



Computer Vision

Mubarak Shah
shah@eecs.ucf.edu



Computer Vision

- The ability of computers to see.
 - Image Understanding
 - Machine Vision
 - Robot Vision
 - Image Analysis
 - Video Understanding

A picture is worth a thousand words.



Gali Tibbon / AFP

A word is worth a thousand
pictures.



A HUNT



Image

- 2-D array of numbers (intensity values, gray levels)
 - Gray levels 0 (black) to 255 (white)
 - Color image is 3 2-D arrays of numbers
 - Red
 - Green
 - Blue
- Resolution (number of rows and columns)
 - 128X128
 - 256X256
 - 512X512
 - 640X480

34	23	58	89	106	97	89	83	83	81
97	39	23	67	75	89	89	89	89	81
129	73	26	67	67	58	75	81	81	75
171	147	97	106	64	7	23	58	81	83
56	89	147	155	114	73	48	58	73	81
23	64	115	148	155	114	48	26	48	73
23	56	74	81	73	64	73	81	89	89
73	56	45	62	57	56	73	81	82	82
97	64	81	103	106	97	89	82	82	82
97	81	89	86	89	97	81	78	82	97

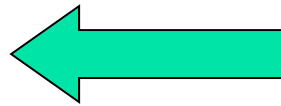




Image Formats

- TIF
- PGM
- PBM
- GIF
- JPEG



Video

- Sequence of frames
- 30 frames per second

- Formats
 - AVI
 - MPEG
 - Quick Time



Video Clip



Sequence of Images

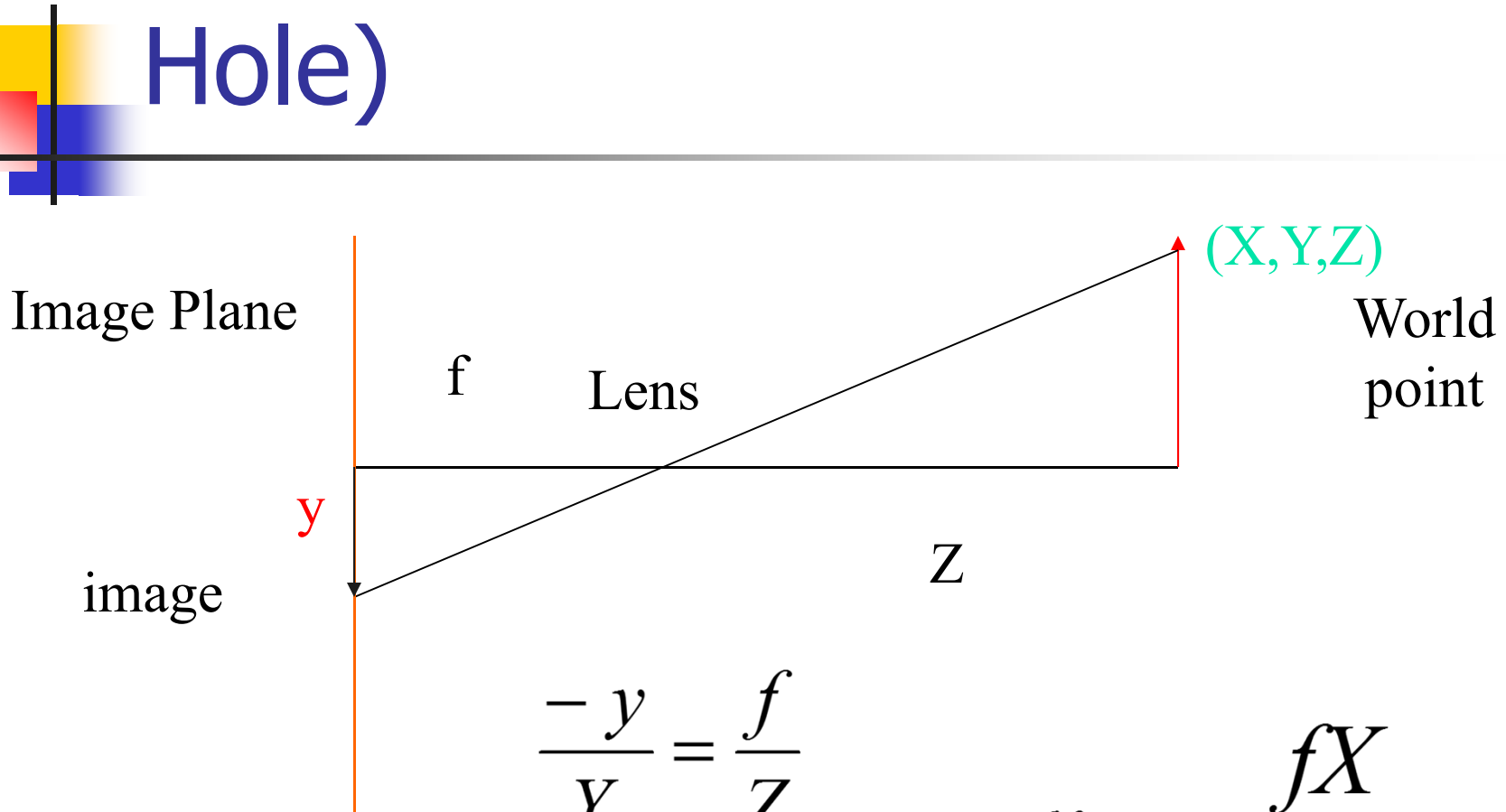




Image Formation

- Light Source
- Camera (extrinsic and intrinsic parameters)
- Scene (Surface reflectance, Surface shape)

Perspective Projection (Pin Hole)



$$\frac{-y}{Y} = \frac{f}{Z}$$

$$y = -\frac{fY}{Z}$$

$$x = -\frac{fX}{Z}$$

Orthographic Projection

Image Plane

image

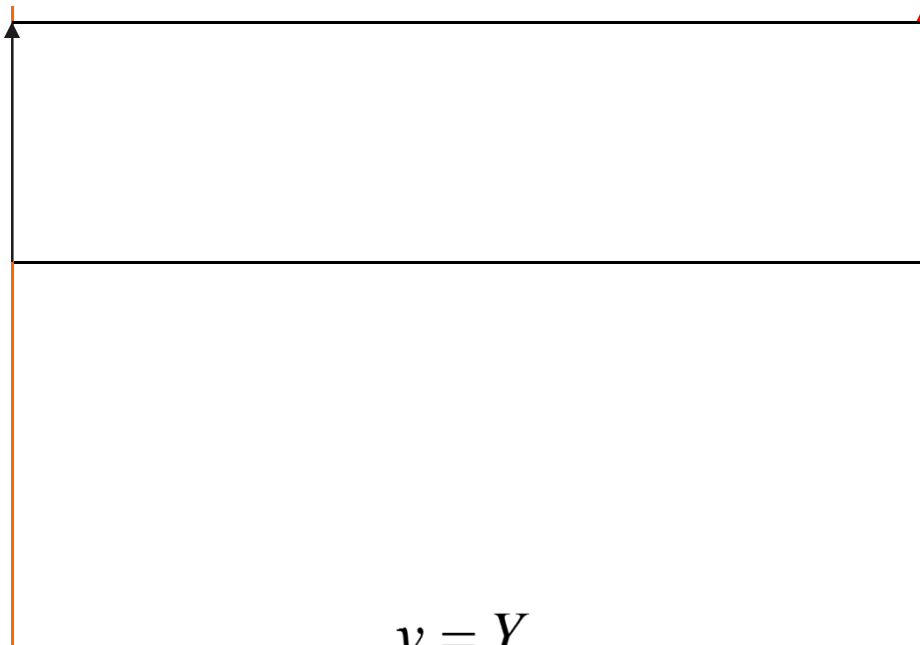
y

$$y = Y$$

$$x = X$$

(X, Y, Z)

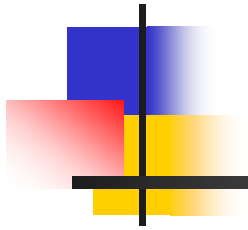
World
point





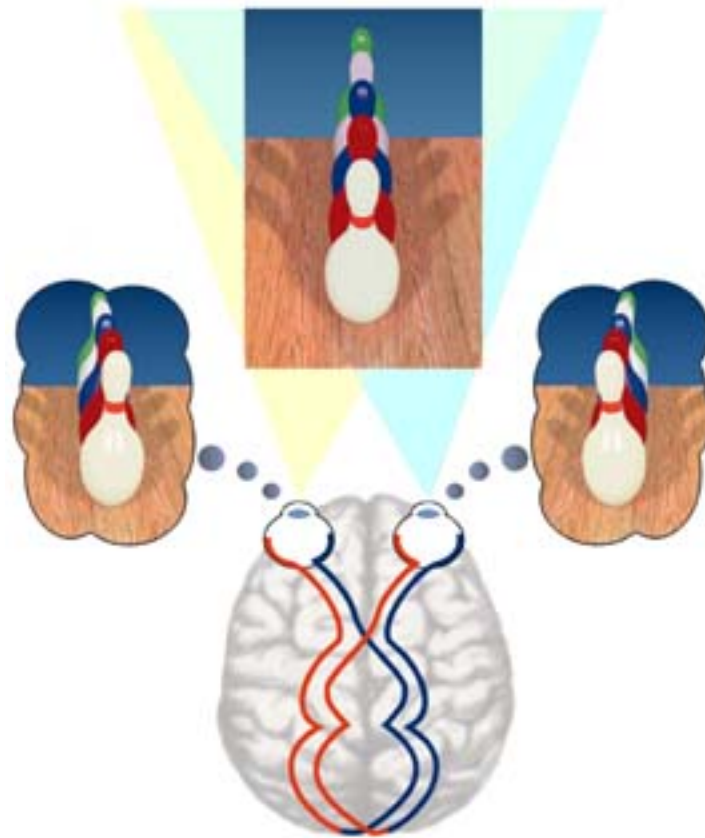
Shape from X

- Recover 3-D shape from 2-D image(s)
 - Stereo
 - Motion
 - Shading
 - Texture
 - Contours



Stereo

<http://www.vision3d.com/stereo.html>

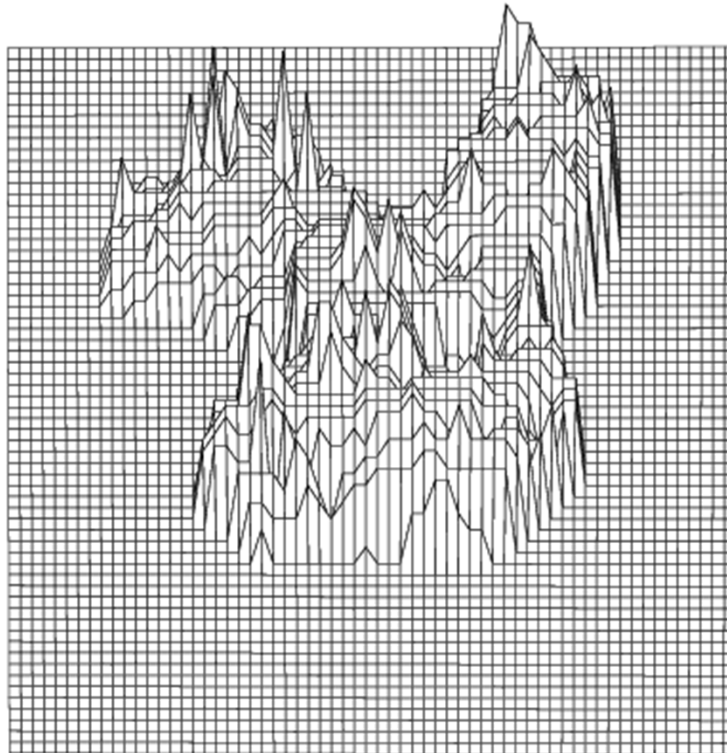


Renault Stereo Pair





Depth Map

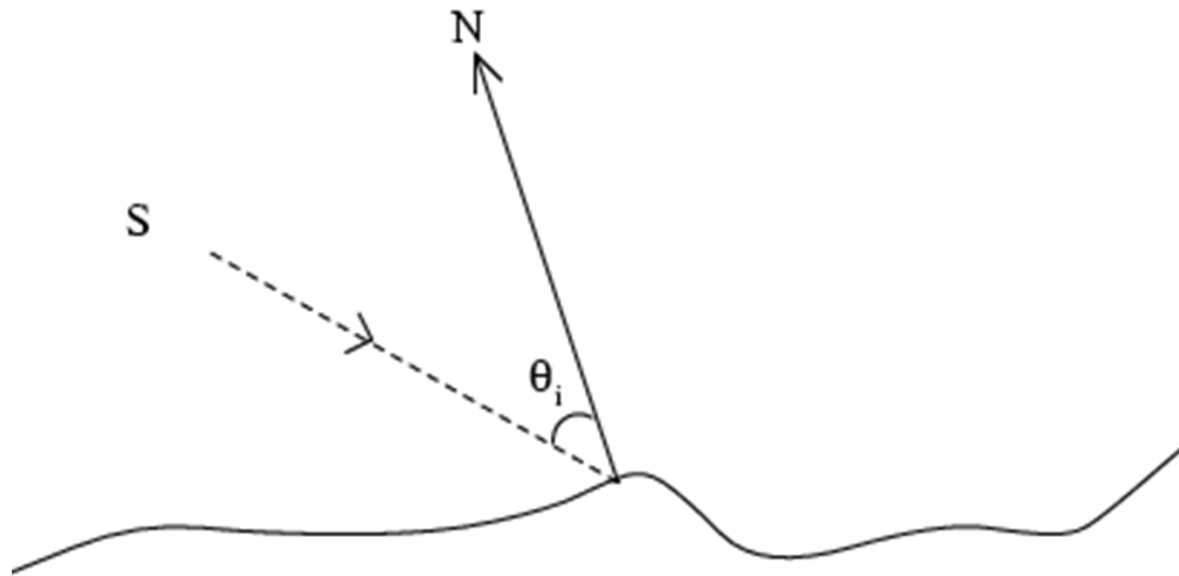


Shape from Shading





Lambertian Model

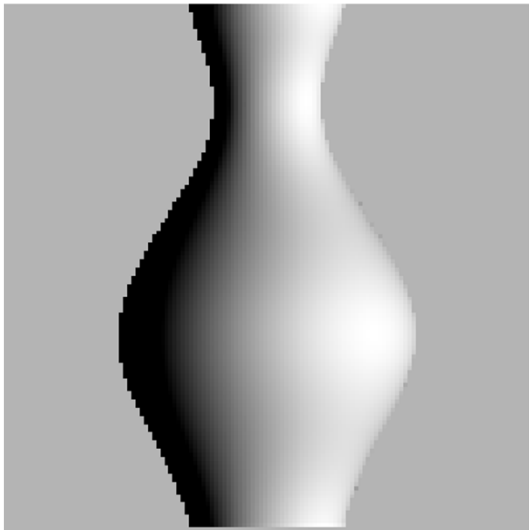


$S=L$, light
source

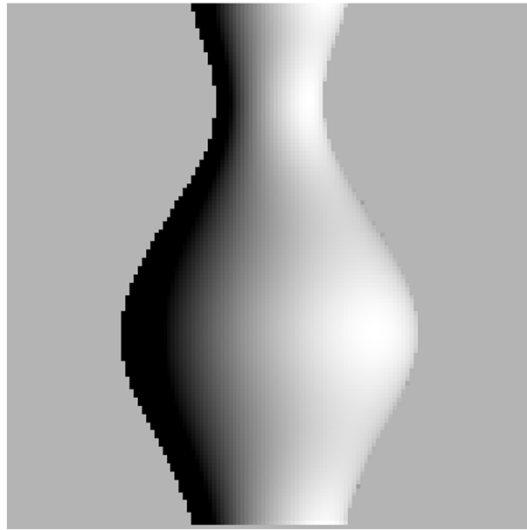
$$I=S \cdot N$$



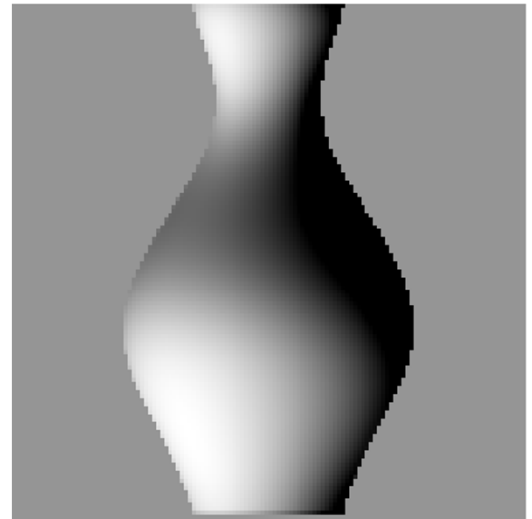
Vase



$(1, 0, 1)$

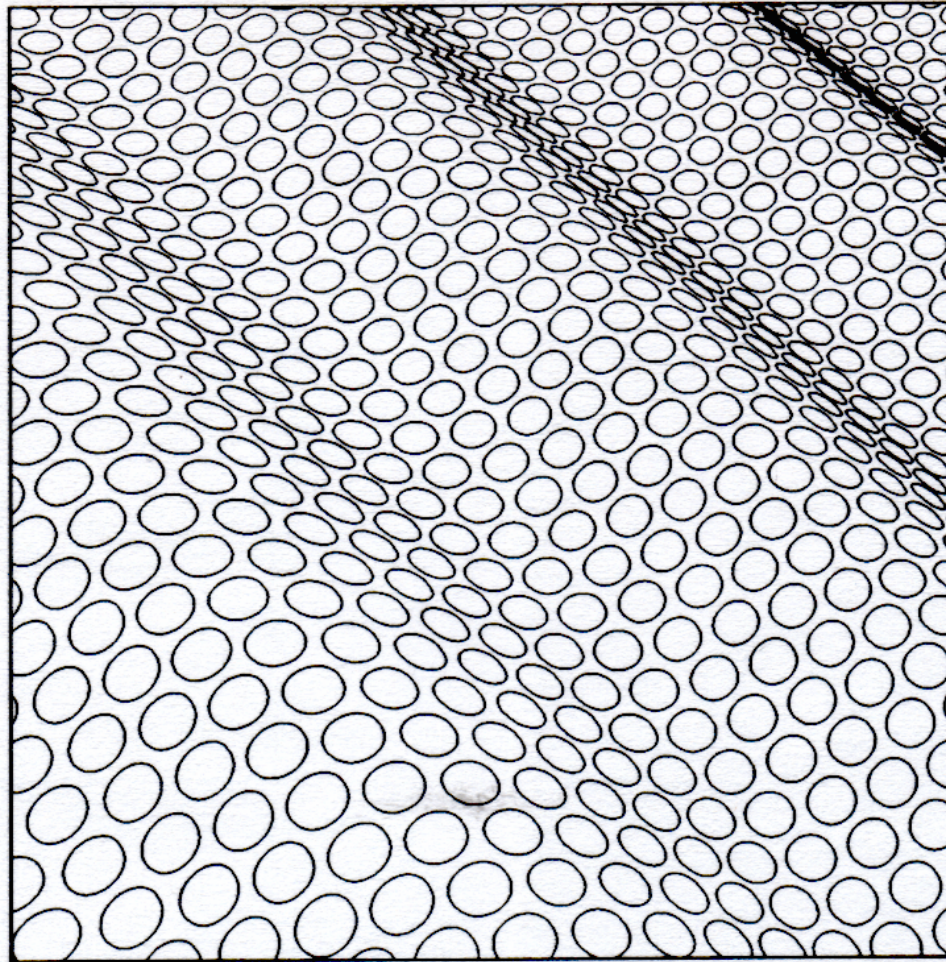


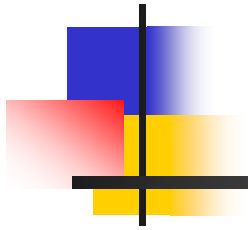
$(-1, 1, 1)$



$(-1, -1, 1)$

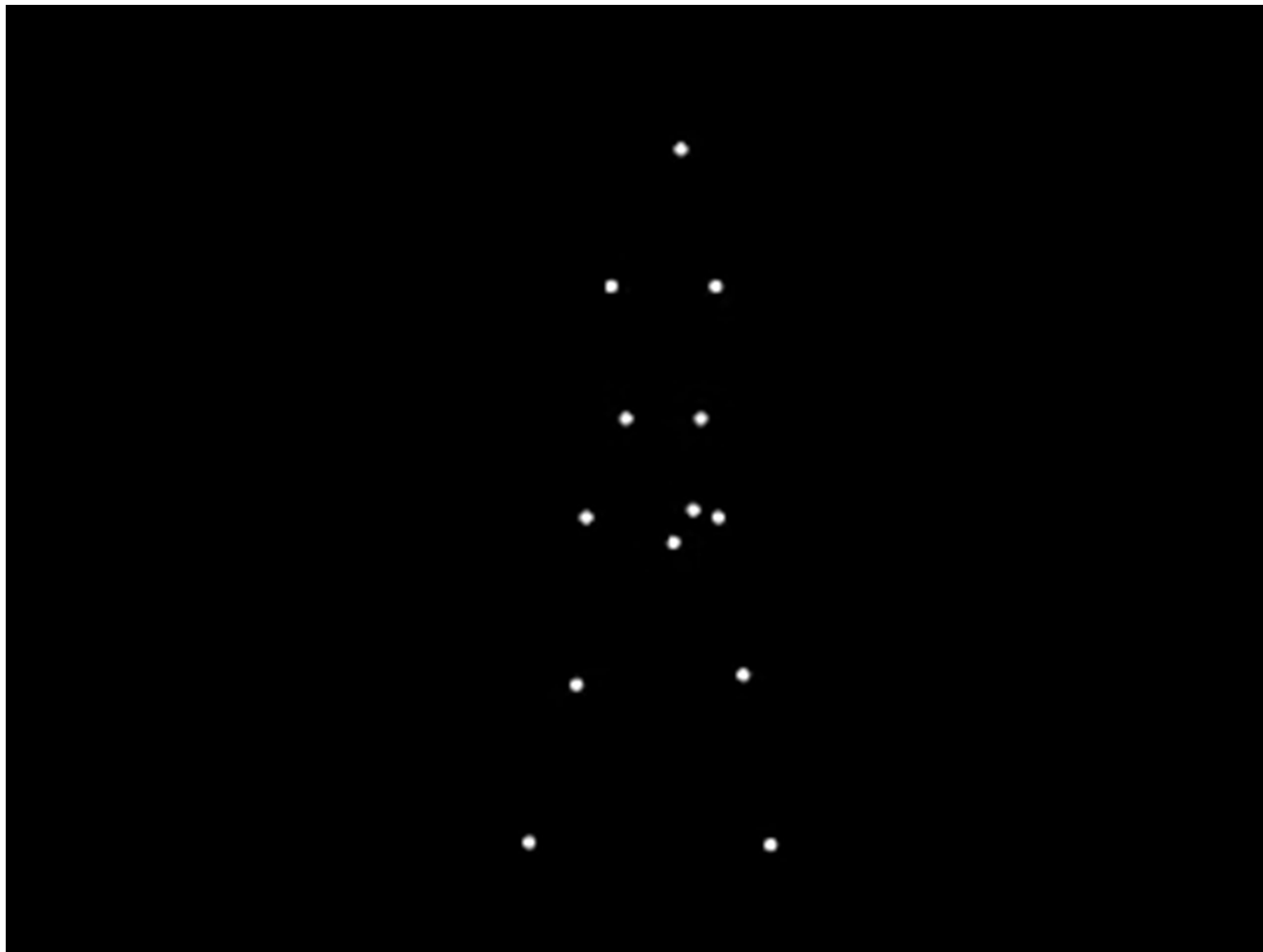
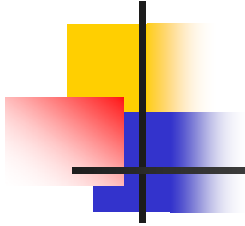
Shape from Texture





Visual Motion

Shape from Motion: Moving Light Display



Shape from Motion



(a)



(b)



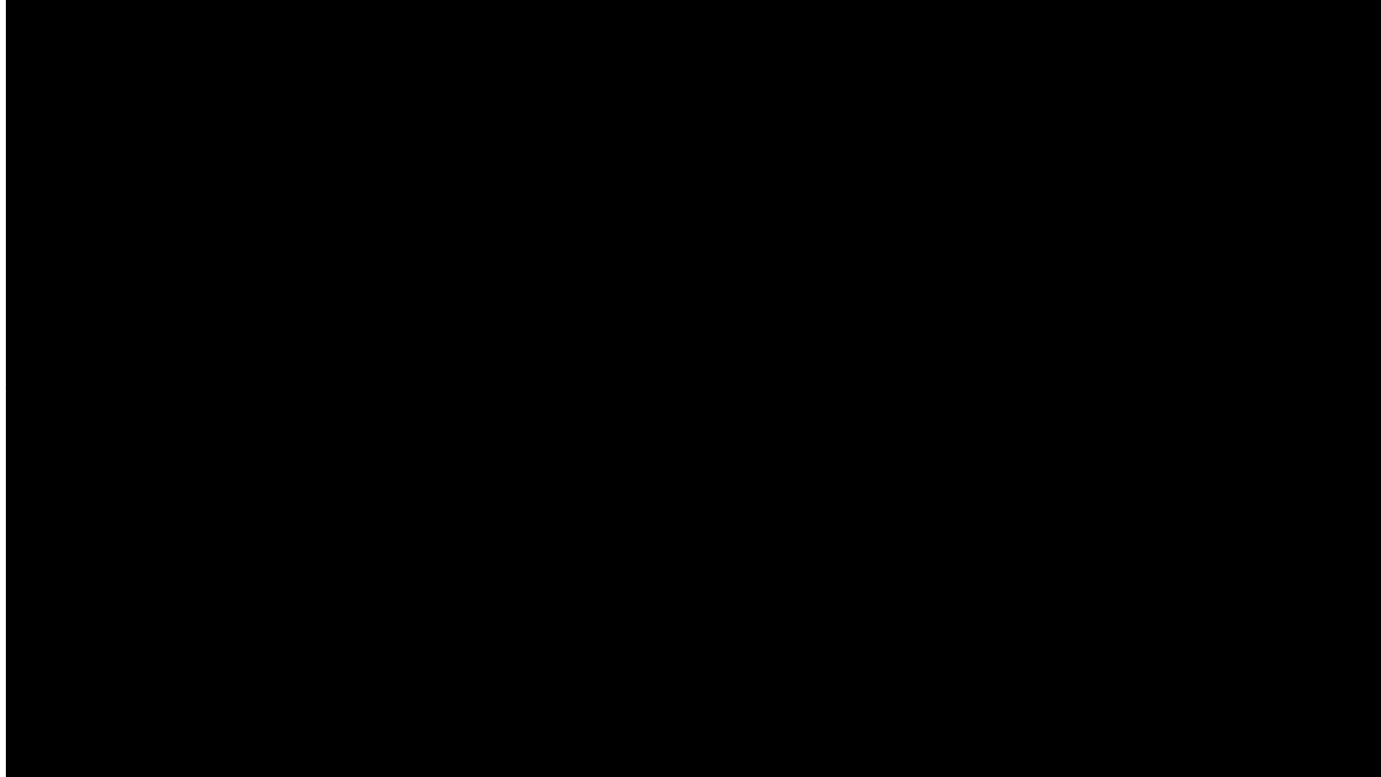
(c)



(d)



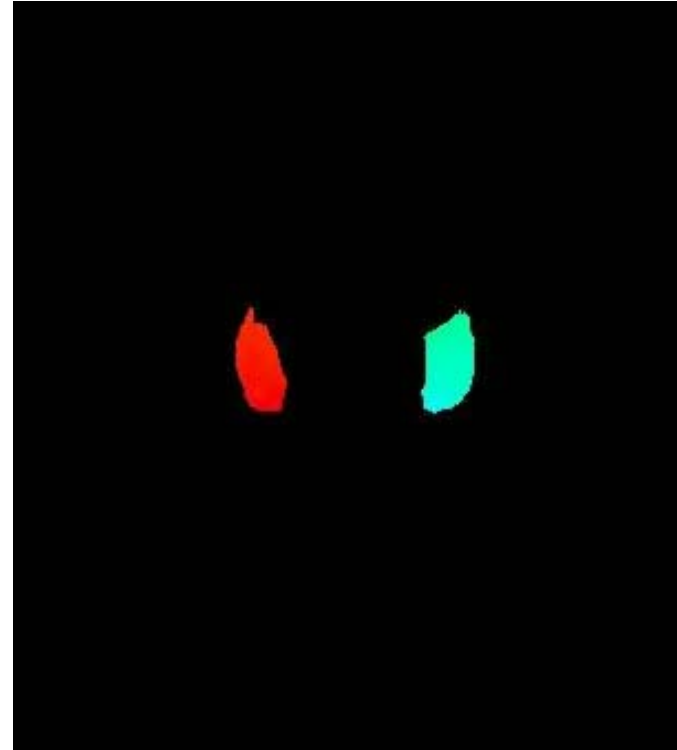
Photosynth



Sequence



Raw Optical flow





Video Clip & Mosaic





Applications of Computer Vision

- Face Recognition
- Object Recognition
- Video Surveillance and Monitoring
 - Object detection, tracking and behavior analysis
- Remote Sensing: UAVs
- Robotics
- Computer Graphics



Object Recognition

Finding People in images

Problem 1: Given an image I

Question: Does I contain an image of a person?

"Yes" Instances



Phil Noble / AP



Mike Hewitt / Allsport



Patrick Gardin / AP



Andy Barron / Reno Gazette-Journal



Sydney Morning Herald

"No" Instances



Eric Miller / Reuters



Mark Garkfinkel / The Boston Herald



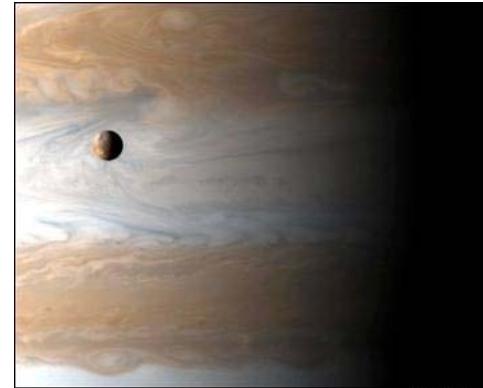
Jeff J. Mitchell / Reuters



Monroe County Sheriff's Department / Newsmakers

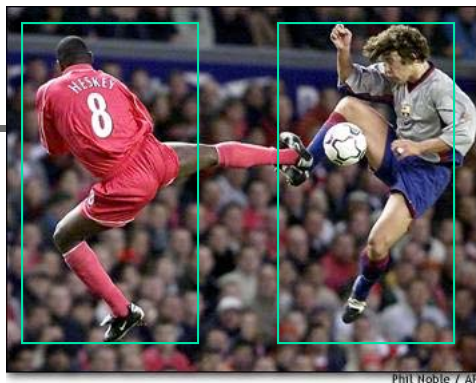


Uno Andersson / AP



NASA via AFP

Localize People (Human Detection)



Phil Noble / AP



Mike Hewitt / Allsport



Patrick Gardin / AP

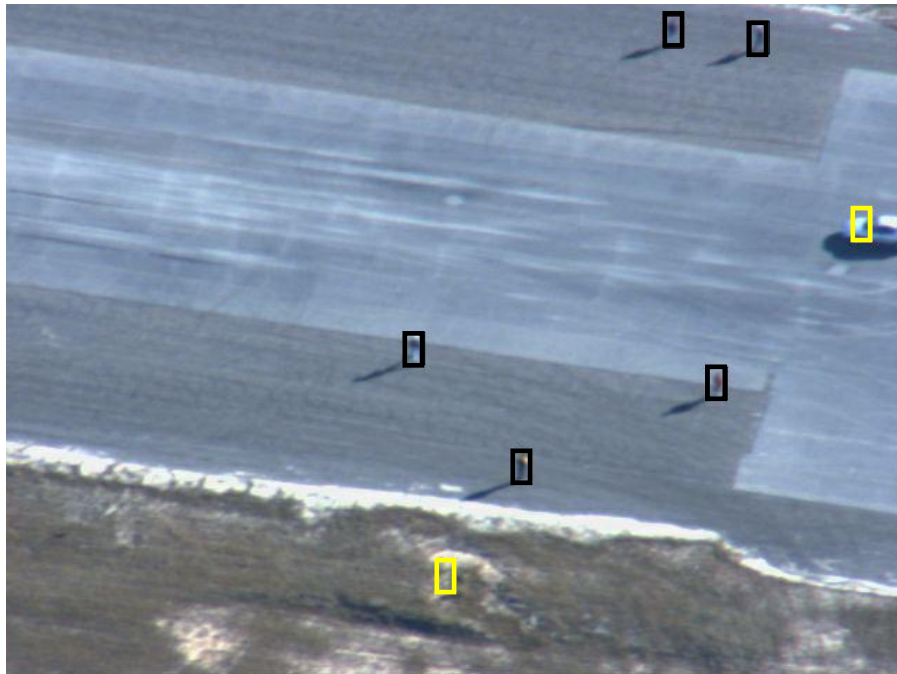


Andy Barron / Reno Gazette-Journal



Sydney Morning Herald

Human Detection



Individuals within small groups of people

Airplanes

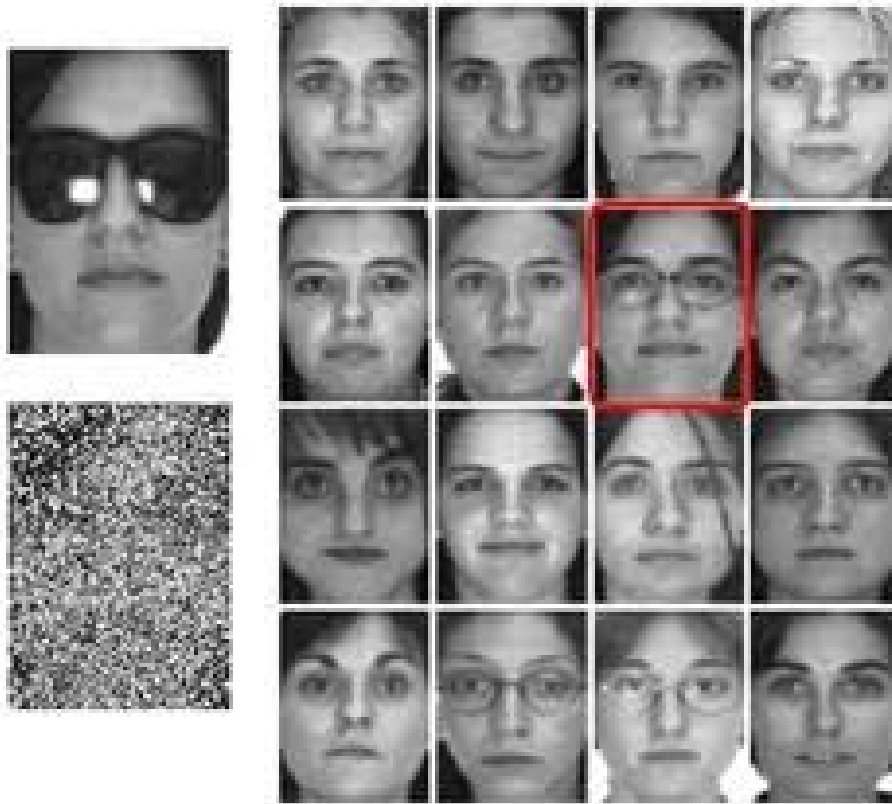


Motor Cycles





Face Recognition



FACIAL EXPRESSIONS

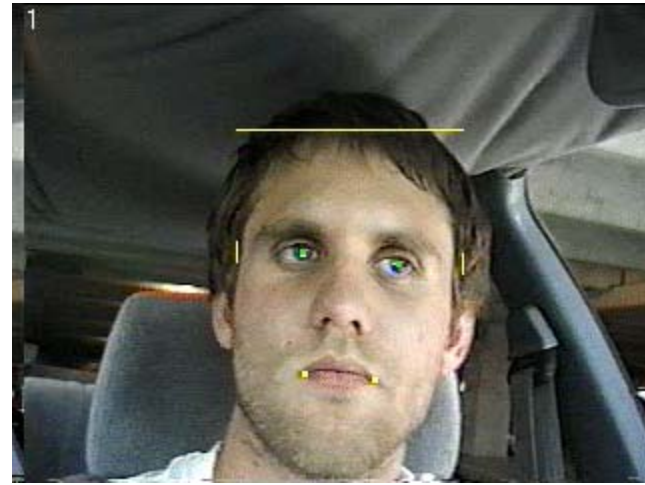
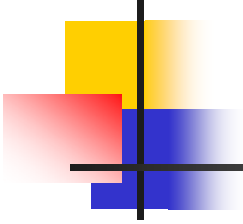


RAISE EYE BROWS

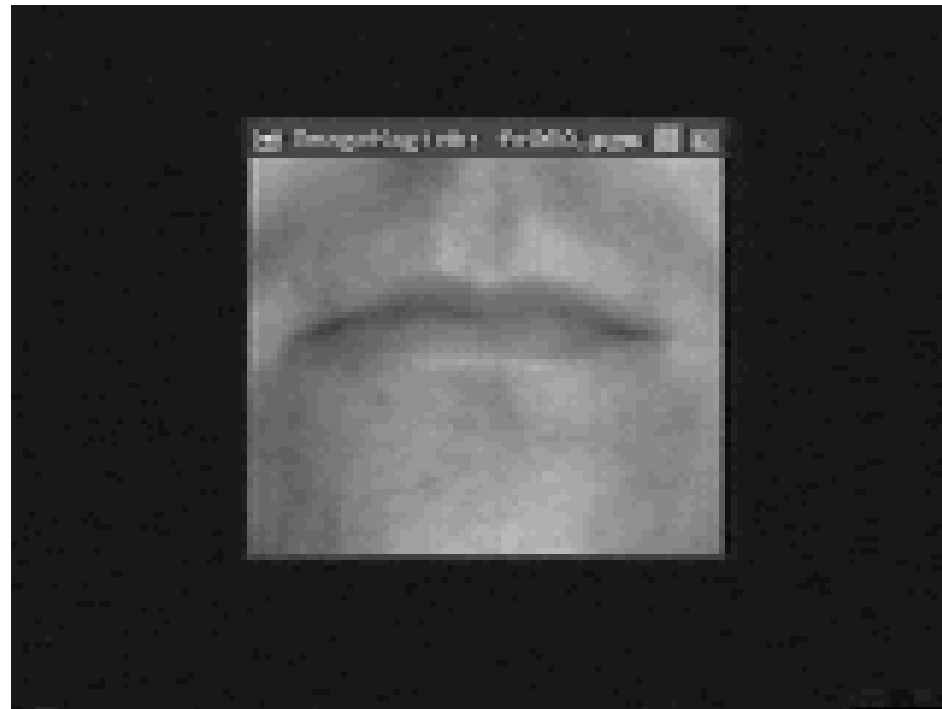


SMILE

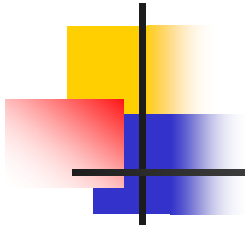
Detecting Driver Alertness



Lipreading



Video Surveillance and Monitoring



Object detection



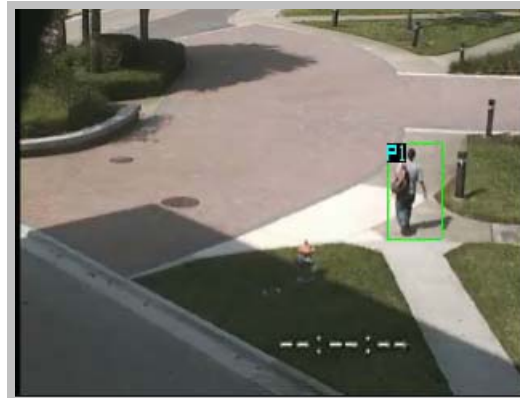
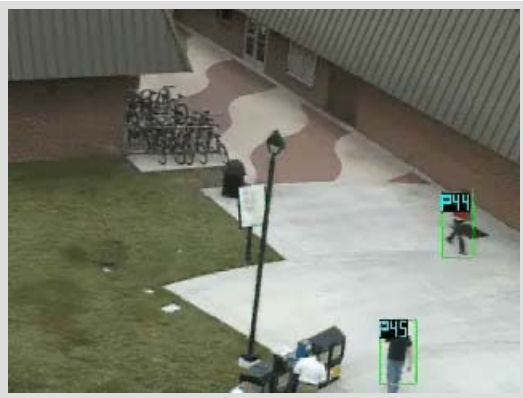
Object tracking

Object categorization
and classification



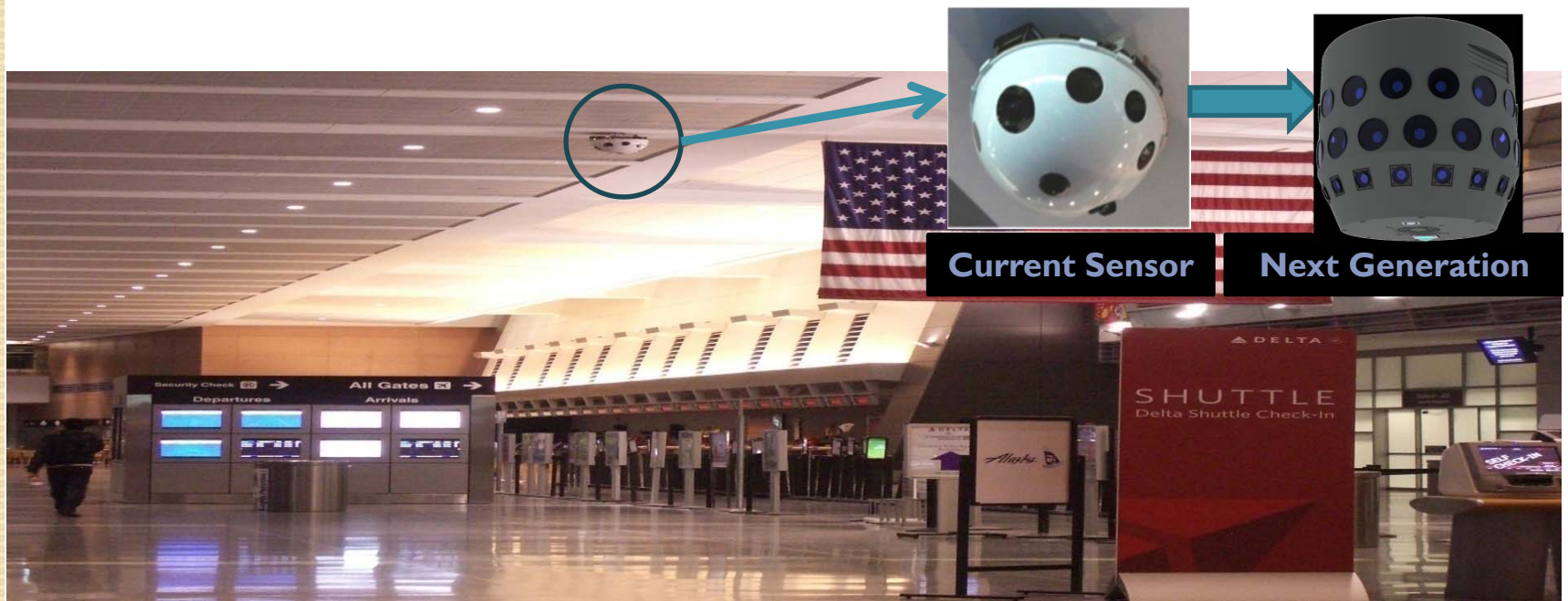
Event or Activities
Recognition

- Automated Surveillance System (Detection & Tracking)



NONA: Project Overview

- Part of the WAS (wide area surveillance) project executed by the Homeland Security Advanced Research Project Agency (HSARPA)
- **Current Sensor**
 - 8 high-resolution cameras
 - provide a 100 mega-pixel, 360° field of view
 - frame rate: 5 frames per second
- **Next Generation**
 - 48 cameras with significantly higher resolution
 - smaller size



NONA System—Airport Sequence I





UAV: Unmanned Aerial Vehicle

UAVs: Unmanned Aerial Vehicles (Drones)



Global Hawk



Predator



Microdrone

KINGFISHER AEROSTAT BALLOON

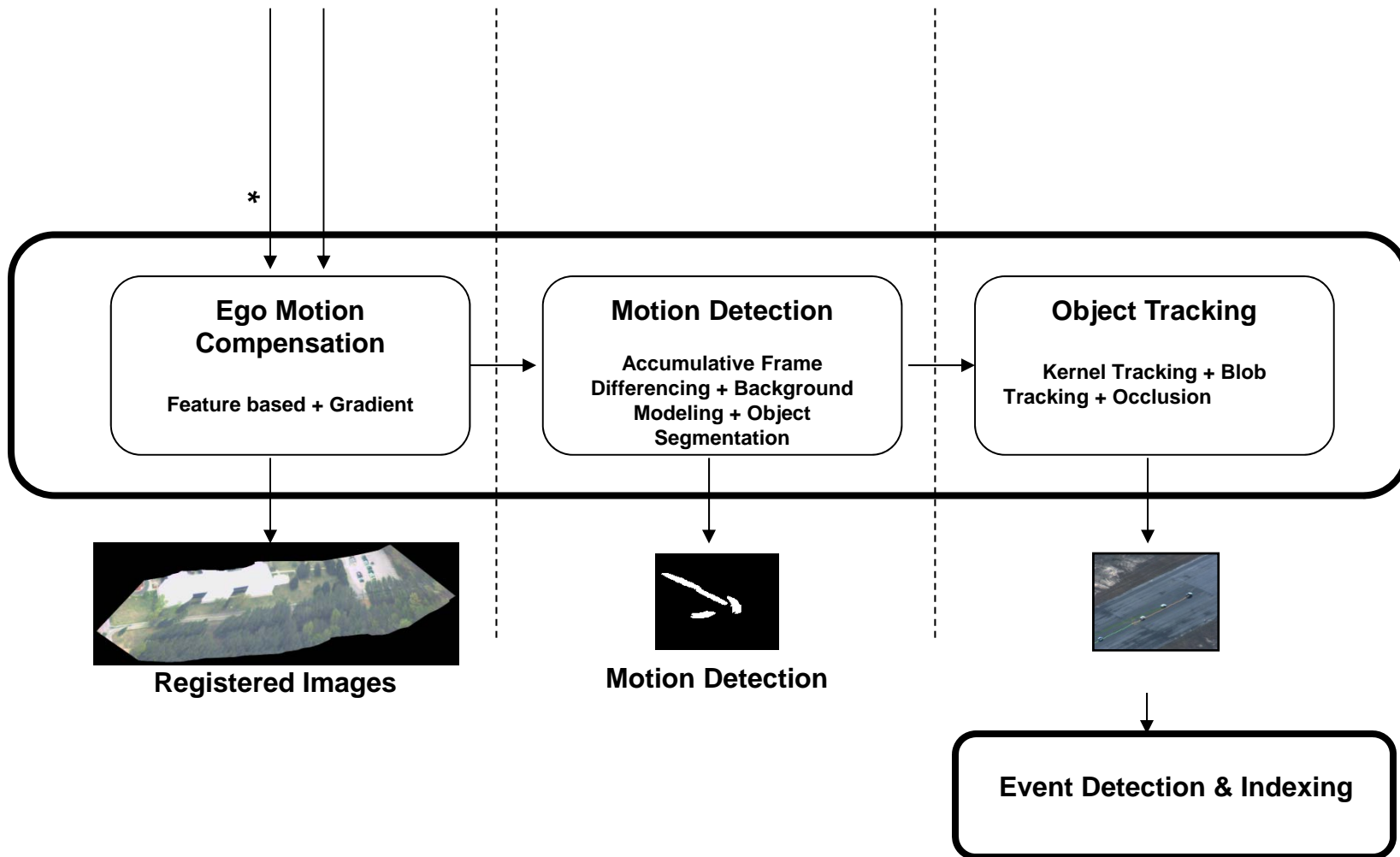


8/21/2012

Computer Vision Lab, UCF

46

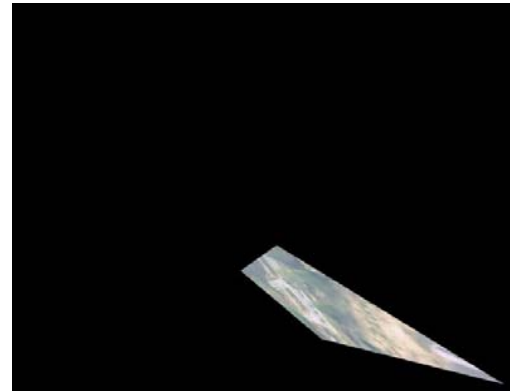
COCOA – System Flow



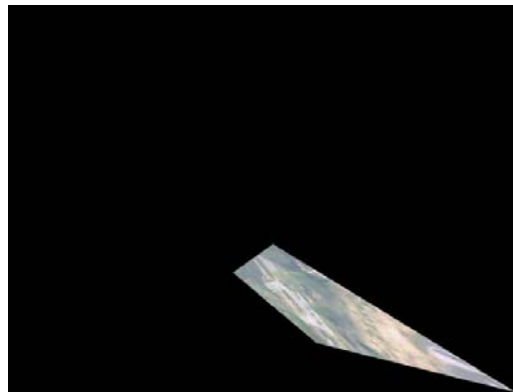
Registration Result - I



Aerial Video - EO



Mosaic

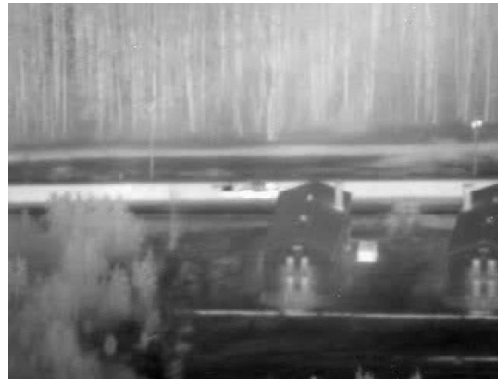


Alignment



Mask

Registration Result - II



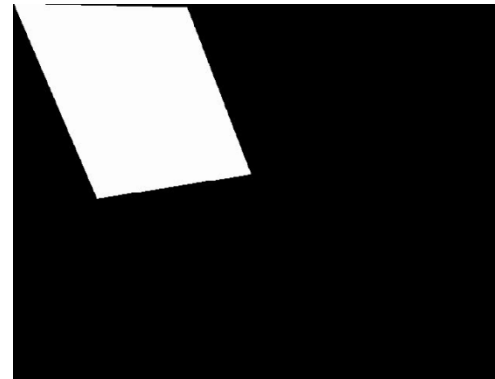
Aerial Video - IR



Mosaic

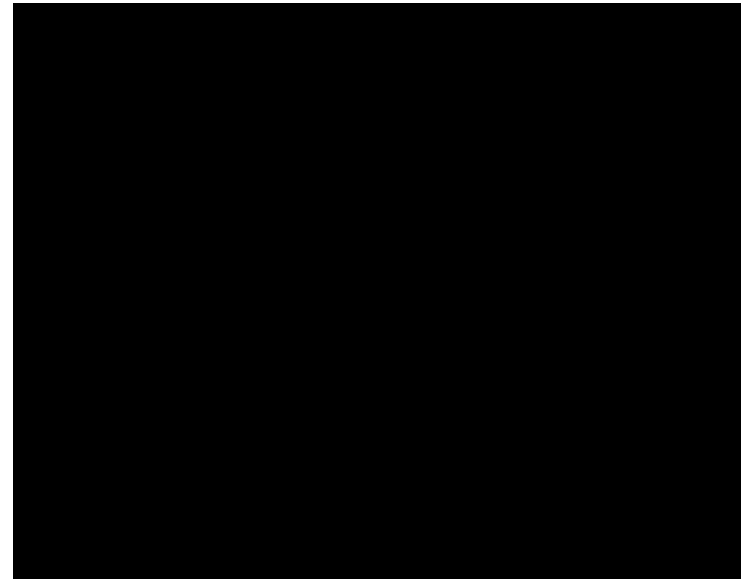


Alignment



Mask

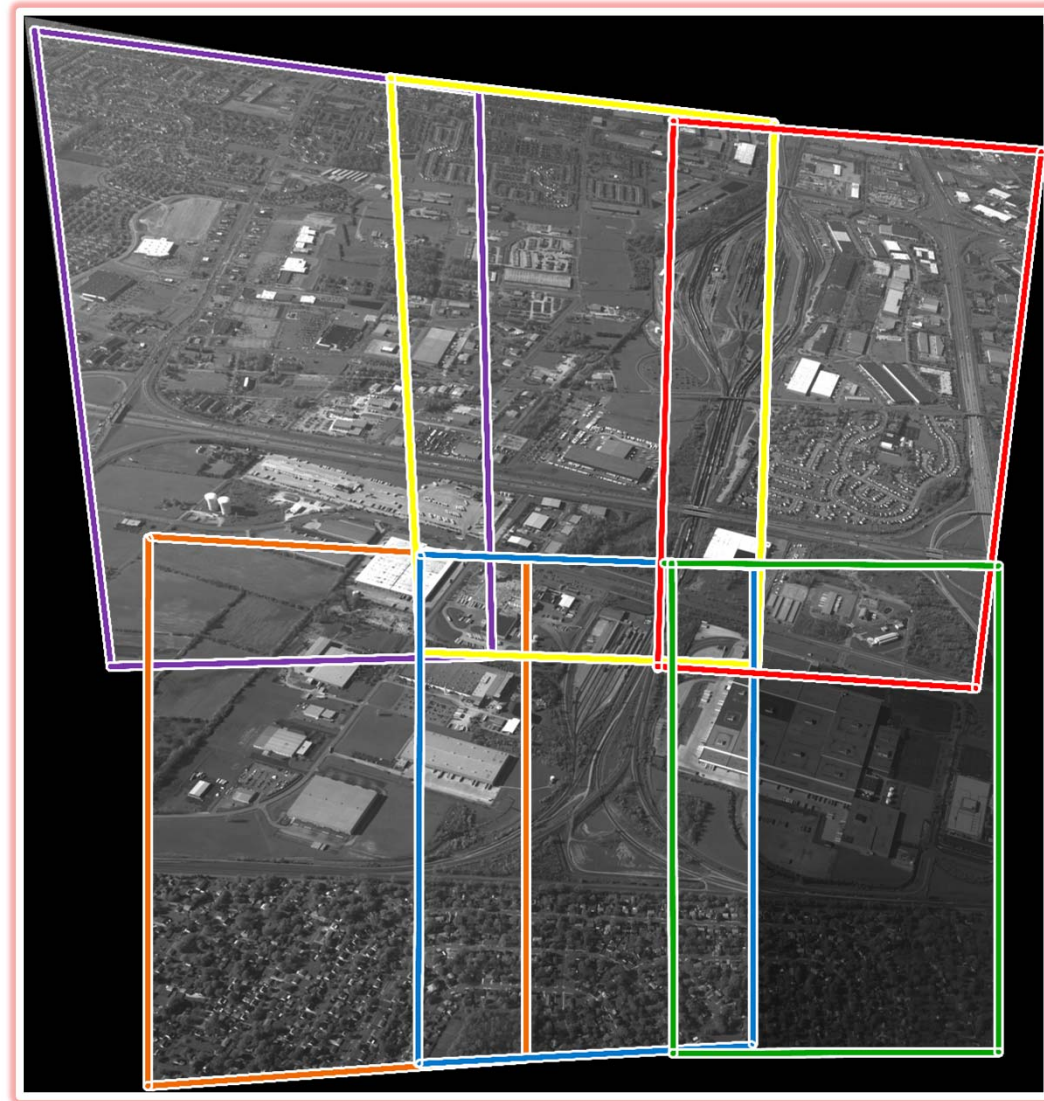
Detection Results



Tracking Results



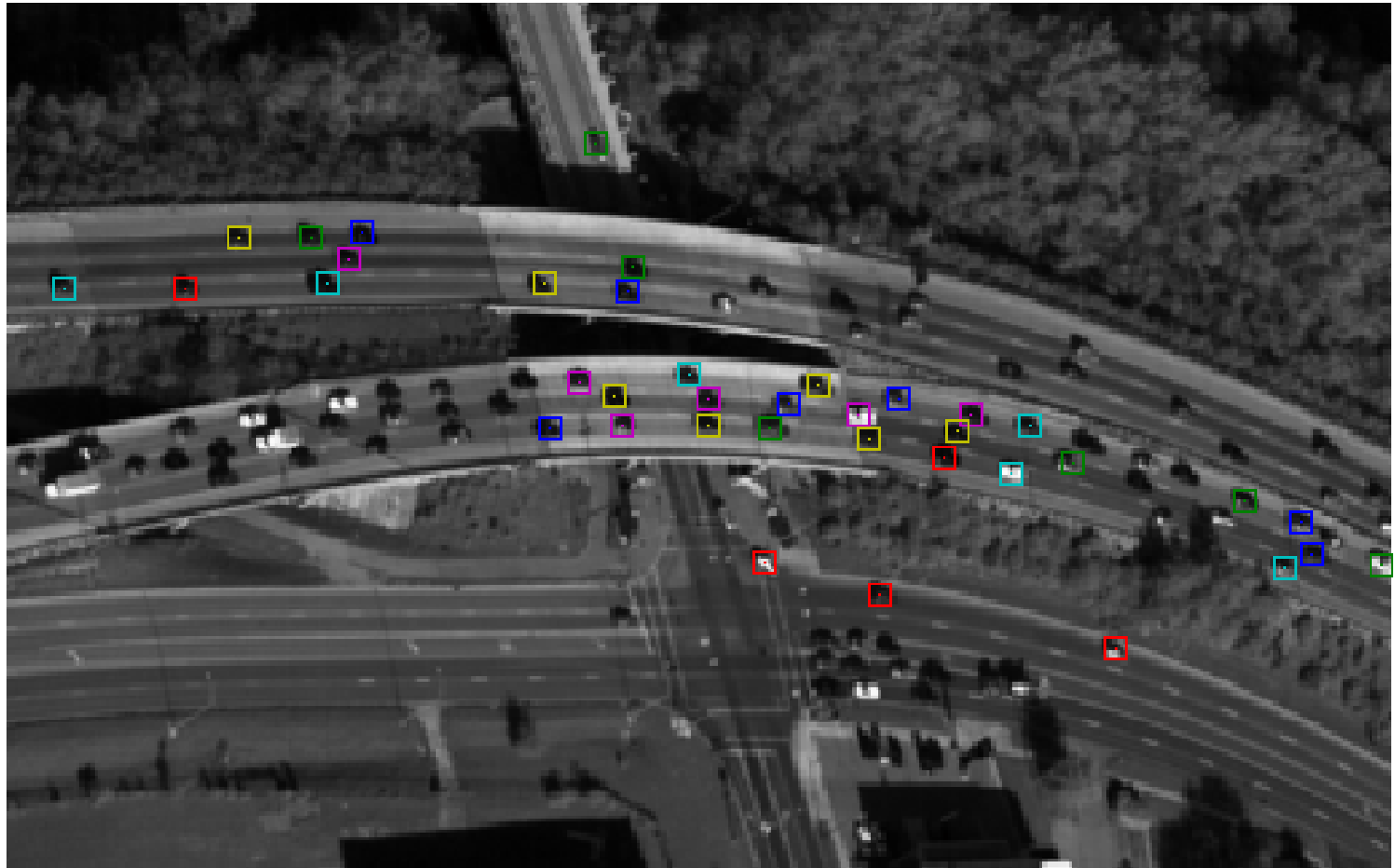
Wide Area Surveillance



Wide Area Surveillance



Tracking Results



Robot Vision (Unmanned Ground Vehicle)



UGV



UGV



Human Action Recognition

Events, Actions, Activities,

- Action
- Event
- Movement
- Activity
- Interaction
- Verb
-

Weizmann Action Dataset

- 10 actions
- 9 actors per action



KTH Data Set

- Six Categories, 25 actors, 4 instances, 600. clips



Boxing



Hand Clapping



Hand Waving



Jogging



Running



Walking

UCF Sports Action Dataset

9 actions, 142 videos.



Bench Swing



Dive



Swing



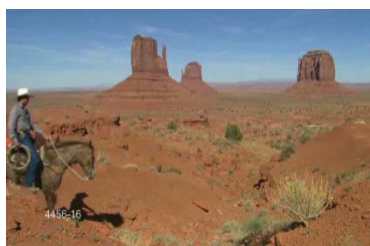
Run



Kick



Lift



Ride



Golf Swing



Skate



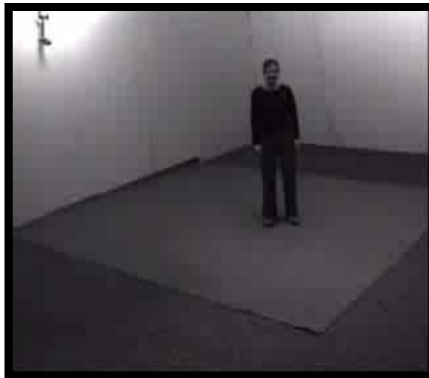
IXMAS Multi-view Data Set

- 13 action categories, 4 camera views, 10 actors, 3 instances.

View 1



View 2



View 3



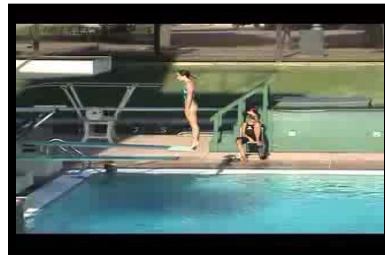
View 4



UCF YouTube Action Dataset (UCF-11)



Cycling



Diving



Golf Swinging



Riding



Juggling



Basketball Shooting



Swinging



Tennis Swinging



Volleyball



Trampoline Jumping



Walking Dog

UCF50



Baseball Pitch



Basketball Shooting



Bench Press



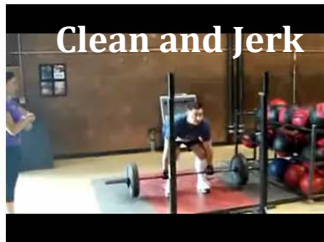
Biking



Billiards



Breaststroke



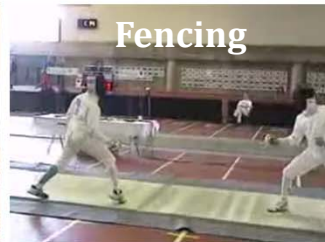
Clean and Jerk



Diving



Drumming



Fencing



Golf Swing



High Jump



Horse Race



Horse Riding



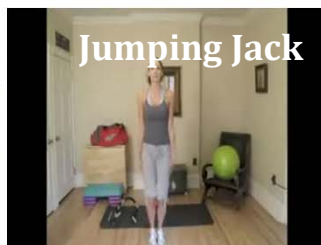
Hula Hoop



Javelin Throw



Juggling Balls



Jumping Jack



Jump Rope



Kayaking



Lunges



Military Parade



Mixing Batter

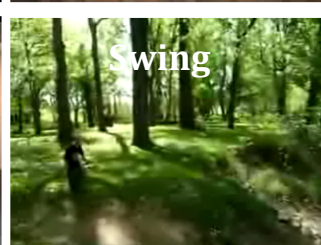
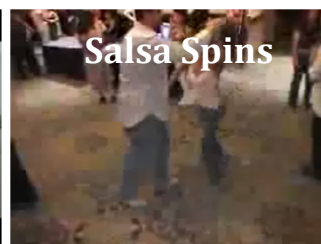
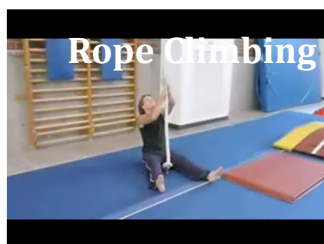
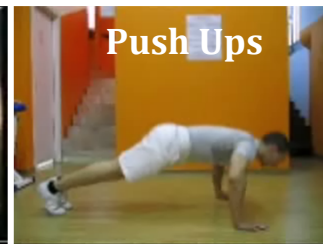
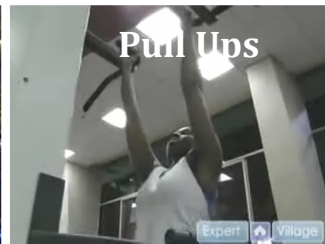


Punching Bags



Pizza Tossing

UCF50



Microsoft Kinect sensor

- Data Captured using Microsoft Kinect sensor

RGB Camera

IR Camera 1



IR Camera 2

- Approximately 50,000 gesture samples

Gesture Lexicons



Diving Signals



Referee Signals



Nurse Gesture



Re

Music Notes



Gestures from Depth camera ▲



▼ Gestures from RGB camera

Discovered Primitives



Left arm moving down



Left arm waving up



Left shoulder moving to right



Right arm moving up



Left arm moving up



Right arm moving away from body



Left arm moving to right



Left arm moving up



Left hand moving forward



Right arm moving laterally up



Left arm moving down



Left arm moving down



Left arm moving to body



Left arm moving away from body



Left shoulder moving down





Right arm moving down

Representative Motion Primitives (out of 136) for different batches





Test case 1: Torso motion adds noise (devel 01– 10 gestures)



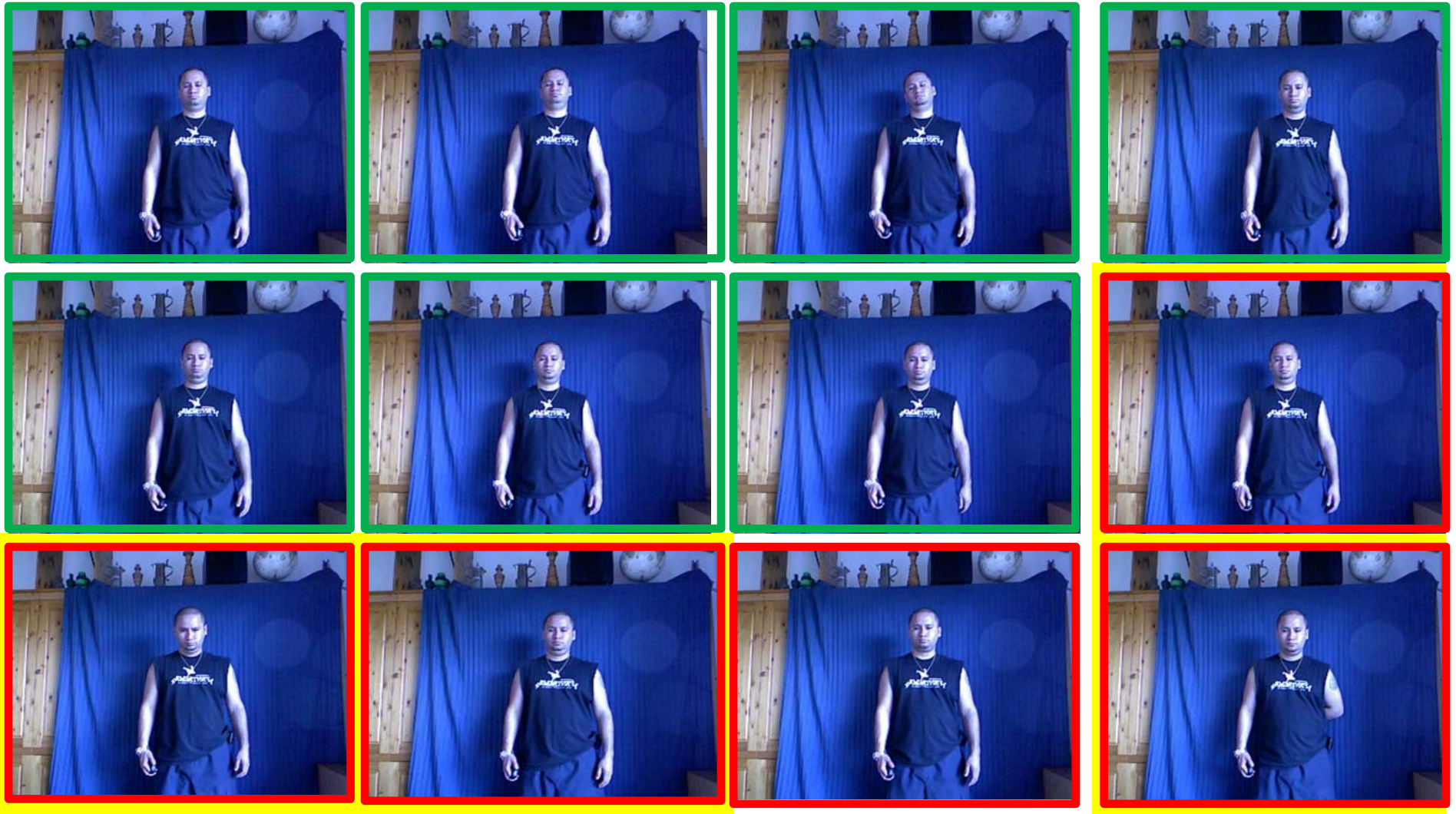
 Instances of Successful Recognition
 Instances of Failed Recognition



Test Case 2: Improvisations (devel 06 – 9 gestures)



 Instances of Successful Recognition
 Instances of Failed Recognition

Test Case 3: Subtle differences (level 09 – 10 gestures)



-  Instances of Successful Recognition
-  Instances of Failed Recognition

COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

OF THE

12/2011 VOL.54 NO. 12

Visual Crowd Surveillance Is Like Hydrodynamics

The Legacy of
Steve Jobs

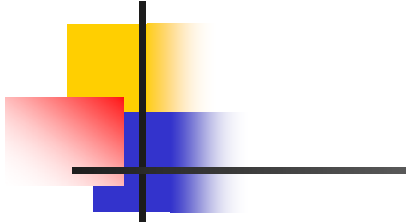
Answer Set
Programming

How Will Astronomy
Archives Survive the
Data Tsunami?

Formal Analysis of
MPI-based Parallel
Programs

Brave NUI World

Association for
Computing Machinery



High Density Crowded Scenes



Political Rallies

Religious Festivals

Marathons

High Density
Moving Objects

Tracking in Crowds



- Average chip size 14 x 22 pixels
- 492 Frames
- Selected 199 athletes for tracking
- Successfully tracked 143 athletes

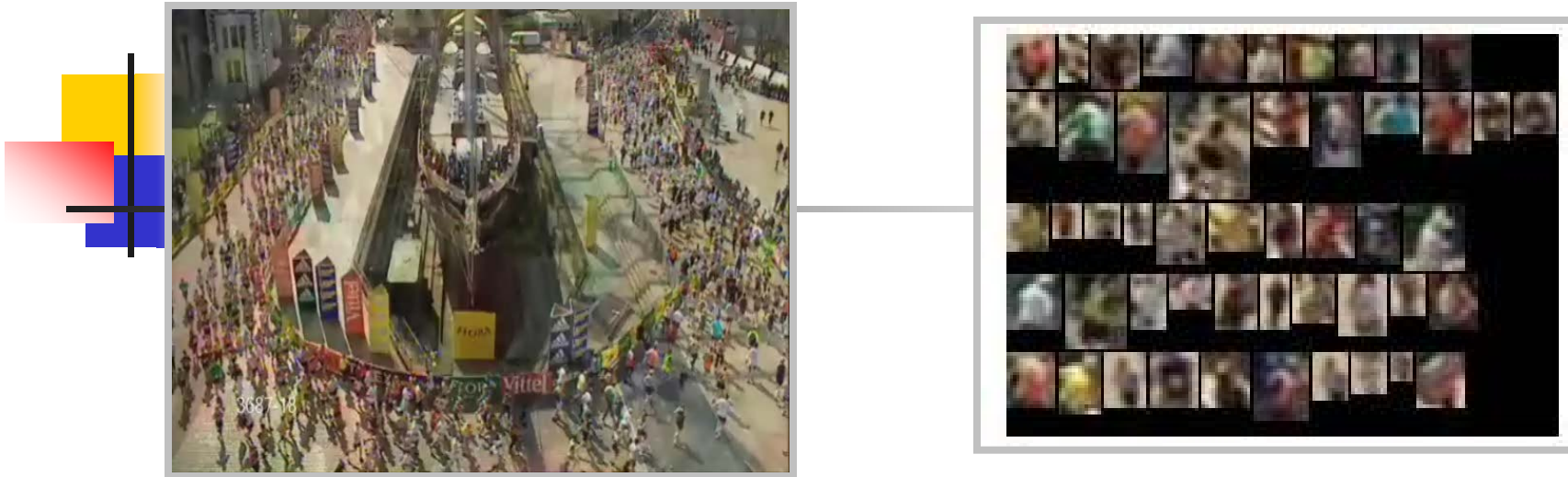
Results



Experiment – 1

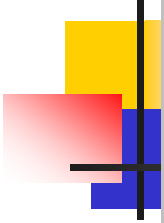


Experiment-3



- Average chip size 14 x 17 pixels
- 453 Frames
- Selected 50 athletes for tracking

Experiment – 3



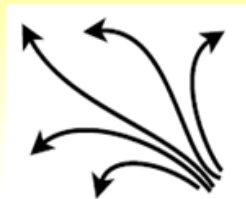
Behaviors in Crowded Scenes



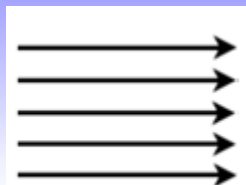
Bottleneck



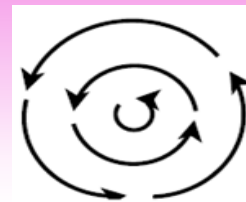
Departure



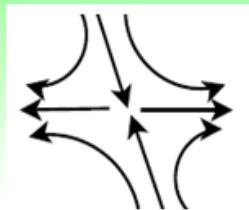
Lanes

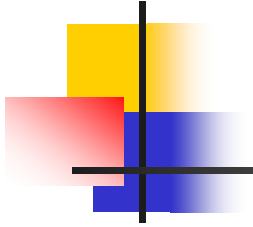


Arch/Rings



Blockings





Where Am I?

“Where Am I?”

Problem:

Accurate Image Localization

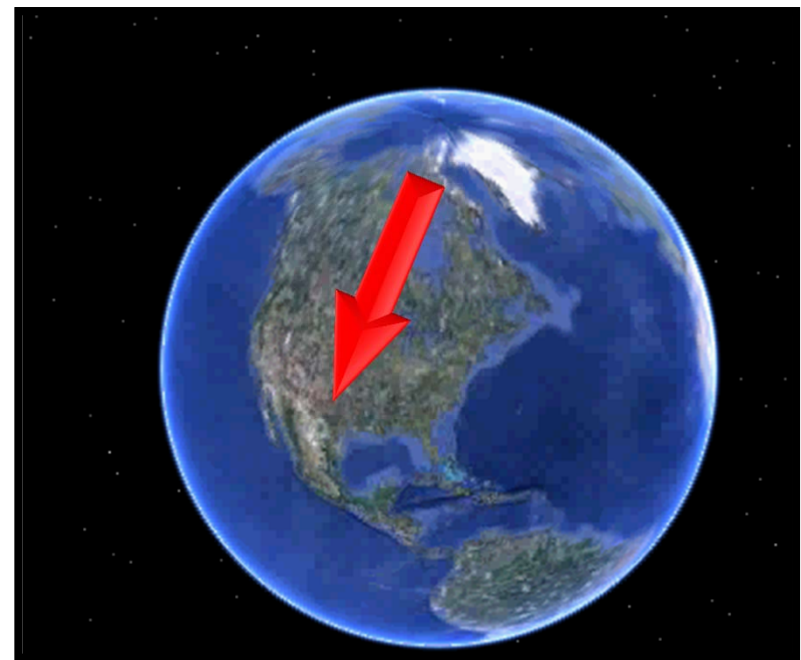
Input



Mere Visual Information(Images)



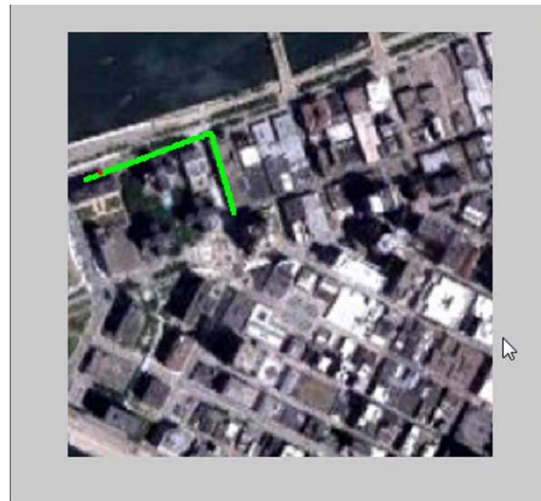
Output



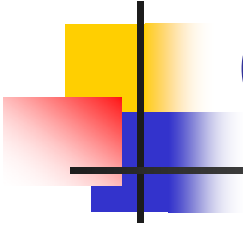
Location in Terms of λ (Lon.) and φ (Lat.)

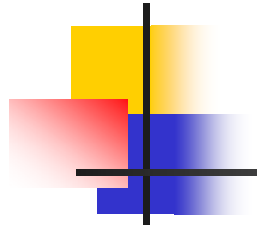
$\varphi=40.4419$, $\lambda=-79.9986$

Geospatial Trajectory Extraction



Computer Vision for Computer Graphics

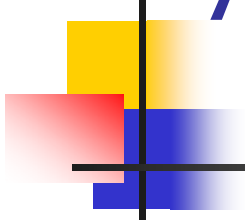




Video Completion

Other Results

Layer Based Video Composition



Results of Doll



Results of Mom-Daughter





Multimedia: Segmentation of Moving-Sounding Objects



Audio Source
Localization

Accepted in IEEE Transactions on
Multimedia



New York Times, August 19, 2012



Robot arms like those at a Philips Electronics factory in the Netherlands can perform the same tasks as hundreds of low-skill workers.

