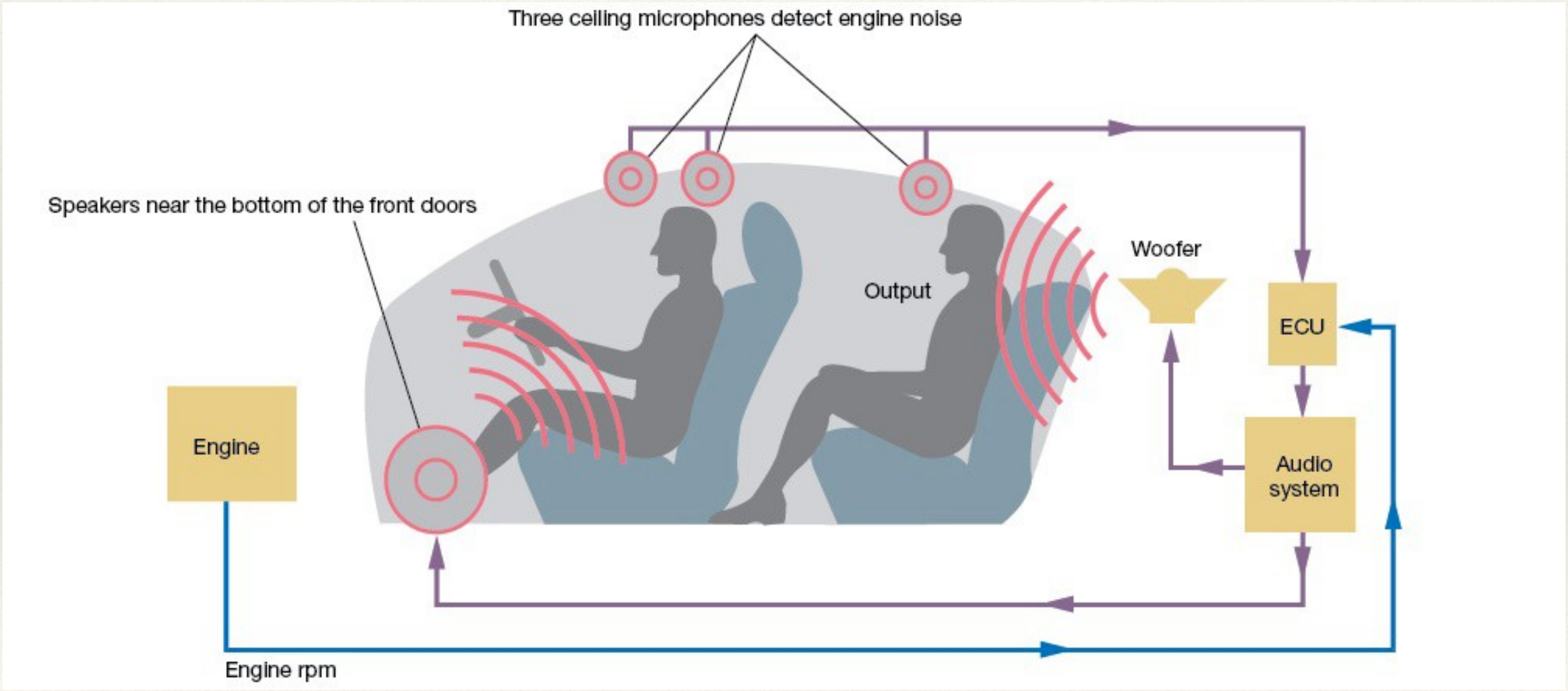
Noise Cancellation

- Noise Cancellation Headphone



Noise Cancellation

- Noise Cancellation in Cars



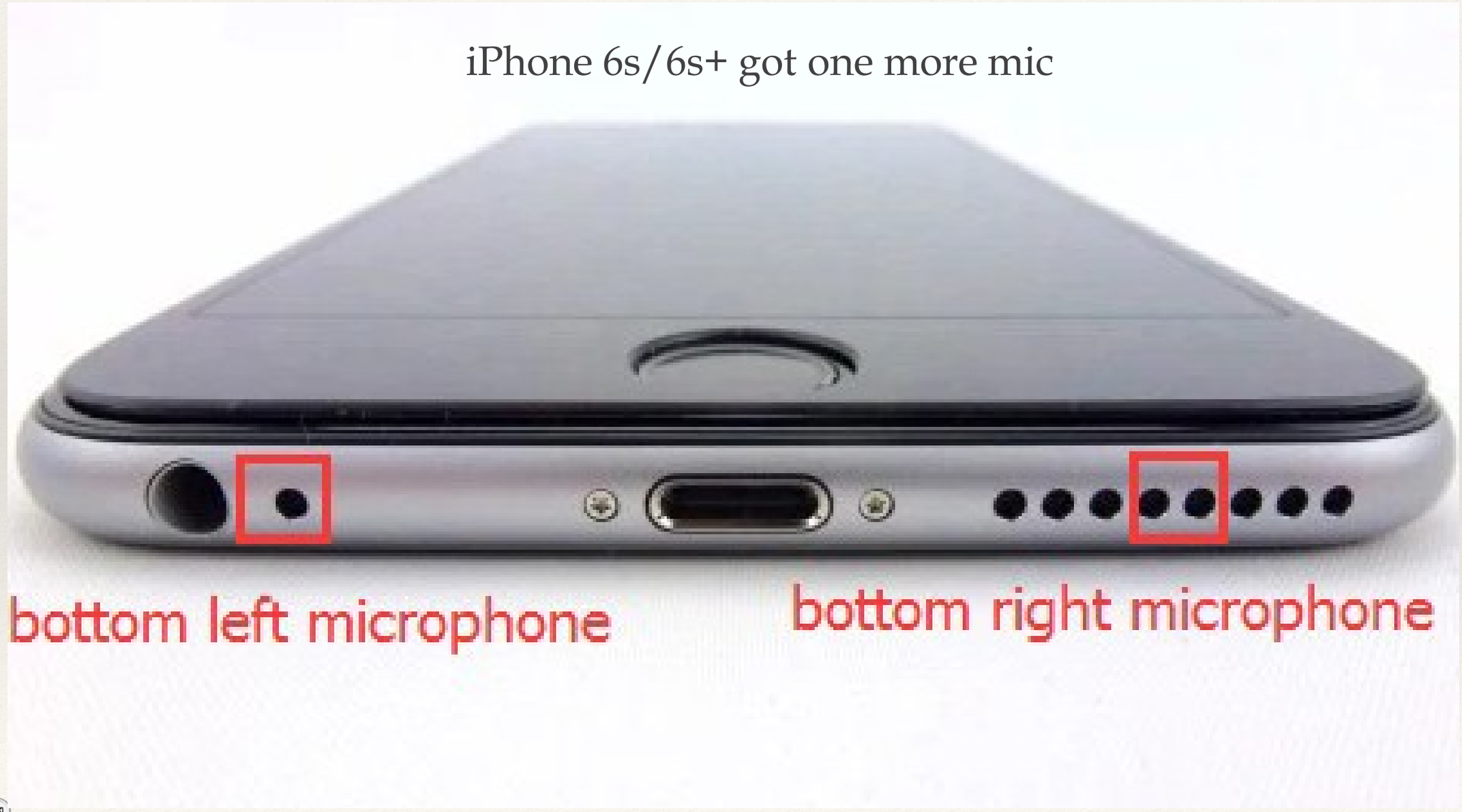
Noise Cancellation

- Noise Cancellation Microphone

- iPhone 4/4s has 2 microphones

Noise Cancellation (Cont'd)

- iPhone 5s/6/6+ have 3 microphones, why?



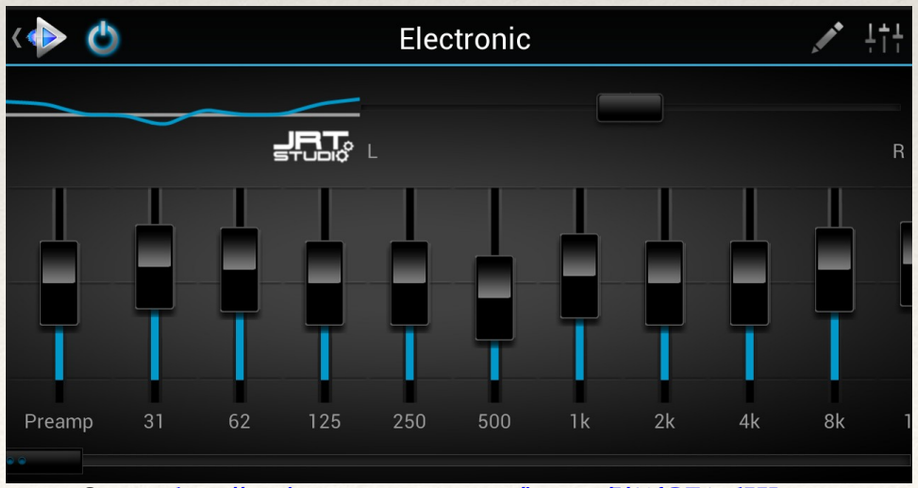
Voice assistants being hampered by poor microphone tech

- Since the launch of the iPhone 5 in 2012, microphone technology hasn't advanced significantly, IHS Markit analyst Marwan Boustany explained to Bloomberg. As a result, mics still have difficulty picking up distant voices and filtering out background sounds. Even without these issues, keeping a mic on all the time — needed for voice triggers like "Hey Siri" — can sometimes consume too much battery life.
- Companies like Apple are said to be looking for better mics from suppliers to fix these problems, as well as achieve a higher acoustic overload point less likely to be tripped. At the same time, size and power consumption need to be kept under control.
- Partly to compensate for poor pickup, device makers have gradually been adding more microphones in recent years. While the first iPhone had a single mic, there are three in the iPhone 6, and four in the iPhone 6s. The Amazon Echo has seven, being a device that needs to hear users from anywhere in a room.
- Some possible solutions may include mics with built-in audio processing algorithms, or ones using piezoelectric technology.
- Microphones could become extremely important to Apple if the company is indeed developing an Echo competitor, whether in the form of a standalone product or an upgraded Apple TV. It's unknown how many mics the "iPhone 7" might be equipped with.

Source: http://appleinsider.com/articles/16/08/25/apples-siri-rivals-being-hampered-by-poor-microphone-tech

Digital Audio Equalizer

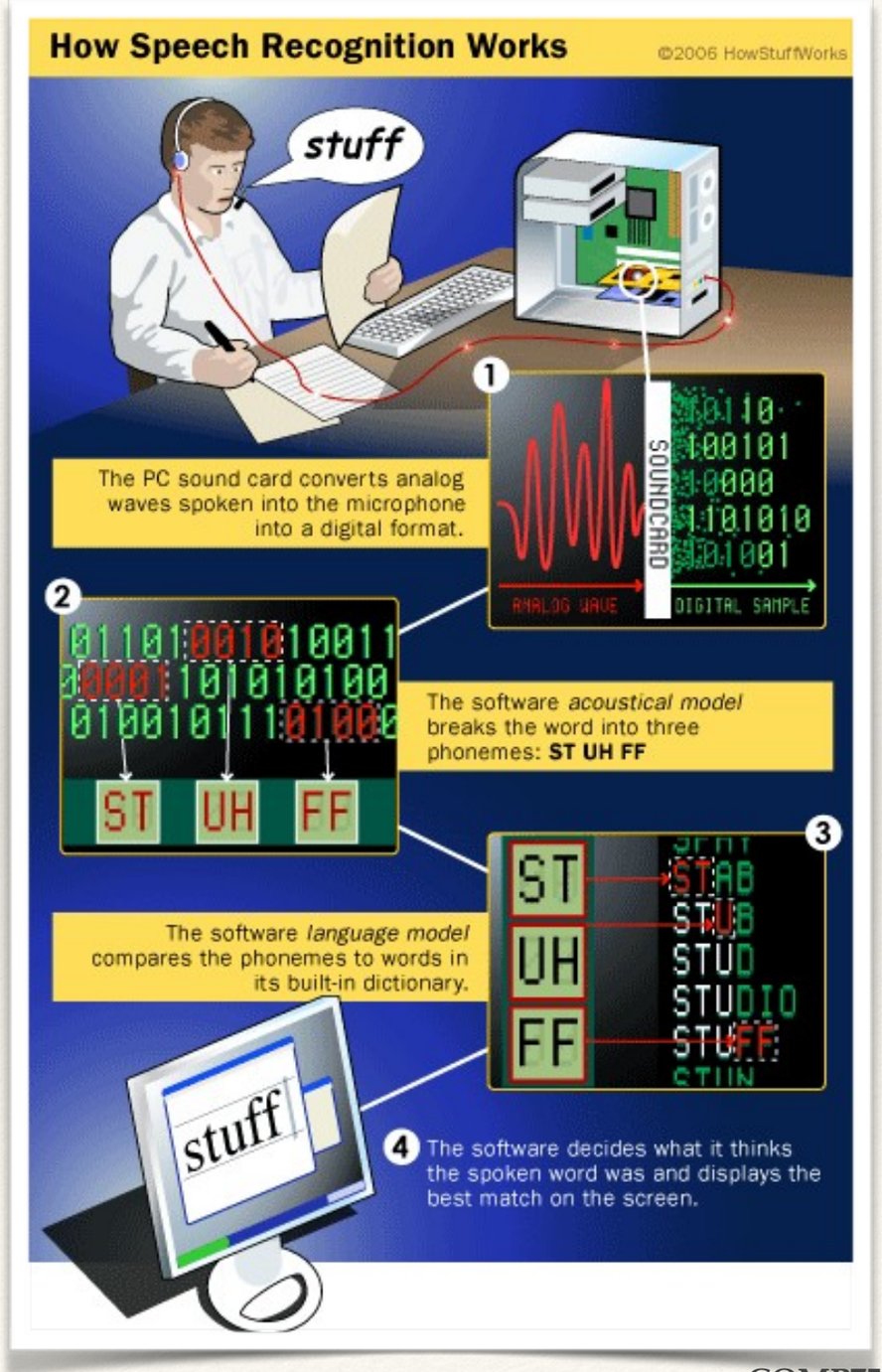
- Allow fine tuning of the gain or attenuation of different frequency bands for the audio signal
  - Frequency bands cover the full audio range of 20Hz to 20Khz
    - Number of frequency bands could be more than 200 bands for high end system
    - Gain/Attenuation ranges will vary from +/- 6 dB to +/- 24 dB.
- Equalizer preset can be used instead of tuning individual gain
  - Bass Booster, Classical, Jazz, R&B, Rock, etc.



Source: http://ecx.images-amazon.com/images/I/61fGZ4pd55L.png

Speech-to-Text

- A/D translates analog wave into digital data
- Noise reduction and signal normalisation is performed first
- Digitized audio data is divided into small segments (as short as a few hundredths/ thousandths of a second)
- Audio segments are then matches against known ***phonemes*** (~40 phonemes in English)
- Map phonemes to works through complex statistical model



Shazam - Audio Search

- Shazam is a service that takes short sample of music, and identifies the song
- Use time-frequency graph (spectrogram) to work out the "fingerprint" of each song in the catalog
- User then take 10 second sample of audio, work out the fingerprint, and Shazam service would match this against its catalog of music
- Matching is based on spectrogram peaks
  - Robust to noise, room reverb, equalisation, overlapping sounds
- Original paper can be found at
  - http://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf

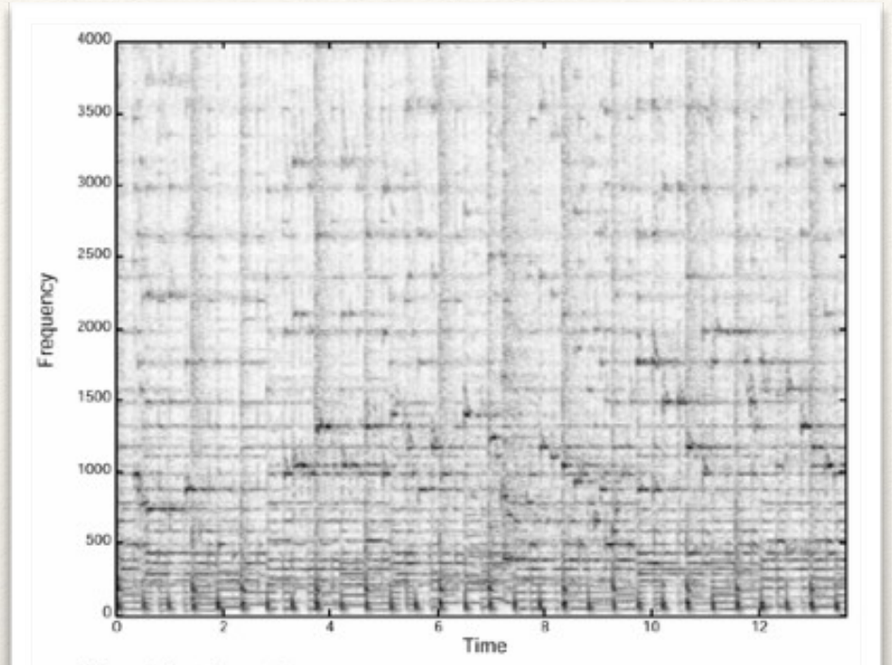
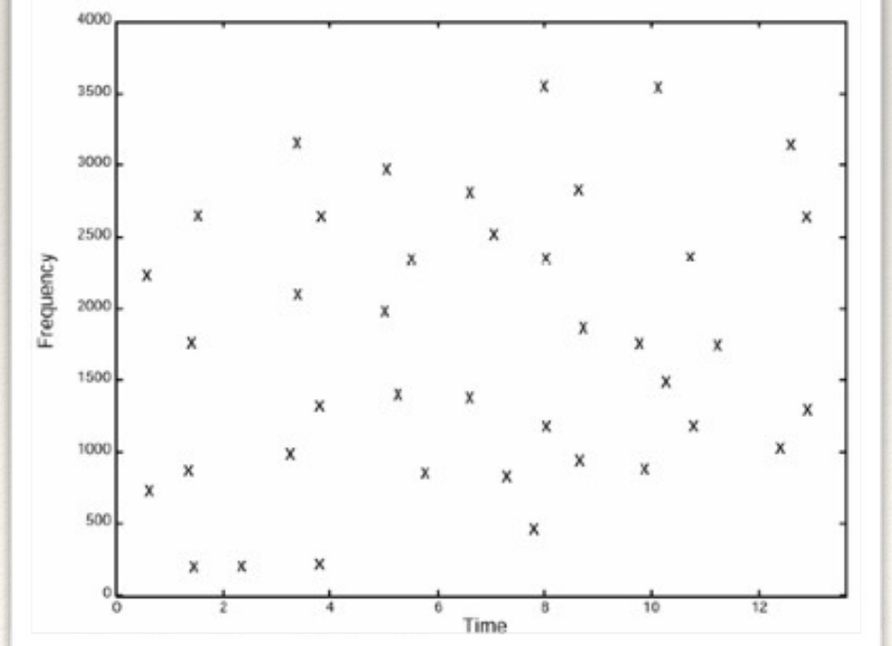
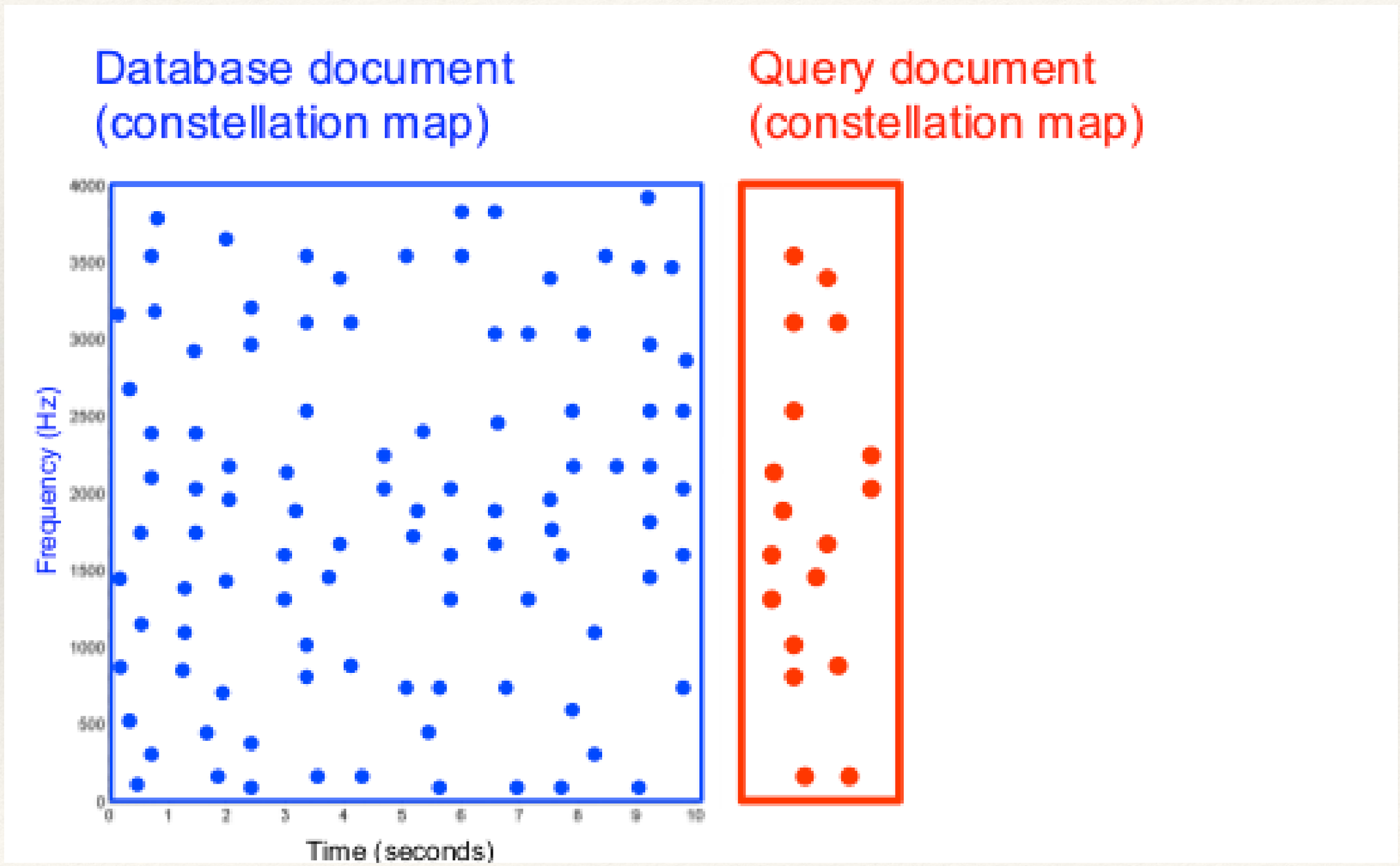


Fig. 1A - Spectrogram

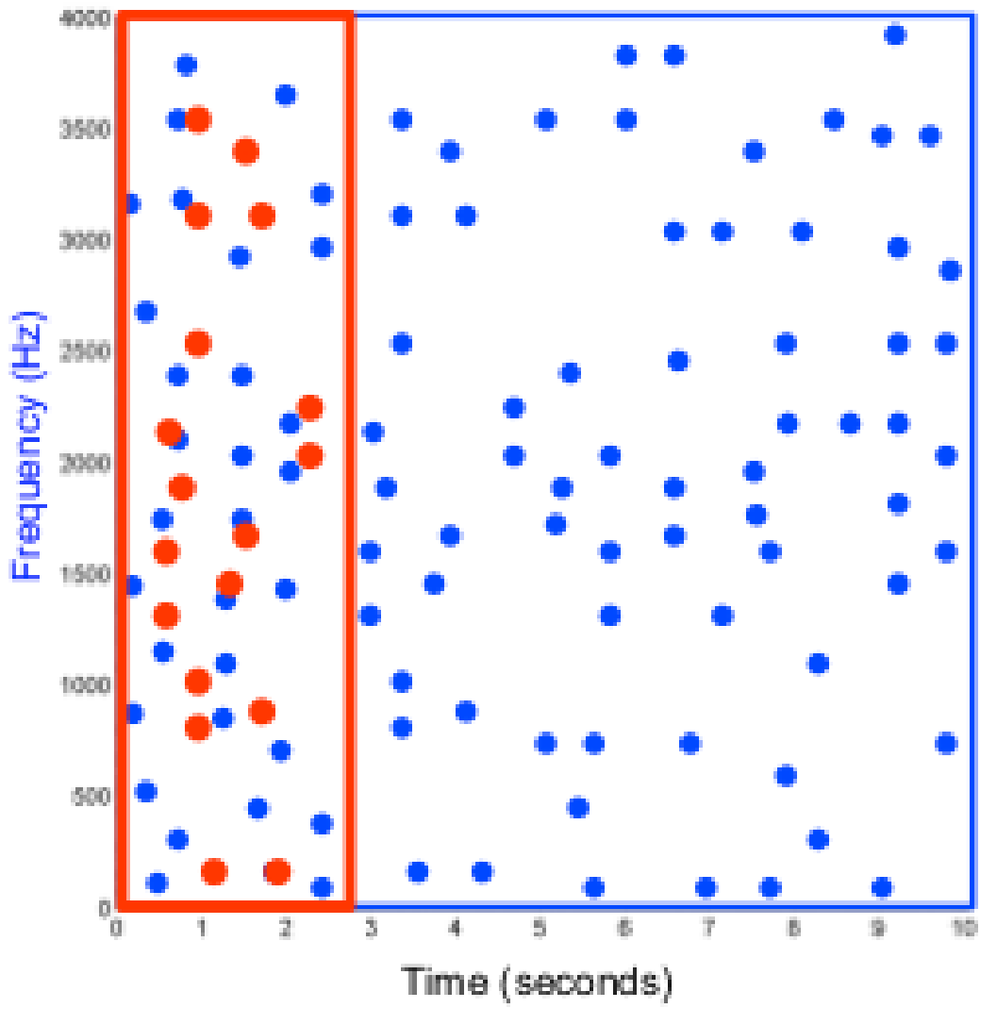


Shazam - Audio Search (Cont'd)



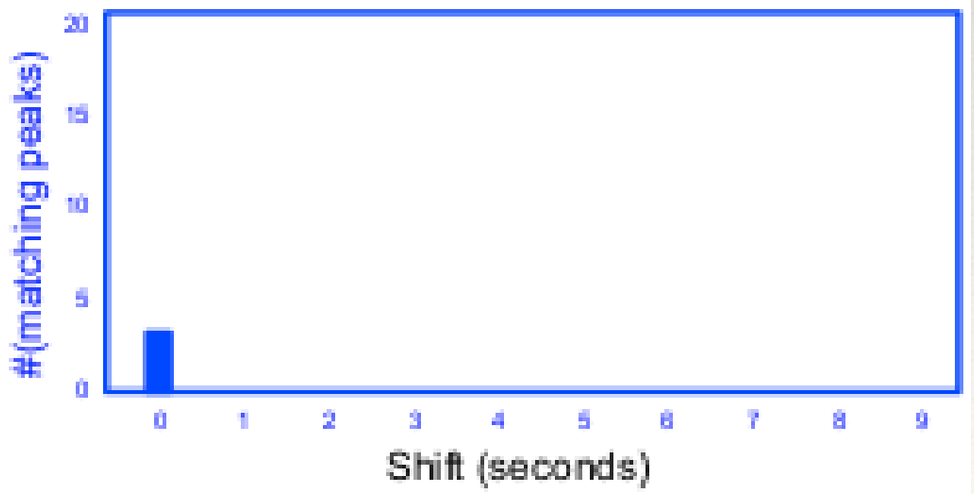
Shazam - Audio Search (Cont'd)

Database document
(constellation map)



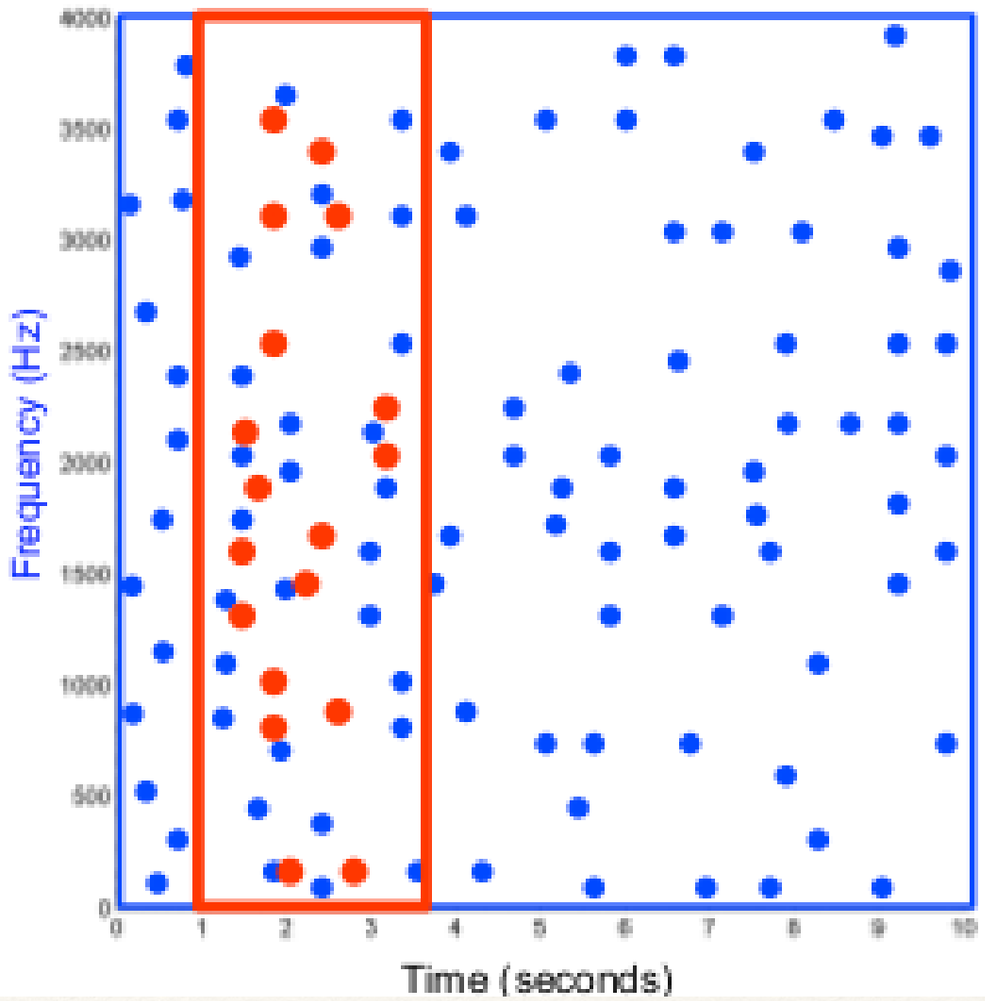
Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks



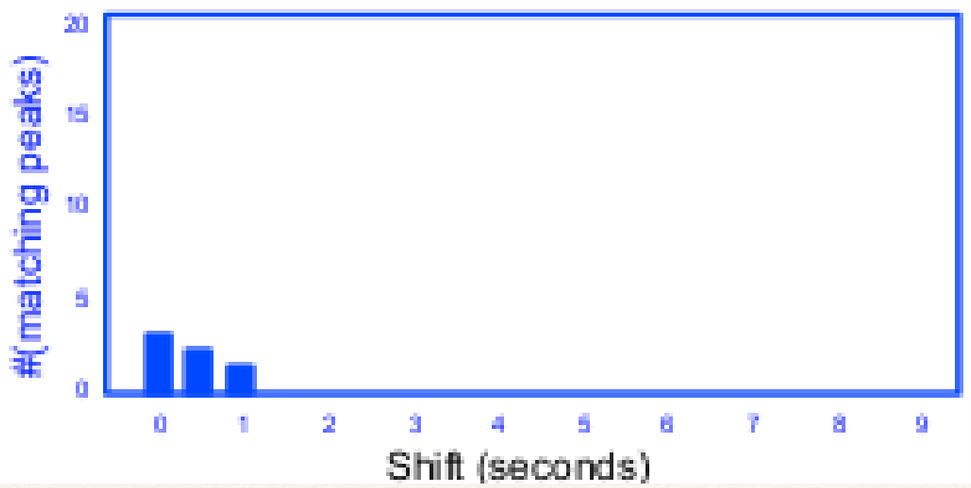
Shazam - Audio Search (Cont'd)

Database document
(constellation map)



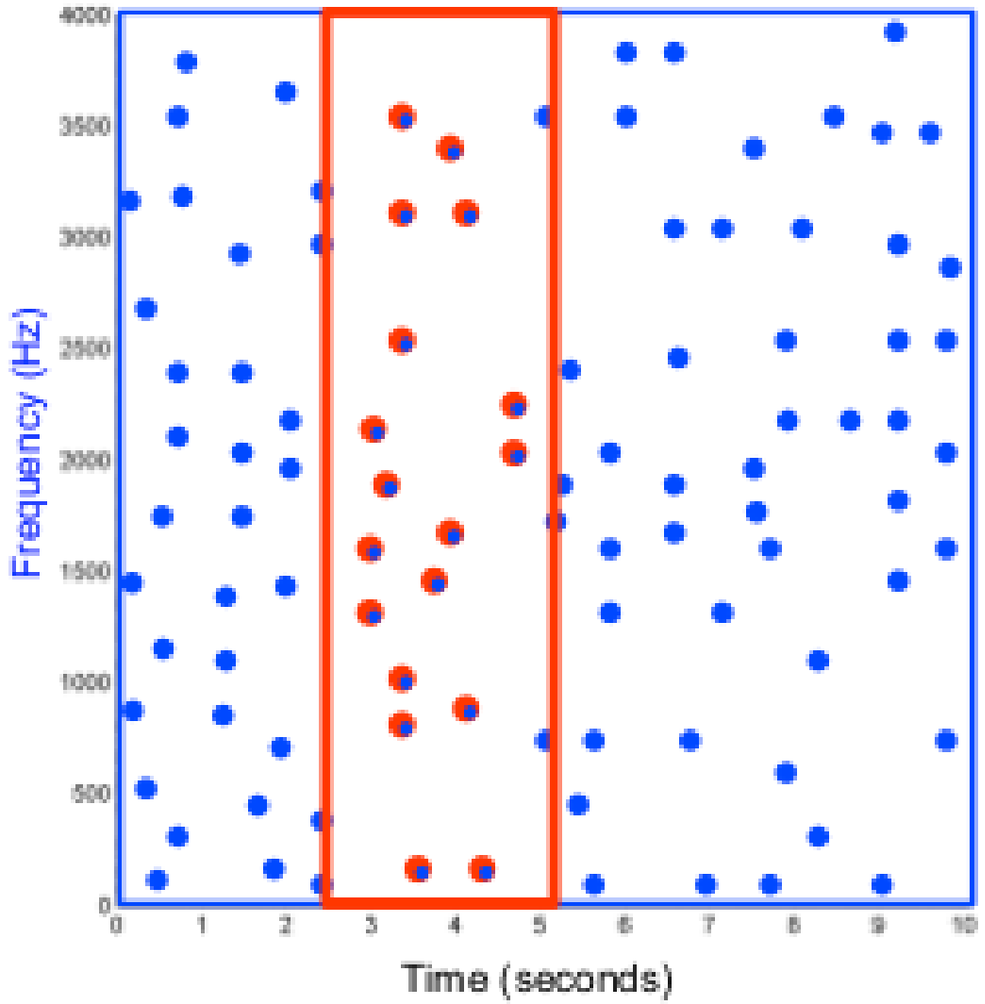
Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks



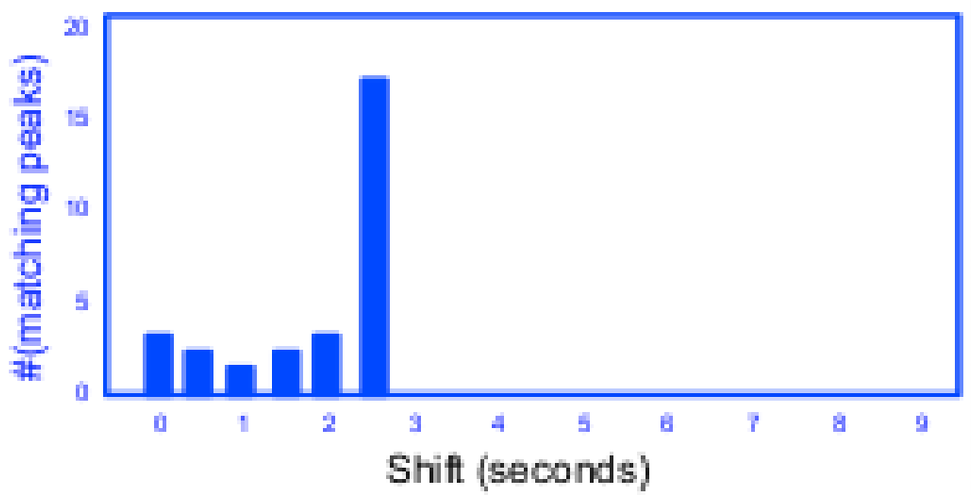
Shazam - Audio Search (Cont'd)

Database document
(constellation map)



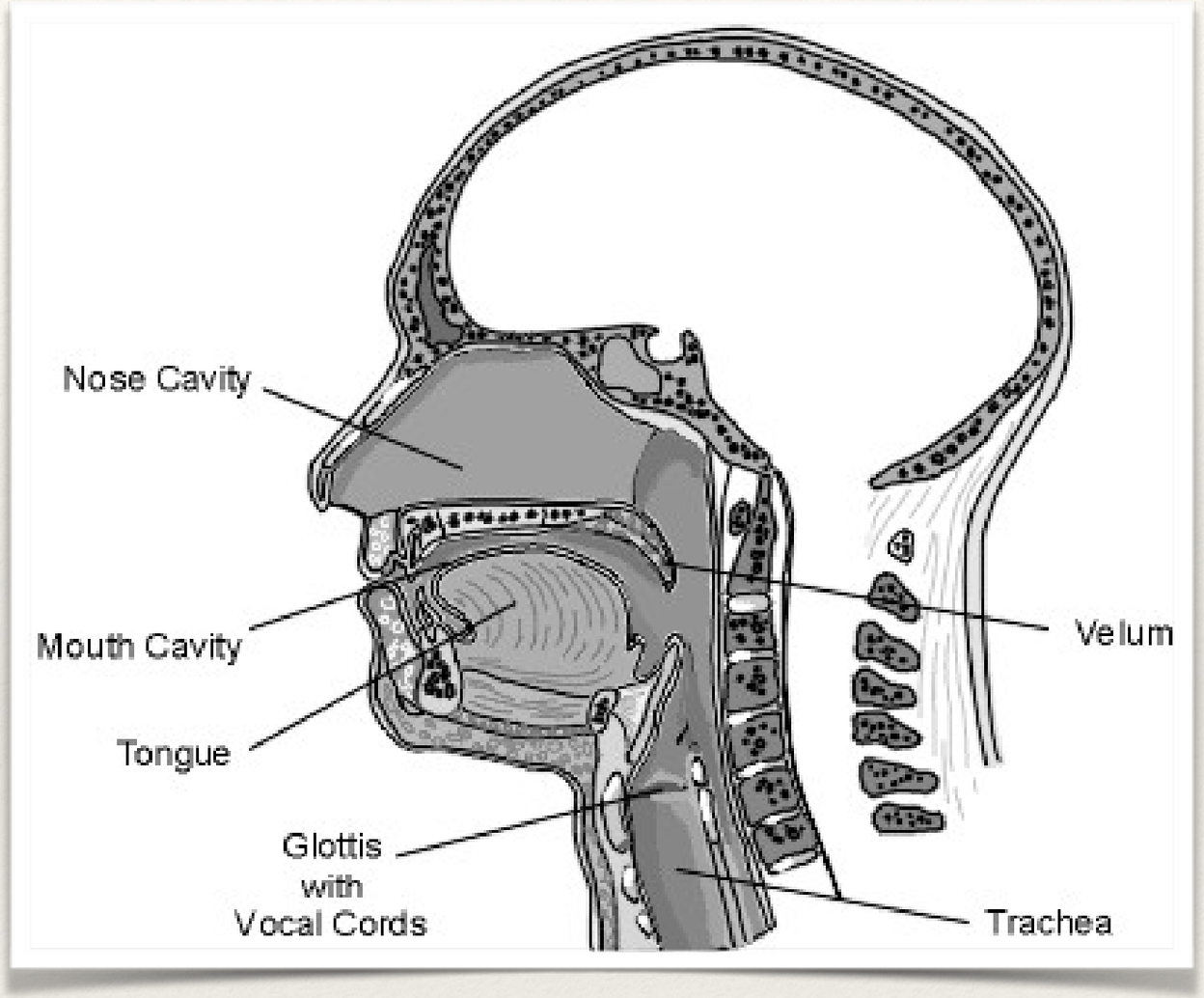
Query document
(constellation map)

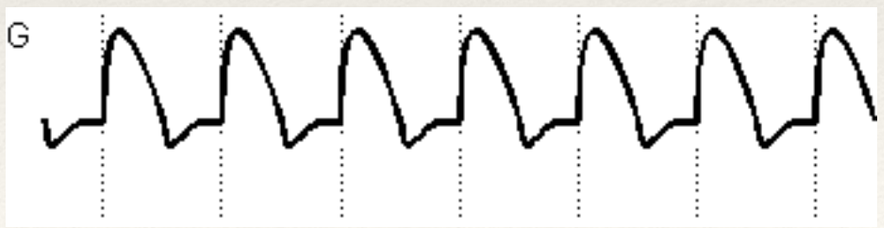
1. Shift query across database document
2. Count matching peaks



Audio Compression

- LPC Vocoder
  - Based on the human vocal model
  - Vocoder - named based on the concatenation of the terms ‘voice’ and ‘coder’
- Speech is produced by a pressing air through epiglottis, which causes vibration of vocal cords (voiced) and/or resonance in articulation tract (mouth and nose cavity)

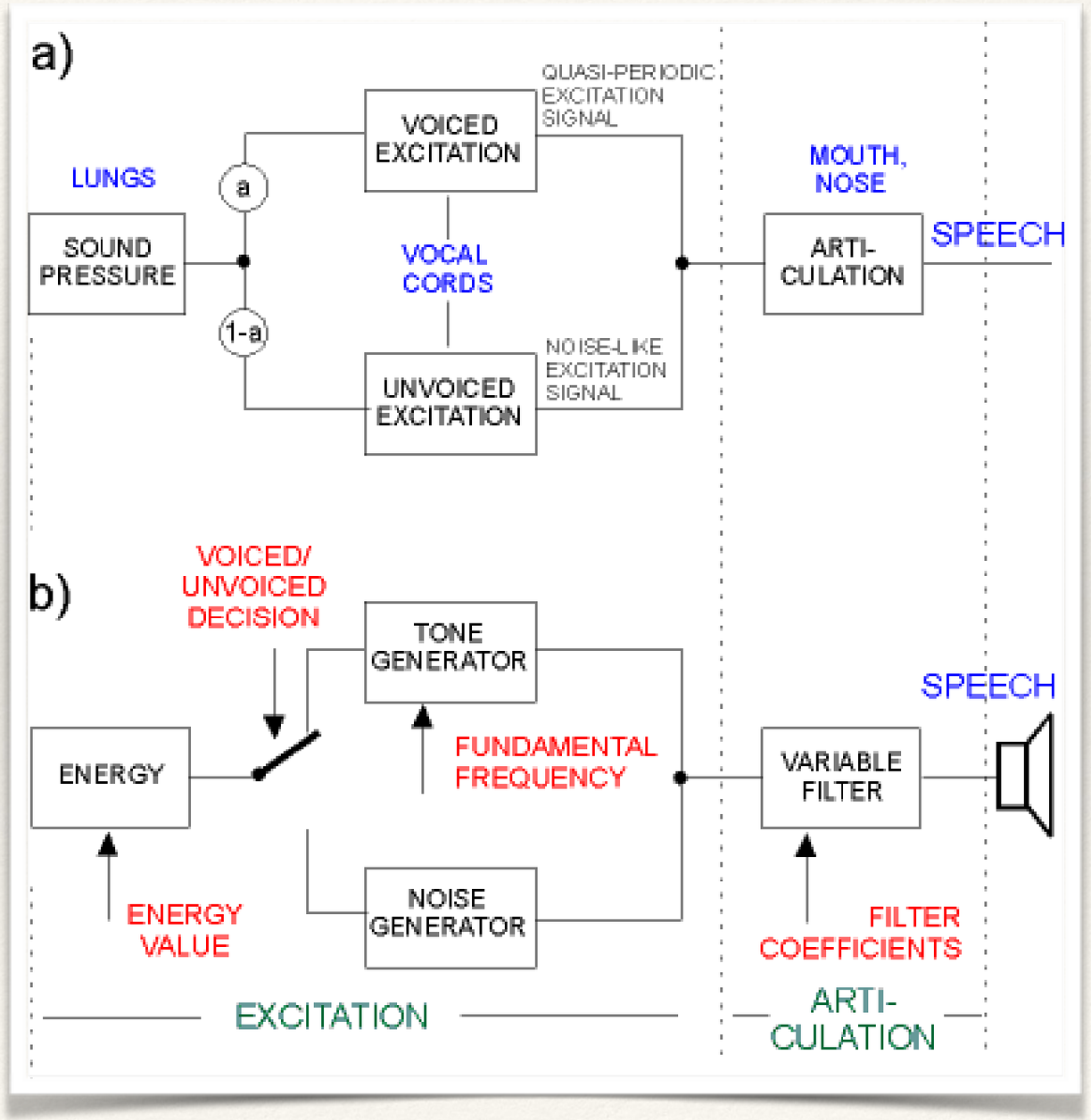




Descriptions and photos from this slides can be found at
http://www.iro.umontreal.ca/~mignotte/IFT3205/APPLETS/HumanSpeechProduction/HumanSpeechProduction.html

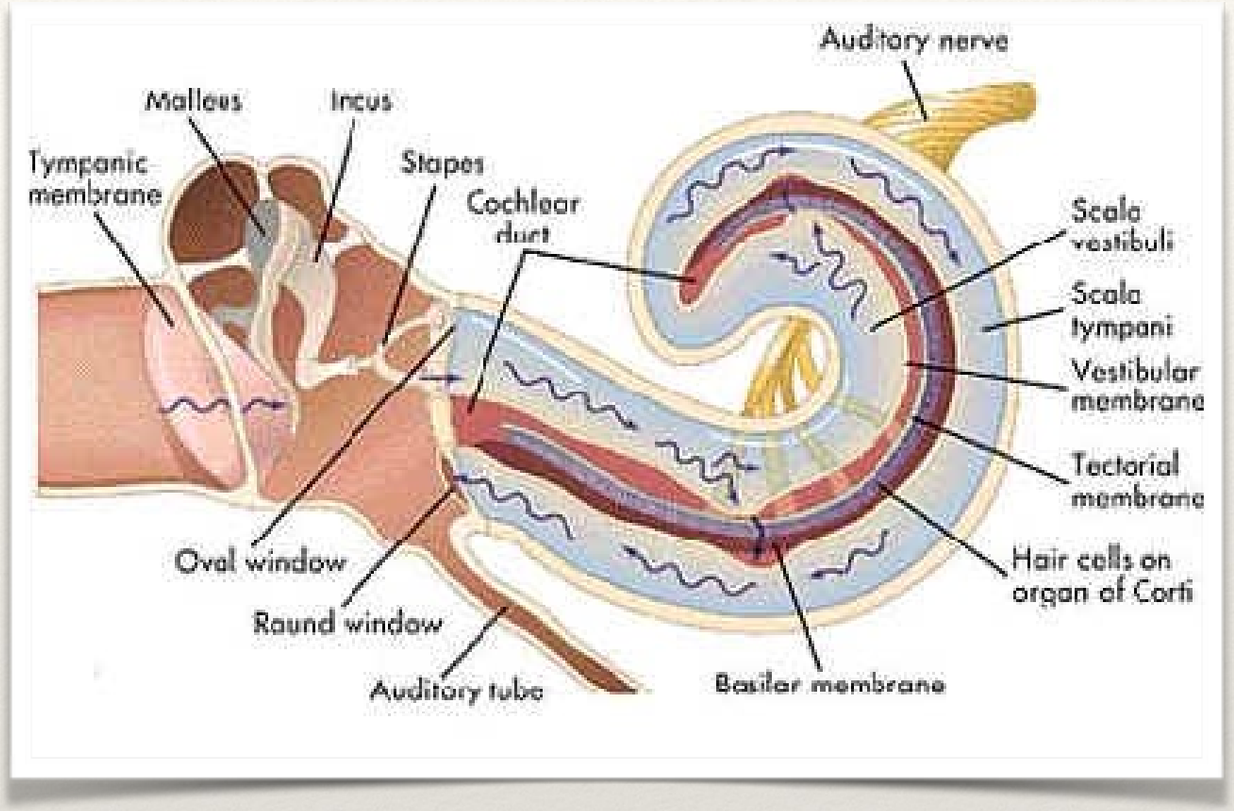
LPC Vocoder

- Compression of speech can be done by estimating speech parameters
  - The filter coefficients that models the articulation track
  - Energy, Pitch, Voiced/ Unvoiced
- GSM voice encoding is based on LPC Vocoder framework
- What is the other use of this Vocoder?

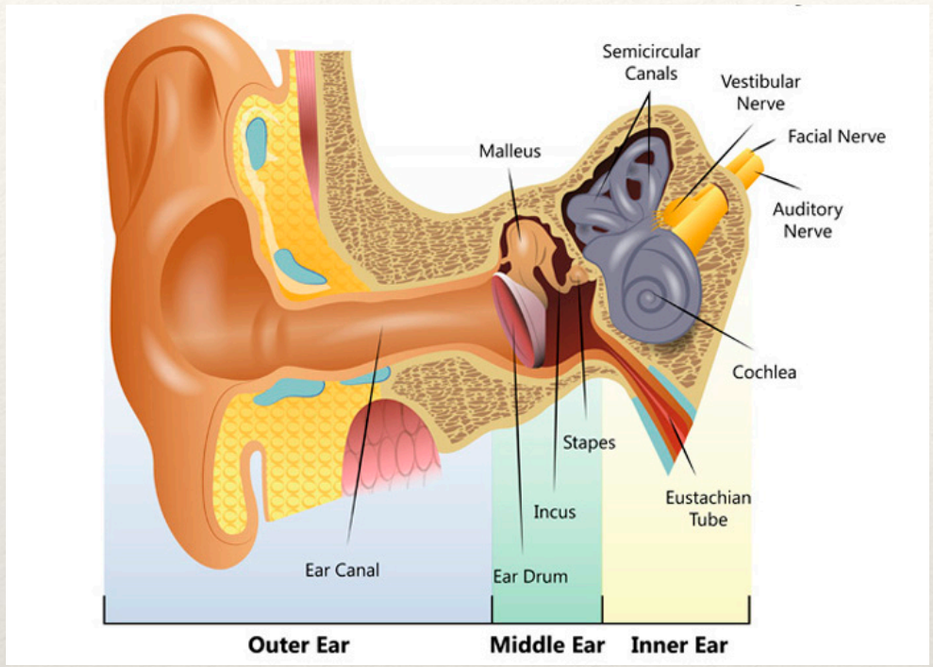


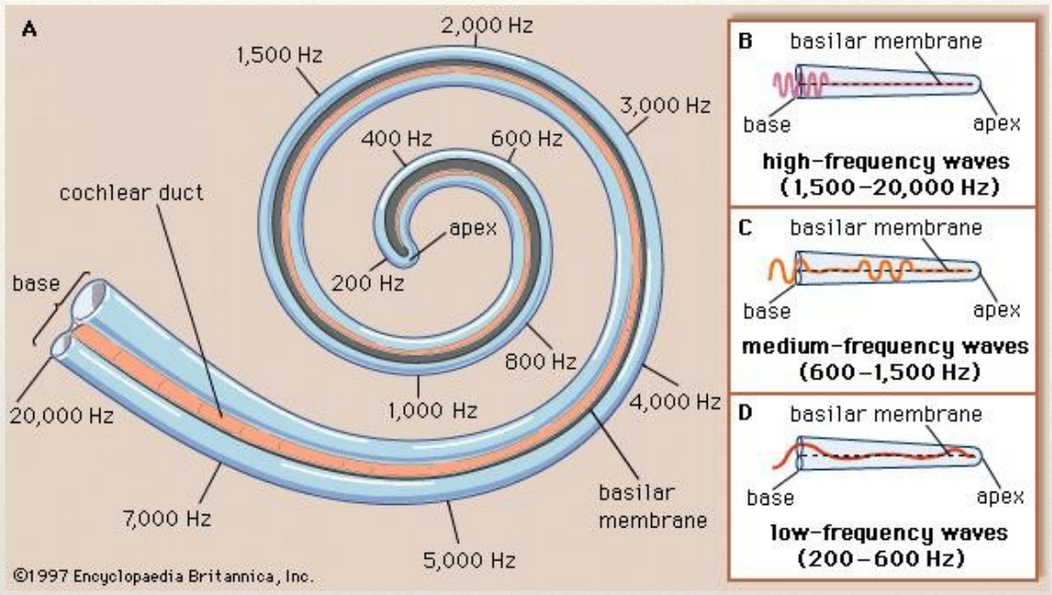
Audio Compression - MPEG Audio

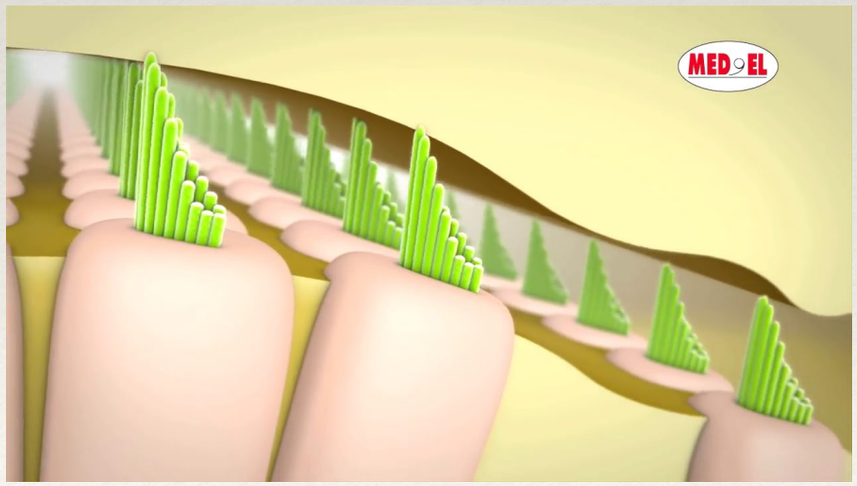
- MPEG Audio
  - Lossy Encoding
  - Based on Frequency Masking under psychoacoustic model
- How human ear works?
  - Mechanical vibration of ear drum transforms to hydraulic pressure in cochlea, which is filled with fluid
  - Pressure waves in cochlea causes stimulation to 20,000+ hair-like nerve cells
  - High tones stimulates more for short hair, and vice versa
  - Human ear senses the spectrogram of sound



Audio Compression - MPEG Audio



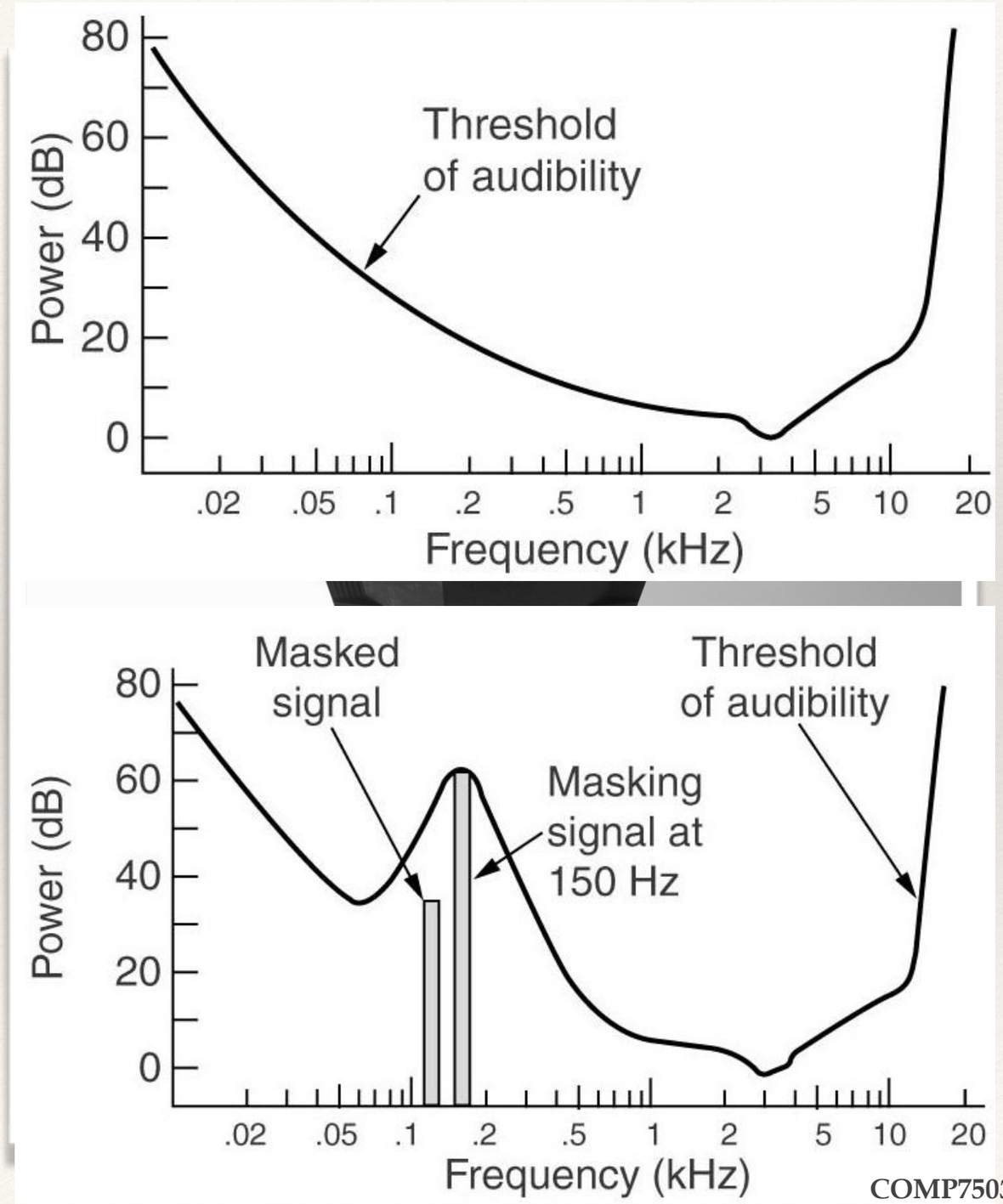




https://www.youtube.com/watch?v=flIAxGsV1q0

MPEG Audio

- Frequency Masking
  - The ability of a loud sound in one frequency band to hide a softer sound in another frequency band that would have been audible in the absence of the loud sound.
- Temporal Masking
  - After a loud sound stops, a quiet sound will be inaudible for a short period of time because the ear turns down its gain when it starts and it takes a finite time to turn it up again.



MPEG Audio (Cont'd)

- Three layers, each layer applies additional optimisations
  - Layer 1: Basic Scheme (used in DCC digital tapes)
  - Layer 2: Advanced Bit Allocation
  - Layer 3: Adds hybrid filters
- MP3 encoding
  - Process samples in groups of 1152 (about 26 msec worth)
  - Divide the samples into 32 frequency bands
  - Determine masked frequencies (using some psychoacoustic model)
  - Use variable length coding to explore coding redundancies

Audio Processing

- Noise Cancellation
- Digital Audio Equalizer
- Speech-To-Text
- Shazam
- Audio Compression

Video Processing and Analysis

- Image/Video Enhancement
- Video Stabilization
- Motion Deblurring
- Face Detection
- Head Tracker
- 3D Face Reconstruction
- Human Gait Recognition
- Video Retrieval

Image Enhancement

- Night Mode Enhancement
  - by capturing multiple frames shot at various shutter speeds. These frames are then intelligently combined in order to both recover shadow detail and prevent motion blur if your subject happens to be moving.

Ref: Apple Special Event Sep 2019

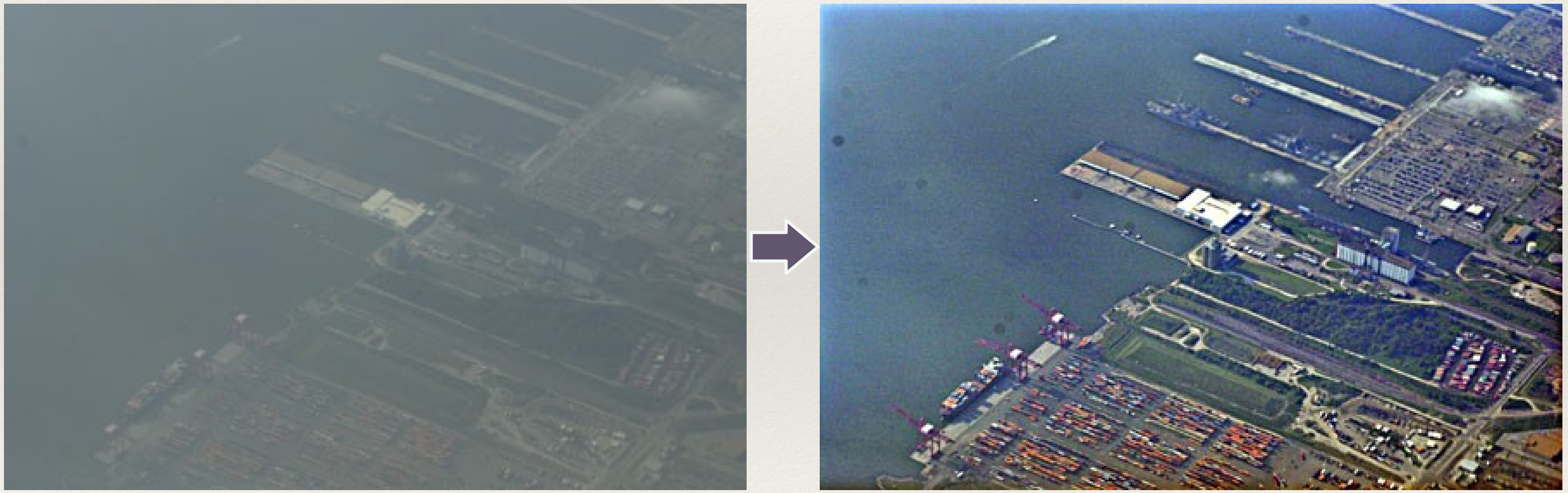
Image Enhancement

- Deep Fusion
  - Each time you press the shutter on the new 11 Pro or 11 Pro Max, the camera captures nine total images: four short images, one long exposure, and four secondary images. These images are then intelligently blended together "to optimize for detail and low noise."

Ref: Apple Special Event Sep 2019

Video Enhancement

- Retinex



Video Stabilization

- Stabilises shaky video footage

https://www.youtube.com/watch?v=flIAxGsV1q0

Motion Deblurring

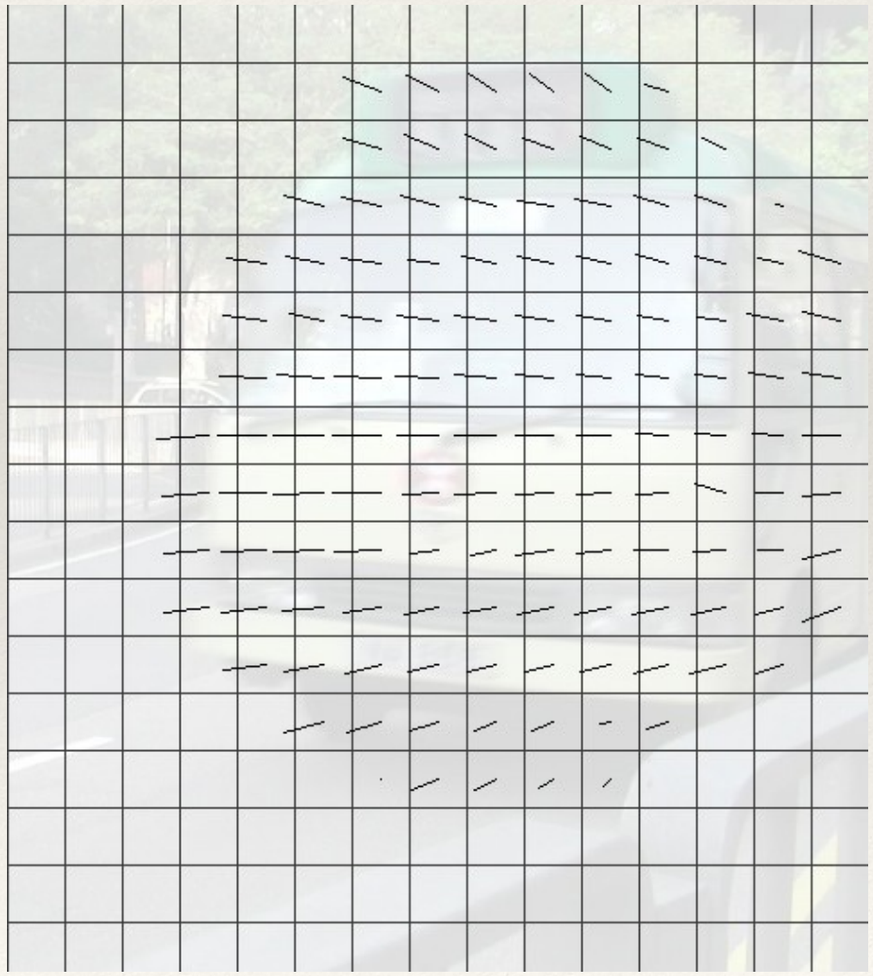
- Motion Blurring Effect

Due to global camera motion

Due to fast moving objects

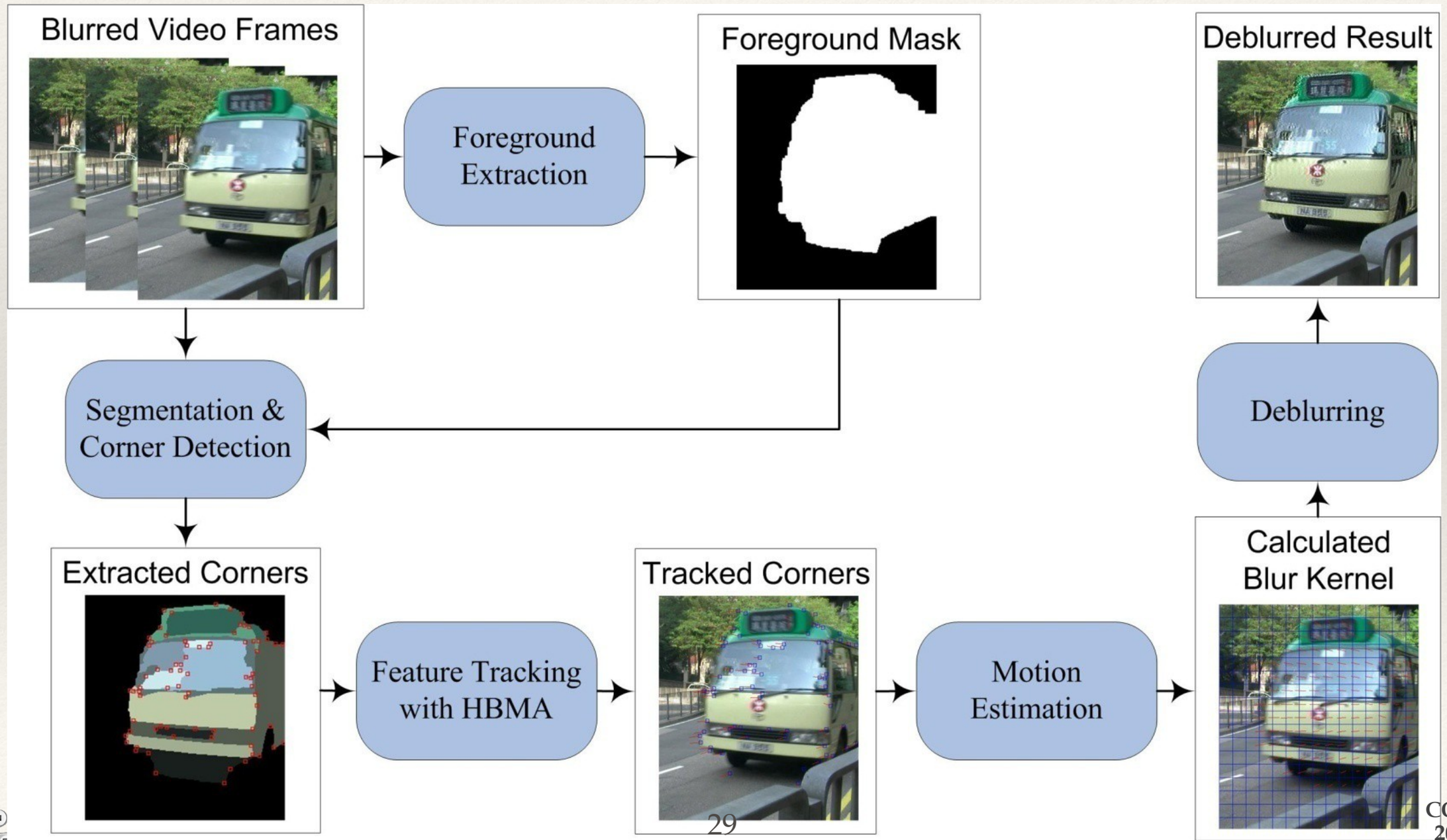
Motion Deblurring (Cont'd)

- Different portions from the same object might have different velocities w.r.t. camera
- Cannot assume single global motion for deblurring



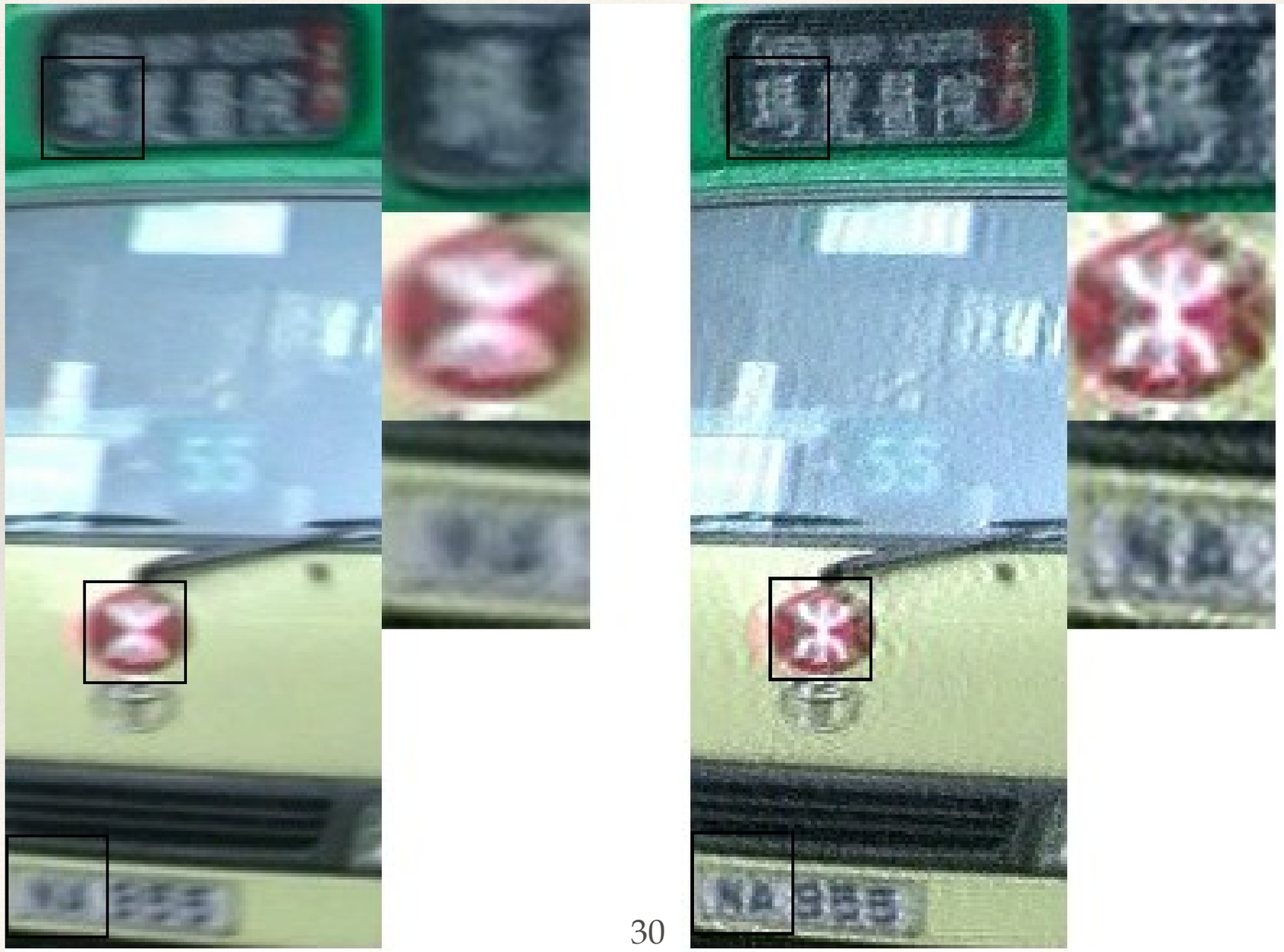
Motion Deblurring (Cont'd)

- Deblurring Processing Workflow



Motion Deblurring (Cont'd)

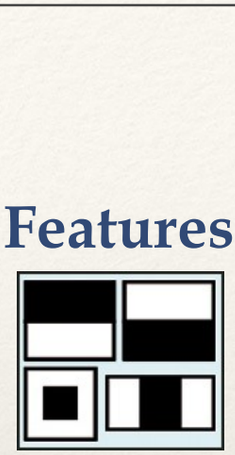
- Deblurring Example

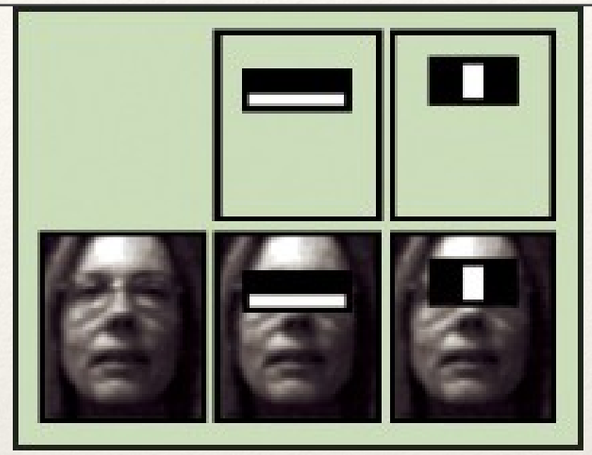


Face Detection (Feature based)

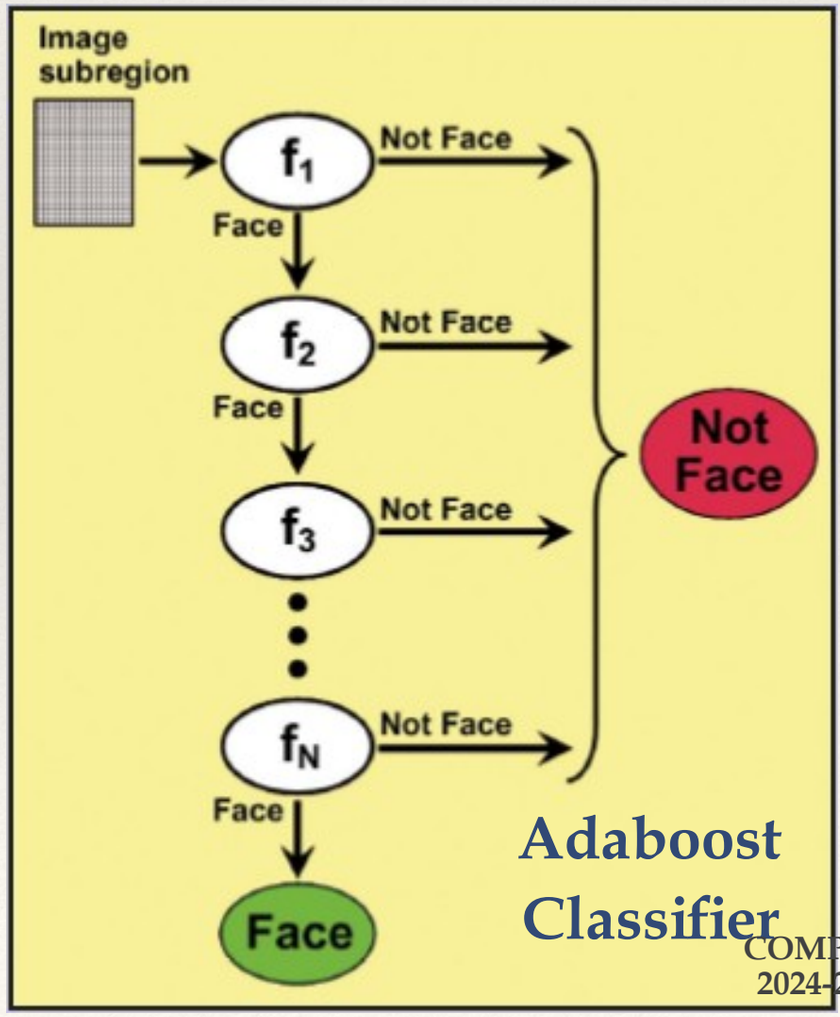
- Haar feature based
  - Proposed by Viola and Jones
  - Can be computed efficiently

Haar Features



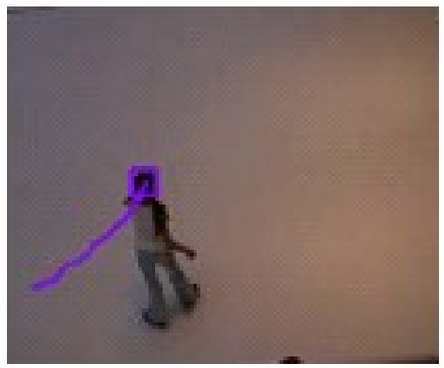


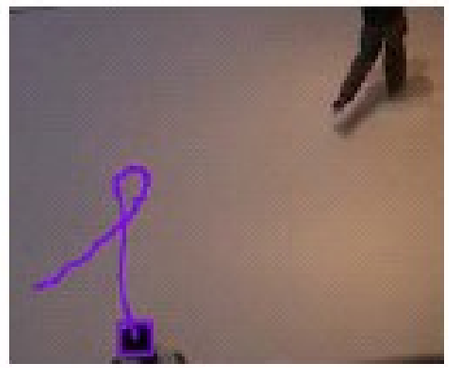
- Adaboost Classifier
  - Combine 'weak' features into a 'stronger' classifier
- For each image, the face detector scans each location to look for presence of face
  - May need to try scaling up or down for different face sizes

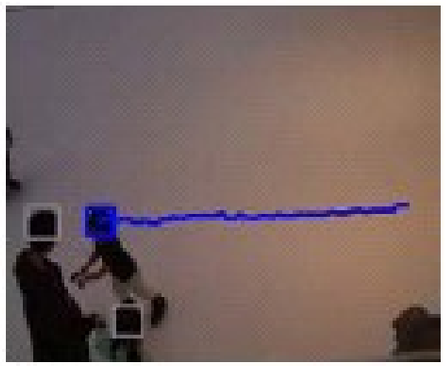


Head Tracker (model based)

(a) frame 153

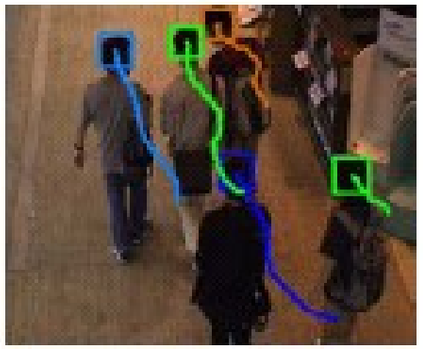






(e) frame 1663

(a) frame 42



(b) frame 69

(c) frame 498

(d) frame 529

(e) frame 549

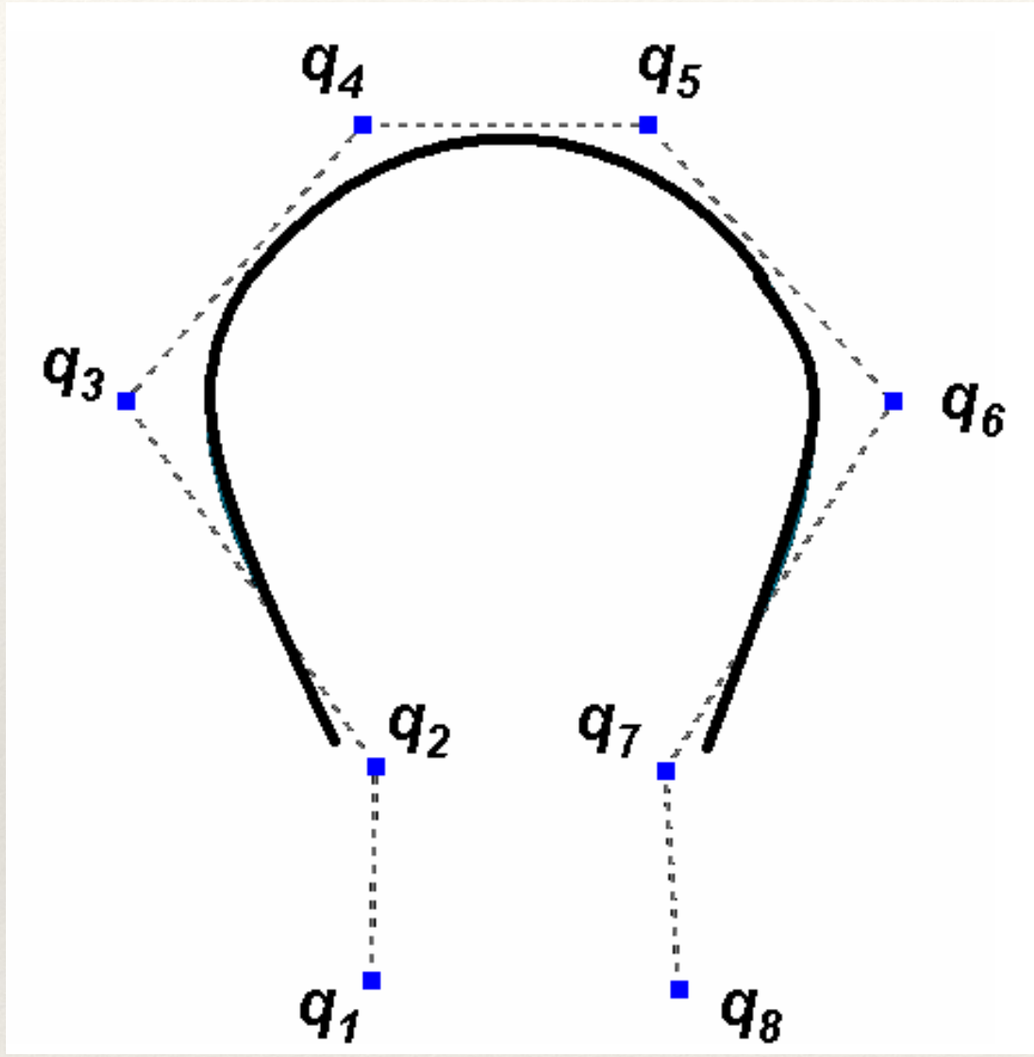
(f) frame 589

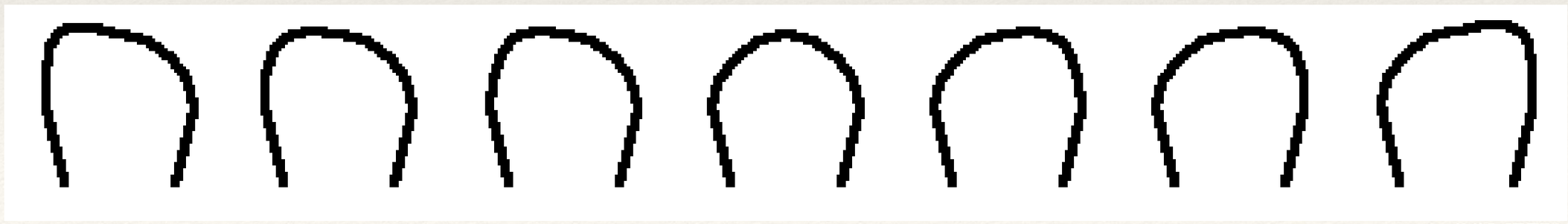
(g) frame 628

(h) frame 644

Head Tracker (Cont'd)

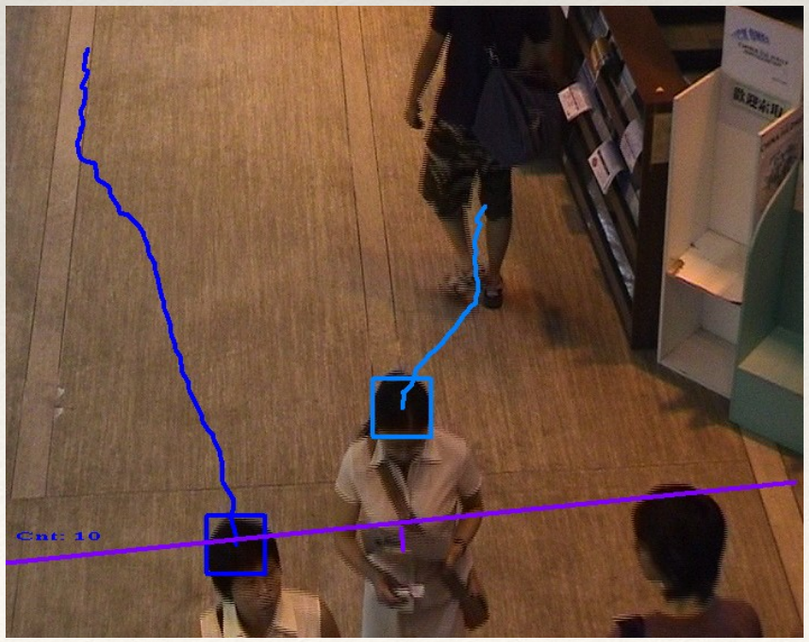
- Side View
- Modelling View from Left to Right



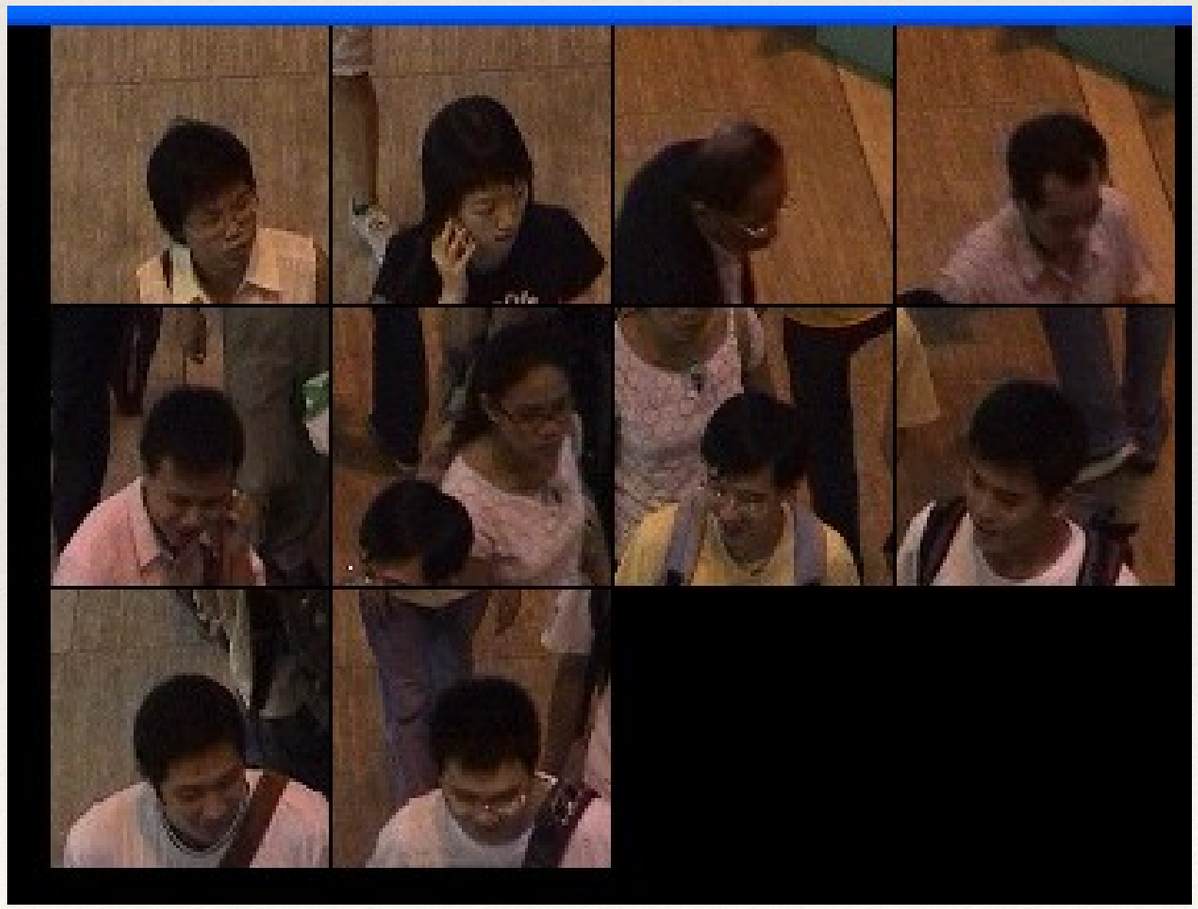


Head Tracker (Cont'd)

People Counting by Tripwire

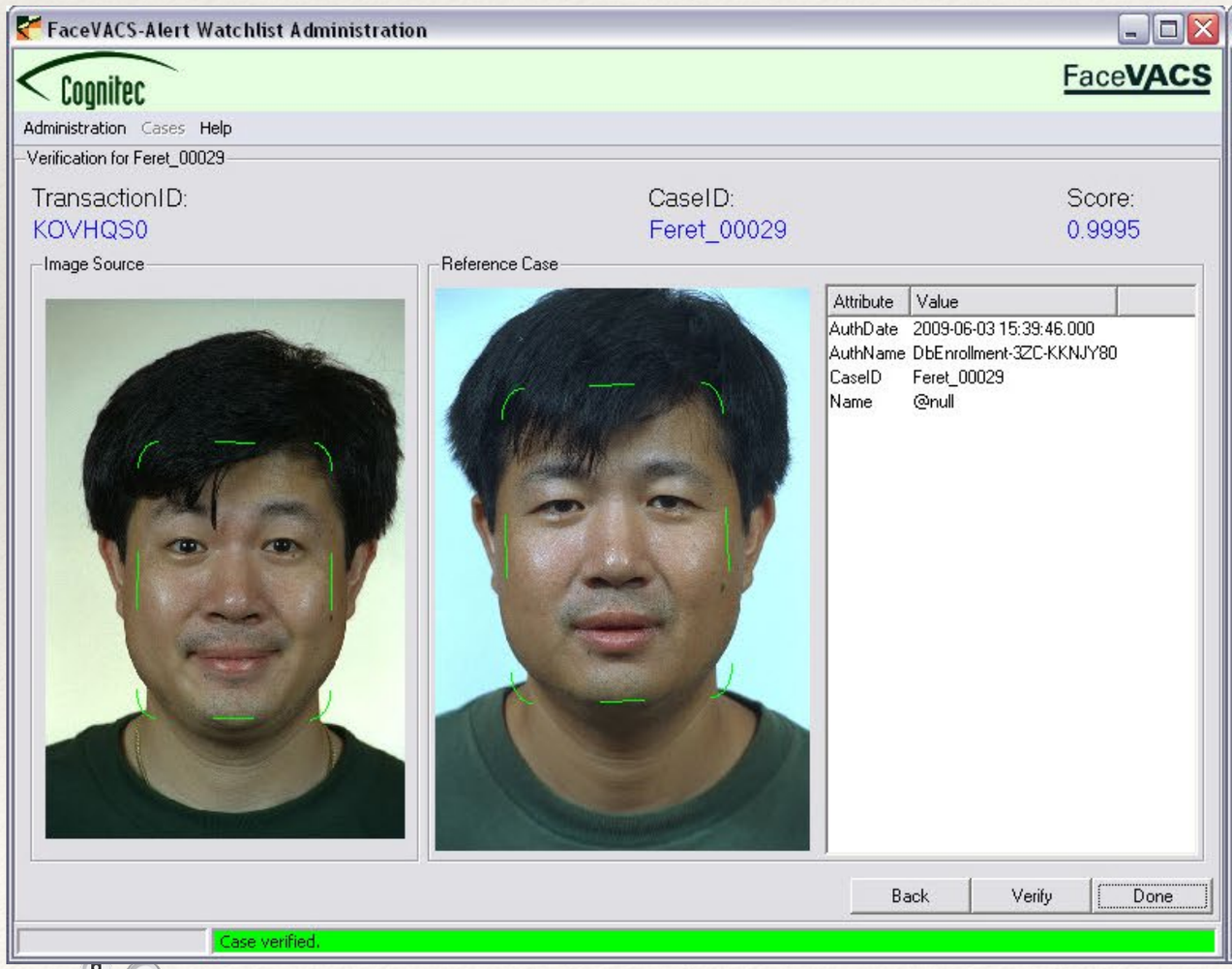


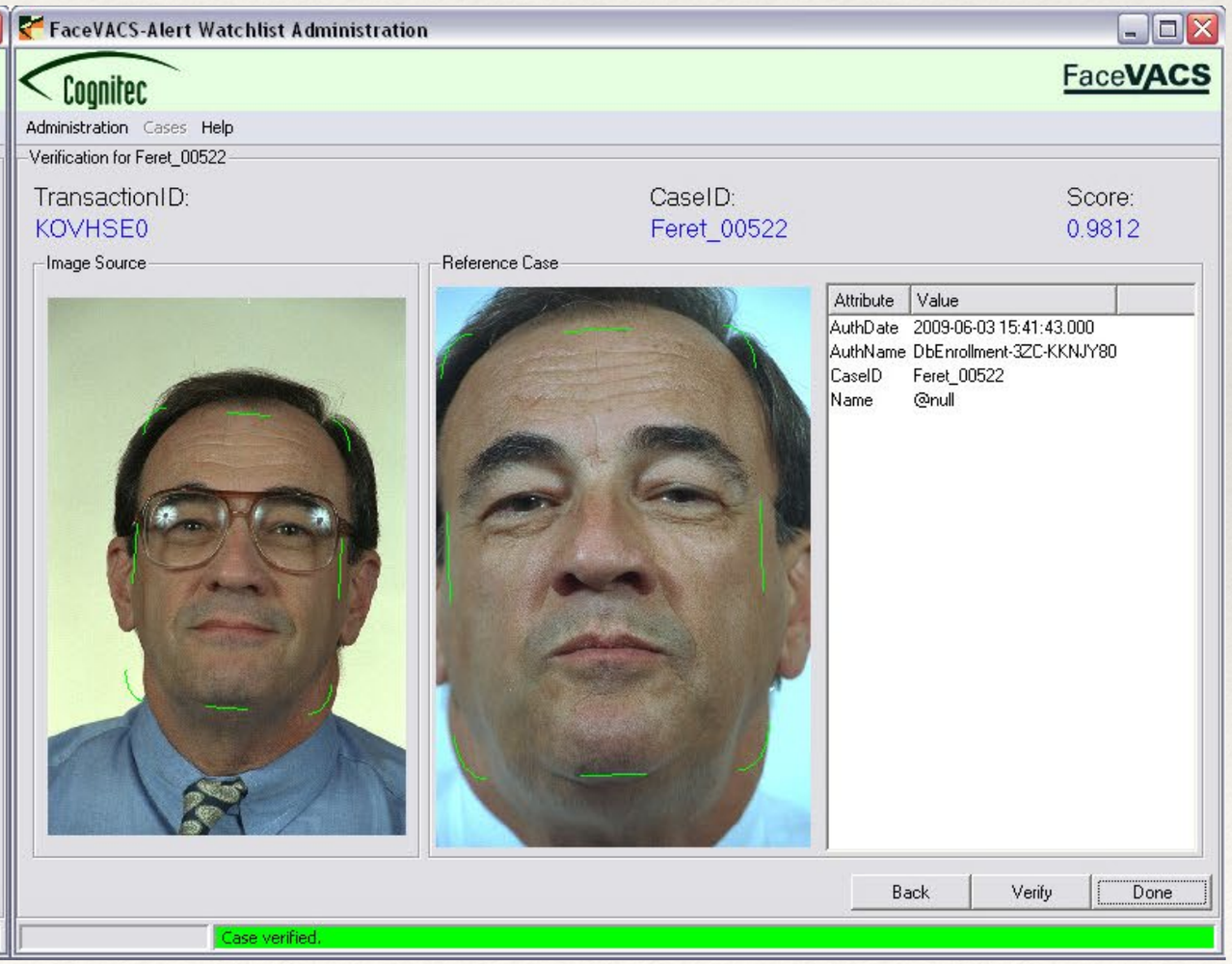
Head Captured when passing the tripwire



3D FaceReconstruction (Model based Cont'd)

- Face recognition works well with frontal face images only

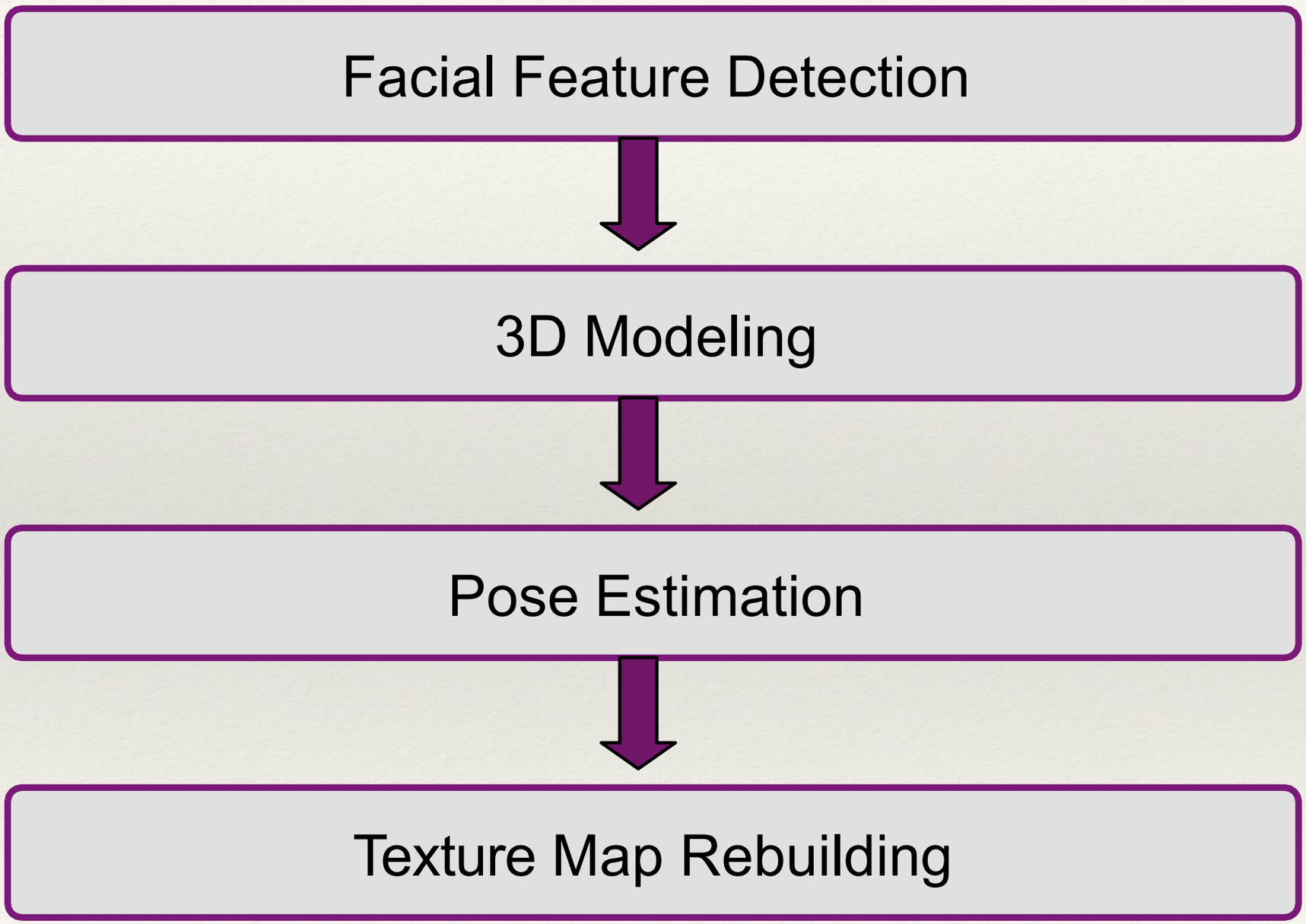




3D Face Reconstruction (Cont'd)

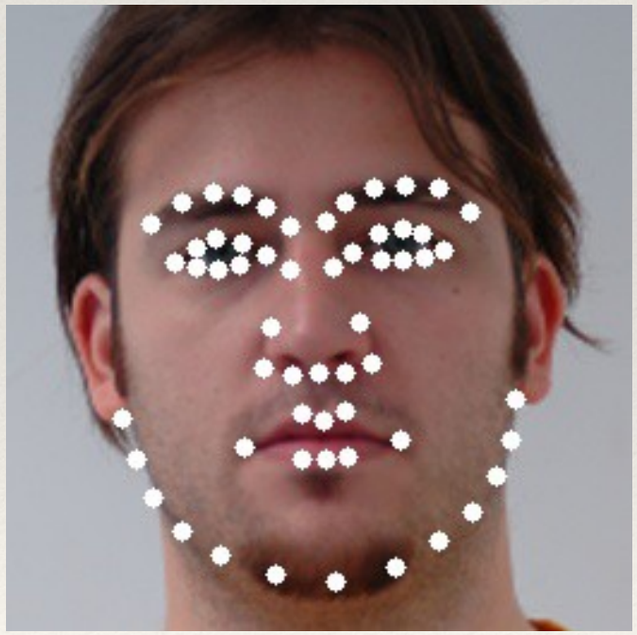
- But face recognition does not work well with profile face image

3D Face Reconstruction (Cont'd)

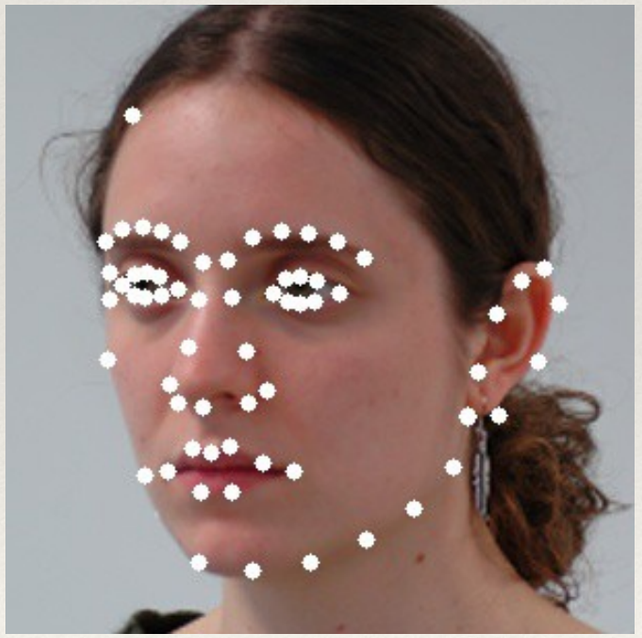


3D Face Reconstruction (Cont'd)

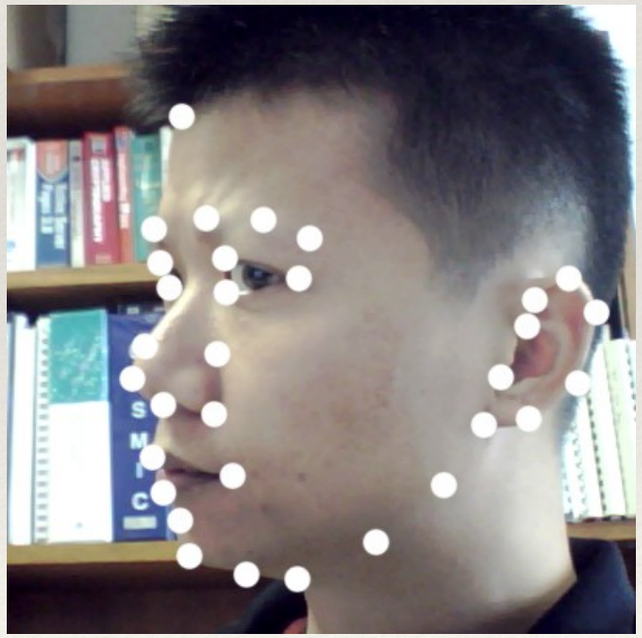
- Using active appearance model (AAM) to locate detailed facial features
- Using three distinct appearance models, which are frontal, half-profile and profile
- Individual models are then applied to match the input face image and search for the best fit. The one with minimum residual error is adopted



Frontal



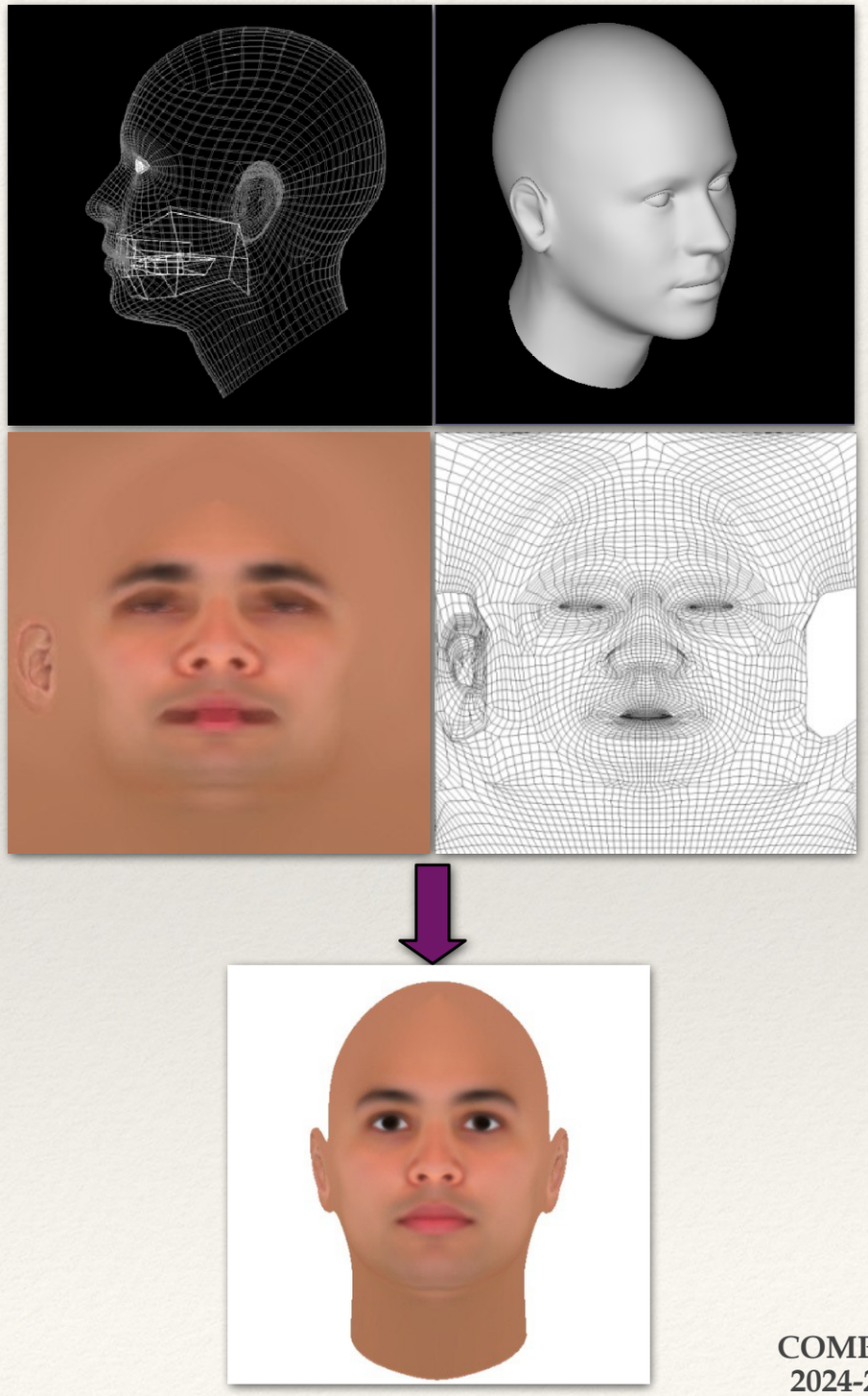
Half-Profile



Profile

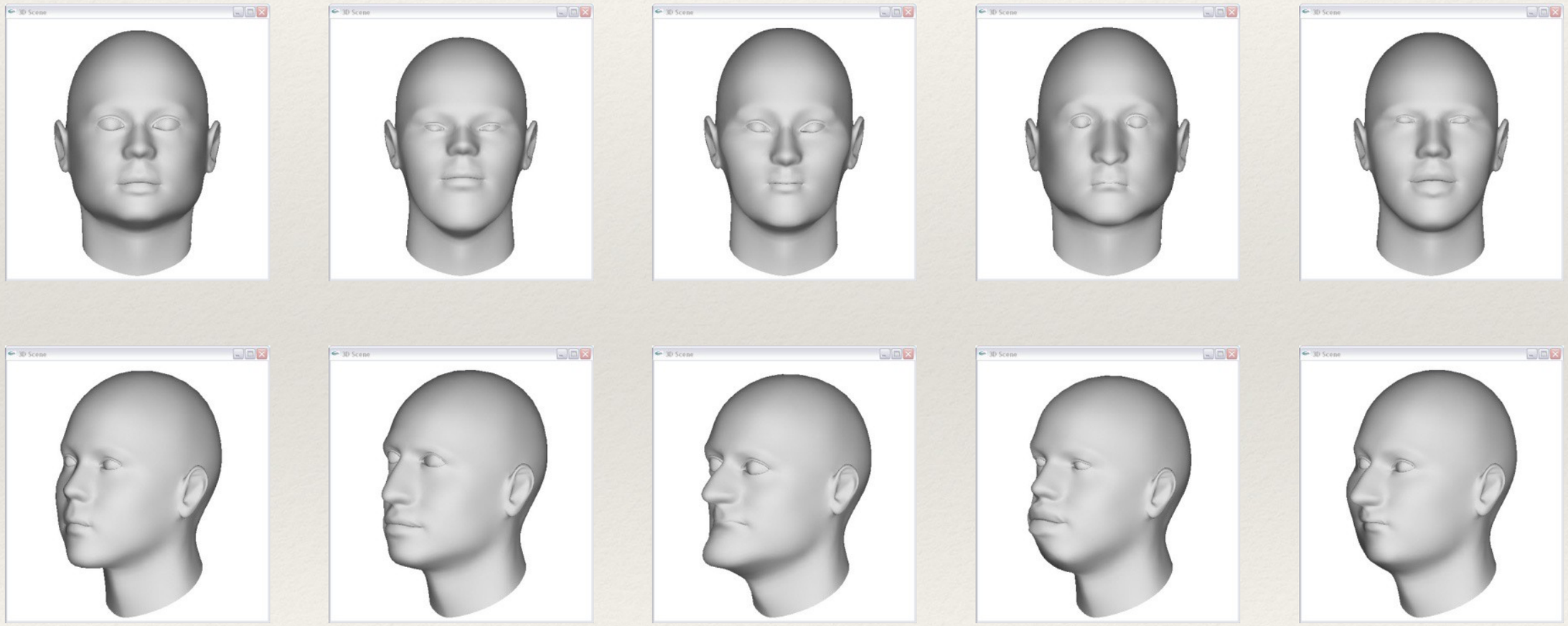
3D Face Reconstruction (Cont'd)

- Employed 3D face mesh has 6292 vertices and 6152 facet
- A texture map contains all the texture of a human face
- Texture coordinates determine how the texture is mapped from the 2D texture map onto the final 3D mesh



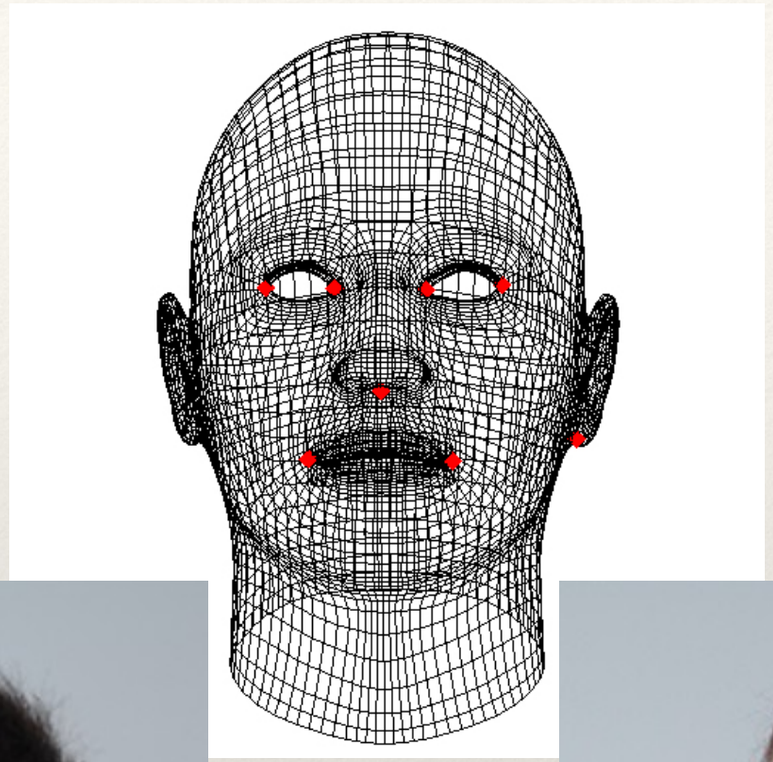
3D Face Reconstruction (Cont'd)

- Face Model Modification



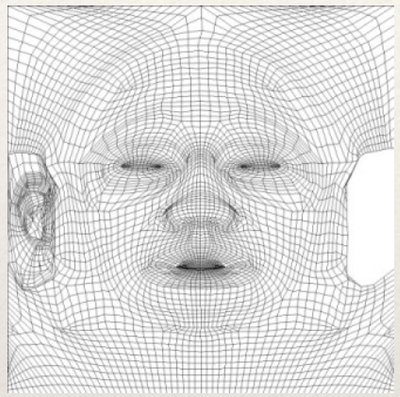
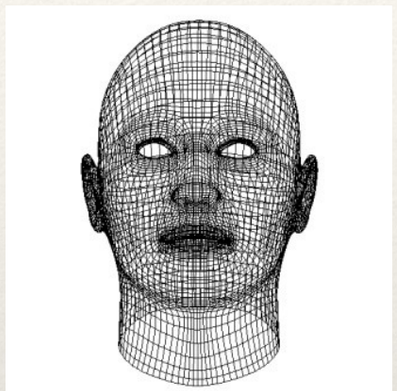
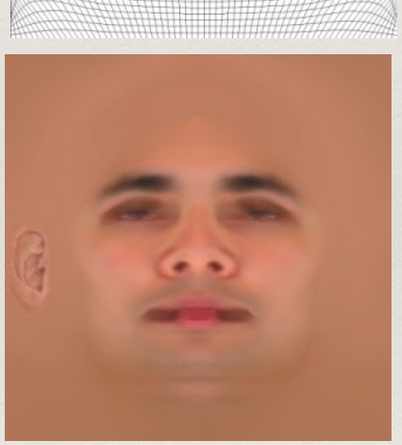
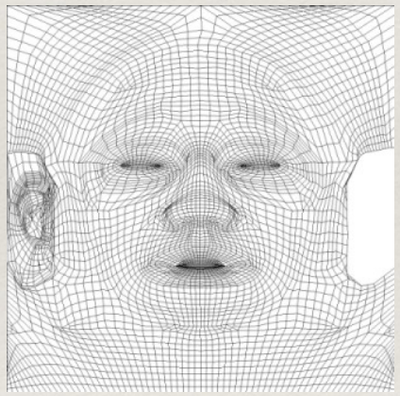
3D Face Reconstruction (Cont'd)

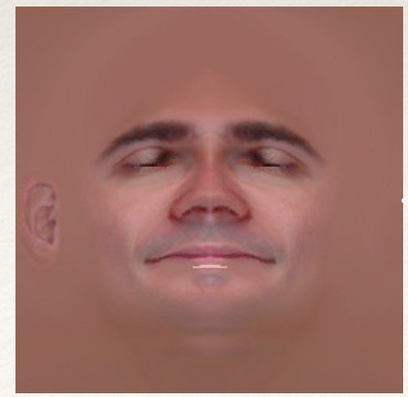
- Pose Estimation

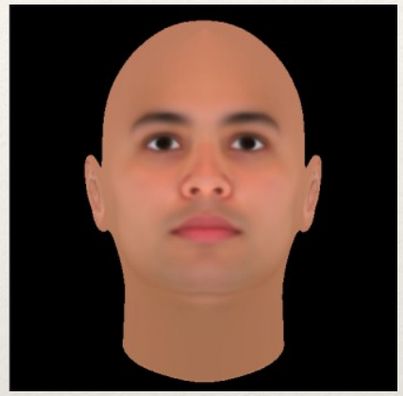


3D Face Reconstruction (Cont'd)

- Texture Map Rebuilding



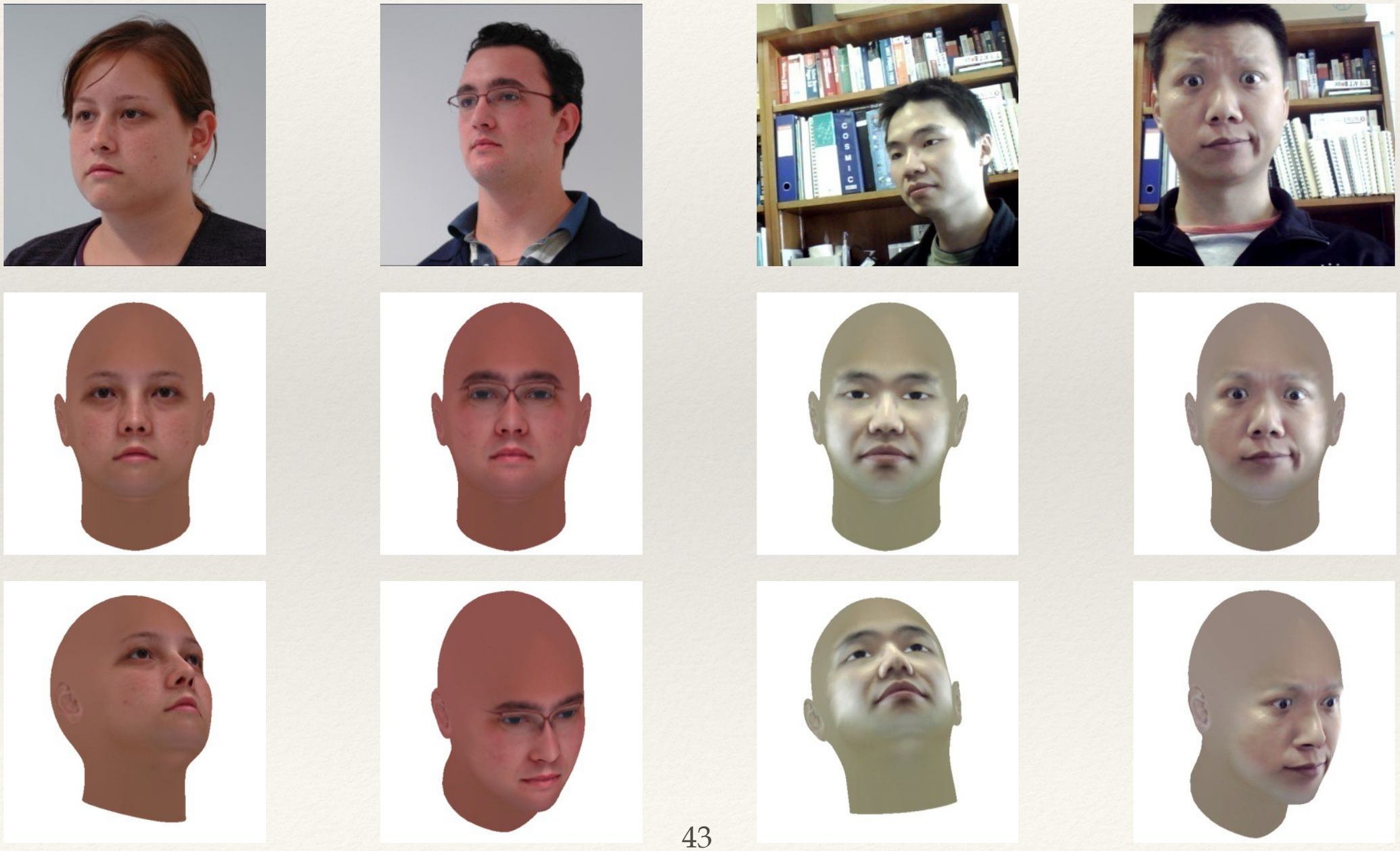
Rendering

Fill & Interpolation

Postprocessing

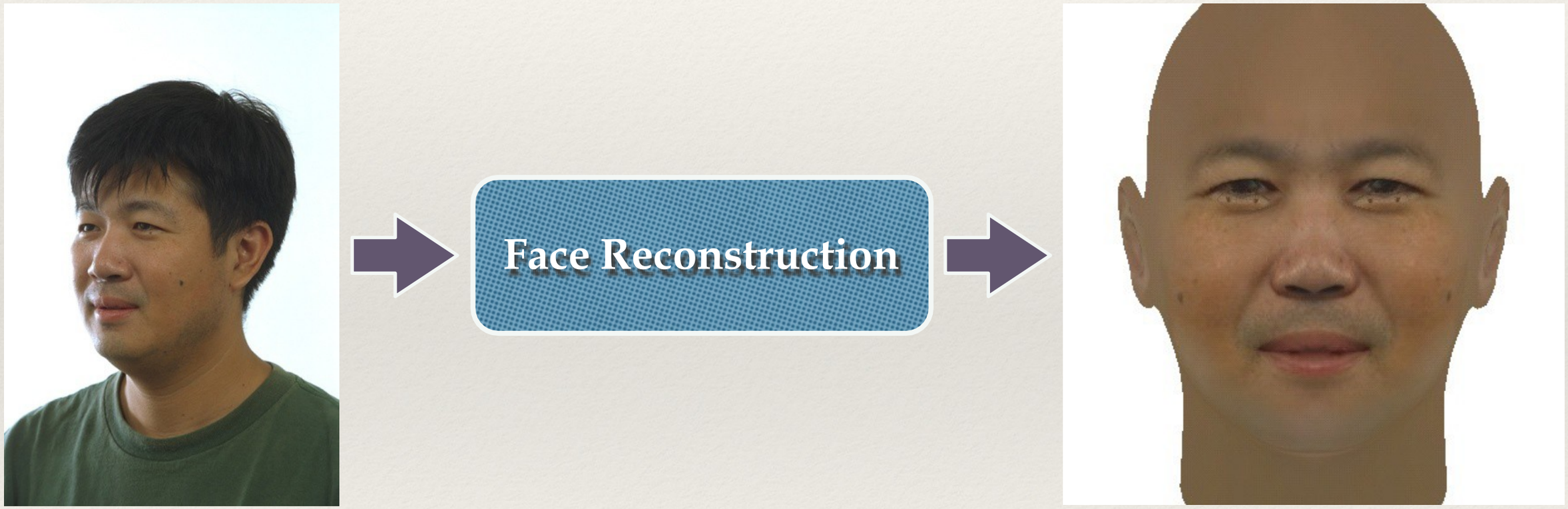
3D Face Reconstruction (Cont'd)

- Reconstruction Examples



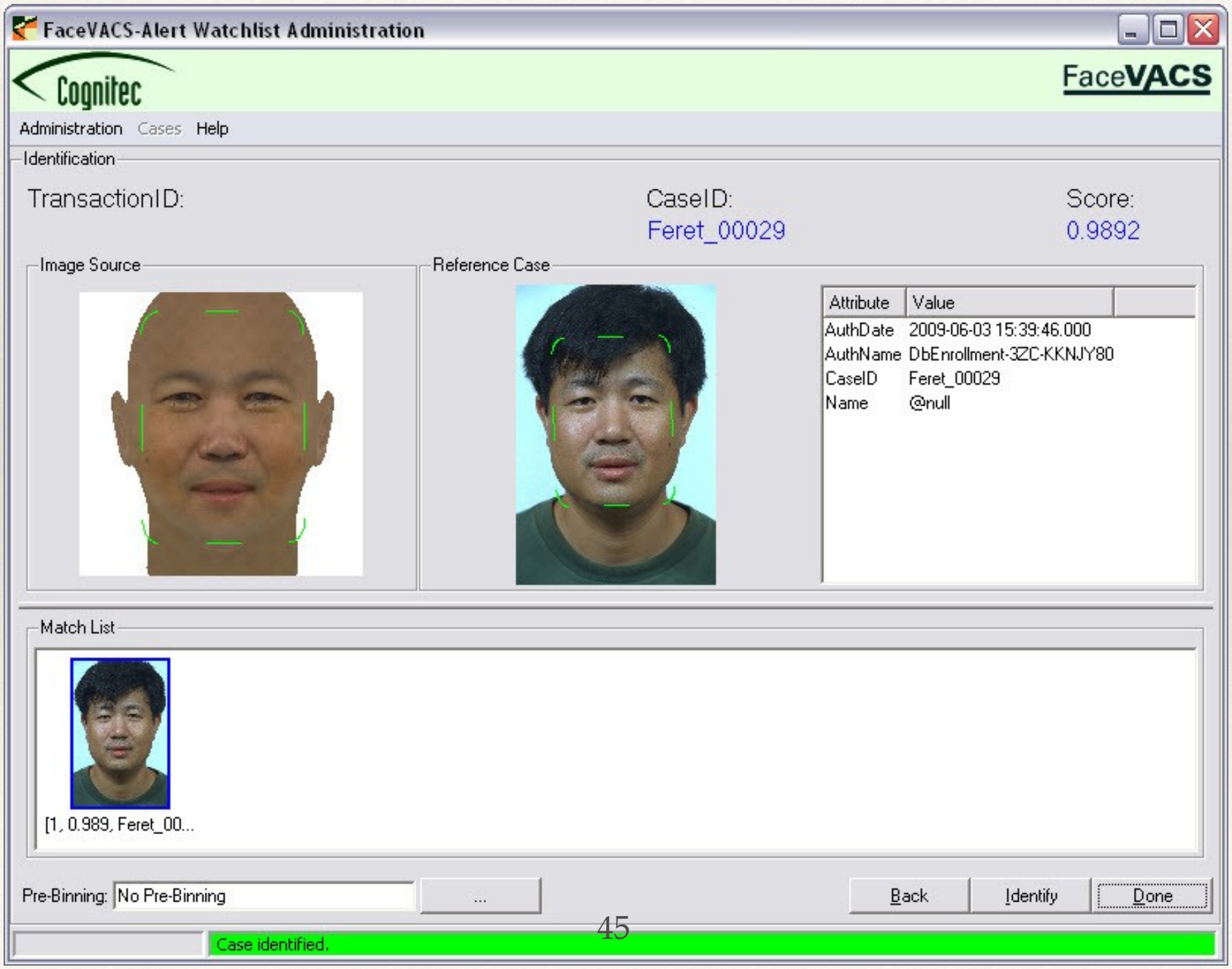
3D Face Reconstruction(Cont'd)

- Transform non-frontal face to frontal face image



3D Face Reconstruction (Cont'd)

- **The profile face can be recognised properly after it is transformed to frontal face by Face Reconstruction**



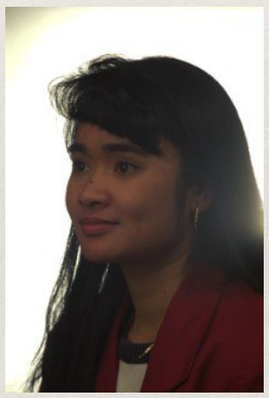
3D Face Reconstruction (Cont'd)

Gallery image with frontal faces




Non-frontal faces






Similarity to gallery image

0.7714 | 0.3341 | 0.1151 | 0.0000




Similarity to gallery image

0.8720 | 0.5469 | 0.6988 | 0.8474

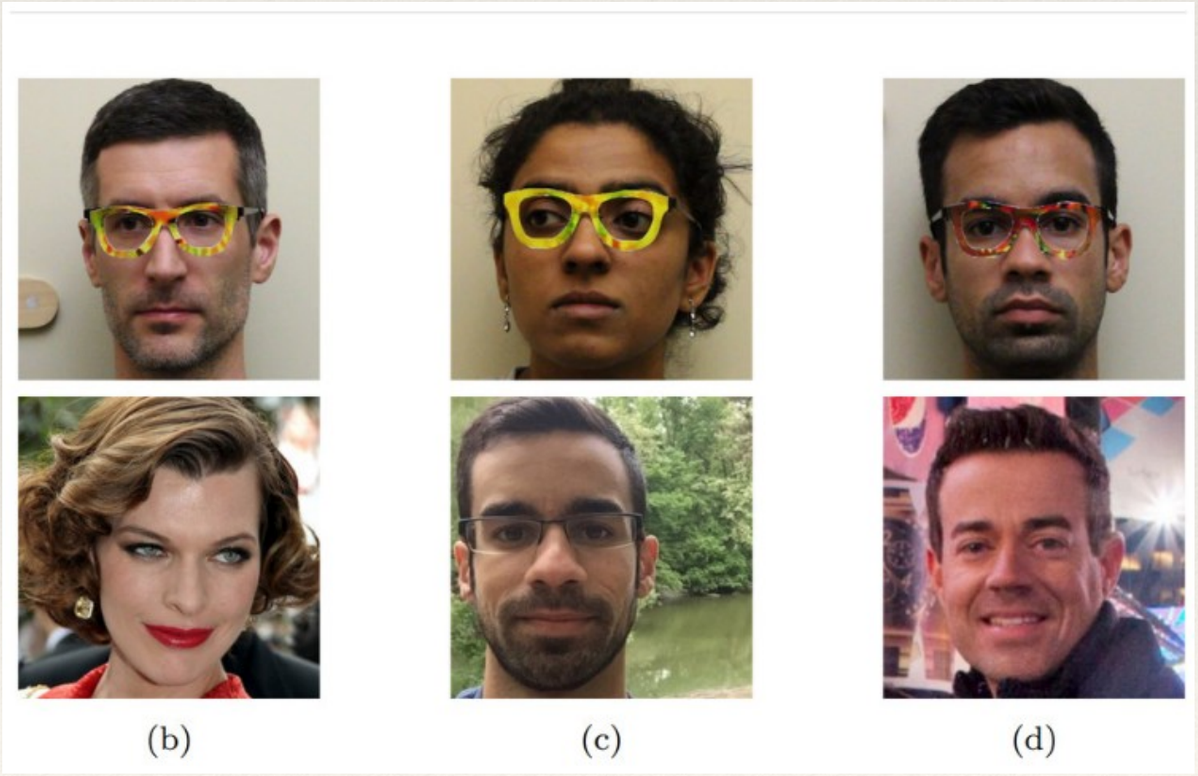
Face Recognition Limitation

- Researchers from Carnegie Mellon University have shown that specially designed spectacle frames can fool even state-of-the-art facial recognition software. Not only can the glasses make the wearer essentially disappear to such automated systems, it can even trick them into thinking you're someone else.
  - Reference: https://www.theverge.com/2016/11/3/13507542/facial-recognition-glasses-trick-impersonate-fool

These glasses trick facial recognition software into thinking you're someone else

by James Vincent | @jjvincent | Nov 3, 2016, 11:04am EDT

SHARE TWEET LINKEDIN

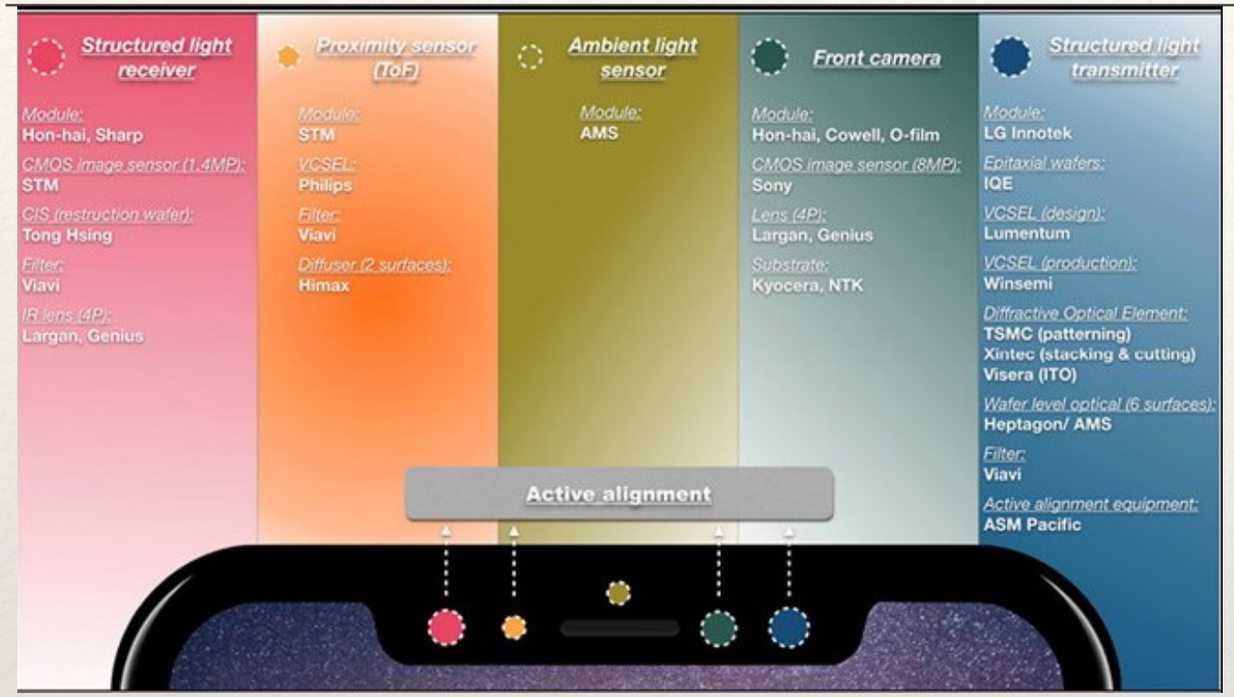


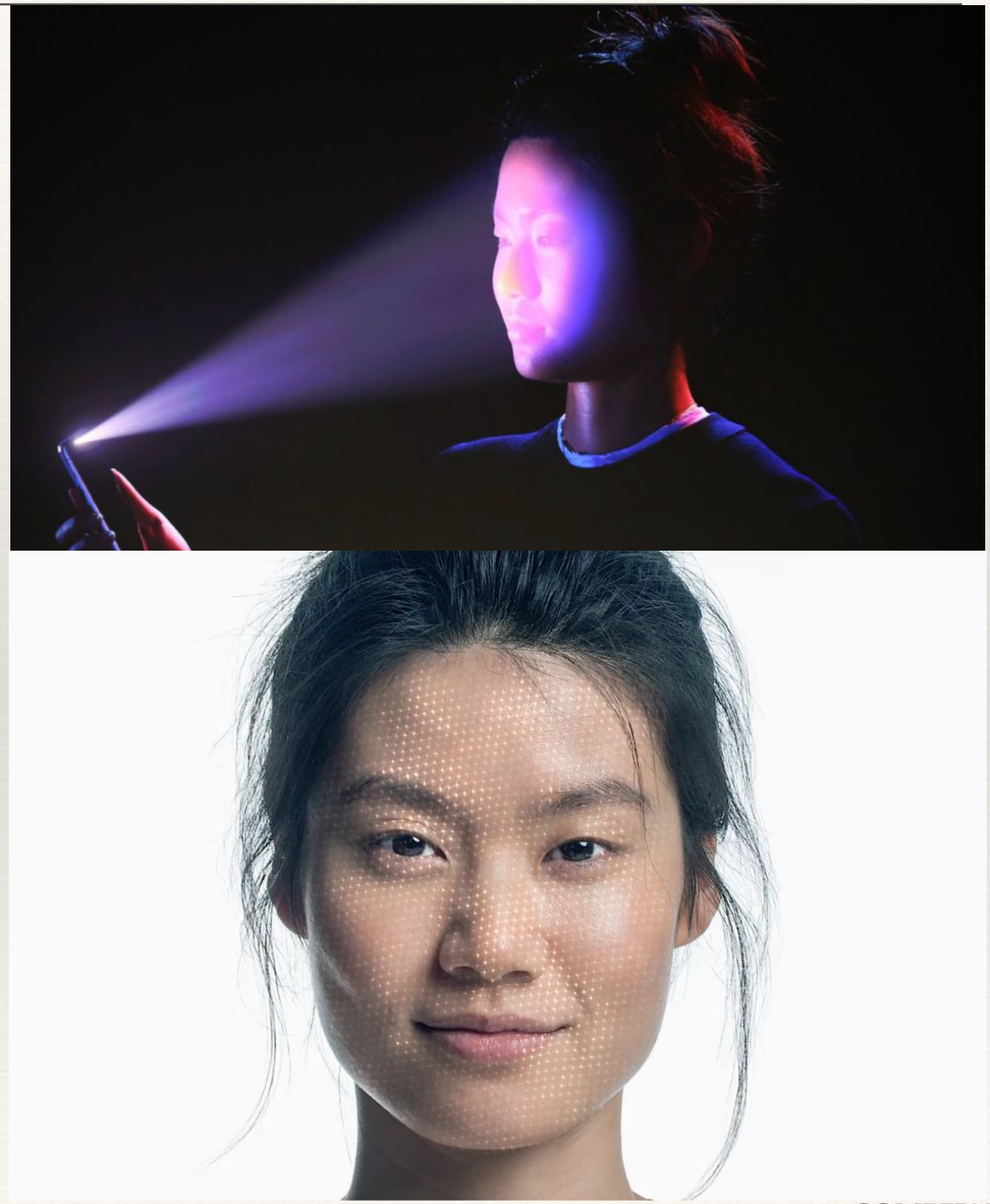
(b) (c) (d)

Face Recognition Limitation

- Researchers at the University of York showed, even the smallest change to someone's facial appearance, like wearing glasses, can shift our ability to identify them if we don't know them
  - Reference: http://www.cnn.com/2016/08/31/health/superman-glasses-disguise-facial-recognition/index.html

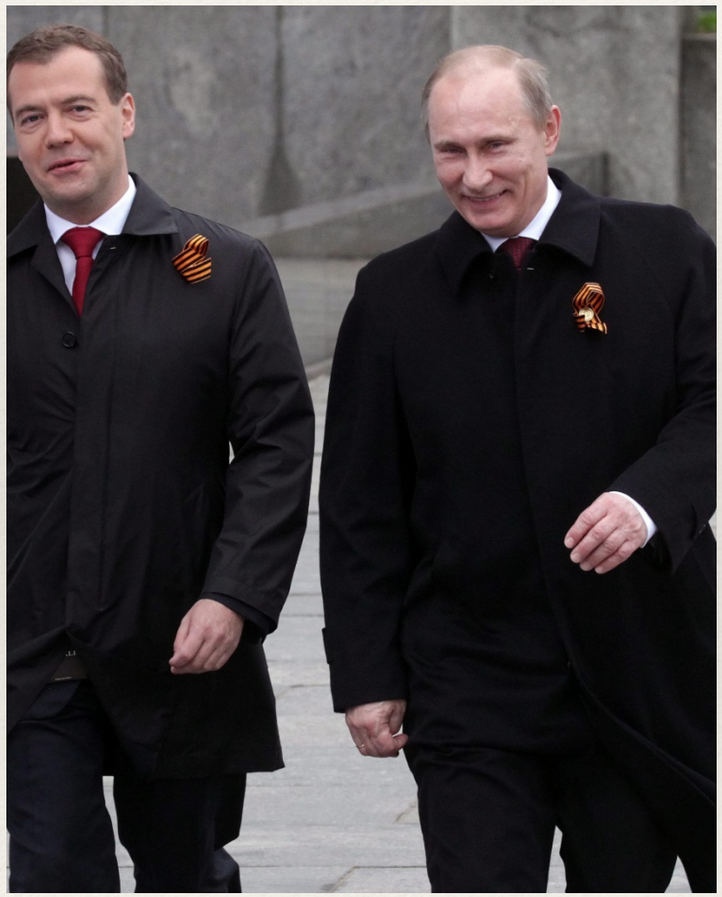
iPhone X~14 FaceID





Human Gait Recognition

- Gait is one of the most important biometrics.

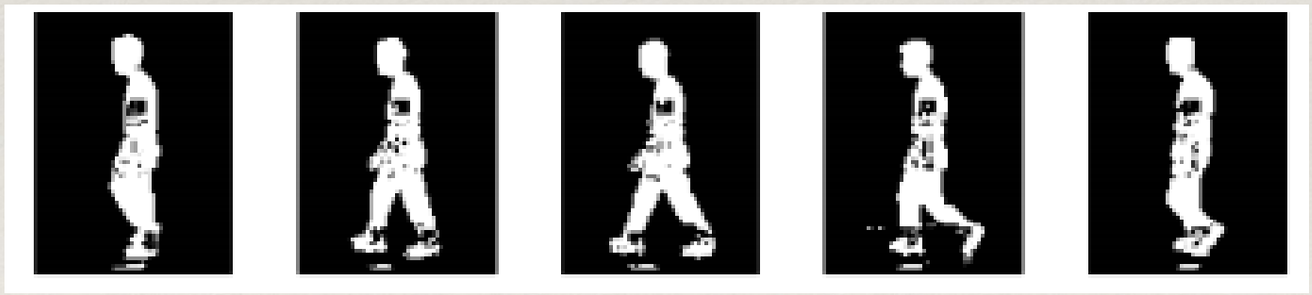


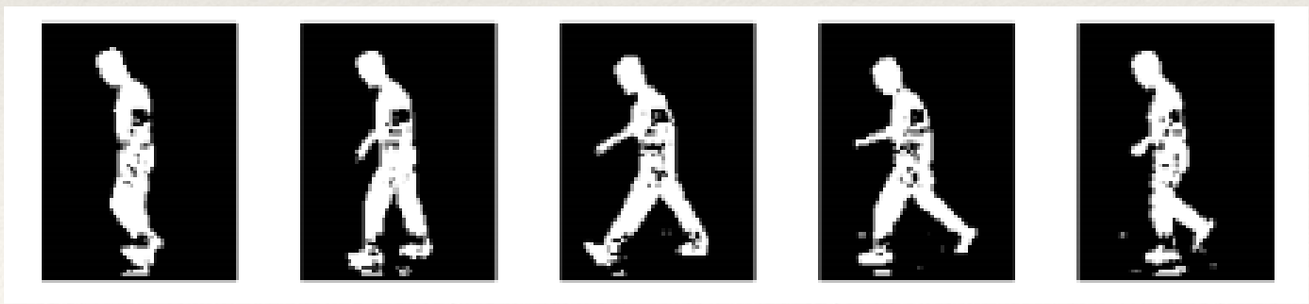
KGB trained gun style

https://www.theguardian.com/world/shortcuts/2015/dec/16/walk-like-the-kgb-get-vladamir-putins-gunslinger-gait

Human Gait Recognition (Cont'd)

- Taking gait as a biometric offers potential for identifying human at low resolution when the subject only consists of few image pixels





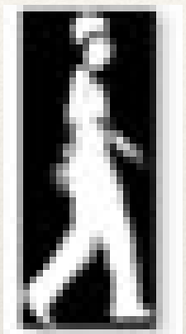
Human Gait Recognition (Cont'd)

- Sample of code words

Human Gait Recognition (Cont'd)

- Bag-of-Gait Representation

Input Silhouette

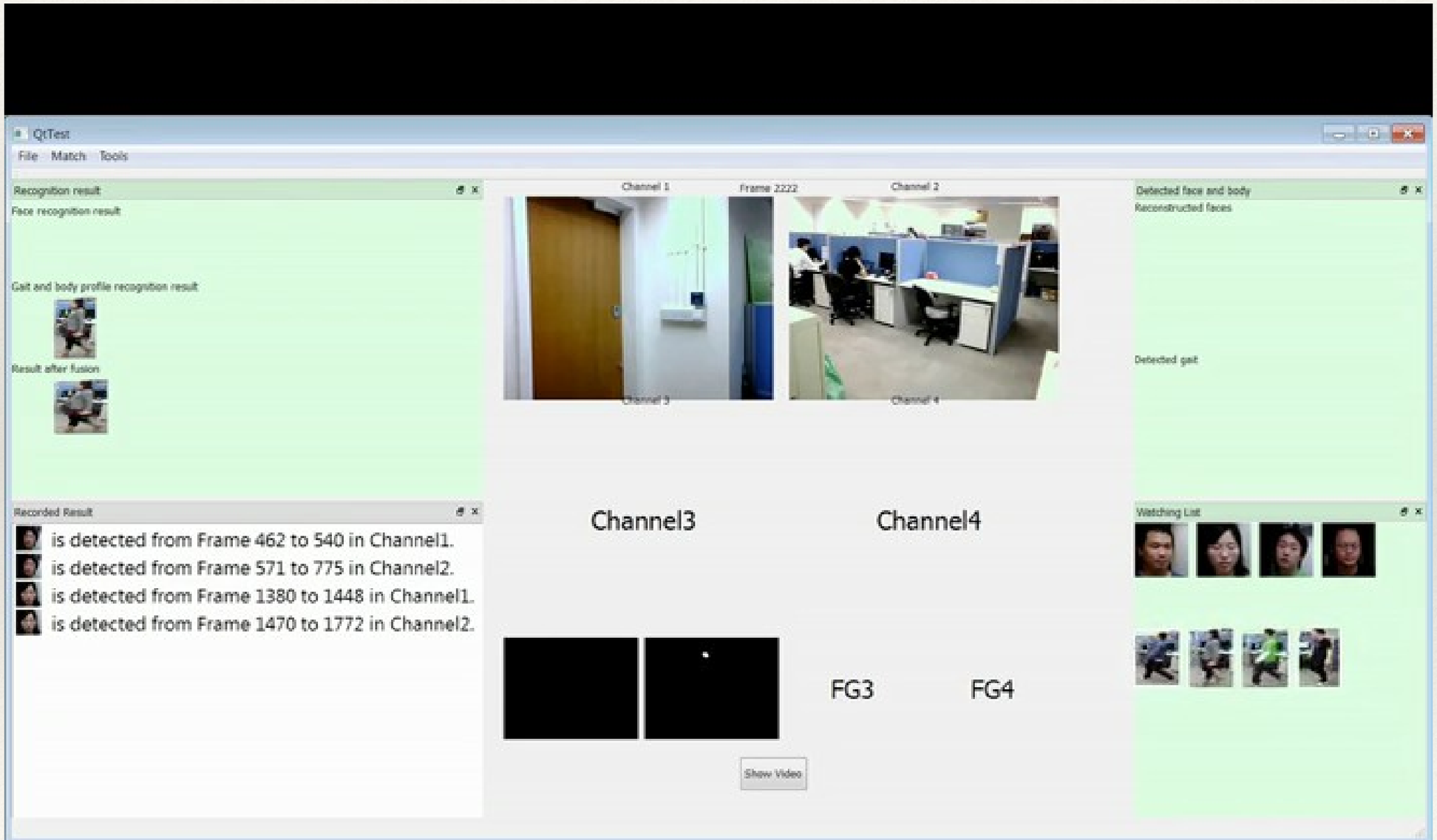


Feature vector

010000

Human Gait Recognition (Cont'd)

- Gait Tracking Demo

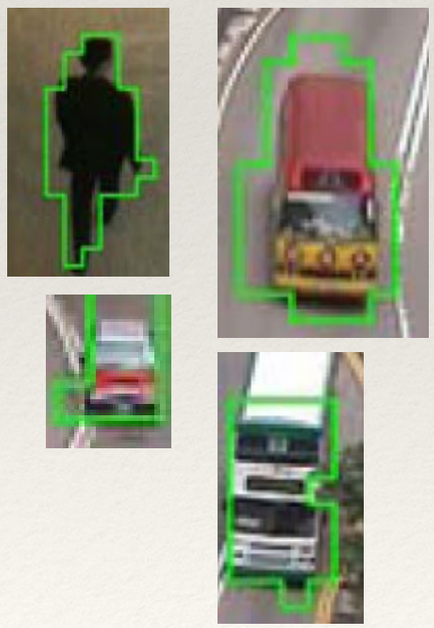


Video Retrieval

- Overview

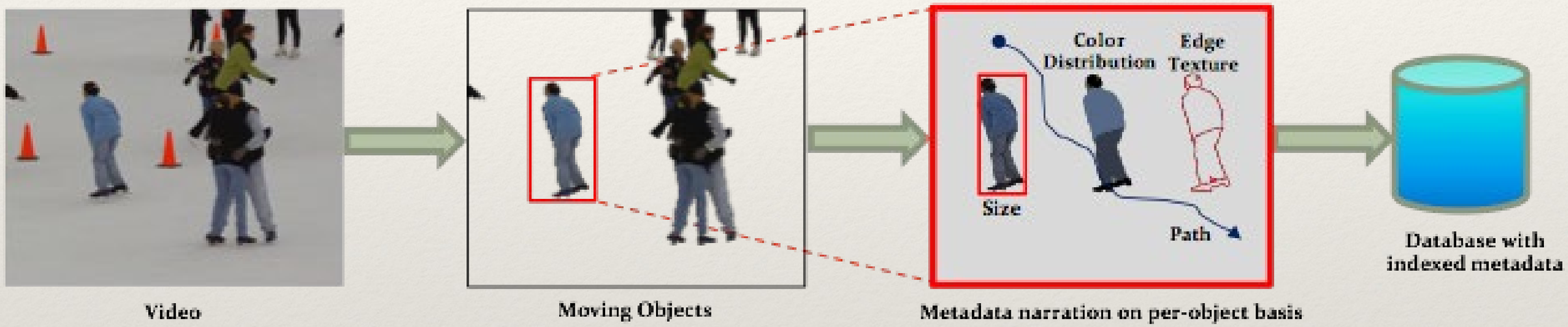
Video Source → H.264/MPEG-4 Encoder → Compressed Video Streams

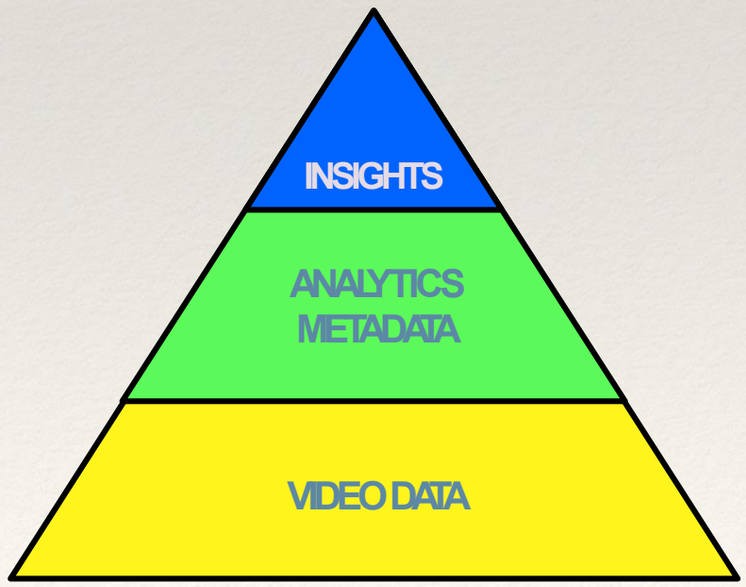
Motion Vectors and Background Model → Moving Object Segmentation → Tracking → Feature Extraction For Each Tracked Object → DB ← Retrieval Request



Video Retrieval (Cont'd)

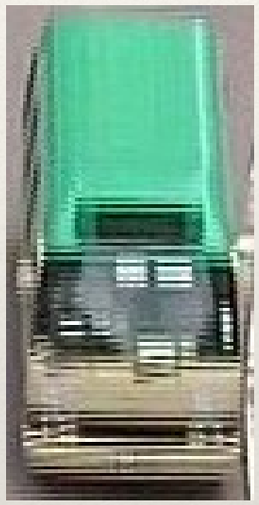
- Video Indexing

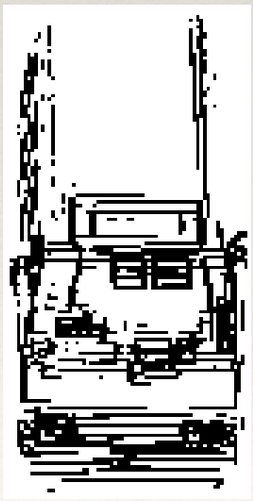


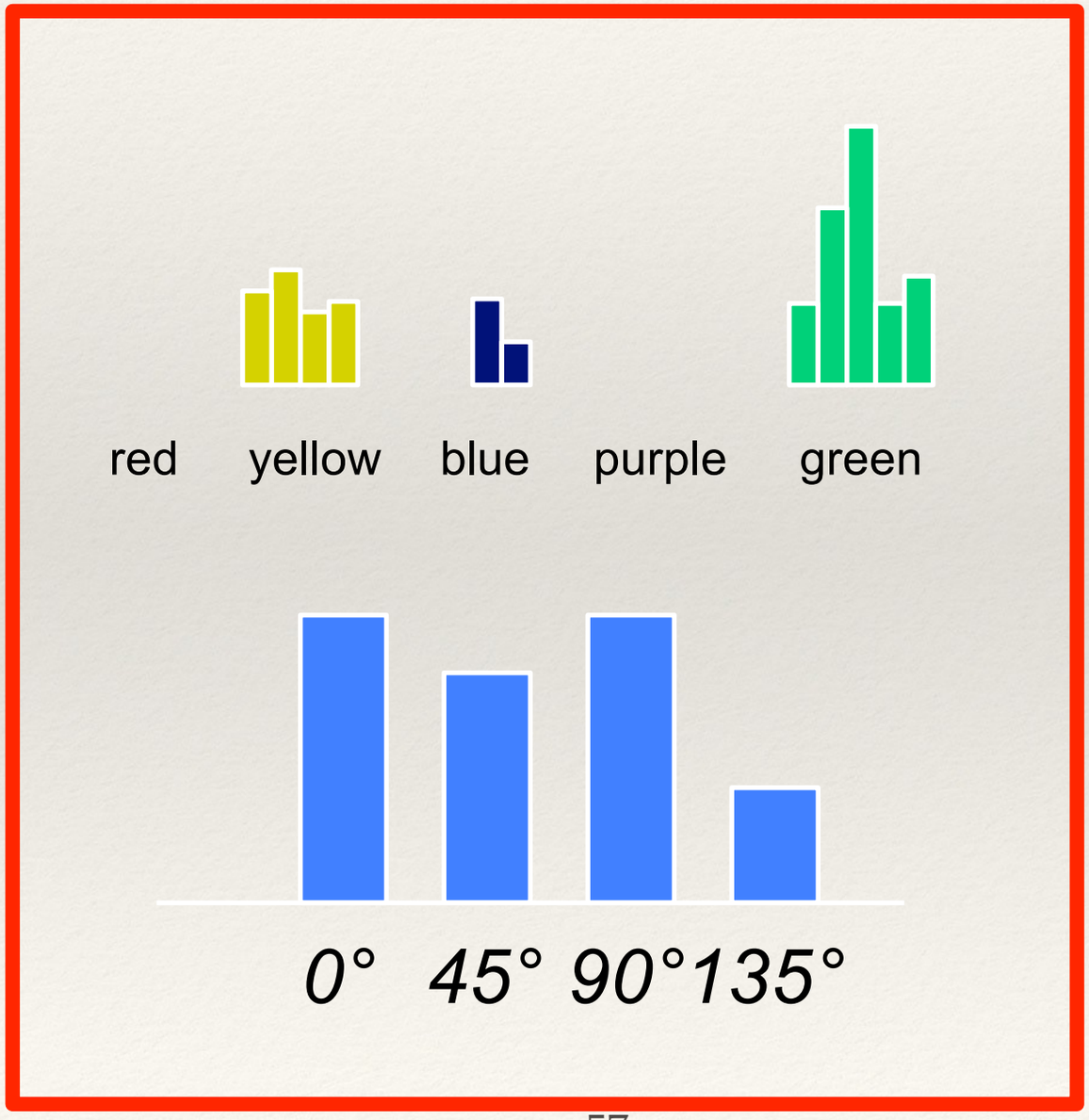


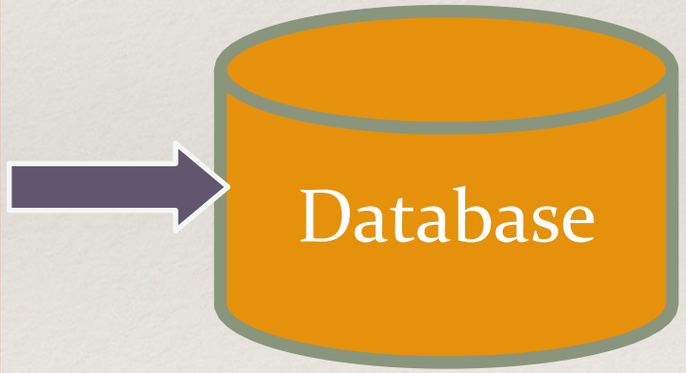
Video Retrieval (Cont'd)

- Color and Edge Features



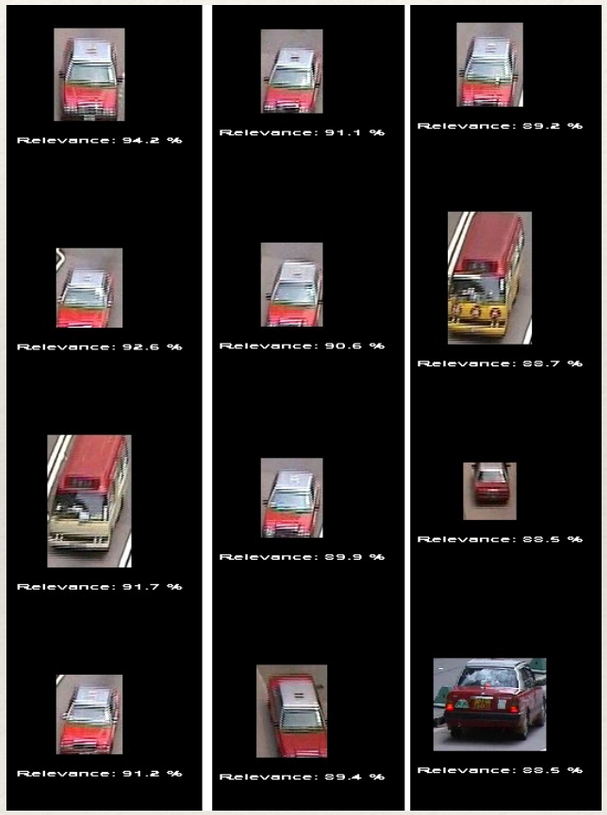


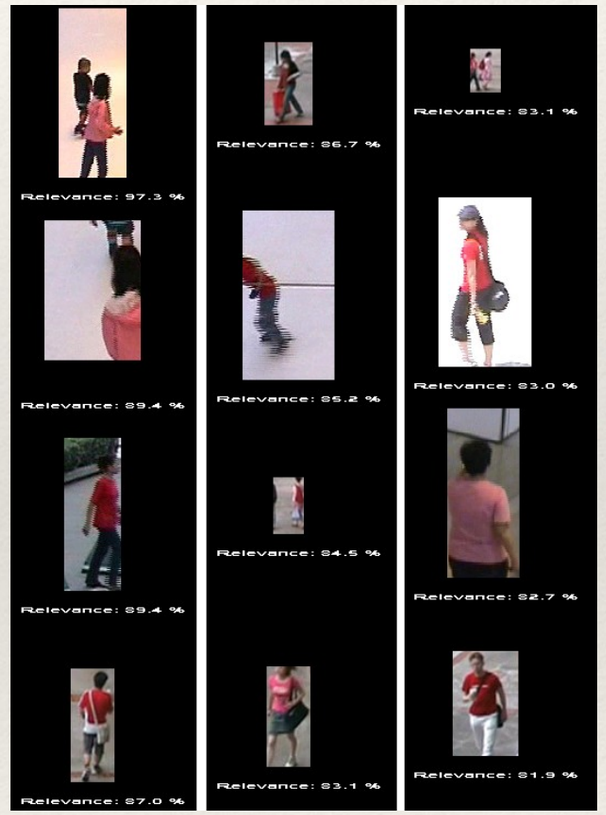


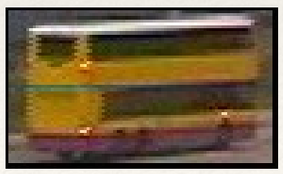


Video Retrieval (Cont'd)

- Retrieval Examples

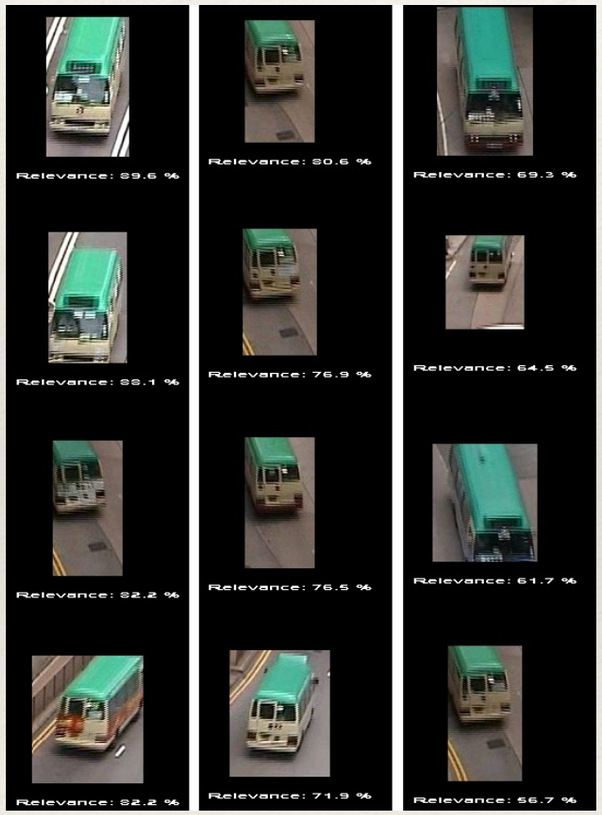


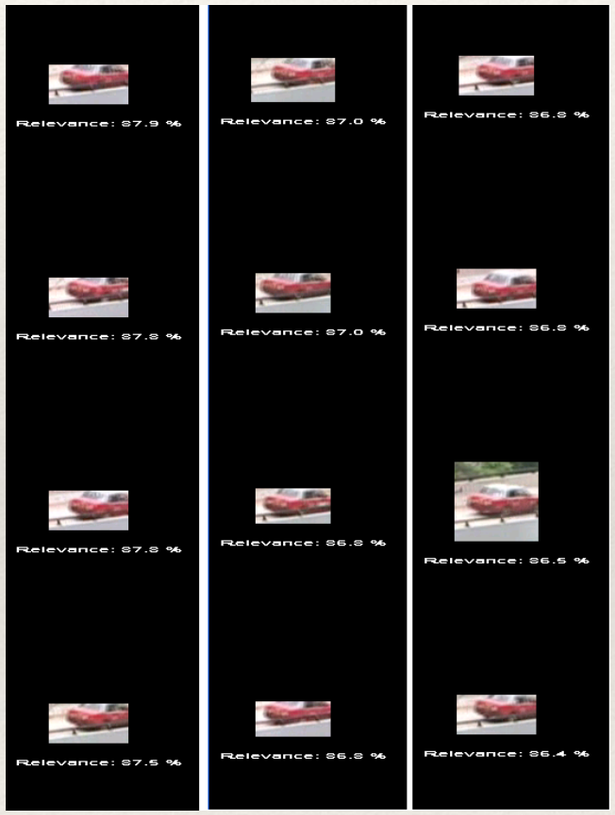


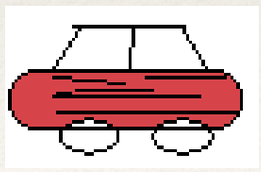


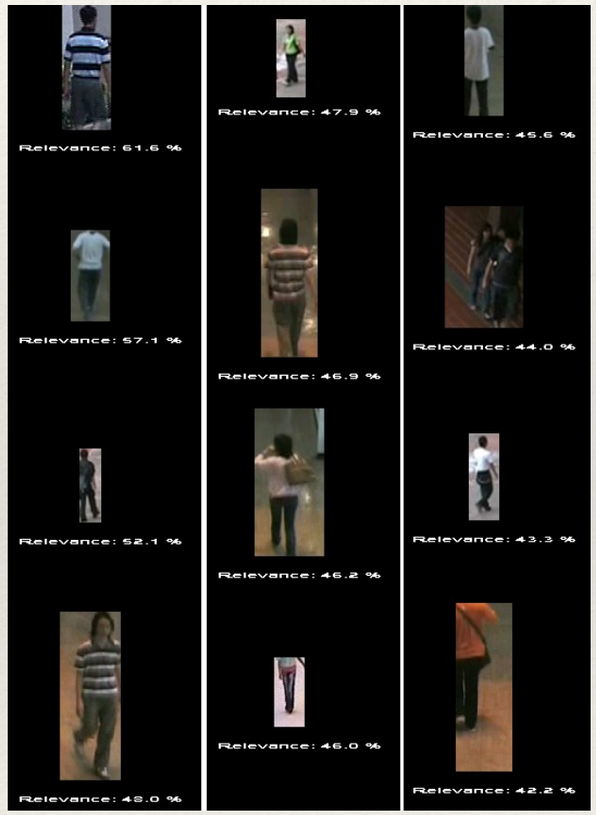
Video Retrieval (Cont'd)

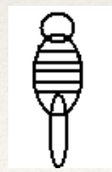
- Retrieval Examples











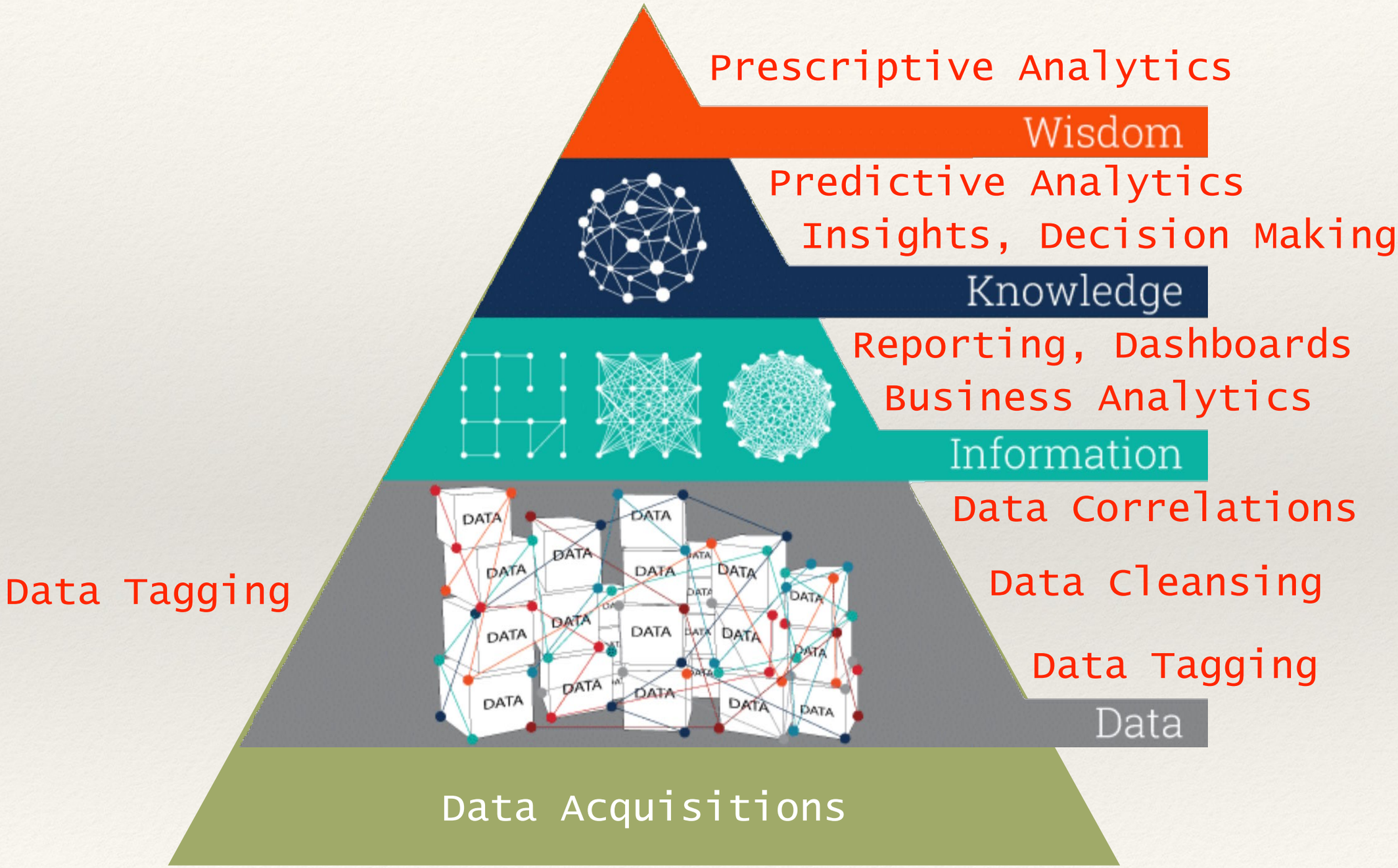
Dialogue in Metaverse

Lex Fridman

Mark Zuckerberg

Ref: https://www.youtube.com/watch?v=MVYrJJNdrEg

Data, Information, Knowledge, Wisdom (DIKW) Pyramid



Ref: https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/

Analytics in Daily Lives

Ref: https://www.caterpillar.com/en/company/innovation/customer-solutions/data-analytics.html

Smart City

https://www.pcgamesn.com/minecraft/super-mario