# *Telling What-Is-What in Video*

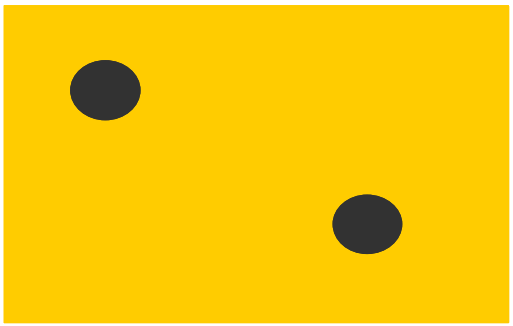
***Gerard Medioni***
***medioni@usc.edu***

USC Viterbi
School of Engineering

IMSC
Integrated
Media Systems
Center

# Tracking

- Essential problem
- Establishes correspondences between elements in successive frames
- Basic problem easy



# Many issues

- One target (pursuit) vs.

- A few objects vs.

- Lots of objects

# More issues: motion type

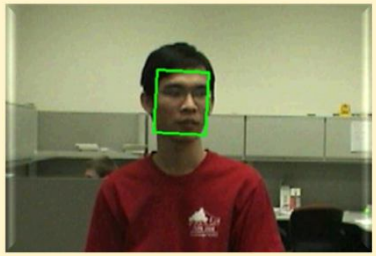
– Rigid

– Articulated

– Non rigid (face expression)

# Tag & Track - The problem

Select any object and follow it in real time



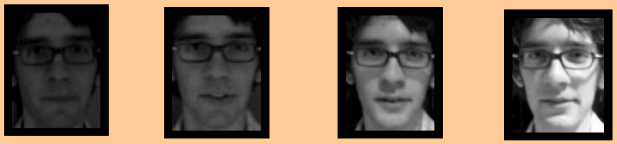
**Object tracking problem**

# Challenges

Unknown type of object

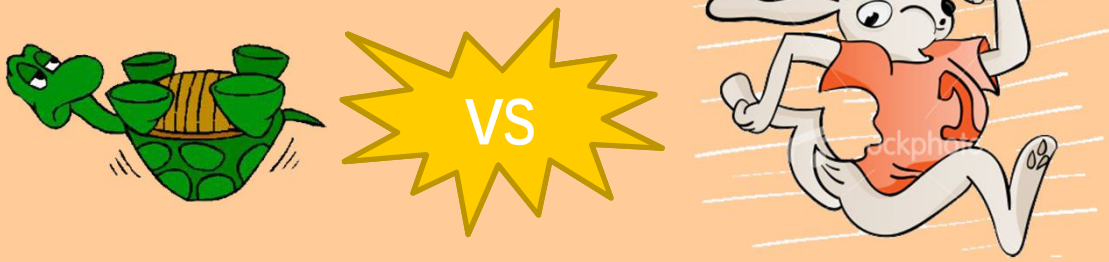
Changes in viewpoint

Changes in lighting



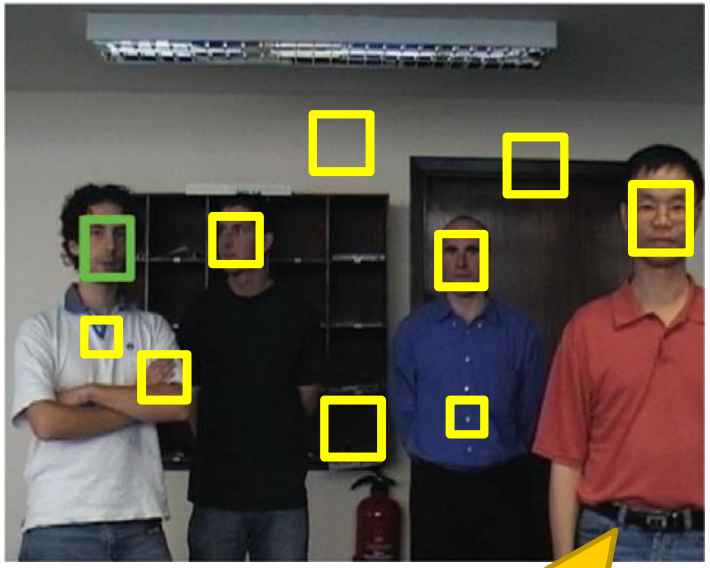
Cluttered background

Running time



# Context Tracker

- Motivation



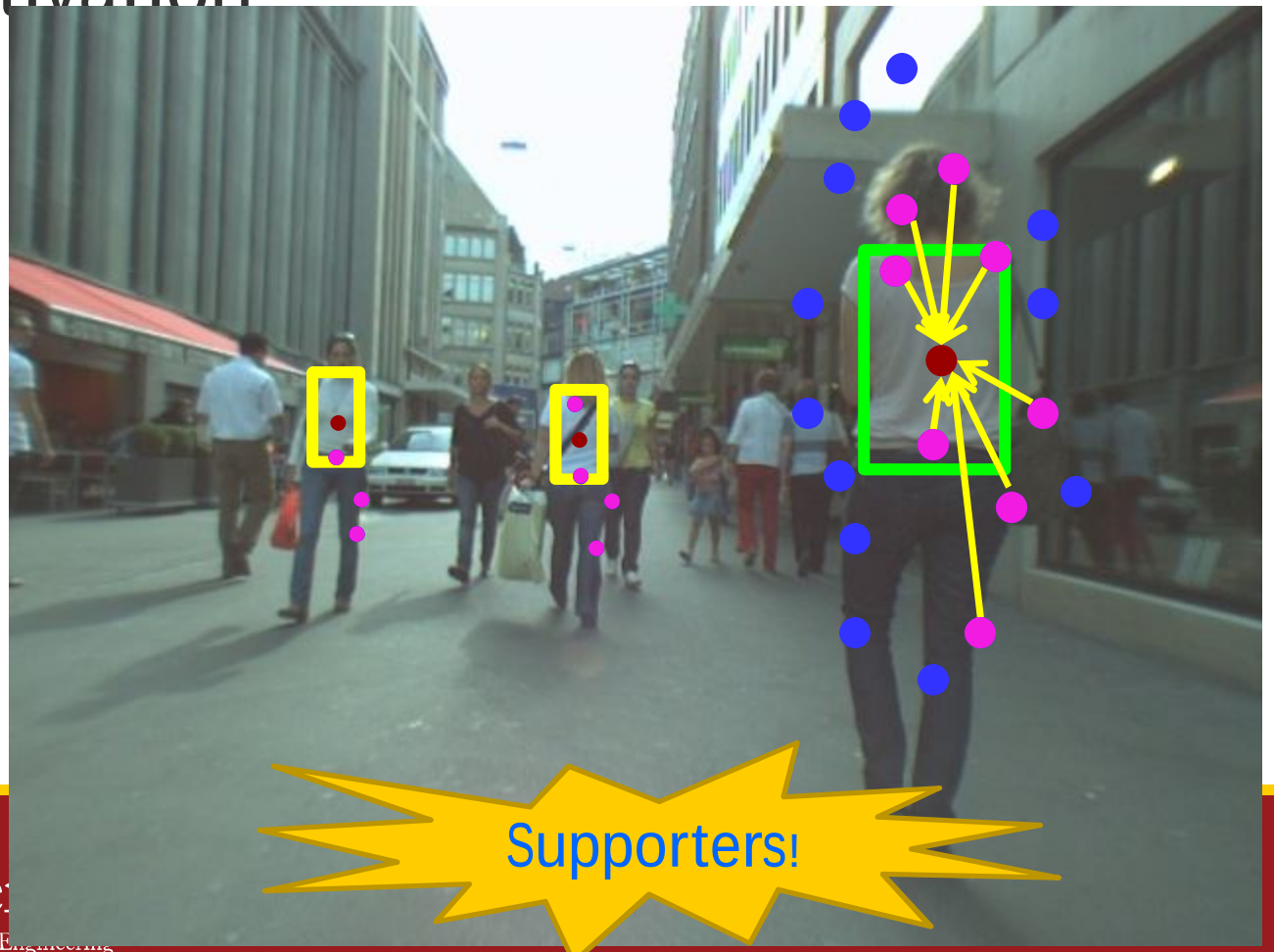
- Context inform[illegible] is overlooke[illegible]line pro[illegible] requirement, speed trade-off
  - + Focus in buildin[illegible] de of background info
  - → Requ[illegible]
  - + Treat every re[illegible]

Explore Distracters and pay more attention to them

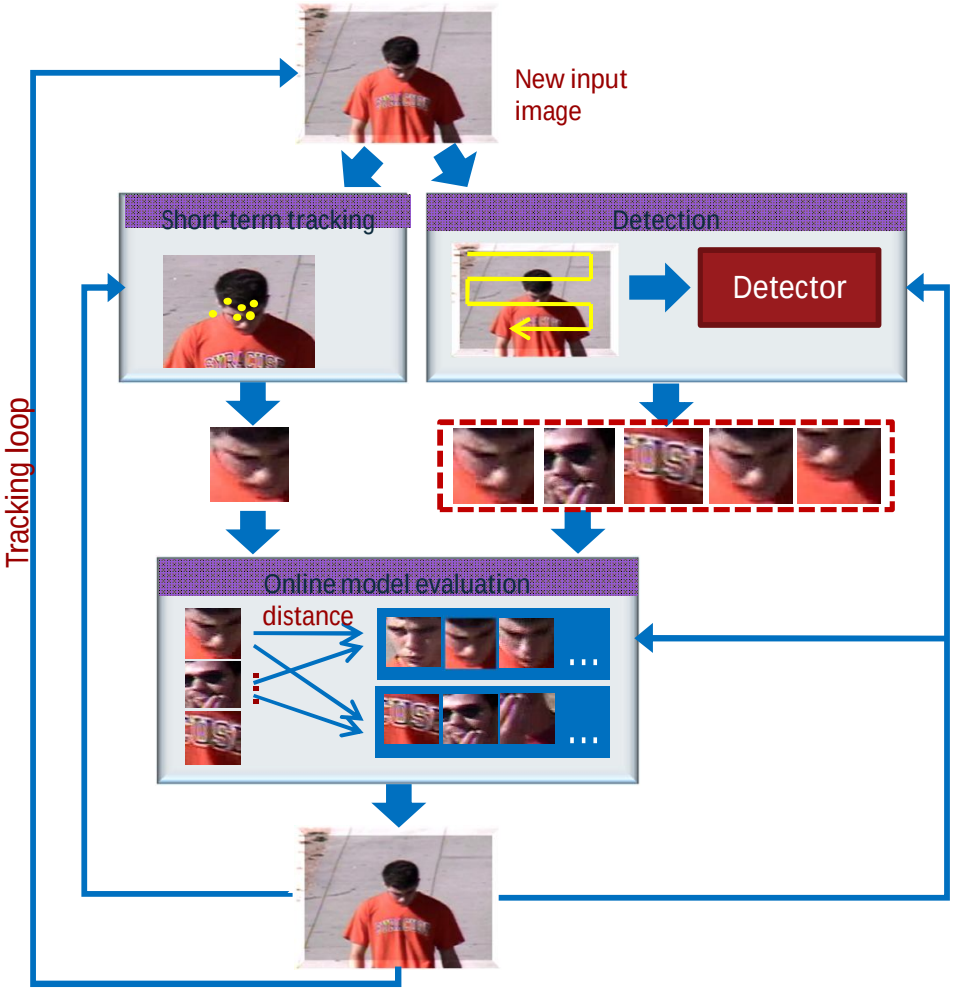
# Context Tracker

- Motivation

What else to explore?



# Context Tracker



# Context Tracker

- Detection:
  - Pass the classifier (share the same classifier)
  - High confidence (look similar to our object)
- Tracking:
  - Same as tracking our target BUT will be killed when being lost or look different from our target
  - Heuristic data association: the higher confidence has higher priority in the association queue

# Context Tracker

- Experiment settings
  - 8 ferns and 4 6bitBP features
  - Minimum search region 20x20

| Video Sequence | Frames | FT | MILT | CoTT | DNBS | VTD | PNT | Ours |
|---|---|---|---|---|---|---|---|---|
| Animal | 72 | 69 | 9 | *8* | 19 | **6** | 37 | 9 |
| Carchase | 5000 | lost @ 355 | lost @ 355 | lost @ 409 | lost @ 364 | lost @ 357 | lost @1645 | **24** |
| Clutter | 1528 | lost @ 1081 | lost @ 413 | 9 | *6* | *6* | **4** | *6* |
| ETHPedestrian | 874 | lost @ 95 | lost @ 95 | lost @ 95 | lost @ 635 | lost @ 95 | **10** | *16* |
| Girl | 502 | lost @ 248 | 30 | **14** | 39 | 69 | 19 | *18* |
| Liquor | 1407 | lost @ 47 | lost @ 288 | 30 | lost @ 404 | lost @ 404 | *21* | **10** |
| Motocross | 2665 | lost @137 | lost @ 485 | lost @ 591 | lost @ 10 | lost @ 10 | **10** | *12* |
| Multifaces | 1006 | lost @ 64 | lost @ 64 | lost @ 394 | lost @ 64 | lost @ 64 | lost @ 97 | **26** |
| Scale | 1911 | 8 | 11 | 6 | lost @ 269 | *3* | 6 | **2** |
| Vehicle | 946 | lost @ 679 | lost @ 481 | *9* | lost @ 517 | lost @ 517 | **8** | **8** |
| Speed (fps, on 320x240) | | 1.6 | 14 | 2 | 7 | 0.2 | 12* | 10 |

# Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments

Paper ID: 1935

# Context Tracker vs. Others - Motocross


# Active Surveillance

Combine

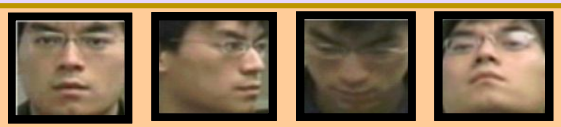
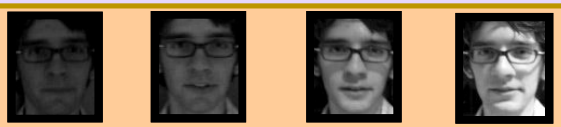
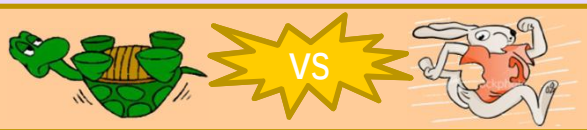
- Real Time tracker and
- Camera Control

– To keep object of interest in the field of view of the camera

– To zoom in (on the face)

# Challenges

Tracking

- Unknown type of object 
- Changes in viewpoint 
- Changes in lighting 
- Cluttered background 
- Running time 

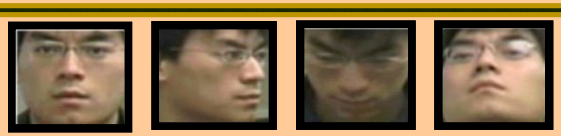

**Control**

- Limited support from commercial cameras with discrete speed control due to the use of stepping motors. 
- Delay because of communication through TCP/IP Network → abrupt motion and motion blur 

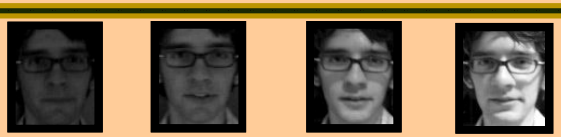
# Challenges

Unknown type of object

Changes in viewpoint



Changes in lighting



Cluttered background

Tracking

Running time

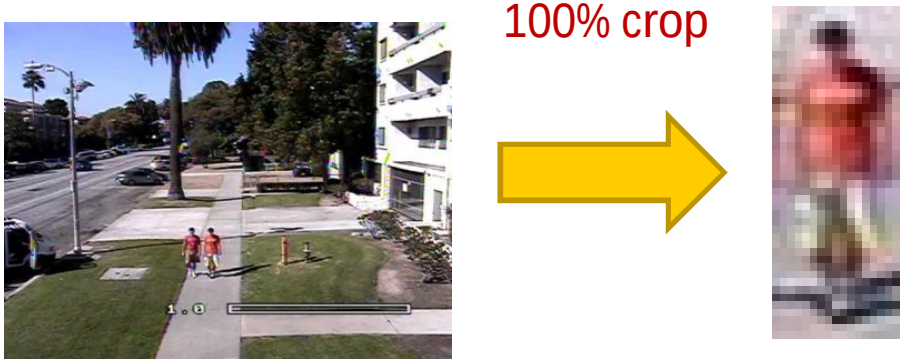

Limited support from commercial cameras with discrete speed control due to the use of stepping motors.

Delay because of communication through TCP/IP Network → abrupt motion and motion blur

Control

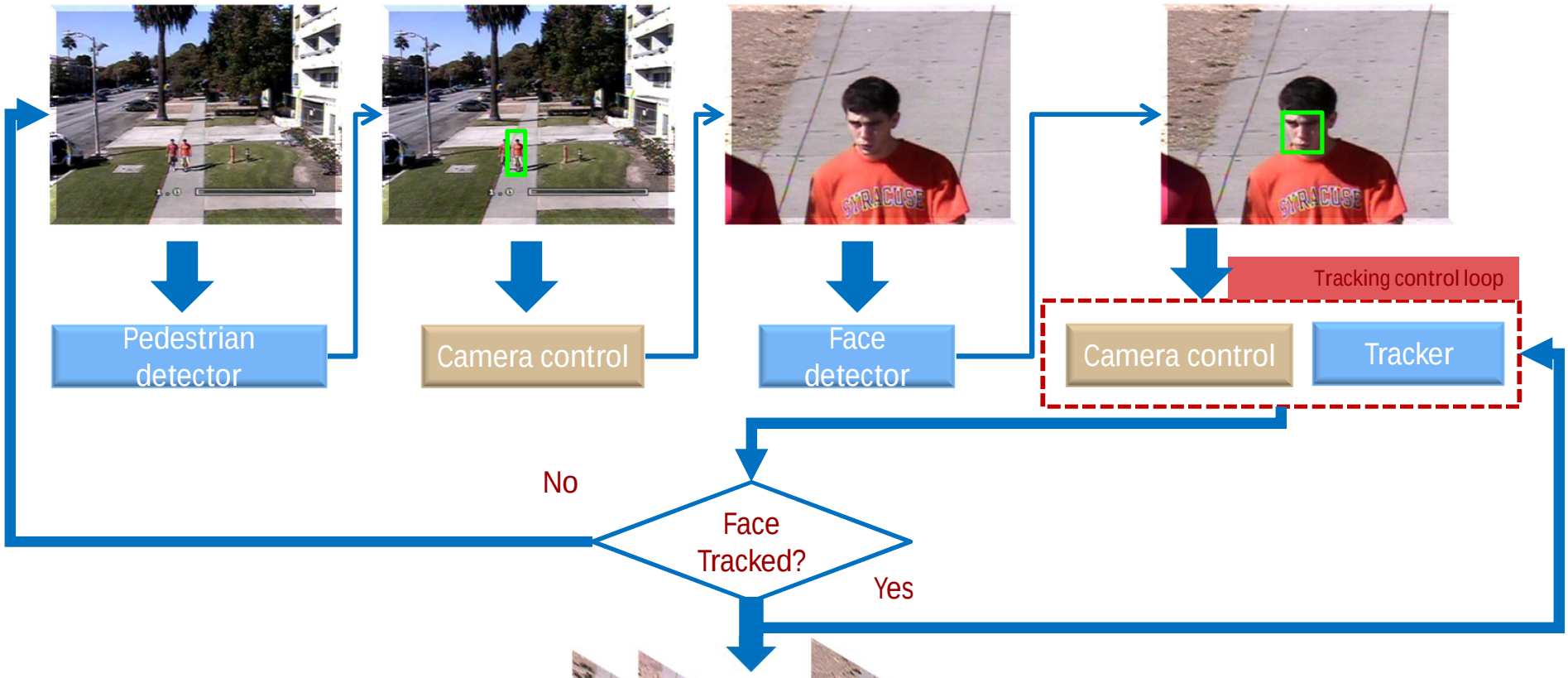
# Challenges

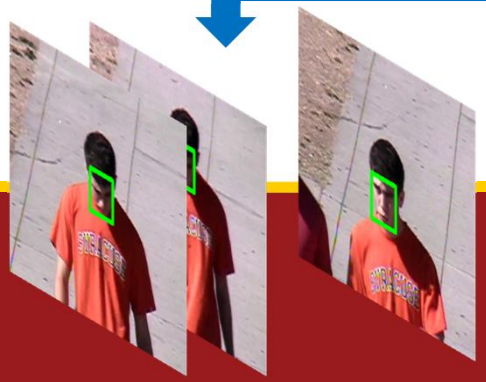
- Practical issues
  - Pedestrians far away (face covers few pixels)




  - In long focal length, people may get out of FOV with a little movement.

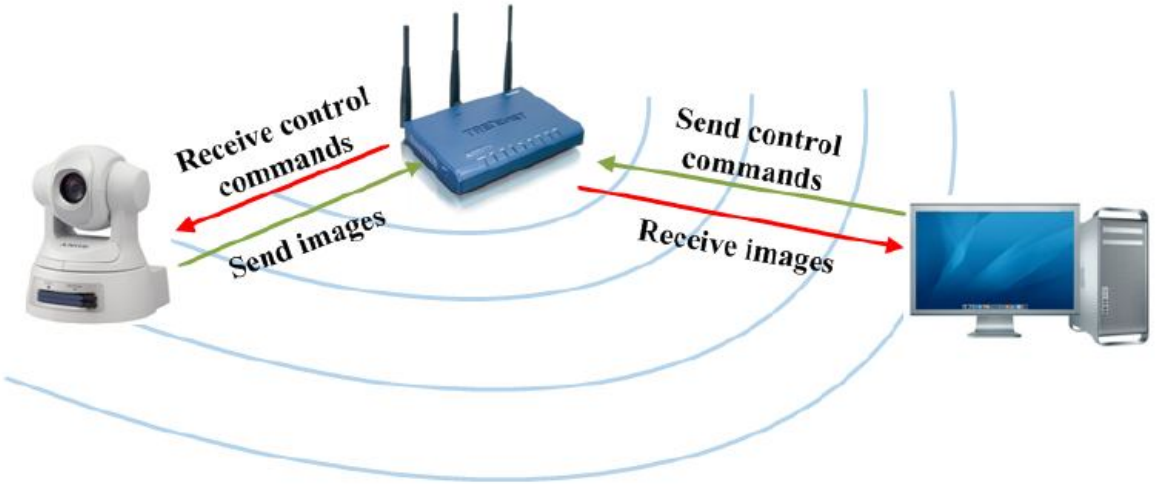
# Overview



Tagged high resolution face sequences

# Experimental setup




- Settings
  - Sony PTZ Network Camera SNC-RZ30N with wireless card
  - 14 levels of speed control for panning and 18 levels for tilting
  - 25x optical zoom, 300x digital zoom
  - Pan angle: -170 to +170 degrees
  - Tilt angle: -90 to +25 degrees

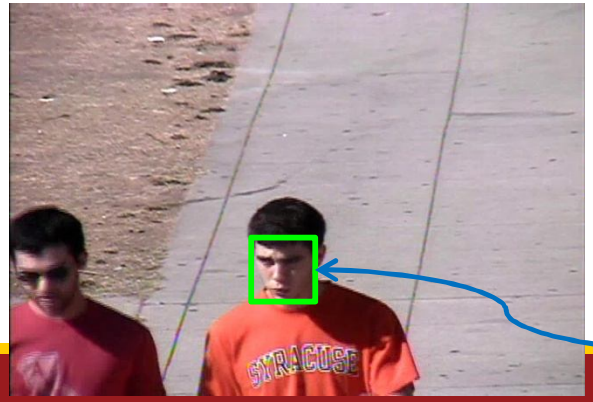
# Results

# Tracking from security PTZ Camera @ USC

Cannot see the face from 100% cropped image

Pedestrian detector

Zooming (11x)



Tracking

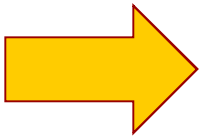
Face track

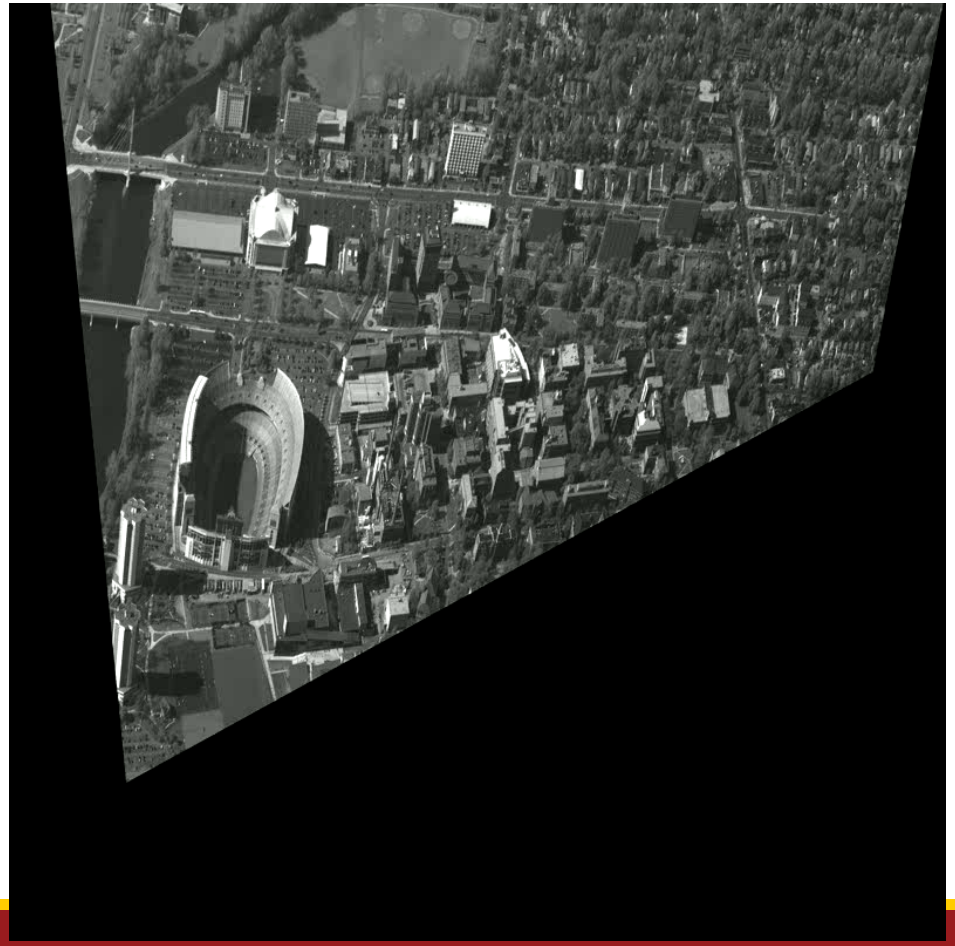
Frontal face detector

# Tracking many objects

- Useful for persistent surveillance
- WAAS (Wide Area Aerial Surveillance)
- Very large images (60MPix-1GPix)
- 2 frames per second

# Video Stabilization





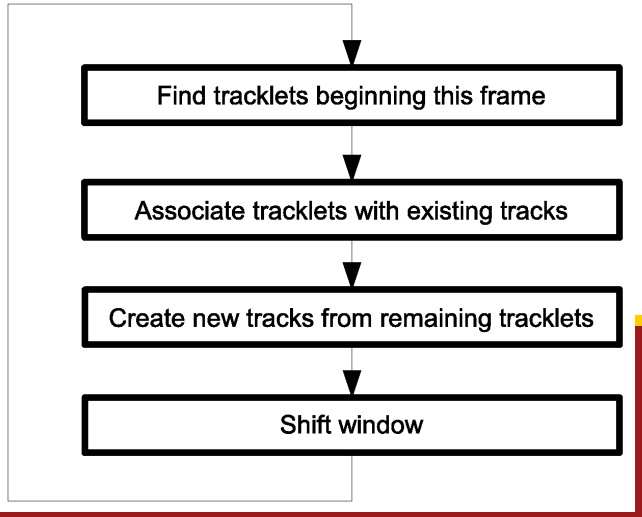
# Video Stabilization Results Close Up

# Tracking

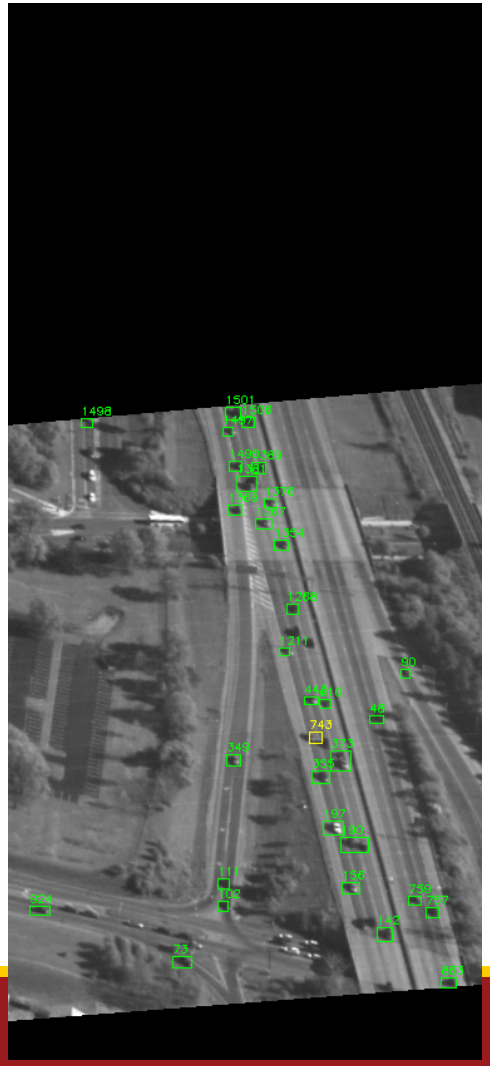
- Motivation
  - Moving objects tell us a lot about the “life” in the geographic area
  - Important for activity recognition
- Challenges
  - Small number of pixels on target
  - Large number of targets

# Approach

- Goal: infer tracklets, each representing one object, over a sliding window of frames
  - 4-8 second window (depends on frame rate)
  - Input: object detections (from background subtraction or otherwise)



# Results (CLIF 2006)



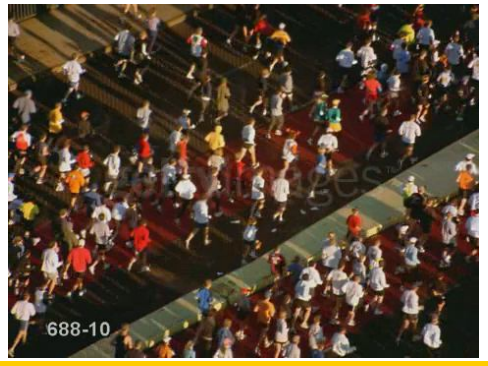
# Tracking Results (CLIF 2006)

| Object Detection Rate | False Alarm Rate | Normalized Track Fragmentation | ID Consistency |
|---|---|---|---|
| 0.72 | 0.04 | 1.01 | 0.84 |

- **Manually generated ground truth**
  - 168 tracks, 80 frames
- **Low track fragmentation**
- **Low false alarm rate**
- **Efficient**
  - > 40 objects tracked at 2 fps
- **Comparison with MCMC tracker (Yu 2009)**
  - Did not converge to a reasonable solution
    - Requires good initialization
  - Does not scale to our domain

# Tracking VERY MANY Objects

- With the development of surveillance system, we will pay more and more attention to analyzing people in crowded scenes. (Sports, political gathering, etc.)


# Crowded Scenes

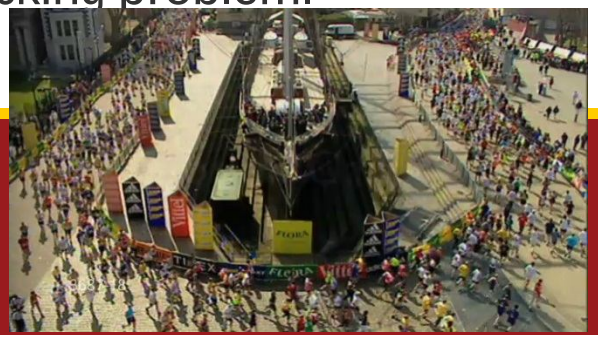
- Challenges
  - Hundreds of similar objects
  - Cluttered background
  - Small object size
  - Occlusions

❑ Detect-then-track method fails: appearance based detector and background modeling based motion blob detector fail

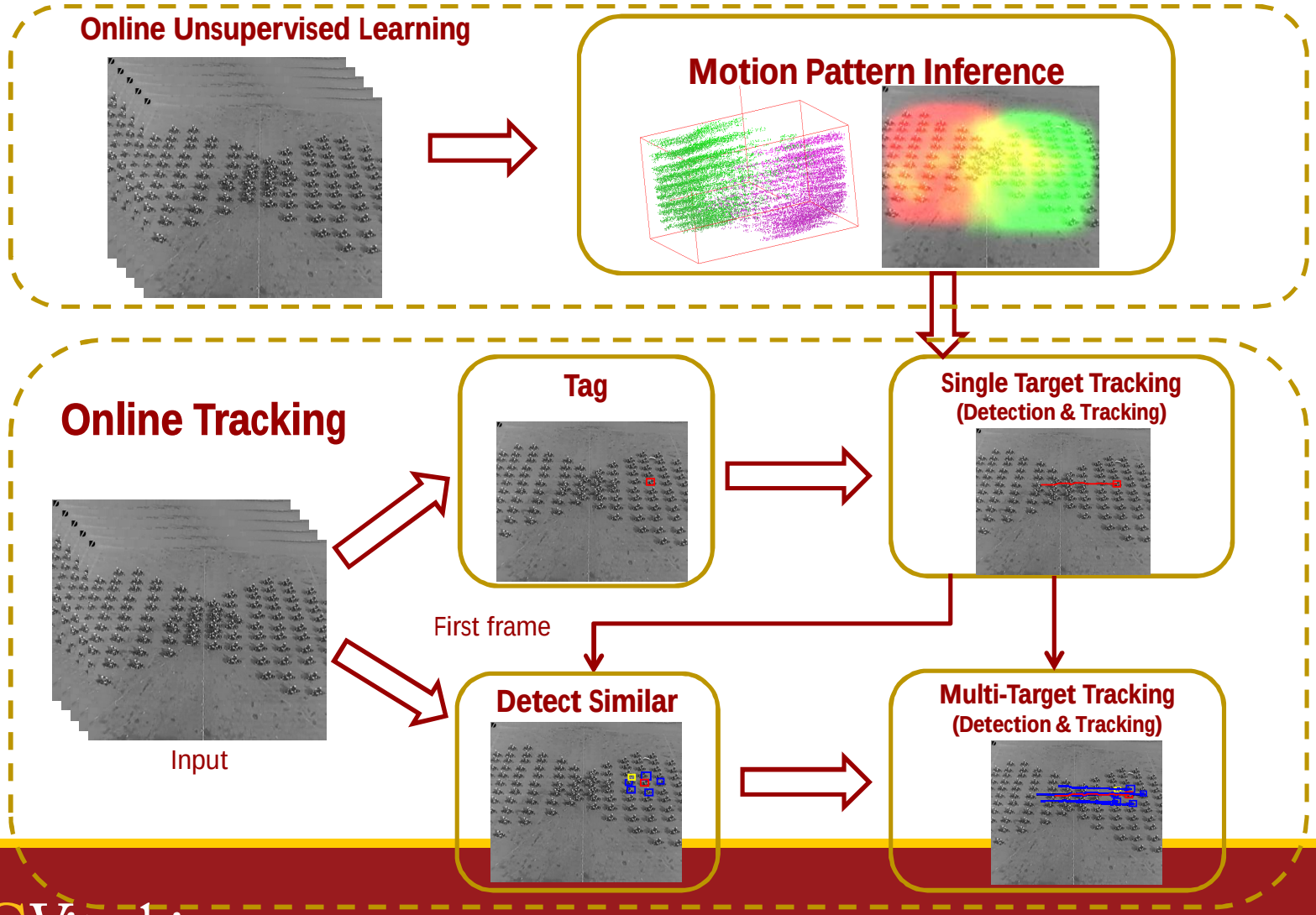
# Tracking Using Motion Patterns for Very Crowded Scenes

We solve the problem of tracking in *structured* crowded scenes using Motion Structure Tracker (MST)

- MST is a combination of visual tracking, motion pattern learning and multi-target tracking.
- In MST, tracking and detection are performed jointly, and motion pattern information is integrated in both steps to enforce scene structure constraint.
- MST is initially used to track a single target, and further extended to solve a simplified version of the multi-target tracking problem.



# An Overview of Motion Structure Tracker



# Motion Structure Tracker for Single Target Tracking

- Tag & Track
- Results for Temporally Stationary Scenes (motion pattern do not change with time)

Marathon-1

Marathon-2


Marathon-3

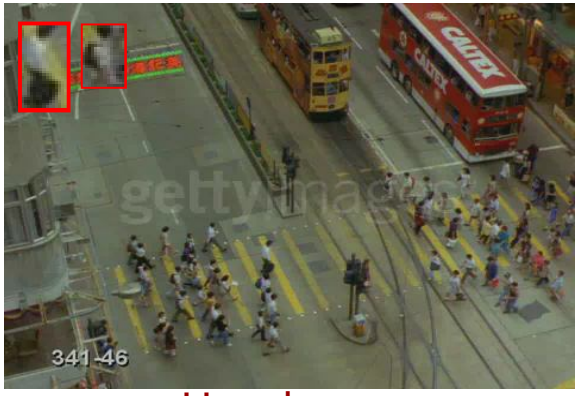
| Sequence | Method | ATR | ACLE |
|---|---|---|---|
| Marathon-1 | IVT Tracker<br>P-N Tracker<br>**Ours** | 35.21%<br>56.16%<br>**81.40%** | 62.8<br>35.1<br>**6.7** |
| Marathon-2 | IVT Tracker<br>P-N Tracker<br>**Ours** | 33.47%<br>68.60%<br>**73.12%** | 86.5<br>56.4<br>**28.5** |
| Marathon-3 | IVT Tracker<br>P-N Tracker<br>**Ours** | 40.03%<br>67.16%<br>**92.08%** | 64.1<br>33.9<br>**4.8** |

ATR : Average Track Ratio

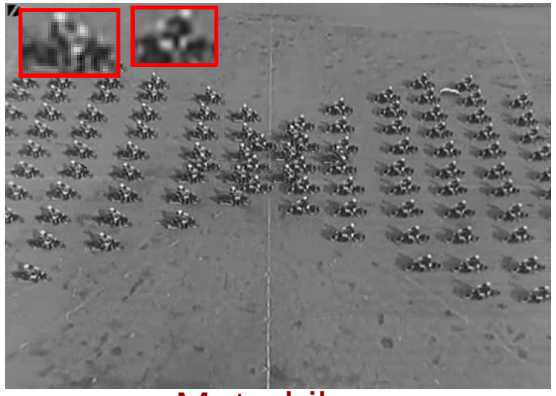
ACLE: Average Center Location Error (ACLE)

# Motion Structure Tracker for Single Target Tracking

- Results for Temporally Non-Stationary Scenes (motion pattern change with time)



Hongkong



Motorbike

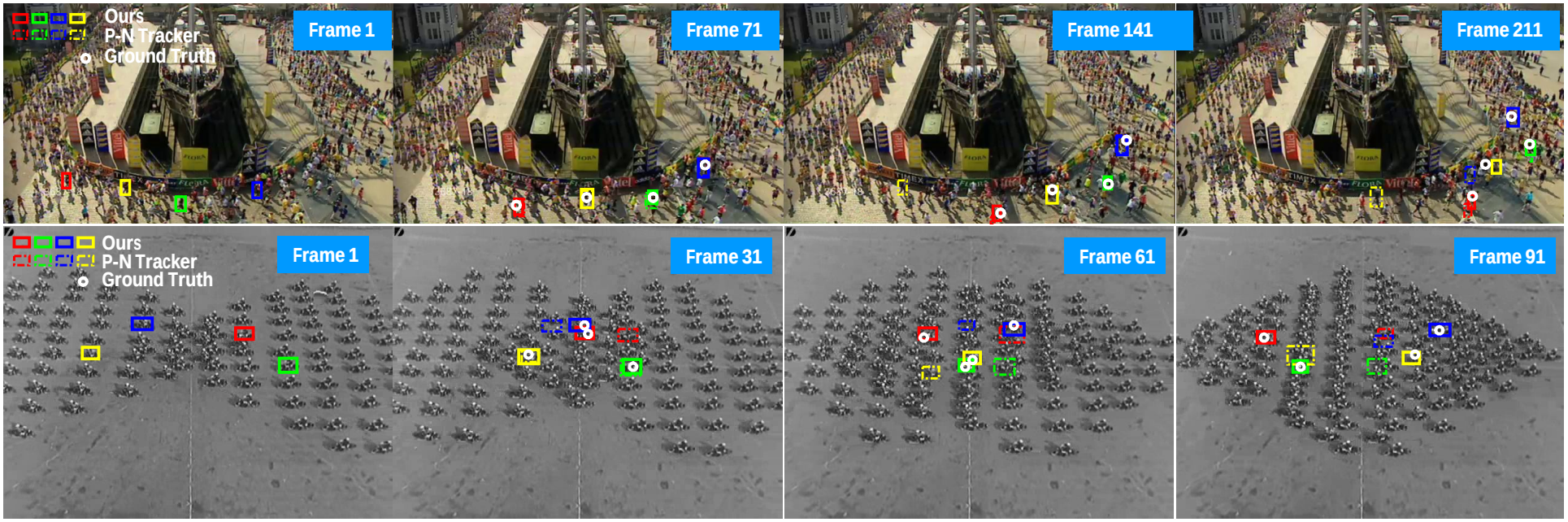
| Sequence | Method | ATR | ACLE |
|---|---|---|---|
| Hongkong | IVT Tracker | 27.63% | 58.9 |
| | P-N Tracker | 39.58% | 42.3 |
| | **Ours** | **62.31%** | **28.5** |
| Motorbike | IVT Tracker | 31.56% | 69.7 |
| | P-N Tracker | 47.22% | 55.4 |
| | **Ours** | **90.75%** | **5.6** |

ATR : Average Track Ratiio

ACLE: Average Center Location Error (ACLE)

# Motion Structure Tracker for Multi-Target Tracking

- Once a user labels a target in the first frame, find similar objects and track all of them




Examples of tracking results comparison. First row: temporally stationary scenes. Second row: temporally non-stationary scenes.

# Tracking Using Motion Patterns for Very Crowded Scenes

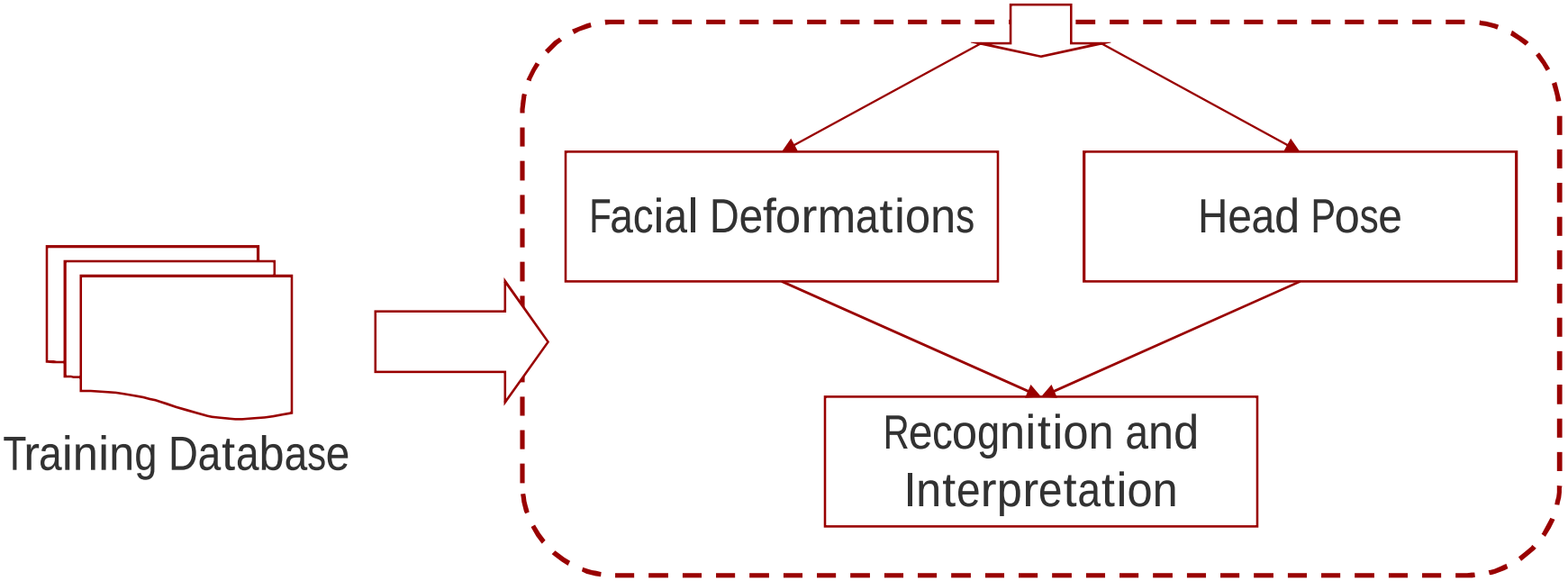
CVPR 2012 Submission

Paper ID 924

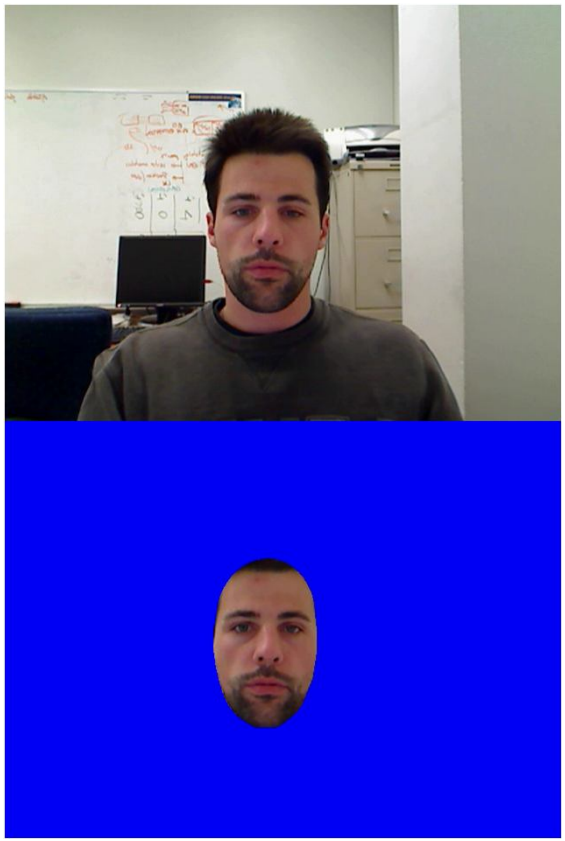
# Expression Analysis

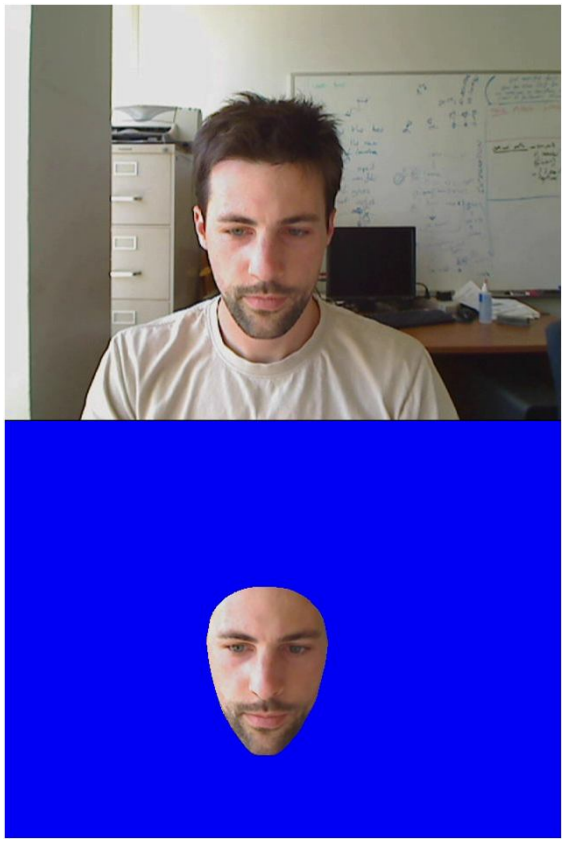
- Understanding facial gestures
  - By analyzing facial motions
  - Facial motion induces detectable appearance changes
- Two classes of facial motions
  - Global, rigid head motion
    - From head pose variation
    - Indicate subject's attention
  - Local, nonrigid facial deformations
    - From facial muscle activation
    - Indicate subject's expression

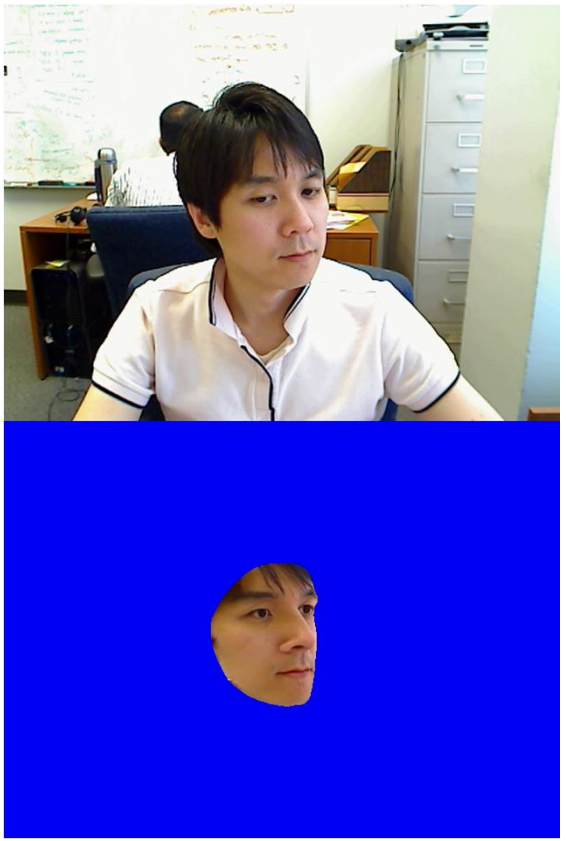
# Overview


Facial Deformations
Head Pose
Recognition and Interpretation
Training Database

# Results ( Rigid tracking, real-time)





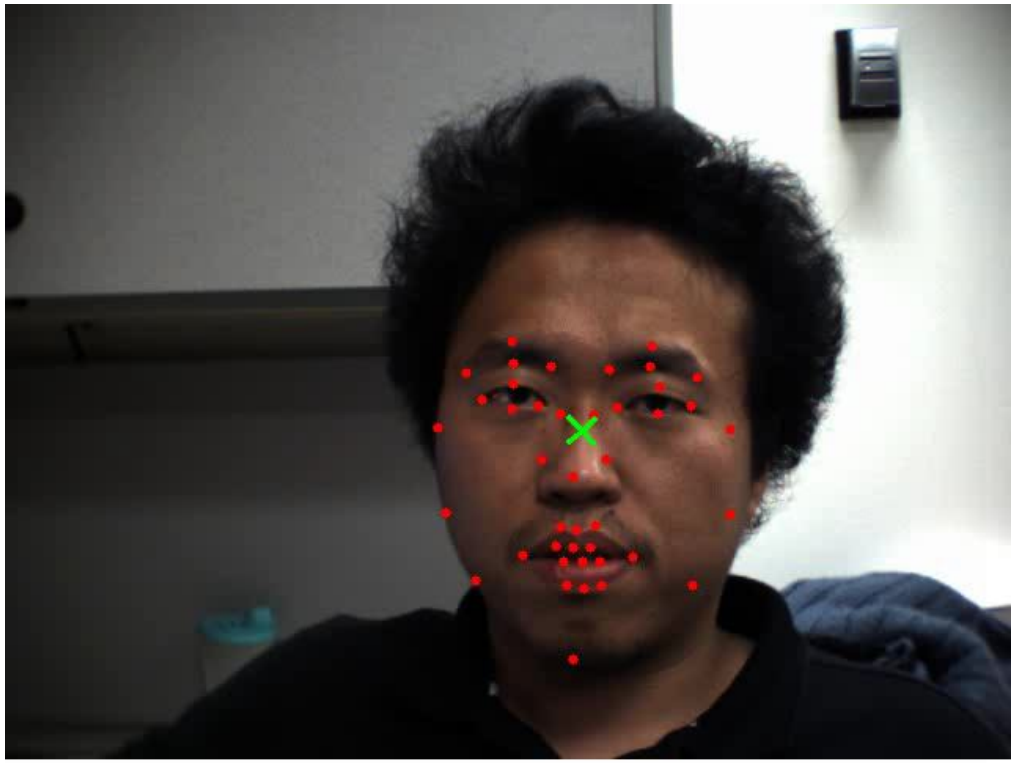


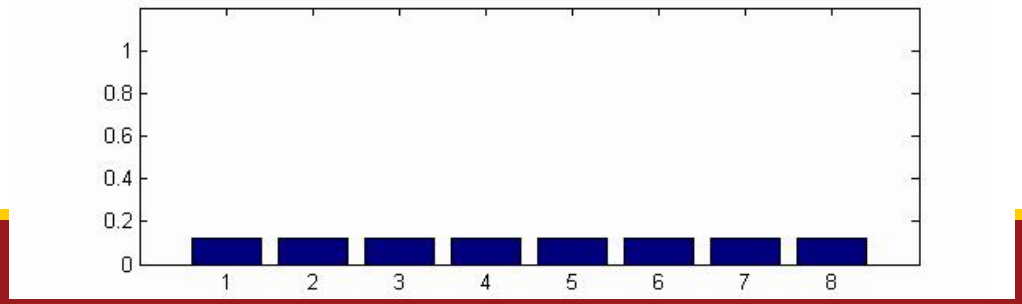
Rotation, translation, & scale

Fast motion

Live webcam

# Expression Analysis





# Summary

- Tracking is a multi-faceted problem
- Many axes of complexity
  - Resolution
  - Number of objects
  - Type of motion
  - …
- Significant progress being achieved