ICSI-533/433 Multimedia Computing

Multimedia Data Analytics for Surveillance

Pradeep Atrey
SUNY at Albany

Topics

- Introduction, goals, requirements and challenges
- Face detection and recognition
- Background modeling, blob detection and tracking
- Event detection
- Multimedia assimilation, sensor selection
- Privacy issues
- Summary and concluding remarks

Topics

- Introduction, goals, requirements and challenges
- Face detection and recognition
- Background modeling, blob detection and tracking
- Event detection
- Multimedia assimilation, sensor selection
- Privacy issues
- Summary and concluding remarks

Public Safety is Important

9/11 Terrorist attack (2001)

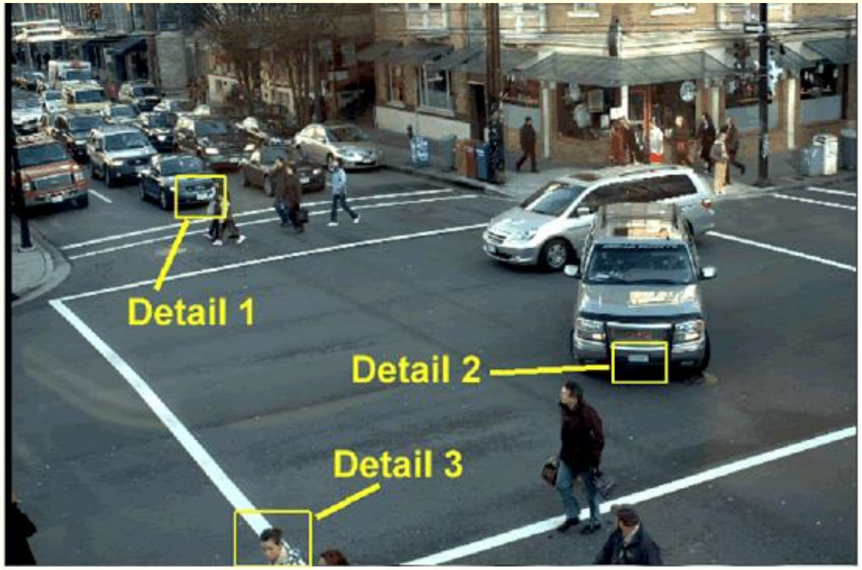
London bombing (2005)

Mumbai attack (2008)

Mumbai serial blast (2011)

Surveillance

- Large number of CCTV cameras
  - "4.2million CCTV cameras in Britain", and the "person can be captured on 300 different cameras in a day"



http://www.nione-security.com/news_view.asp?id=727

http://p10.hostingprod.com/@spyblog.org.uk/blog/cctv-surveillance-cameras/
http://www.dvbhardware.com/index.php?cPath=9

Multimedia Surveillance

Motion sensor

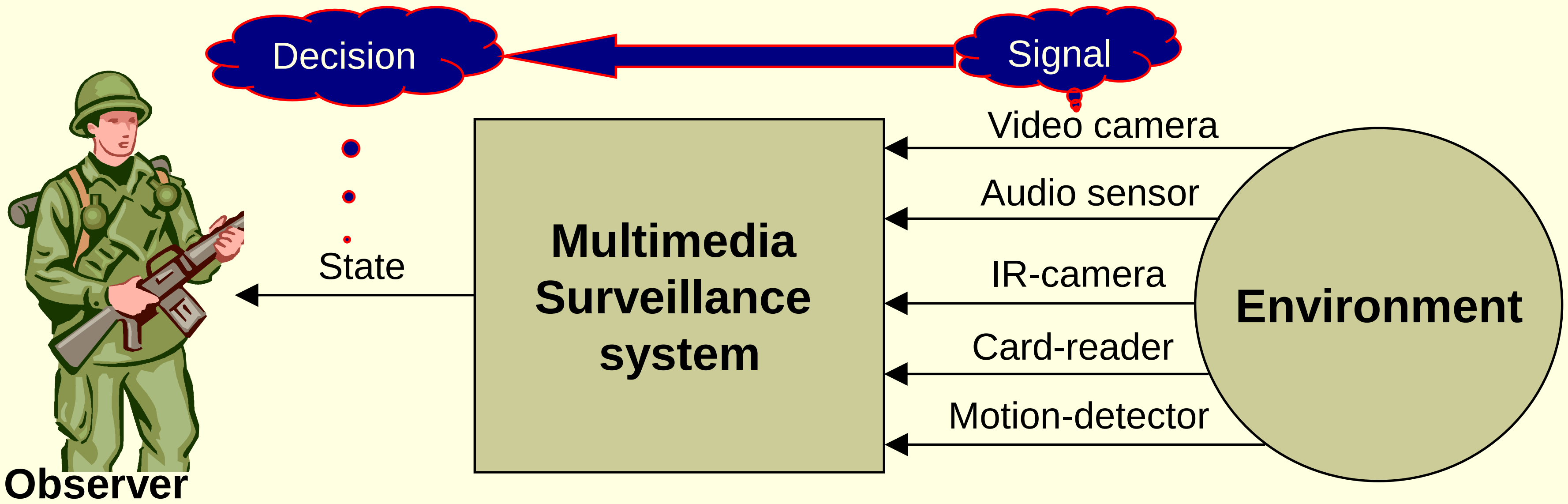
Audio sensor

RFID sensor

Night vision camera

IP camera

What is Multimedia Surveillance?

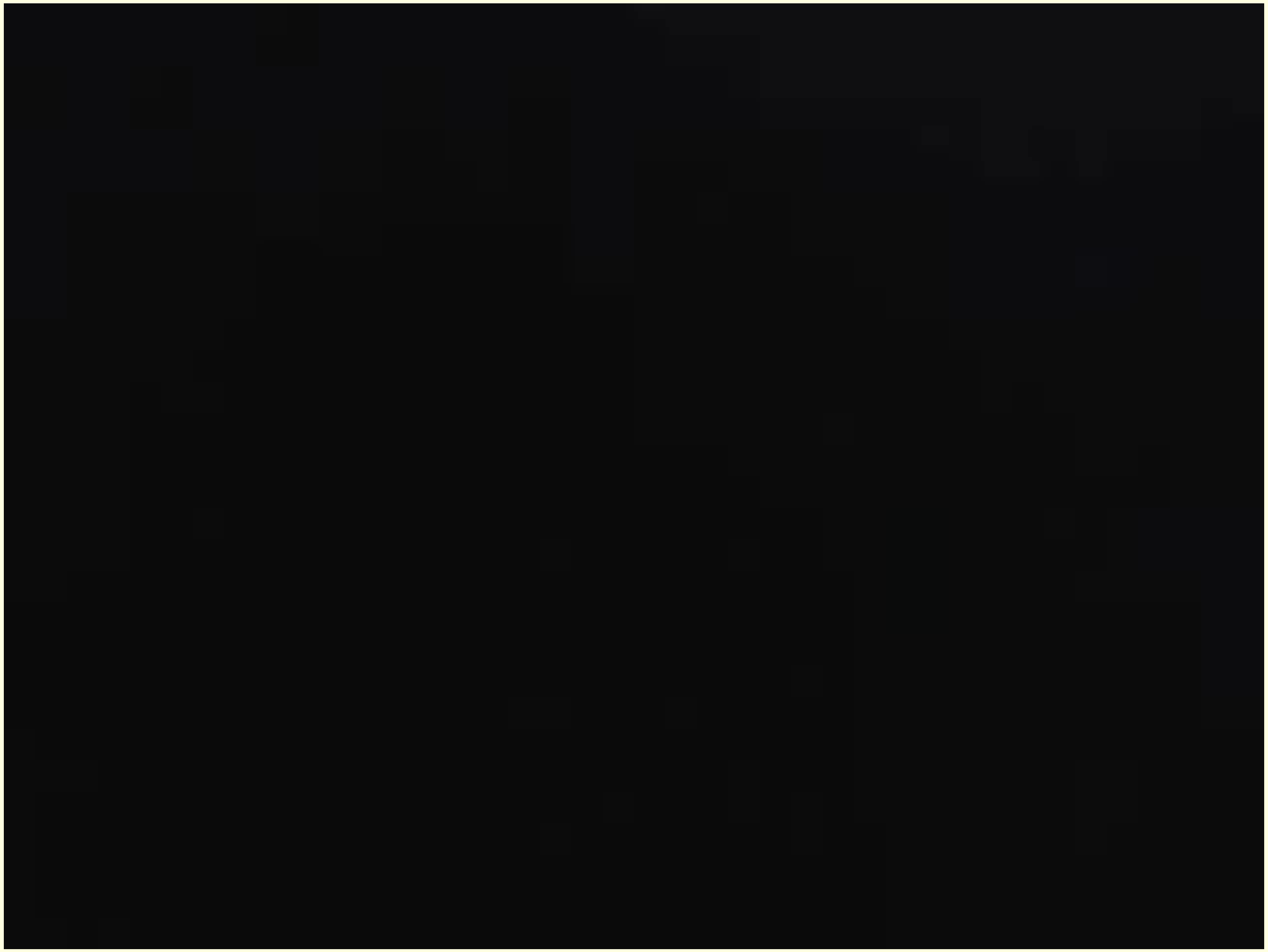


- Multimedia: Multiple Media (Sensors)
- Surveillance: To keep a watch on people's activities and behaviors for the purpose of security
- Multimedia Surveillance: Performing surveillance using multimedia system that consist of multiple sensors

Why multiple media?

- Better accuracy in signal-to-decision mapping
- Use of correlation (Index of agreement)
- Use of complementarities and cooperativeness
- Better reliability and scalability
- Better Timeliness
- Lower amortized cost

A motivating example



Multimedia Fusion for Surveillance: Characteristics

- Different media are usually captured in different formats and at different rates. Therefore, the fusion process needs to address this asynchrony to better accomplish a task.
- The processing time of different types of media streams are dissimilar, which influences the fusion strategy that needs to be adopted.
- The modalities may be correlated or independent. The correlation can be perceived at different levels.
- The different modalities usually have varying confidence levels in accomplishing different tasks.
- The capturing and processing of media streams may involve certain costs, which may influence the fusion process.

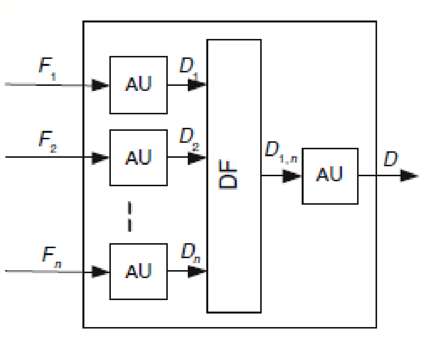
Fundamental Problems in Multimedia Surveillance

- Person detection and identification
  - Face detection
  - Face recognition

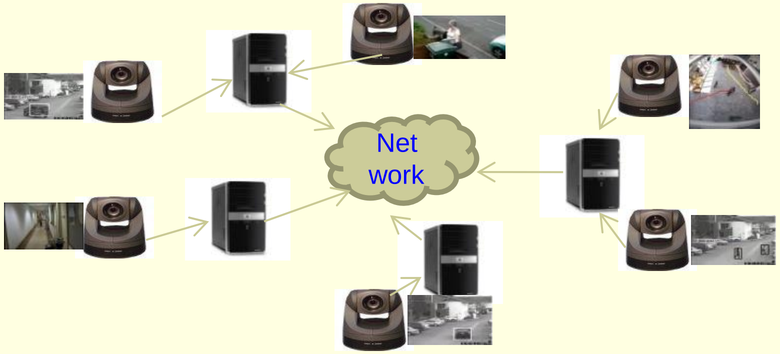
- Object detection and tracking
  - Background modeling
  - Blob detection
  - Blob tracking
  - Event detection classification
    - normal vs. suspicious

Fundamental Problems in Multimedia Surveillance

- Multimodal information assimilation
  - When, what and how to fuse?

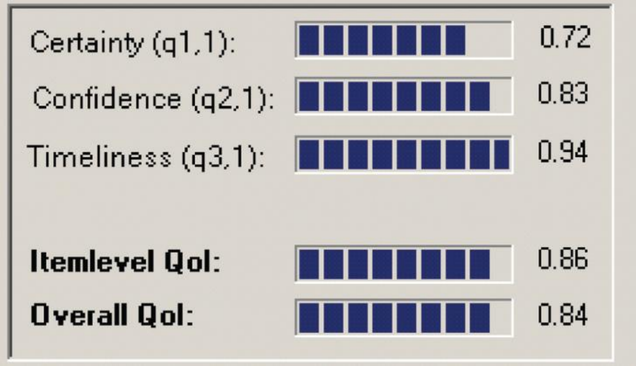


- System design
  - How many and what types of media sensors?
  - Quality of chosen media sensors?
  - Sensor placement?
  - System architecture?
  - Functionality delegation?



Fundamental Problems in Multimedia Surveillance

- Performance modeling
  - Quality of Information (QoI) assessment
  - Accuracy or timeliness?



- Privacy vs. Security
  - Two contradictory goals
  - How to achieve the goal of security while preserving the privacy of people?
  - What is privacy? What causes privacy loss?

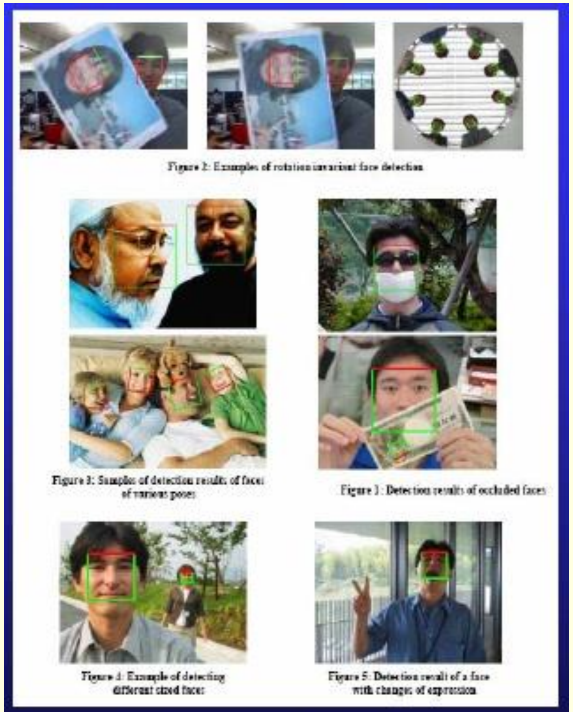
- And so on…

Topics

- Introduction, goals, requirements and challenges
- Face detection and recognition
- Background modeling, blob detection and tracking
- Event detection
- Multimedia assimilation, sensor selection
- Privacy issues
- Summary and concluding remarks

Face Detection and Recognition

- Face Detection: A Solved Problem?
  - Identify and locate human faces in an image regardless of their
    - position
    - scale
    - in-plane rotation
    - orientation
    - pose (out-of-plane rotation)
    - and illumination



Omron's face detector
[Liu et al. 04]

Face Detection Approaches

- Knowledge-based methods
  - Encode human knowledge of what constitutes a typical face (usually, the relationships between facial features)
- Feature invariant approaches
  - Aim to find structural features of a face that exist even when the pose, viewpoint, or lighting conditions vary
- Template matching methods
  - Several standard patterns are stored to describe the face as a whole or the facial features separately
- Appearance-based methods
  - The models (or templates) are learned from a set of training images which capture the representative variability of facial appearance

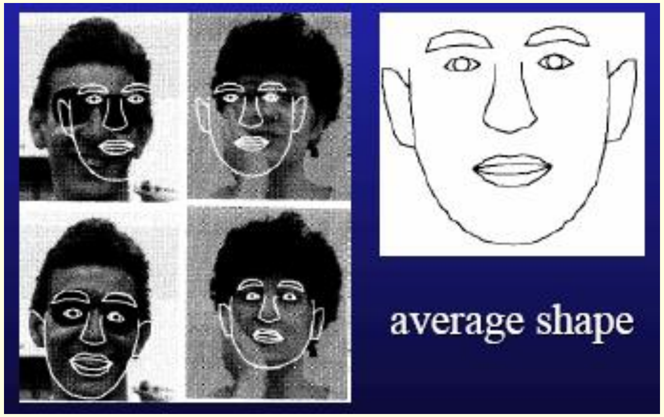
Knowledge-based Methods

- Characteristics
  - Top-down approach: Represent a face using a set of human-coded rules
  - Example:
    - The difference between the average intensity values of the center part and the upper part is significant
    - A face often appears with two eyes that are symmetric to each other, a nose and a mouth
    - Use these rules to guide the search process
- Pros:
  - Easy to come up with simple rules to describe the features of a face and their relationships
  - Based on the coded rules, facial features in an input image are extracted first, and face candidates are identified
  - Work well for face localization in uncluttered background
- Cons:
  - Difficult to translate human knowledge into rules precisely: detailed rules fail to detect faces and general rules may find many false positives
  - Difficult to extend this approach to detect faces in different poses: implausible to enumerate all the possible cases

Feature-Based Methods

- Characteristics
  - Bottom-up approach: Detect facial features (eyes, nose, mouth, etc) first
  - Facial features: edge, intensity, shape, texture, color, etc
  - Aim to detect invariant features
  - Group features into candidates and verify them
- Pros:
  - Features are invariant to pose and orientation change
- Cons:
  - Difficult to locate facial features due to several corruption (illumination, noise, occlusion)
  - Difficult to detect features in complex background

Template Matching Methods



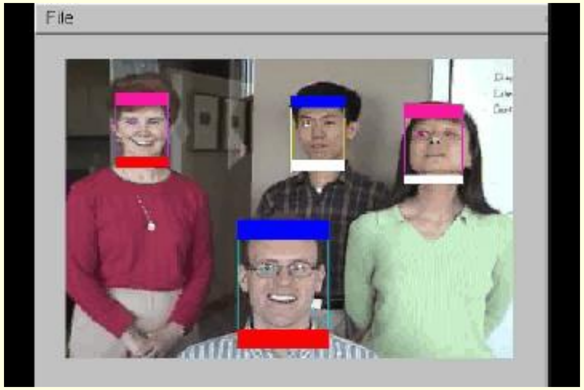
- Characteristics
  - Store a template
    - Predefined: based on edges or regions
    - Deformable: based on facial contours
  - Templates are hand-coded (not learned)
  - Use correlation to locate faces
- Pros:
  - Simple
- Cons:
  - Templates needs to be initialized near the face images
  - Difficult to enumerate templates for different poses (similar to knowledge based methods)

Appearance-Based Methods

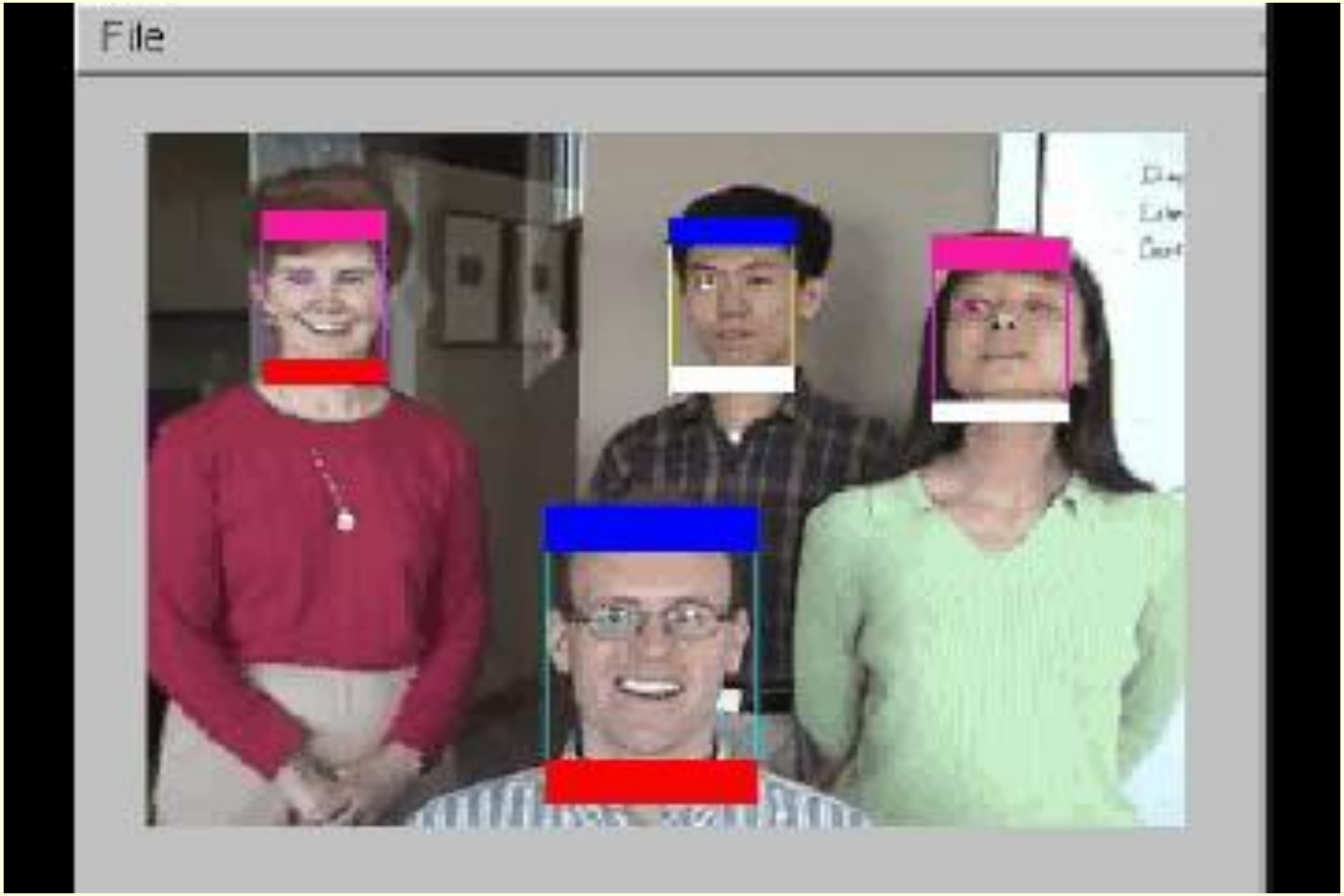
- Characteristics
  - Train a classifier using positive (and usually negative) examples of faces
  - Many face detectors are proposed
- Popular one: Adaboost-Based Detector [Viola and Jones 01]
- Pros:
  - Use powerful machine learning algorithms
  - Has demonstrated good empirical results
  - Fast and fairly robust
  - Extended to detect faces in different pose and orientation
- Cons
  - Usually needs to search over space and scale
  - Need lots of positive and negative examples
  - Limited view based approach

Other Face Detectors

- Color-based face detector
  - Based on Skin and Non-Skin Color Model
- Video-Based Face Detector
  - Motion cues:
    - Frame differencing
    - Background modeling and subtraction
  - Can also use depth cue (e.g., from stereo) when available
  - Reduce the search space dramatically



Face Detection Demo



Face Recognition Demo

Topics

- Introduction, goals, requirements and challenges
- Face detection and recognition
- Background modeling, blob detection and tracking
- Event detection
- Multimedia assimilation, sensor selection
- Privacy issues
- Summary and concluding remarks

Challenges

- Object detection
  - Challenges
  - GMM based method
- Tracking
  - Challenges
  - Particle filter based tracking

Object Detection

- Goal: To detect the regions of an image that are semantically important to us:
  - People
  - Vehicle
  - Buildings

- Application
  - Crowd management
  - Traffic management
  - Video compression, video surveillance, vision-based control, human-computer interfaces, medical imaging, augmented reality, and robotics…

Object Detection in Images

- Subjectively defined
- Generally template based
- Mainly done by image segmentation

Object Detection in Videos

- Relatively Moving – Object
- Relatively Static – Background

The goal here is to differentiate the moving object from background!

Surveillance Video

- Static camera
- Background relatively static
- Subtract the background image from current image
- Ideally this will leave the moving objects
- This is not an ideal world…

Feature Based

- Objects are modeled in terms of features
- Features are chosen to handle changes in illumination, size and orientation
  - Shape based – Very hard
  - Color based – Low cost but not accurate

Template Based

- Example template are given
- Object detection becomes matching features
- Image subtraction, correlation

Motion Based

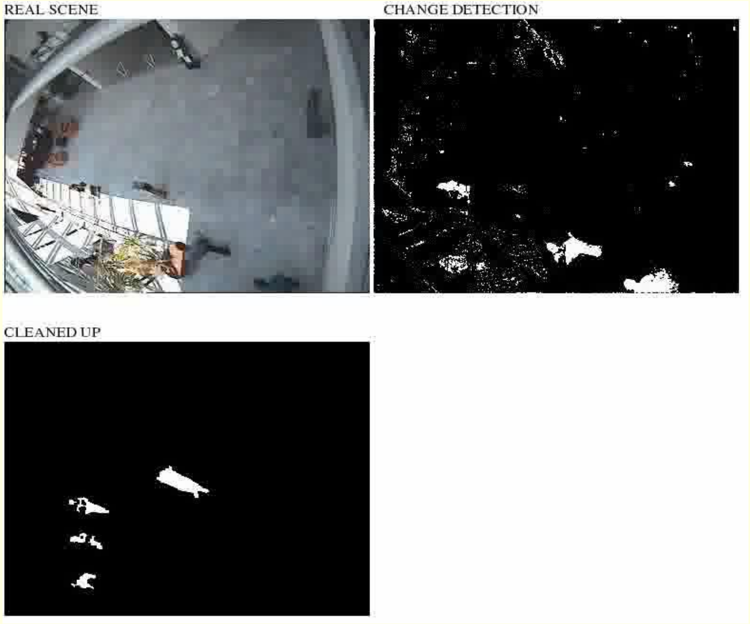
- Model background
- Subtract from the current image
- Left are moving objects ☺
- Remember! This is not a ideal world…

Problems in Modeling Background

- Acquisition noise
- Illumination variation
- Clutter
- New object introduced into background
- Object may not move continuously

Outline of Object Detection

- Determine the background and foreground pixels
- Draw contours around foreground pixels
- Use heuristics to merge these contours

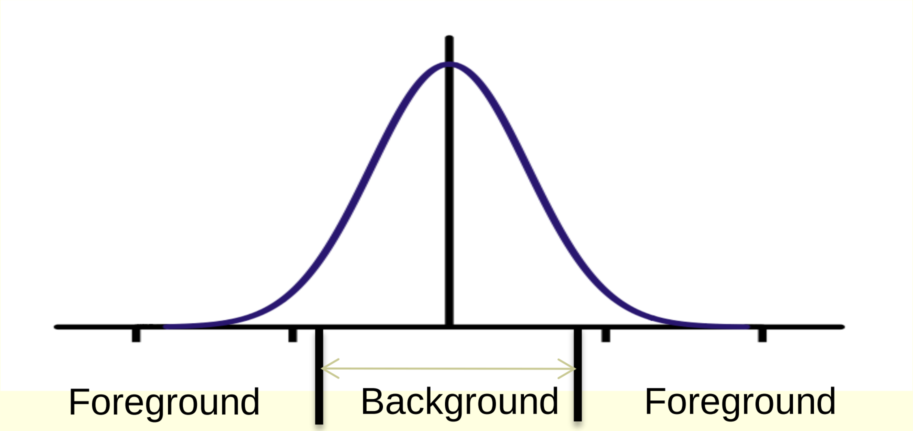


Ideal World

- Single value modeling of background
- Anything different is foreground

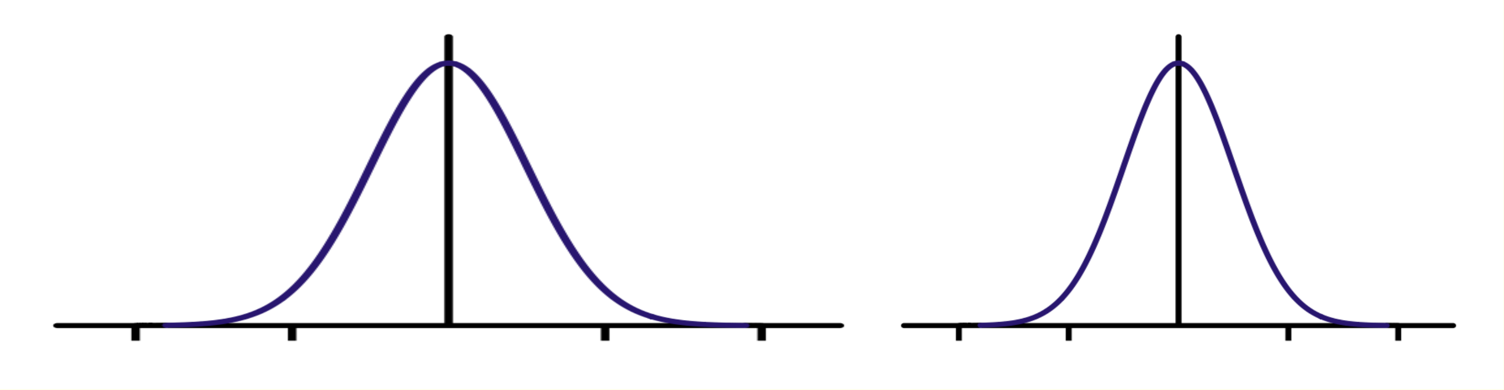
Static Background

- Each pixel resulted from a particular surface under particular lightening
  - Single Gaussian is enough (μ, σ)
- If $|P_t - \mu| < 2.5 * \sigma$
  - Pixel belongs to background, else foreground



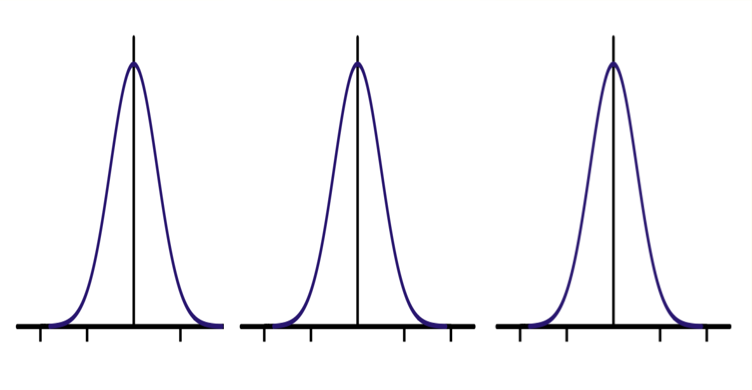
Illumination Variation

- Whenever a pixel matches the background Gaussian, update the background model i.e.
  - If $\left|P_t - \mu_t\right| < 2.5 * \sigma t$
  - Then $\mu_{t+1} = (1-\alpha)\mu_t + \alpha\mu_t$
  - Standard deviation updated accordingly



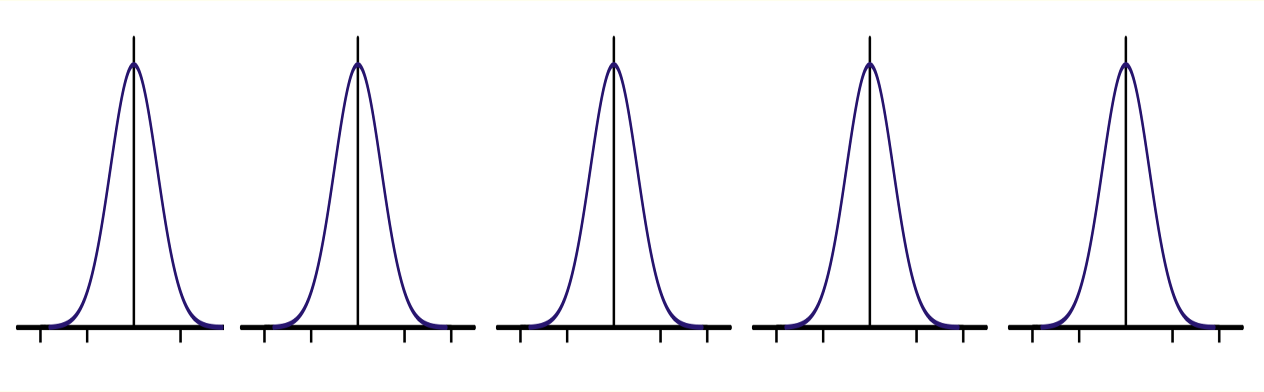
Clutter

- Think of tree leaves…
- Multiple surfaces, still part of background
- Gaussian Mixture Model
- Update each Gaussian after matching



Static Object Introduced

- Think of flower pot…
- Background model should adapt to this change
- Use Gaussian for new surface as well
  - Few extra Gaussians for the foreground



Background Selection

- A background Gaussian will have
  - More persistence – high *w*
  - Less variation – low σ_t
  - Sort Gaussians wrt w / σ_t
- Pick top k Gaussians as background such that

$$\arg\min_k \left(\sum_{i=1}^{k} w_i > T \right)$$

If pixel belongs to one of these, it's a background pixel

Adaptive Background Model

- Every pixel is modeled as mixture of Gaussians
- More persistent Gaussians belong to background and others to foreground
- The Gaussians are updated after each frame

Connecting the Dots

- The output of background modeling is a binary image
- Dilation/Erosion can further reduce noise
- Contour drawing
- Bounding boxes

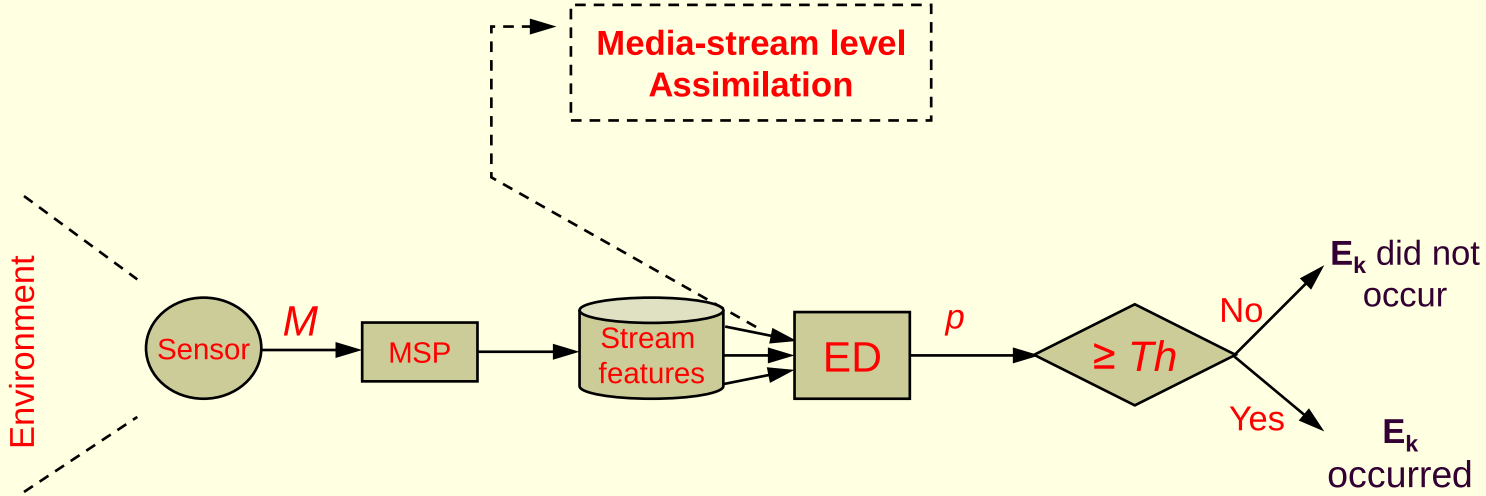
Revisit the problems

- Problems
  - Slow moving background – clutter
  - New object introduced into background
  - Illumination variation
  - Object may not move continuously

Topics

- Introduction, goals, requirements and challenges
- Face detection and recognition
- Background modeling, blob detection and tracking
- Event detection
- Multimedia assimilation, sensor selection
- Privacy issues
- Summary and concluding remarks

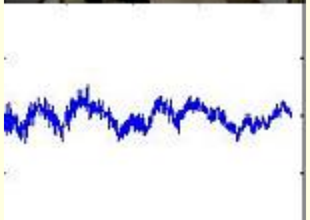
Single Medium based Event Detection



MSP = Media Stream Processor, ED = Event detector (classifier),
p = probability of occurrence of event E_K

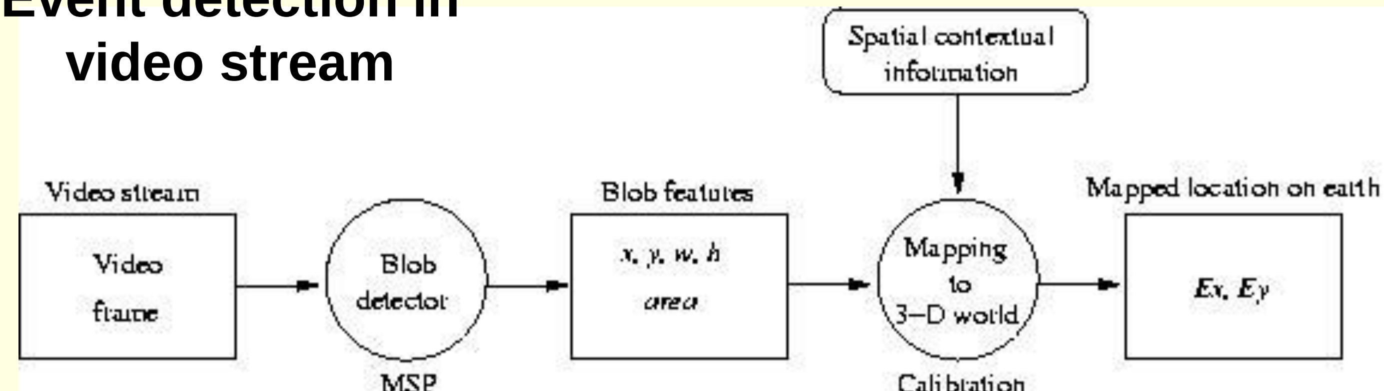
Video event detection

Audio event detection



Single Modality based Event Detection

Event detection in video stream

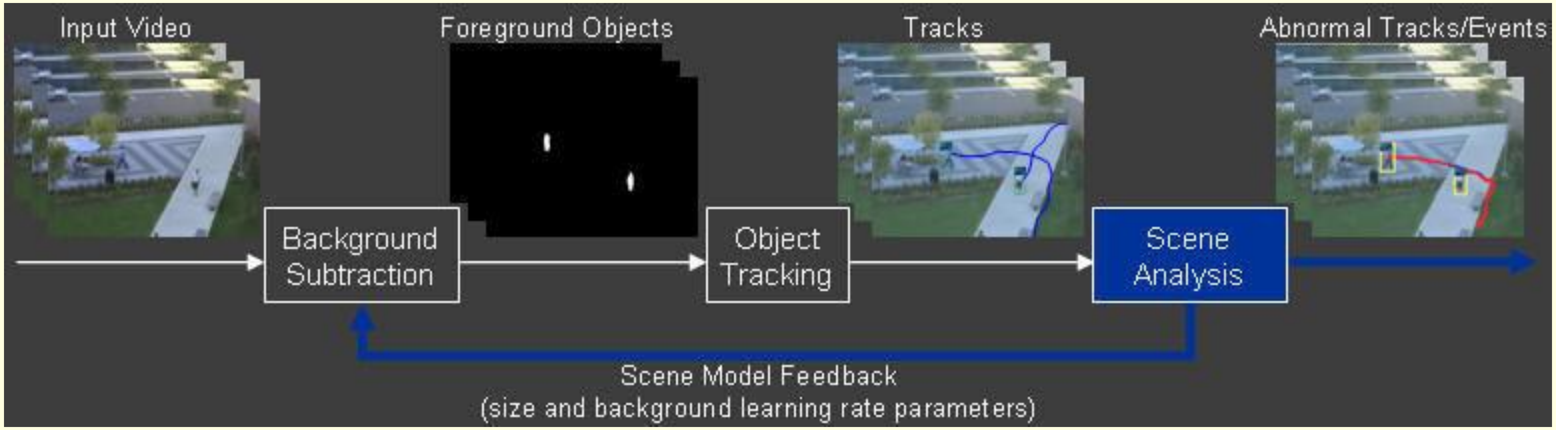


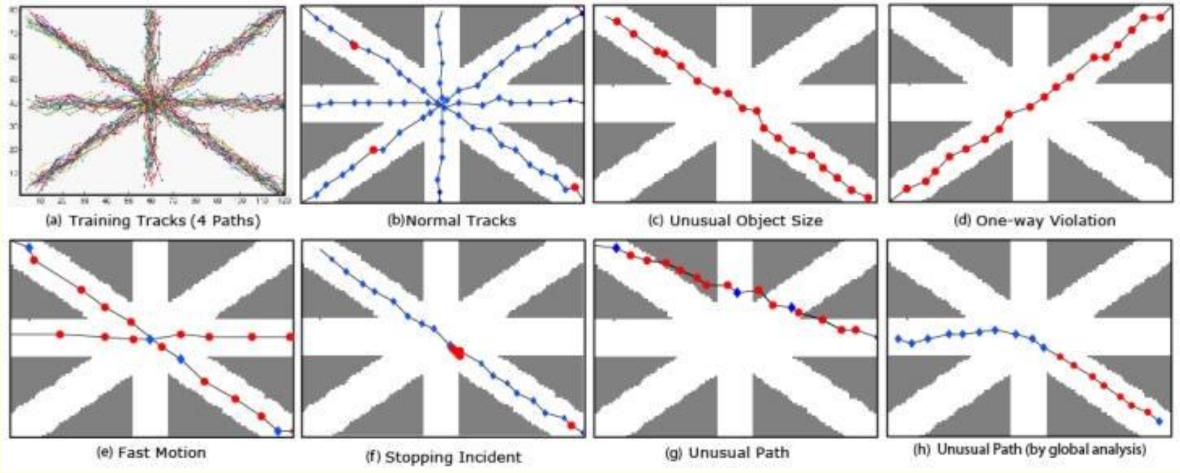
$$[Ex, Ey] = T \times [Ix, Iy]$$

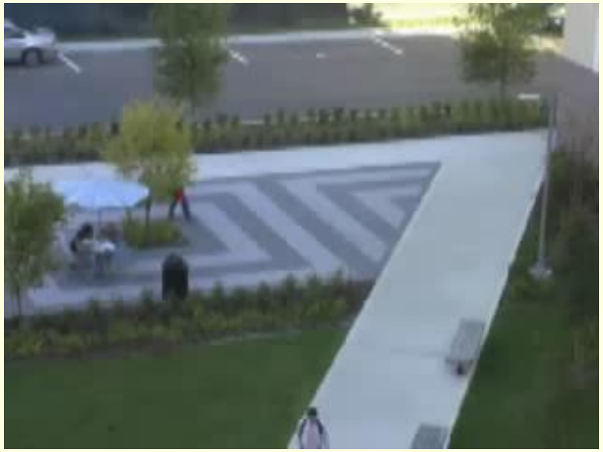
[*Ex, Ey*] : Coordinate on ground plane
T : Transformation matrix
[*Ix, Iy*] = Coordinate on image plane

Single Modality based Event Detection

Video event detection – single camera







(Unusual path)

Arslan Basharat, Alexei Gritai, and Mubarak Shah, University of Central Florida, USA [2008]

Single Modality based Event Detection

Video event detection – Multi-camera **(Abandoned baggage)**

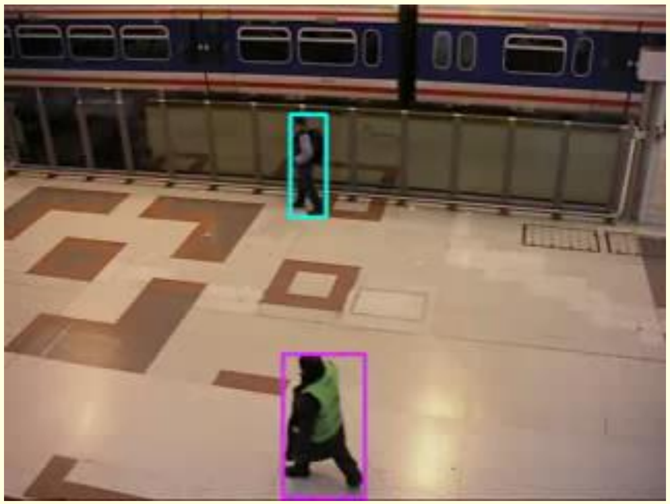
Vivek Singh, Pradeep Atrey, and Mohan Kankanhalli, NUS, Singapore [2007]

(Multi-camera multi-person tracking)

Wei Qu, Dan Schonfeld, and Magdi Mohamed, University of Illinois, Chicago, IL, USA [2006]

Single Modality based Event Detection

Video event detection – Commercial systems



(Abandoned baggage)

Indect (http://www.indect-project.eu/)

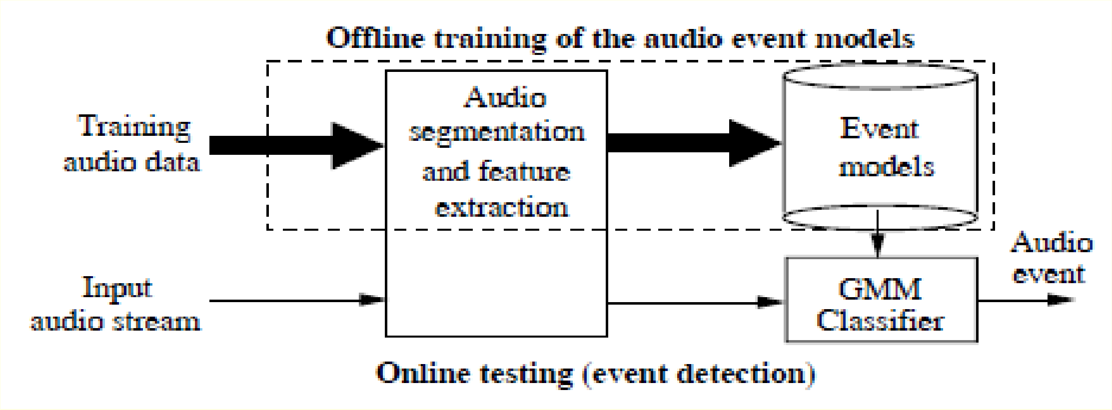
VideoTec (http://www.cctvalbert.com/en/page_700.html)

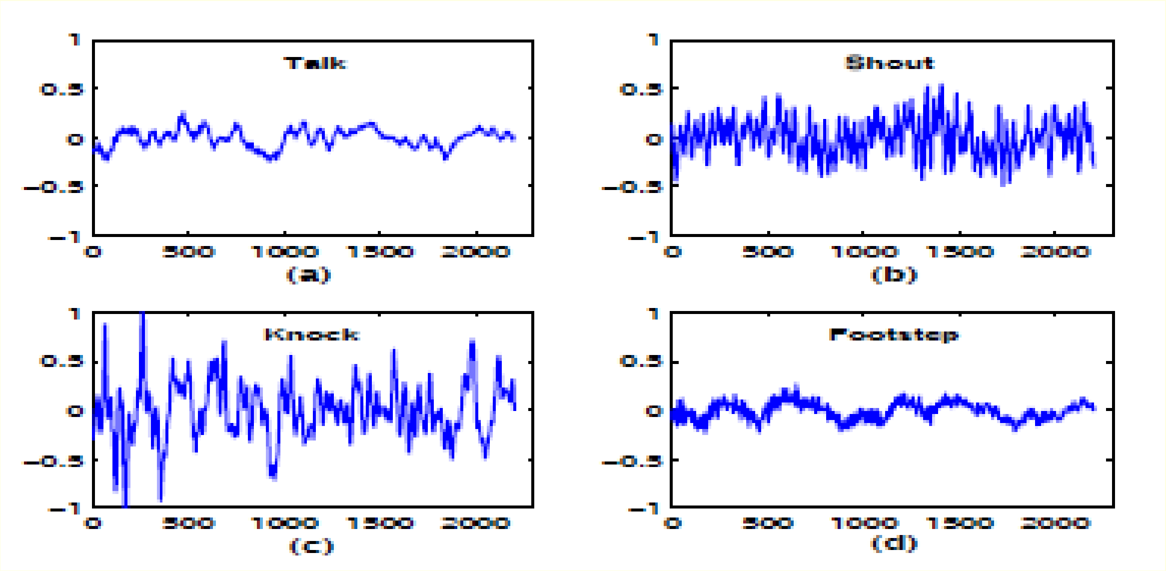
albert
Intelligent Video Analysis

(Pedestrian tracking)

Single Modality based Event Detection

Event detection in audio stream

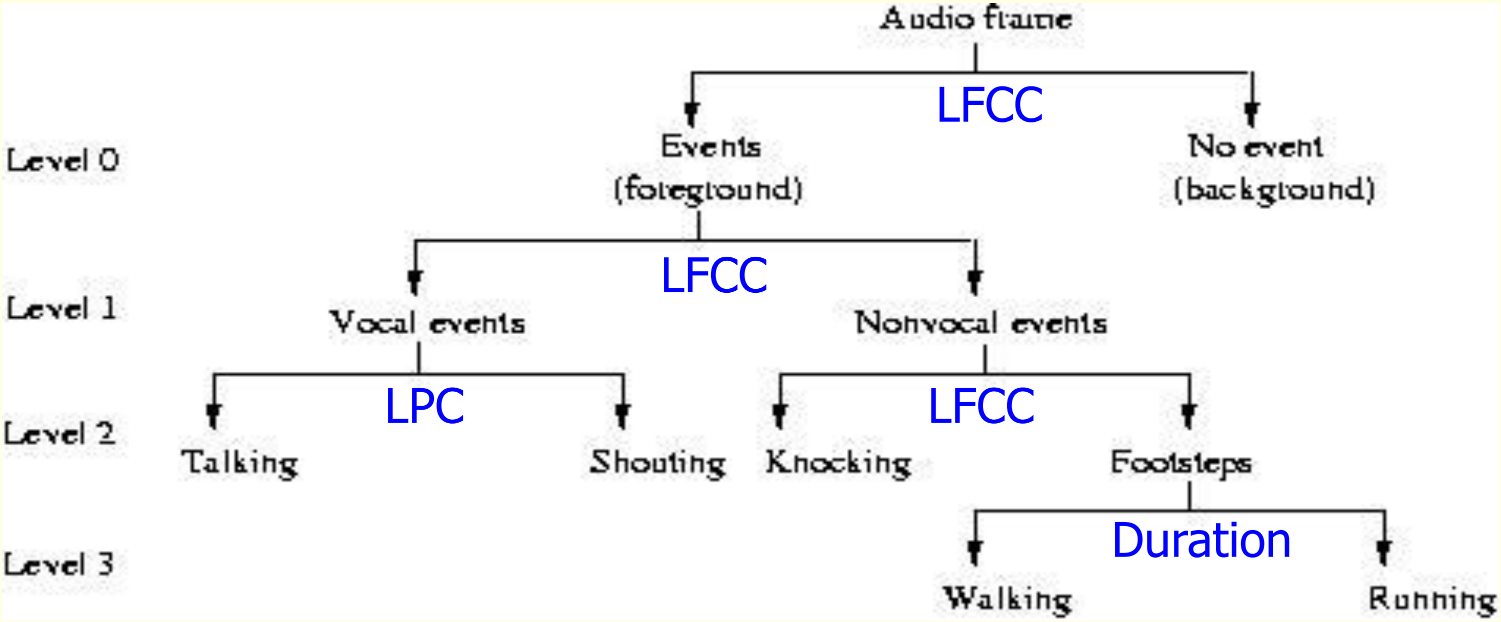




Pradeep Atrey, Namunu Maddage, and Mohan Kankanhalli, NUS, Singapore [2006]

Single Modality based Event Detection

Event detection in audio stream



LPC = Linear Predictor Coefficient
LFCC = Log Frequency Cepstral Coefficient

Pradeep Atrey, Namunu Maddage, and Mohan Kankanhalli, NUS, Singapore [2006]

Single Modality based Event Detection

- Audio events detected with decent accuracies
  - Walking
  - Running
  - Talking
  - Shouting
  - Door knocking
  - Gunshot
  - Coughing
  - Etc.

- Single audio sensor as well as multiple audio sensors

Multiple Modalities based Event Detection

- Multiple modalities (video, audio, motion sensor, etc.)
- Other issues
  - "When" (along a timeline) to assimilate?
  - "What" streams to assimilate?
  - "How" streams to assimilate?

Next…

- Information Assimilation Framework for Multimedia Surveillance – [Atrey and Mohan Kankanhalli 2006]