# INTRODUCTION TO REAL TIME AUDIO-VISUAL SENSING SYSTEMS

**Dr TIAN Jing**

**tianjing@nus.edu.sg**

# Module objective

Module: Introduction to audio and video sensing systems

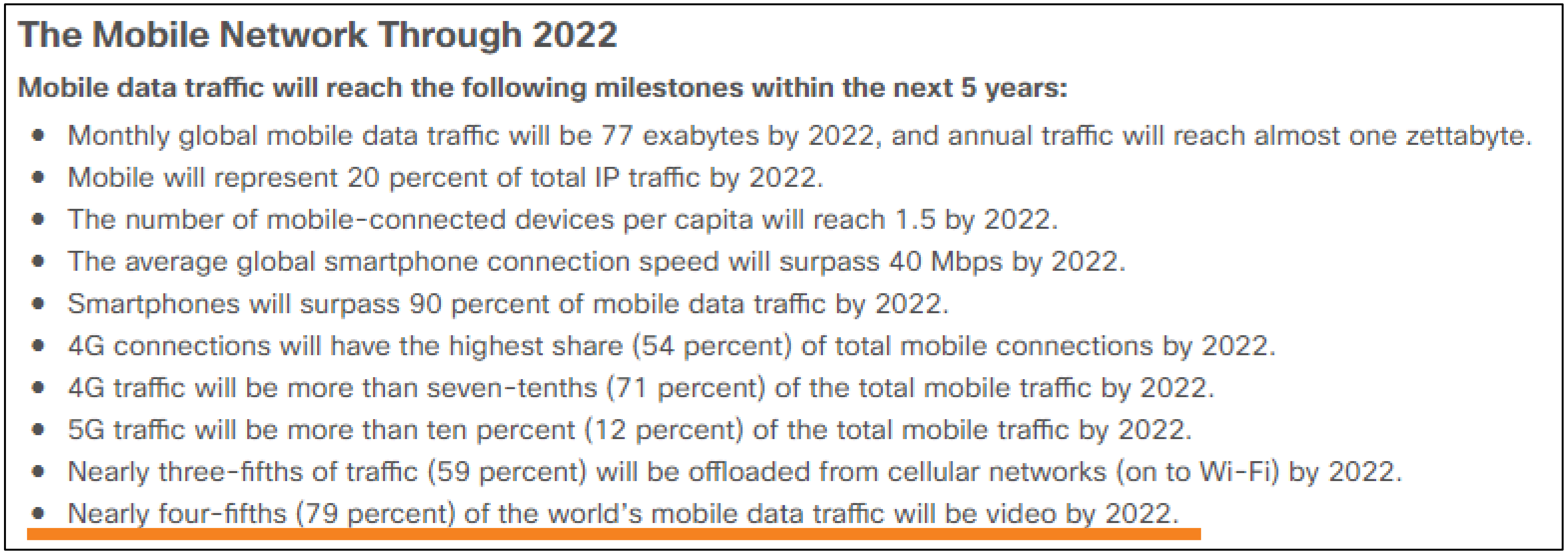
Knowledge and understanding

- Overview of audio-visual processing concepts
- Business applications of audio-visual sensing technology and sense making methods

Key skills

- Identify needs and challenges of audio-visual sensing technology in various industrial applications

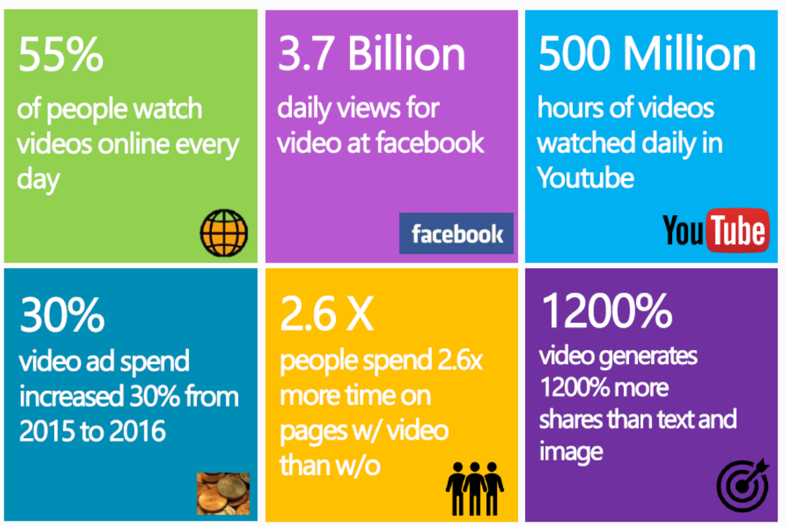
# Audio and video data



### The Mobile Network Through 2022

**Mobile data traffic will reach the following milestones within the next 5 years:**

- Monthly global mobile data traffic will be 77 exabytes by 2022, and annual traffic will reach almost one zettabyte.
- Mobile will represent 20 percent of total IP traffic by 2022.
- The number of mobile-connected devices per capita will reach 1.5 by 2022.
- The average global smartphone connection speed will surpass 40 Mbps by 2022.
- Smartphones will surpass 90 percent of mobile data traffic by 2022.
- 4G connections will have the highest share (54 percent) of total mobile connections by 2022.
- 4G traffic will be more than seven-tenths (71 percent) of the total mobile traffic by 2022.
- 5G traffic will be more than ten percent (12 percent) of the total mobile traffic by 2022.
- Nearly three-fifths of traffic (59 percent) will be offloaded from cellular networks (on to Wi-Fi) by 2022.
- Nearly four-fifths (79 percent) of the world's mobile data traffic will be video by 2022.



55% of people watch videos online every day

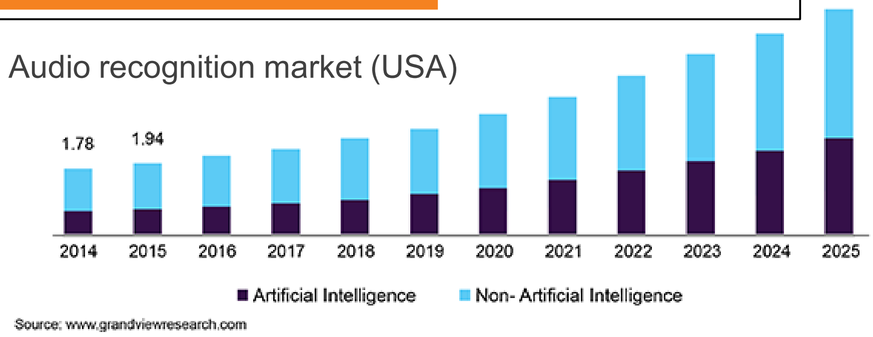
3.7 Billion daily views for video at facebook

500 Million hours of videos watched daily in Youtube

30% video ad spend increased 30% from 2015 to 2016

2.6 X people spend 2.6x more time on pages w/ video than w/o

1200% video generates 1200% more shares than text and image



Audio recognition market (USA)

Source: www.grandviewresearch.com

There is a camera installed for every 29 people on the planet, and in developed nations, the number rises to a camera for every 8 people.

Reference: https://www.forbes.com/sites/miketempleman/2017/09/06/17-stats-about-video-marketing; Voice and Speech Recognition Market Report, https://www.grandviewresearch.com/industry-analysis/voice-recognition-market, https://www.computer.org/publications/tech-news/research/real-time-video-analytics-for-camera-surveillance-in-edge-computing

# Motivation: Security

- Changi Airport pilots a Multi-Signal Surveillance Platform which combines audio with video analytics to monitor security incidents.
- Reduce reliance on security manpower and reduce fatigue, patrolling tasks and operation costs in managing a site
- Increase response times of security officers on-site

Changi Airport to use audio, video analytics to monitor security

- Airport Immigration project in some Europe airports, where travellers are given an automated lie detection test.
- Questions such as “What is in your suitcase” are asked by a virtual agent.
- Micro-expressions are scored for each response. Travellers who failed the test will be referred to human assessors.

Passengers to face AI lie detector tests at EU airports

CNN travel

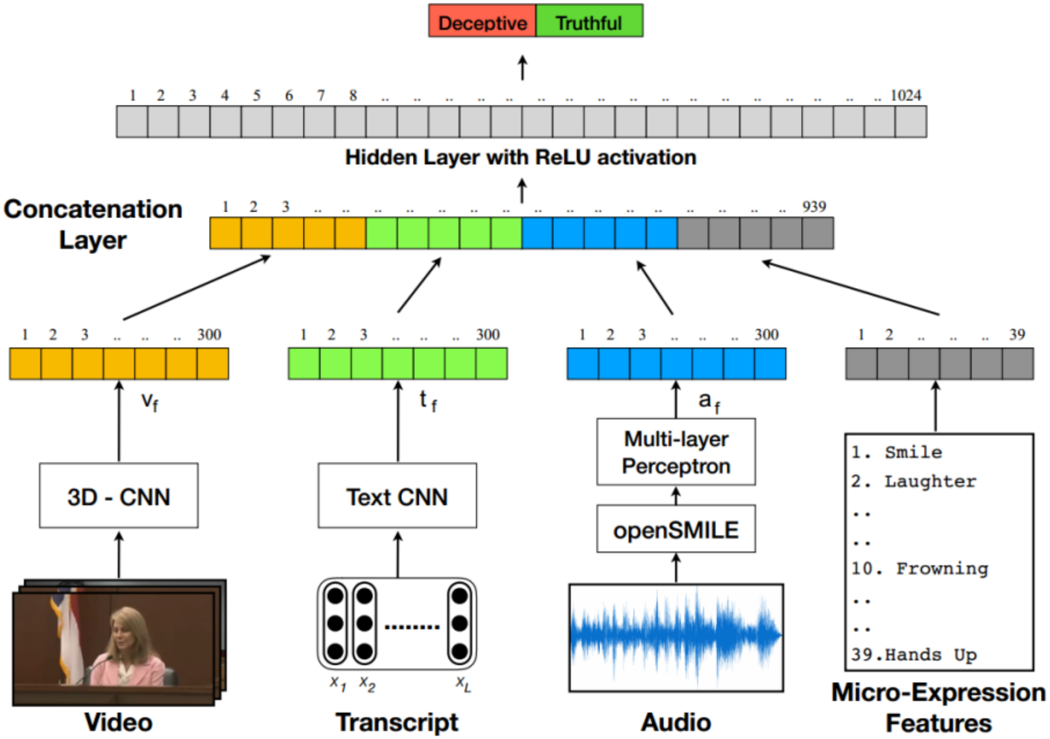
Rob Picheta, CNN • Updated 2nd November 2018

Reference:

- https://www.straitstimes.com/singapore/changi-airport-to-use-audio-video-analytics-to-monitor-security
- https://edition.cnn.com/travel/article/ai-lie-detector-eu-airports-scli-intl/index.html

# Motivation: Insurance

- A digitalized rating process where a facial recognition system is used to verify the identity of applicants automatically.
- Facial expressions of applicants are also analysed to determine their willingness to repay the loans.



Reference: https://www.scmp.com/business/banking-finance/article/2117469/big-data-could-help-bring-micro-lending-millions-left-out

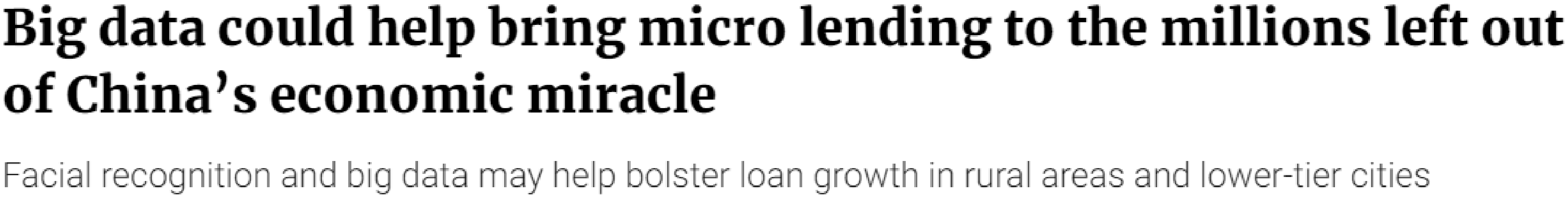
# Motivation: Multimodal deception detection

Title: Deception Detection using Real-life Trial Data [1]

**Introduced a Multimodal Deception Detection dataset with video recordings taken in court trials and 1-1 interviews**

- Videos
  - 61 deceptive, 60 truthful
  - Average length: 28 s
  - Subject profile: 21 unique female, 35 unique male
  - Age group: 16 – 60 years old
- Manually Transcribed speech
- Manually annotated micro-expressions & hand gestures

**Sample Video**

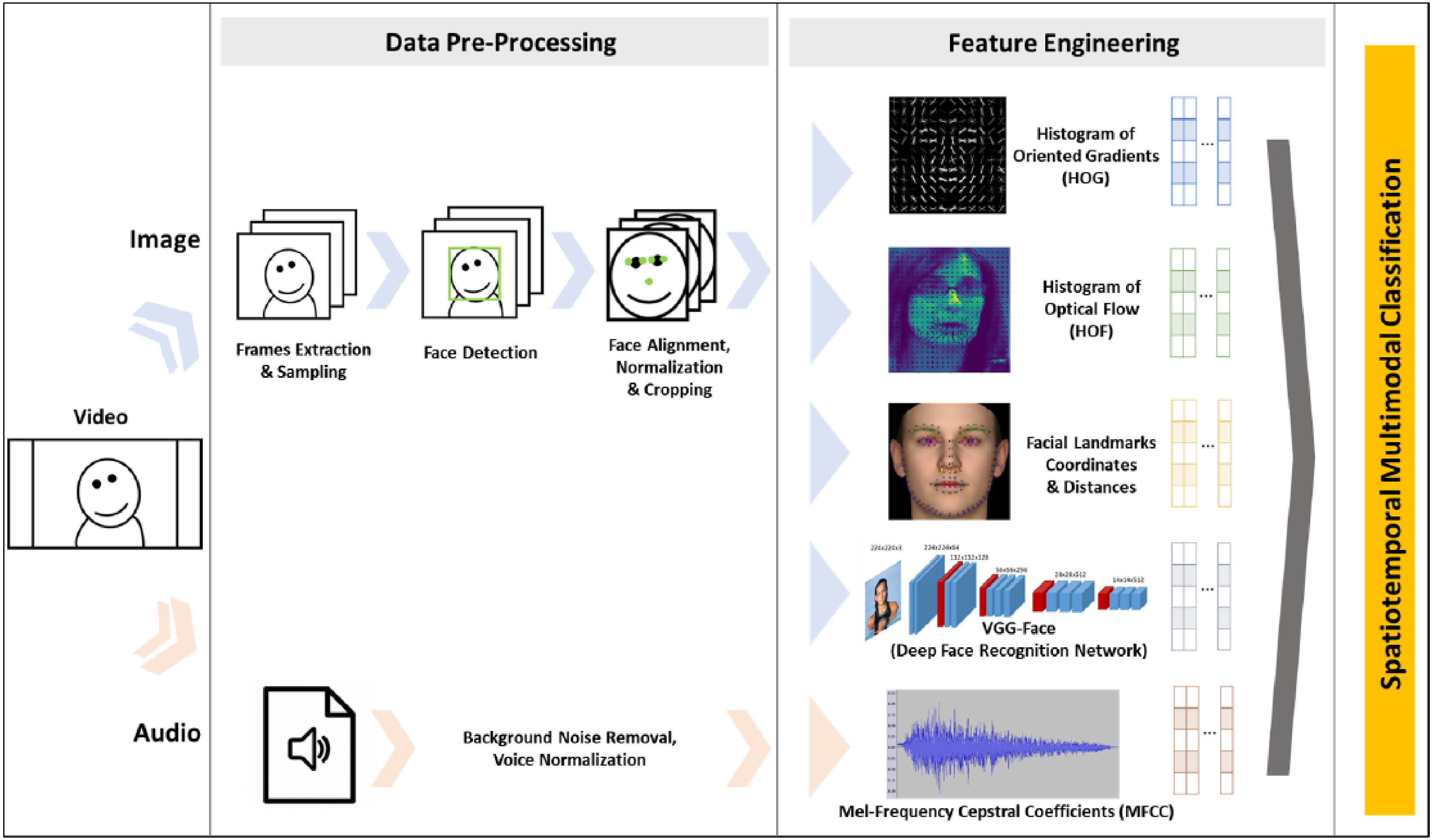


**Big data could help bring micro lending to the millions left out of China’s economic miracle**

Facial recognition and big data may help bolster loan growth in rural areas and lower-tier cities

Reference
- Photo: https://paperswithcode.com/task/deception-detection/codeless
- Deception detection, https://lit.eecs.umich.edu/deceptiondetection/
- https://www.scmp.com/business/banking-finance/article/2117469/big-data-could-help-bring-micro-lending-millions-left-out

# Motivation: Multimodal deception detection



Data Pre-Processing

Feature Engineering

Image

Video

Audio

Frames Extraction & Sampling

Face Detection

Face Alignment, Normalization & Cropping

Background Noise Removal, Voice Normalization

Histogram of Oriented Gradients (HOG)

Histogram of Optical Flow (HOF)

Facial Landmarks Coordinates & Distances

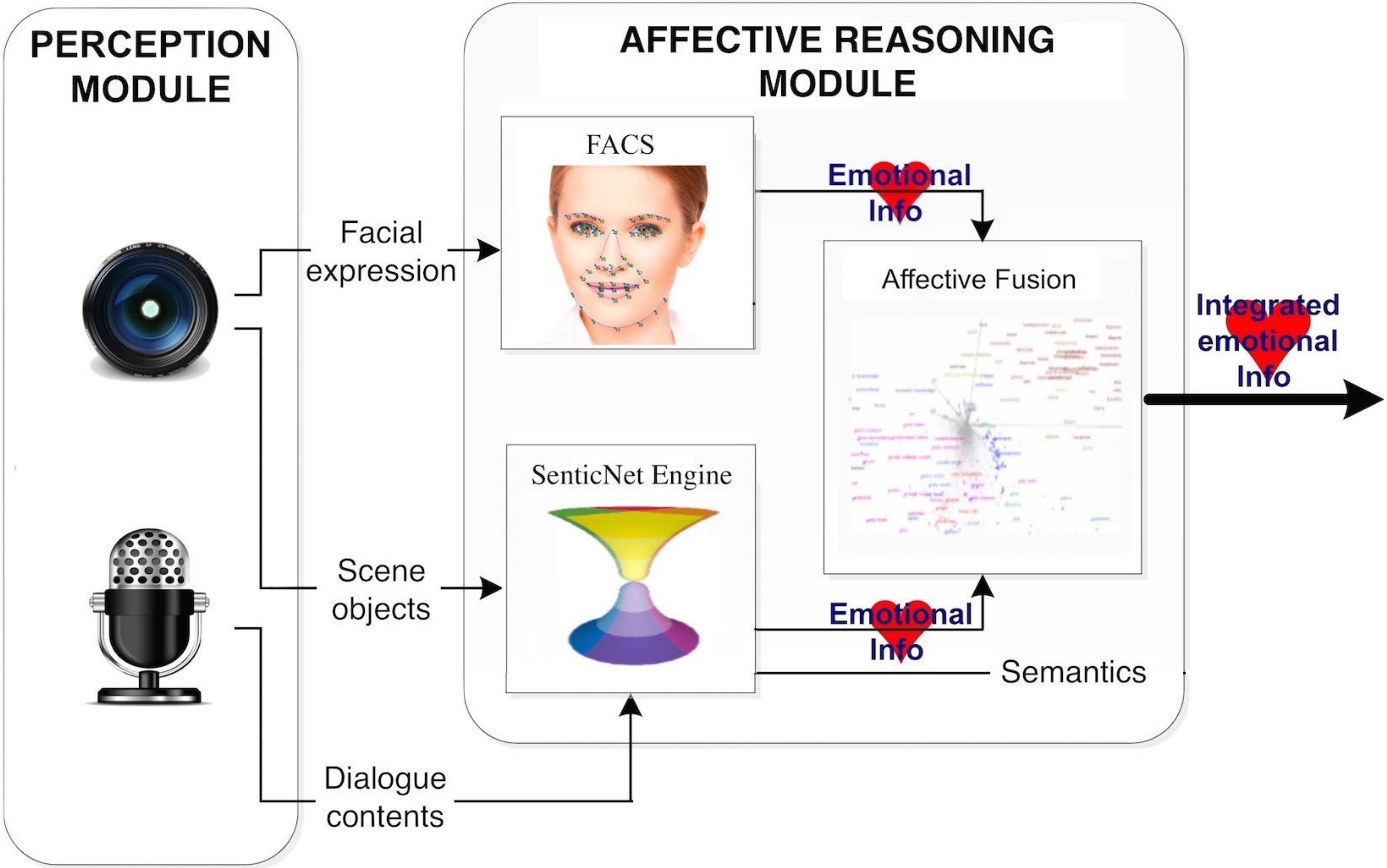
VGG-Face (Deep Face Recognition Network)

Mel-Frequency Cepstral Coefficients (MFCC)

Spatiotemporal Multimodal Classification

Reference: ISS student FYP project, KENNETH ANTHONY, KWEK GUANG JIE, BRYAN, TAN ZIYING ALYSA, TEH VIVIAN

# Motivation: Sentiment



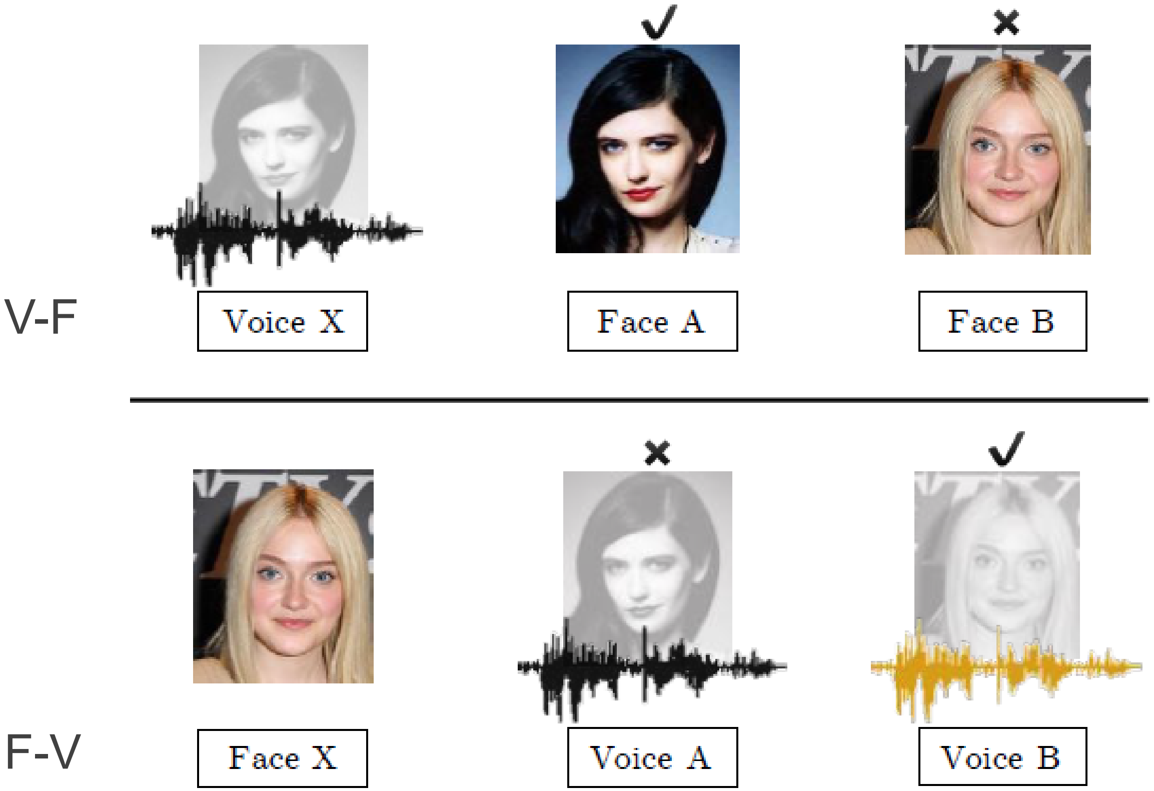
Reference: http://citic.ugr.es/pages/formacion/jornadas_sefori/recursos_sefori2016/lineas_investigacion_en_emociones/%21

# Motivation: Cross-modal biometric matching

Cross-modal biometric matching

- V-F: given an audio clip of a voice and two or more face images/videos, select the face image/video that corresponds to the voice.
- F-V: given an image or video of a face, determine the corresponding voice.

- Can you recognize someone’s face if you have only heard
- their voice?
- Can you recognize their voice if you have only seen their face?

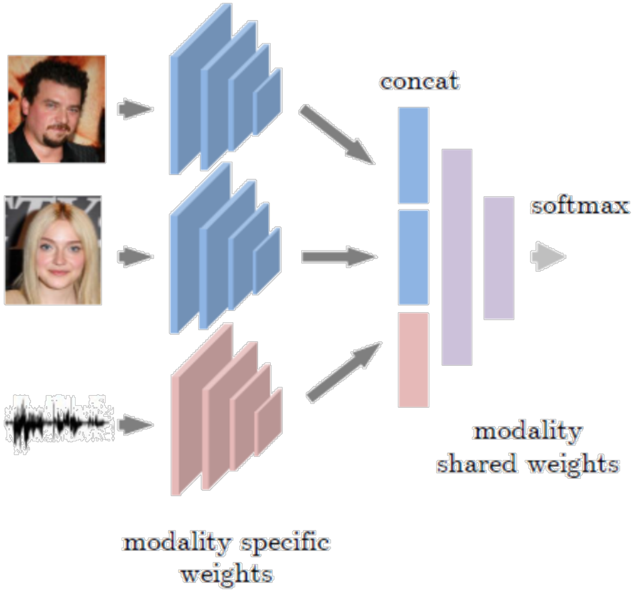


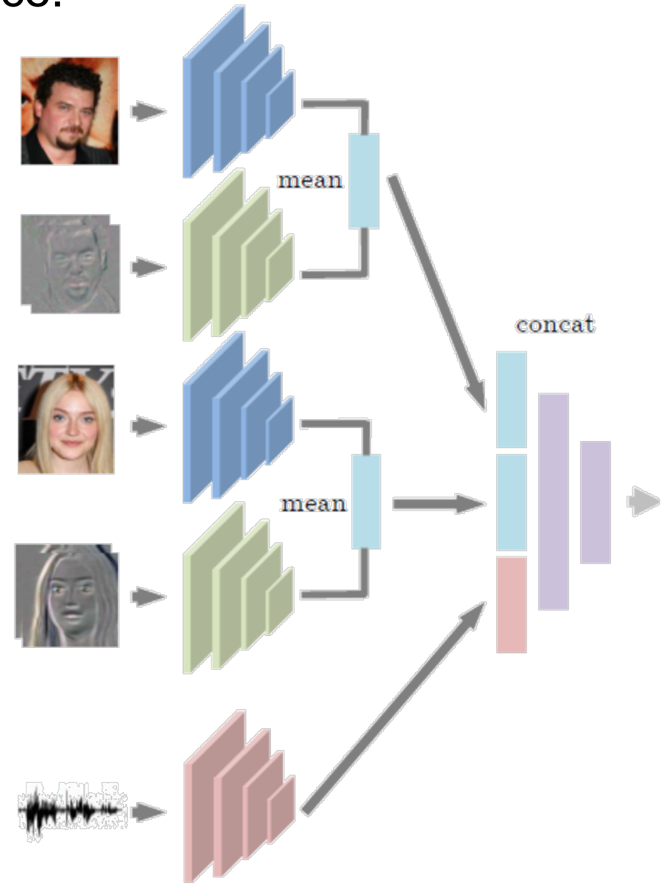
# Motivation: Cross-modal biometric matching

(1) Static (left figure): The static 3-stream CNN architecture consisting of two face sub-networks and one voice network.

(2) Dynamic (right figure): A 5-stream dynamic-fusion architecture with two extra streams as dynamic feature subnetworks.

Output: Predict the position of the positive face.





Reference: A. Nagrani, S. Albanie, A. Zisserman, Seeing voices and hearing faces: Cross-modal biometric matching, CVPR 2018, http://www.robots.ox.ac.uk/~vgg/research/CMBiometrics/

# Motivation: Speech-based video generation

Generate a video of a talking face. The method takes as inputs: (i) still images of the target face, and (ii) an audio speech segment; and outputs a video of the target face lip synched with the audio.

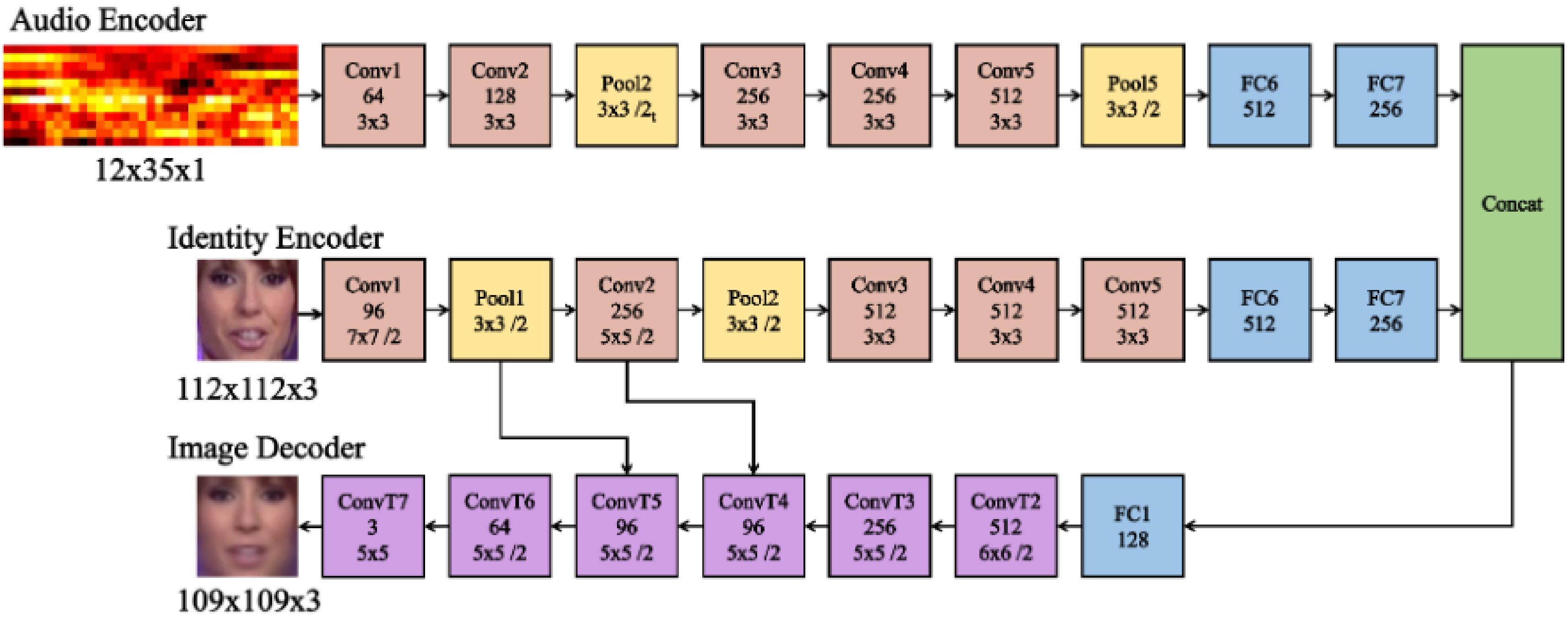
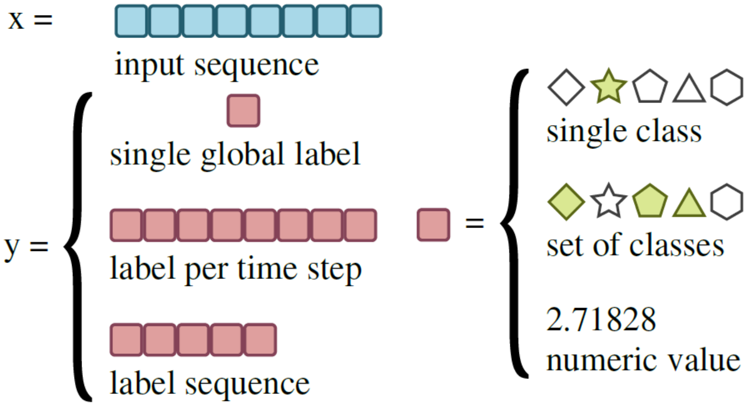


Figure 6: From top to bottom: (i) audio encoder, (ii) identity encoder with a single still image input, and (iii) image decoder. “/2” refers to the stride of each kernel in a specific layer.

Reference: Joon Son Chung, Amir Jamaludin, Andrew Zisserman, You said that? BMVC 2017, https://arxiv.org/abs/1705.02966

# Key audio sensing tasks

- Analyzing
  - speech
  - music
  - environmental sound
- Synthesis and transformation of audio
  - source separation
  - speech enhancement
  - audio generation



x =

input sequence

y =

single global label

label per time step

label sequence

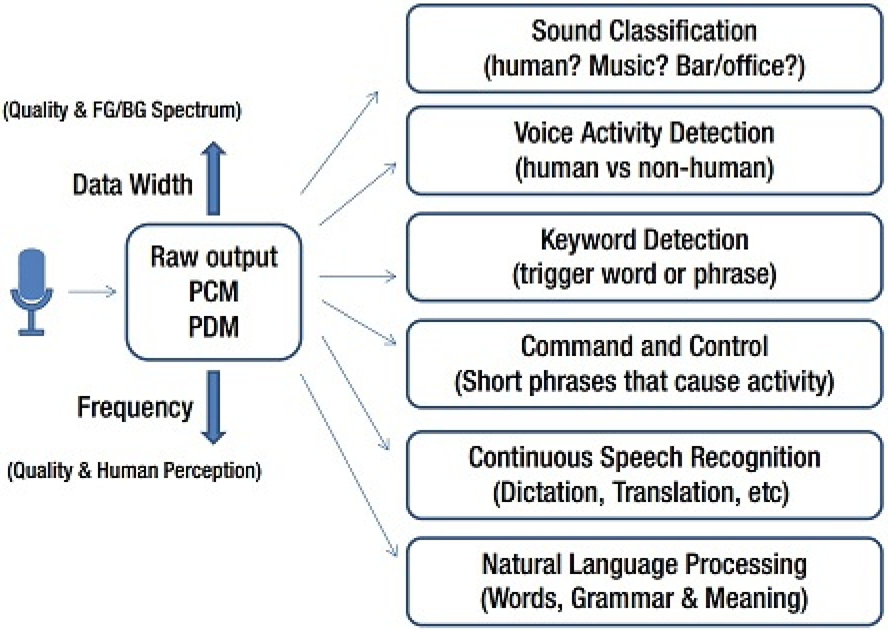
=

single class

set of classes

2.71828

numeric value



(Quality & FG/BG Spectrum)

Data Width

Raw output
PCM
PDM

Frequency

(Quality & Human Perception)

Sound Classification
(human? Music? Bar/office?)

Voice Activity Detection
(human vs non-human)

Keyword Detection
(trigger word or phrase)

Command and Control
(Short phrases that cause activity)

Continuous Speech Recognition
(Dictation, Translation, etc)

Natural Language Processing
(Words, Grammar & Meaning)

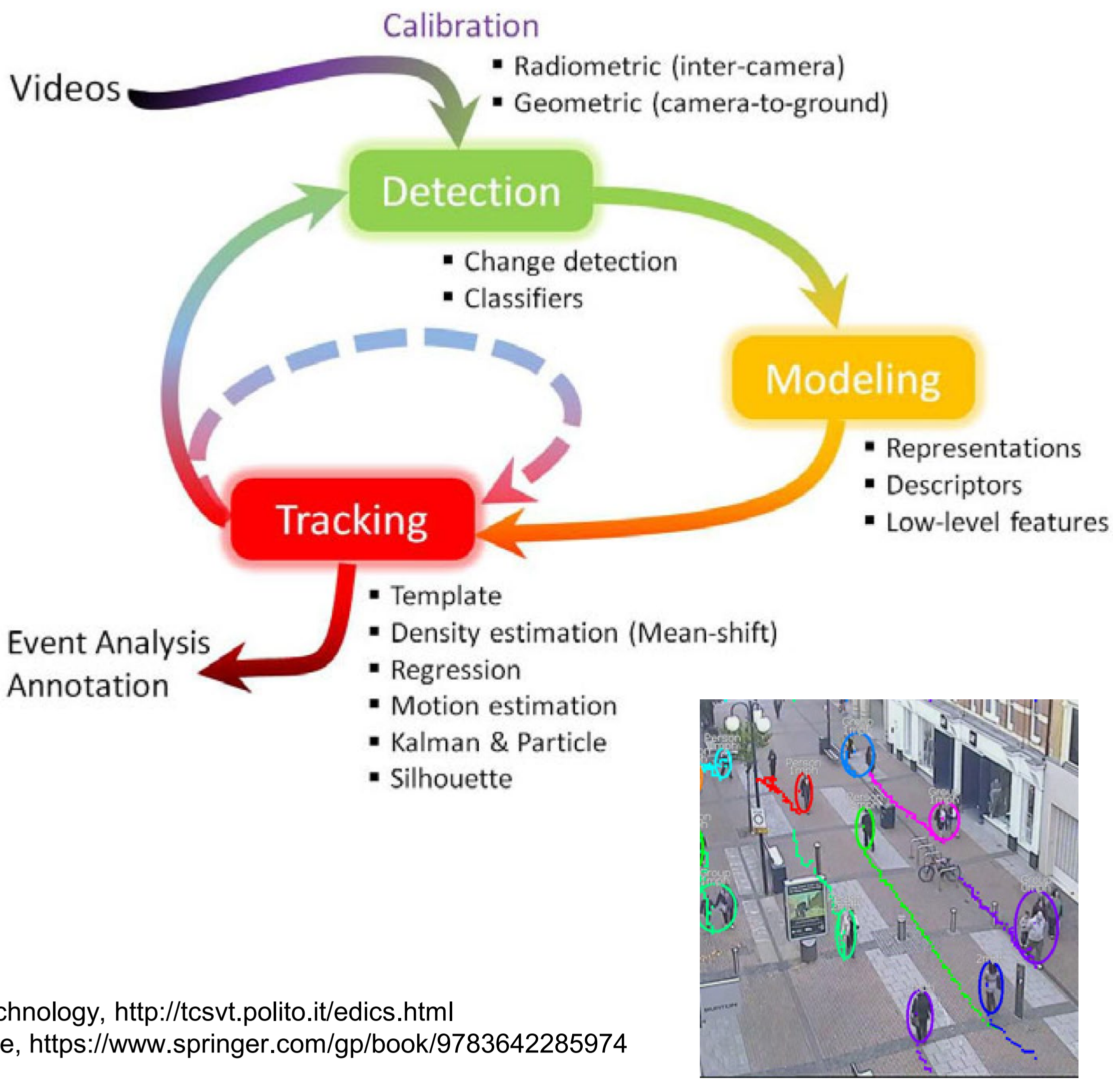
Reference

- H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep Learning for Audio Signal Processing, IEEE Journal of selected topics of signal processing, May, 2019, https://arxiv.org/abs/1905.00078
- https://www.apress.com/gp/blog/all-blog-posts/making-sense-of-sensors/12253808

# Key video sensing tasks

- Processing
- Analytics
- Compression
- Communications
- Search and retrieval
- Applications for X, error concealment, super-resolution, tracking, trajectory, action.

An example in object detection and tracking



Videos

Calibration
- Radiometric (inter-camera)
- Geometric (camera-to-ground)

Detection
- Change detection
- Classifiers

Modeling
- Representations
- Descriptors
- Low-level features

Tracking
- Template
- Density estimation (Mean-shift)
- Regression
- Motion estimation
- Kalman & Particle
- Silhouette

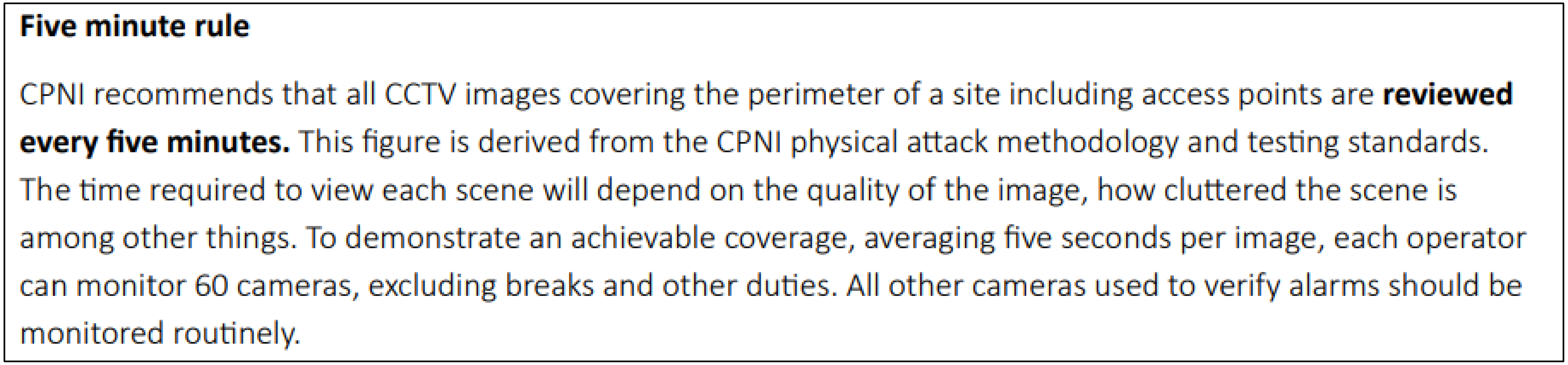
Event Analysis
Annotation

Reference
- EDIC code for IEEE Trans. circuit systems for video technology, http://tcsvt.polito.it/edics.html
- C. Shan, et al., Video Analytics for Business Intelligence, https://www.springer.com/gp/book/9783642285974

# Why real time?

When used in real-time mode, each frame of the video stream is analysed as soon as it is captured and alarms are generated whenever pre-defined triggers are encountered. When used in forensic mode (post-event), analysis software can be used to search through recorded video for pre-defined triggers, or search for points in the video where alerts have been generated.

Reference: Singapore video analytics within video surveillance systems, 2019, https://www.singaporestandardseshop.sg/Product/GetPdf?fileName=191003125456TR%2069-2-2019%20Preview.pdf&pdtid=d3d8770b-56fb-4ccf-b238-d963c4b76a28



**Five minute rule**

CPNI recommends that all CCTV images covering the perimeter of a site including access points are **reviewed every five minutes.** This figure is derived from the CPNI physical attack methodology and testing standards. The time required to view each scene will depend on the quality of the image, how cluttered the scene is among other things. To demonstrate an achievable coverage, averaging five seconds per image, each operator can monitor 60 cameras, excluding breaks and other duties. All other cameras used to verify alarms should be monitored routinely.

For maximum situational awareness for an operator this function should be enabled. It is recommended that 5 seconds of pre alarm footage and 10 seconds of post alarm footage are displayed automatically on the generation of an alarm.

Reference: UK centre for protection of national infrastructure CCTV for CNI Perimeter Security guidance, https://www.cpni.gov.uk/system/files/documents/ff/2f/CCTV%20for%20CNI%20Perimeter.pdf

# What is real time?

## Real-time in Perceptual Sense

- Real-time in the perceptual sense is used mainly to describe the interaction between a human and a computer device for a near instantaneous response of the device to an input by a human user.
- Ref. [1]: “the result of processing appears effectively ‘instantaneously’ (usually in a perceptual sense) once the input becomes available.”
- Ref. [2]: “digital processing of an image which occurs seemingly immediately; without a user-perceivable calculation delay.”

## Real-time in Signal Processing Sense

- Ref. [3]: “completing the processing within the allowable or available time between samples.”

Reference:

1. A. Bovik, “Introduction to Digital Image and Video Processing,” Handbook of Image & Video Processing, Amsterdam: Elsevier Press, 2005.
2. N. Guy, Photonotes Dictionary of Photography, http://www.photonotes.org/, 2004.
3. N. Kehtarnavaz, Real-Time Digital Signal Processing Based on the TMS320C6000. Amsterdam: Elsevier, 2004.

# Thank you!

Dr TIAN Jing
Email: tianjing@nus.edu.sg