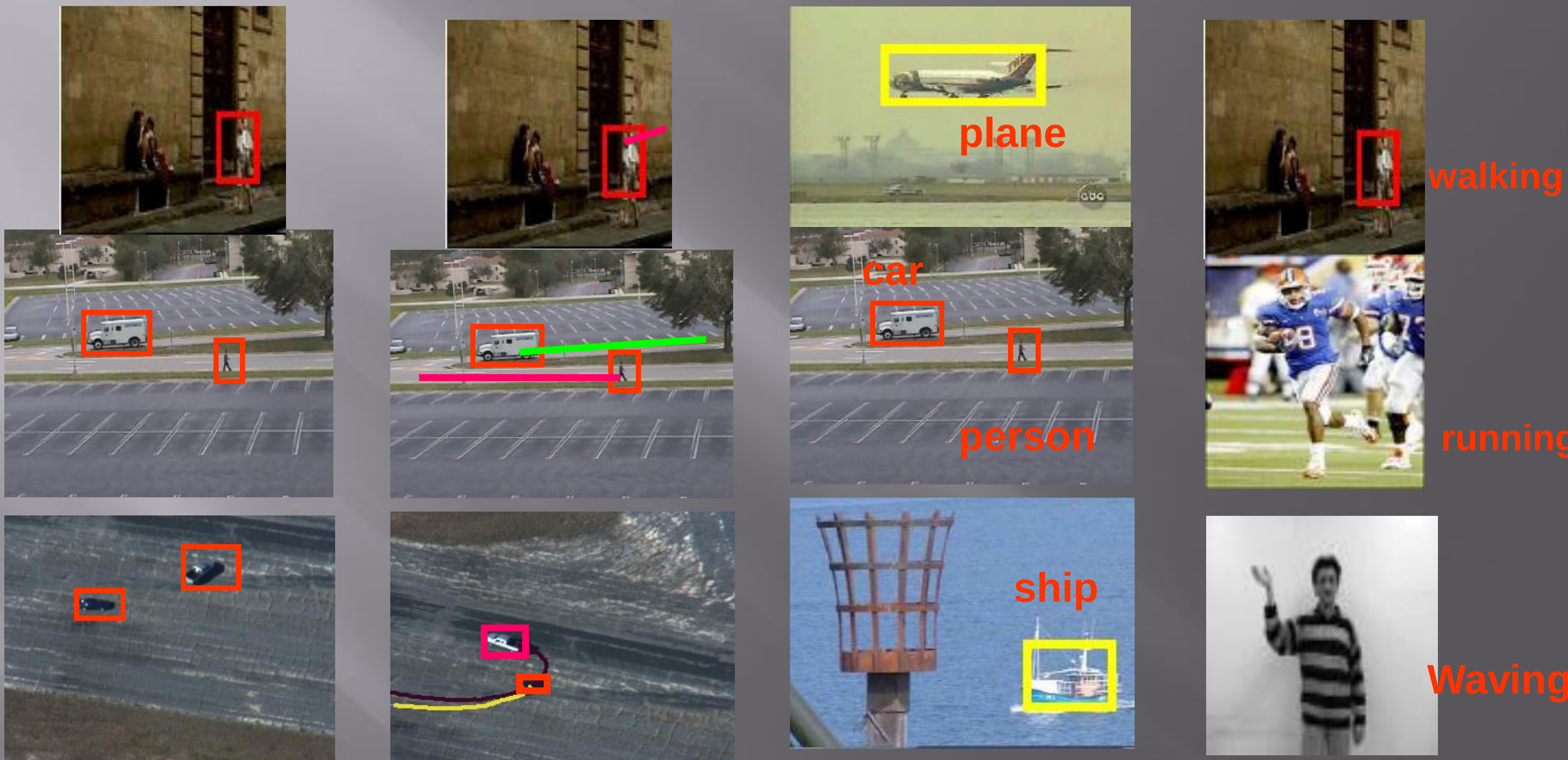
VIDEO SURVEILLANCE & MONITORING

Mubarak Shah
Computer Vision Lab
University of Central Florida

University of Central Florida VISION

Main Steps



Object detection → Object tracking → Object categorization and classification → Events or Activities Recognition

University of Central Florida VISION

BAYESIAN MODELING OF DYNAMIC SCENES FOR OBJECT DETECTION

Omar Javed, Khurram Shafique, and Mubarak Shah

IEEE Workshop on Motion & Video Computing

2002

University of Central Florida VISION

Interest Region Detection Using Background Subtraction

- Aim:
  - Mark pixels in the image corresponding to interesting objects
  - Completely unsupervised learning

- General Approach:
  - Build a per-pixel model of background
  - Find deviations from the model

Background Subtraction

- Important Problems in Realistic situations:
  - Quick illumination changes

  - Relocation of background objects.

Background Subtraction

– Initialization with moving objects

– Shadows

University of Central Florida VISION

Background Modeling
THE HIERARCHICAL APPROACH

- Which features to use?
  - Color
  - Gradient

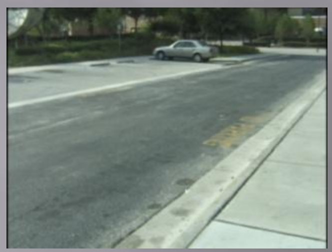


Image-1

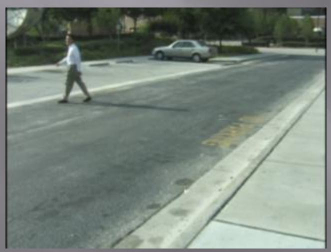
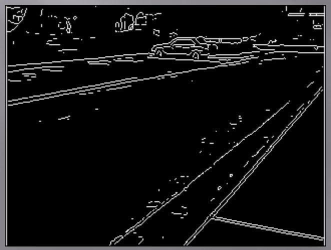
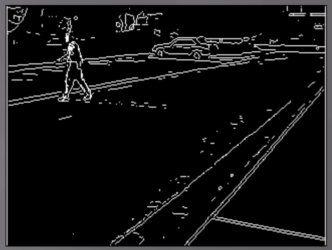


Image-153

Color based (Image 153)



Gradient Image-1



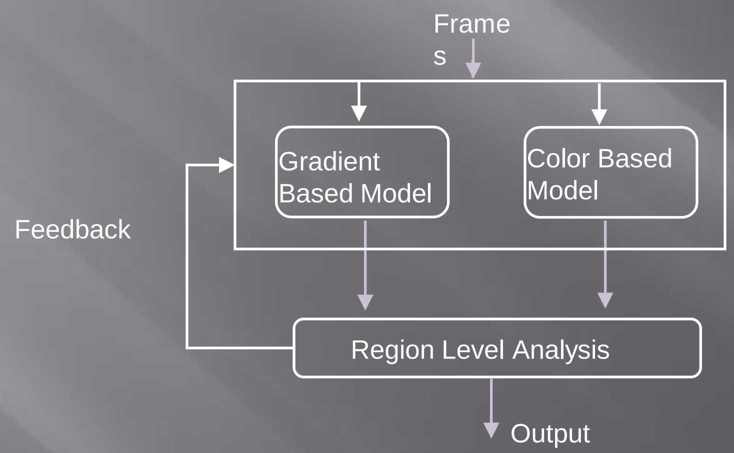
Gradient Image-153

Gradient based (image 153)

University of Central Florida VISION

Background Modeling
THE HIERARCHICAL APPROACH

- Fusion of features
  - Validation of pixel level results at the regional level.



  - Updating of pixel level models based on feedback from regional level process.

Pixel Level Subtraction

- Color based subtraction
  - Per-pixel mixture of Gaussians.
- Gradient based subtraction
  - Gradient feature vector, $\Delta = [\Delta_m, \Delta_d]$

    where

    $$\Delta_m = \sqrt{f_x^2 + f_y^2}$$

    $$\Delta_d = \tan^{-1}\left(\frac{f_y}{f_x}\right)$$

  - Distribution of the gradient feature vector?

Subtraction in the Gradient Domain

- Let $x_{i,j}^{t}$ be the latest value that matched k^{th} distribution belonging to background at pixel (i,j)
- The gray level value will be given as

$$g_{i,j}^{t} = \alpha R + \beta G + \gamma B$$

- Assuming independence of color channels,

$$g_{i,j}^{t} \sim N(\mu_{i,j}^{t}, (\sigma_{i,j}^{t})^{2})$$

University of Central Florida VISION

Subtraction in the Gradient Domain

- Let us define

$$f_x = g^t_{i+1,j} - g^t_{i,j}$$

$$f_y = g^t_{i,j+1} - g^t_{i,j}$$

- Assuming independence between neighboring pixels.

$$f_x \sim N(\mu_{f_x}, (\sigma_{f_x})^2)$$

$$f_y \sim N(\mu_{f_y}, (\sigma_{f_y})^2)$$

- The covariance is given by

$$Cov(f_x, f_y) = Cov(g^t_{i+1,j} - g^t_{i,j}, g^t_{i,j+1} - g^t_{i,j}) = (\sigma^t_{i,j})^2$$

Subtraction in the Gradient Domain

- Distribution of feature vector $[\Delta_m, \Delta_d]$

$$F(\Delta_m, \Delta_d) = \frac{\Delta_m}{2\pi\sigma_{fx}\sigma_{fy}\sqrt{1-\rho^2}} \exp\left(-\frac{z}{2(1-\rho^2)}\right)$$

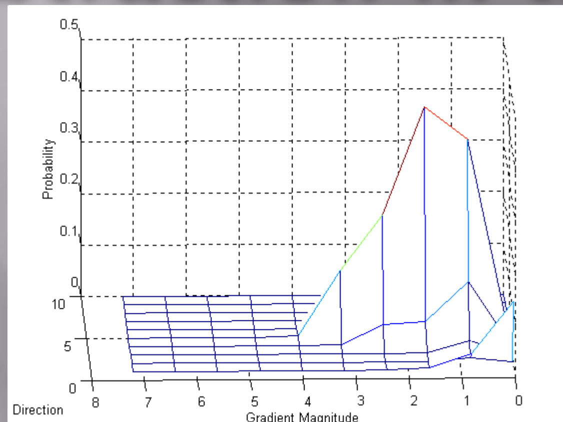
  - where

$$z = \left(\frac{\Delta_m \cos\Delta_d - m_{fx}}{\sigma_{fx}}\right)^2 - 2\rho\left(\frac{\Delta_m \cos\Delta_d - m_{fx}}{\sigma_{fx}}\right)\left(\frac{\Delta_m \sin\Delta_d - m_{fy}}{\sigma_{fy}}\right) + \left(\frac{\Delta_m \sin\Delta_d - m_{fy}}{\sigma_{fy}}\right)^2$$

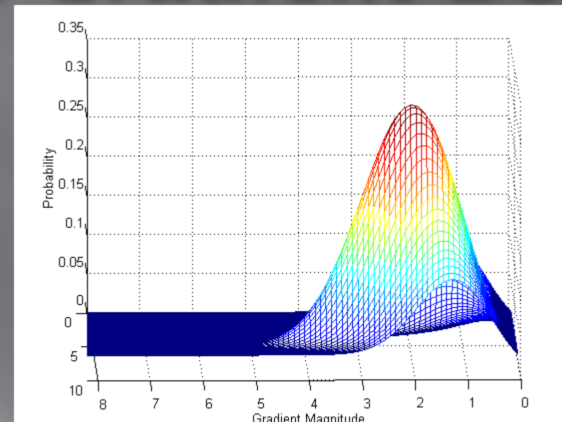
$$\rho = \frac{\sigma_{i,j}^2}{\sigma_{fx}\sigma_{fy}}$$

- If $F(\Delta_m, \Delta_d) < T_g$ then pixel is marked as foreground.

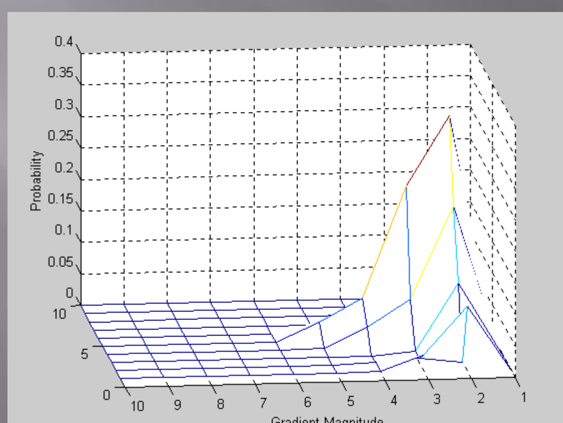
Subtraction in the Gradient Domain



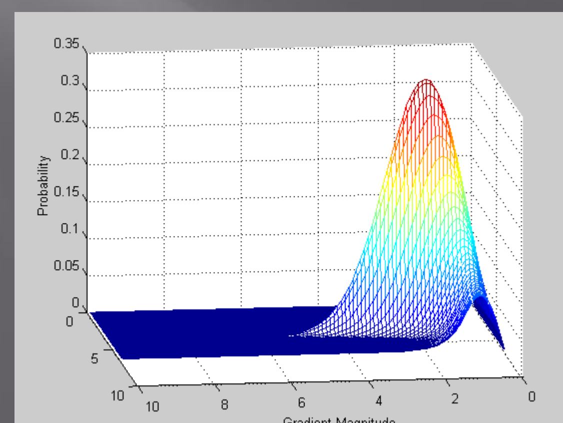
Sample Histogram of $\Delta = [\Delta_m, \Delta_d]$



Parametric Distribution of $\Delta = [\Delta_m, \Delta_d]$



Sample ogram of $\Delta = [\Delta_m, \Delta_d]$



Parametric Distribution of $\Delta = [\Delta_m, \Delta_d]$

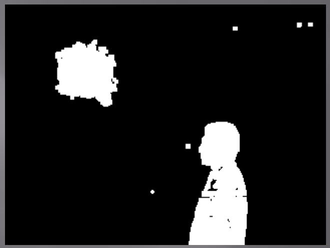
University of Central Florida VISION

Pixel Level Subtraction

- Examples of pixel wise subtraction in the color and gradient domain.

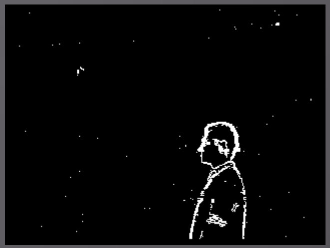
Image-1

Image-85



Color based

(Image 85)



Gradient based

(Image 85)

Region Level Processing

- A region with edges on its boundary that are different from the background is a valid region.

- A region *R* is accepted as a valid region if

$$\frac{\sum_{(i,j)\in\partial R}(\nabla I(i,j)G(i,j))}{|\partial R|} \geq \rho_B$$

Where

- ∂R is the set of boundary pixels of a connected region *R* in color based results
- $G(i,j)$ is the gradient based subtraction output at pixel (i,j)
- $\nabla I(i,j)$ is the edge map

University of Central Florida VISION

Region Level Processing

• For each color based region, presence of“edge difference” pixels at the boundaries is checked.

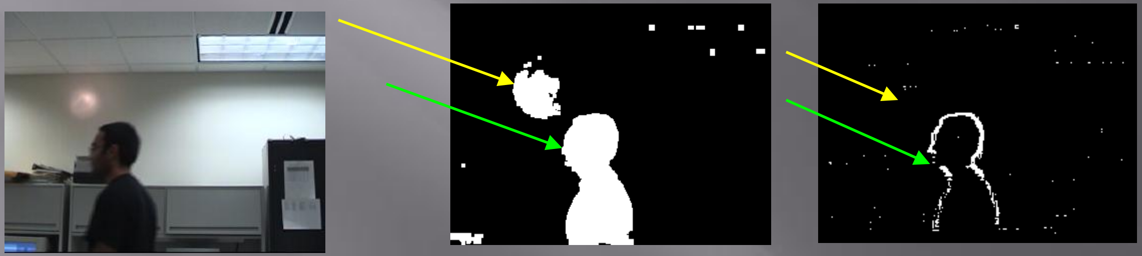


Image   Color based   Gradient

• Regions with small number of edge difference pixel are removed, color model is updated.

Final

University of Central Florida VISION

Background Modeling
THE HIERARCHICAL APPROACH

- Local Illumination Change.

Mixture of Gaussians (S&G) Method

Hierarchical Subtraction

Background Modeling
THE HIERARCHICAL APPROACH

- Relocation of background object.

Mixture of Gaussians (S&G) Method

Hierarchical Subtraction

Background Modeling
THE HIERARCHICAL APPROACH

- Moving with initialization & illumination change.

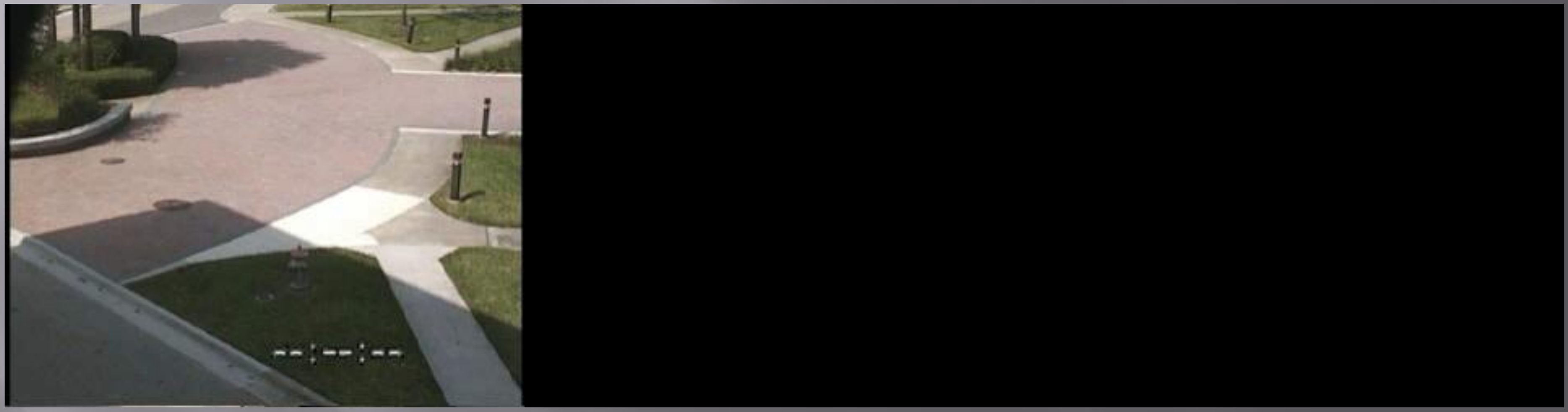
Mixture of Gaussians (S&G) Method

Hierarchical Subtraction

University of Central Florida VISION

Background Modeling
THE HIERARCHICAL APPROACH

- Quick Illumination change.



Mixture of Gaussians (S&G) Method

Hierarchical Subtraction

University of Central Florida VISION

BAYESIAN MODELING OF DYNAMIC SCENES FOR OBJECT DETECTION

Yaser Sheikh and Mubarak Shah

IEEE Conference on
Computer Vision and Pattern Recognition
2005

University of Central Florida
VISION

What is a Dynamic Scene?
(and why should I be interested?)

Periodic

Temporal texture

Nominal camera motion

Temporal texture

University of Central Florida VISION

Overview

- Modeling the Background
  - Non-parametric density estimation
  - Joint Domain-Range Feature Space
- Modeling the Foreground
  - Competitive detection (back ground vs foreground)
- MAP-MRF estimation framework
  - Efficient minimization using graph cuts

The Background Distribution

- **Object Detection:** Given an image, what is the probability of observing a pixel color at a certain location?
  P(**pixel** | **background video**) instead of
  P(**pixel** | **background pixel**)
- Analysis on $\mathbf{x}_i \in \mathbb{R}^5$, the feature space: $[R,G,B,X,Y]$
- This is our background *model*
- Use Kernel Density Estimation on this 5 dimensional space
- Probabilistic Low Level Descriptor

Kernel Density Estimation
(a.k.a. Parzen Windows)

- Definition of KDE

$$f(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-x_i}{h_i}\right)$$

where *K* is a *d*-variate kernel function usually satisfying .

- The Gaussian kernel is a common choice.

Kernel Density Estimation

- Characteristics
  - Nonparametric technique
  - Effective multi-modal data representation
- Background model is represented in the 5-space with the set $\psi_b=\{\mathbf{y}_1,\mathbf{y}_2\ldots\mathbf{y}_n\}$, where n pixels have been observed thus far.

$$\Pr(\mathbf{x}|\psi_b) = n^{-1}\sum_{i=1}^{n}\varphi_H(\mathbf{x}-\mathbf{y}_i)$$

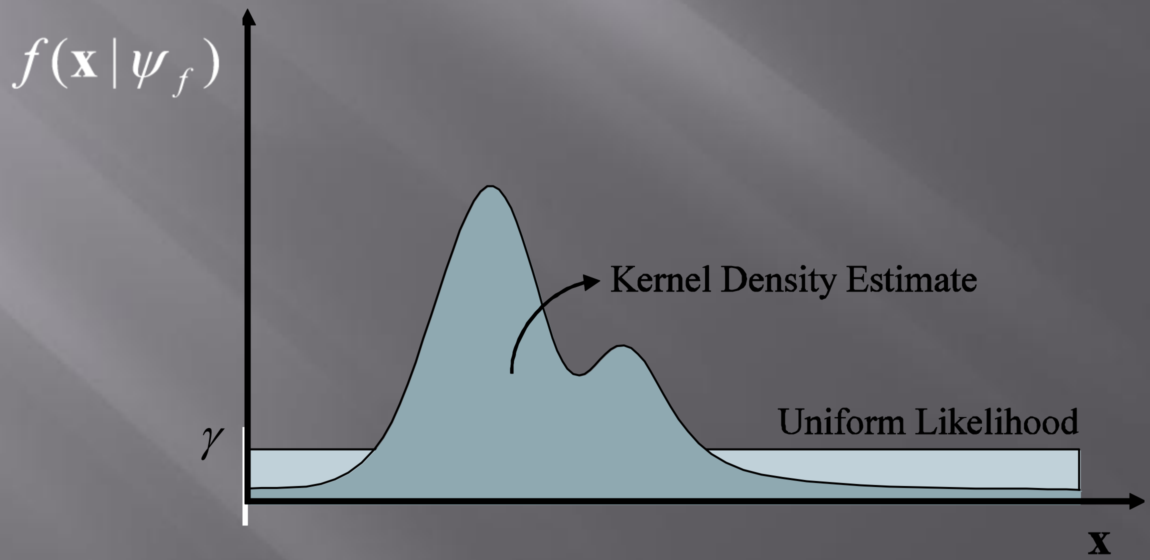
Temporal Persistence

- **Intuition:** Objects tend to maintain constant colors, and tend to persist in the same proximity (smooth motion)
- Frame at t-1 contains substantial evidence for detection at t
- Given the detected foreground in previous frames, what is the probability of the object belonging to that foreground?

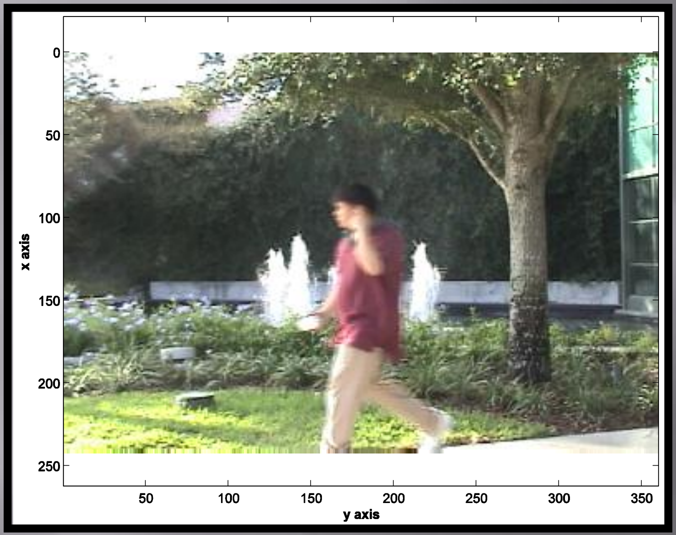
Foreground Model

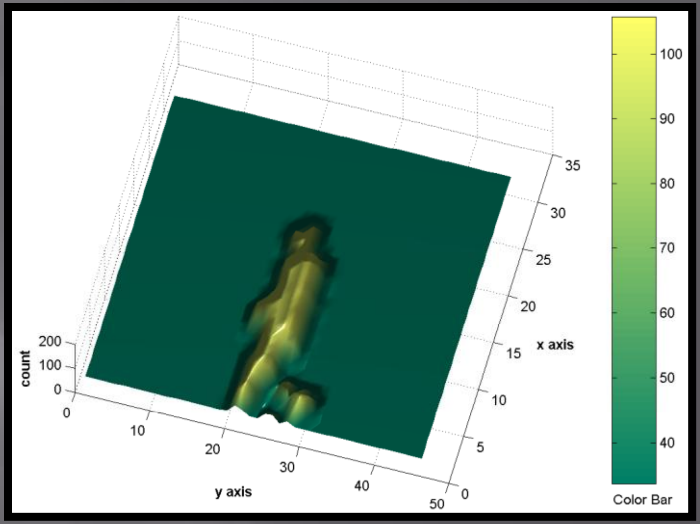
- Foreground Density Estimator

$$f(\mathbf{x}|\psi_f) = \alpha\gamma + (1-\alpha)\frac{1}{m}\sum_{i=1}^{m}\varphi_H(\mathbf{x}-\mathbf{y}_i)$$

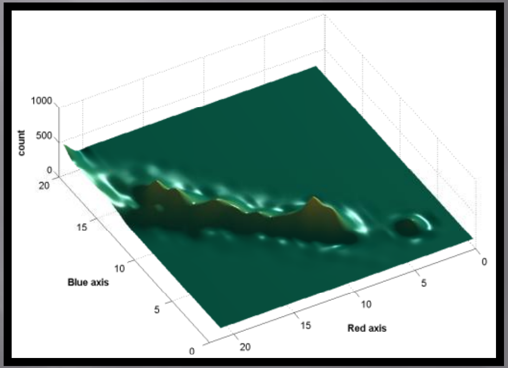


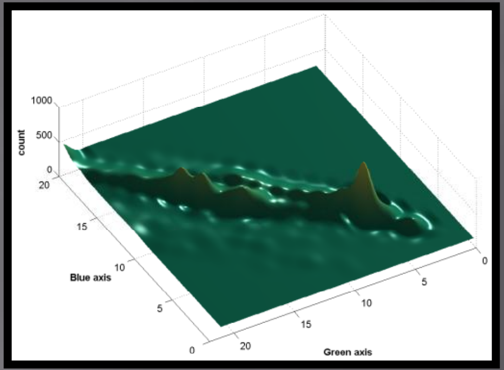
Marginals of the Foreground Distribution

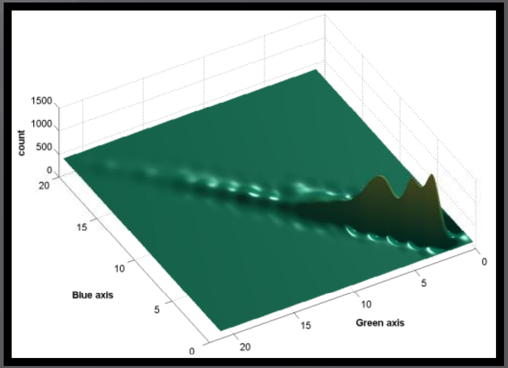




BR
BG
GR







Classification vs Detection

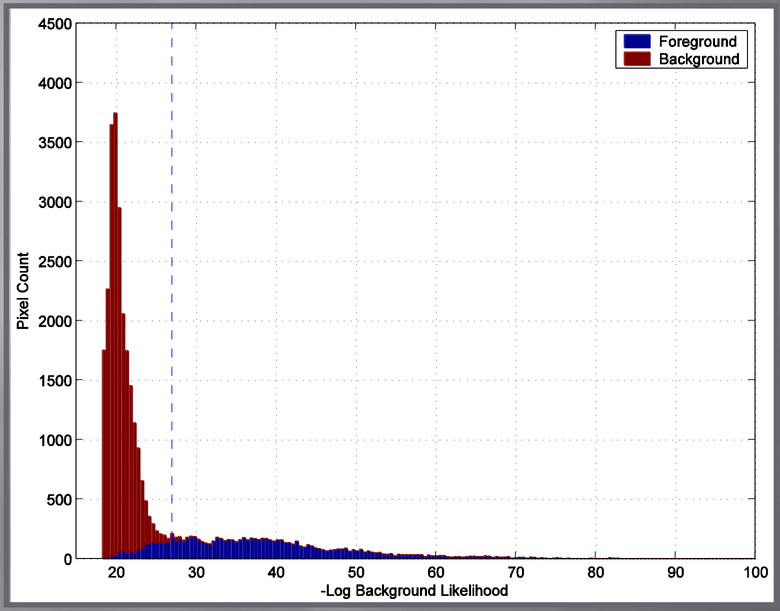
foreground

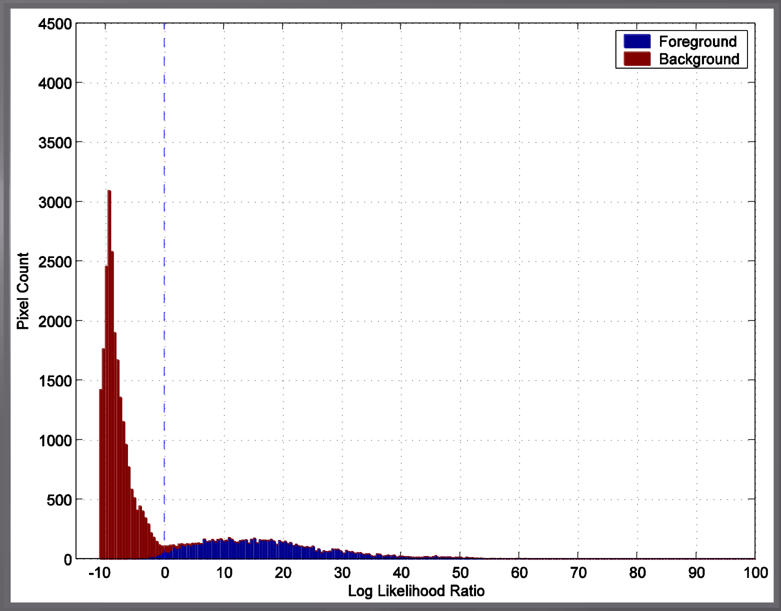
ratio

background

University of Central Florida VISION

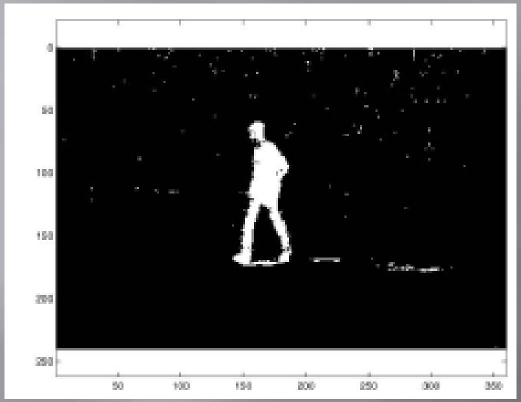
Improved Discrimination



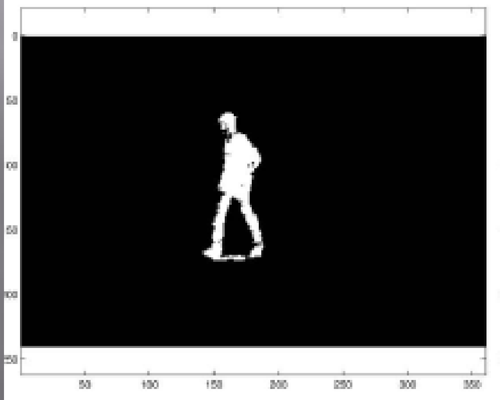


University of Central Florida VISION

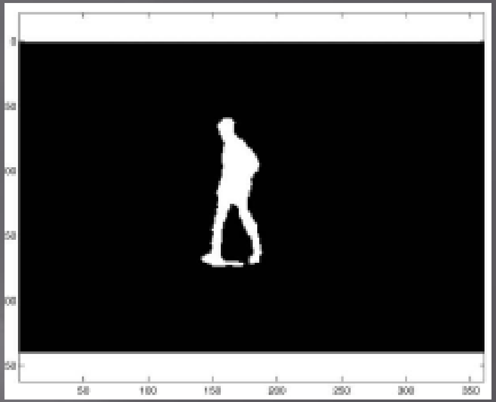
Detection



Thresholding on background model



Thresholding on Likelihood model



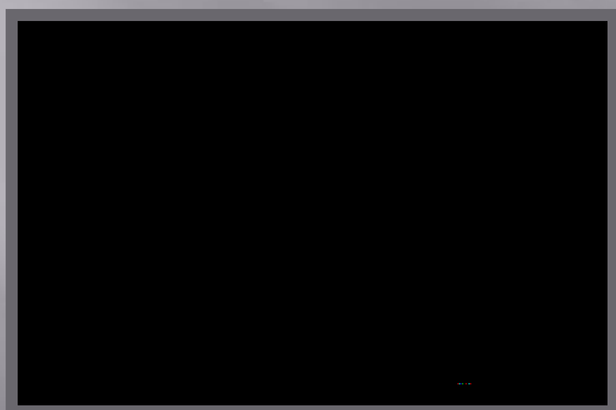
MAP MRF

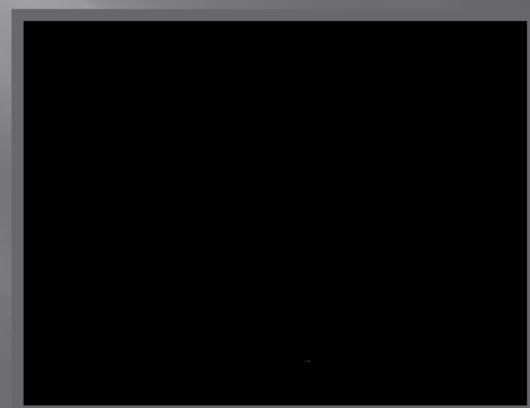
The Watery Sequence

Mixture Of Gaussians

Proposed

The Fountain Sequence





Proposed

Mixture Of Gaussians

Nominal Motion Sequence

Proposed

Ground

Truth

Mixture Of

Gaussians

Quantitative Analysis
(Pixel Level)

- Manually segmented 500 frames of the nominal motion sequence

- Pixel-wise comparison between manual detection and proposed detection scheme

University of Central Florida VISION

Nominal Motion Sequence

Proposed

Ground

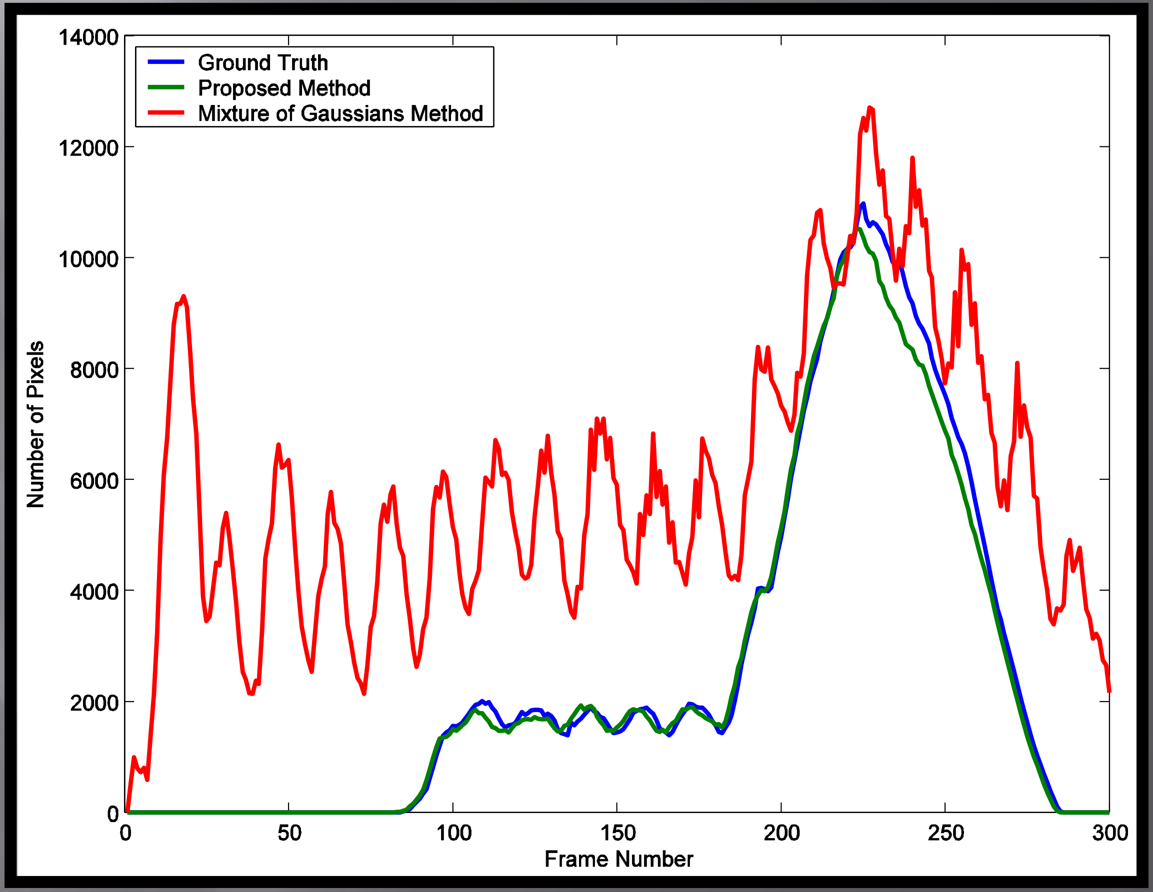
Truth

Mixture Of

Gaussians

University of Central Florida VISION

Pixel-wise Analysis



University of Central Florida VISION

Quantitative Analysis
Object Level

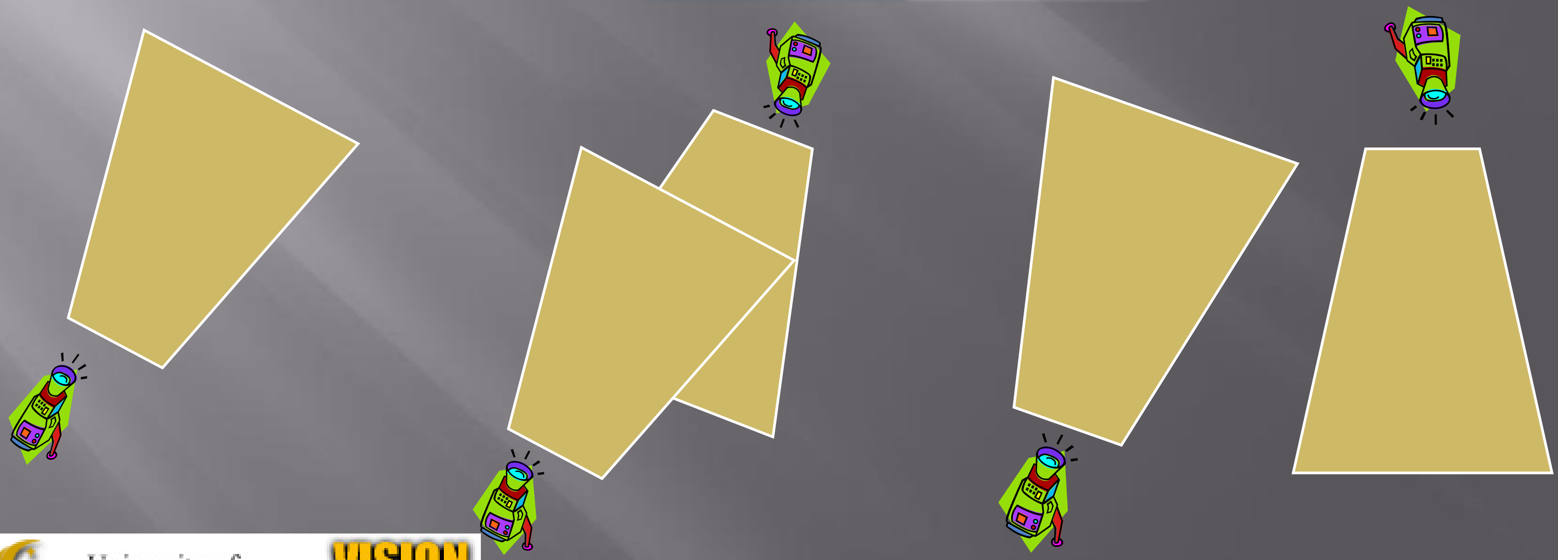
	Objec ts	Detected	Mis-detected	Detection Rate	Mis-Detection Rate
Sequence 1	84	84	0	100.00%	0.00%
Sequence 2	115	114	1	99.13%	0.87%
Sequence 3	161	161	0	100.00%	0.00%
Sequence 4	94	94	0	100.00%	0.00%
Sequence 5	170	169	2	99.41%	1.18%

University of Central Florida VISION

VISUAL TRACKING

Camera Configurations

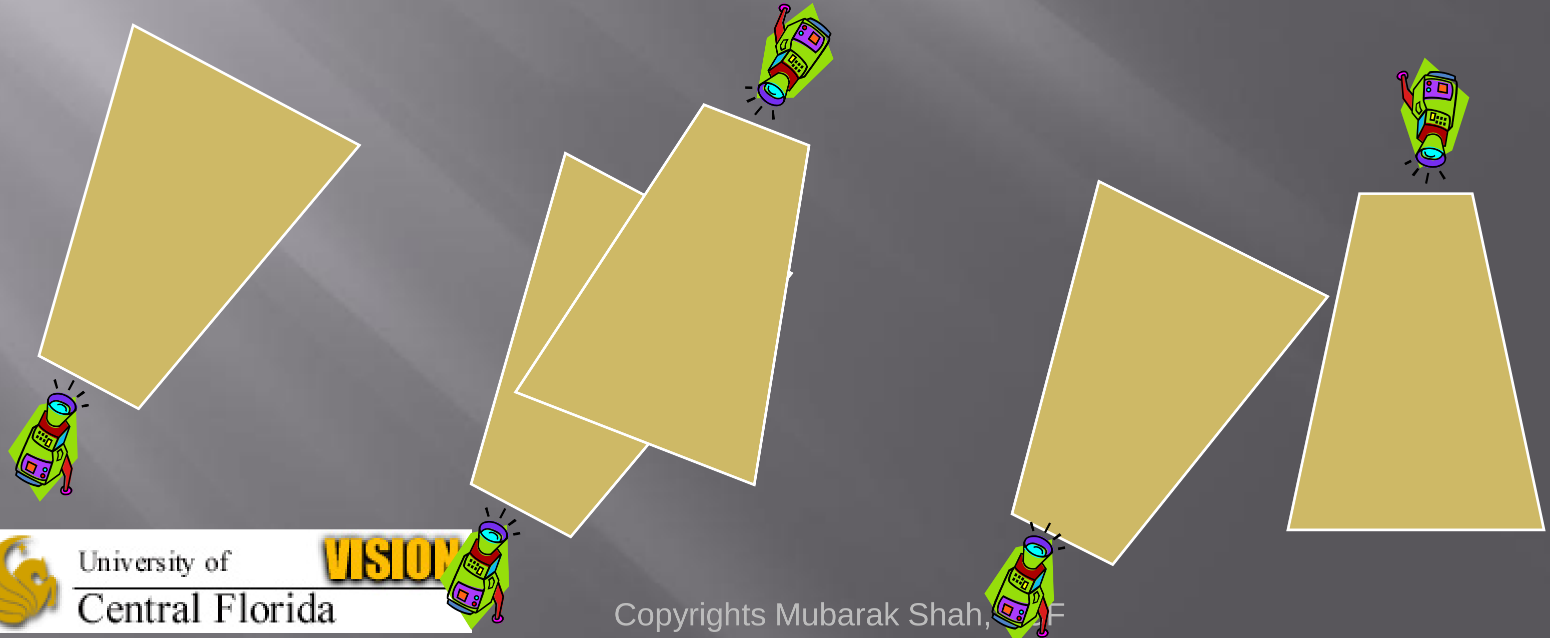
- Stationary Camera
  - Single camera (KNIGHT)
    - Javed et al. ECCV 2002
  - Multiple cameras with overlapping field of view
    - Khan et al. PAMI 2003, Khan et al ECCV 2006
  - Multiple cameras with non-overlapping field of view
    - Javed et al. ICCV 2003, CVPR 2005, AAAI-2007



University of Central Florida VISION

Camera Configurations

- Moving Cameras
  - Single camera
    - Yilmaz et al. PAMI 2004, Khan et al CVPR 2007
  - Multiple cameras with overlapping field of view
    - Sheikh et al. ICCV 2005, Yilmaz et al ICCV2005
  - Multiple cameras with non-overlapping field of view
    - Sheikh et al. CVPR 2007



Tracking IN Single Fixed Camera

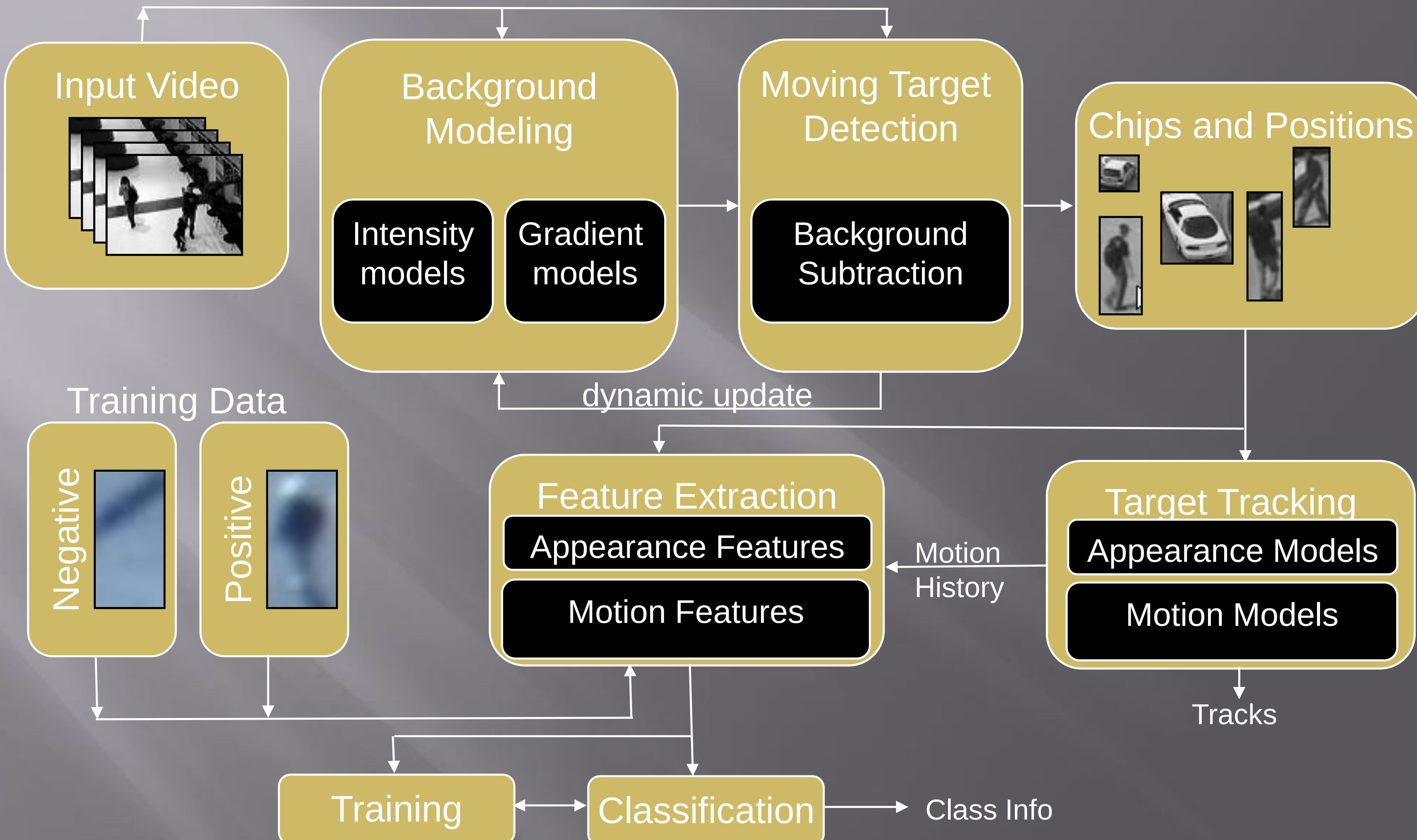
KNIGHT

Crime Scene Detection System for The Orlando Police Department

Views

University of Central Florida VISION

KNIGHT: Video Surveillance System



KNIGHT: Single Fixed Camera Tracking

University of Central Florida VISION

KNIGHT: Single Fixed Camera Tracking (Occlusion)

Visual Monitoring of Railroad Crossings

- Detected Violators

03-01-2005, 11:32:33

03-01-2005, 11:32:19

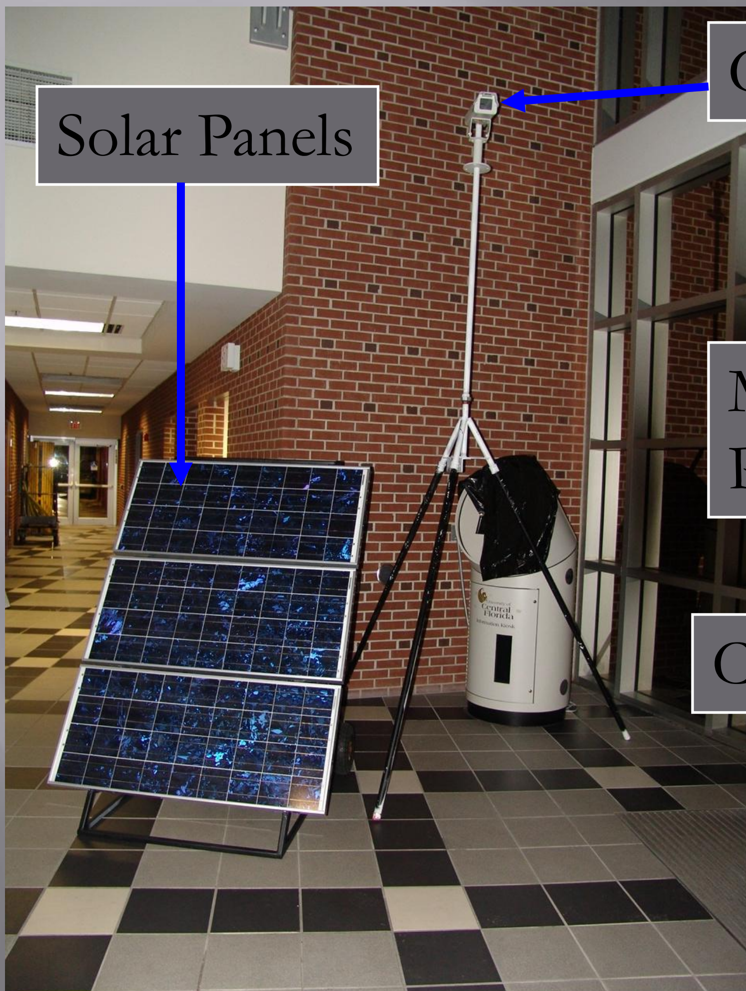
03-08-2005, 13:18:23

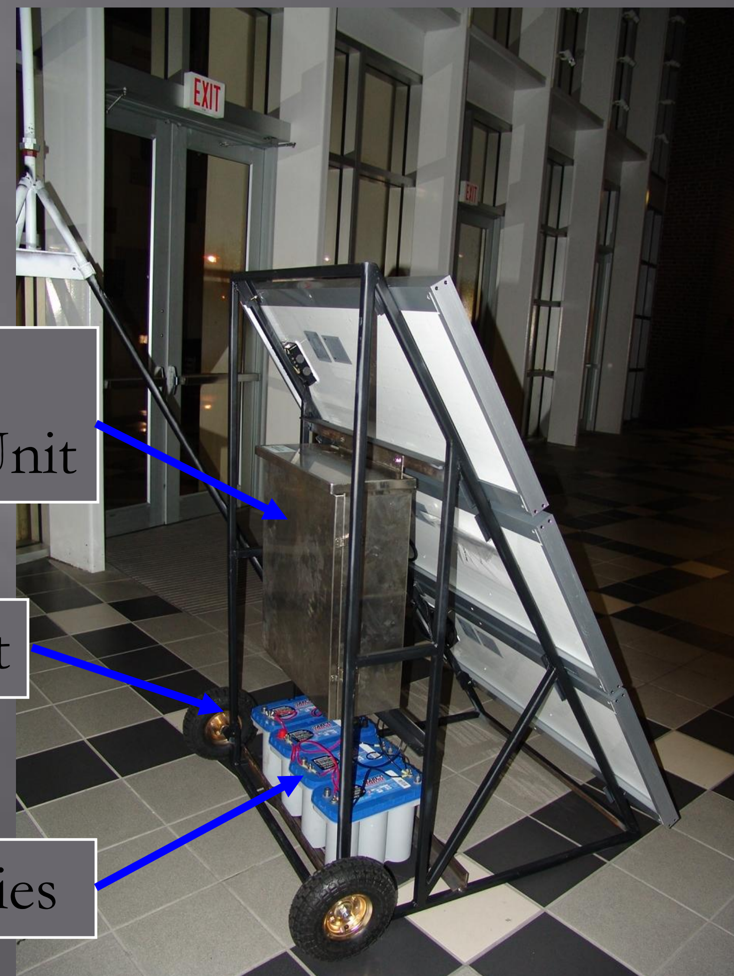
02-22-2005, 11:01:42

03-02-2005, 12:42:57

03-11-2005, 15:17:53

The Portable System

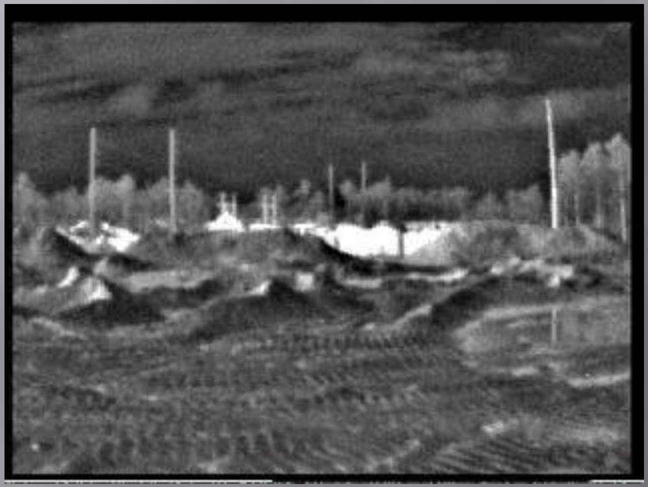




University of Central Florida VISION

Nighttime Video Surveillance

DARPA Phase II STTR



Single Moving Camera Tracking

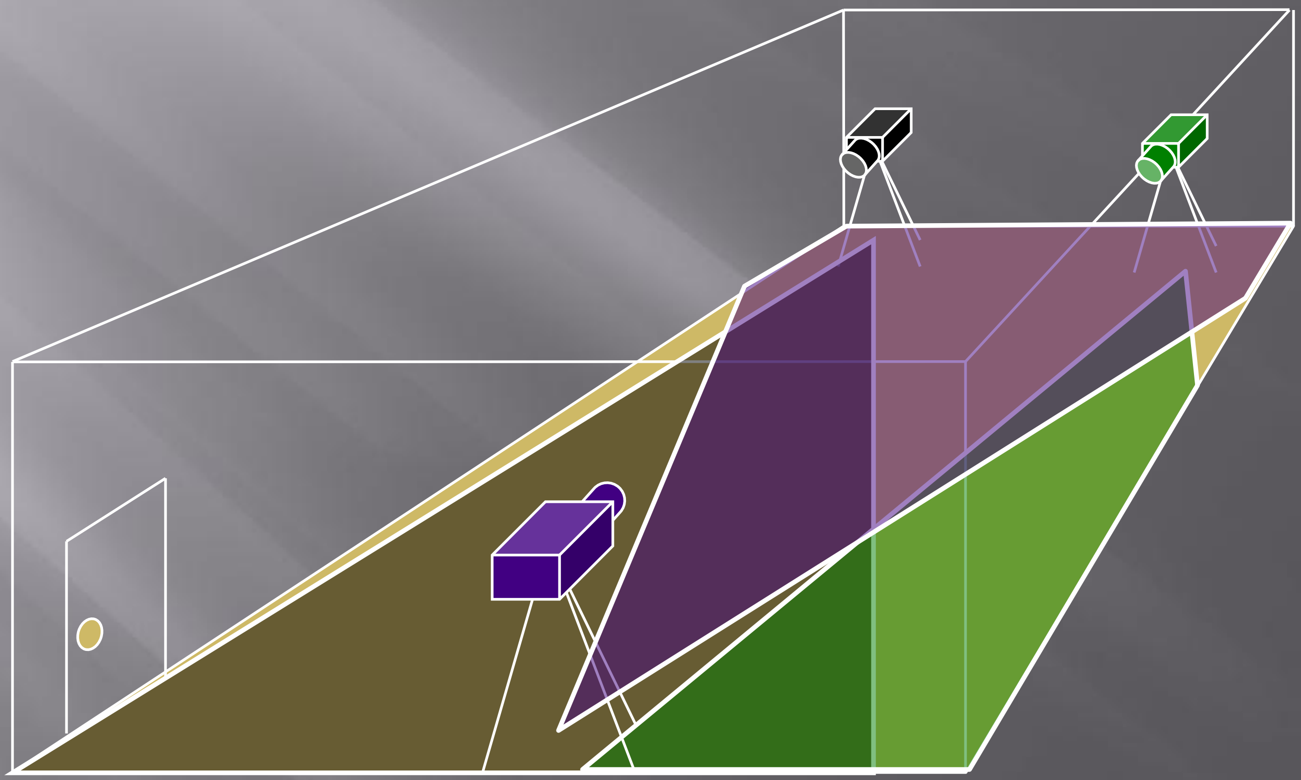
MULTIPLE CAMERA TRACKING

TRACKING ACROSS MULTIPLE FIXED OVERLAPPING CAMERA

Sohaib Khan and Mubarak Shah

PAMI 2003

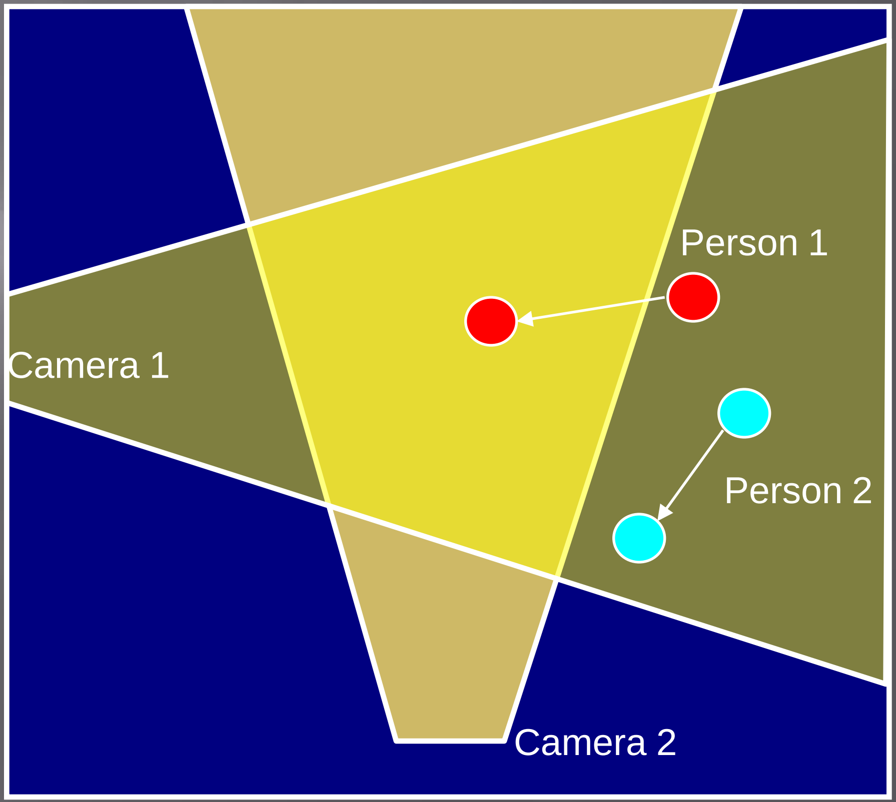
Completely Covering an Environment



University of Central Florida VISION

Ambiguity in Consistent Labeling

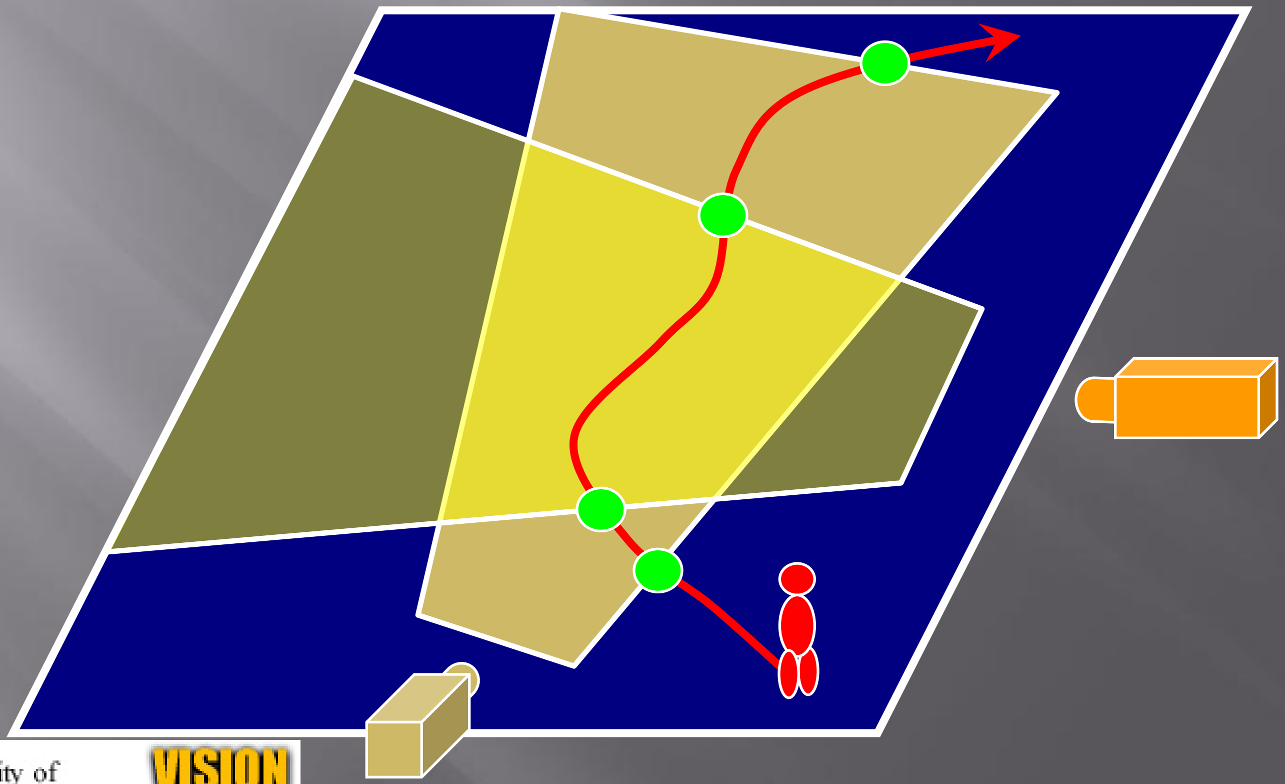
- What is the label of new person entering FOV of Camera 2?



University of Central Florida VISION

View Events

- A view-event is an instant in time when an object enters or leaves the FOV of a camera



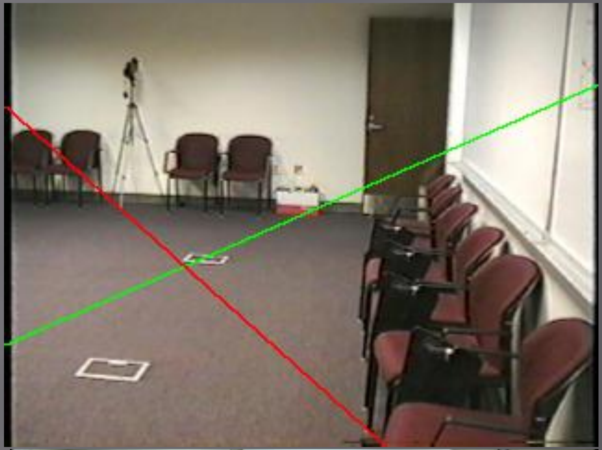
University of Central Florida VISION

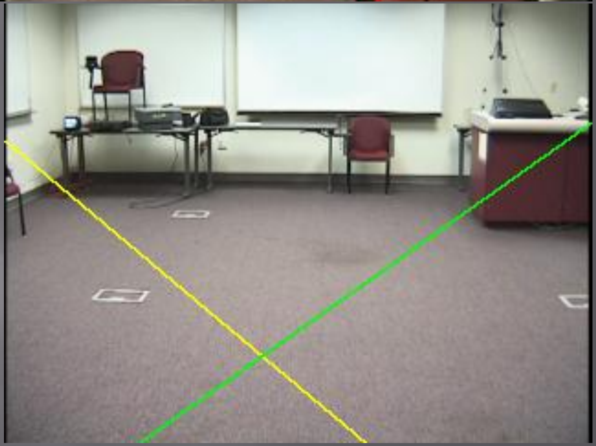
FOV lines

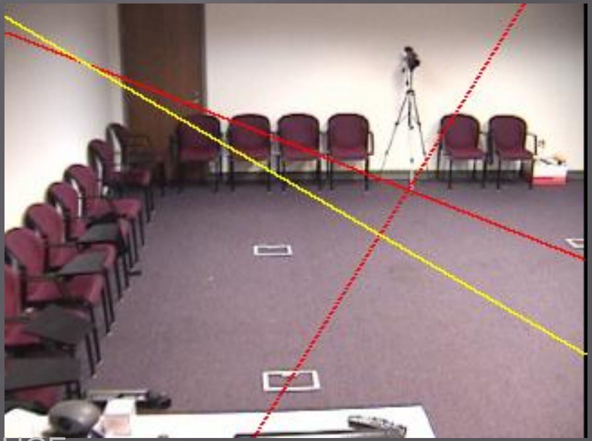
- Track the bottom of the bounding box
- Two non-ambiguous correspondences can mark edge of FOV line
- This line can be used for future ambiguities

Experiments - 1

- Three Cameras, indoor environment
- Training:
  - One person walked in the room for about 20 sec

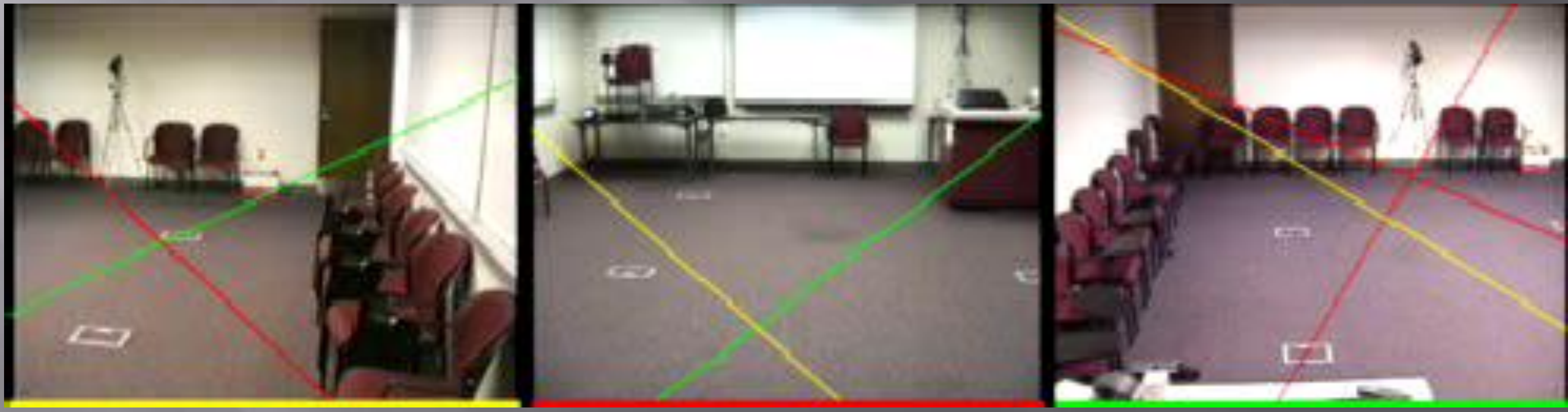






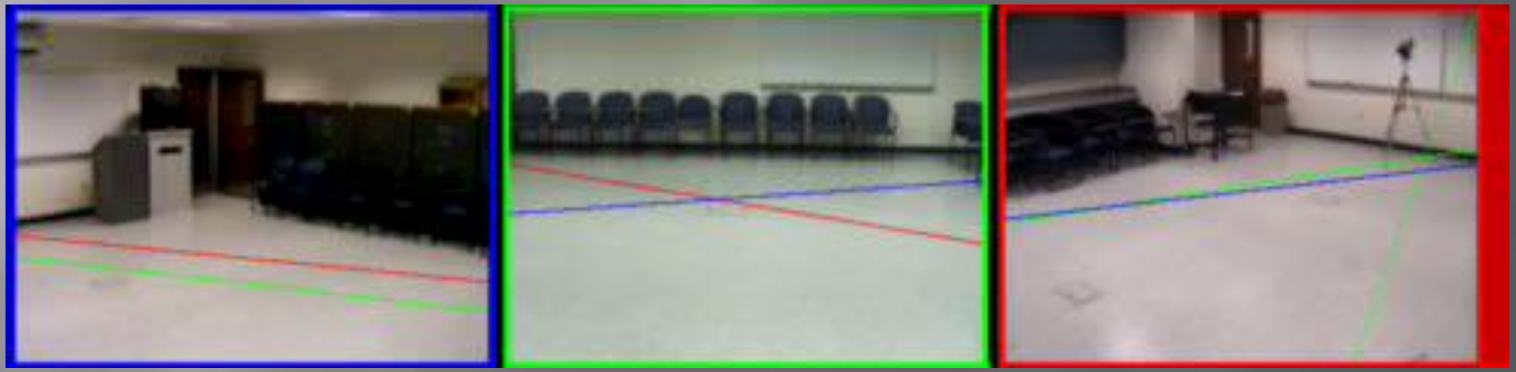
Results

Camera1 Camera2 Camera3

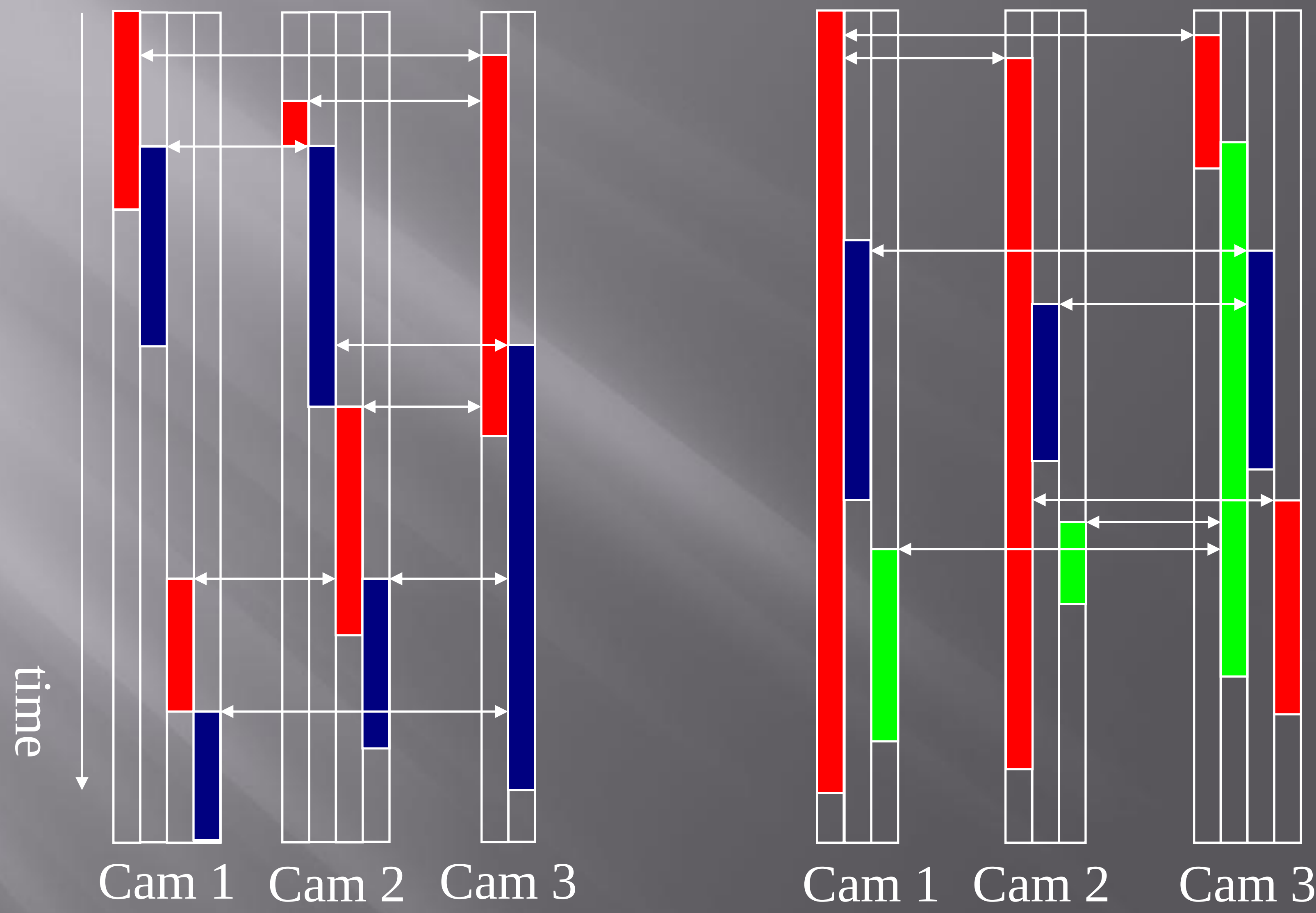


Multiple Fixed & Overlapping Cameras Tracking

Camera1 Camera2 Camera3



University of Central Florida VISION


time
Cam 1
Cam 2
Cam 3
Cam 1
Cam 2
Cam 3

TRACKING ACROSS MULTIPLE FIXED NON-OVERLAPPING CAMERA

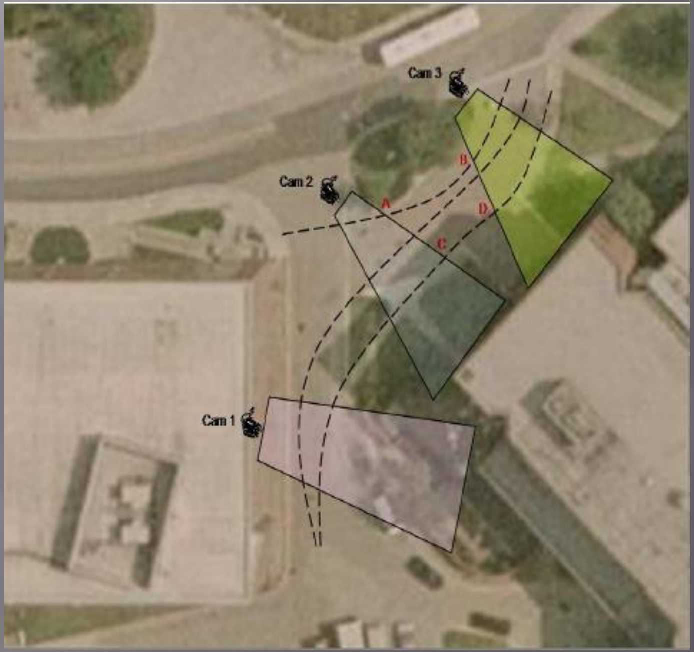
Omar Javed, Khurram Shafique and Mubarak Shah

ICCV2003, CVPR2005

University of Central Florida VISION

Tracking Across Multiple Cameras

- Task Definition:
  - To maintain the identity of objects as they move across multiple cameras



University of Central Florida VISION

Motivation

- Wide area surveillance requires tracking over disjoint views.
  - Camera Resolution
  - Occlusion due to scene structures
- Luxury of calibrated cameras is not available in most cases.

Introduction

- Problems
  - Successive observations of object might be widely separated in space and time.
  - Appearance of objects change across cameras

    –Difference in illumination

    –Difference in camera parameters (focal length, gain, response function

    –Difference in pose of object

University of Central Florida VISION

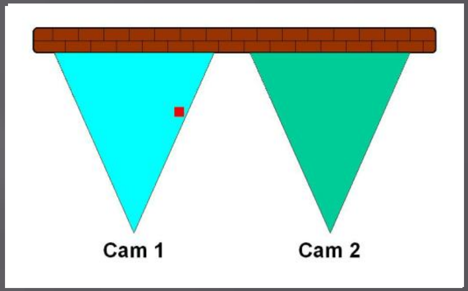
Key Observations

- Due to physical and practical constraints, some paths are more likely to be taken than others, e.g., roads, walkways, corridors.
- The observed motion pattern provides clues about inter-camera relationship.
- The transformation between the color distributions of two views of an object lies in a low dimensional space.

University of Central Florida VISION

Features for Multi-Camera Tracking

- Location of exit and entry from one camera to another.
- Time Interval between observations
  - Magnitude of motion of object.
  - Direction of motion
  - Location of exit and entry
- Appearance.



University of Central Florida VISION

Probabilistic Framework

- We have a system of l cameras $C_1, C_2, \ldots C_l$.
- There are n objects $p_1, p_2, \ldots p_n$ that generate a sequence of tracks in the system of cameras in successive time instances.
- $O_j = \{O_{j,1}, O_{j,1}, \ldots O_{j,n}\}$ be the set of observations that were observed by C_j.
- Each observation $O_{j,a}$ is composed of two independent feature sets, which are appearance $O_{j,a}(A)$ and space-time (location,time,velocity) features $O_{j,a}(ST)$.
- Single camera tracks are available

University of Central Florida VISION

Probabilistic Framework

MAP Solution: $K' = \arg\max_{K \in \Sigma} P(K \mid O)$

For $K = \{k_{a,b}^{c,d}\}$ in solution space Σ

$$P(K \mid O) = P(K \mid O_1, O_2, \ldots, O_r) = \prod_{k_{i,a}^{j,b} \in K} P\left(k_{i,a}^{j,b} \mid O_{i,a}, O_{j,b}\right)$$

Assuming independence

$$P(K \mid O) = \prod_{k_{i,a}^{j,b} \in K} \frac{P(O_{i,a}(A), O_{j,b}(A) \mid k_{i,a}^{j,b}) P(O_{i,a}(ST), O_{j,b}(ST) \mid k_{i,a}^{j,b}) P(k_{i,a}^{j,b})}{P(O_{i,a}, O_{j,b})}$$

Using Bayes Law and Assuming Independence
between appearance and spatio-temporal observations

Probabilistic Framework

- Maximizing the following term will give us the solution

$$K' = \arg\max_{K \in \Sigma} \sum_{k_{i,a}^{j,b} \in K} P(O_{i,a}(A), O_{j,b}(A) \mid k_{i,a}^{j,b}) P(O_{i,a}(ST), O_{j,b}(ST) \mid k_{i,a}^{j,b}) P(C_i, C_j)$$

University of Central Florida VISION

Learning Probability Density Functions

- Learning phase
  - Assumption of known correspondences.
    - One solution is to use only appearance matching for correspondences. Note that only week or ambiguous matches can be discarded during the training phase.
  - Estimate spatio-temporal and appearance pdfs from the observed data.

University of Central Florida VISION

Estimating Spatio-Temporal Pdf

- Features:
  - Entry and Exit Locations and Cameras
  - Velocity
  - Inter-Camera Travel Time
- Parzen Windows for density estimation
- For a sample S, consisting of '*n*' data points $x_1, x_2, ..x_n$, the Parzen estimate is given as

$$\hat{p}(x) = \frac{1}{n} |H|^{-\frac{1}{2}} \sum_{i=1}^{n} K(H^{-\frac{1}{2}}(x - x_i))$$

University of Central Florida VISION

Estimating Appearance Pdf

- Goal: Learn the change in appearance (color) of an object as it moves from one camera to other.

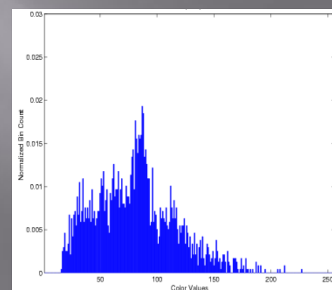
Same person in two different cameras

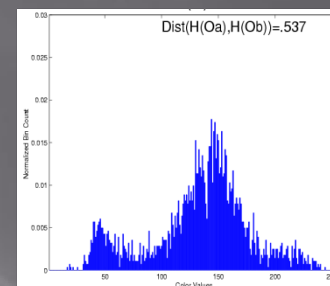
Effectiveness of Subspace Learning

$O_{i,a}$ $O_{j,b}$

Same person in cameras C_i & C_j

$H(O_{i,a})$ $H(O_{i,b})$





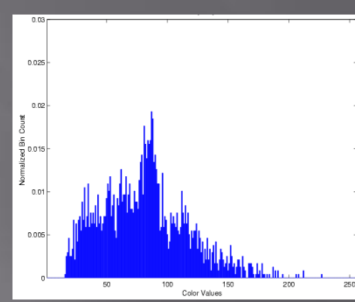
Red Channel Histogram Similarity

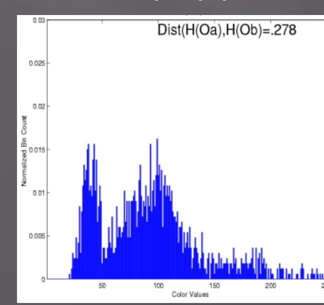
$$BhattacharrayDis\tan ce = d = 0.537$$

$O_{i,a}$ $O_{j,c}$

Different persons in cameras C_i & C_j

$H(O_{i,a})$ $H(O_{i,c})$





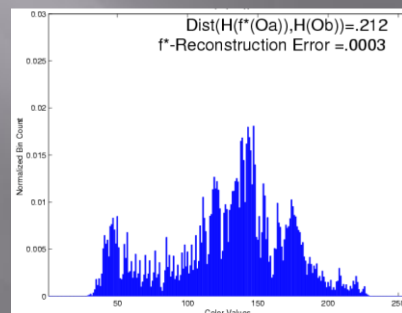
Red Channel Histogram Similarity

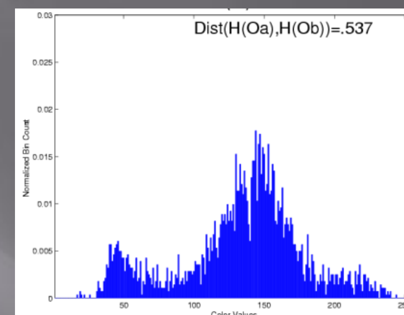
$$d = 0.278$$

University of Central Florida VISION

Effectiveness of Subspace Learning

$O_{i,a}$ $O_{j,b}$ $H(f_1(O_{i,a}))$ $H(O_{i,b})$



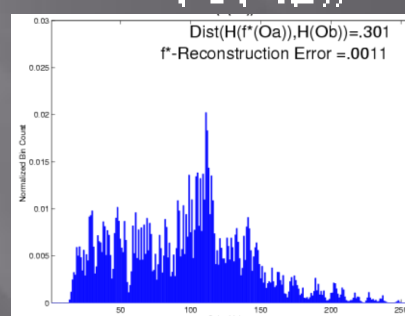


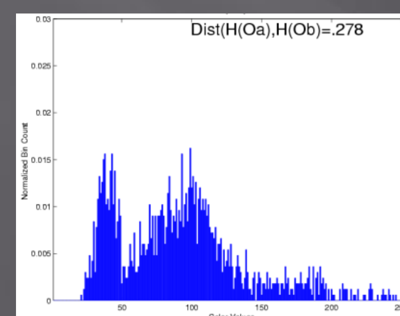
Histogram Similarity after Transformation

$d = 0.212$

Same person in cameras C_i & C_j

$O_{i,a}$ $O_{j,c}$ $H(f_2(O_{i,a}))$ $H(O_{i,c})$





Histogram Similarity after Transformation

Different persons in cameras C_i & C_j

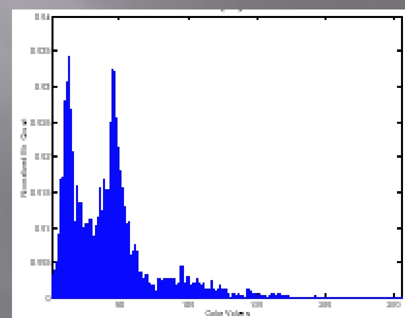
$d = 0.301$

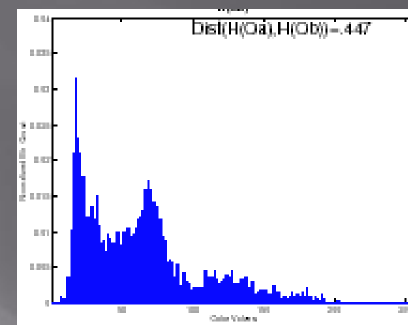
University of Central Florida VISION

Effectiveness of Subspace Learning

$O_{i,a}$ $O_{j,b}$

$H(O_{i,a})$ $H(O_{j,b})$



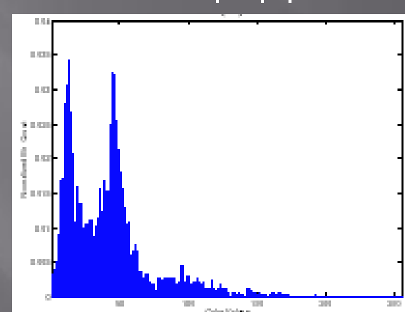


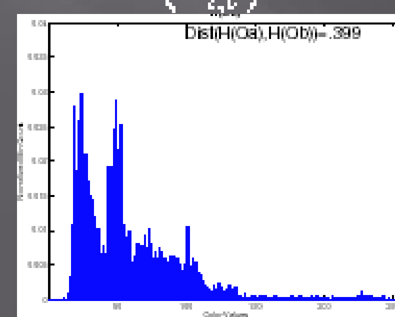
Blue Channel Histogram Similarity

Same person in cameras C_i & C_j

$d = 0.447$

$H(O_{i,a})$ $H(O_{i,c})$





Blue Channel Histogram Similarity

Different persons in cameras C_i & C_j

$d = 0.399$

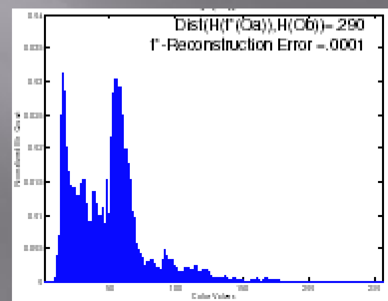
University of Central Florida VISION

Effectiveness of Subspace Learning

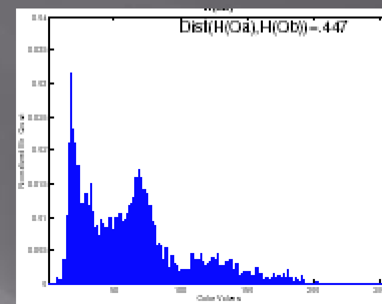
$O_{i,a}$ $O_{j,b}$

Same person in cameras C_i & C_j

$H(f_1(O_{i,a}))$



$H(O_{i,b})$



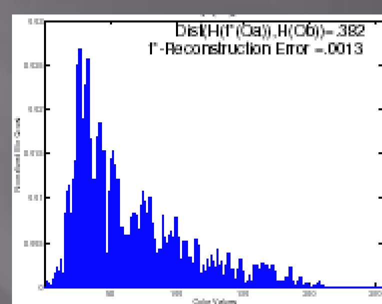
Histogram Similarity after Transformation

$d = 0.290$

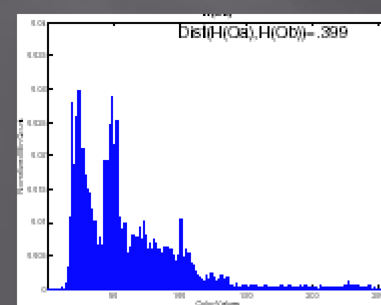
$O_{i,a}$ $O_{j,c}$

Different persons in cameras C_i & C_j

$H(f_2(O_{i,a}))$



$H(O_{i,c})$



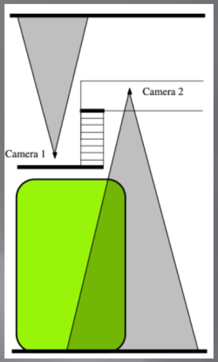
Histogram Similarity after Transformation

$d = 0.382$

University of Central Florida VISION

Results

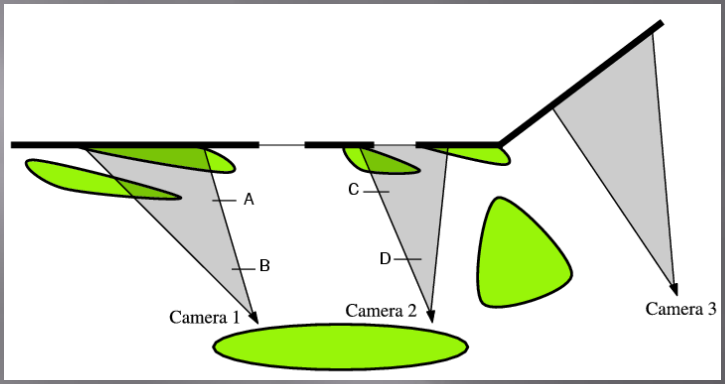
Camera Setup
for ExpeAriment # 1



A Clip from the test sequence

Results

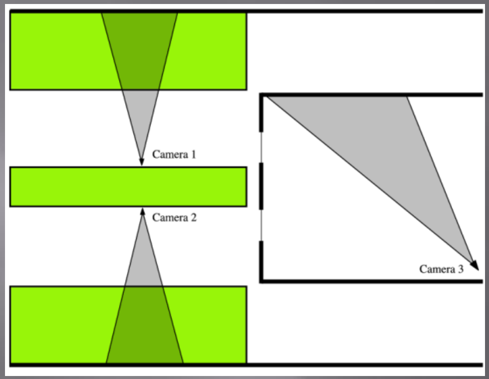
Camera Setup
for Experiment # 2



A Clip from the test sequence

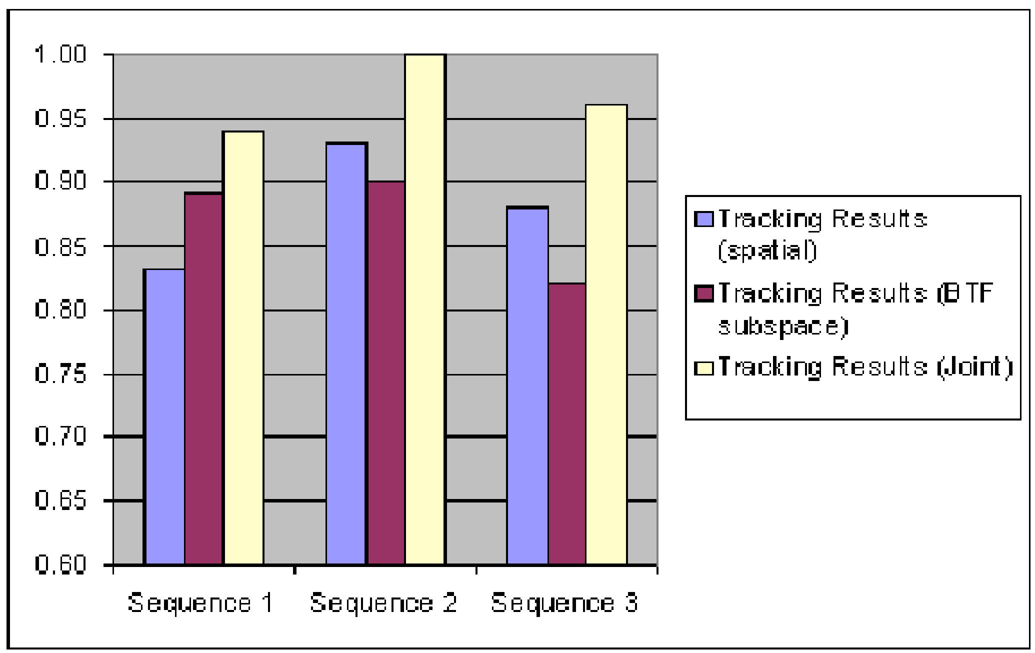
Results

Camera Setup
for Experiment # 2

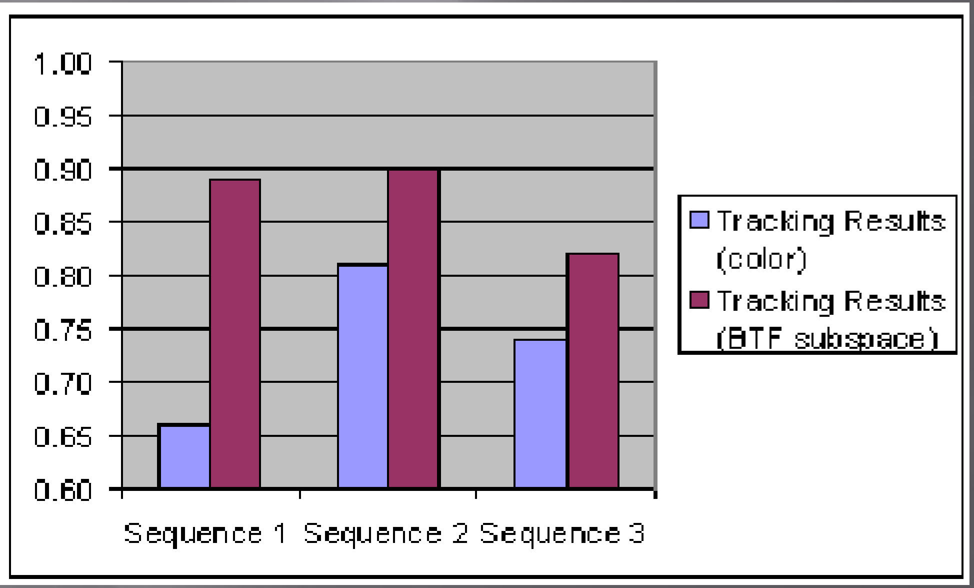


A Clip from the test sequence

Results (Tracking Accuracy)



Results (Comparison with direct Color Matching)

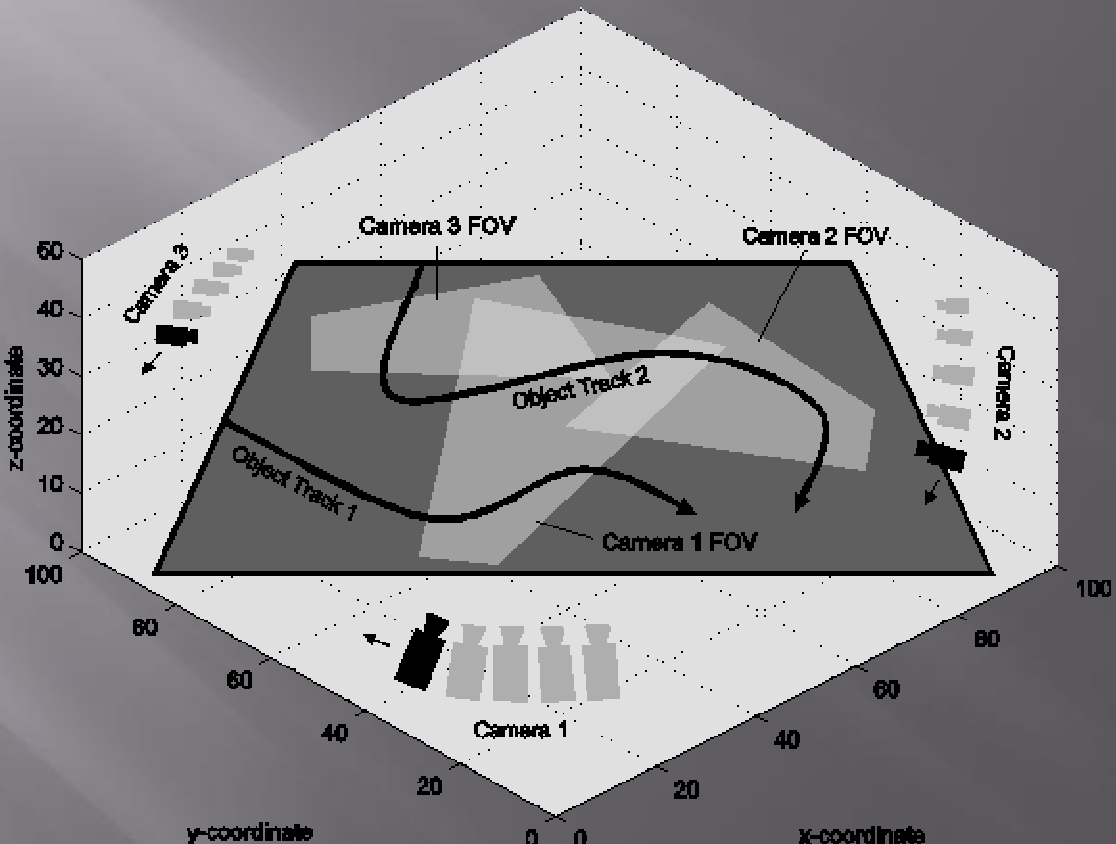


TRACKING ACROSS MULTIPLE MOVING AIRBORNE CAMERAS

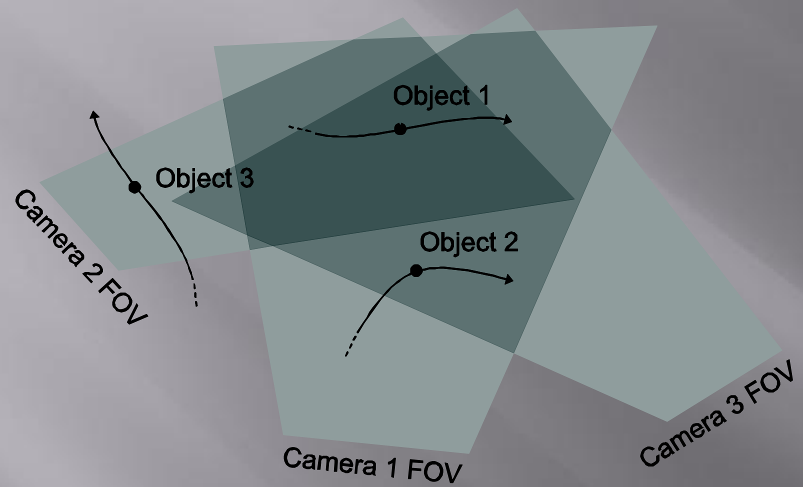
Yaser Sheikh & Mubarak Shah
ICCV 2005

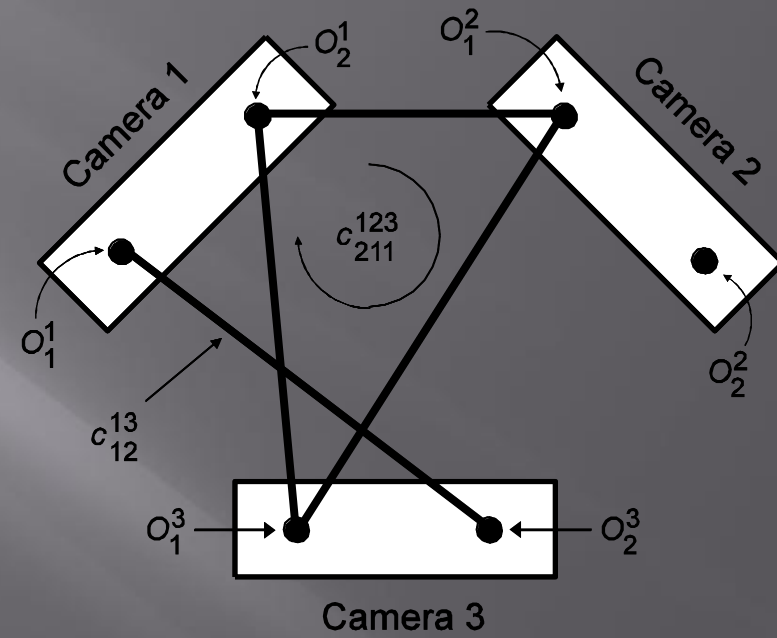
University of Central Florida VISION

Motion in the Forest



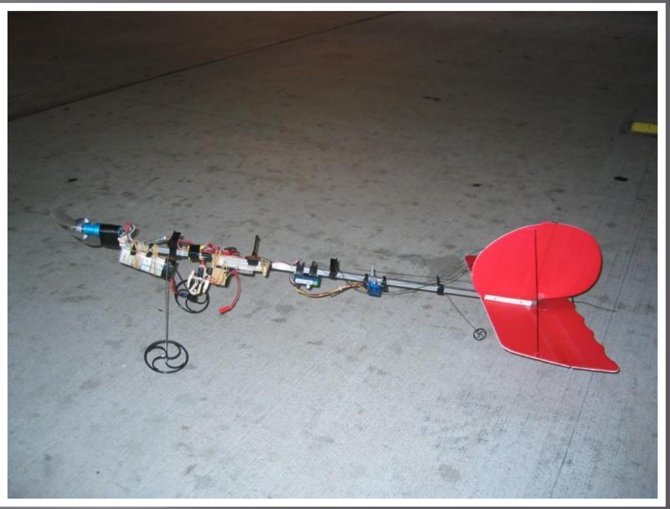
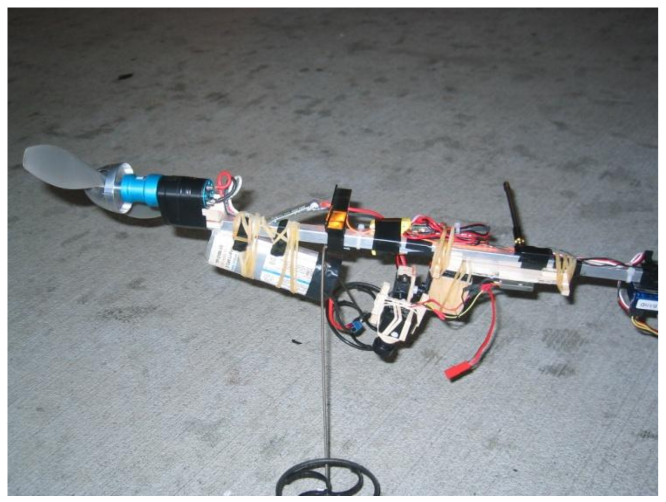
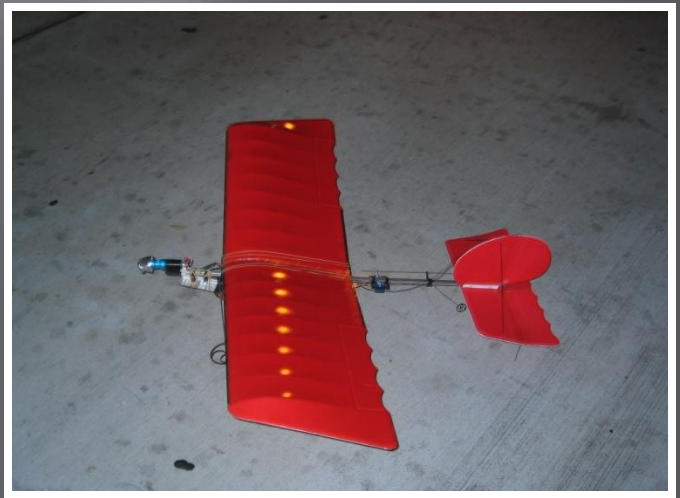
Graph Theoretic Formulation



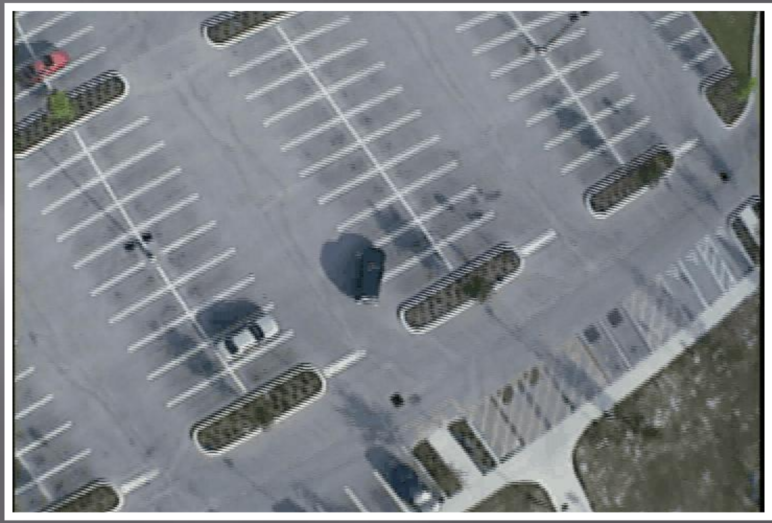


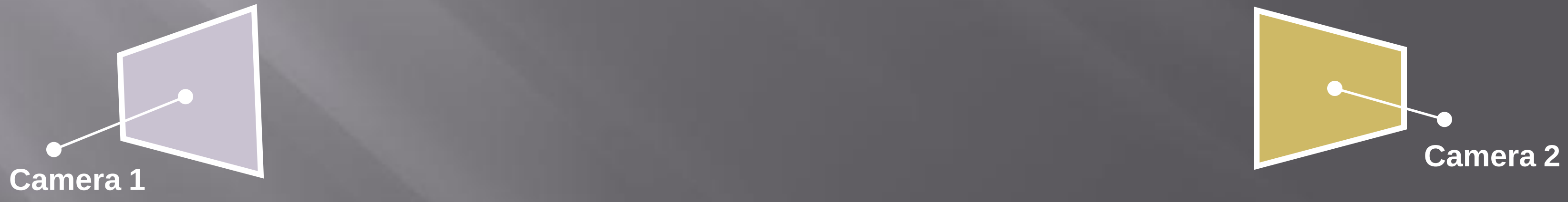
- **Three Objects, three cameras**
  Object 1 is visible in all three cameras
  Object 2 is visible in Camera 1 and Camera 3
  Object 3 is visible only in Camera 2

UCF's UAVs

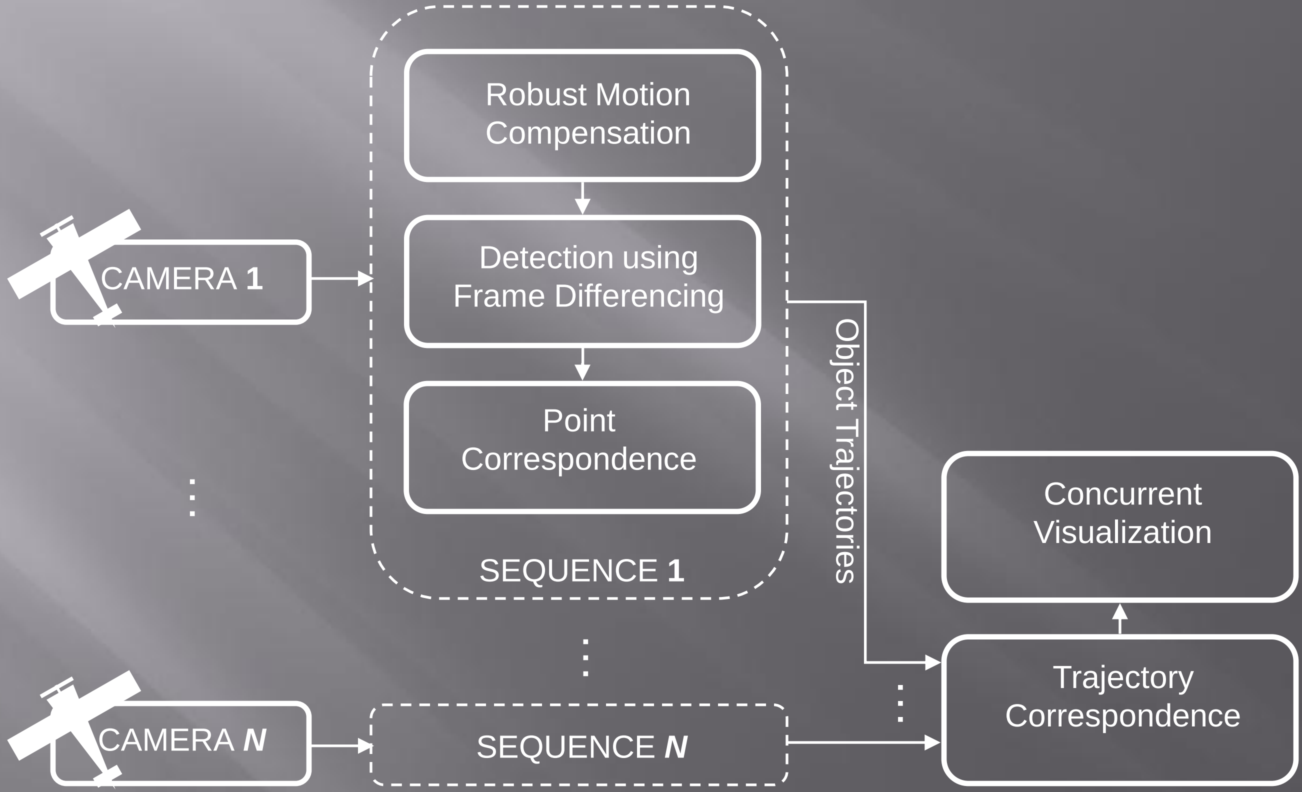





Example Input Data





System Overview

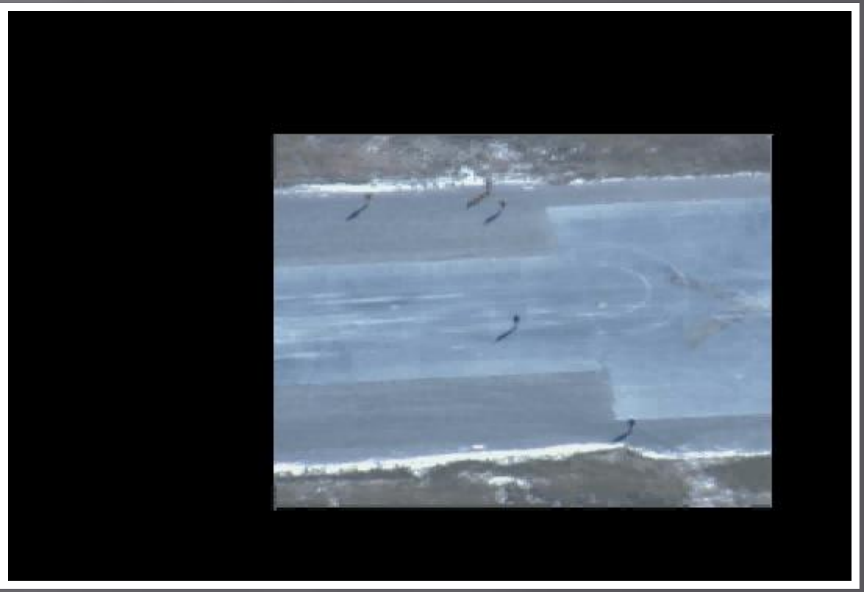


Assumptions

- Assumption of Scene Planarity
  - Validated by altitude of aerial vehicle
  - Reasonable deviations do not affect solution
- Spatiotemporal Overlap of Fields of View
  - Objects are simultaneously visible in two cameras at a time (for all pairs of cameras).

Compensating Camera Motion

Original Sequence

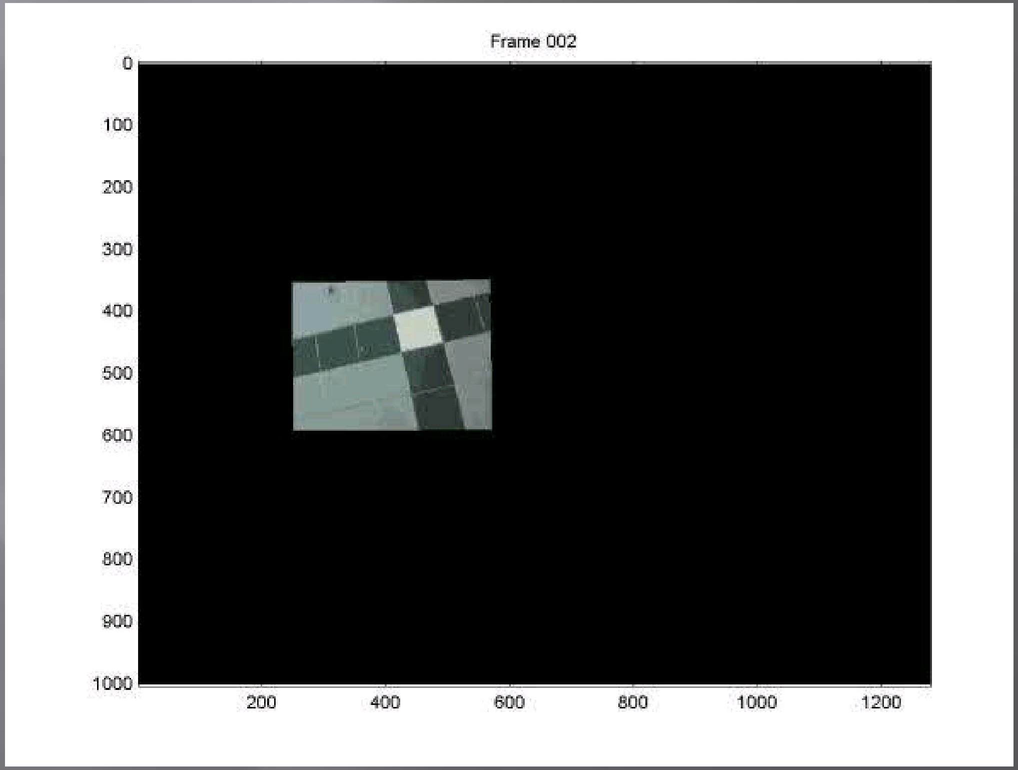


Motion Compensated

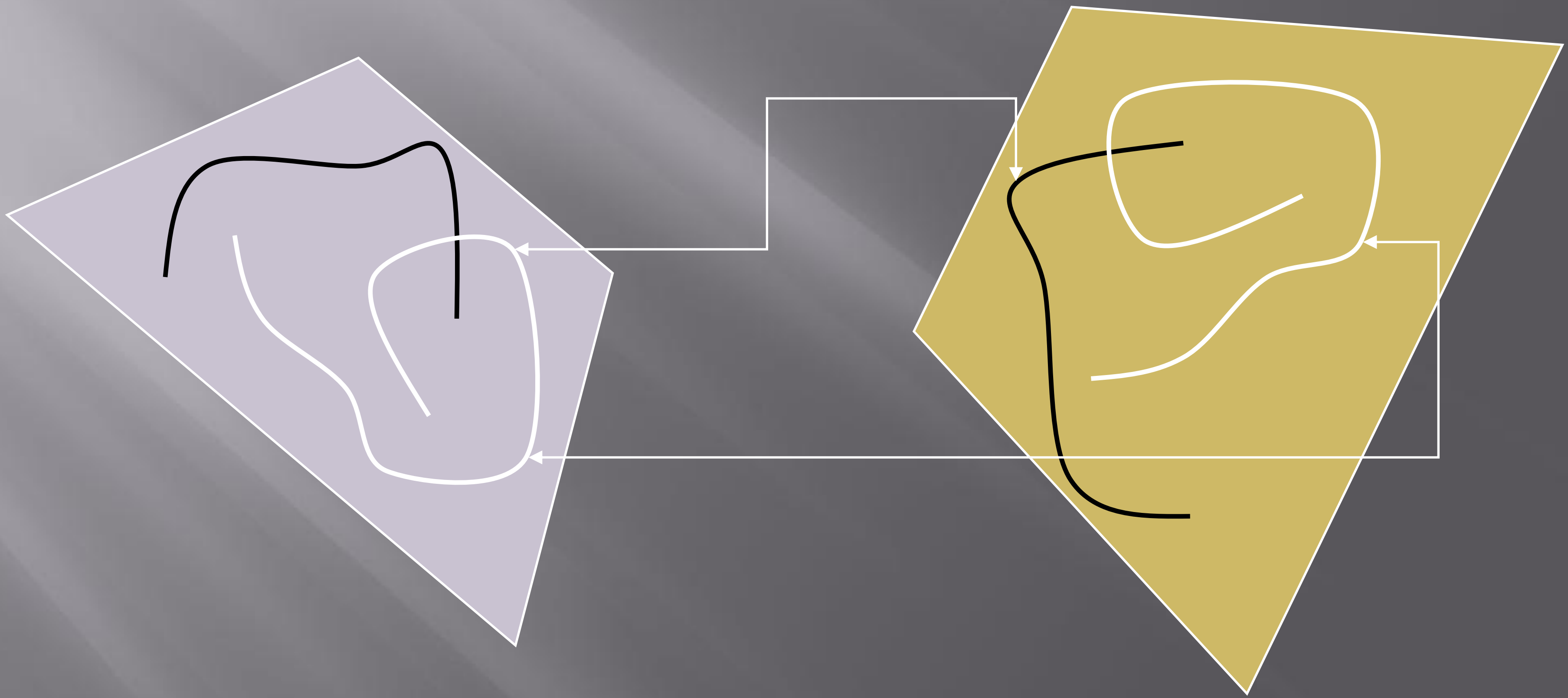
University of Central Florida VISION

Detection and Tracking

- We are not interested in solving the problem of tracking *within* cameras in this work



Similarity Metric



Data Modeling

- **Data** Two trajectories, $\mathbf{x}_1$ and $\mathbf{x}_2$. Each point modeled as a random variables with independent Gaussian noise,

$$x_i = \bar{x}_i + \varepsilon, \quad \varepsilon \sim N(0,\sigma)$$

Homography $\mathbf{H}_{12}$ exists between the two trajectories since we model the scene as a plane.

- The pair of tracks $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ related *exactly* by $\mathbf{H}_{12}$
- Maximum Likelihood Estimate of ***H***

$$\arg\max_{\bar{x}, H} \prod_i \left(\frac{1}{2\pi\sigma^2}\right) e^{-(d(x_{1i},\bar{x}_{1i})^2 + d(x_{2i},\bar{x}_{2i})^2)/2\sigma^2}$$

$$\arg\max_{\bar{\mathbf{x}}_1, H} \prod_i \left(\frac{1}{2\pi\sigma^2}\right) e^{-(d(x_{1i},x_{2i}))^2/2\sigma^2}$$

Correspondence

- We wish to compute,

$$P(x \mid X) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i - \hat{x}_i)^2}{2\sigma^2}} \right)$$

- Taking the log,

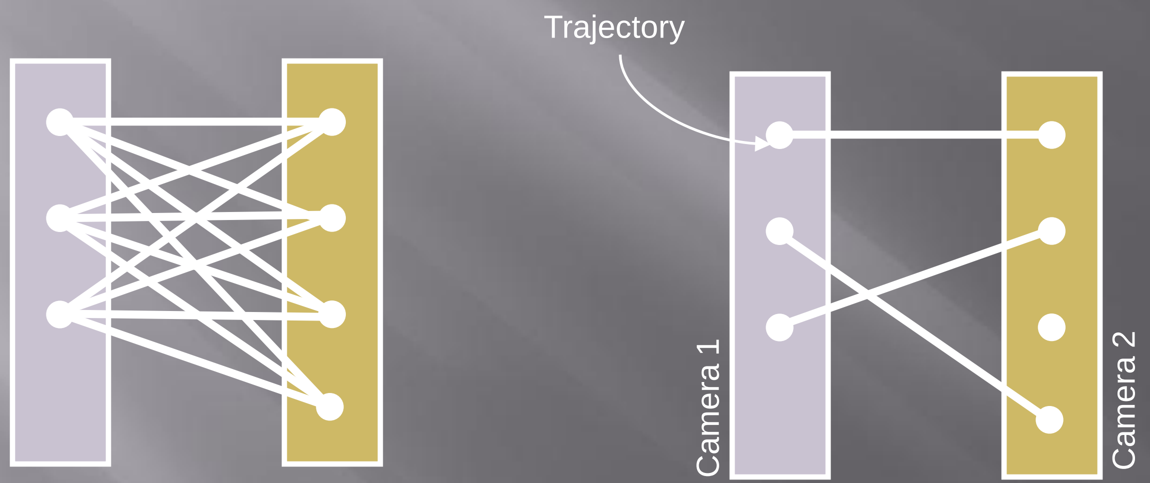
$$\log P(x \mid X) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

  - Intuitively this is using the estimate of the statistical *mean* of the reprojection error at each point.
  - If outliers exists, robust estimators, such as the *median* of the reprojection error can be used

University of Central Florida VISION

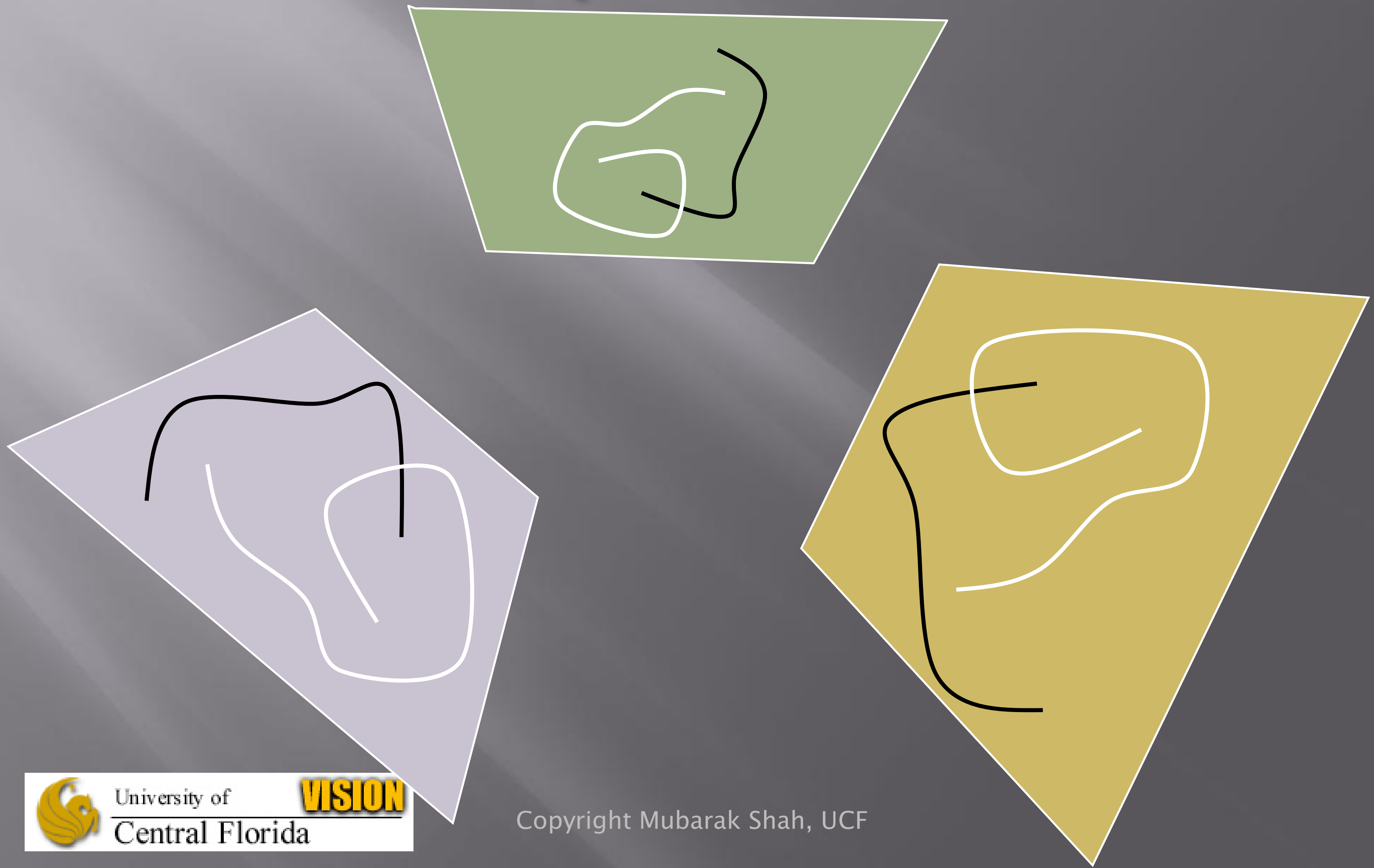
Correspondence Across 2 Cameras

- Edge weight is $p(c_{ij}^{pq} | \mathbf{x}_i^p, \mathbf{x}_j^q)$



- Optimal correspondence in a ML sense can be found using bipartite matching

Trajectory Correspondence Across Multiple Cameras



University of Central Florida VISION

Global Correspondence

- For multiple cameras, we need to find the correspondence C such that,

$$C = \arg\max_{C \in \mathbf{C}} \sum_{c_{ij}^{pq} \in C} \log P(c_{ij}^{pq} \mid X_i^p, X_j^q)$$

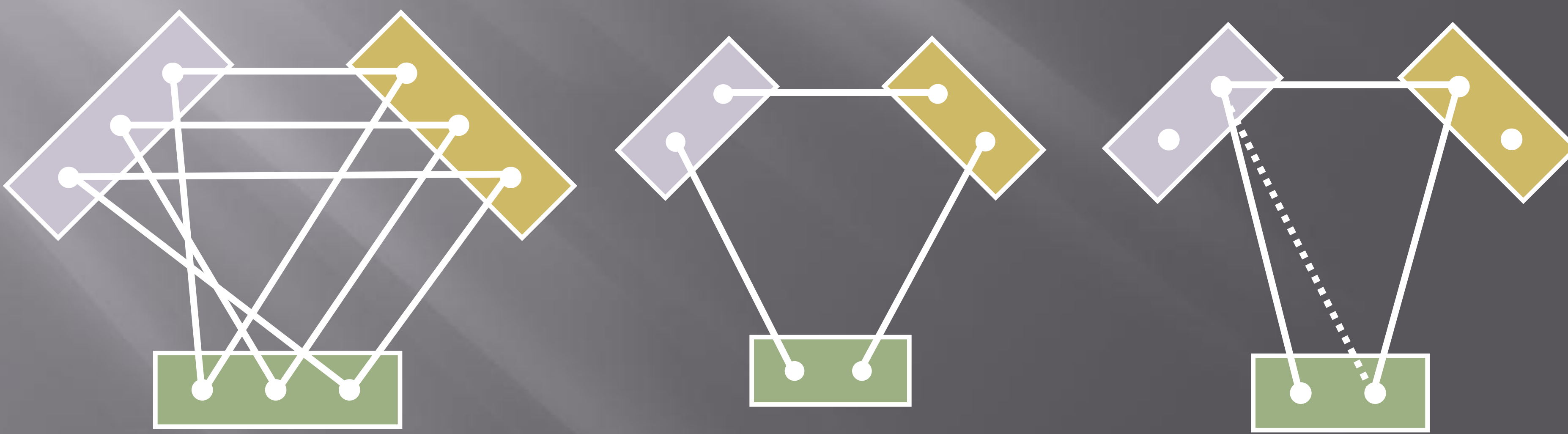
- Where,

$\mathbf{x}_i^p$	Trajectory *i* in camera *p*
c_{ij}^{pq}	Correspondence hypotheses b/w x_i^p and x_j^q
C	Global correspondence hypotheses
$\mathbf{C}$	Space of global correspondence solutions

University of Central Florida VISION

Correspondence Across Cameras

- For greater than two cameras complexity increases
- In addition another constraint needs to be satisfied: transitive closure

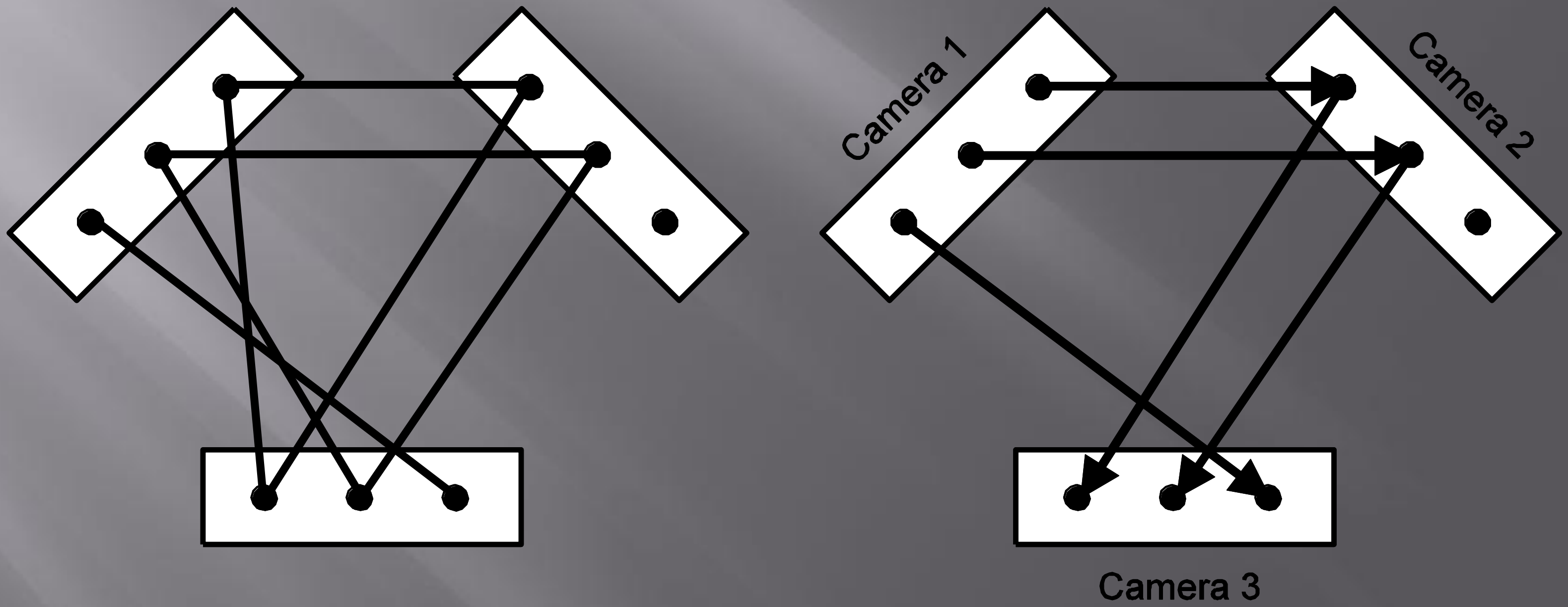


Correspondence Across Cameras

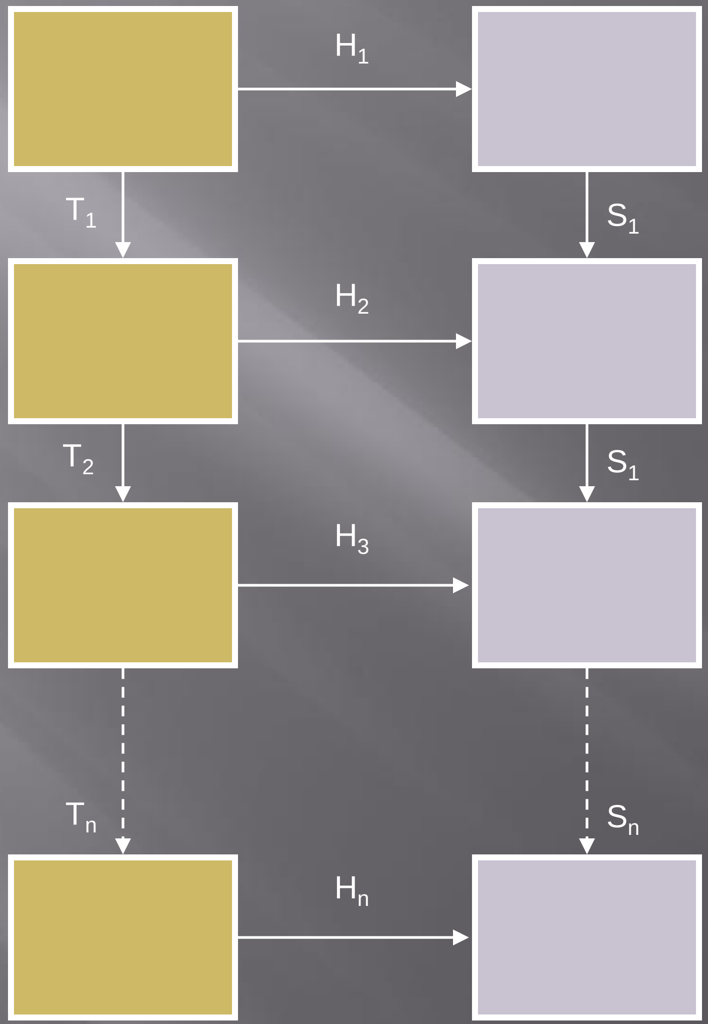
- For two sequences, the optimal solution can be obtained by maximum bipartite matching
- For $k \geq 3$, the problem of finding correspondences is known to be NP Hard.
- Therefore, we consider a reformulation of the problem as a directed graph.
- The direction comes from arbitrarily enumerating the cameras.

Correspondence Across Cameras

- A polynomial time solution exists to approximate the maximum matching (Shafique and Shah, TPAMI 2005)

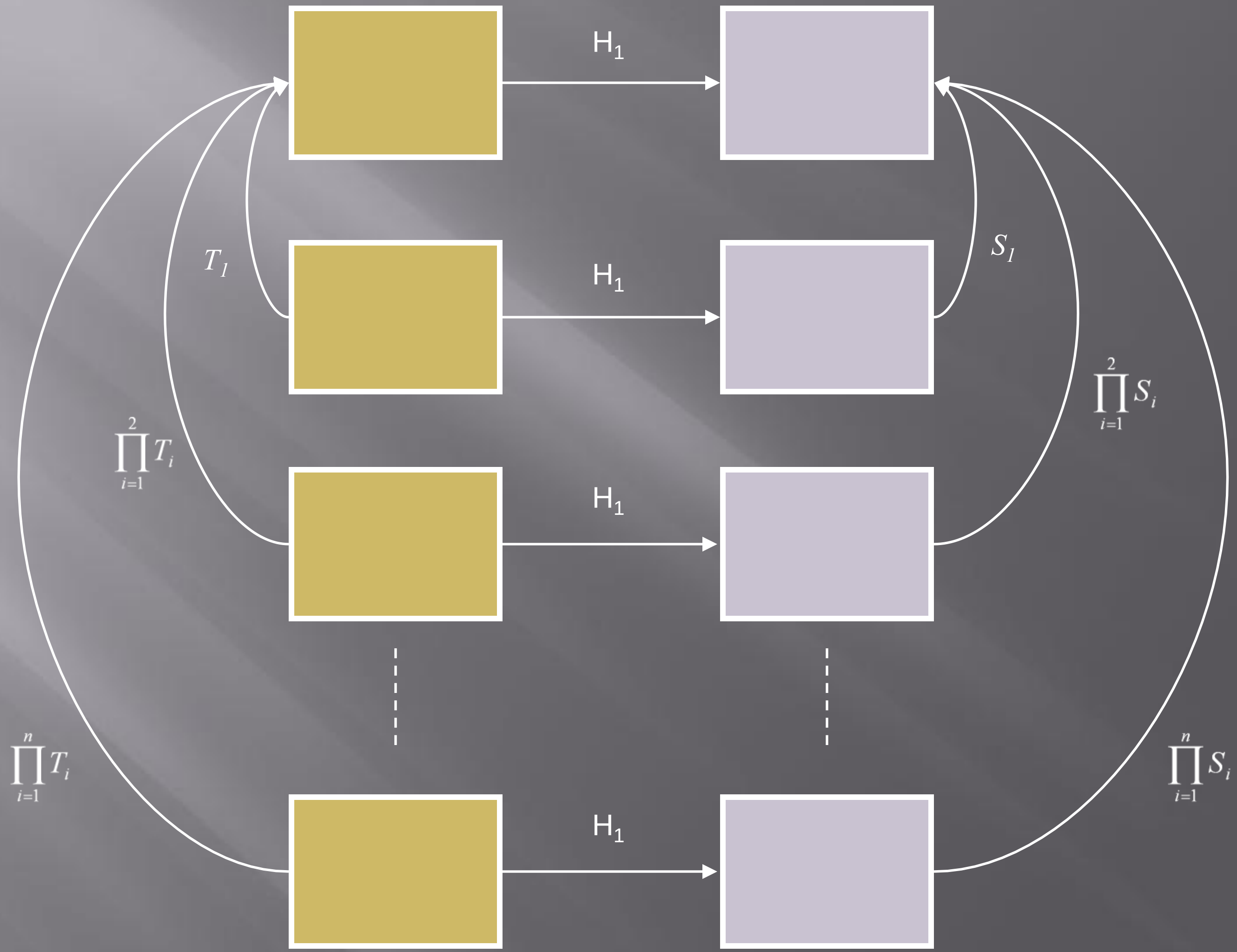


Concurrent Visualization

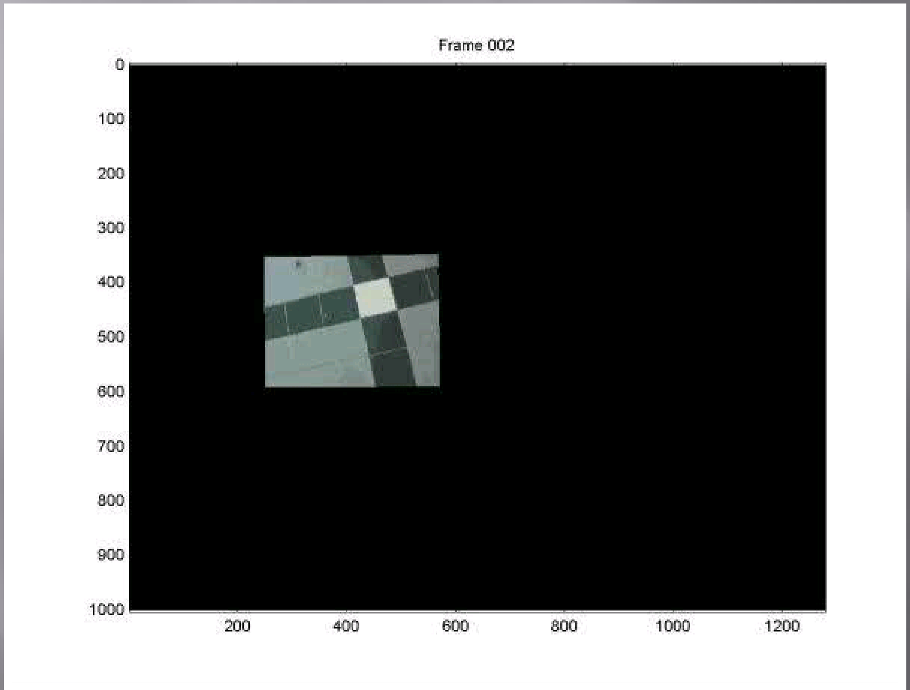


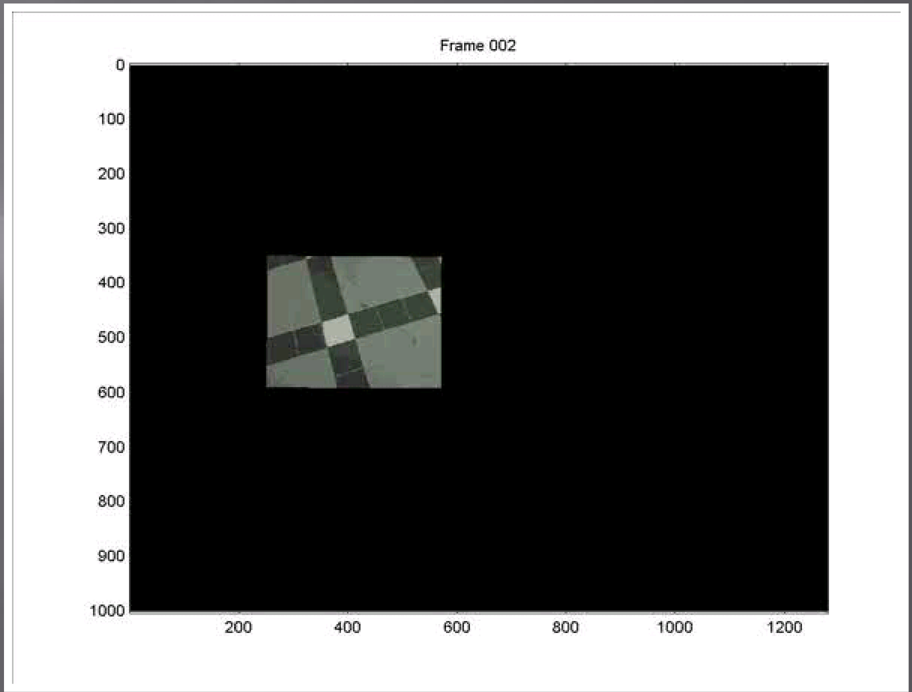
University of Central Florida VISION

Concurrent Visualization



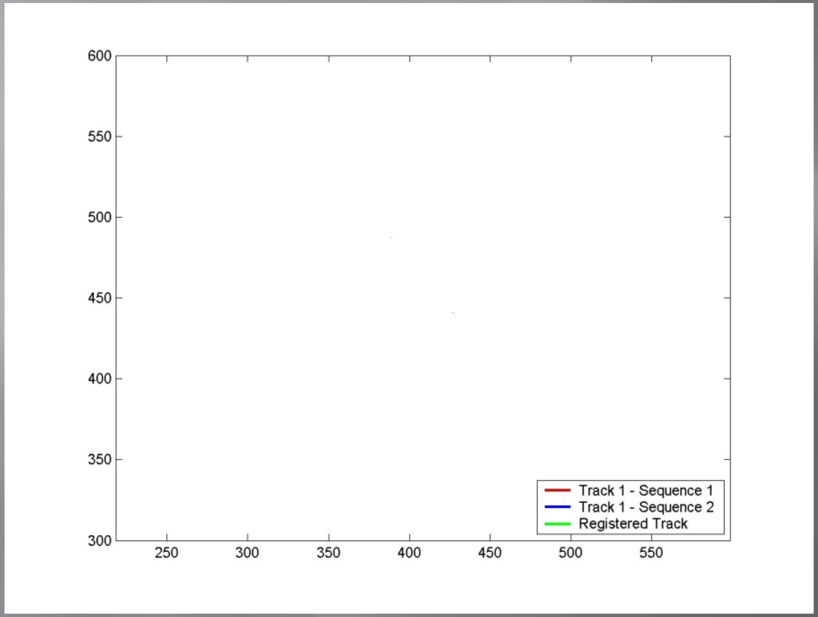
Controlled Sequence 1

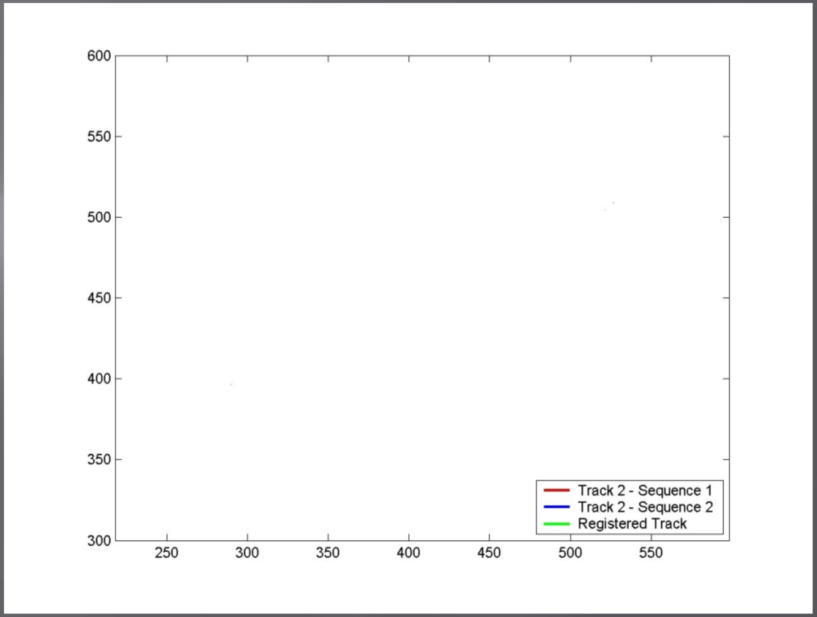




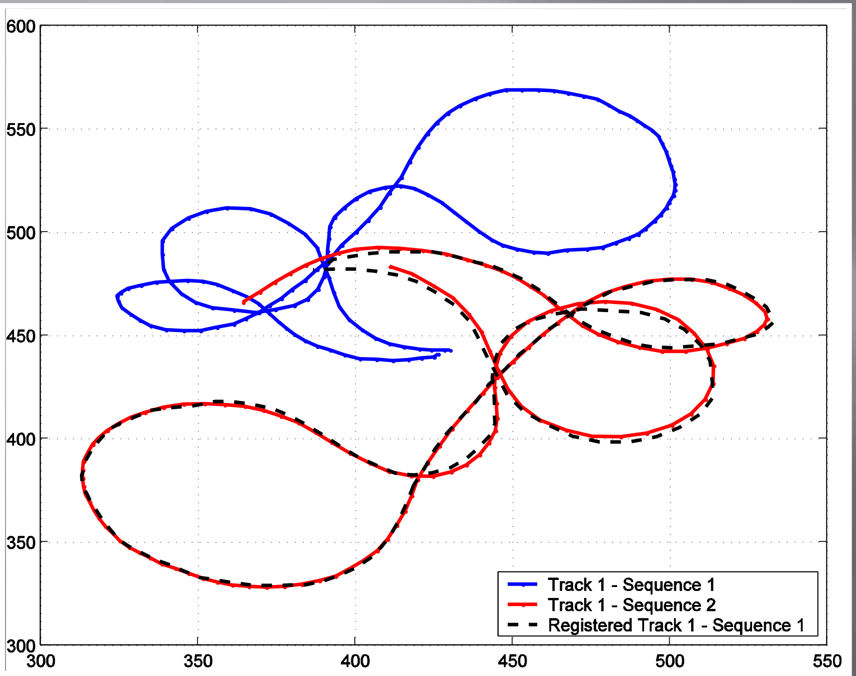
University of VISION Central Florida

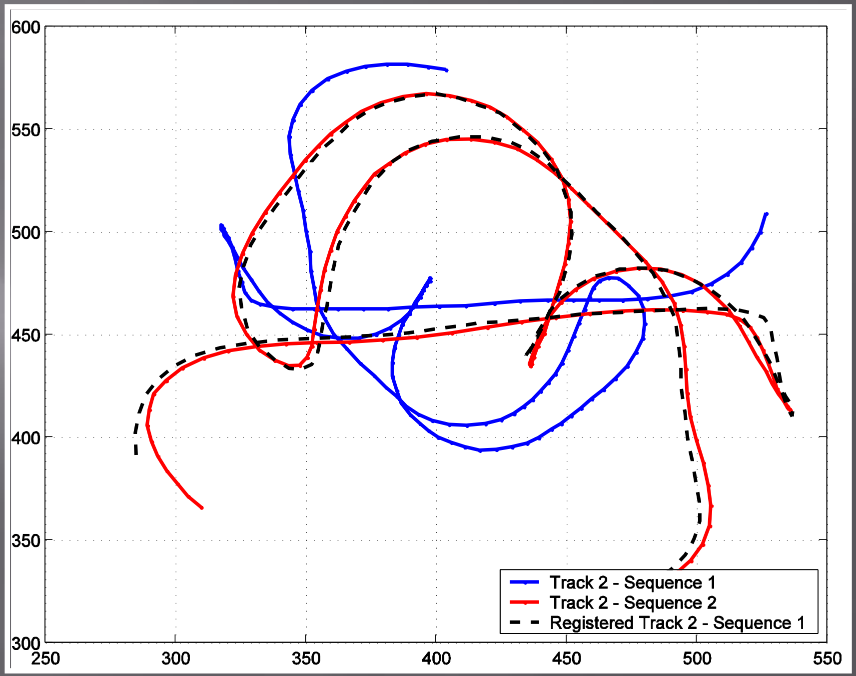
Tracks Corresponded





Still Picture…

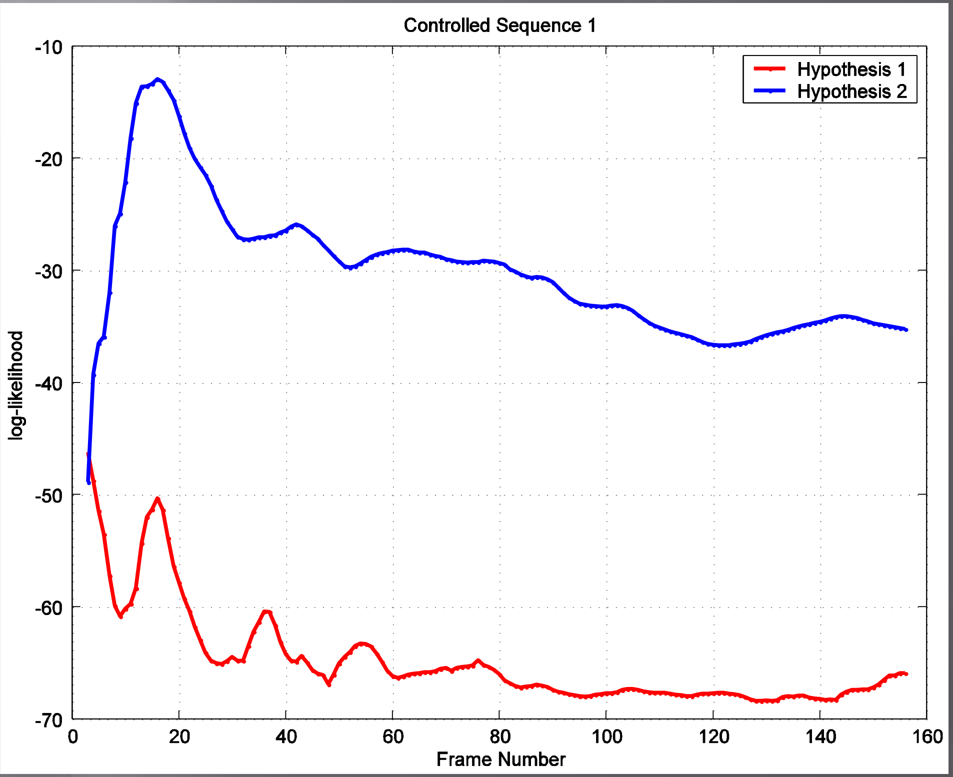




University of Central Florida VISION

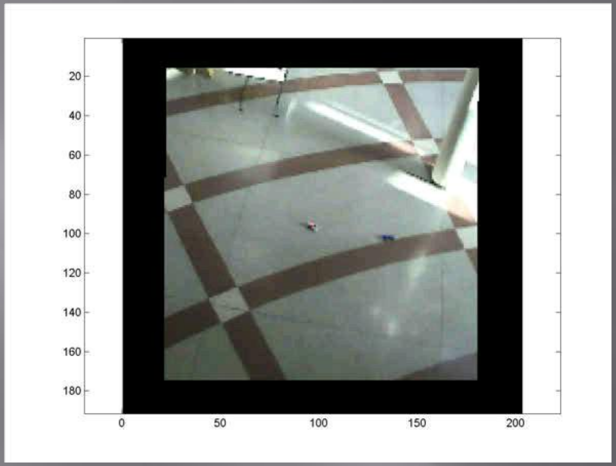
Hypotheses

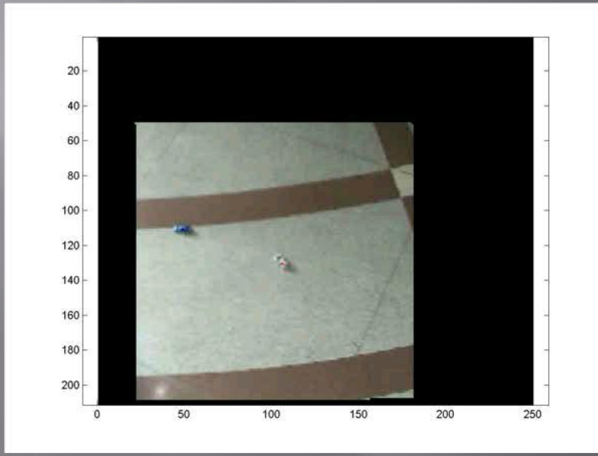
- Two hypotheses are tested

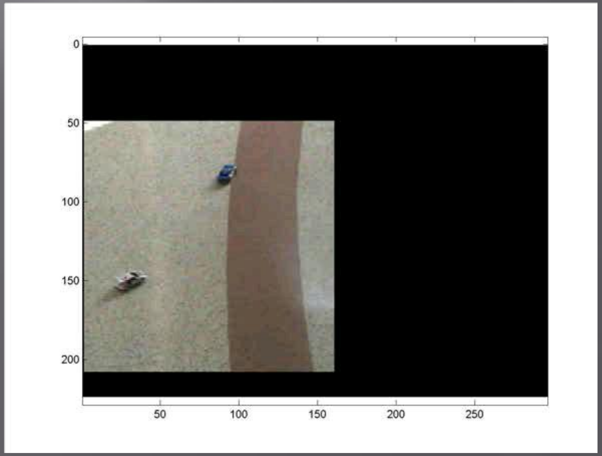


University of Central Florida VISION

Controlled Sequence 2

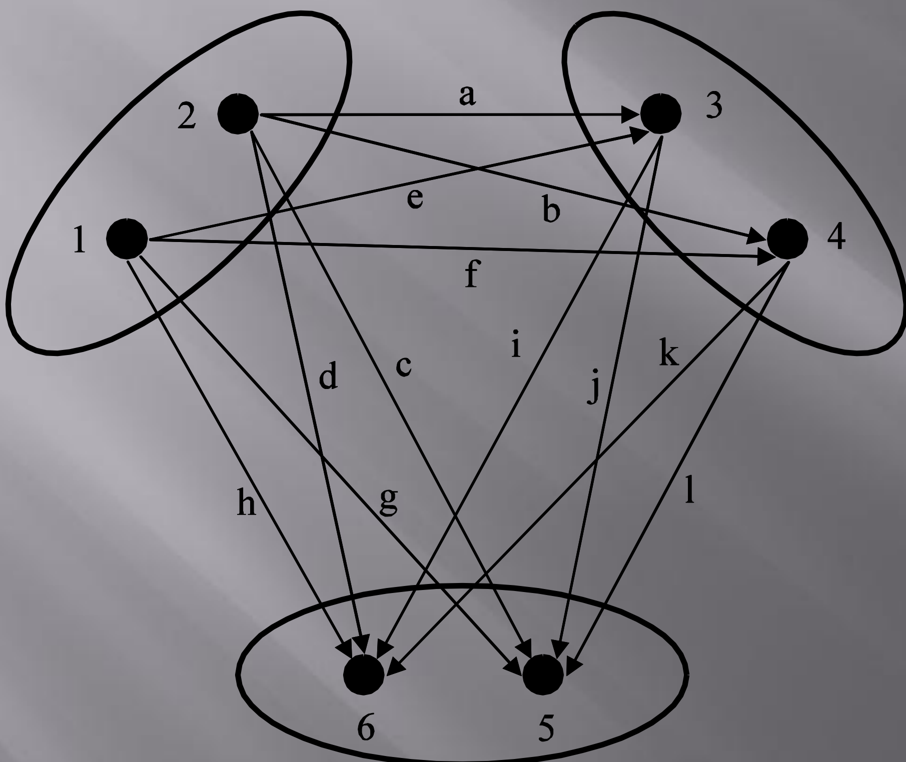


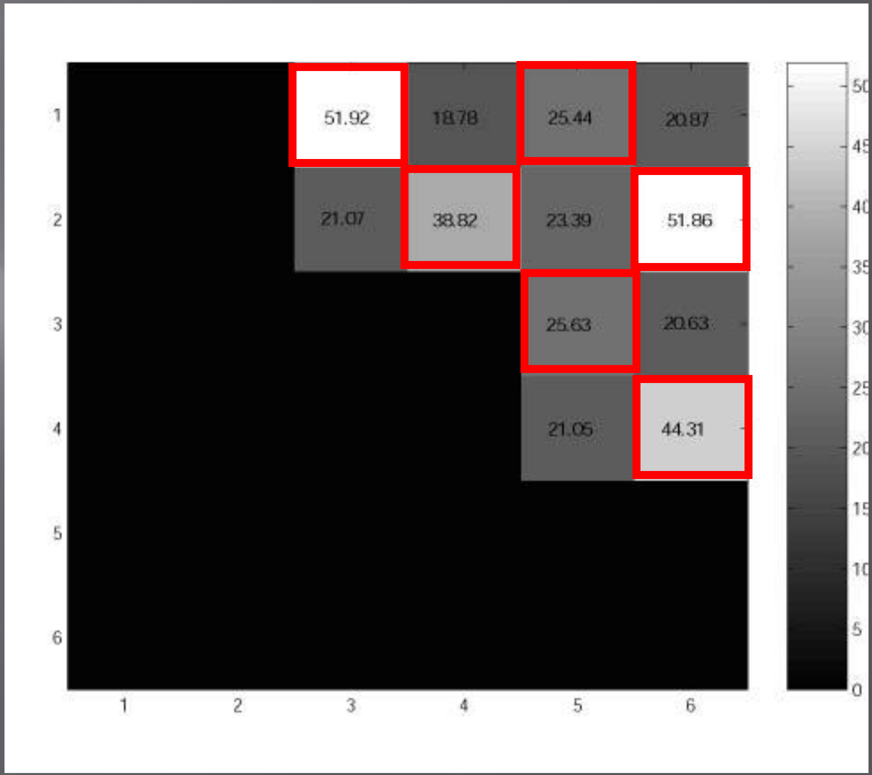




University of Central Florida VISION

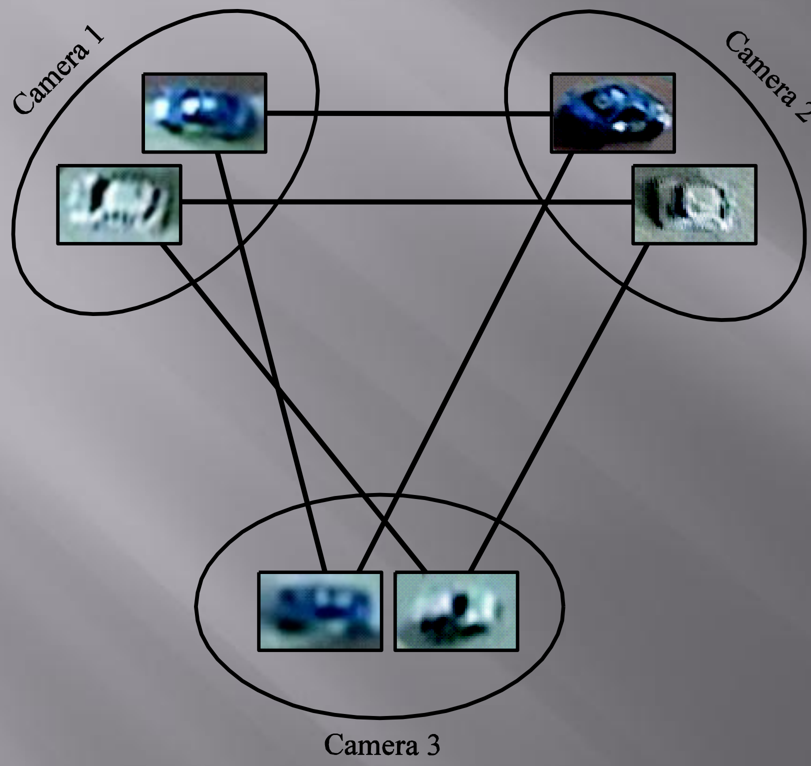
Controlled Sequence 2

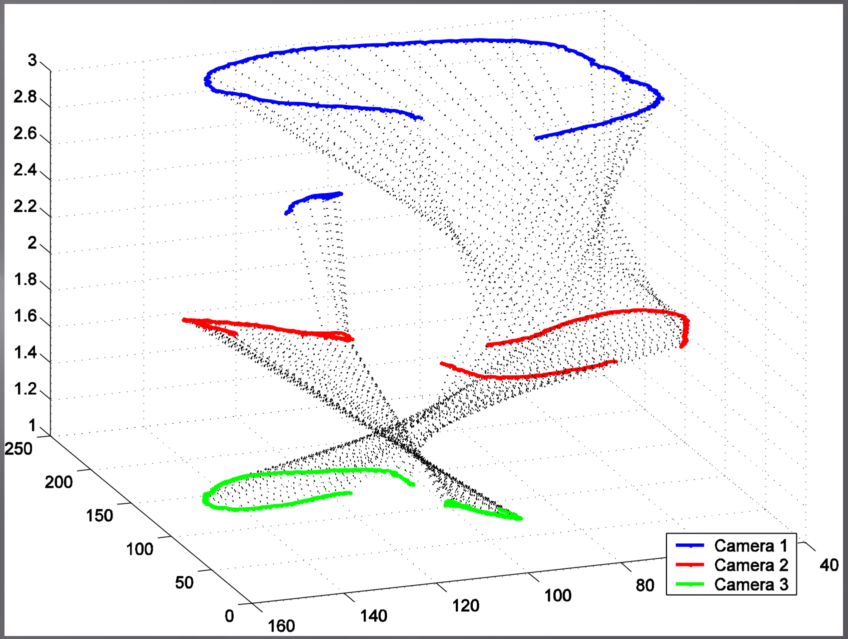




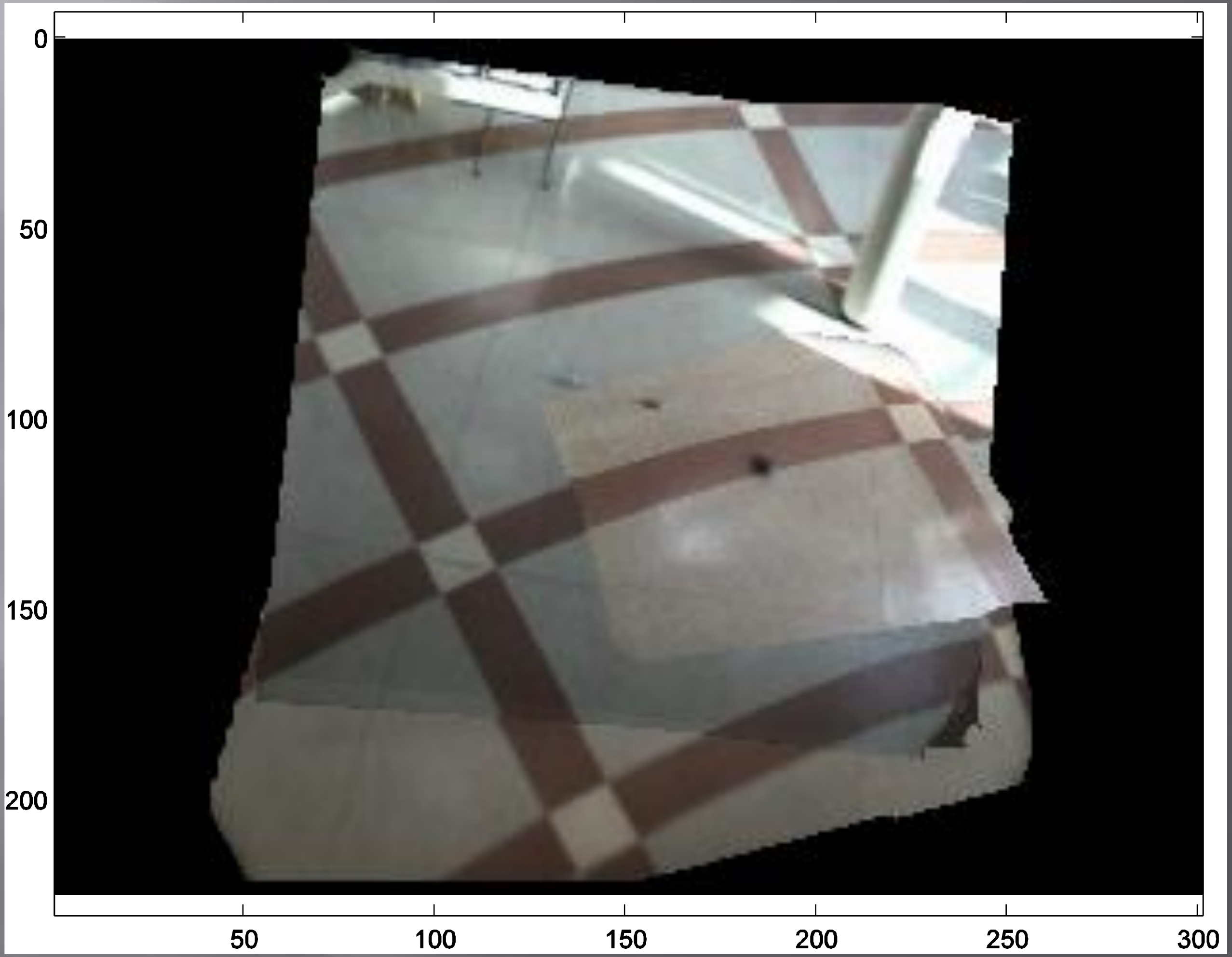
University of Central Florida VISION

Controlled Sequence 2



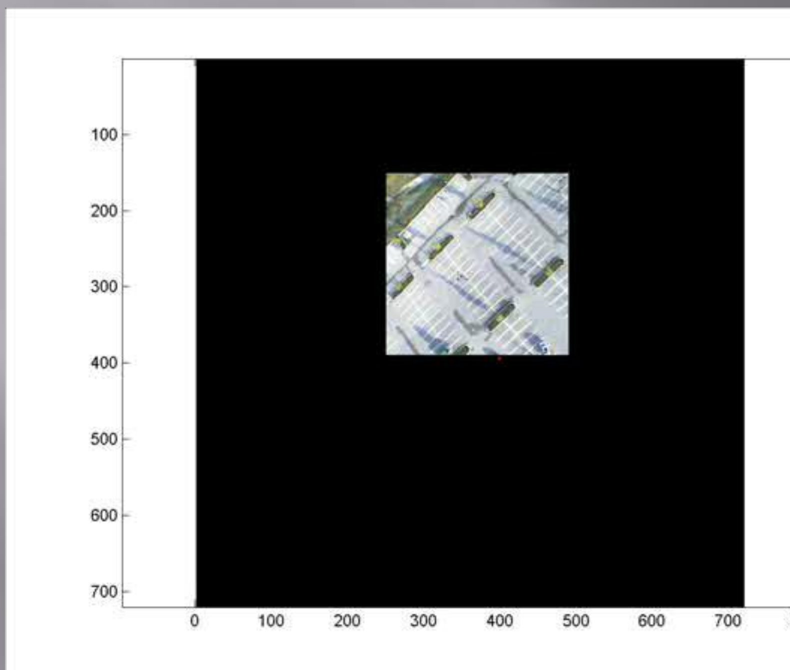


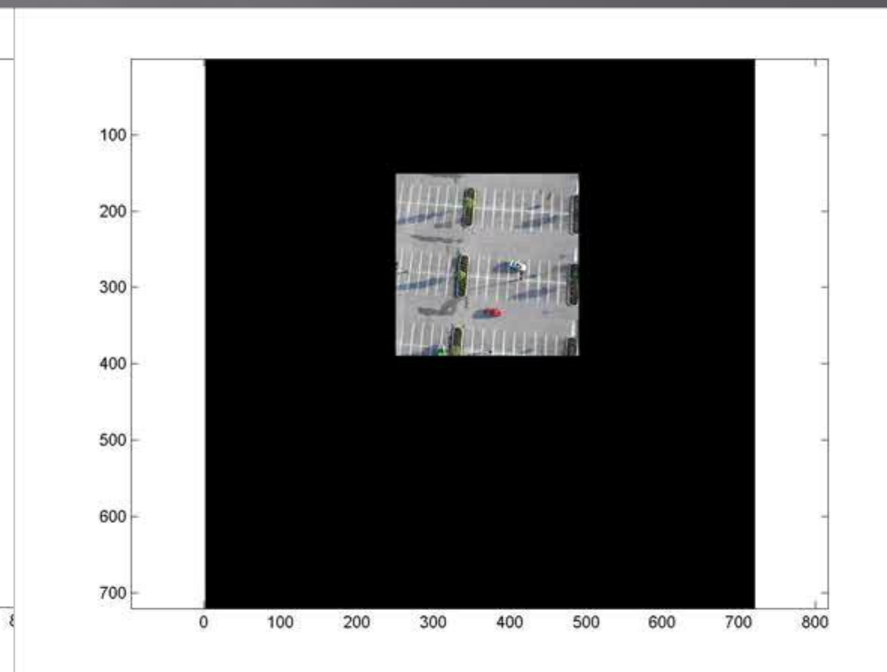
Controlled Sequence 2



University of Central Florida VISION

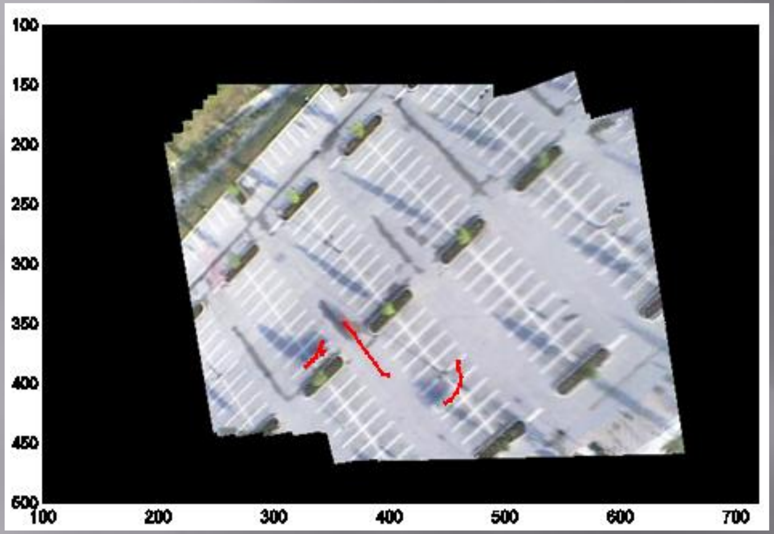
UAV Sequence 1

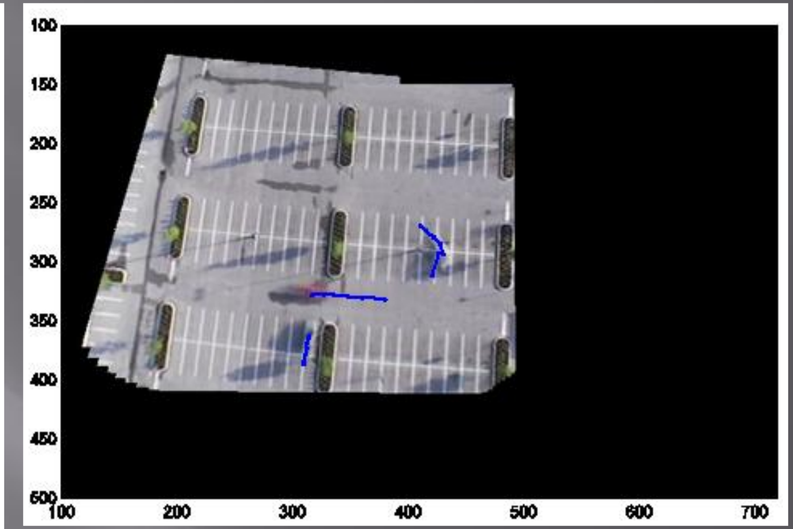


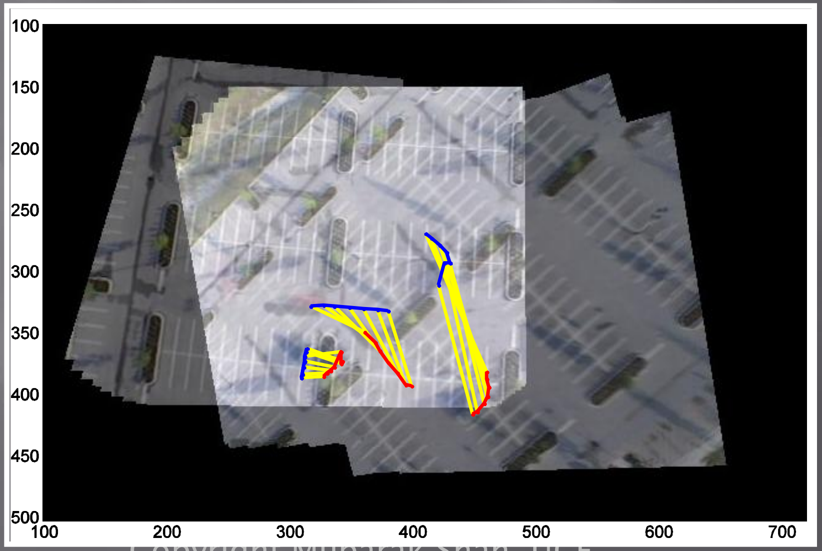


University of Central Florida VISION

Correspondence

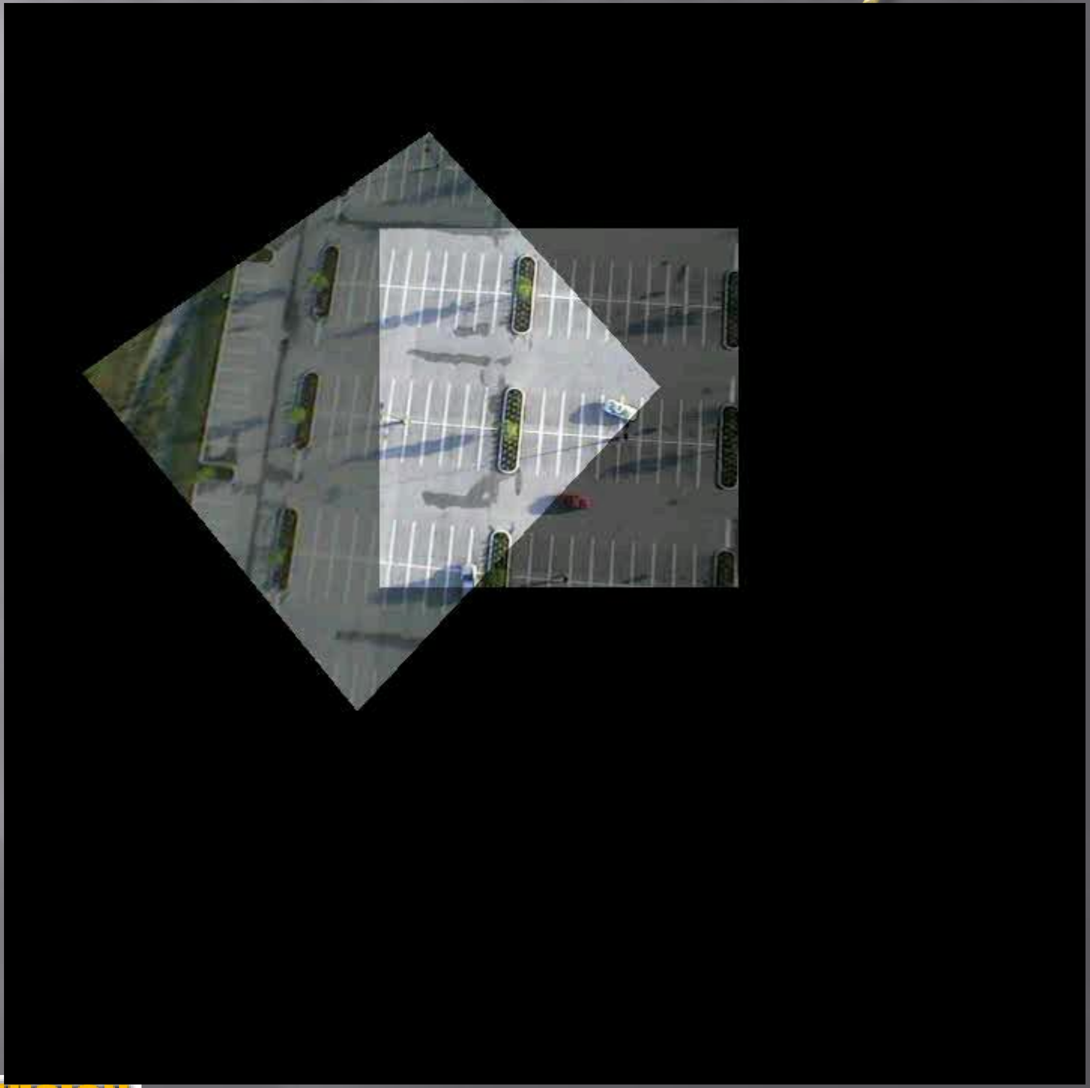






University of Central Florida VISION

Results - Overlayed



Concurrent Visualization

- Mosaics are a compact representation of a single video (planar)
- Concurrent Mosaics can be used for the visualization of multiple videos
- Assuming a Lambertian scene, we can define a color transference function between the mosaics

University of Central Florida VISION

Concurrent Visualization

- We approximate the color transference function by a cubic trivariate polynomial

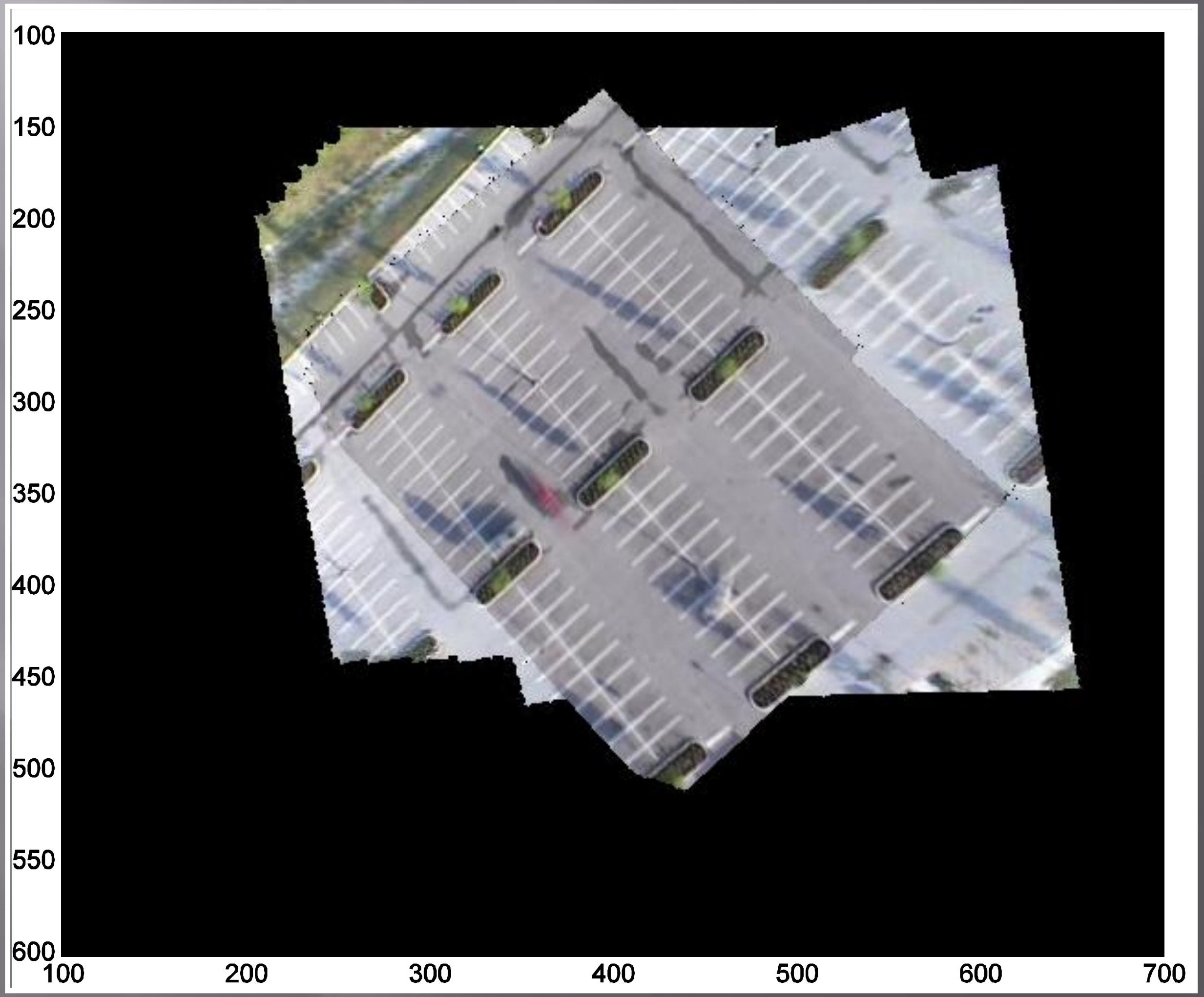
$$r = f_r(r,g,b) = \sum_{i+j+k \leq 3} \alpha^{(r)}_{ijk} r^i g^j b^k + \varepsilon_r$$

$$g = f_g(r,g,b) = \sum_{i+j+k \leq 3} \alpha^{(g)}_{ijk} r^i g^j b^k + \varepsilon_g$$

$$b = f_b(r,g,b) = \sum_{i+j+k \leq 3} \alpha^{(b)}_{ijk} r^i g^j b^k + \varepsilon_b$$

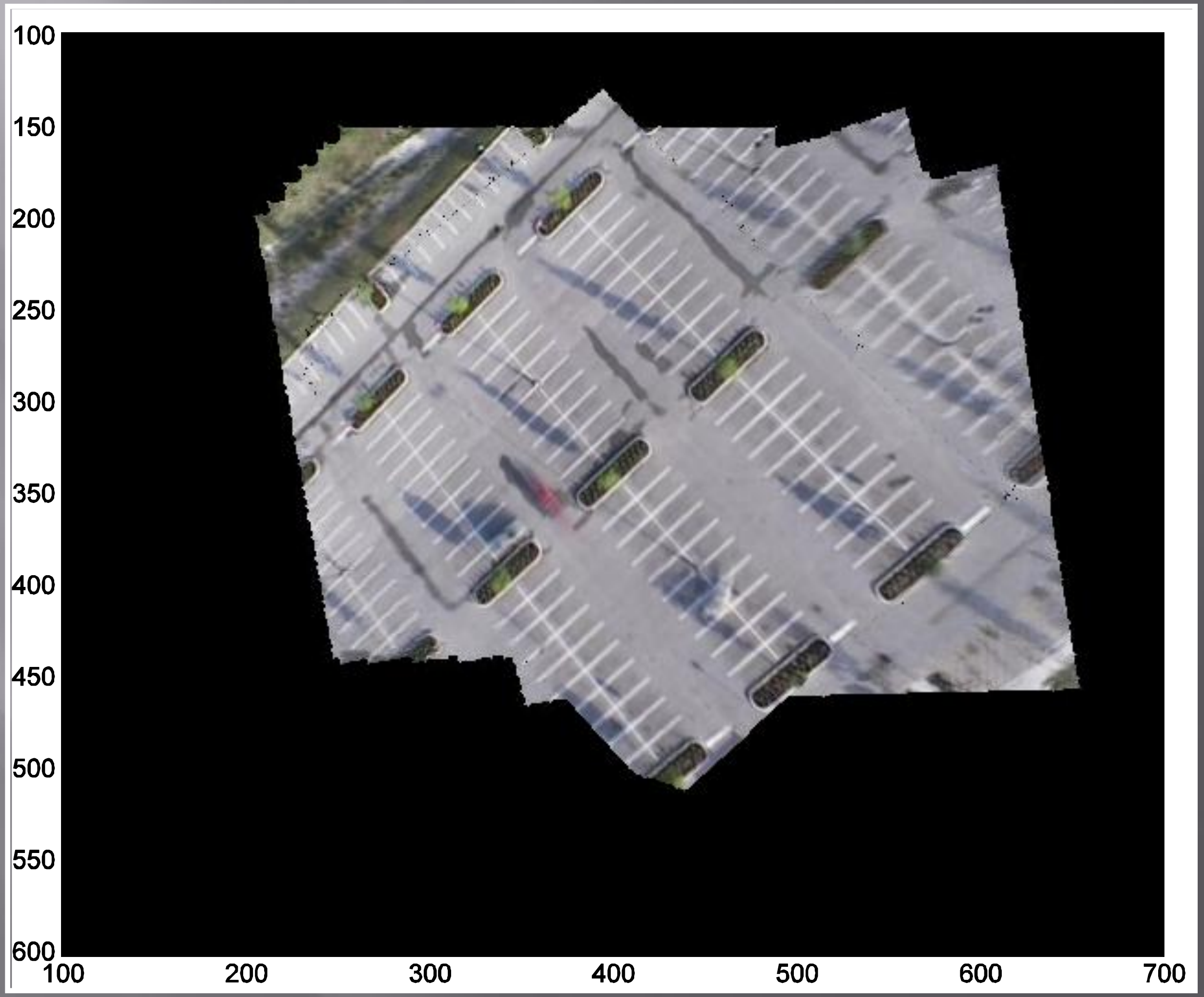
- The transfer functions values are estimated by multiple regression

Concurrent Mosaic



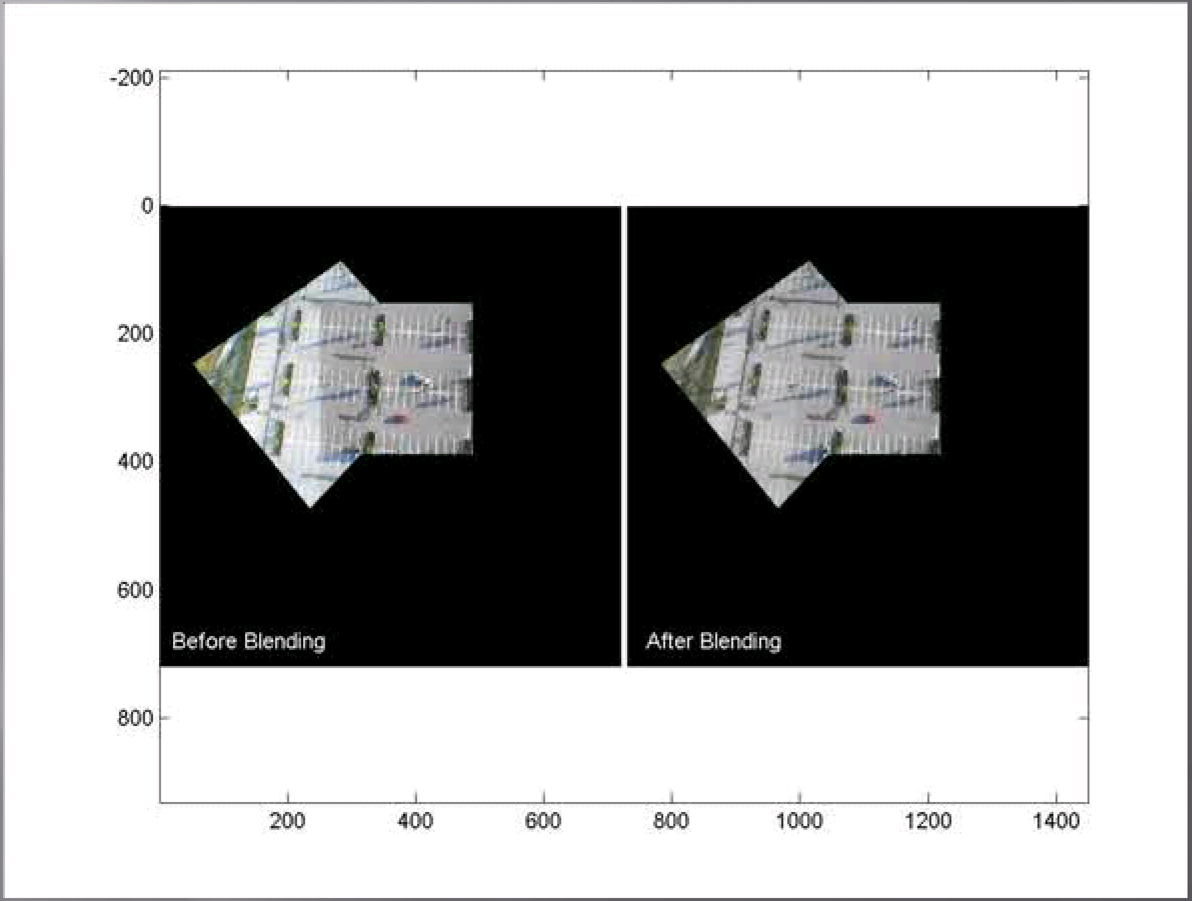
University of VISION
Central Florida

Blended Concurrent Mosaic



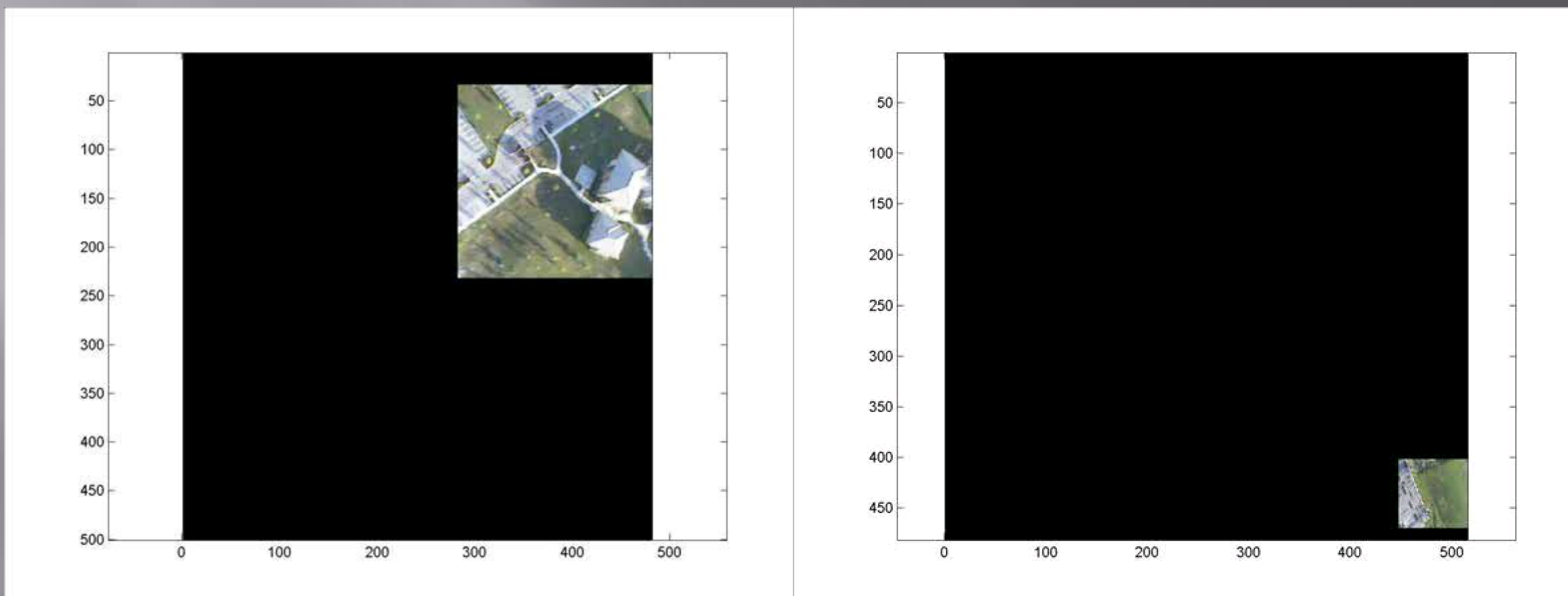
University of Central Florida VISION

Blending

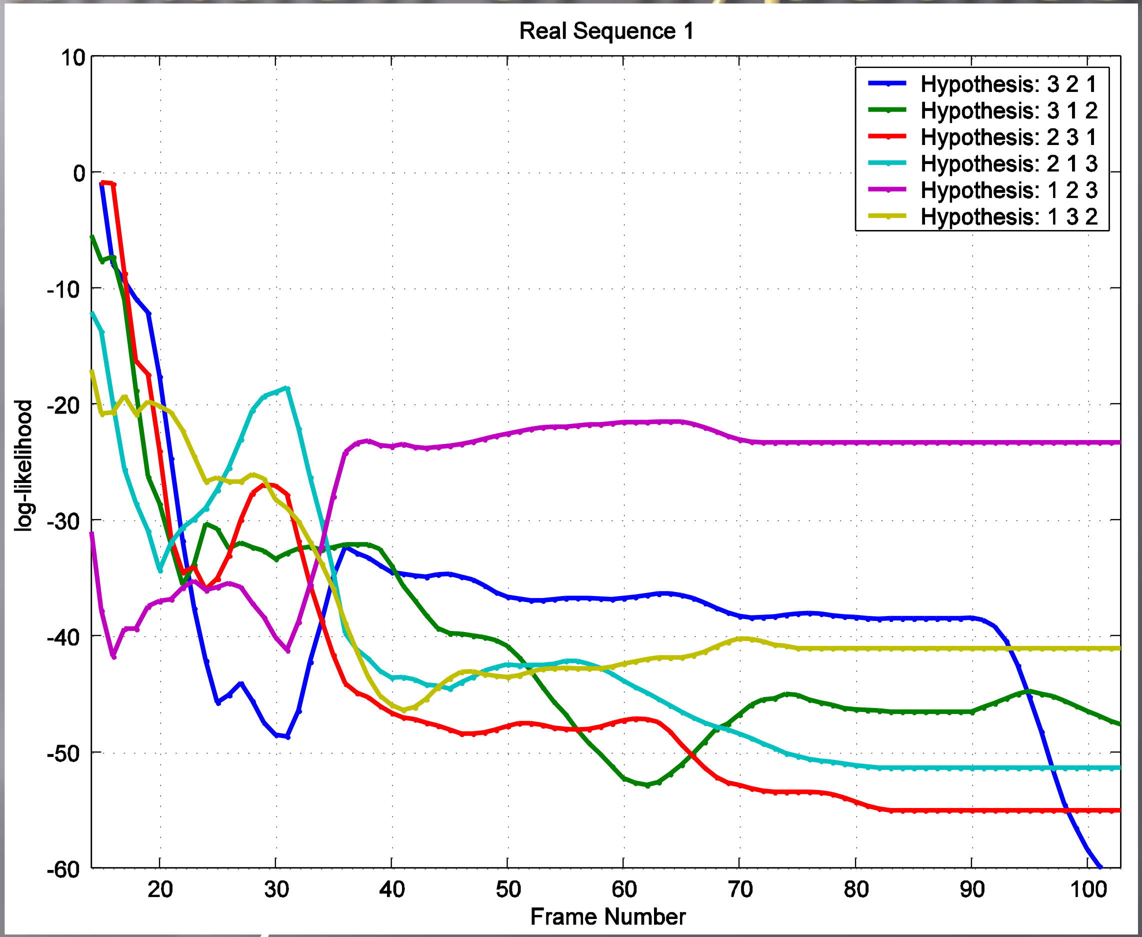


-200
0
200
400
600
800
200
400
600
800
1000
1200
1400
Before Blending
After Blending

UAV Sequence 2



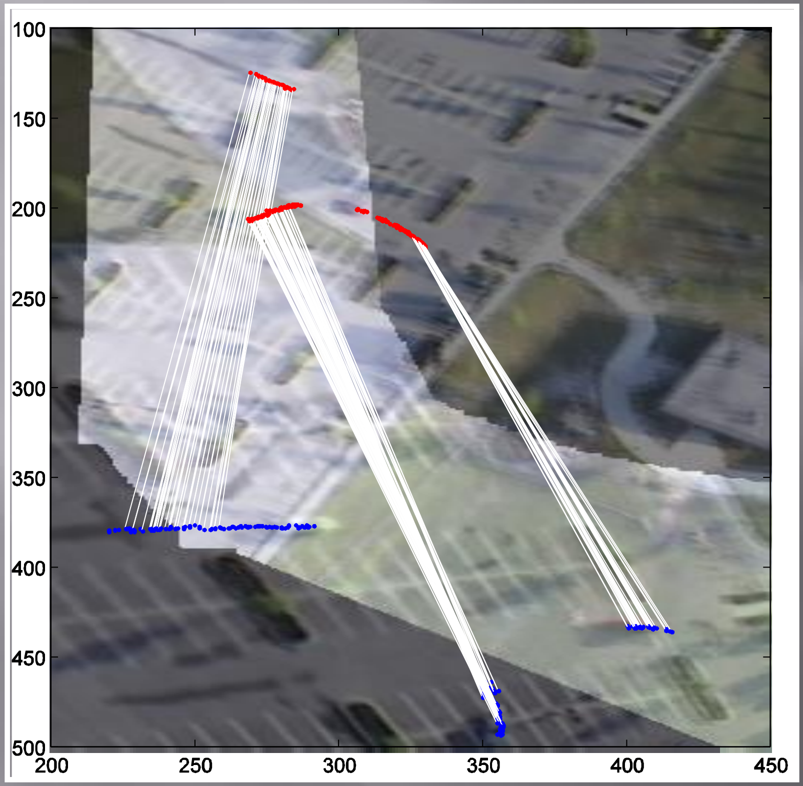
Variation of hypotheses

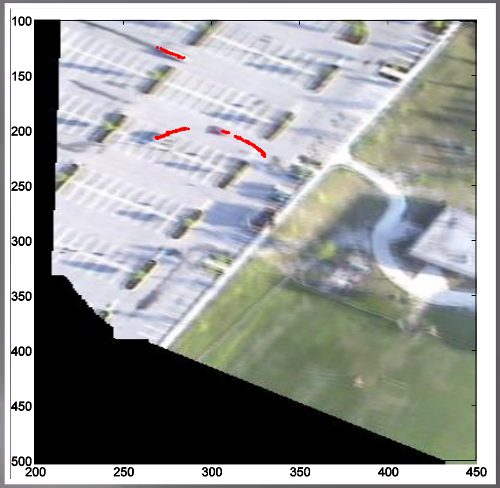


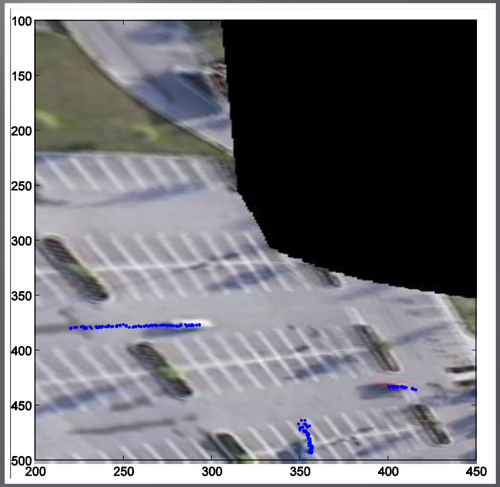
Collinear Motion

University of Central Florida VISION

Correspondence

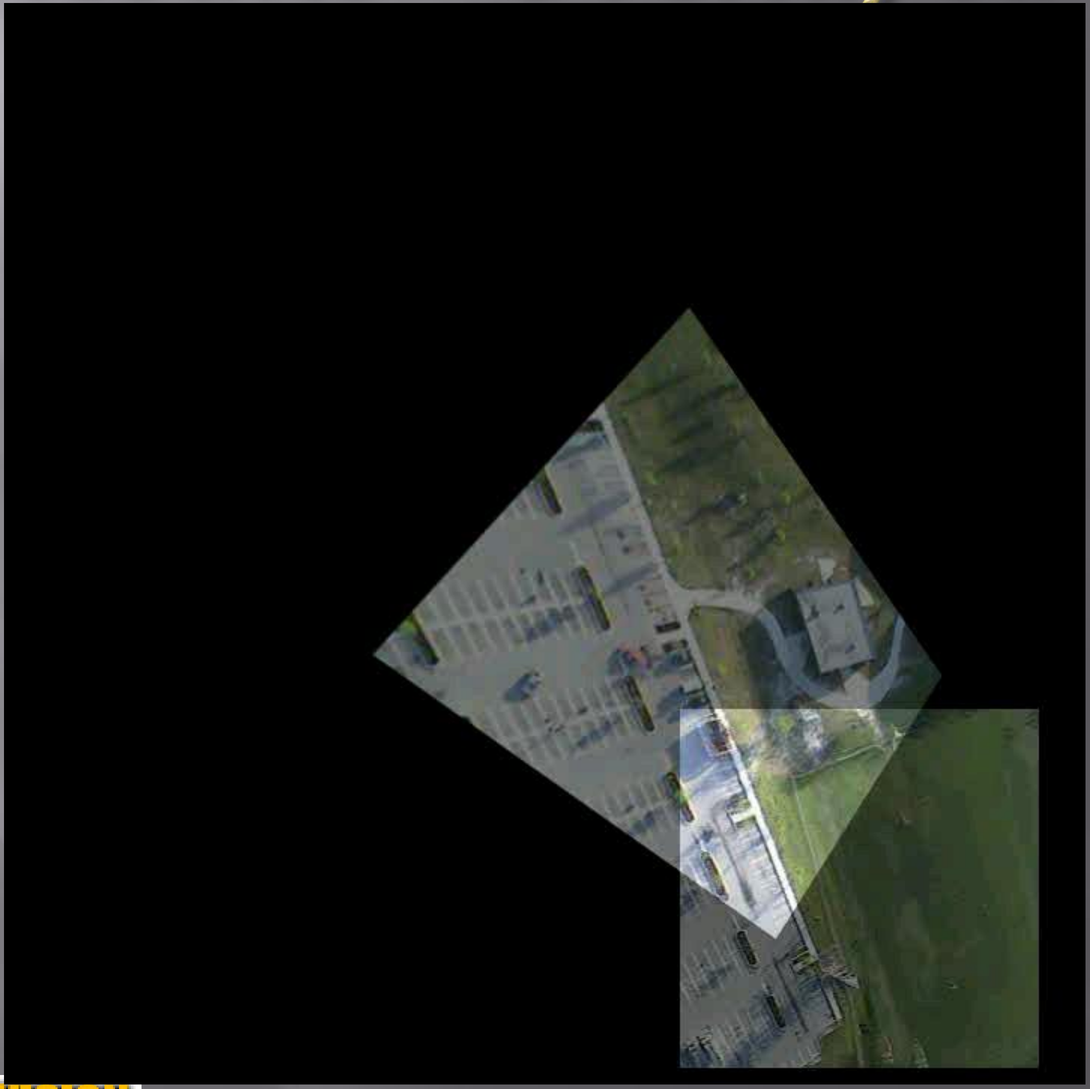




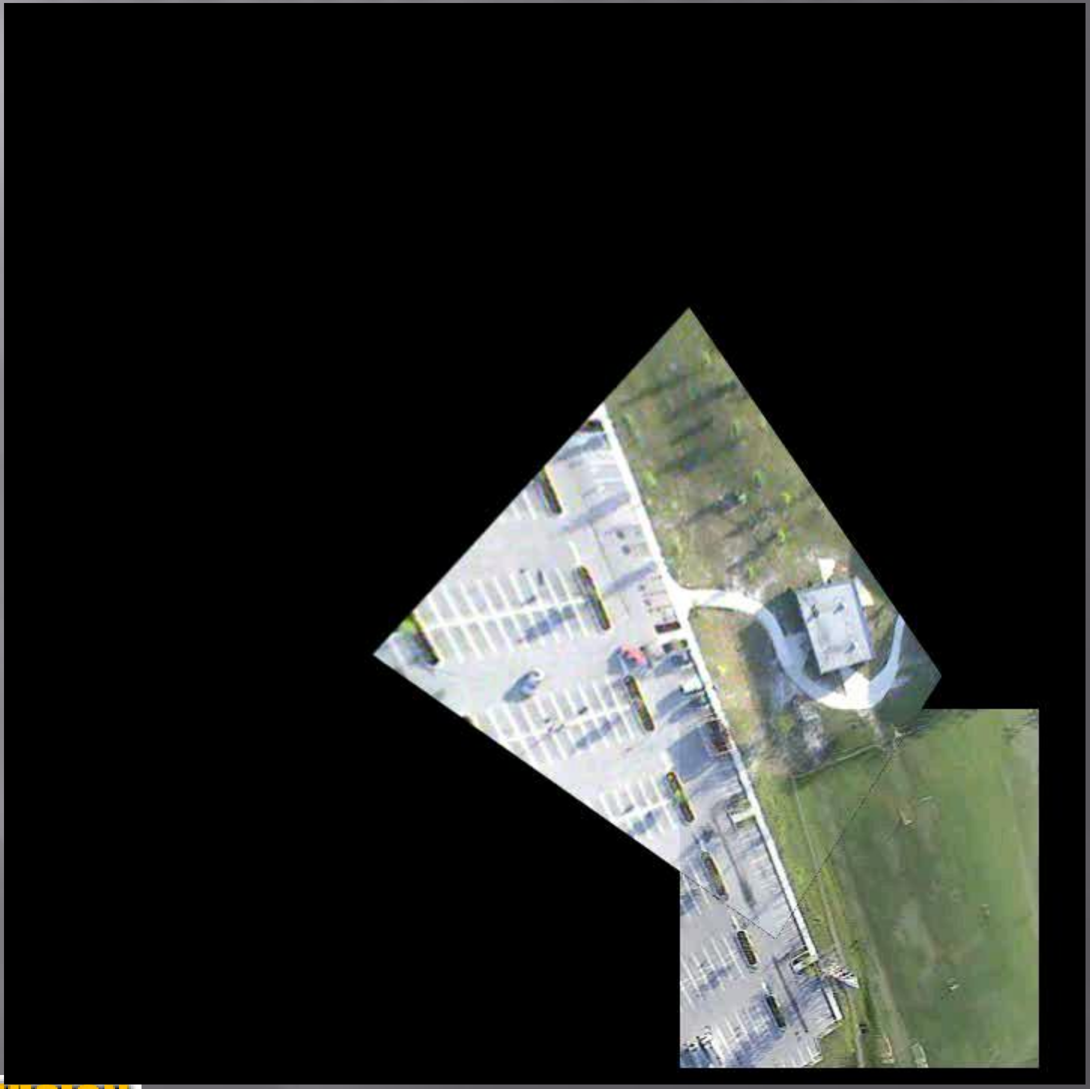


University of Central Florida VISION

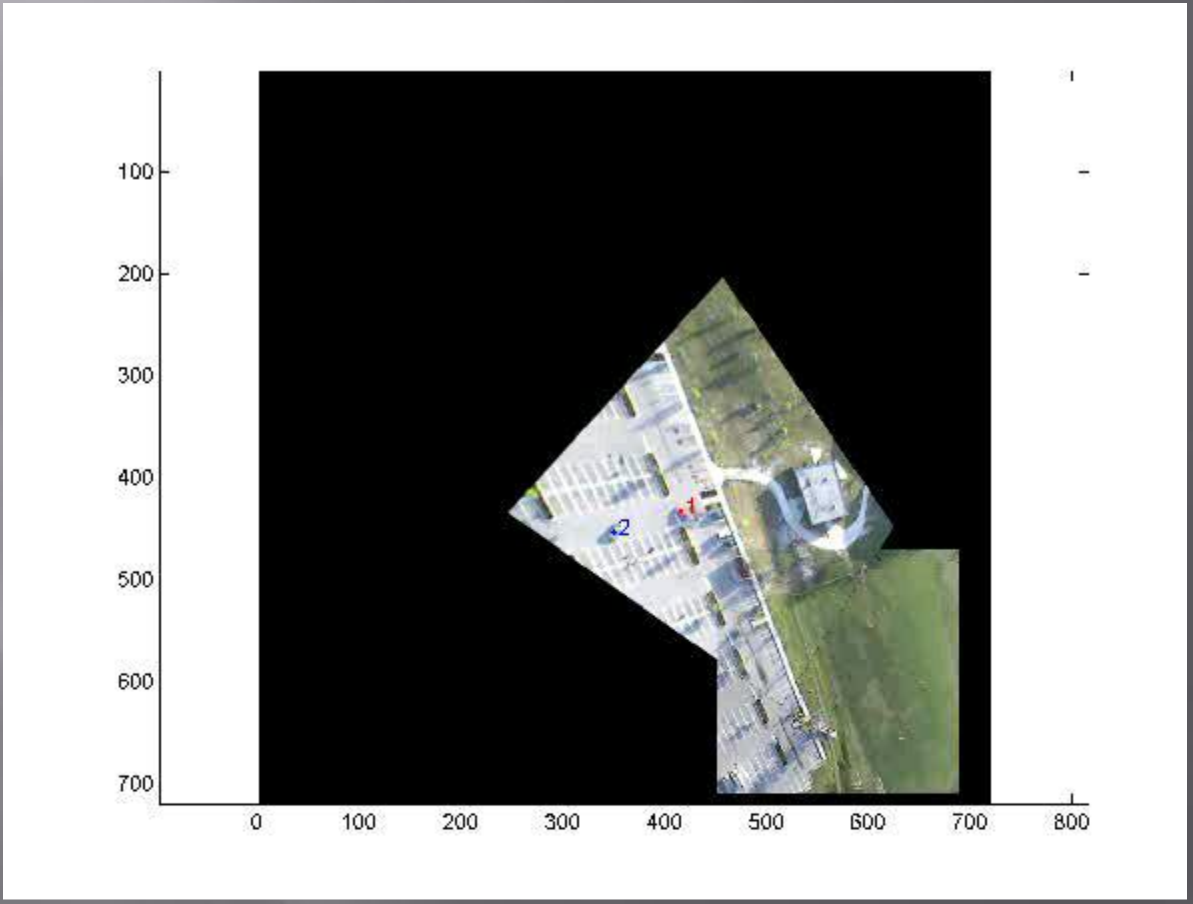
Results - Overlayed



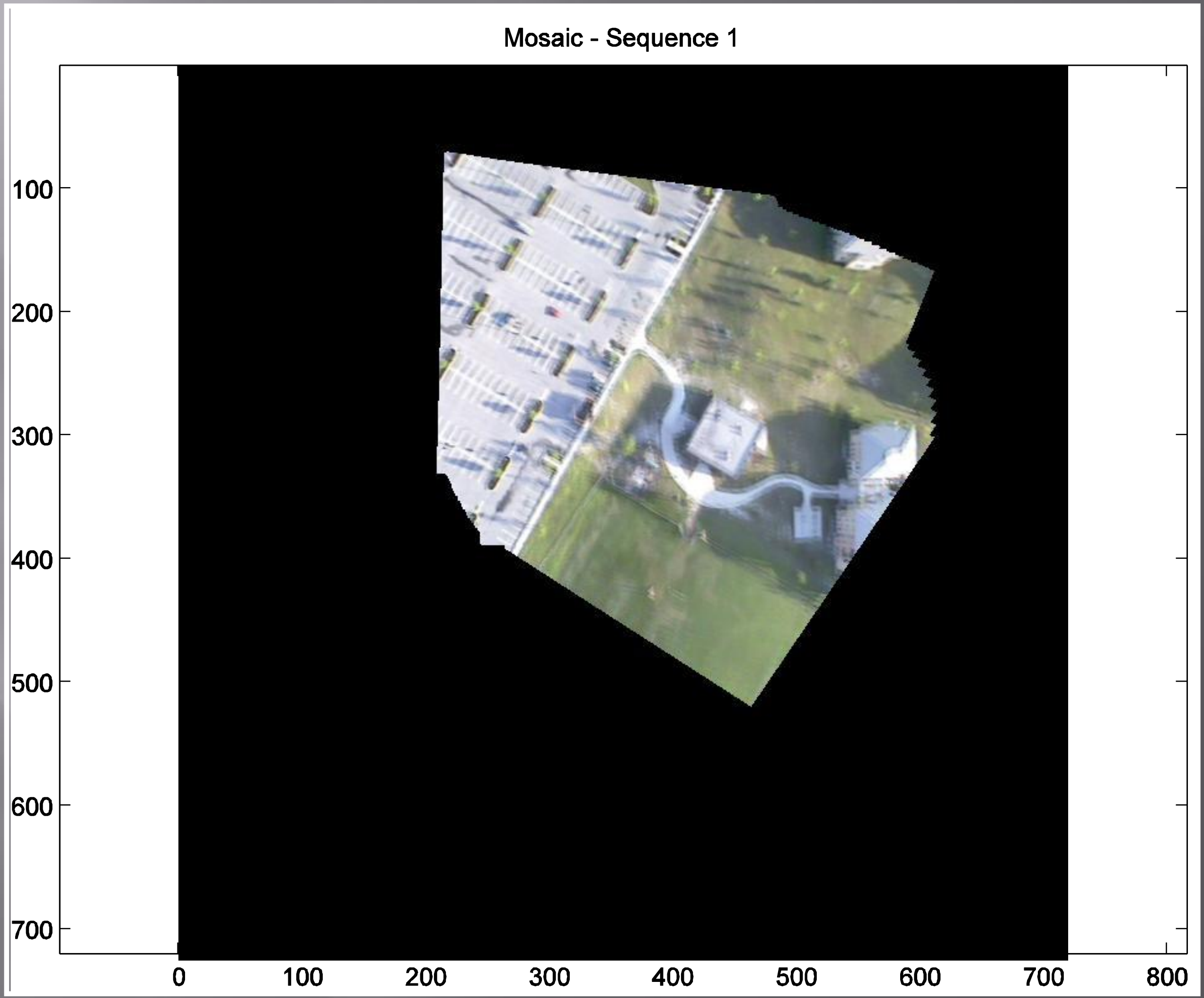
Results - Blended



Corresponded Tracks

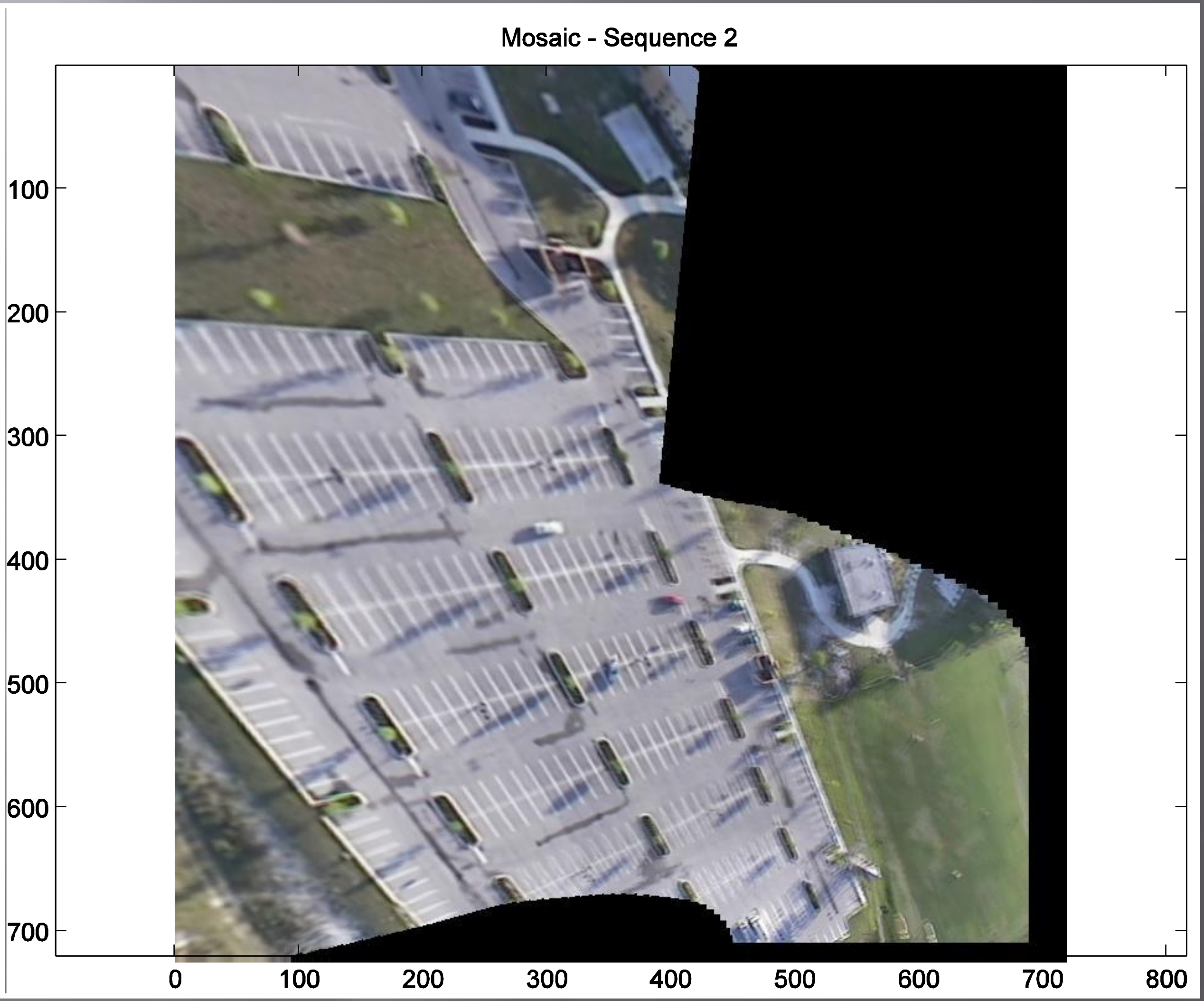


Mosaic 1



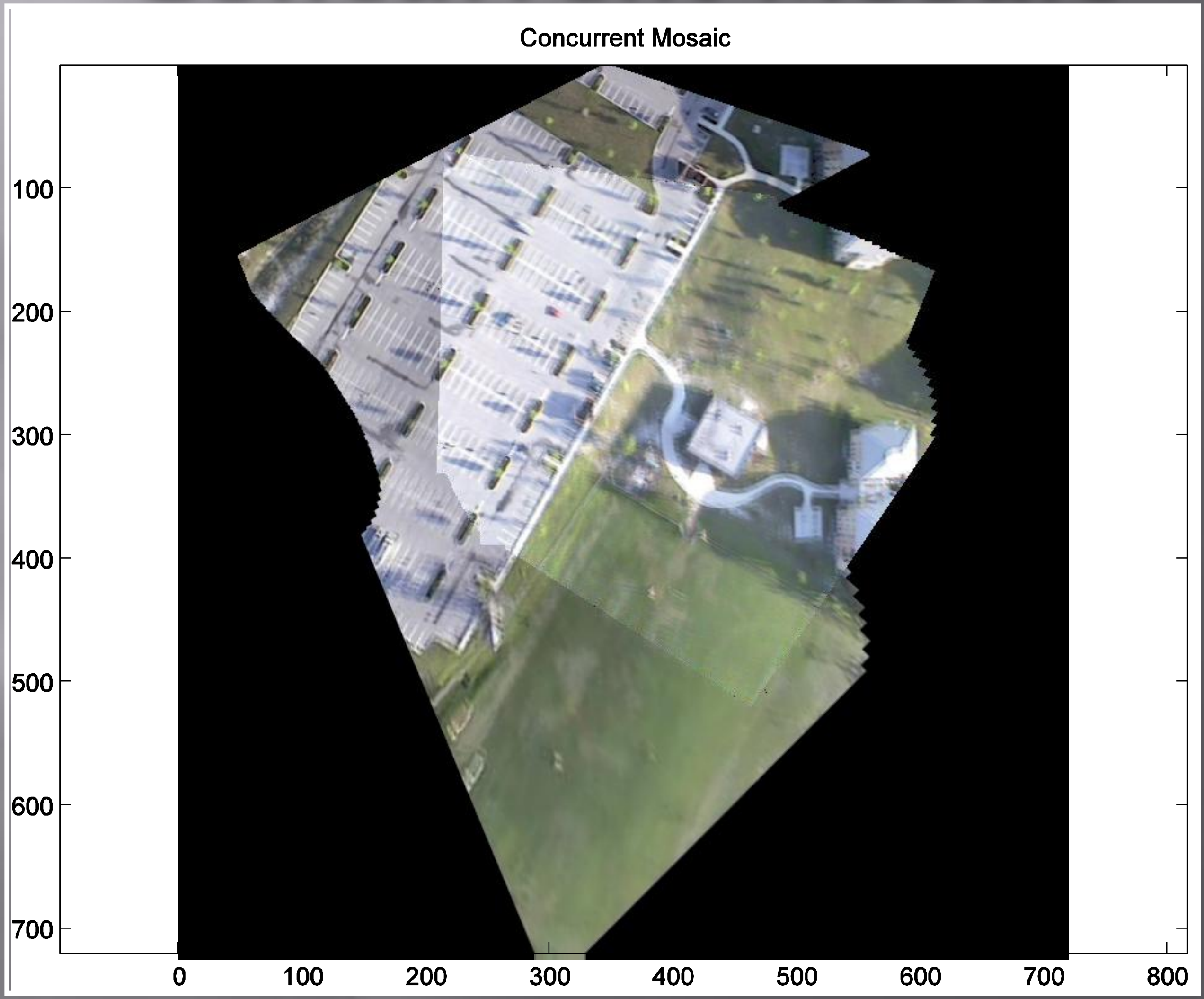
Mosaic - Sequence 1
100
200
300
400
500
600
700
0
100
200
300
400
500
600
700
800

Mosaic 2



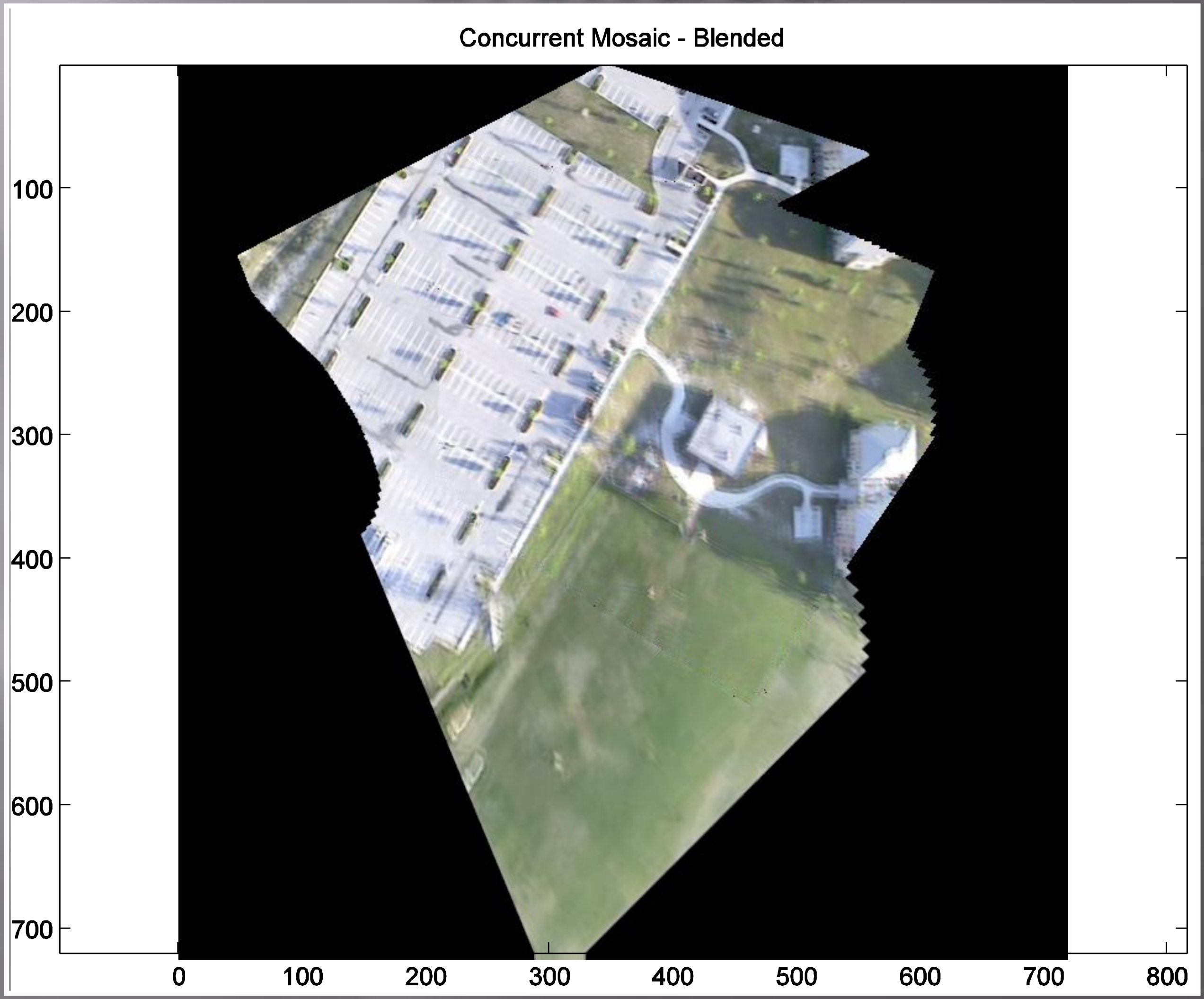
University of Central Florida VISION

Concurrent Mosaic



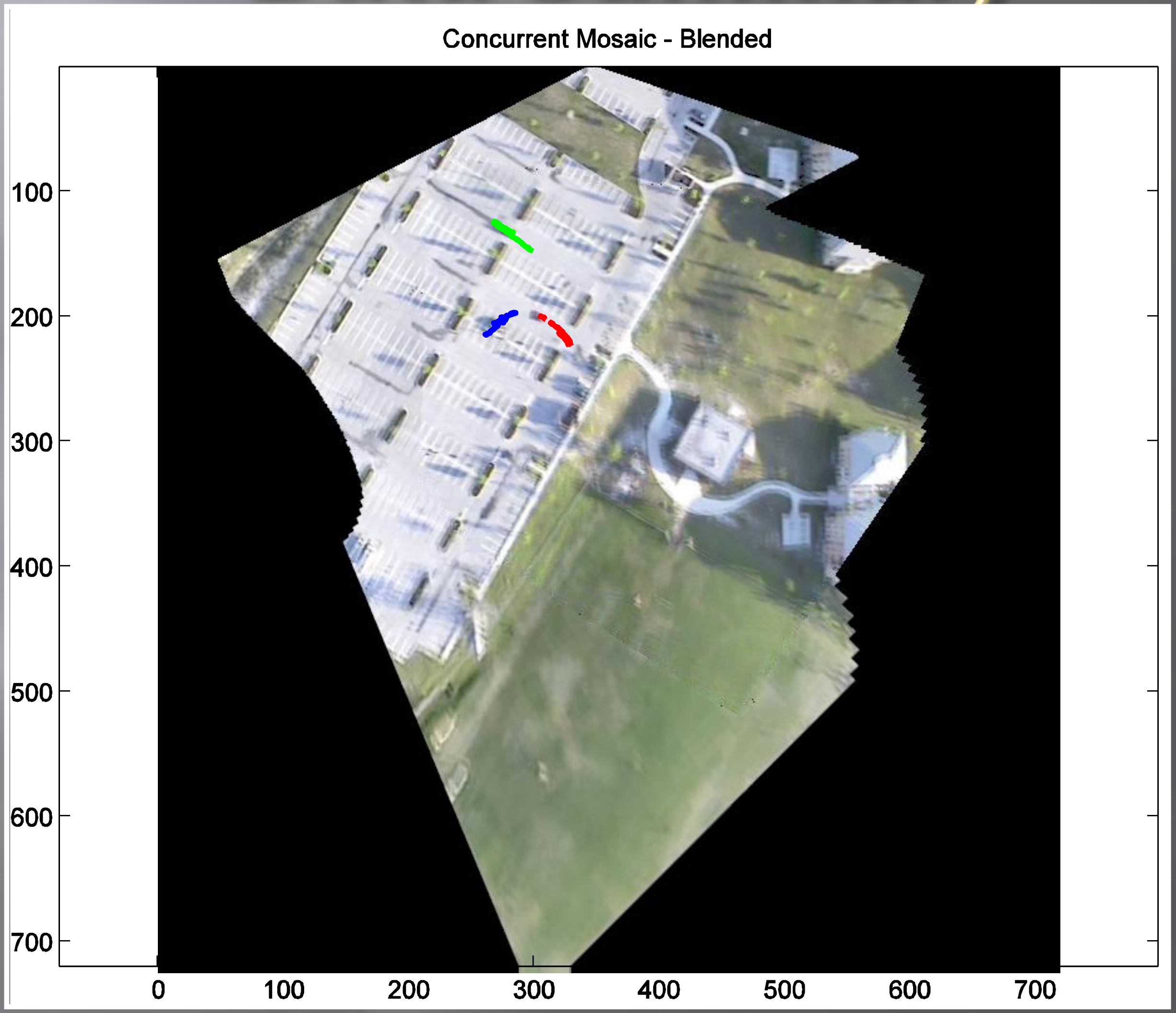
University of Central Florida VISION

Blended



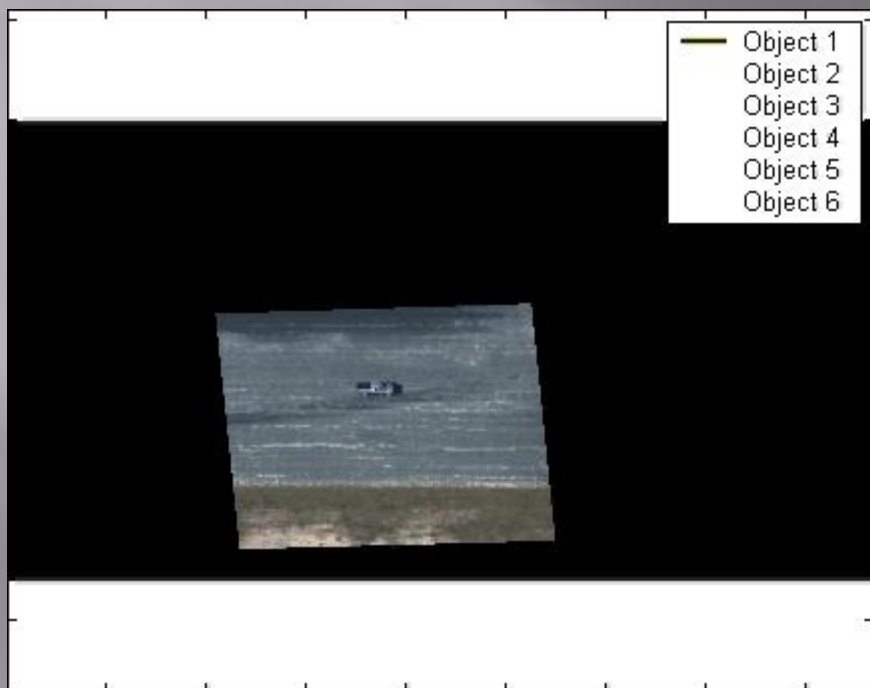
University of Central Florida VISION

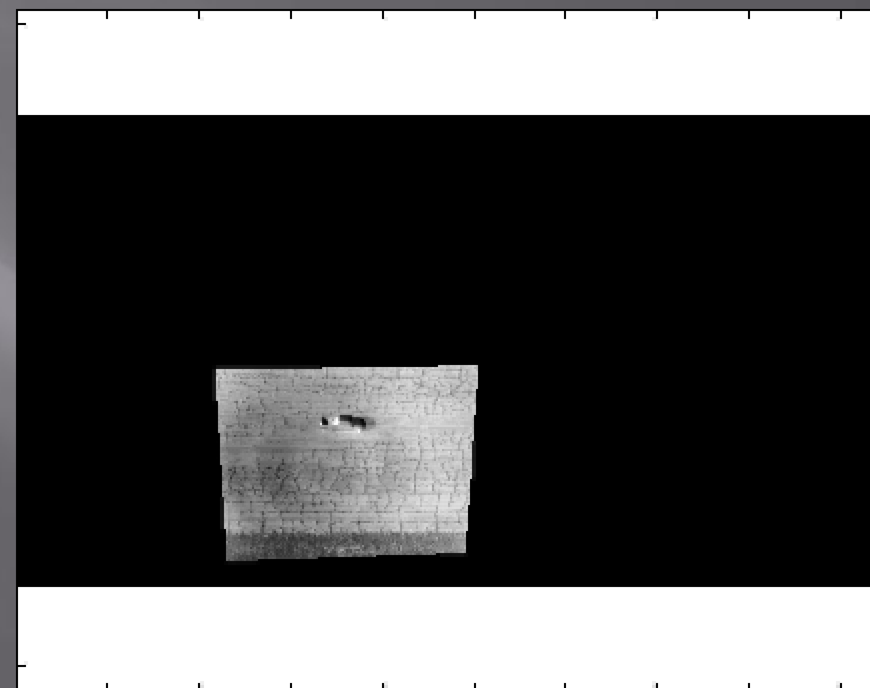
Data Summary



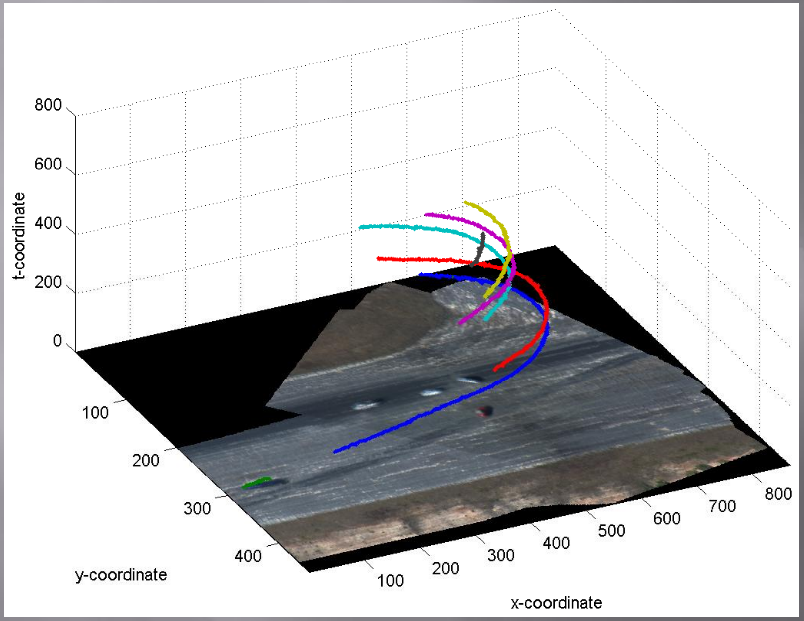
University of VISION Central Florida

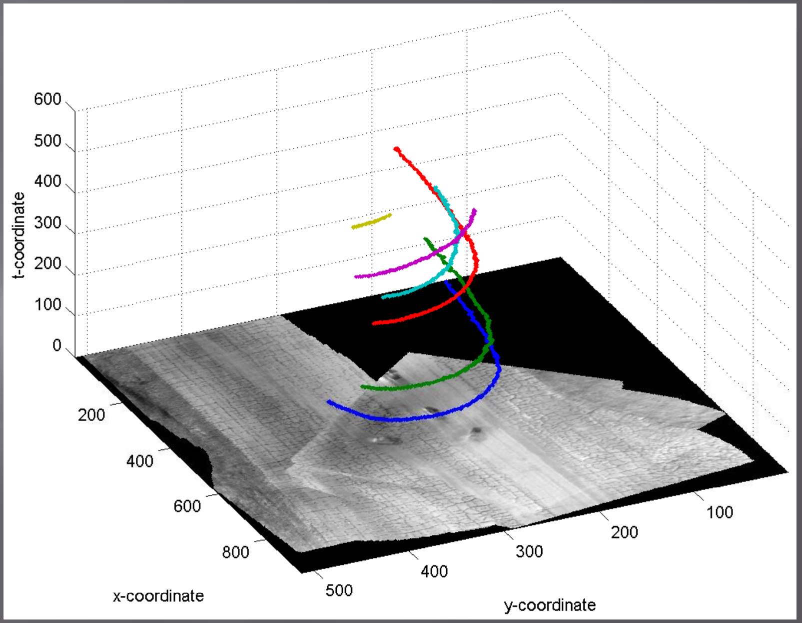
UAV Sequence 3





UAV Sequence 3





Trajectory Association Across Non-overlapping Cameras in Planar Scene

Yaser Sheikh, Xin Li and Mubarak Shah

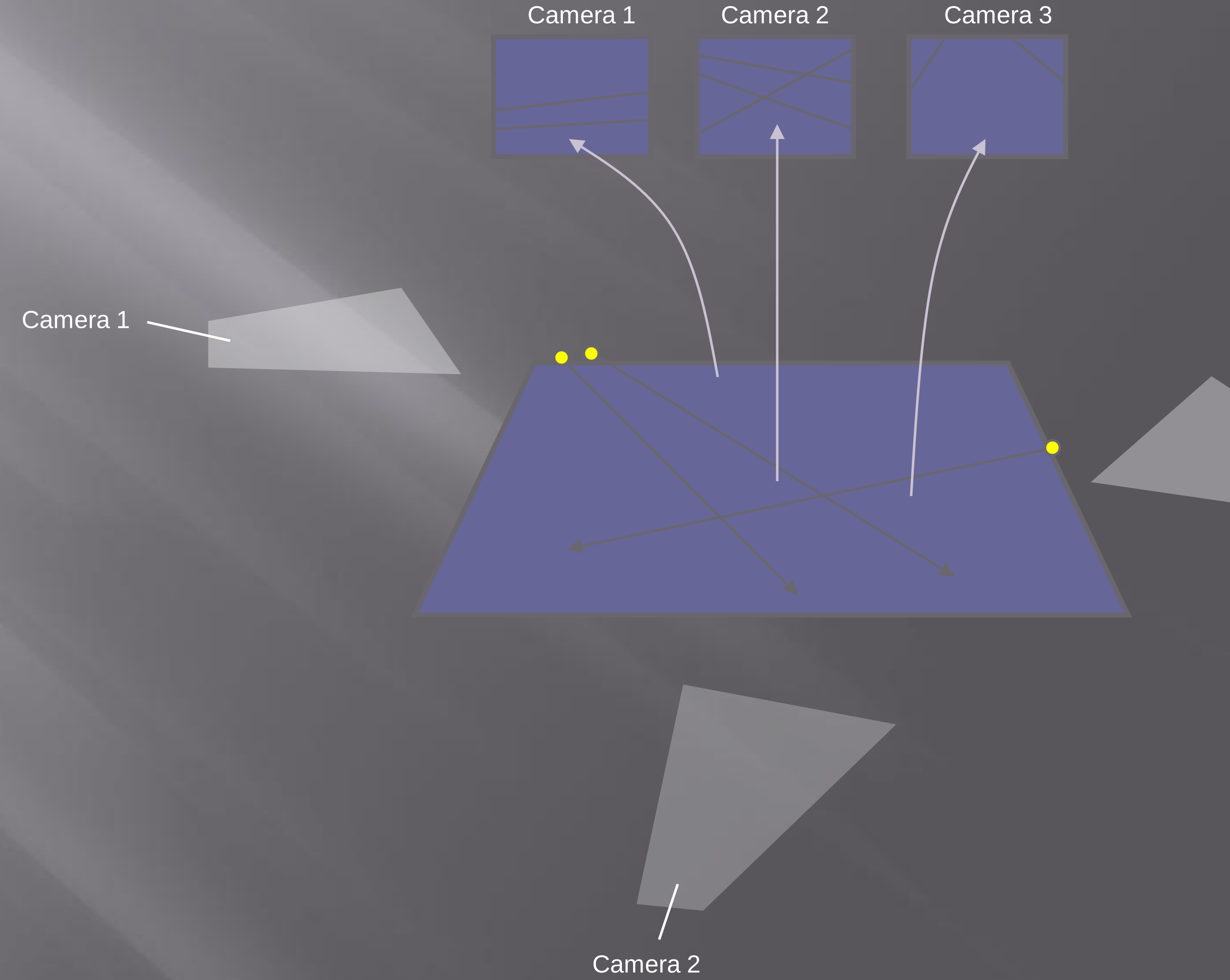
CVPR 2007

University of Central Florida VISION

Data Model

- K: # of objects
- N: # of cameras
- **H**: World Plane to image plane homography
- c^i_j: Association of object *j* in camera *i*

- *Goals*:
  - Association
  - Homographies
  - True Trajectories

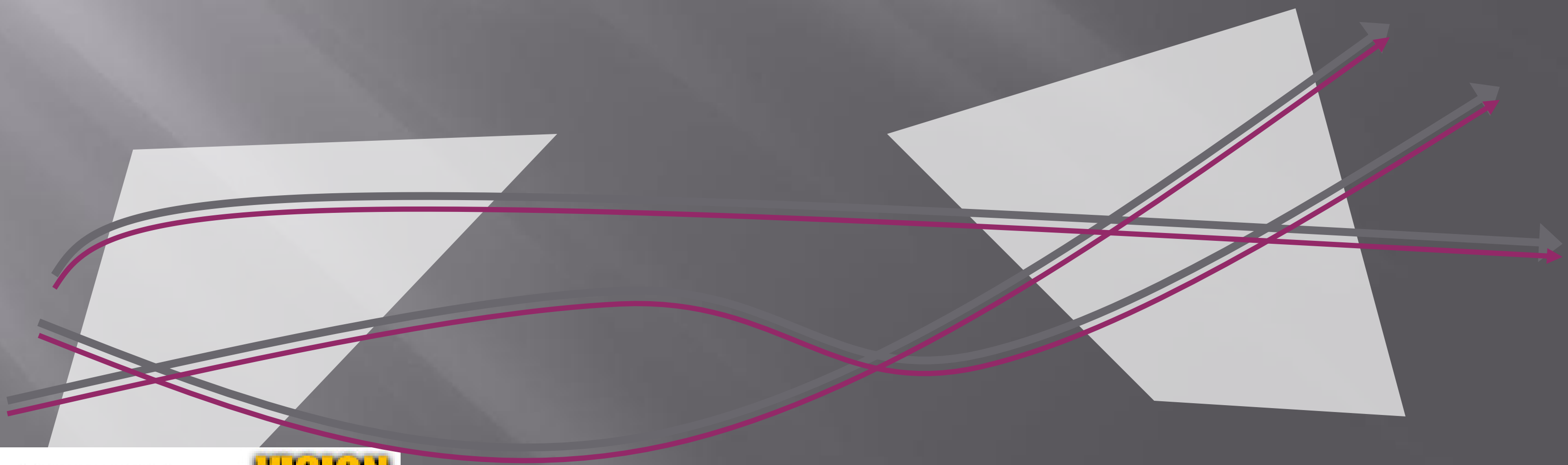


University of Central Florida VISION

Global Association

General Case: Zero spatiotemporal overlap

- The kinematics of the object are modeled
- Allows us to constrain spatial relation of non-overlapping FoVs



University of Central Florida VISION

Kinematic Polynomial Models

- Model: The trajectories of each object define a polynomial in *t*.
- In general,

$$x(t) = \sum_{i=0}^{d} p_i t^i, \quad y(t) = \sum_{i=0}^{d} q_i t^i$$

- Linear

$$x(t) = p_1 t + p_0, \quad y(t) = q_1 t + q_0$$

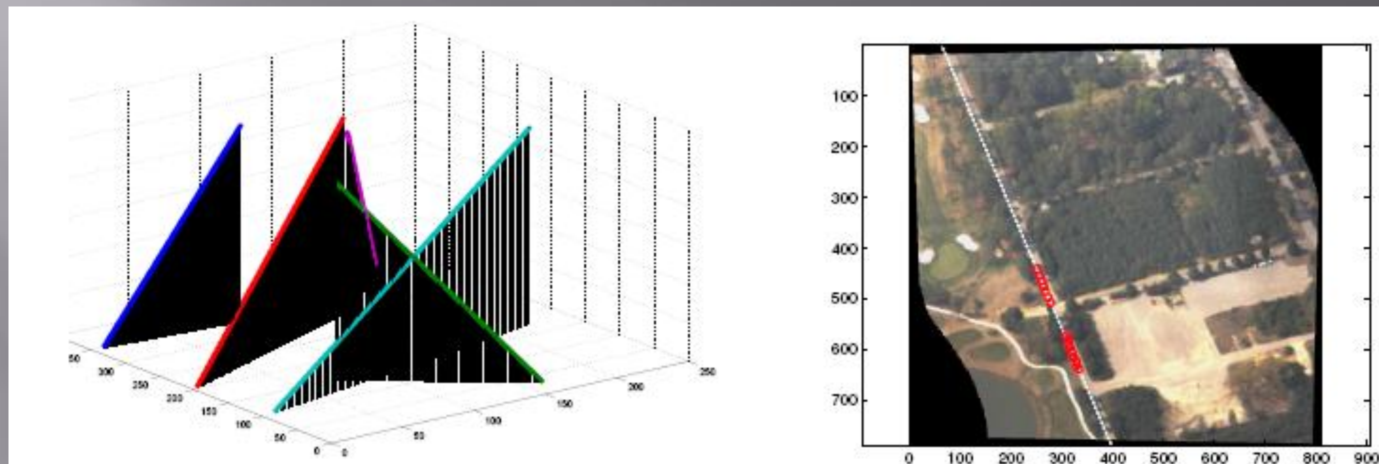
- Quadratic

$$x(t) = p_2 t^2 + p_1 t + p_0, \quad y(t) = q_2 t^2 + q_1 t + q_0$$

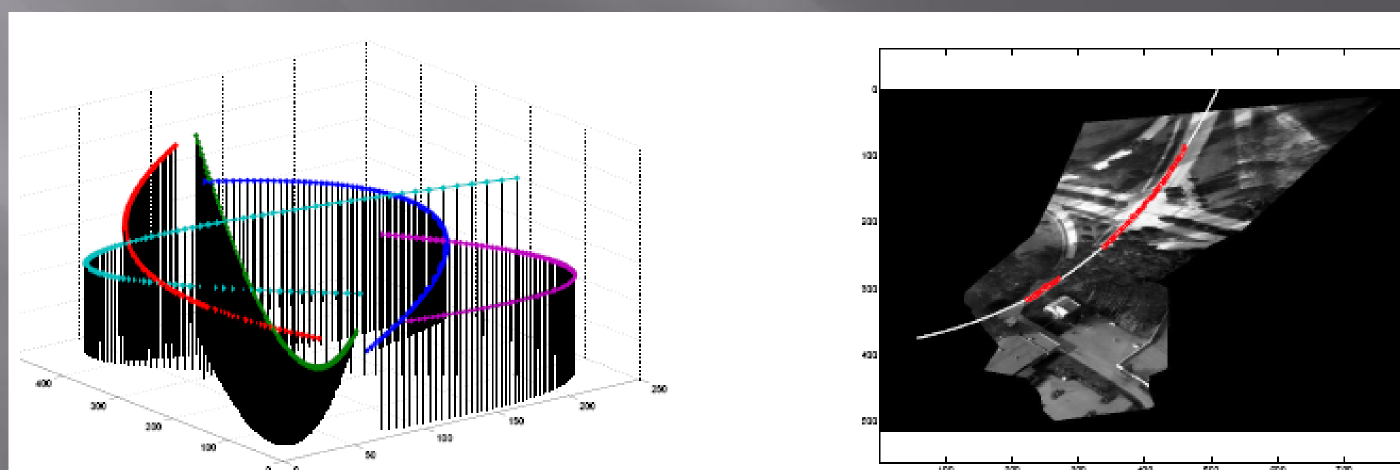
University of Central Florida VISION

Kinematic Polynomial Models

Linear



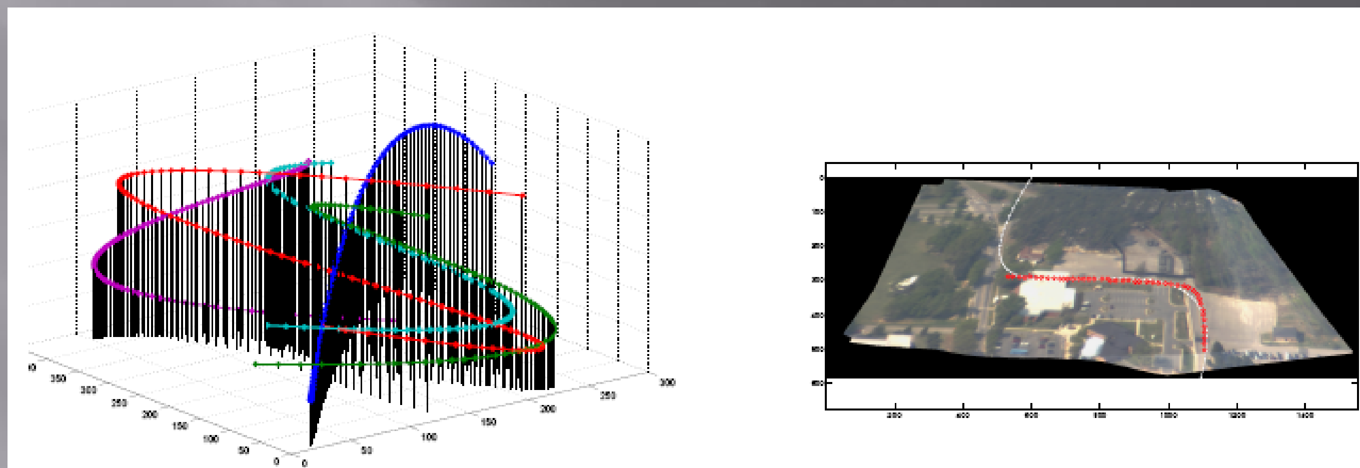
Quadratic



University of Central Florida VISION

Kinematic Polynomial Models

Cubic



Number of Unknowns

- Linear (constant velocity)
  - 4*k+9*N
- Quadratic (constant acceleration)
  - 6*k+9*N
- Cubic
  - 8*k+9*N

Homographies

- To get the imaged point at time *t* we have,

$$\begin{bmatrix} \lambda x_j^i(t) \\ \lambda y_j^i(t) \\ \lambda \end{bmatrix} = \mathbf{H}^i \begin{bmatrix} x_j(t) \\ y_j(t) \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} \lambda x_j^i(t) \\ \lambda y_j^i(t) \\ \lambda \end{bmatrix} = \mathbf{H}^i \begin{bmatrix} p_{1j}t + p_{0j} \\ q_{1j}t + q_{0j} \\ 1 \end{bmatrix} = \mathbf{H}^i \mathbf{P}_j \begin{bmatrix} t \\ 1 \end{bmatrix}$$

University of Central Florida VISION

Maximum Likelihood Parameter Estimation

- **Question:**
  Given the data, under this model, what is the optimal estimate of association, parameters and homographies?
- **Formally:**

  Find the Maximum Likelihood estimates of $\boldsymbol{\Theta} = (\{\mathbf{P}_k\}_K, \{\mathbf{H}^n\}^N)$ and $\mathbf{C} = \{\mathbf{c}\}^N{}_K$ for data $\underline{\mathbf{X}} = \{\underline{\mathbf{x}}\}^N{}_K$.

Problem 1: Define the likelihood function

Problem 2: Provide a maximization algorithm

Results

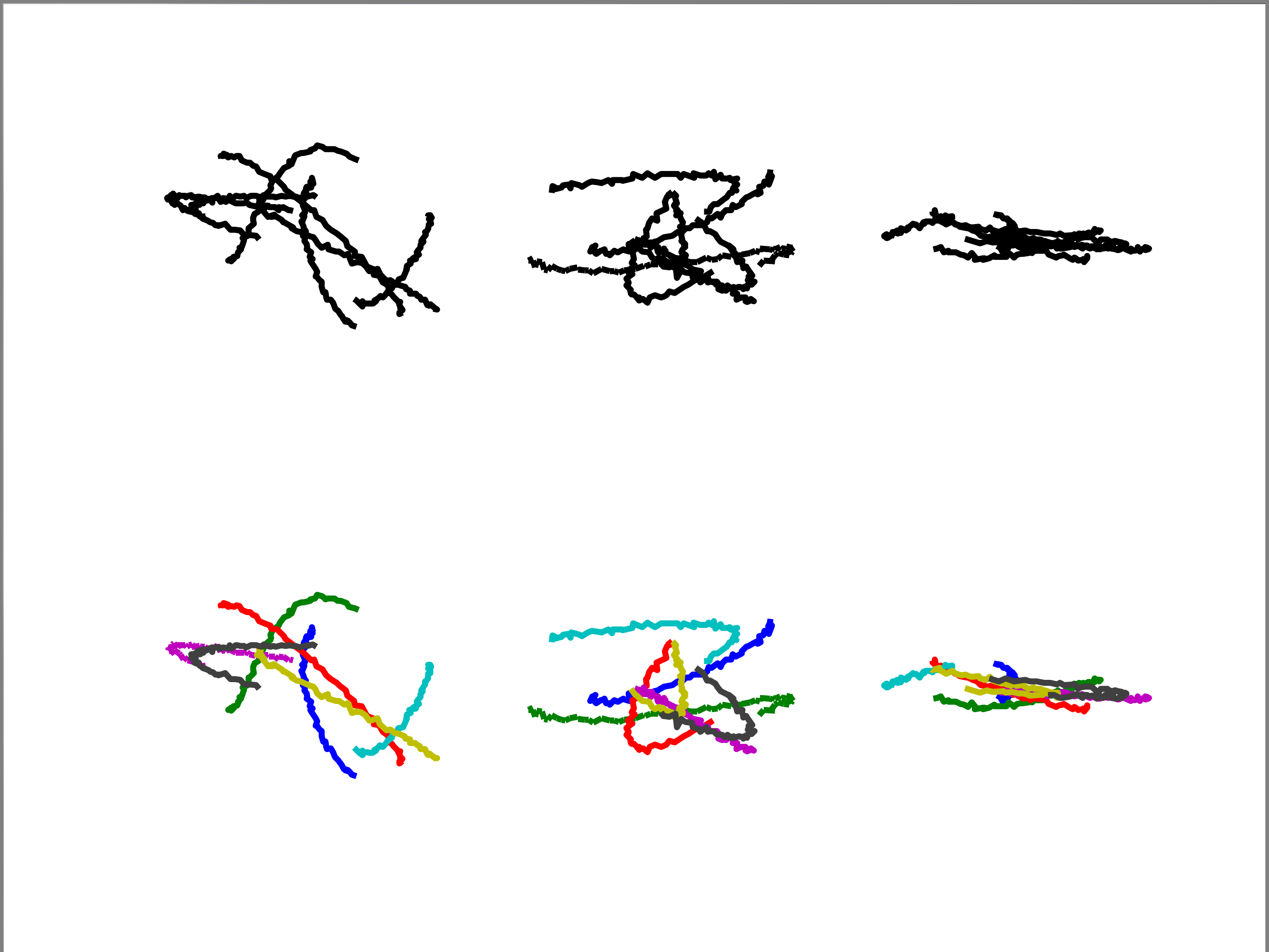
- Simulations
  - Linear, Quadratic and Cubic
- Reacquisition of objects in a single camera
- Association across cameras

University of Central Florida VISION

Simulations

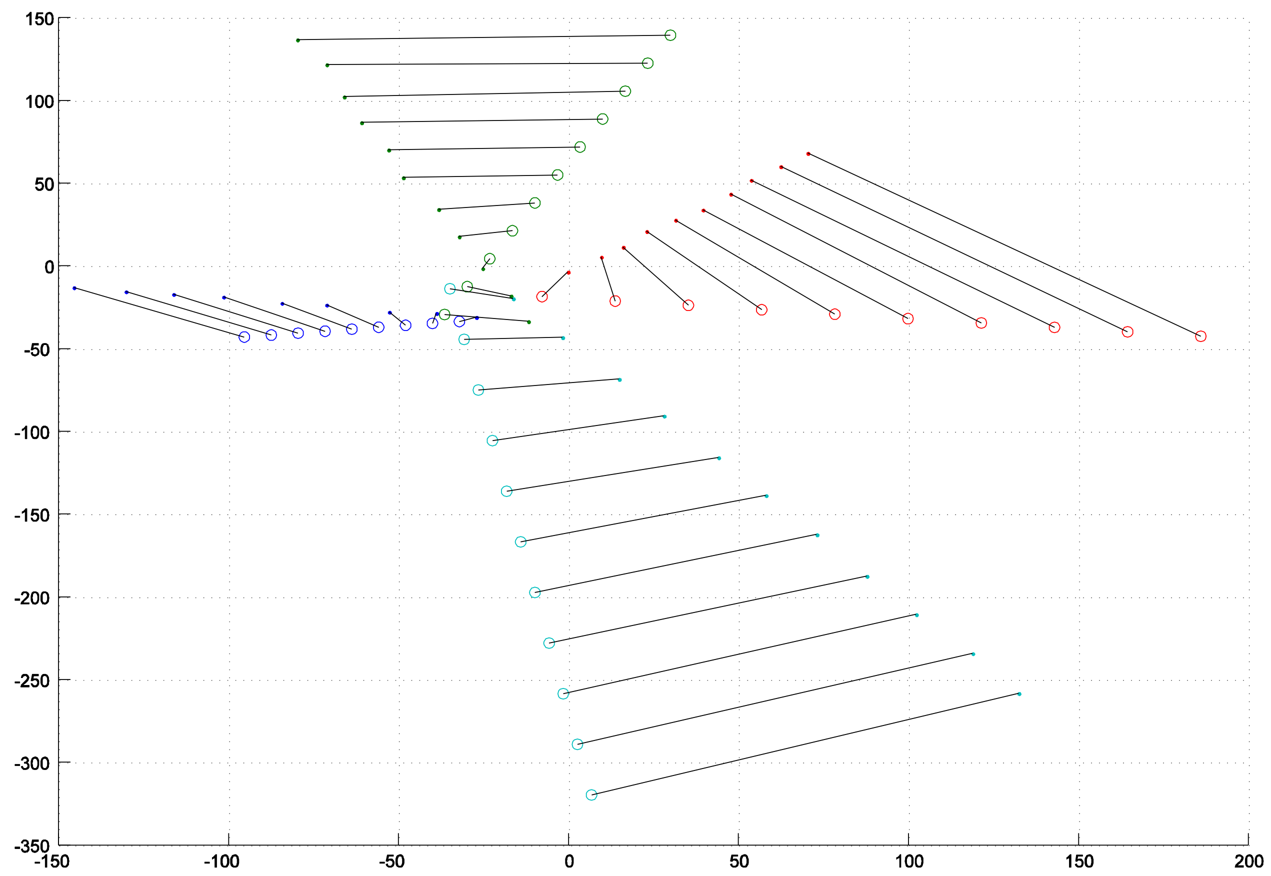
- Measurement noise
- Number of Objects
- Number of Cameras
- Total Time duration
- Observation Start and Duration (per camera)
- Parameter noise

Simulations



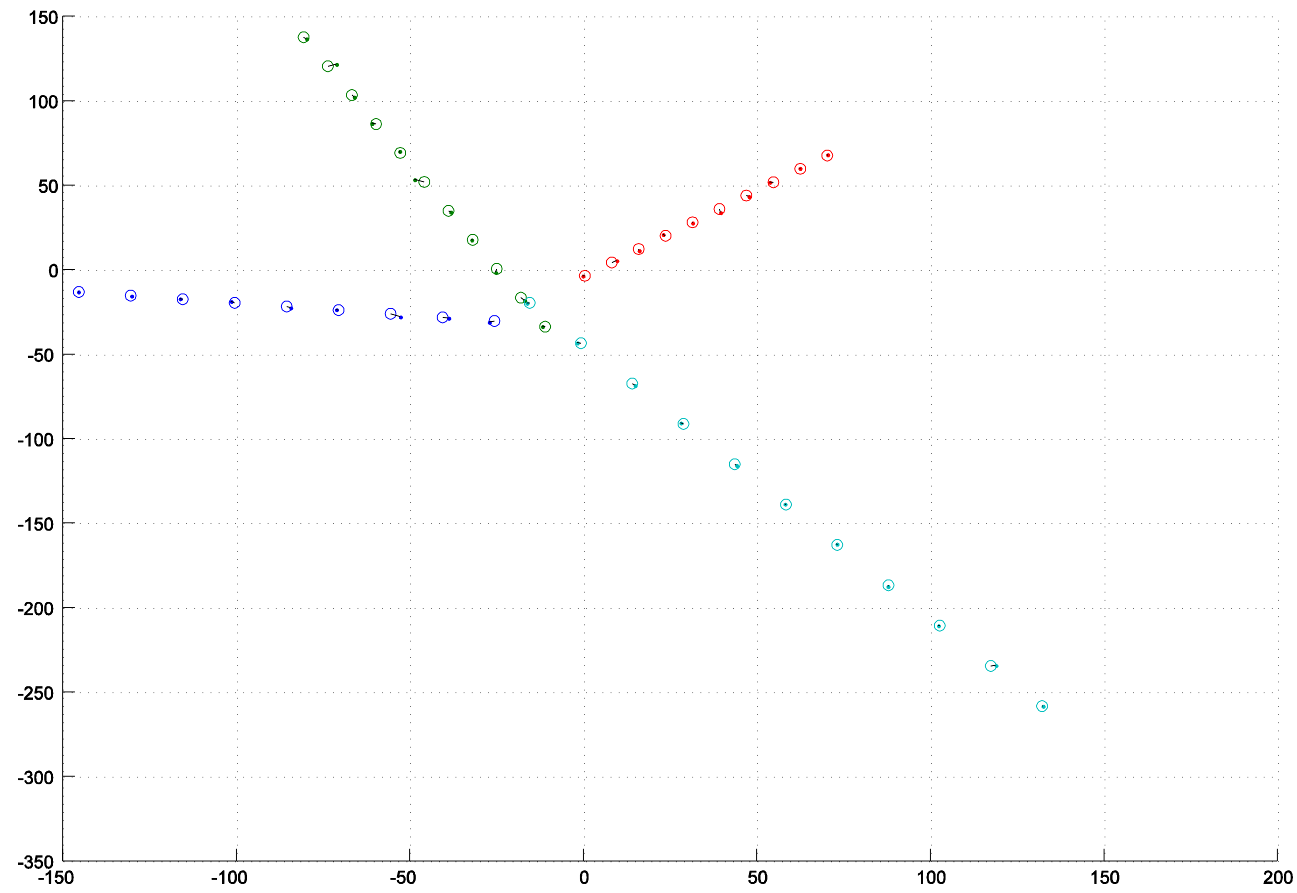
University of Central Florida VISION

Simulations E.g.



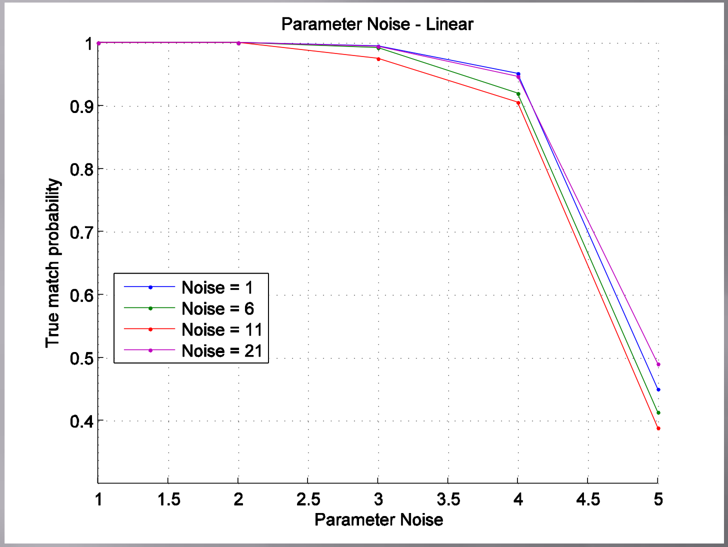
University of Central Florida VISION

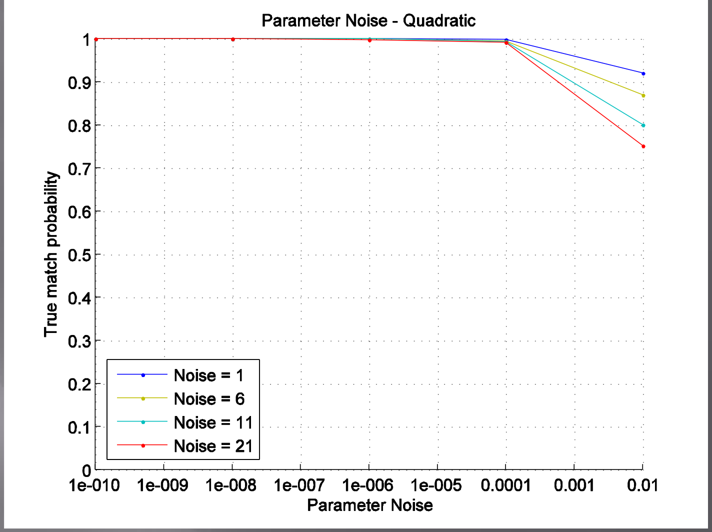
Simulations E.g.

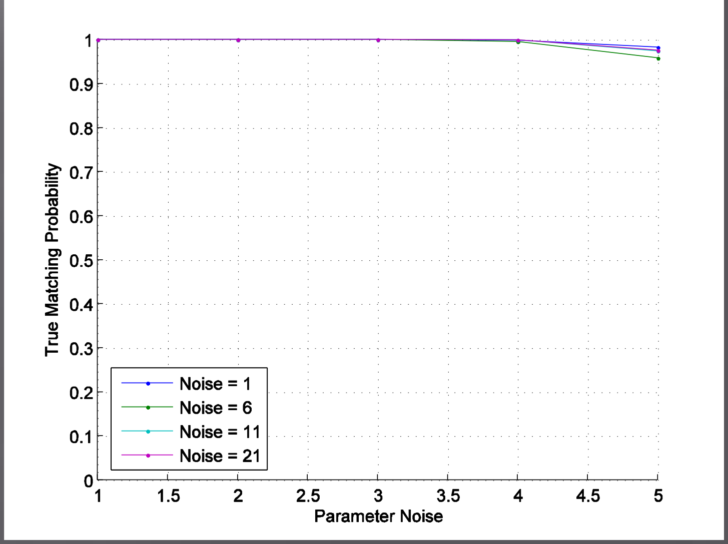


University of Central Florida VISION

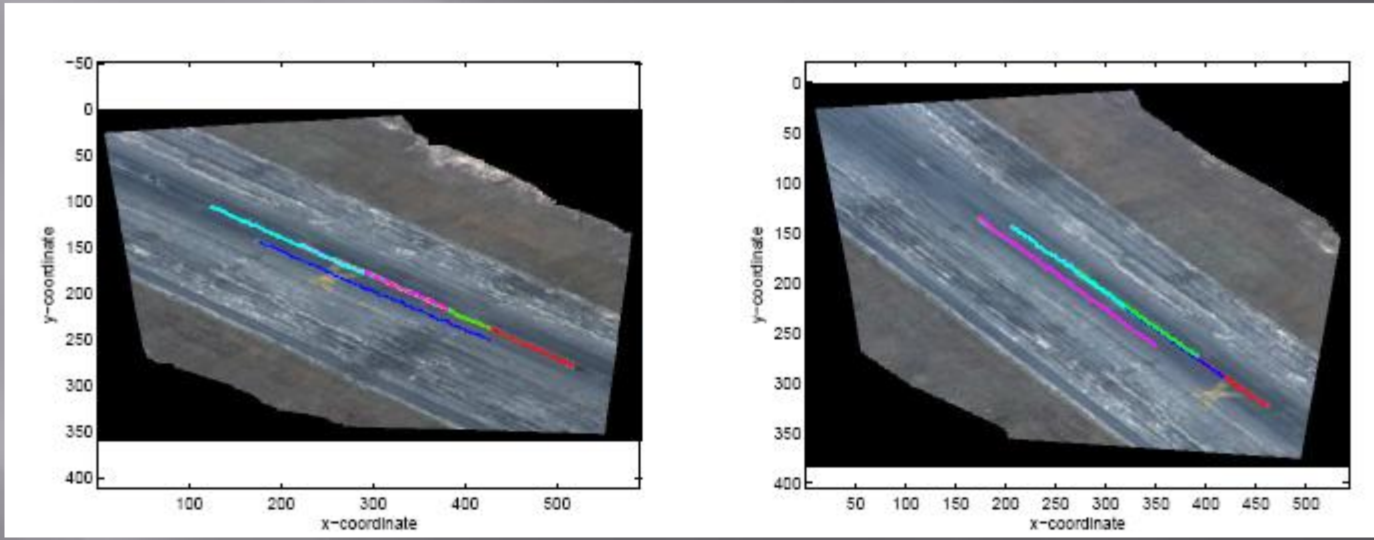
Parameter Noise





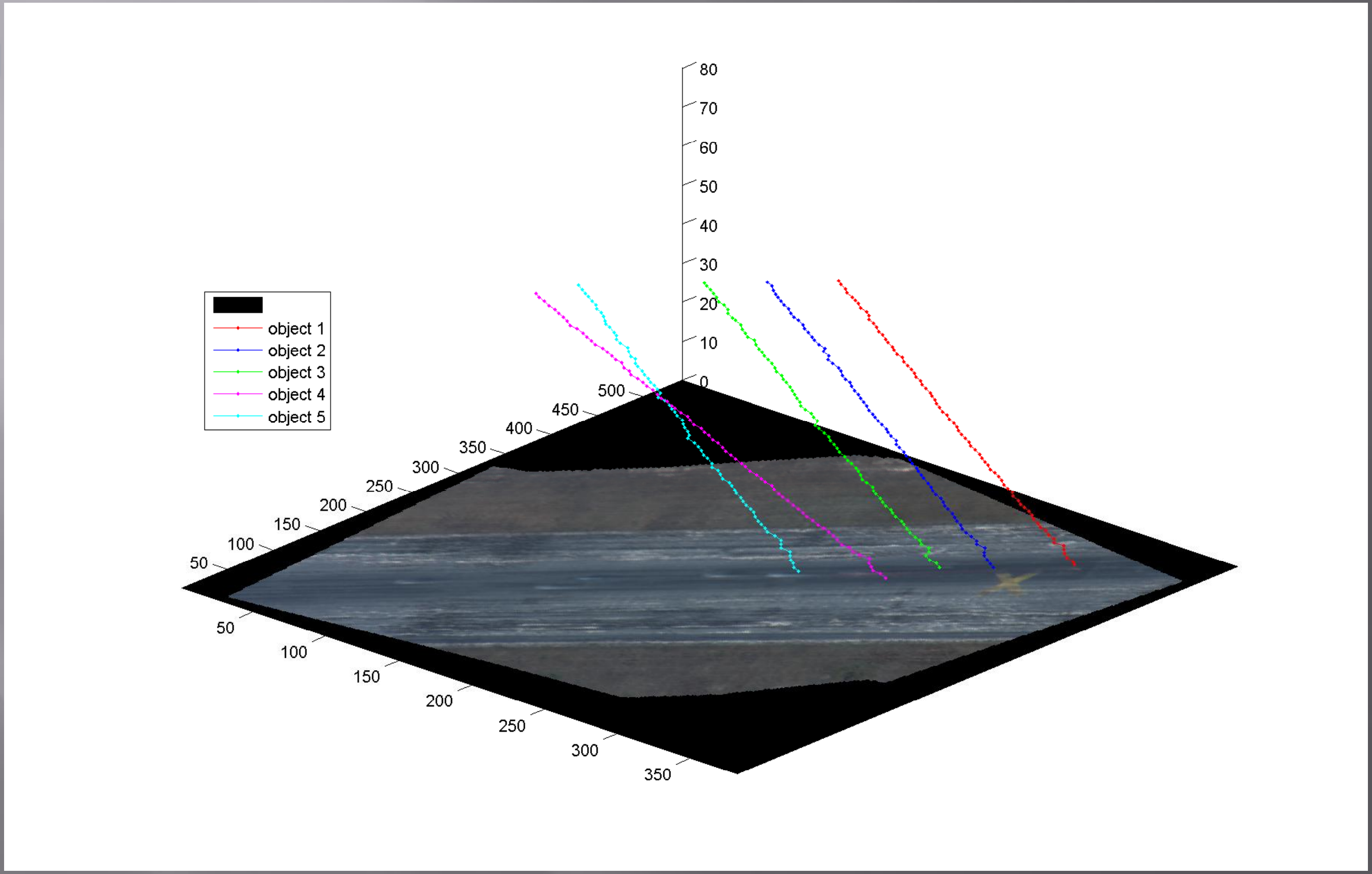


Re-association Experiment I



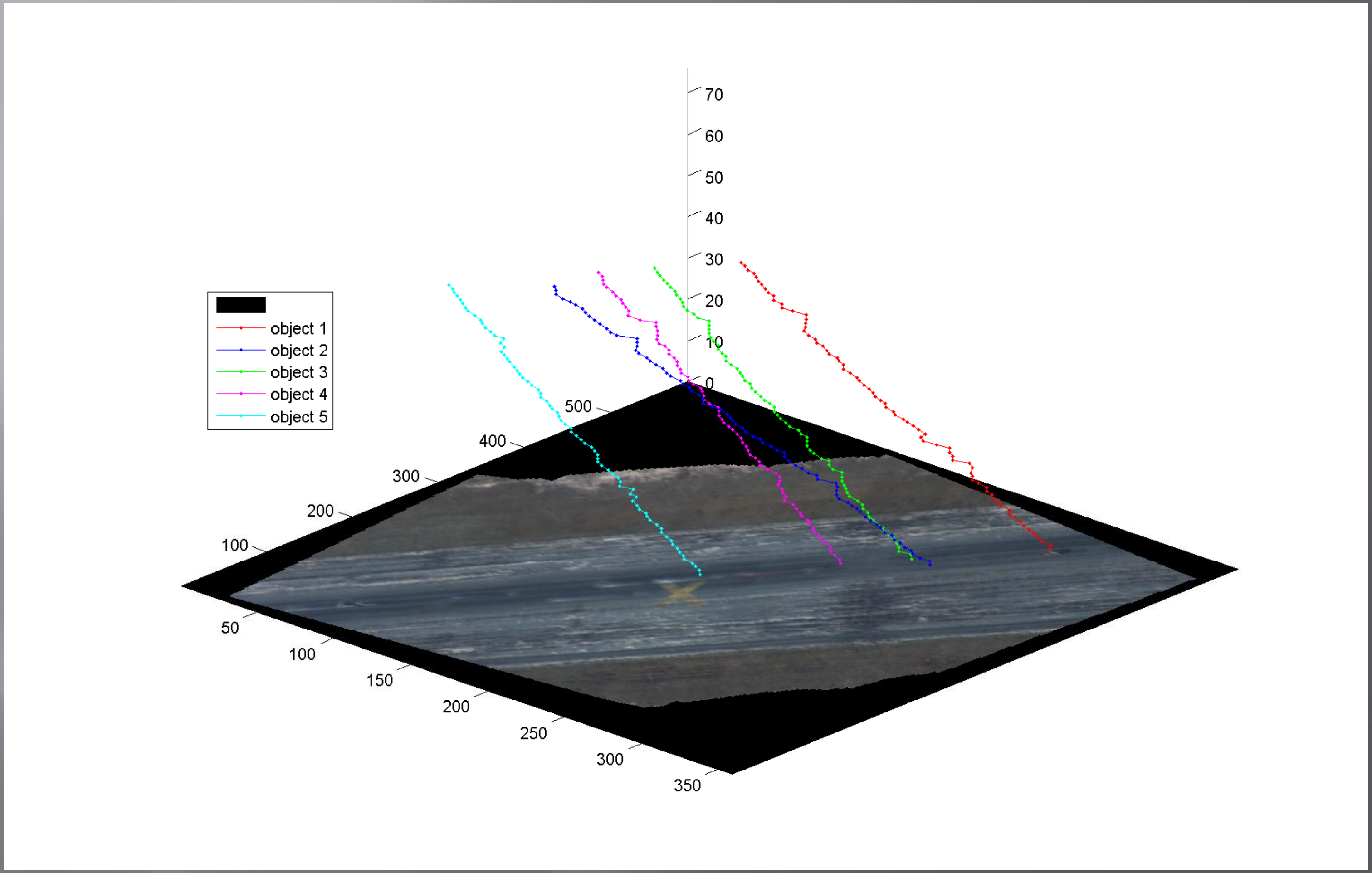
University of Central Florida VISION

Re-association Experiment 1



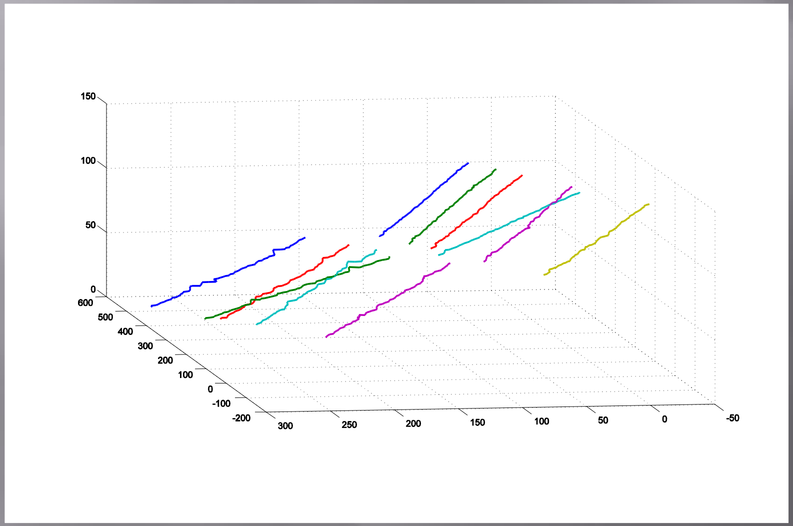
University of Central Florida VISION

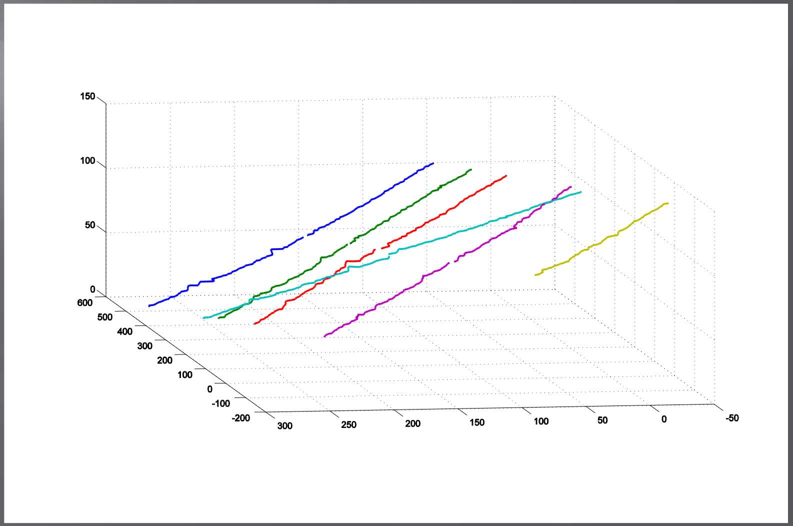
Re-association Experiment I



University of Central Florida VISION

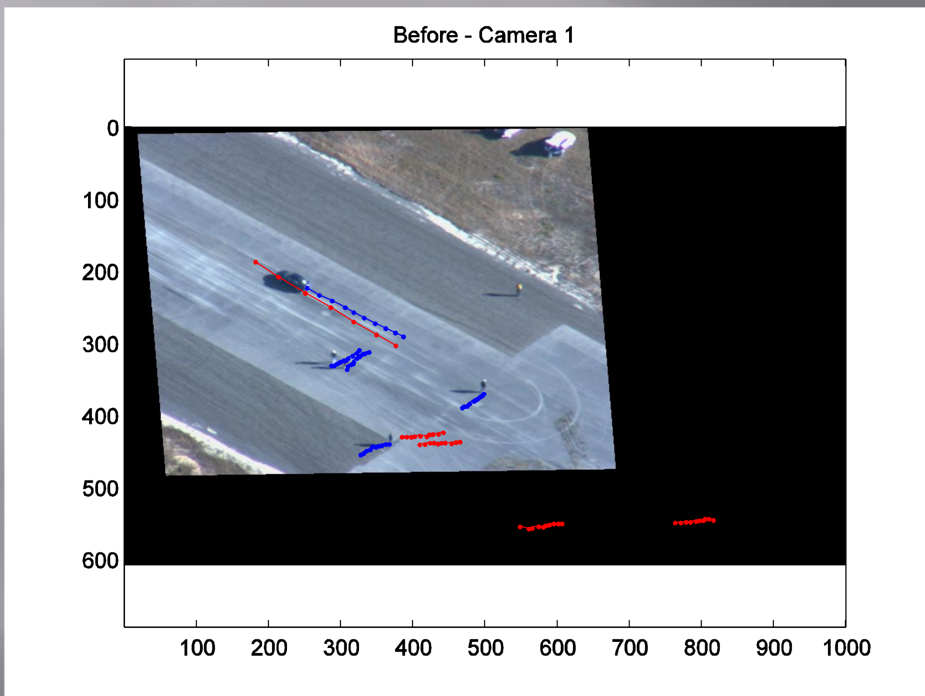
Before and After

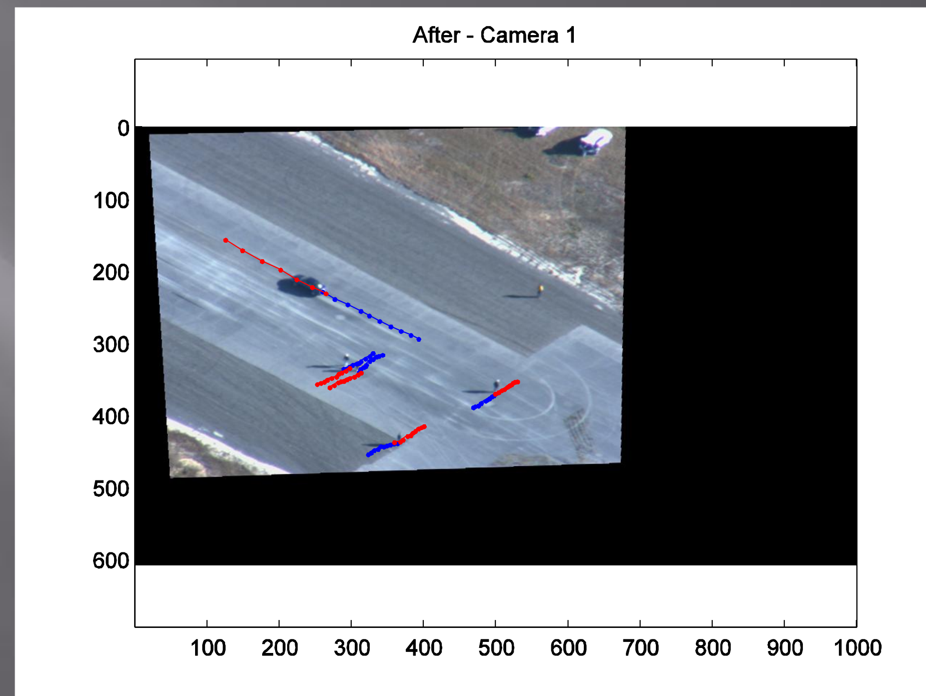




University of Central Florida VISION

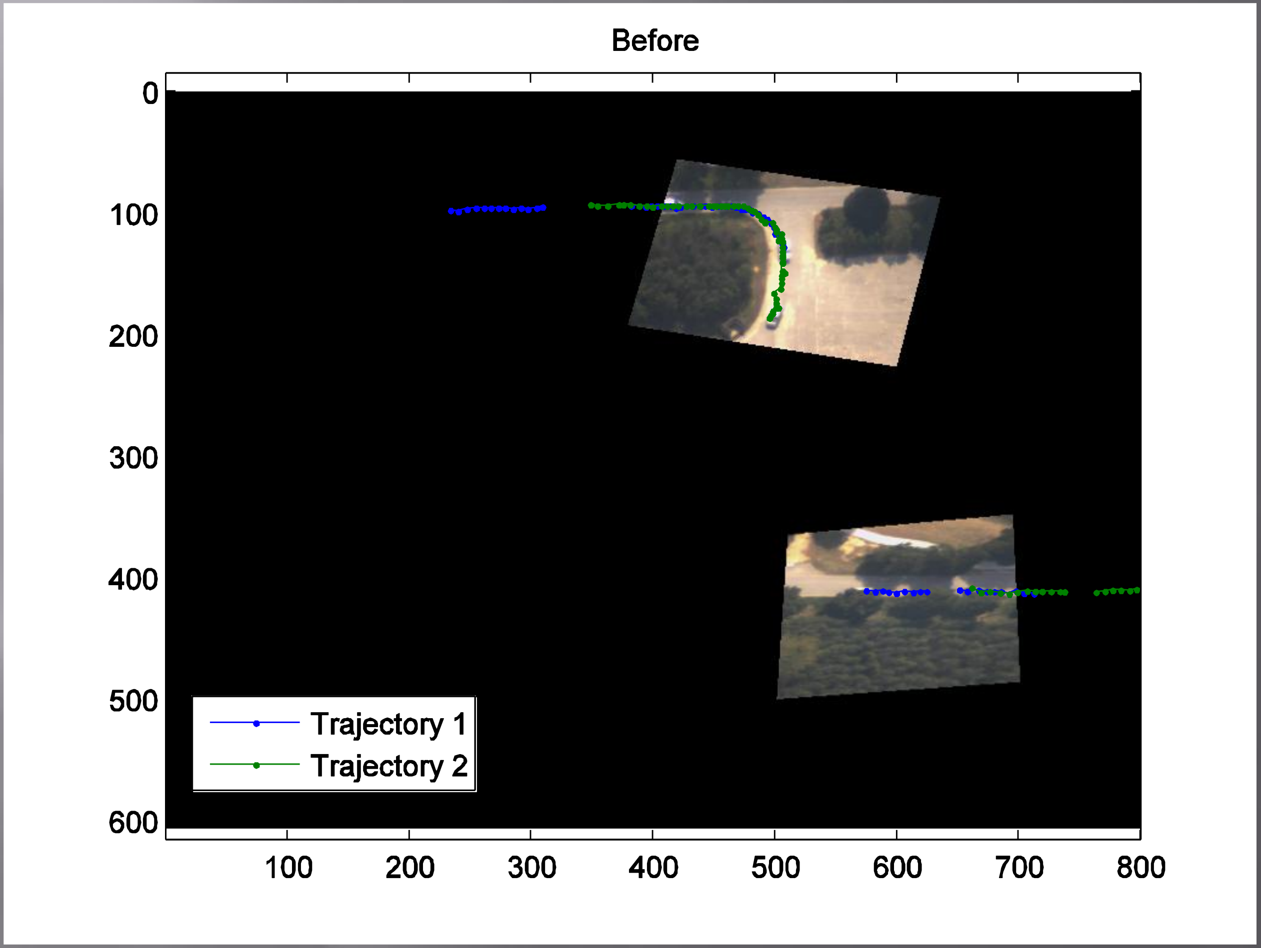
Experiment-2





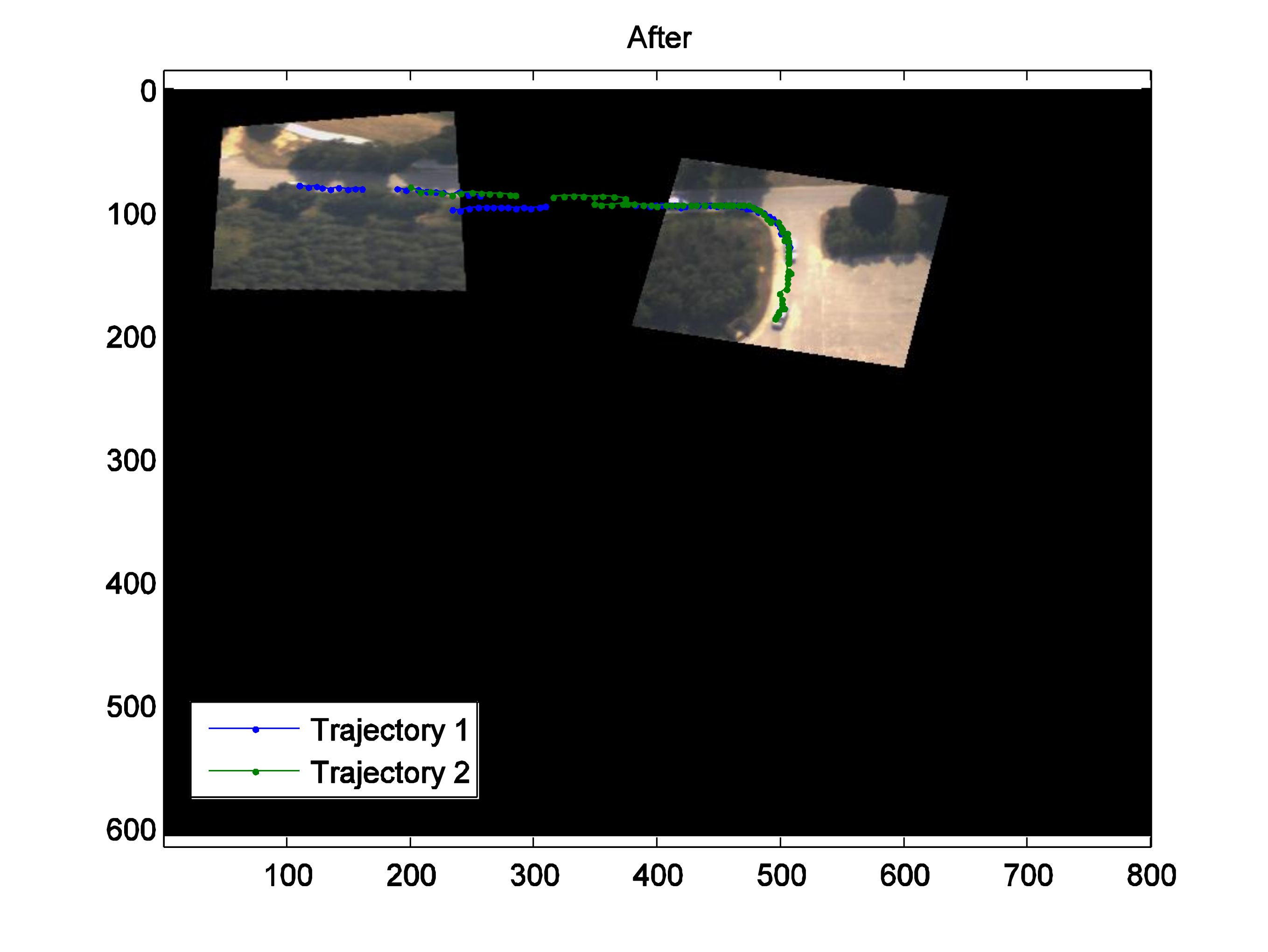
University of Central Florida VISION

Experiment 3



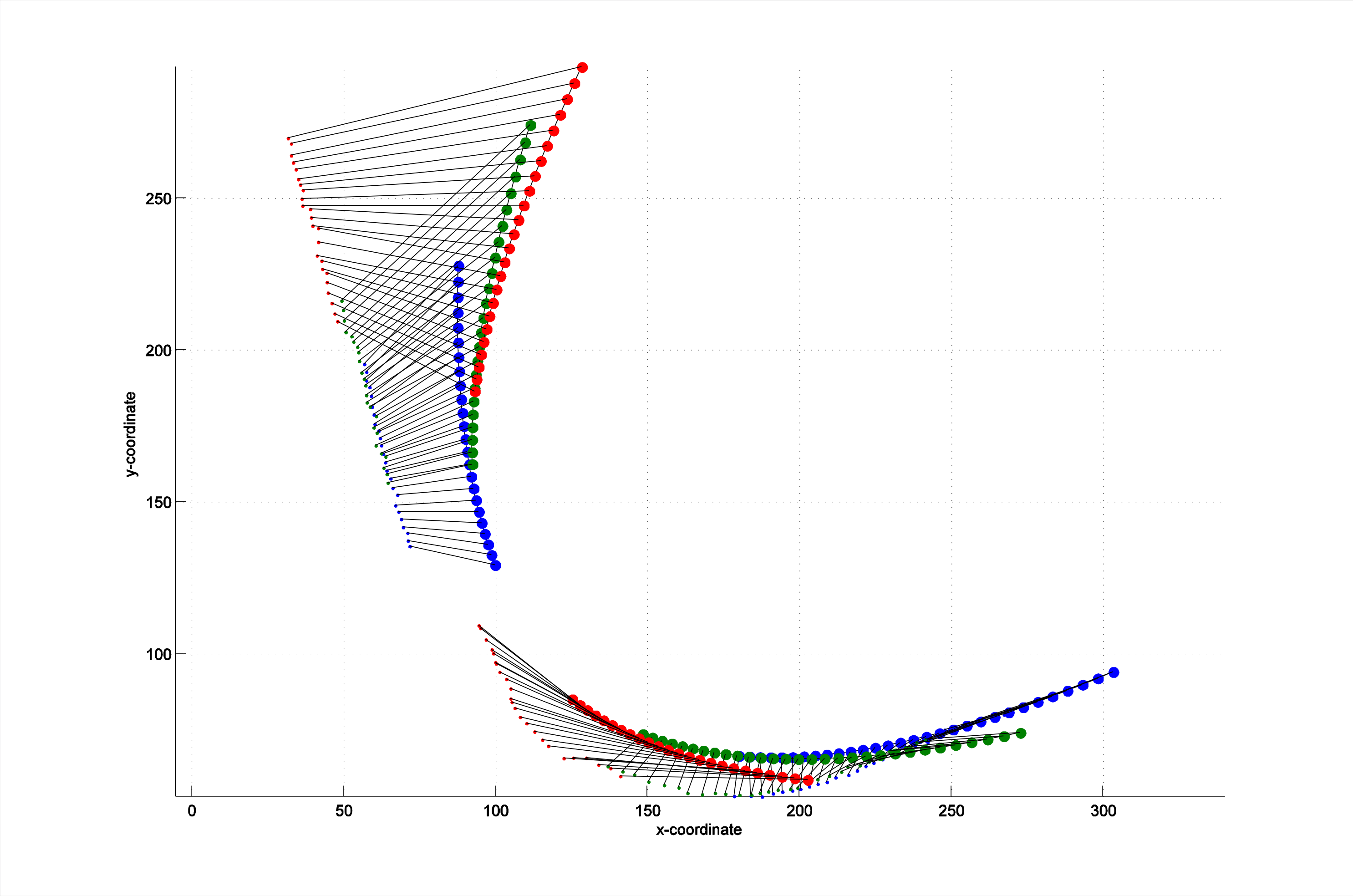
University of Central Florida VISION

Experiment 3



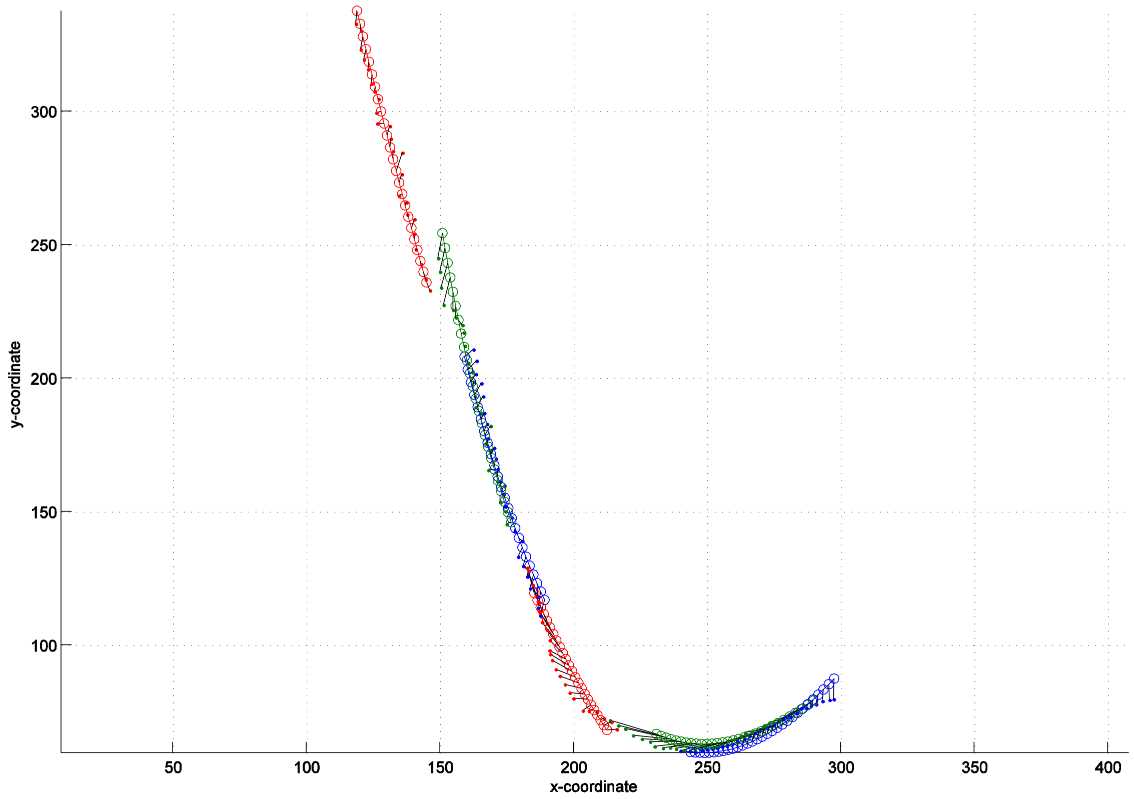
University of Central Florida VISION

Before and After



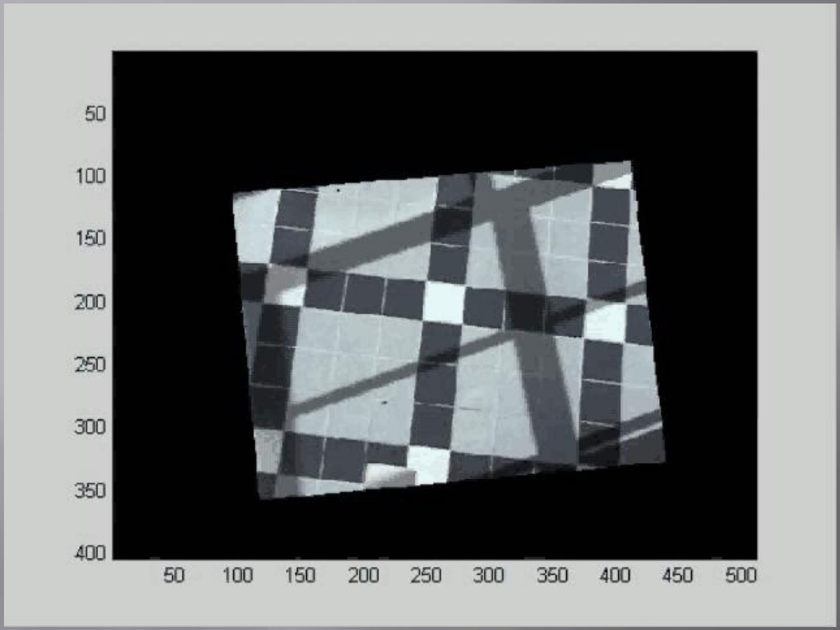
University of VISION Central Florida

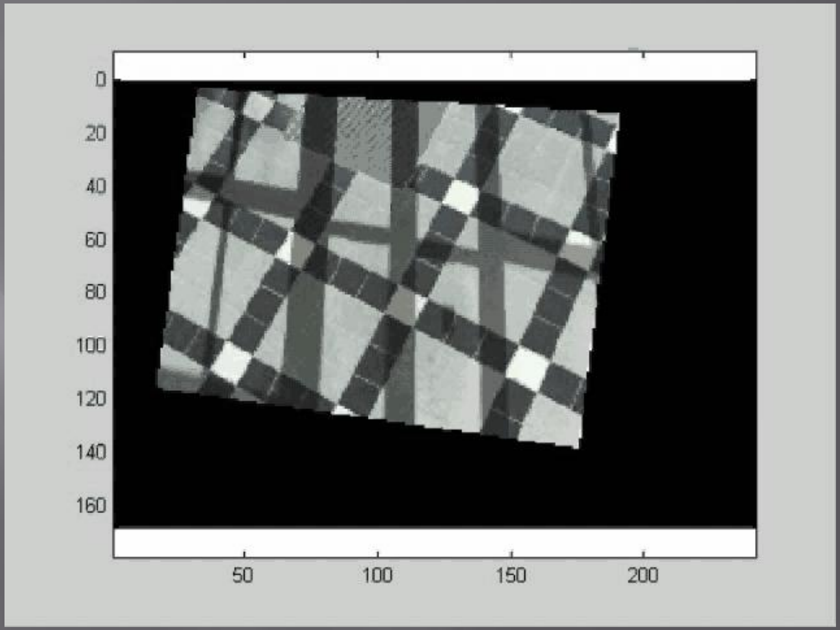
Before and After



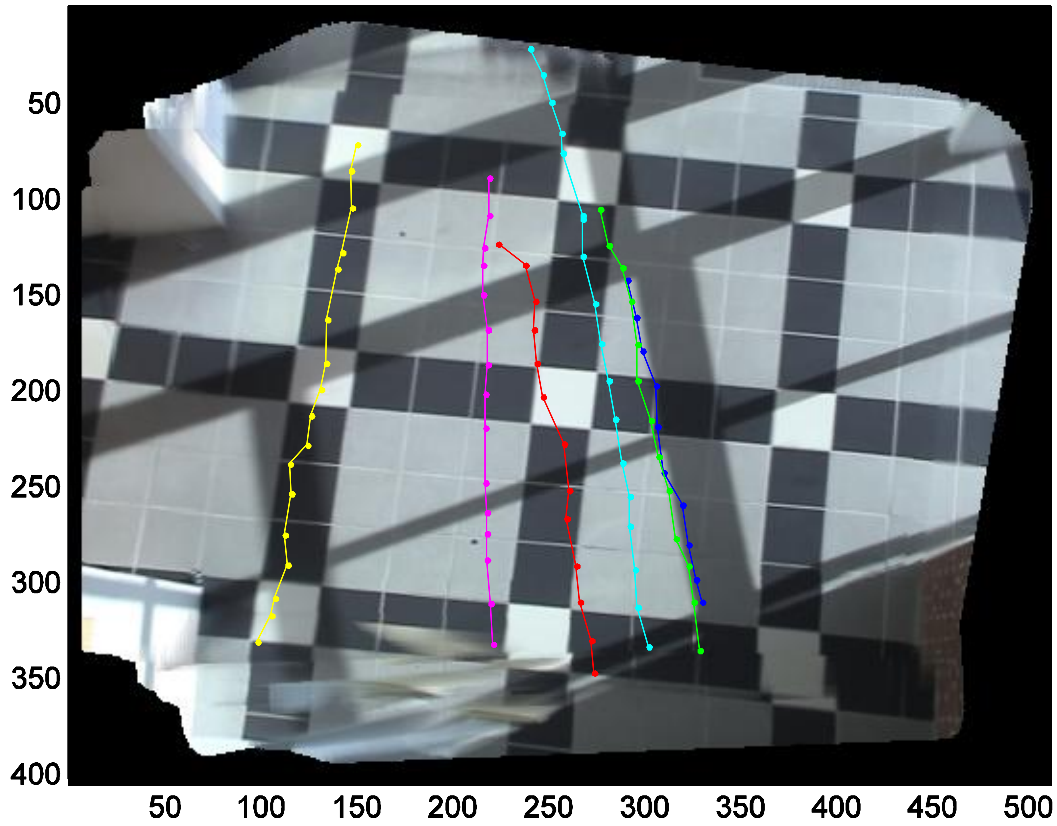
University of Central Florida VISION

CSB Sequence

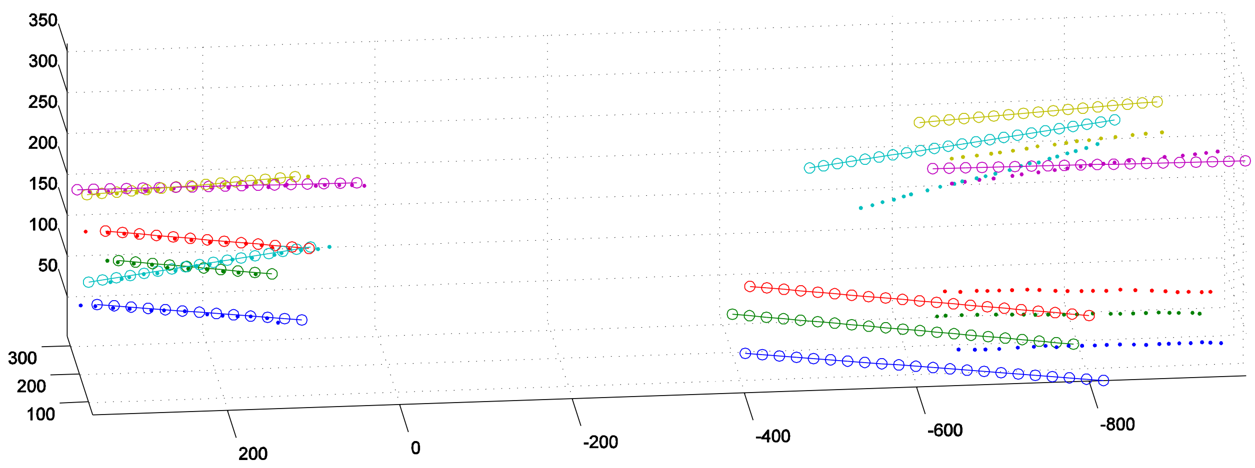




Experiment 5

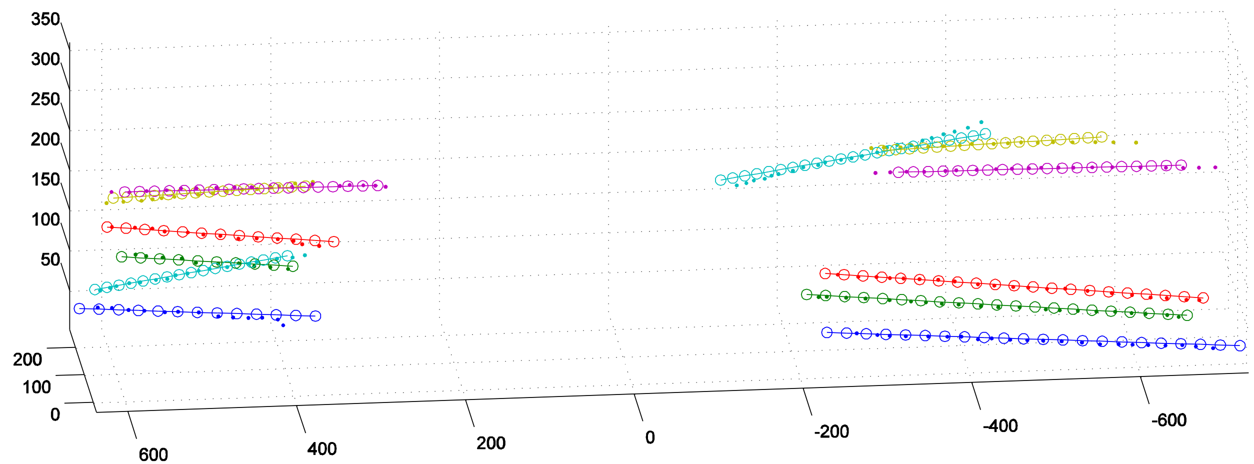


Experiment 5



University of Central Florida VISION

Experiment 5

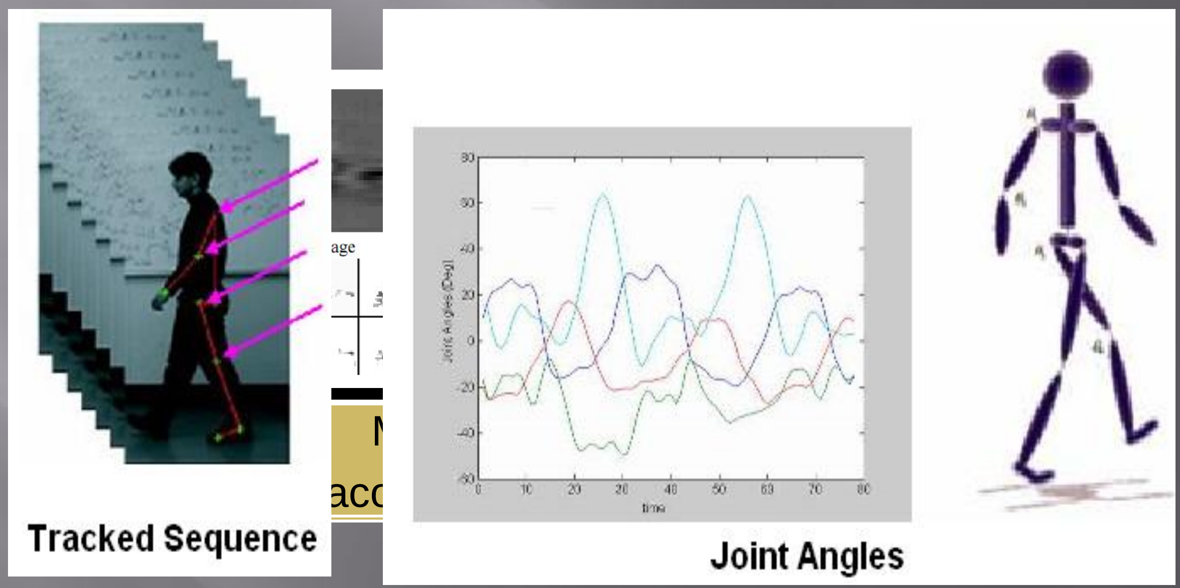


HUMAN ACTION RECOGNITION

Part III

Actions in the Computer Vision Literature

- Categorization on the basis of *representation*



University of Central Florida VISION

Contents

- View Invariant Action Recognition
- Actions As Objects
- Anthropometric Representation for Invariant Action Recognition
- Action Recognition In Two Moving Cameras
- Chaotic Invariants for Human Action Recognition

VIEW INVARIANT REPRESENTATION & RECOGNITION OF ACTIONS

Cen Rao, Alper Yilmaz, Alexi Gritai

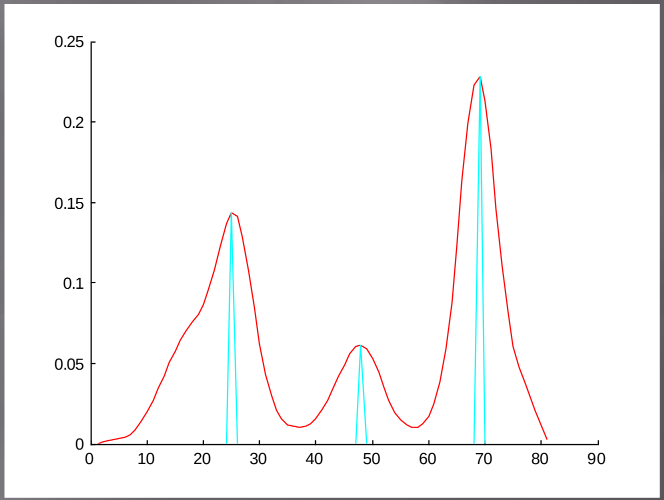
IJCV 2002

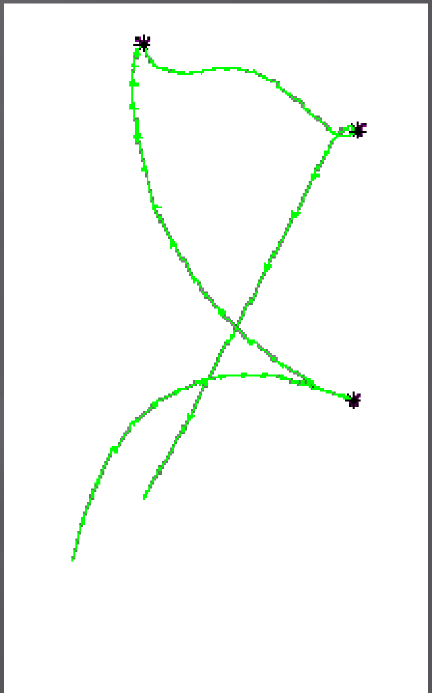
ICCV 2003

University of Central Florida VISION

Representation of Actions

- Dynamic Instants:
  - Maximum in spatiotemporal curvature represents an important change of motion characteristic.
- Intervals

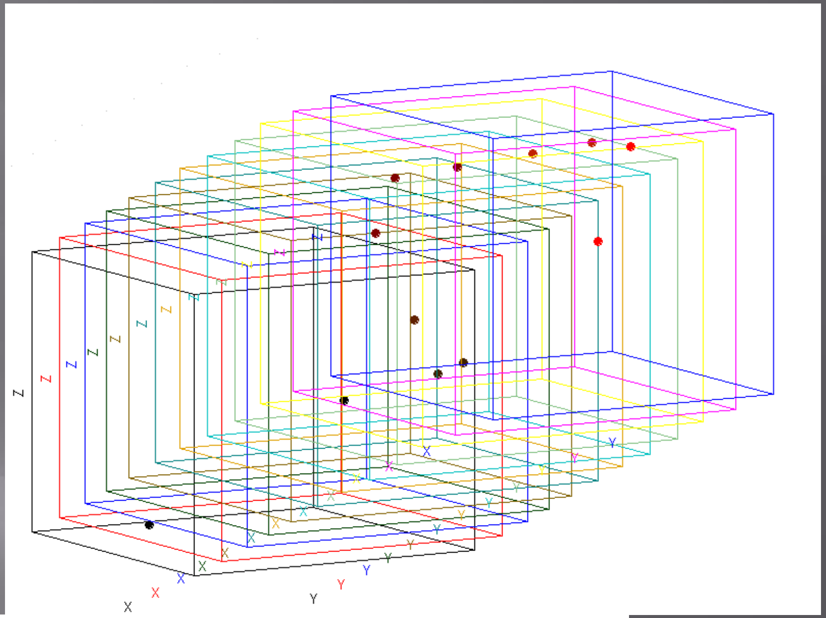




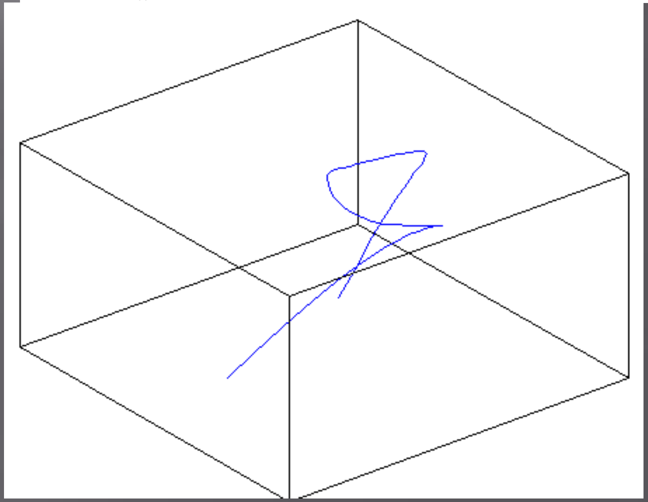
University of Central Florida VISION

Action Trajectory in 4D

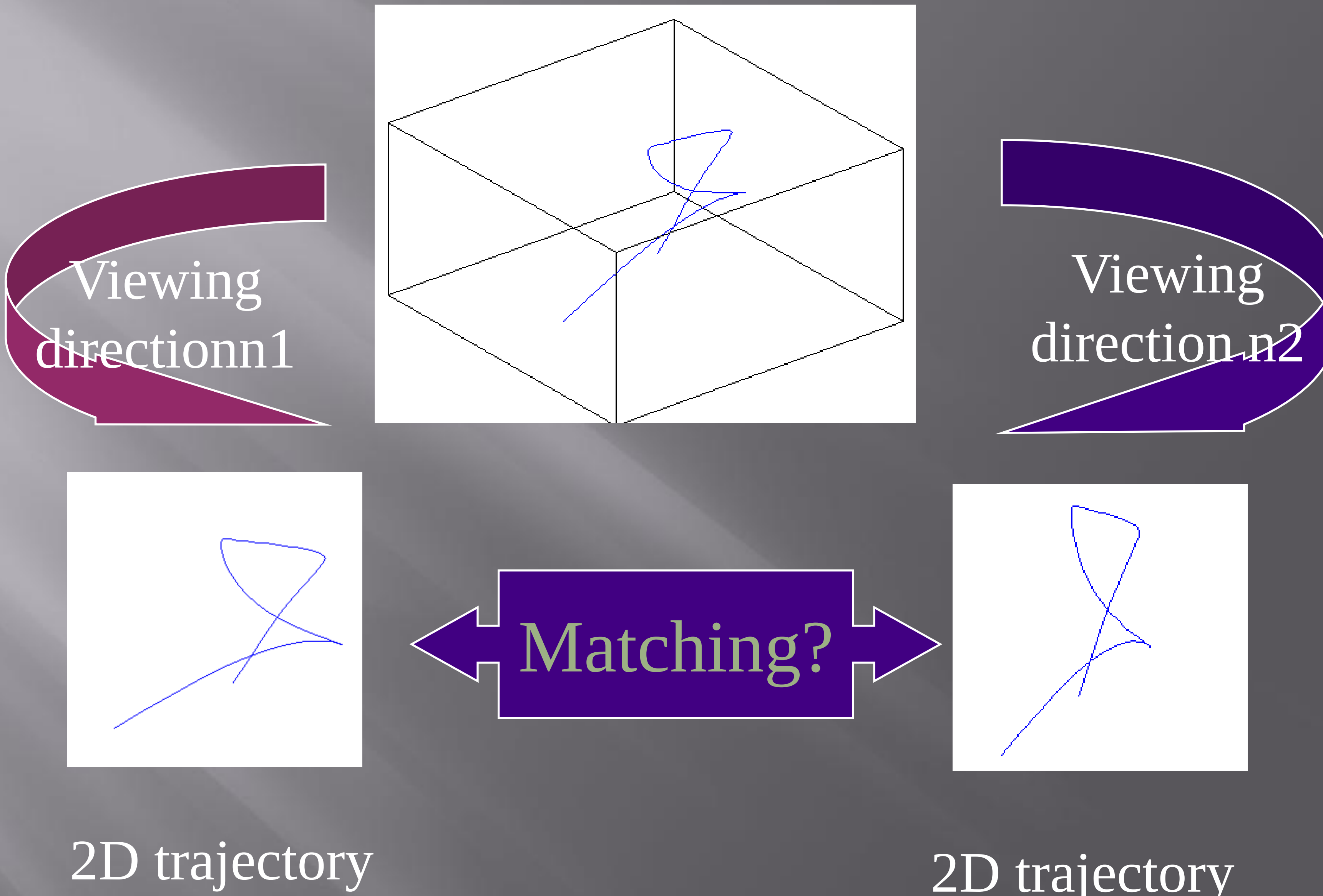
Sampling in time



Ignore the time index



University of Central Florida VISION



University of VISION
Central Florida

Affine View Invariant Matching

Rank Theorem (Tomasi & Kanade)

- S is a set of 3-D points and Πs are projection matrices for different viewpoints, then we can arrange image coordinates of points in an observation matrix, M, as follows:

$$M = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ y_1^1 & y_2^1 & \cdots & y_n^1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ y_1^2 & y_2^2 & \cdots & y_n^2 \end{bmatrix} = PS = \begin{bmatrix} \Pi_1 \\ \Pi_2 \end{bmatrix} \bullet \begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ Y_1 & Y_2 & \cdots & Y_n \\ Z_1 & Z_2 & \cdots & Z_n \end{bmatrix}$$

$$\Pi_v = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \end{bmatrix}$$

***M* is 4 by *n*, *P* is 4 × 3 and *S* is 3 × *n* ,**

then the rank of M is at most 3.

University of Central Florida VISION

Generalized Affine Rank Theorem

- A set of image points match *if and only if* M is of rank at most 3. (Shapiro & Zisserman, Seitz & Dyer)
- A set of "instants" match *if and only if* M of rank at most 3 . Therefore, the similarity measure is:

$$M = \begin{bmatrix} \mu_1^i & \mu_2^i & \cdots & \mu_n^i \\ v_1^i & v_2^i & \cdots & v_n^i \\ \mu_1^j & \mu_2^j & \cdots & \mu_n^j \\ v_1^j & v_2^j & \cdots & v_n^j \end{bmatrix} \qquad dist_{ij} = \sigma_4$$

Perspective View-Invariant Matching

- Fundamental matrix captures the relationship between the corresponding points in two views.

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}^T F \begin{bmatrix} u'_i \\ v'_i \\ 1 \end{bmatrix} = 0, \qquad F = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix}$$

Perspective View-invariant Measure

- Consider the fundamental matrix constraint and rearrange the constraint as following:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}^T F \begin{bmatrix} u'_i \\ v'_i \\ 1 \end{bmatrix} = 0, \qquad M f = \begin{bmatrix} u_1u'_1 & u_1v'_1 & u_1 & v_1u'_1 & v_1v'_1 & v_1 & u'_1 & v'_1 & 1 \\ u_2u'_2 & u_2v'_2 & u_2 & v_2u'_2 & v_2v'_2 & v_2 & u'_2 & v'_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_nu'_n & u_nv'_n & u_n & v_nu'_n & v_nv'_n & v_n & u'_n & v'_n & 1 \end{bmatrix} f = 0$$

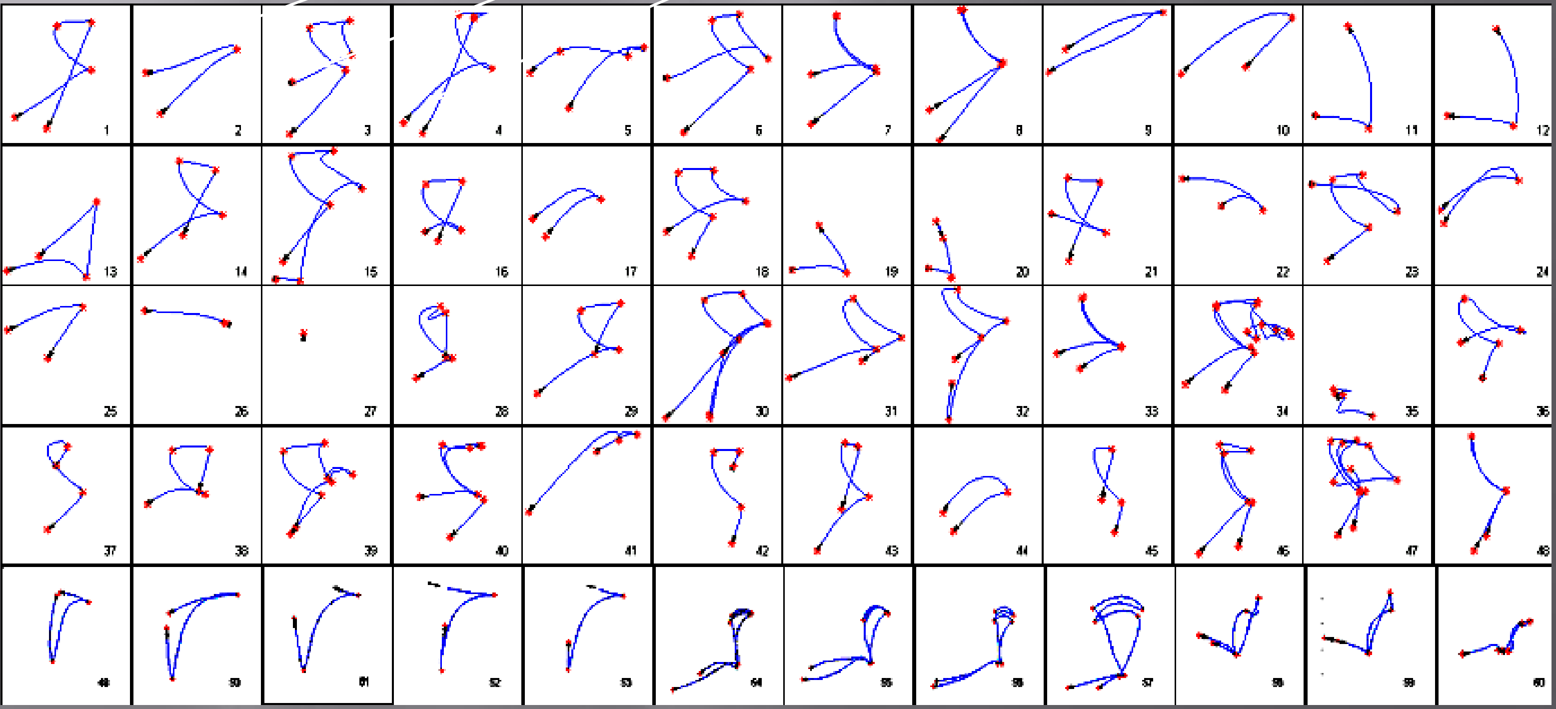
M is 9 by *n* matrix $\qquad f = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{21} & f_{22} & f_{23} & f_{31} & f_{32} & f_{33} \end{bmatrix}^T$

To solve the equation, the rank(M) must be 8.
The 9^{th} singular value of M, σ_9, is the match measure.

University of Central Florida VISION

Experimental Results

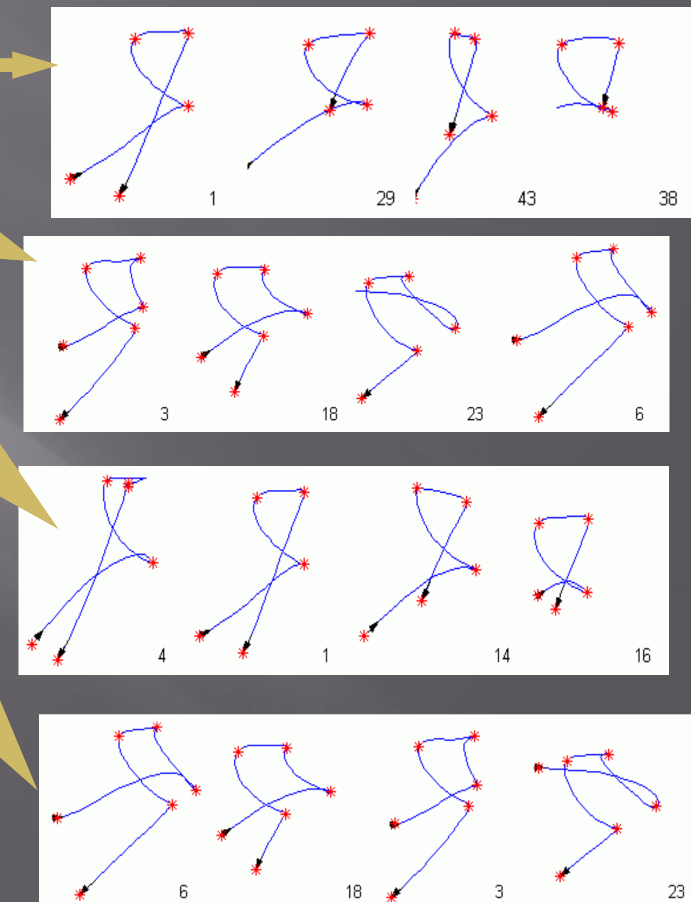
60 Action Trajectories
7 People



University of Central Florida VISION

Experimental Results

Actions	3 Best matches	Evaluation & comments
1	29 43 38	Correct
2	Pick up	Correct
3	18 23 6	Correct
4	1 14 16	One wrong
5		Unique action
6	18 3 23	Correct
7	48 33 8	correct
8	48 33 7	One wrong
9	Pick up	Correct
10	Put down	Correct
11	Pick up	Correct
12	Put down	Correct
13		Unique action
14	43 16 1	Correct
15		Unique action
16	14 29 1	Correct
17	**Pick up**	Incorrect, object hidden
18	6 3 23	Correct
19	Pick up	Correct
20		Unique random motion



Central Florida

Temporal Alignment Results

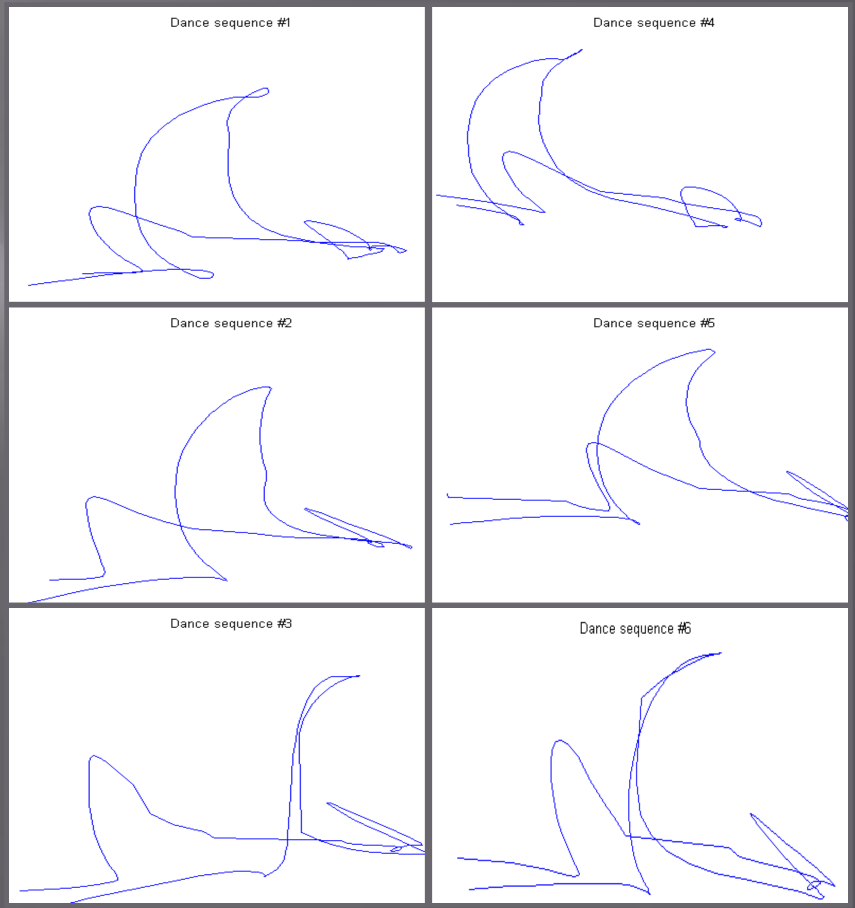
Before Temporal Alignment

After Temporal Alignment

Temporal Alignment of Videos

Input videos:

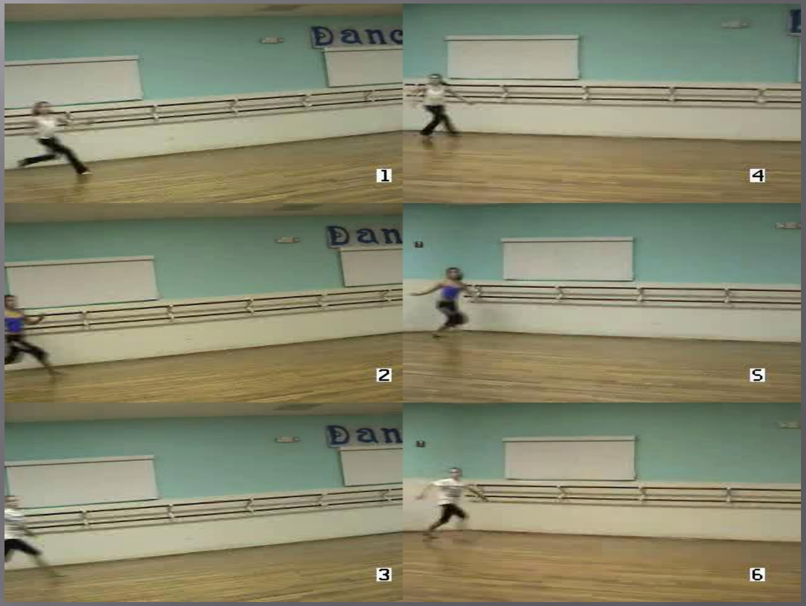
Trajectories of the right foot:



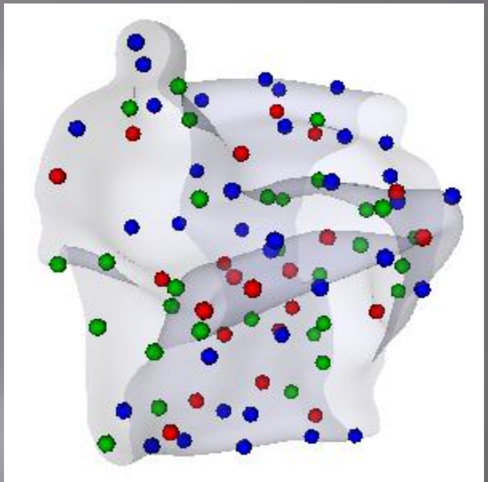
University of Central Florida VISION

Temporal Alignment Results

Synchronized videos:



University of Central Florida VISION



ACTION AS OBJECTS

Alper Yilmaz

- A. Yilmaz and M. Shah "Actions Sketch: A Novel Action Representation," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2005.
- A. Yilmaz and M. Shah "Representing Actions Using Differential Geometry," Computer Vision and Image Understanding (CVIU), submitted 2006.

University of Central Florida VISION

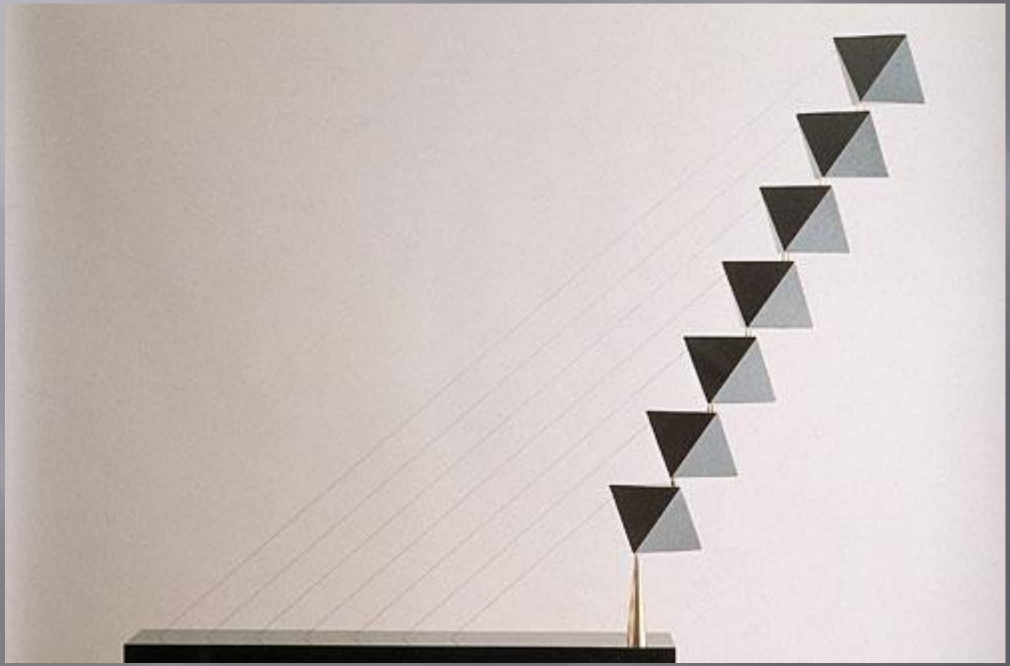
Actions As Objects

When something moves it develops a shape.
Santiago Calatrava
(Sculpture into architecture)

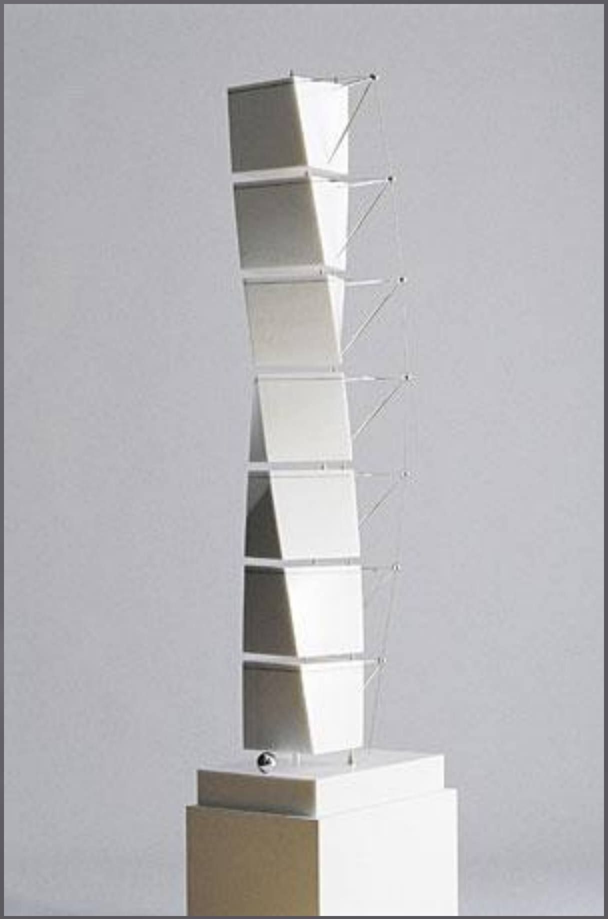
Milwaukee Museum of Art

University of Central Florida VISION

Actions As Objects



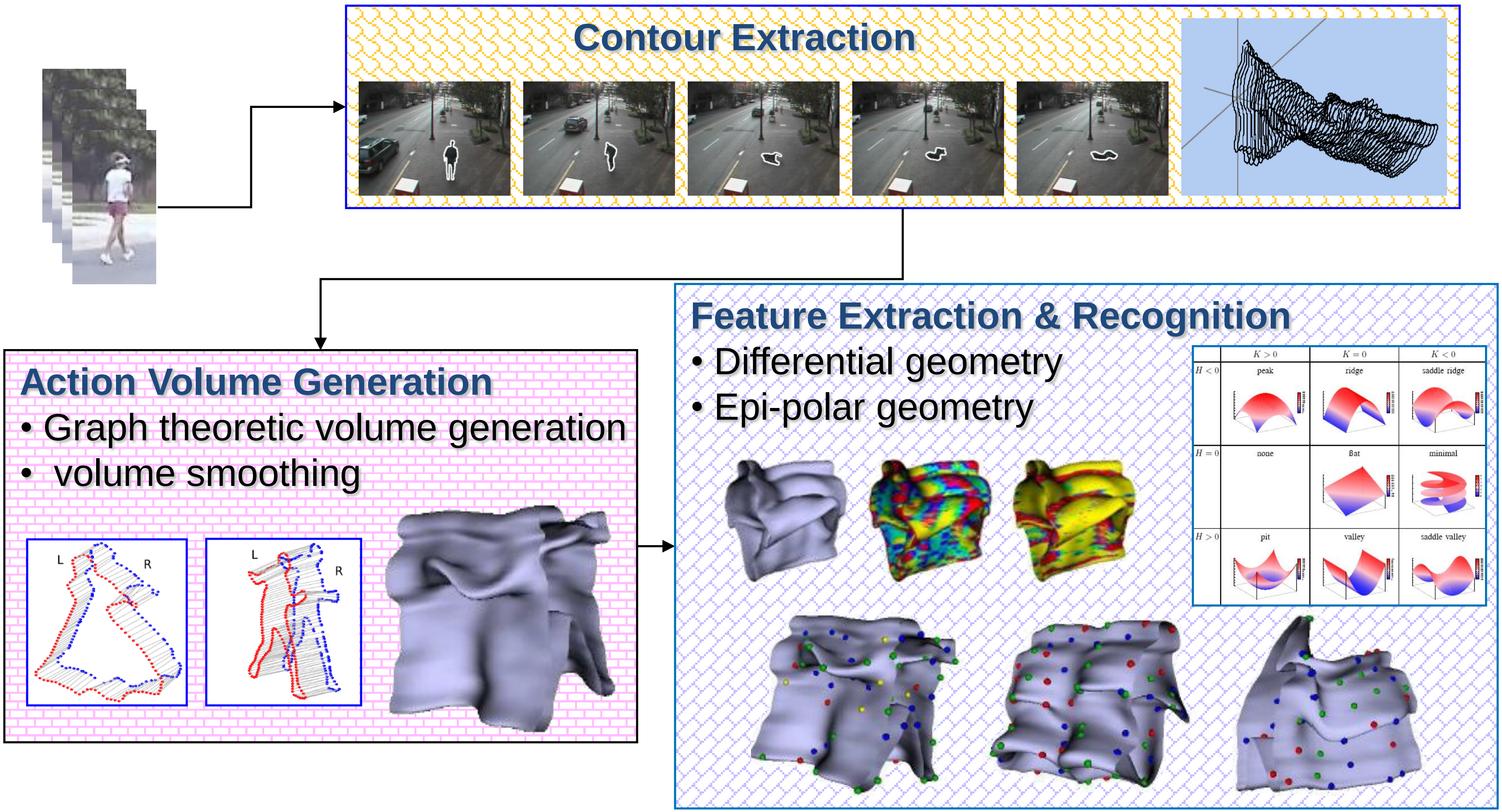
Musical Star



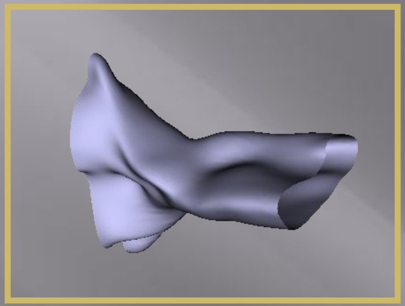
Turning Torso

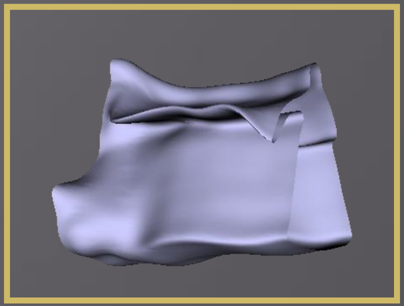
University of Central Florida VISION

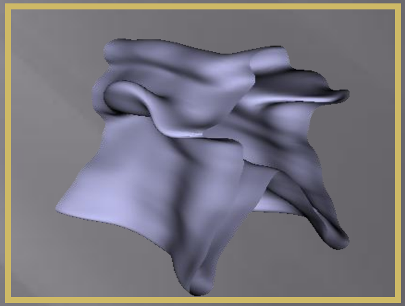
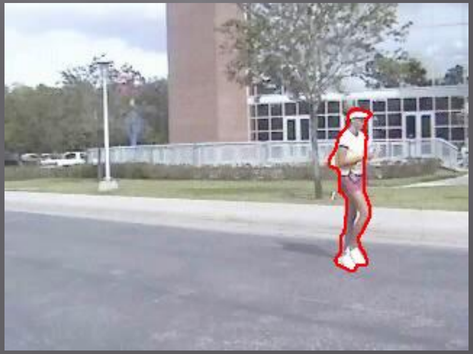
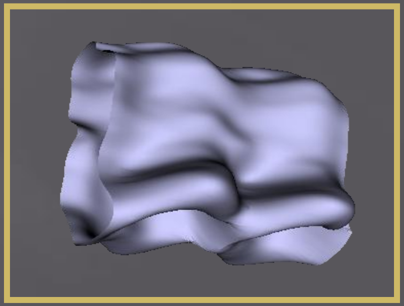
Flow diagram



Resulting Volume

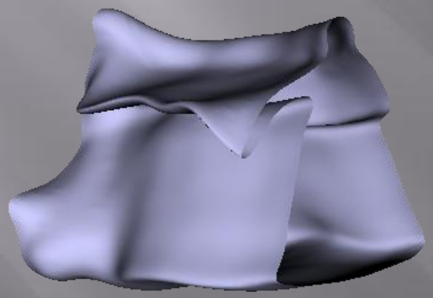


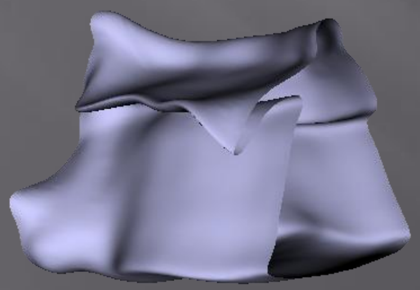
University of Central Florida VISION

Properties of the Action Volume

- Space-time (3D) object
- Encodes shape and motion
- Uses complete object contours instead of a single point on the object.
- Suitable for fine action analysis
- Continuous representation
  - Same volume for same action of different lengths



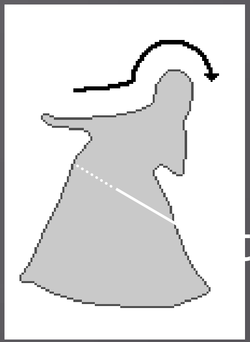
40 frames

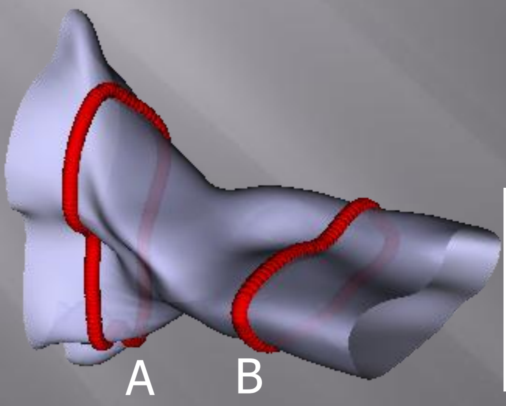


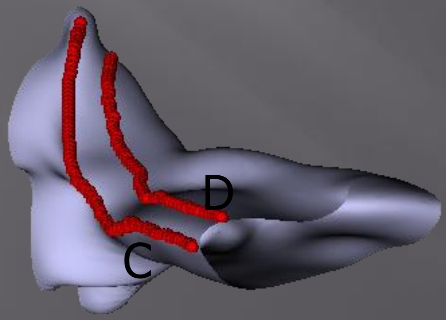
20 random selected frames

Properties of the Action Volume

- Can be represented in 2D
  - Arc length and time
- Can regenerate contour at time *t*
- Can provide spatial trajectory of contour points







What is the Action Sketch?

- Important action descriptors
  - Unique shape and motion characteristics
- Related to differential geometric properties of action volume
  - 1st and 2nd fundamental forms
    - Gaussian and mean curvatures
    - Fundamental surface types

Copyrights Mubarak Shah, UCF

Computing Gaussian (K) and Mean (H) Curvatures

- K and H are two algebraic invariants of Weingarten mapping S.

$$K = \det(\mathbf{S})$$

$$H = \frac{1}{2}\operatorname{trace}(\mathbf{S})$$

from 1st fundamental form

$$\mathbf{S} = \mathbf{g}^{-1}\mathbf{b}$$

from 2nd fundamental form

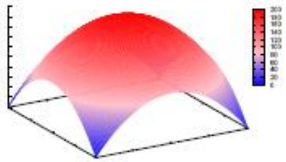
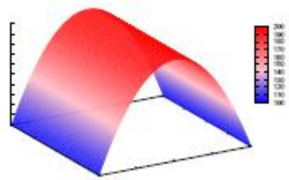
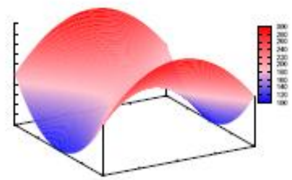
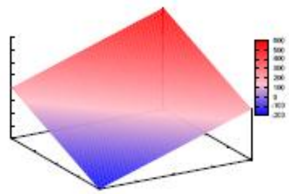
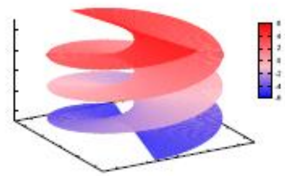
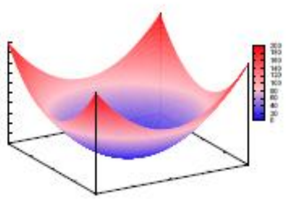
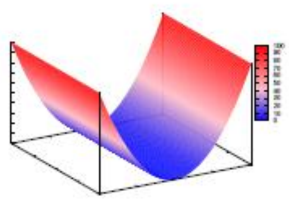
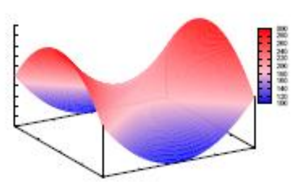
$$\mathbf{g} = \begin{bmatrix} f_s f_s & f_s f_t \\ f_s f_t & f_t f_t \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} f_{ss}\vec{n} & f_{st}\vec{n} \\ f_{st}\vec{n} & f_{tt}\vec{n} \end{bmatrix}$$

where $f(s,t)$ is a point on the volume, n is normal at f

University of Central Florida VISION

Copyrights Mubarak Shah, UCF

Fundamental Surface Types

	$K > 0$	$K = 0$	$K < 0$
$H < 0$	peak 	ridge 	saddle ridge 
$H = 0$	none	flat 	minimal 
$H > 0$	pit 	valley 	saddle valley 

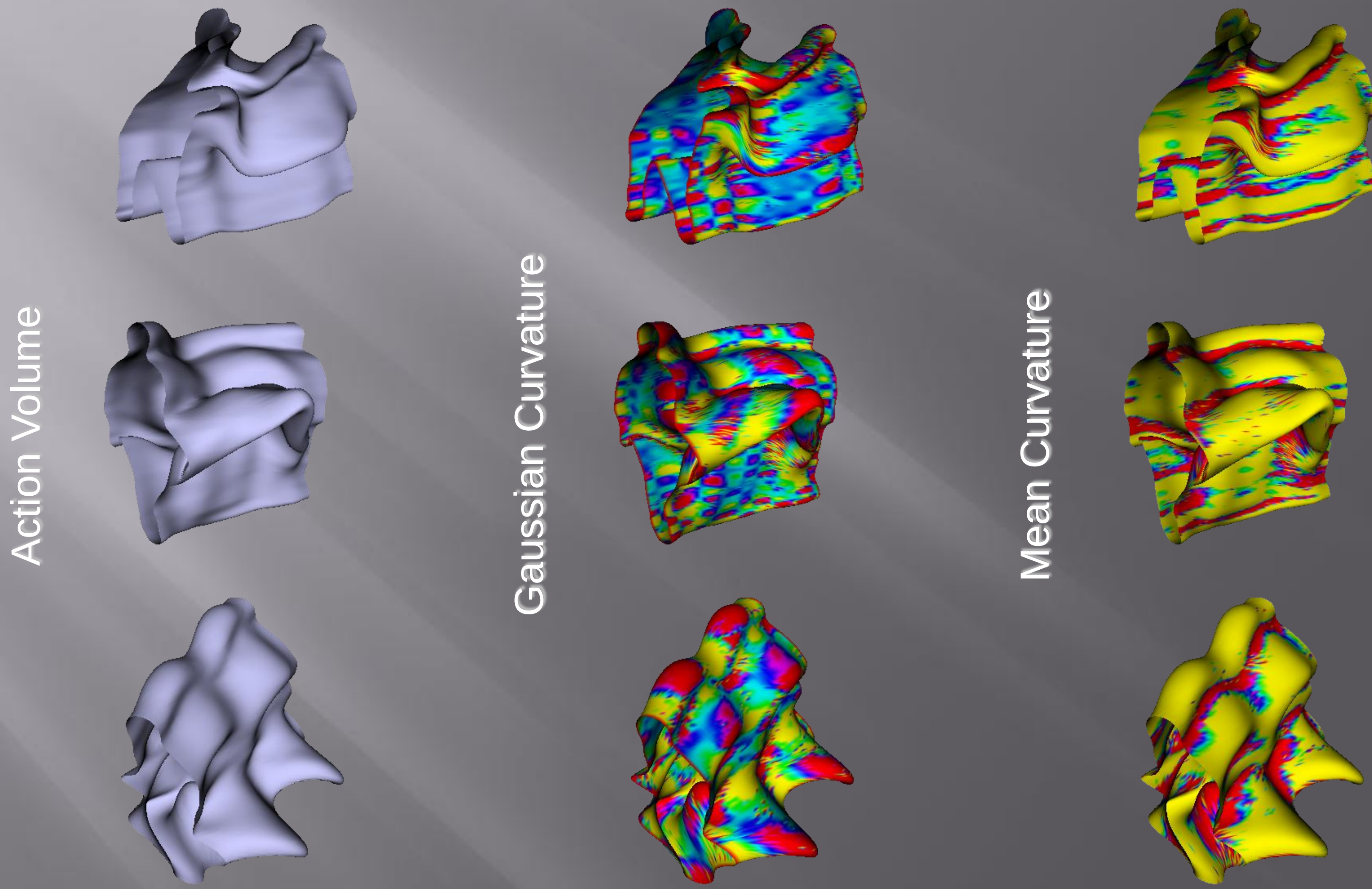
University of Central Florida VISION

Properties of Surface Types

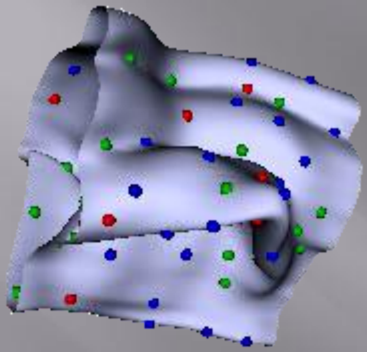
- Rotation and translation invariant in spatio-temporal space.
- Encodes intrinsic properties of surface.
  - Defines the convexity or concavity of surface.
- Related to speed and acceleration.

Copyrights Mubarak Shah, UCF

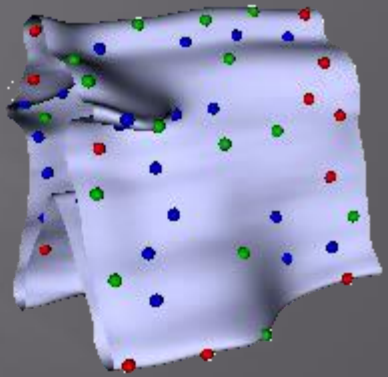
Differential Geometric Surface



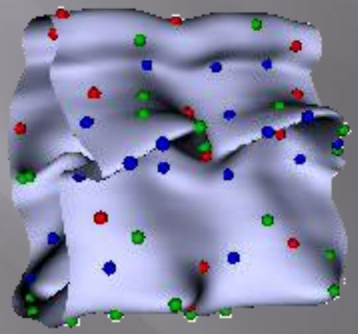
Examples



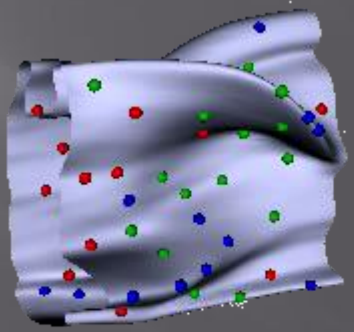
kicking



dance



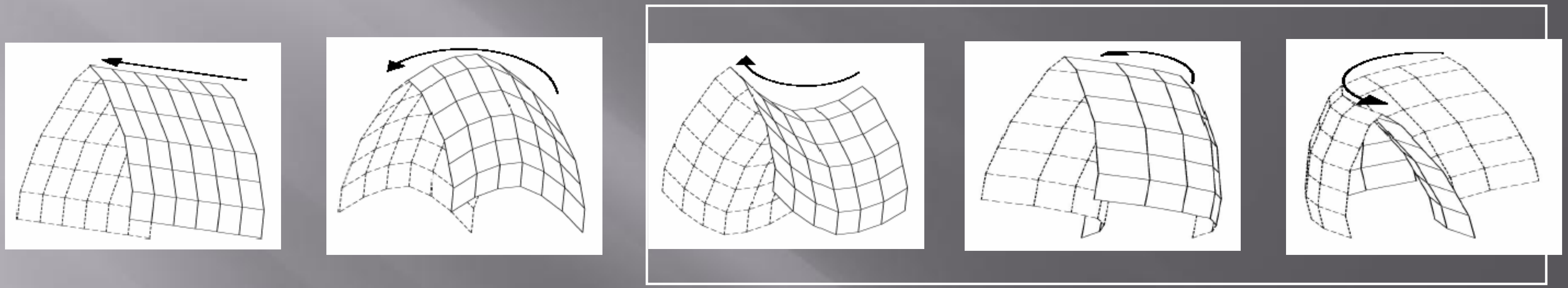
walking



surrender

University of Central Florida VISION

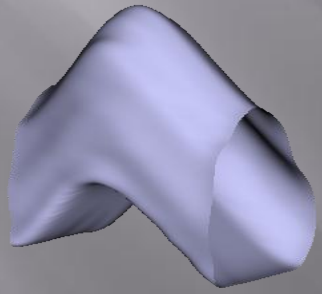
Surface patches & their relation to the object motion



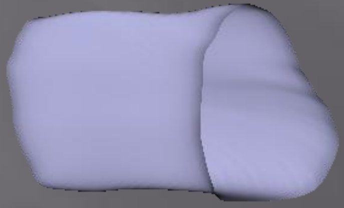
ridge peak saddle ridge

Action Descriptors Relation to Object Motion

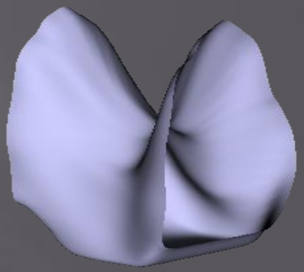
peak



ridge



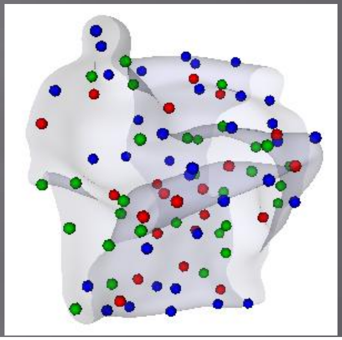
saddle ridge

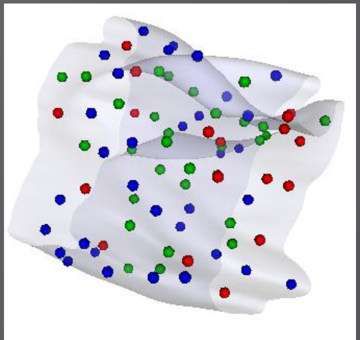


University of VISION Central Florida

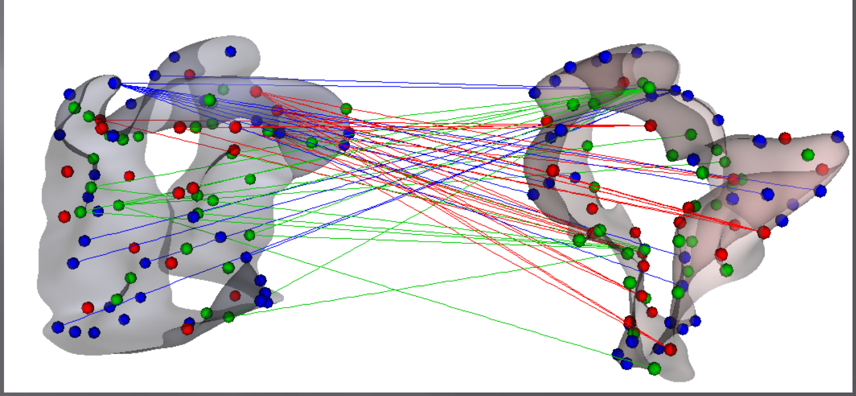
Matching Volumes: Establishing Correspondence

- Generate bipartite action graphs





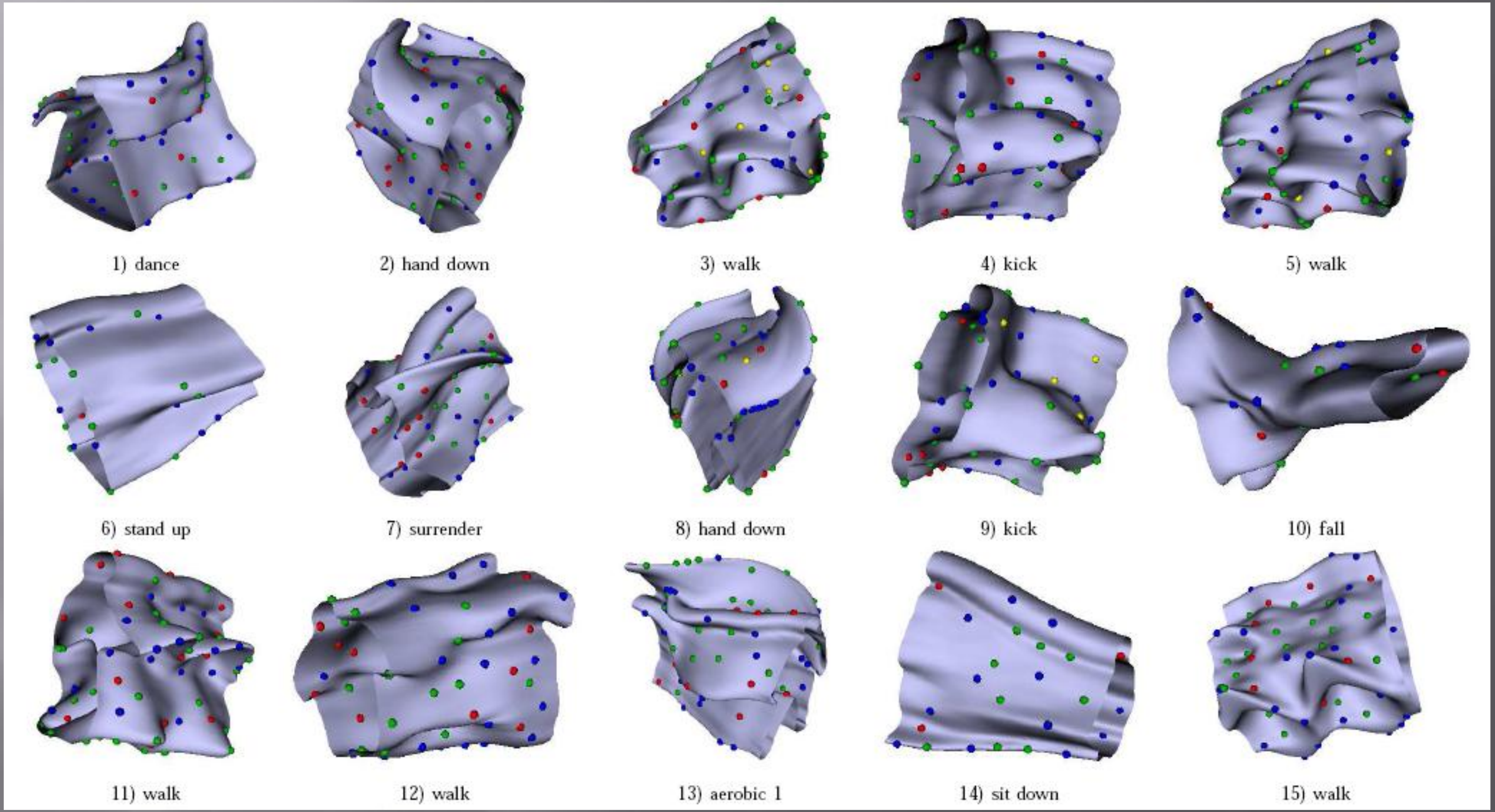
- Define weights by
  - Space-time proximity
  - Shape similarity



- Find Maximum Matching

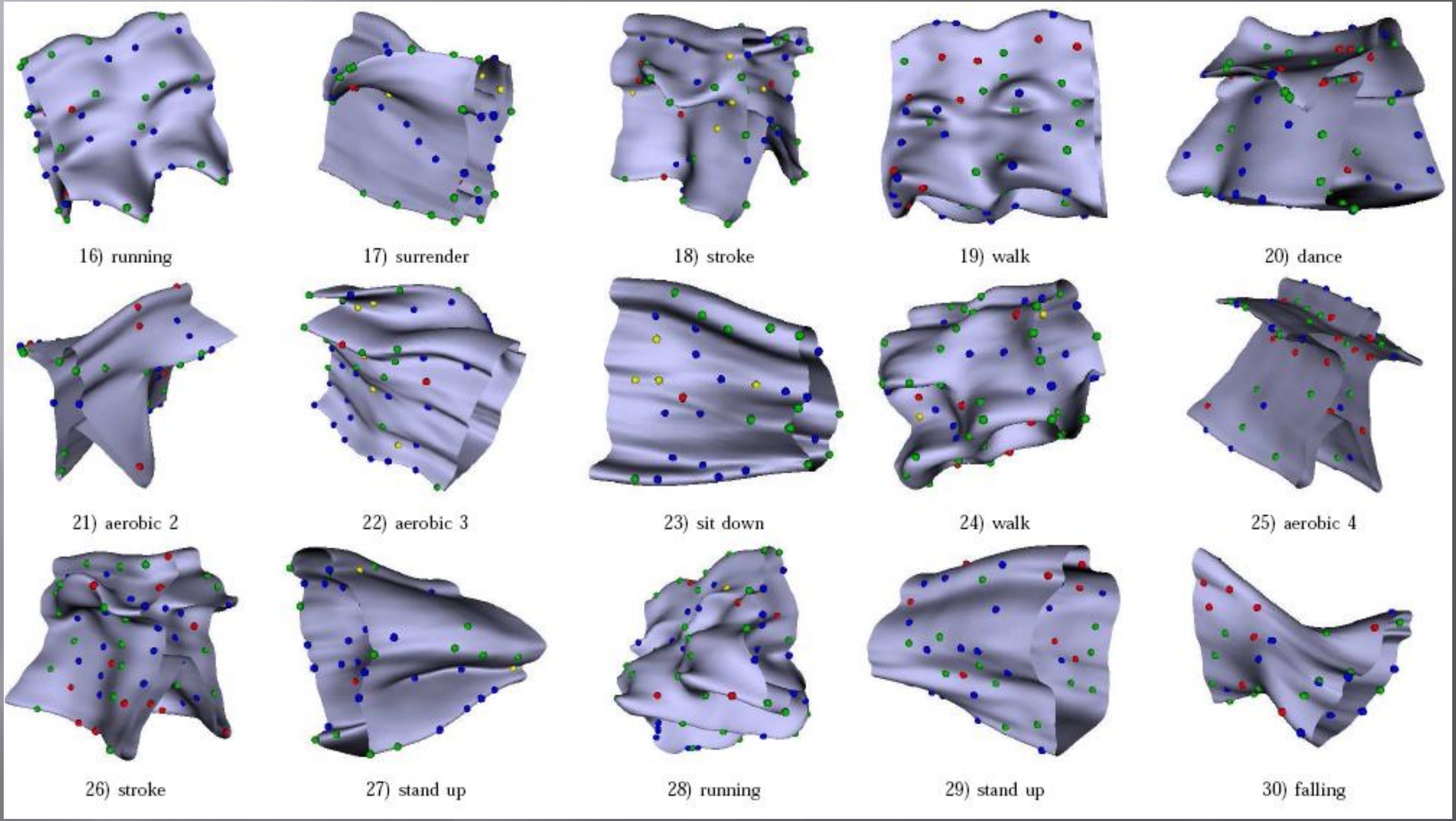
University of Central Florida VISION

Action Volumes



University of Central Florida VISION

Action Volumes



University of Central Florida VISION

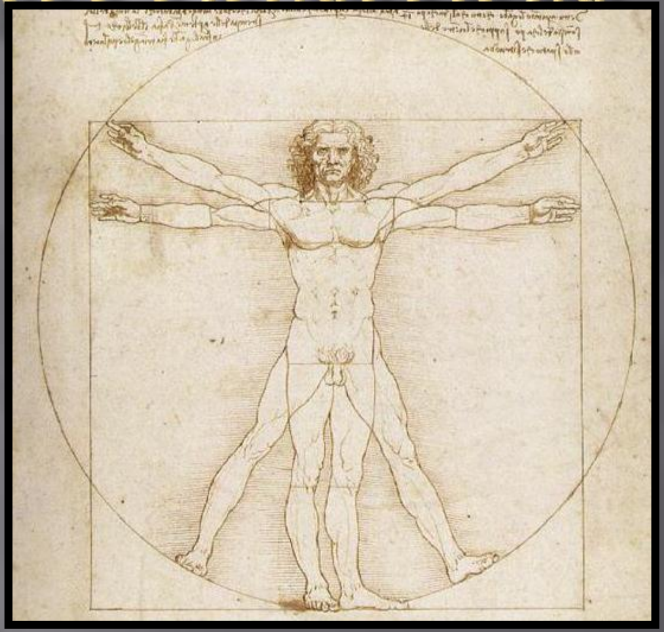
Recognition Results

Input Action	#	Matching action	#
Dance	1	*Dance*	20
Hand down	2	**Stand up**	29
Walking	3	*Walking*	11
Kicking	4	*Kicking*	9
Walking	5	*Walking*	11
Stand up	6	*Stand up*	29
Surrender	7	*Surrender*	17
Hands down	8	*Hands down*	82
Kicking	9	*Kicking*	4
Falling	10	*Falling*	30
Walking	11	*Walking*	11
Walking	12	**Sit down**	23
Sit down	14	*Sit down*	23

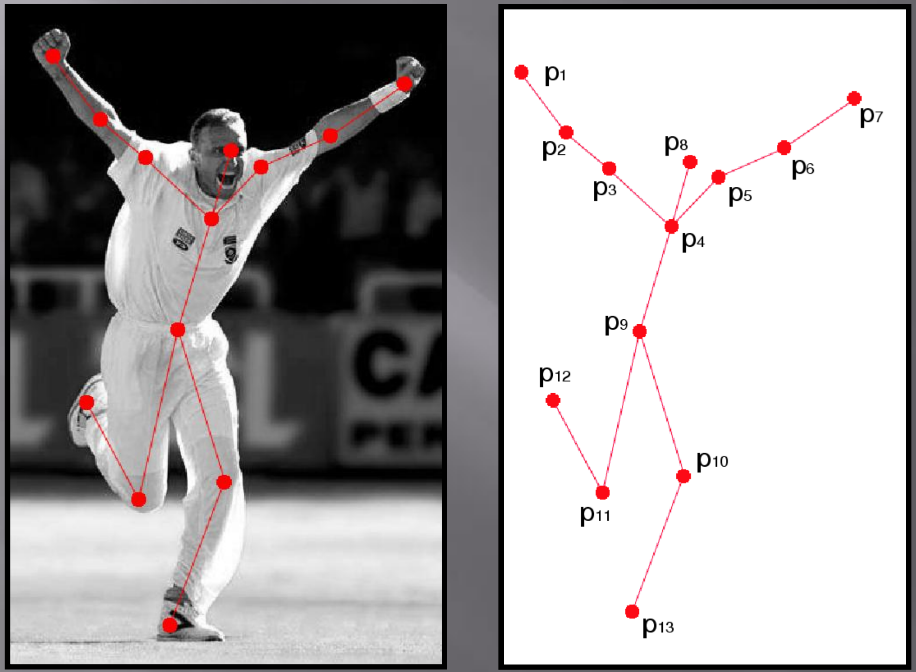
Video	#	Matching action	#
Walking	15	*Walking*	11
Running	16	*Running*	28
Surrender	17	*Surrender*	17
Tennis stroke	18	*Tennis stroke*	26
Walking	19	*Walking*	11
Dance	20	*Dance*	1
Sit down	23	*Sit down*	23
Walking	24	*Walking*	11
Tennis stroke	26	*Tennis stroke*	18
Stand up	27	*Stand up*	29
Running	28	*Running*	16
Stand up	29	**Hands down**	8
Falling	30	*Falling*	10

University of Central Florida VISION

ANTHROPOMETRIC REPRESENTATION FOR INVARIANT ACTION RECOGNITION



Representation of Actors



- Point-based model contains sufficient description for the recognition of human actions, [1].

[1] G. Johanasson. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14(2): 201 – 211, 1993.

Anthropometry

- \An`thro*pom"e*try\, *n.* Measurement of the height and other dimensions of human beings, especially at different ages, or in different races, occupations, etc.

- Variability in human proportion is not *arbitrary*.
- Action Recognition must address this variation.

Pose and Posture

- Posture: The stance an actor has at a time instant
- Pose: The global orientation and position of an actor

Different Poses, Same Posture

Different Postures, Same Pose

University of Central Florida VISION

Anthropometric Constraint

- **Conjecture:** The relationship between points of two actors X and Y in the same posture can be described by a matrix *M*

$$\mathbf{X}_i = M\,\mathbf{Y}_i$$

where i = 1, 2 … n, M is a 4x4 non-singular matrix, $\mathbf{X}_i$ and $\mathbf{Y}_i$ are sets of points describing two actors.

- This transformation simultaneously captures:
  - the different poses
  - difference in size/proportions.

University of Central Florida VISION

Anthropometric Constraint

- This was verified empirically between the 5th percentile woman and 95th percentile man.

- **Mean error of**
  - **227.3 mm before the transformation,**
  - **23.87 mm after the transformation.**

Dimension	Men 5th %ile	Men 50th %ile	Men 95th %ile	Men SD	Women 5th %ile	Women 50th %ile	Women 95th %ile	Women SD
1. Stature	1625	1740	1855	70	1505	1610	1710	62
2. Eye height	1515	1630	1745	69	1405	1505	1610	61
3. Shoulder height	1315	1425	1535	66	1215	1310	1405	58
4. Elbow height	1005	1090	1180	52	930	1005	1085	46
5. Hip height	840	920	1000	50	740	810	885	43
6. Knuckle height	690	755	825	41	660	720	780	36
7. Fingertip height	590	655	720	38	560	625	685	38
8. Sitting height	850	910	965	36	795	850	910	35
9. Sitting eye height	735	790	845	35	685	740	795	33
10. Sitting shoulder height	540	595	645	32	505	555	610	31
11. Sitting elbow height	195	245	295	31	185	235	280	29
12. Thigh thickness	135	160	185	15	125	155	180	17
13. Buttock-knee length	540	595	645	31	520	570	620	30
14. Buttock-popliteal length	440	495	550	32	435	480	530	30
15. Knee height	490	545	595	32	455	500	540	27
16. Popliteal height	395	440	490	29	355	400	445	27
17. Shoulder breadth (bideltoid)	420	465	510	28	355	395	435	24
18. Shoulder breadth (biacromial)	365	400	430	20	325	355	385	18
19. Hip breadth	310	360	405	29	310	370	435	38
20. Chest (bust) depth	215	250	285	22	210	250	295	27
21. Abdominal depth	220	270	325	32	205	255	305	30
22. Shoulder-elbow length	330	365	395	20	300	330	360	17
23. Elbow-fingertip length	440	475	510	21	400	430	460	19
24. Upper limb length	720	780	840	36	655	705	760	32
25. Shoulder-grip length	610	665	715	32	555	600	650	29
26. Head length	180	195	205	8	165	180	190	7
27. Head breadth	145	155	165	6	135	145	150	6
28. Hand length	175	190	205	10	160	175	190	9
29. Hand breadth	80	85	95	5	70	75	85	4
30. Foot length	240	265	285	14	215	235	255	12
31. Foot breadth	85	95	110	6	80	90	100	6
32. Span	1655	1790	1925	83	1490	1605	1725	71
33. Elbow span	865	945	1020	47	780	850	920	43
34. Vertical grip reach (standing)	1925	2060	2190	80	1790	1905	2020	71
35. Vertical grip reach (sitting)	1145	1245	1340	60	1060	1150	1235	53
36. Forward grip reach	720	780	835	34	650	705	755	31

R. Bridger. *Human Performance Engineering: A Guide for system designers,* Prentice Hall, 1982

Postural Constraint

- **Proposition 1:** If x_t and y_t describe the imaged posture of two actors at time t, a Fundamental Matrix can be uniquely associated with ($\mathbf{x}_t$, $\mathbf{y}_t$) if the two actors are in the same posture.

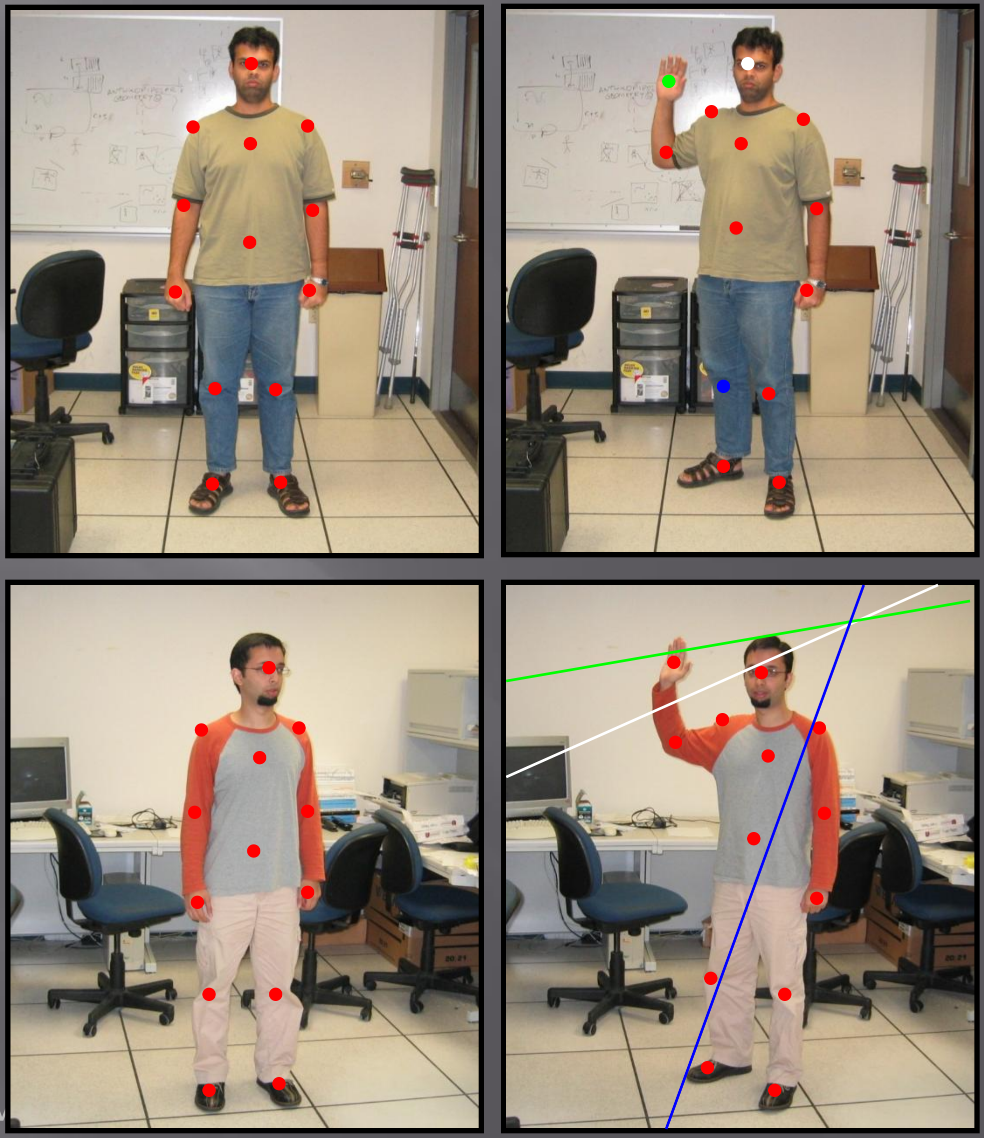
$$x_t^T F y_t = 0$$

- Two actors performing the action instead of two views.
- This is valid for a single time instance.

University of Central Florida VISION

Capturing View Variance

- The fundamental matrix captures the variability in proportion as well as the change in view.



University of Central Florida VISION

Postural Constraint

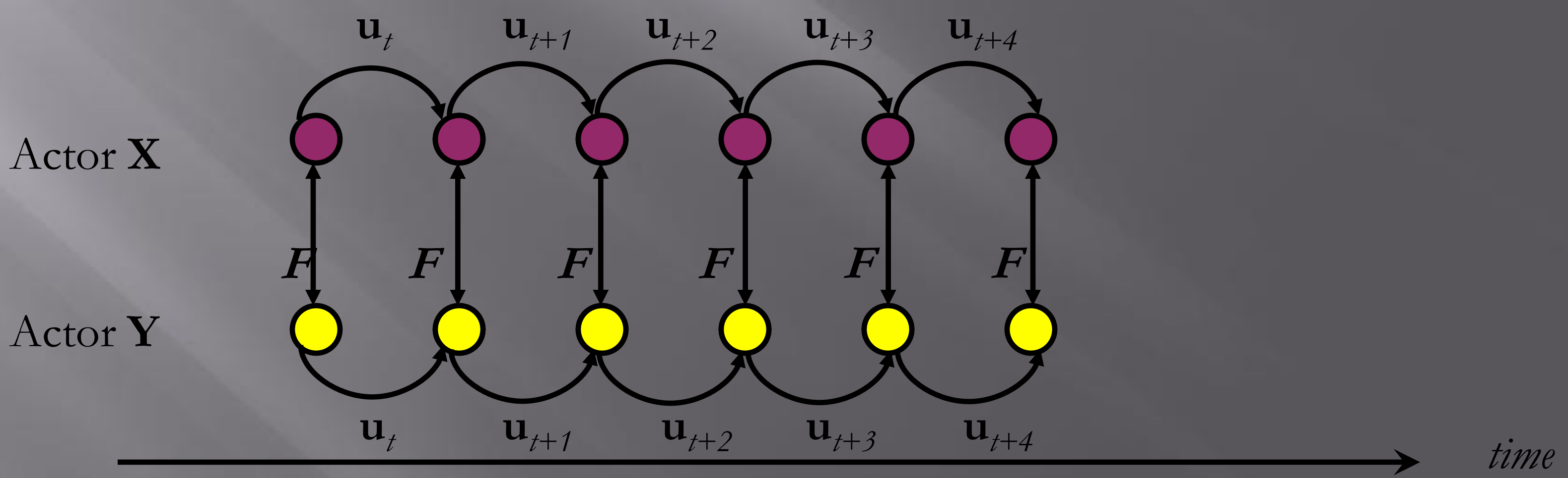
- The similarity of *posture* between two actors can be measured using the **ninth singular** value of a measurement matrix *A*, where $Af = 0$.

$$\begin{bmatrix} x_1'x_1 & \dots & x_n'x_n \\ x_1'y_1 & \dots & x_n'y_n \\ x_1' & \dots & x_n' \\ y_1'x_1 & \dots & y_n'x_n \\ y_1'y_1 & \dots & y_n'y_n \\ y_1' & \dots & y_n' \\ x_1 & \dots & x_n \\ y_1 & \dots & y_n \\ 1 & \dots & 1 \end{bmatrix}^T \begin{bmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{bmatrix} = Af = 0$$

University of Central Florida VISION

Action Constraint

- **Proposition 2:** For an action element $\mathbf{u}_t$, the fundamental matrices associated with (x_t, y_t) and (x_{t+1}, y_{t+1}) are the same if both actors perform the action element defined by $\mathbf{u}_t$.



Measuring Action Similarity

- Since all the *F*s are the same:

$$A_1 f = 0$$
$$A_2 f = 0$$
$$\vdots$$
$$A_k f = 0$$

- Thus the ninth singular value of

$$\boldsymbol{A} = [\, A_1, A_2 \ldots A_k \,]$$

can be used as a view invariant measure.

Experimental Results

- We performed a diverse set of experiments
  - Action Detection
    - Analyzing periodicity
    - Multiple view multiple people
  - Action Synchronization
    - Following the leader
  - Odd one out

Action Detection
Analyzing Periodicity

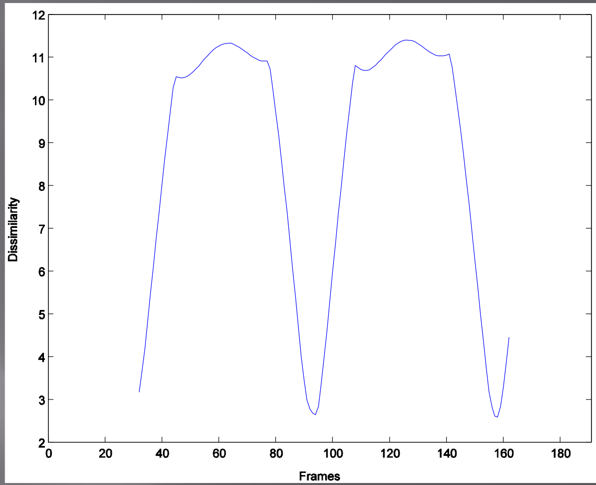
Reference Pattern

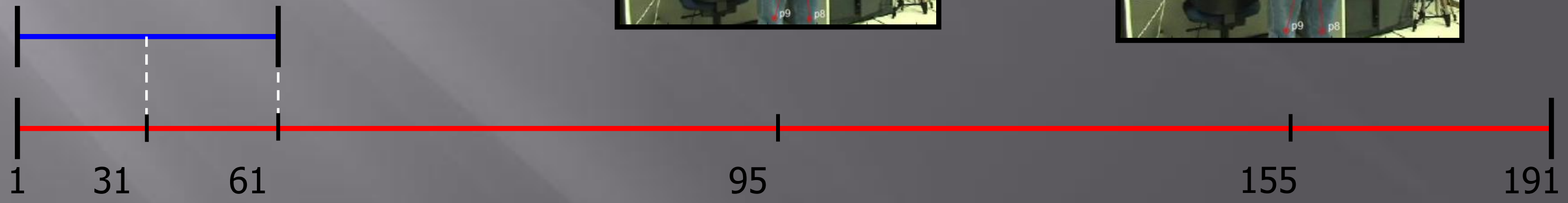
Test Sequence

University of Central Florida VISION

Action Detection

Analyzing Periodicity





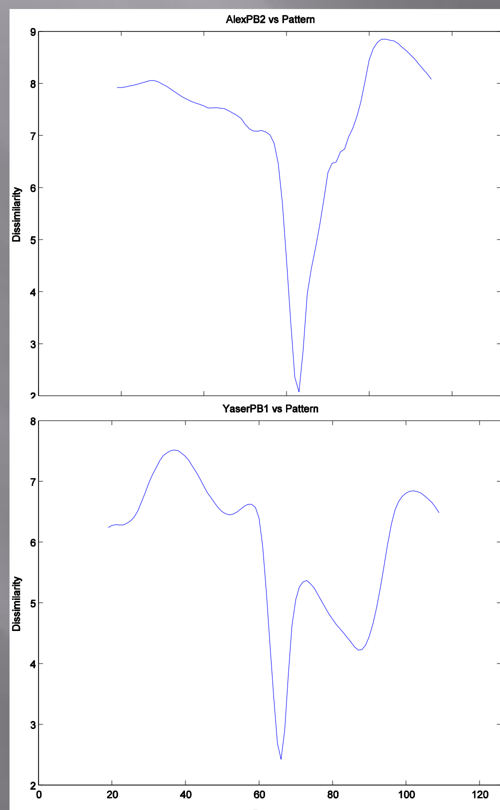
Action Detection: Different approaches, different people, the same action

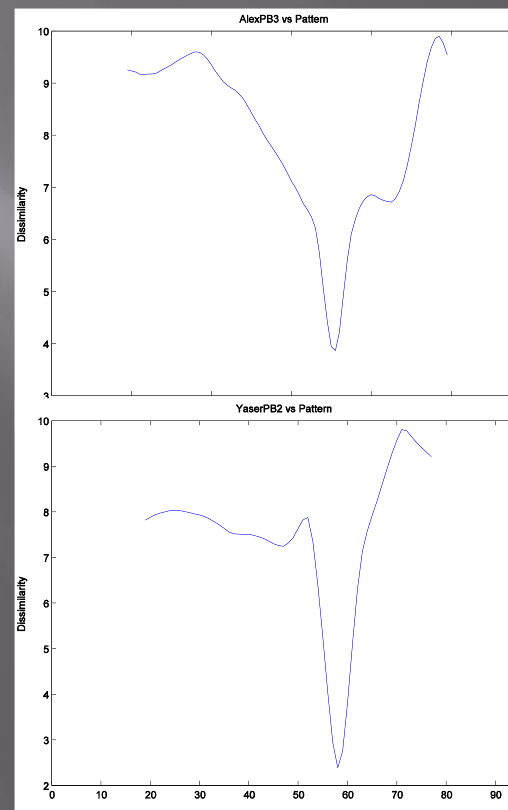
ReferencePattern

Test Sequences

University of Central Florida VISION

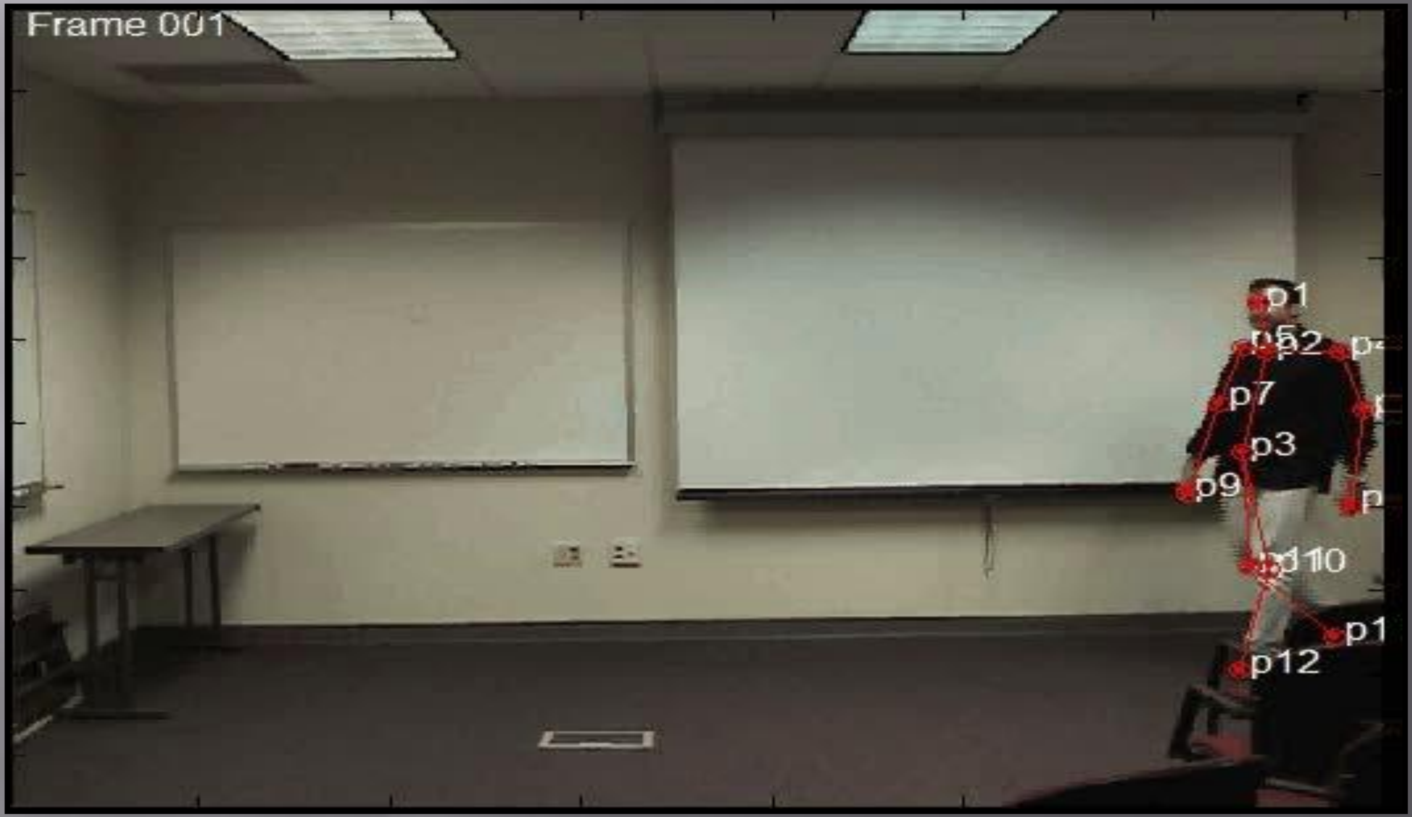
Action Detection: Different approaches, different people, the same action



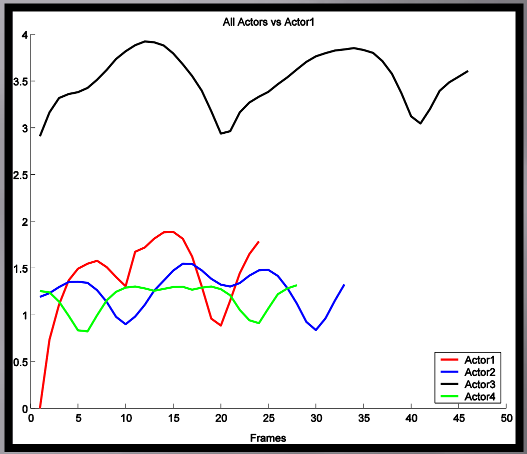


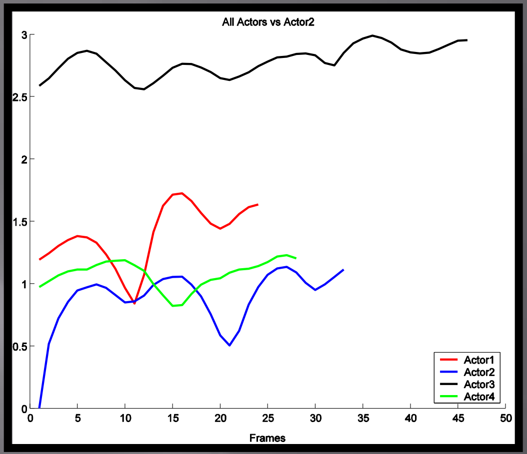
University of Central Florida VISION

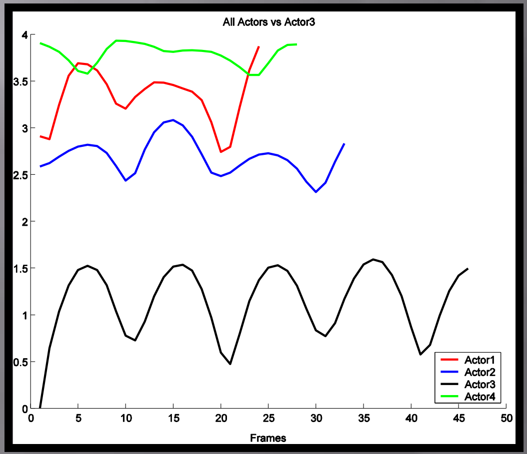
Analyzing Actions
Odd One Out

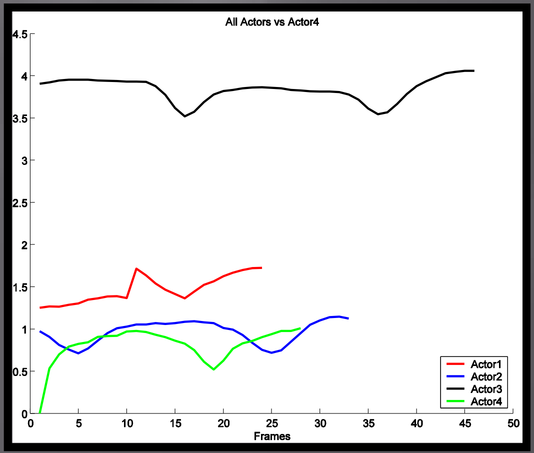


‘Odd One Out’





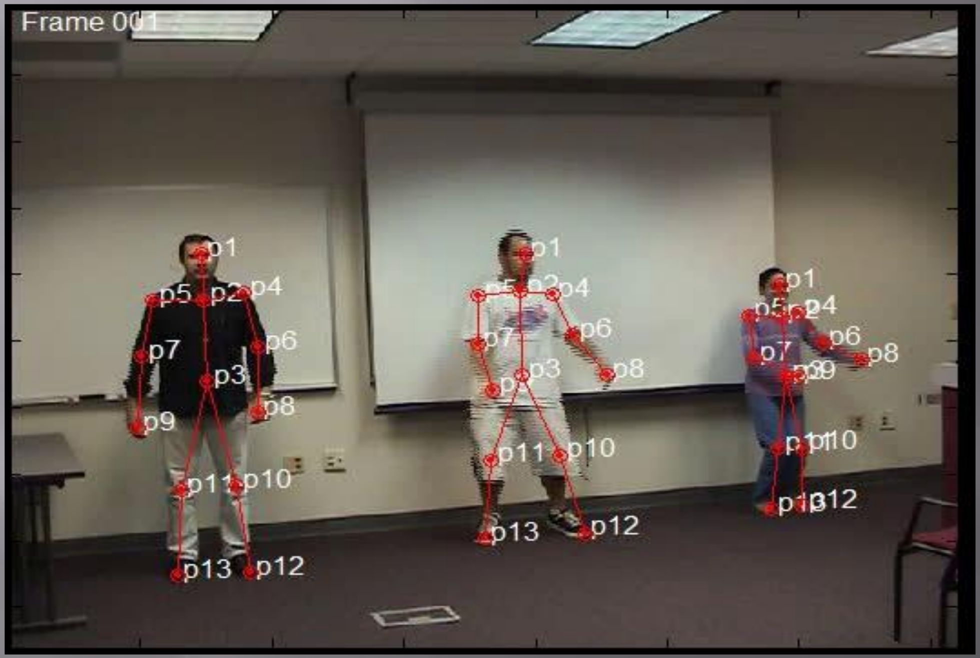




University of Central Florida VISION

Action Synchronization

Following the Leader



University of Central Florida VISION

Action Synchronization
Following the Leader

SPACE-TIME PROJECTION FOR UNIFORMLY MOVING CAMERAS

Yaser Sheikh, Alex Gritai and

Mubarak Shah

CVPR2007

University of Central Florida VISION

Space-time Projection Model

- Notation
  - Cartesian space-time world coordinate $\mathbf{X} = [T \; X \; Y \; Z]^T$
  - Homogeneous space-time world coordinate $\mathbf{U} = [T \; X \; Y \; Z \; 1]^T$
  - Inhomogeneous camera space-time coordinate $\mathbf{x} = [t \; u \; v]^T$
  - Homogeneous camera space-time coordinate $\mathbf{u} = [t \; w \; uw \; vw]^T$
- Projection
  - $\mathbf{x} = \left(\alpha_t T, \; fX + p_u, \; fY + p_v\right)^T$
    - f is the focal length
    - α_t is the reciprocal of the frame-rate of the camera
    - (p_u, p_v) are the principal point offset

Fundamental Constraint Between Galilean Cameras

$$(tu'\ tv'\ t\ u'\ v'\ 1)\,\Gamma\,(tu\ tv\ t\ u\ v\ 1)^T = 0$$

$$\begin{pmatrix} tu' \\ tv' \\ t \\ u' \\ v' \\ 1 \end{pmatrix}^T \begin{pmatrix} 0 & 0 & 0 & f_1 & f_2 & f_3 \\ 0 & 0 & 0 & f_4 & f_5 & f_6 \\ 0 & 0 & 0 & f_7 & f_8 & f_9 \\ f_{10} & f_{11} & f_{12} & f_{13} & f_{14} & f_{15} \\ f_{16} & f_{17} & f_{18} & f_{19} & f_{20} & f_{21} \\ f_{22} & f_{23} & f_{24} & f_{25} & f_{26} & f_{27} \end{pmatrix} \begin{pmatrix} tu \\ tv \\ t \\ u \\ v \\ 1 \end{pmatrix} = 0$$

$$\Gamma = \begin{pmatrix} 0 & \Delta F^T \\ \Delta F & F_{00} \end{pmatrix}$$

University of Central Florida VISION

Epipolar Geometry

- Epipolar surface in one camera corresponding to a point in the other camera is defined by the

$$(\ldots)(\ldots)^T = 0$$

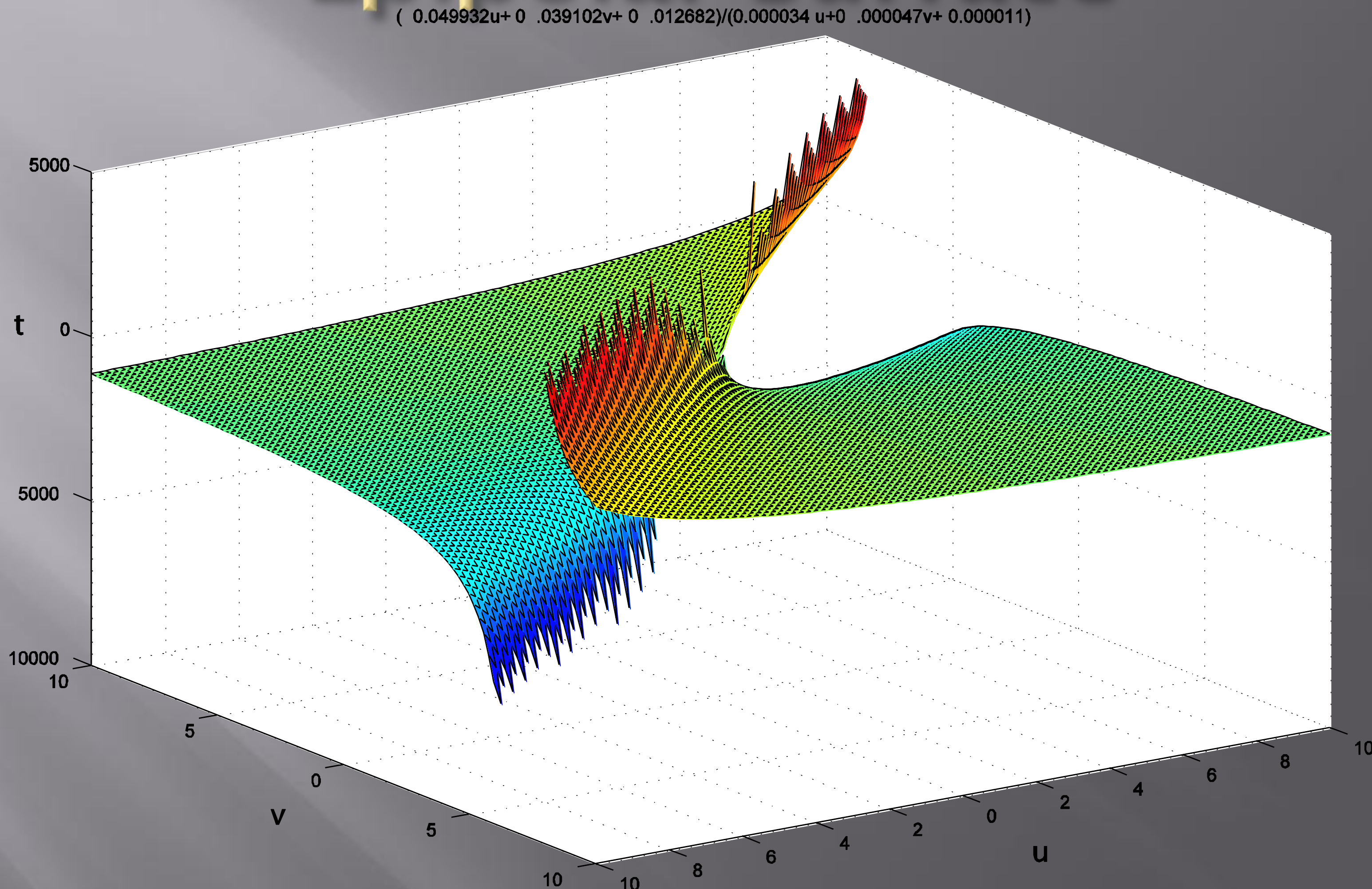
$$s_1 \ldots + s_2 \ldots + s_3 \ldots + s_4 \ldots + s_5 \ldots + s_6 = 0$$

where $\mathbf{s}=[s_1,\ldots,s_6]$ is computed from given

$$\mathbf{s}[\ldots]$$ as

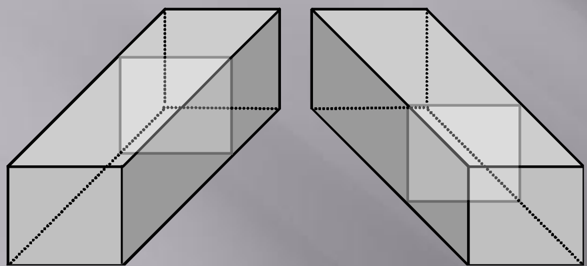
$$(\ldots)$$

Epipolar Surface



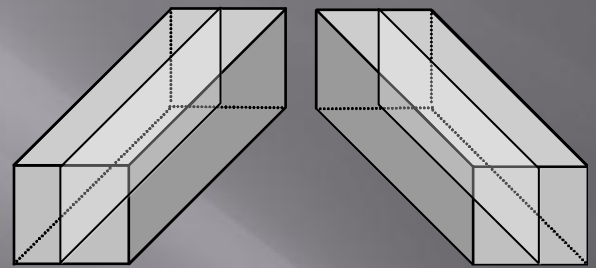
Specializations

Perspective images



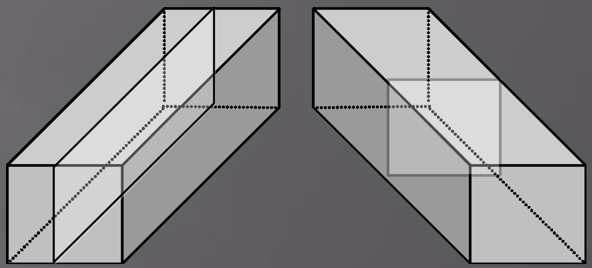
Hartley and Faugeras

Pushbroom images



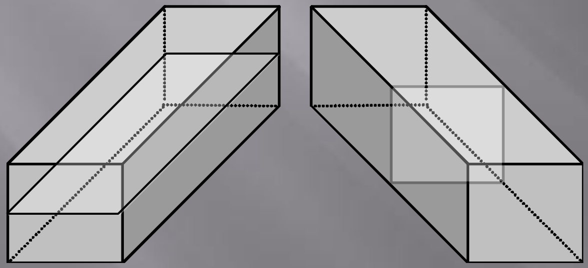
Gupta and Hartley

Pushbroom and Perspective images

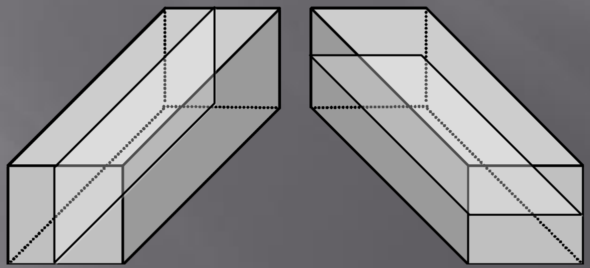


Khan, Rafi and Shah

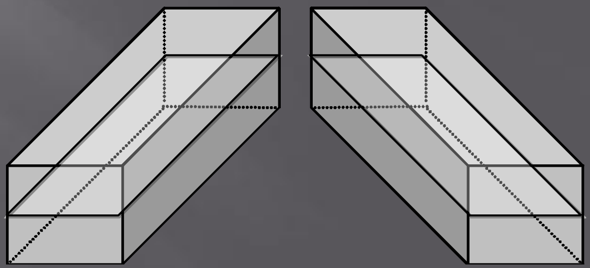
EPI and Perspective images



Pushbroom and EPI images



EPI images



University of Central Florida VISION

APPLICATION

Action Recognition In Two Moving Cameras

Application to Action Recognition (two moving cameras)

- Corresponding points in two videos should satisfy:

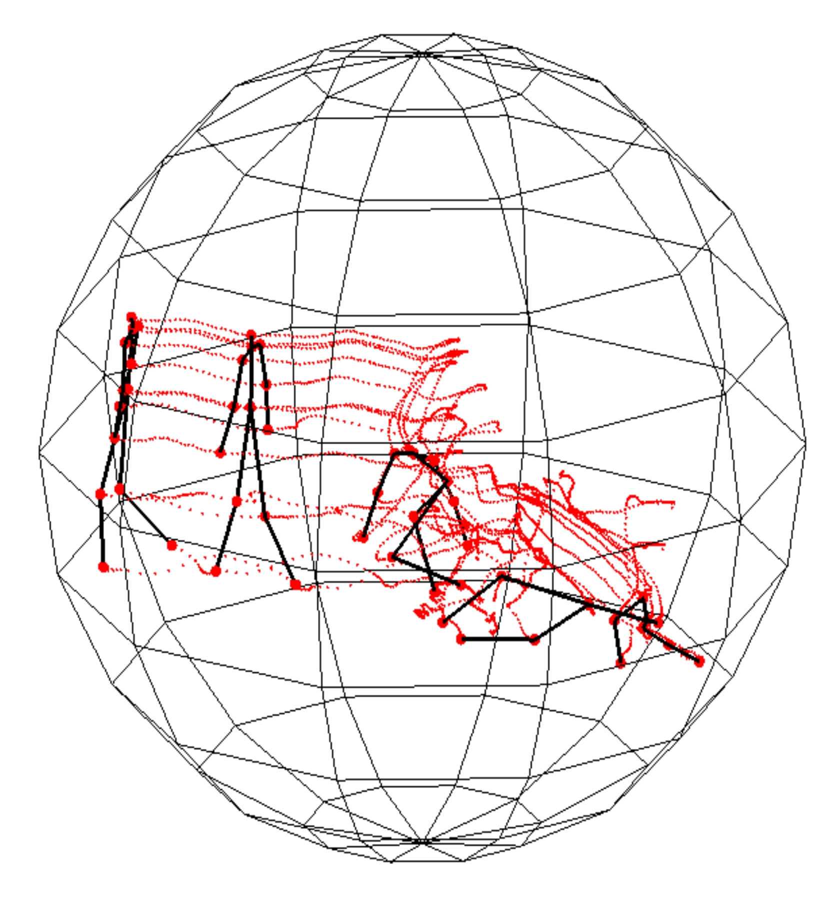
$$(x'_i, y'_i, 1)\, F\, (x_i, y_i, 1)^T = 0$$

$$\begin{bmatrix} x'_1x_1 & \ldots & x'_nx_n \\ x'_1y_1 & \ldots & x'_ny_n \\ x'_1 & \ldots & x'_n \\ y'_1x_1 & \ldots & y'_nx_n \\ y'_1y_1 & \ldots & y'_ny_n \\ y'_1 & \ldots & y'_n \\ x_1 & \ldots & x_n \\ y_1 & \ldots & y_n \\ 1 & \ldots & 1 \end{bmatrix}^T \begin{bmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{bmatrix} = Af^T = 0$$

$$\kappa = \frac{\sigma_{27}}{\sigma_1}$$

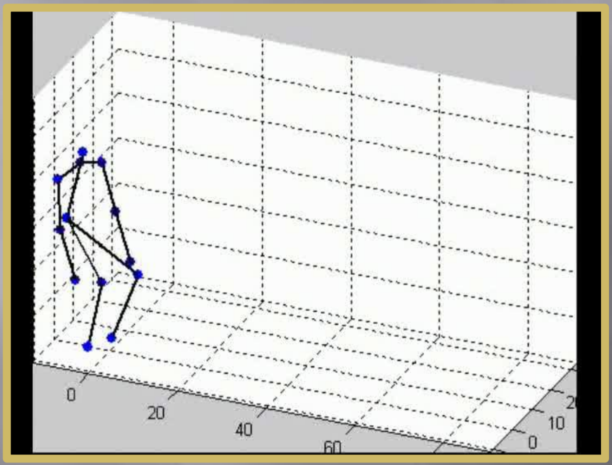
Experimental Results based on Motion Capture Data

- Types of Transformations
  - Viewpoint
  - Anthropometric
  - Time
  - All three together

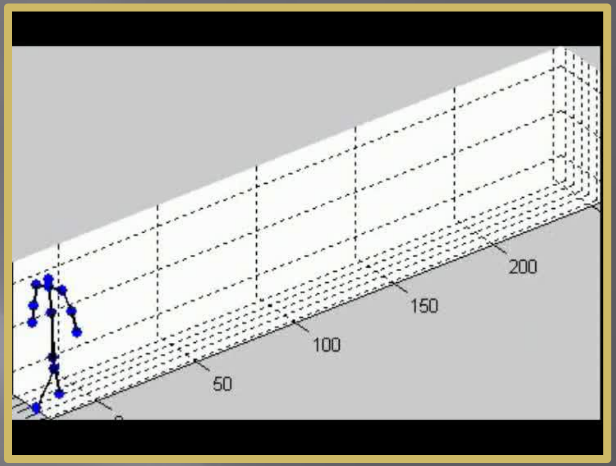


Copyrights Mubarak Shah, UCF

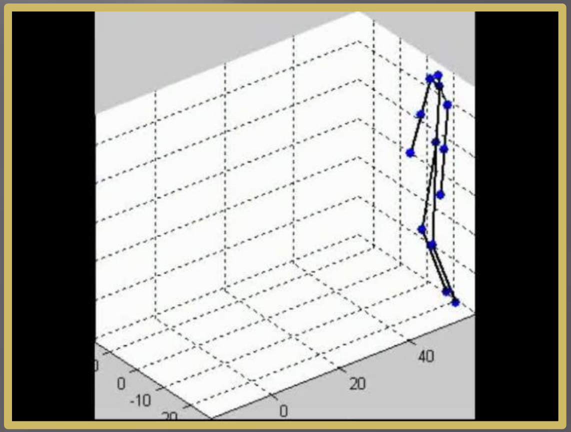
Robustness to View Point Transformations: Different Actions



Standing up



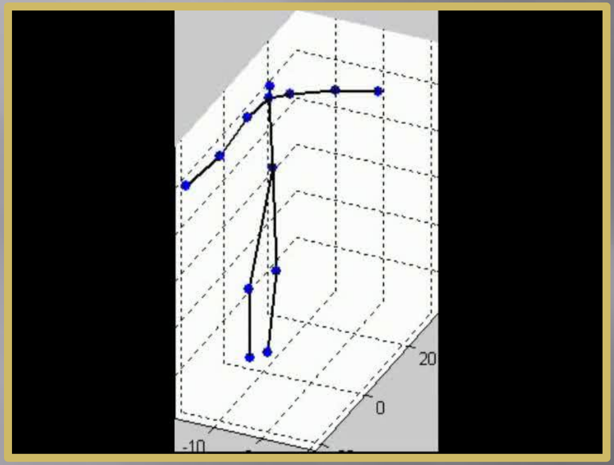
walking

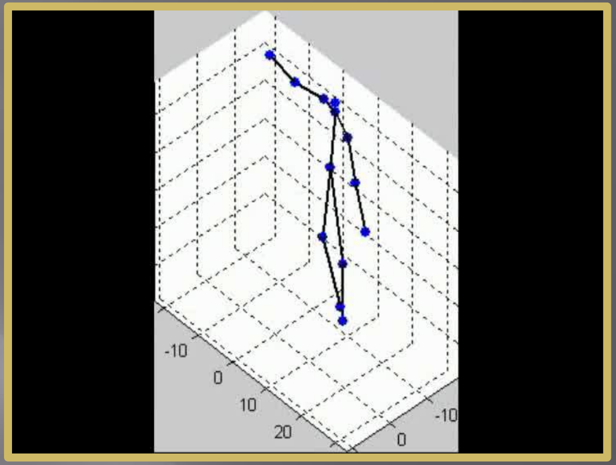


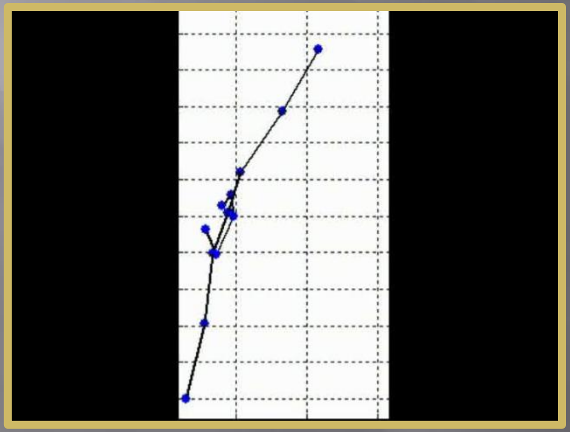
sitting

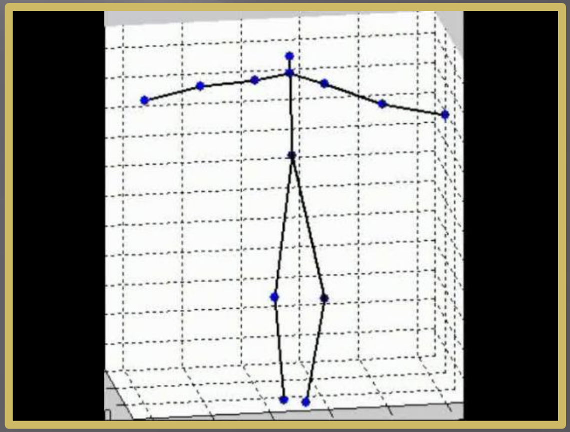
University of Central Florida VISION

Robustness to View Point Transformations: Same Action



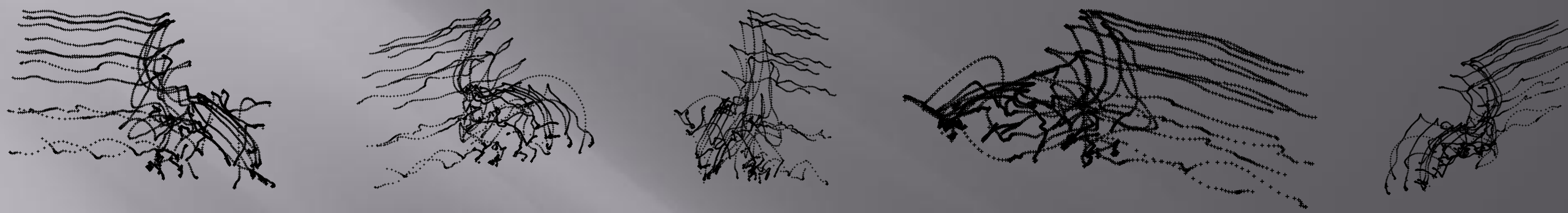




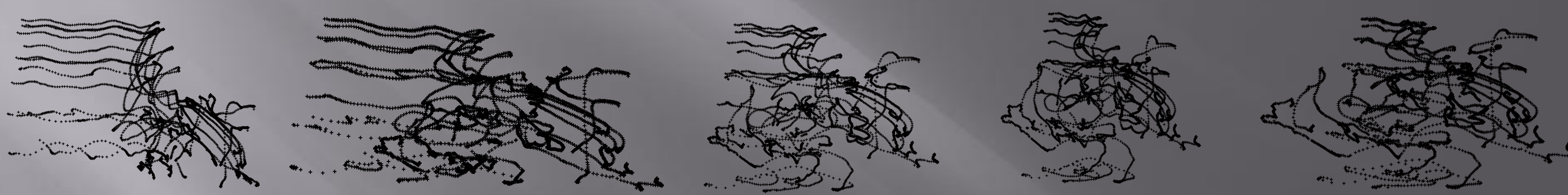


University of Central Florida VISION

Human Actions Under Viewpoint, Anthropometric and Camera Motion Transformations (getting up)



Viewpoint

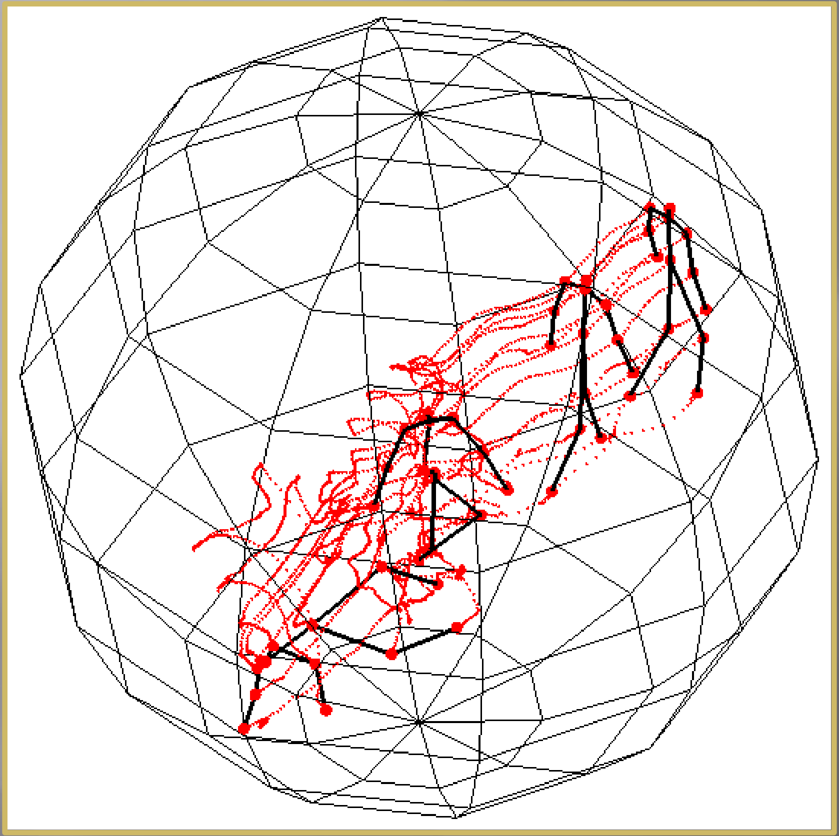


Anthropometric

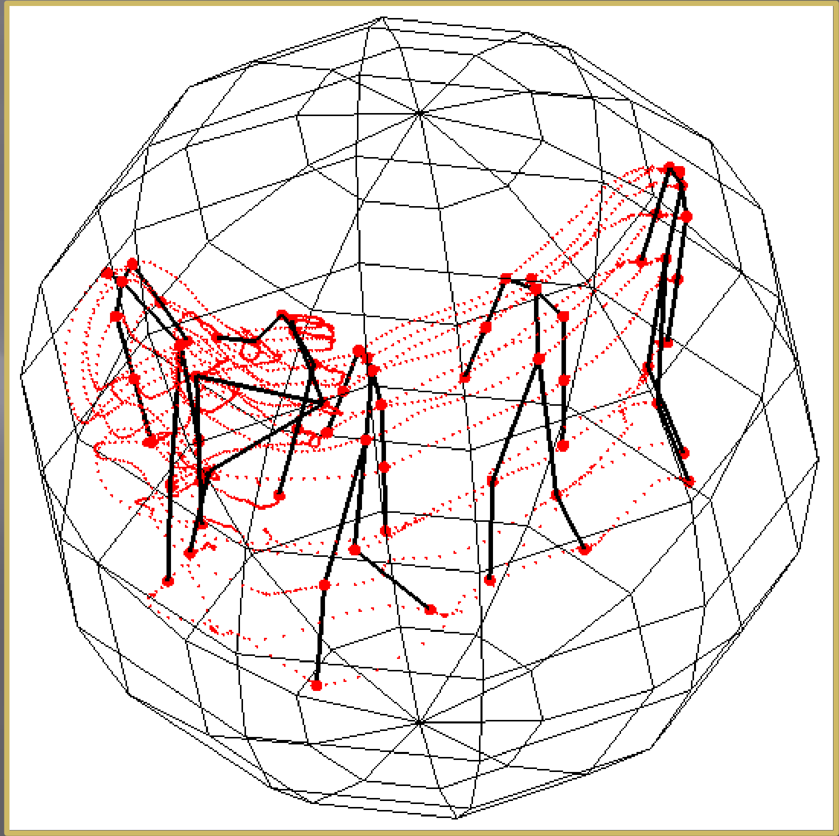
Camera Motion

University of Central Florida VISION

Viewing Sphere (Azimuth 0-350, Elevation 0-90 (interval of 10))



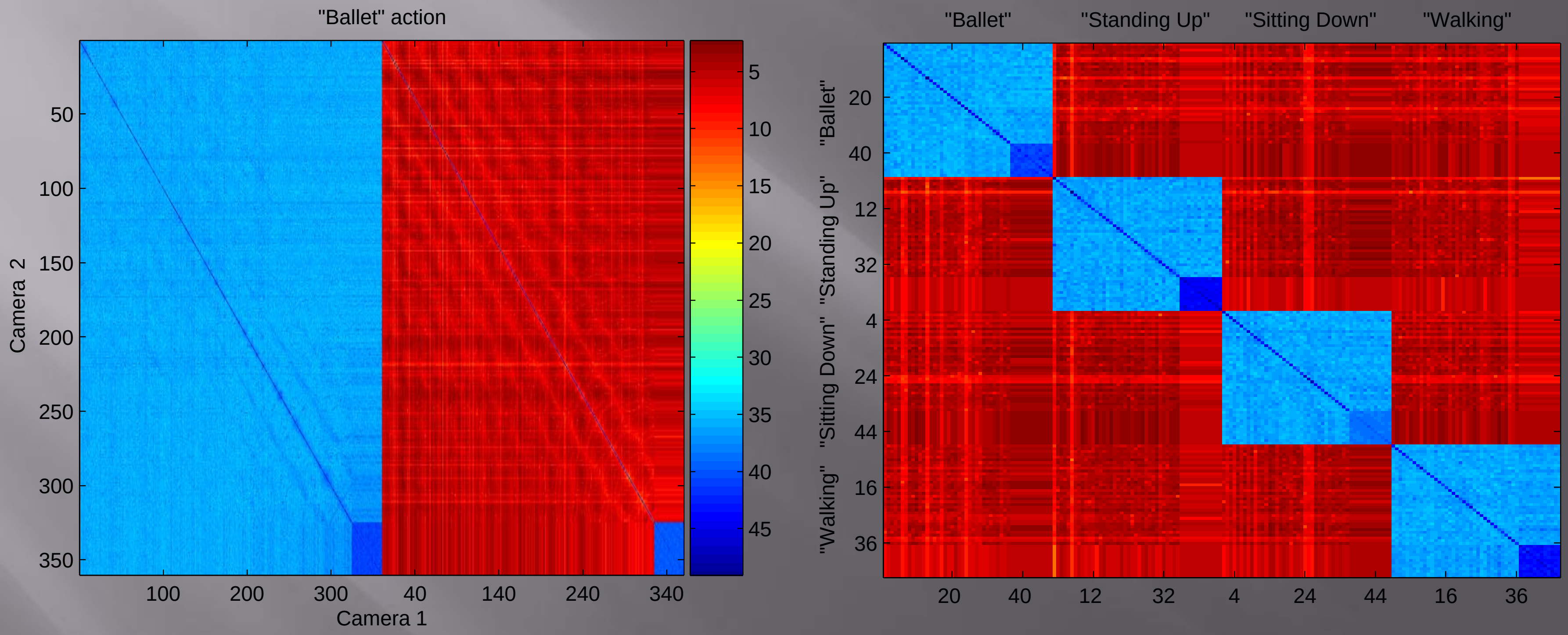
Getting Up



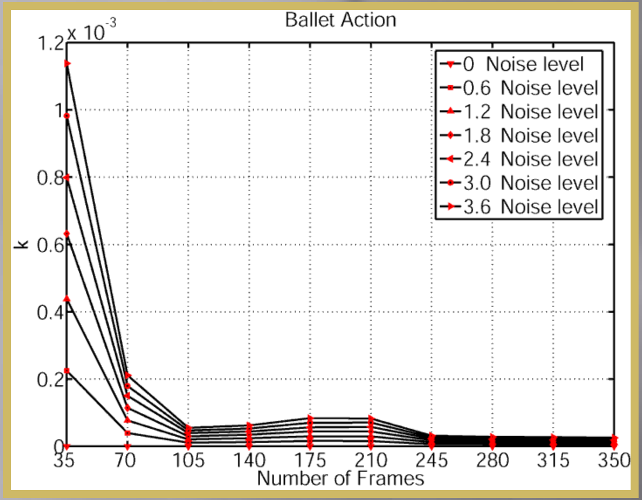
Sitting Down

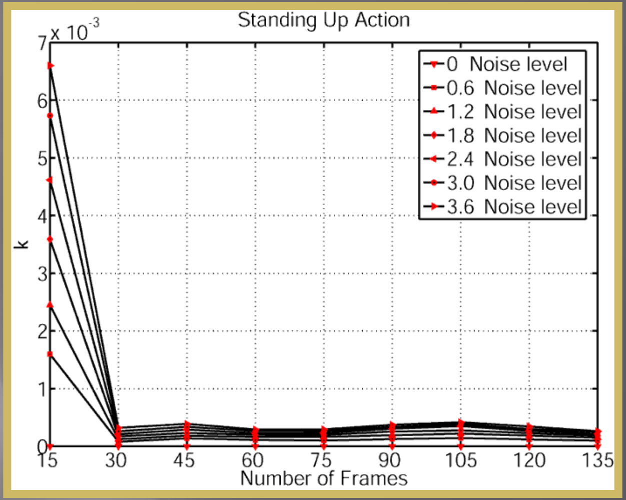
University of Central Florida VISION

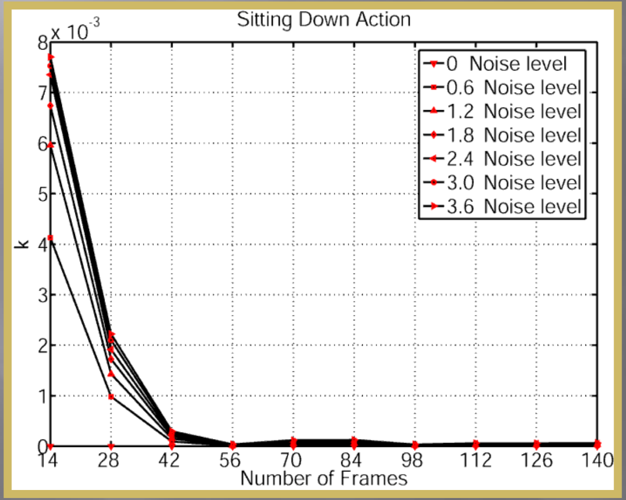
Robustness to View Point Transformations

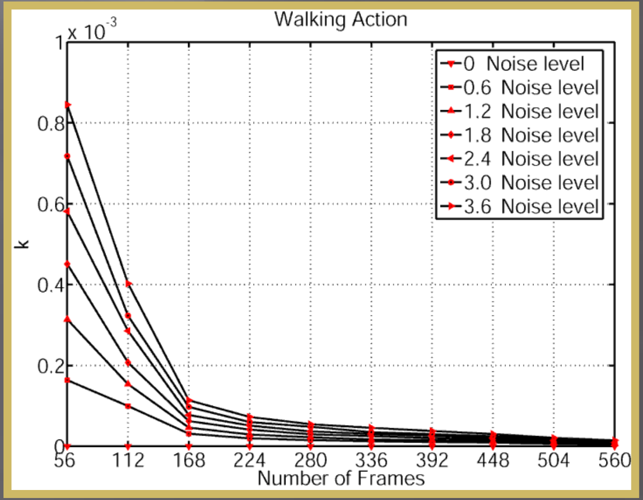


Viewpoint Transformations



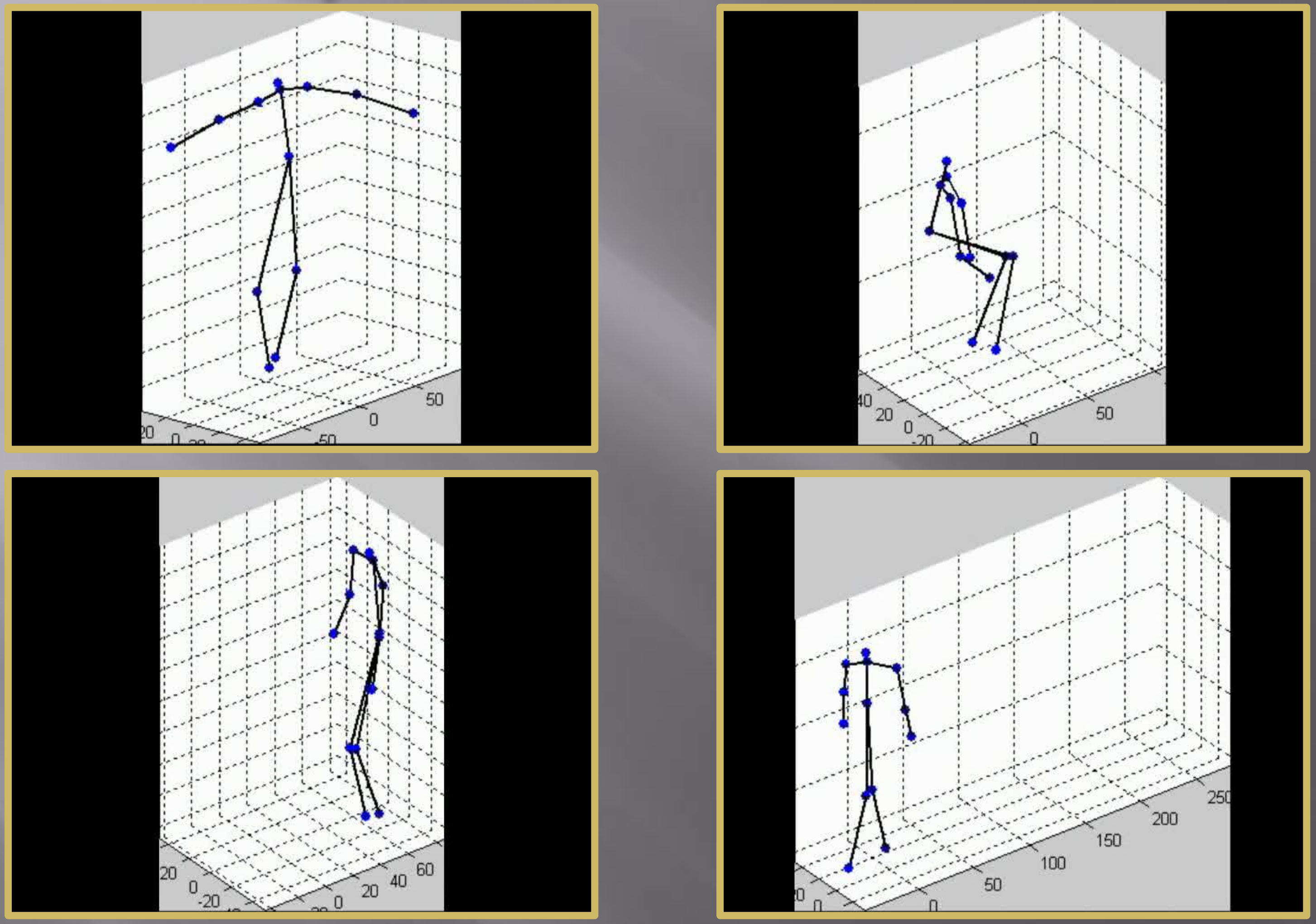




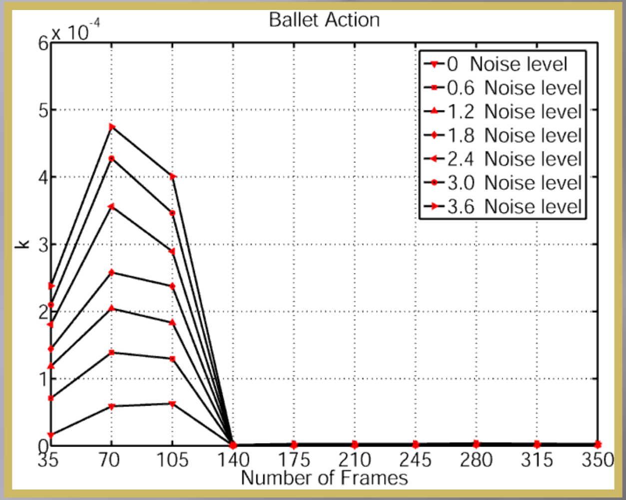


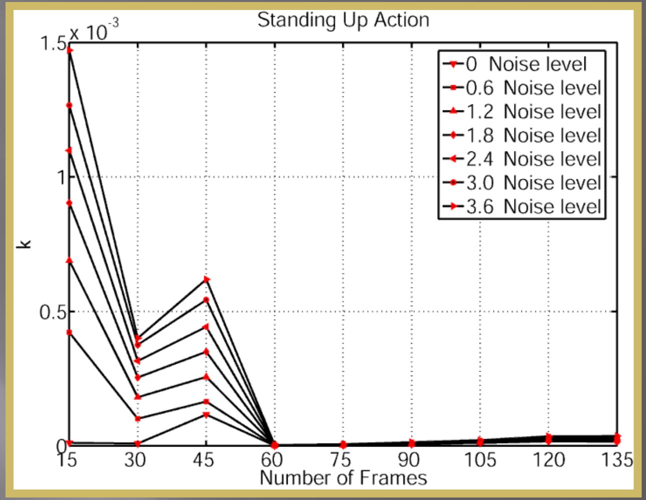
University of Central Florida VISION

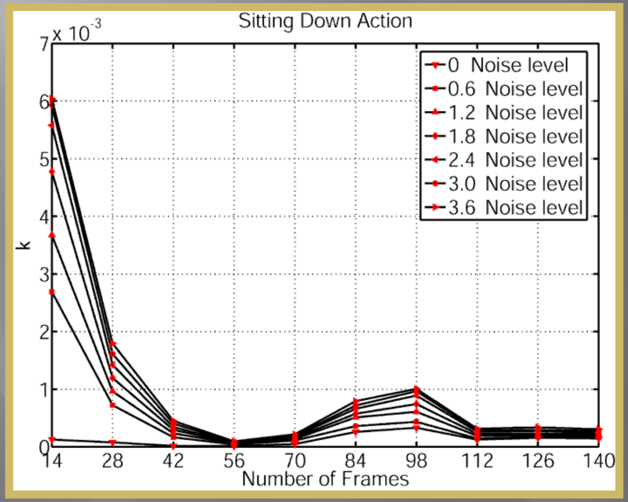
Anthropometric Transformations

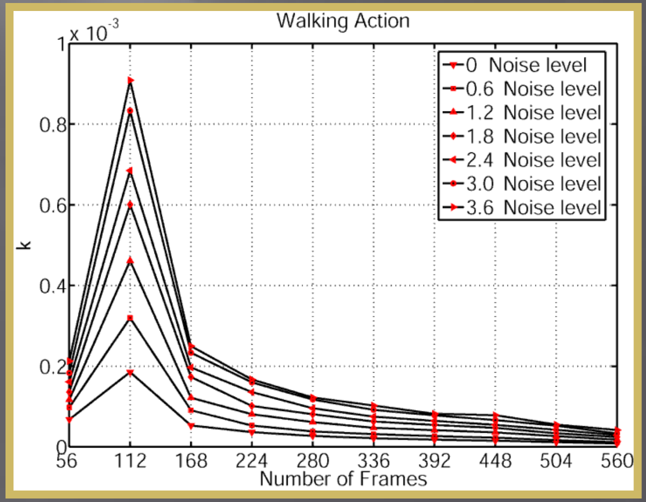


Anthropometric Transformations



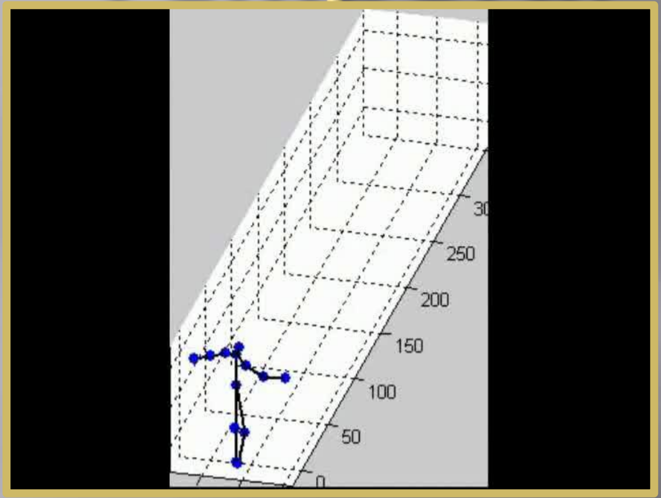


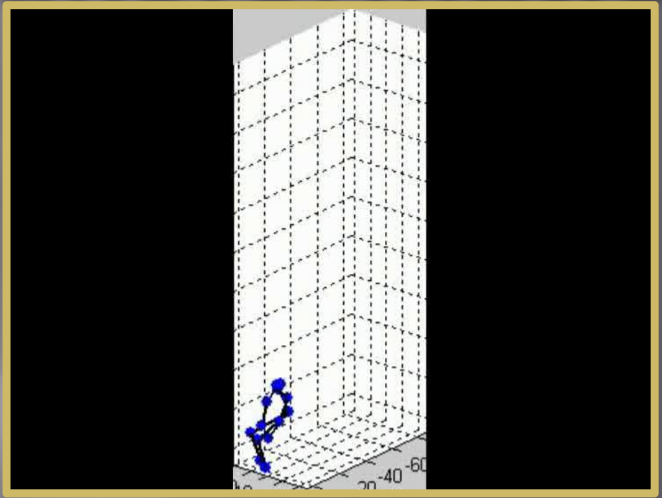


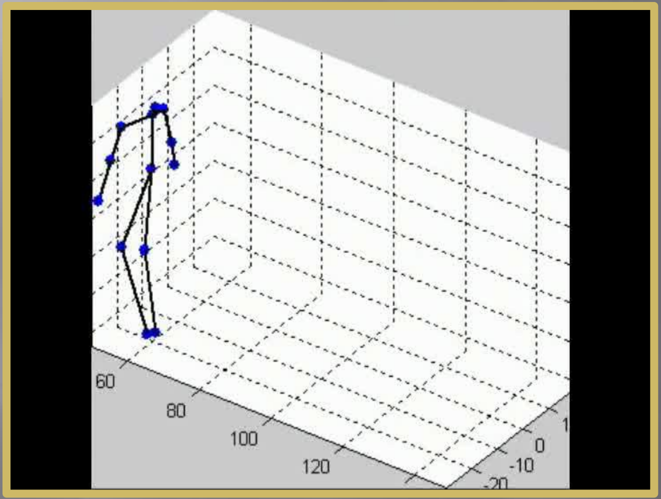


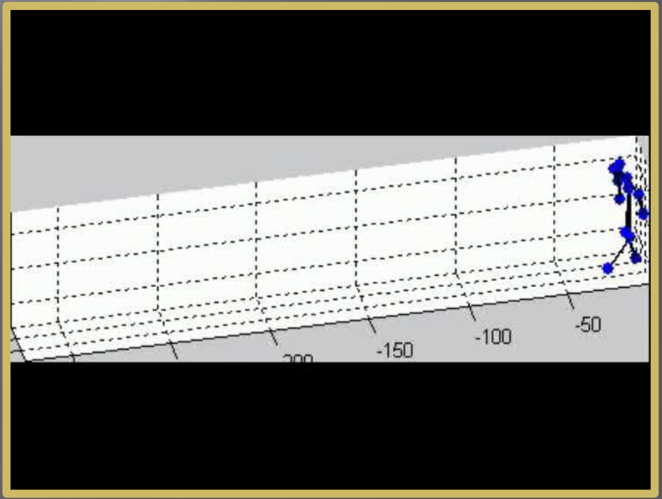
University of Central Florida VISION

Temporal Transformations



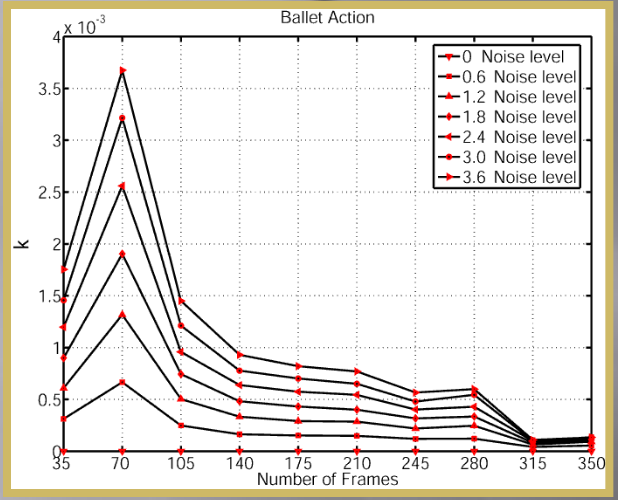


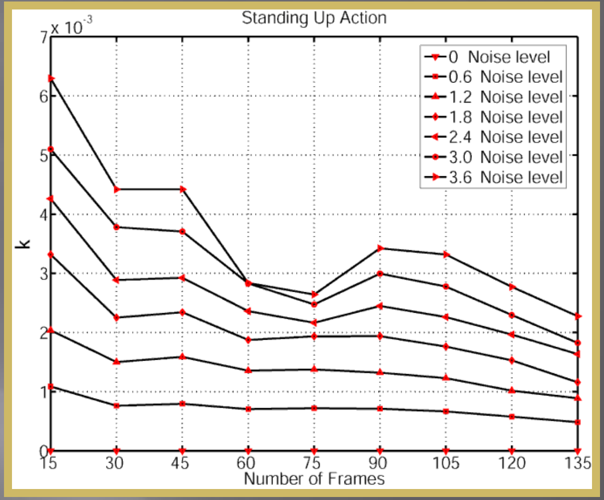


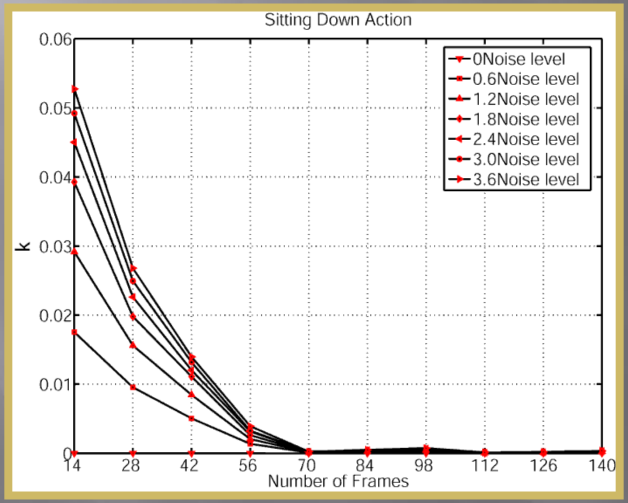


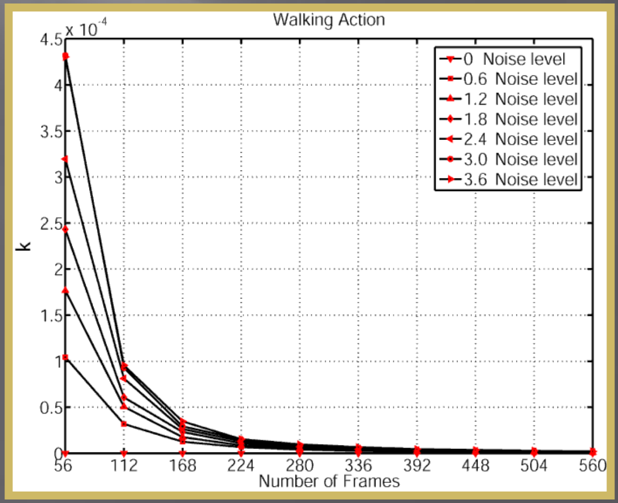
University of Central Florida VISION

Temporal Transformations



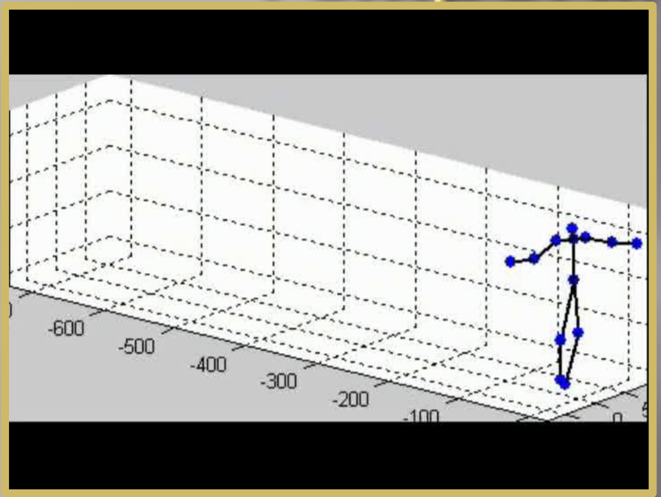


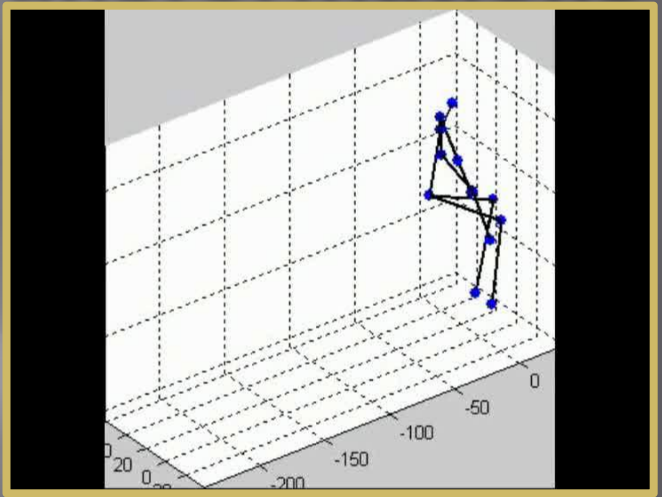


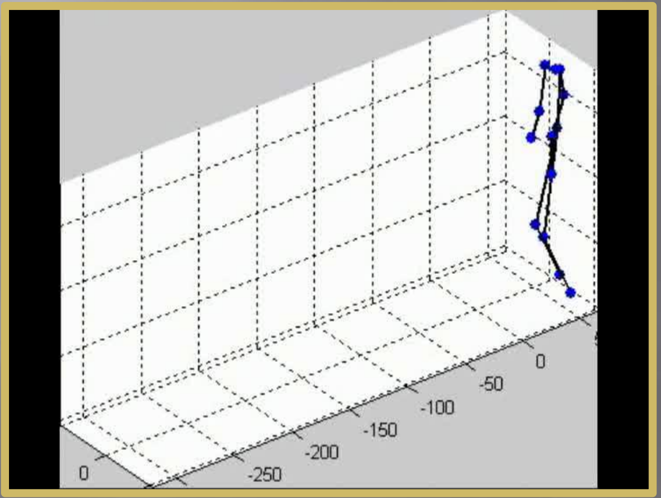


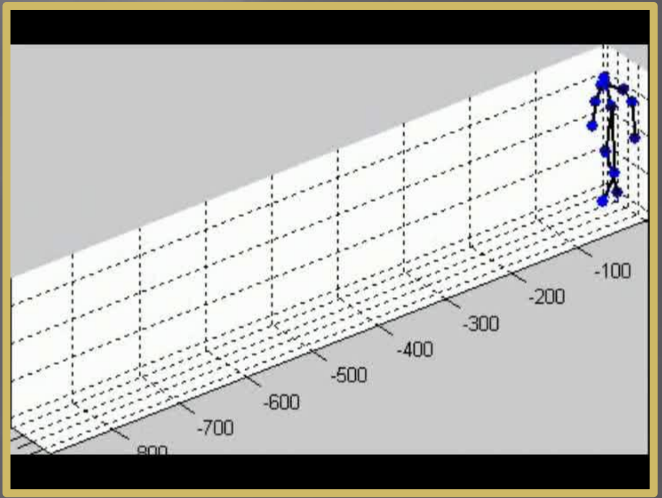
University of Central Florida VISION

Viewpoint, Anthropometric and Temporal Transformations

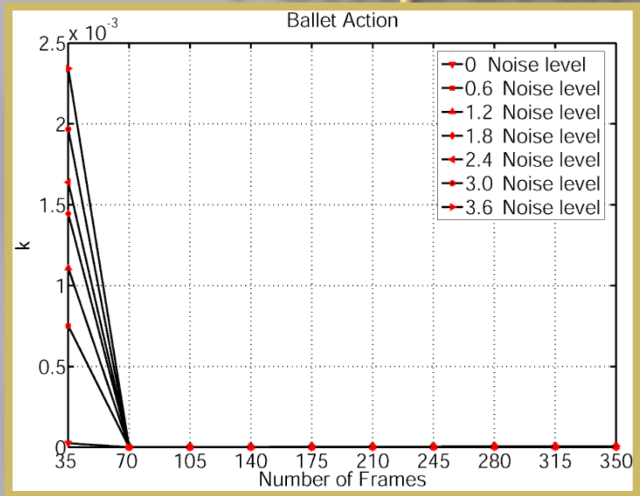


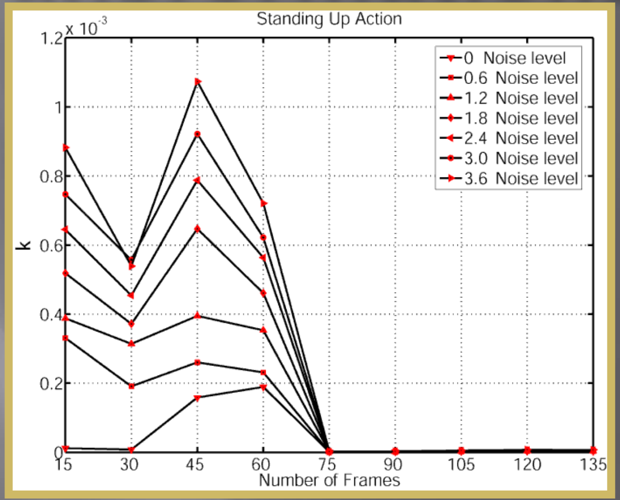


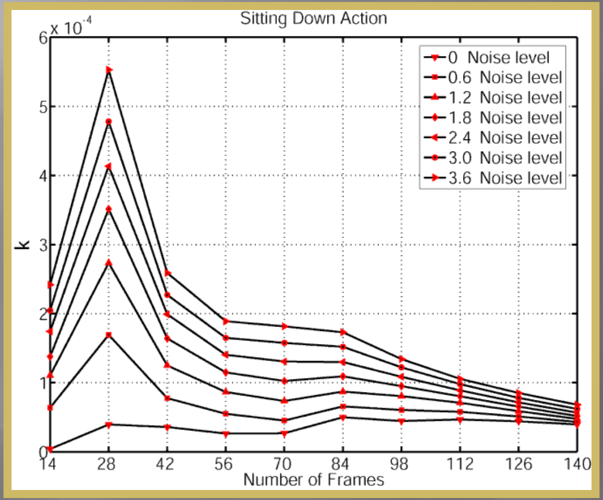


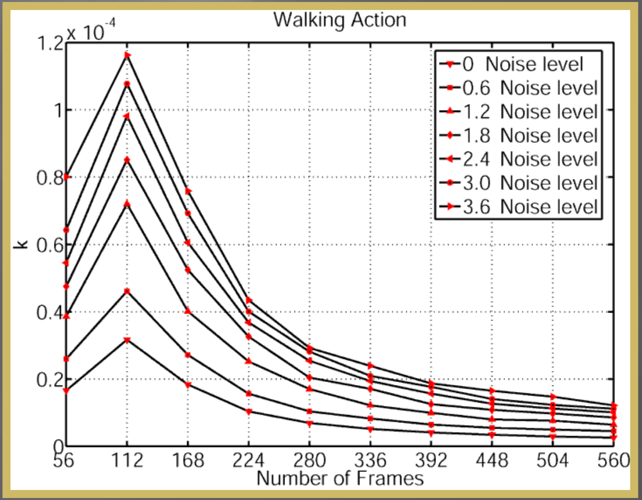


Viewpoint, Anthropometric and Temporal Transformations









CHAOTIC INVARIANTS FOR HUMAN ACTION RECOGNITION

Saad Ali, Arslan Basharat, Mubarak Shah

ICCV 2007

University of Central Florida VISION

Proposed Idea

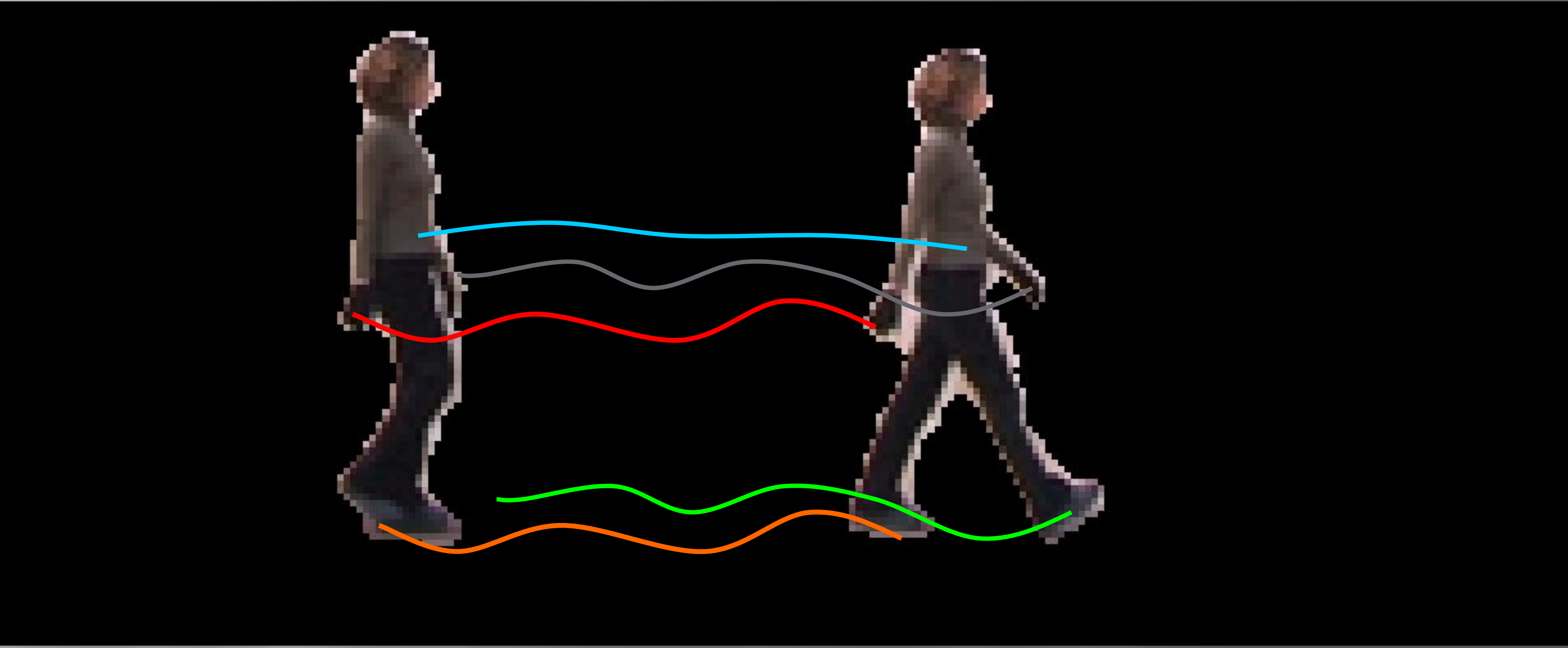
$f(\theta_1^t, \theta_2^t, \ldots, \theta_N^t)$ $f(\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_N^{t+1})$ $f(\theta_1^{t+2}, \theta_2^{t+2}, \ldots, \theta_N^{t+2})$ $f(\theta_1^{t+3}, \theta_2^{t+3}, \ldots, \theta_N^{t+3})$

f $(\theta_1^t, \theta_2^t, \ldots, \theta_N^t)$

Rule Unknown State Space Variables

We have the access to the data generated by the dynamical system controlling this action !

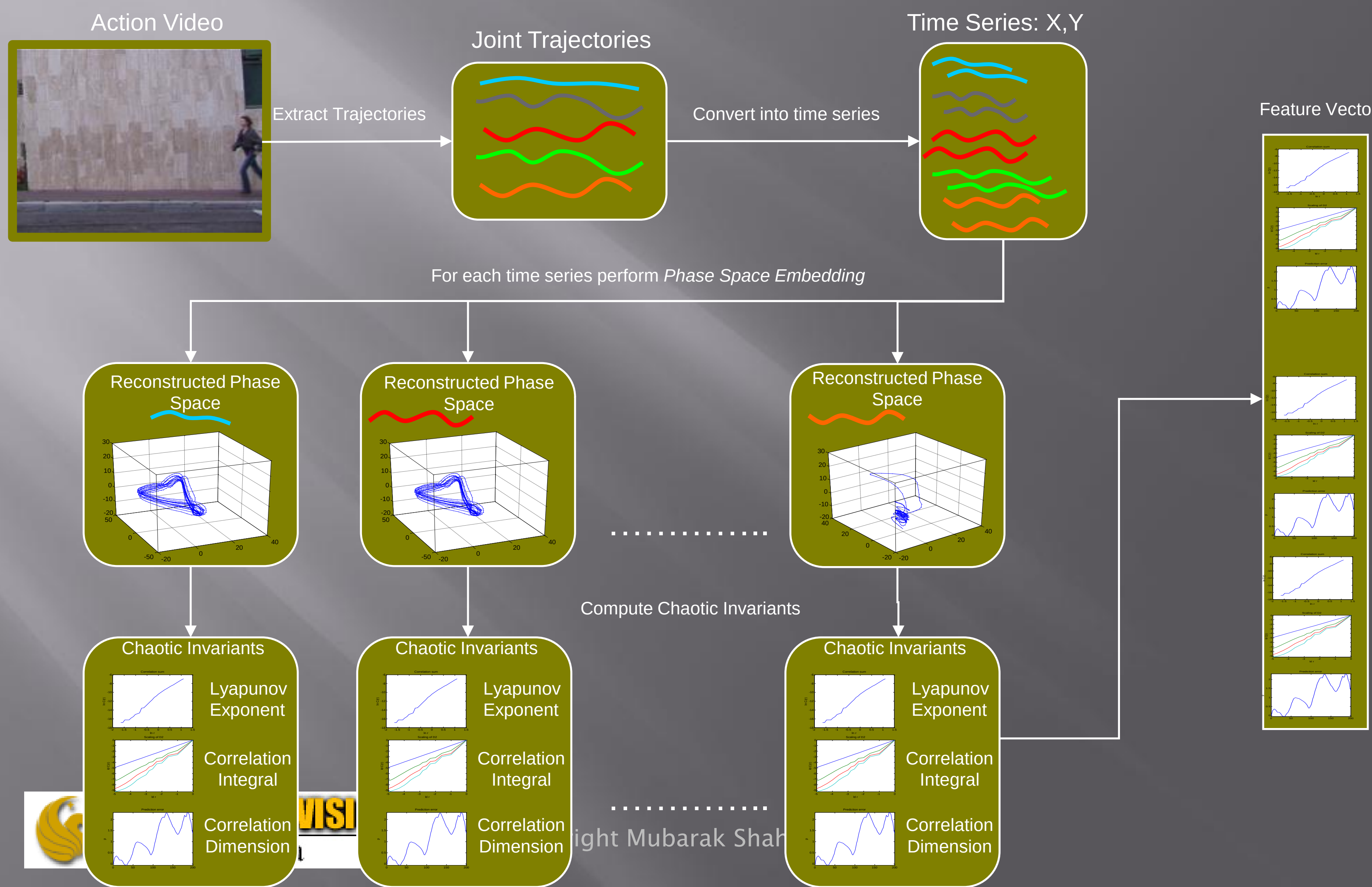
Proposed Idea



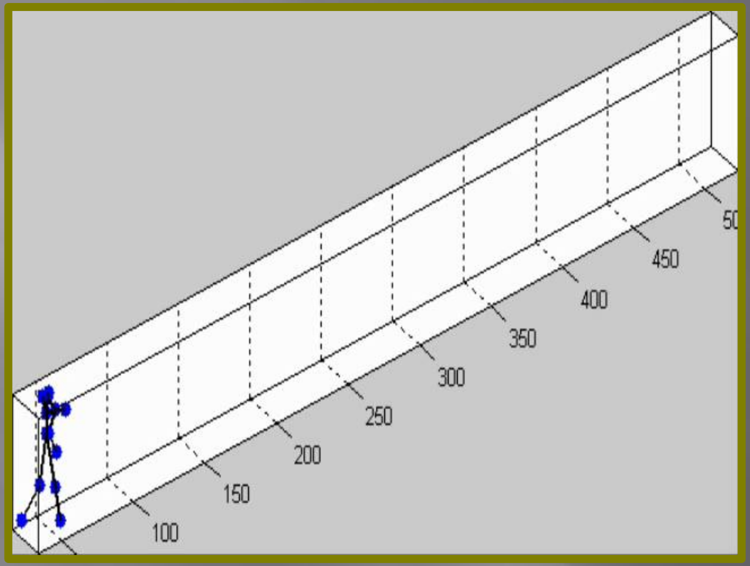
- Experimental Data: Trajectories of body joints
- From this data construct the phase space (and get periodic strange attractors) corresponding to the dynamical system responsible for generating the data.
- That is: Let the data speak to you and tell you what mechanisms are generating chaotic behavior.

University of Central Florida VISION

Algorithmic Overview



Action Representation

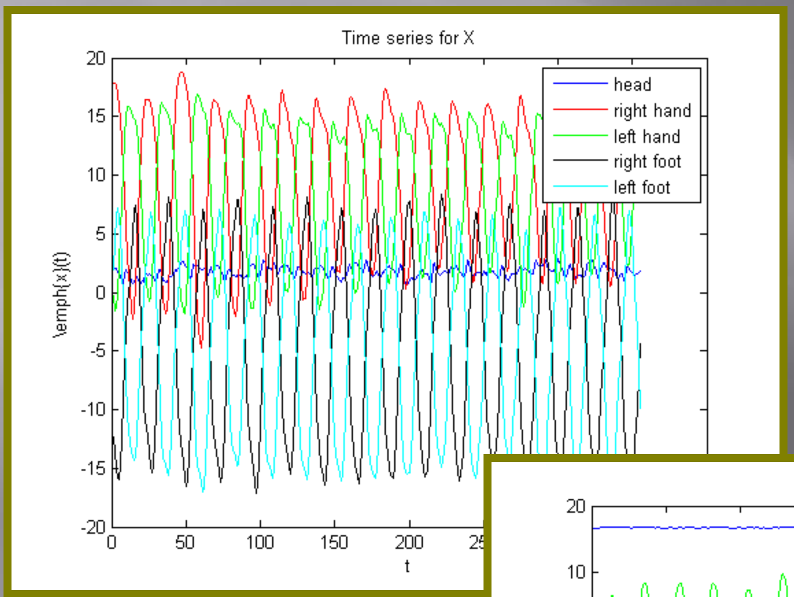


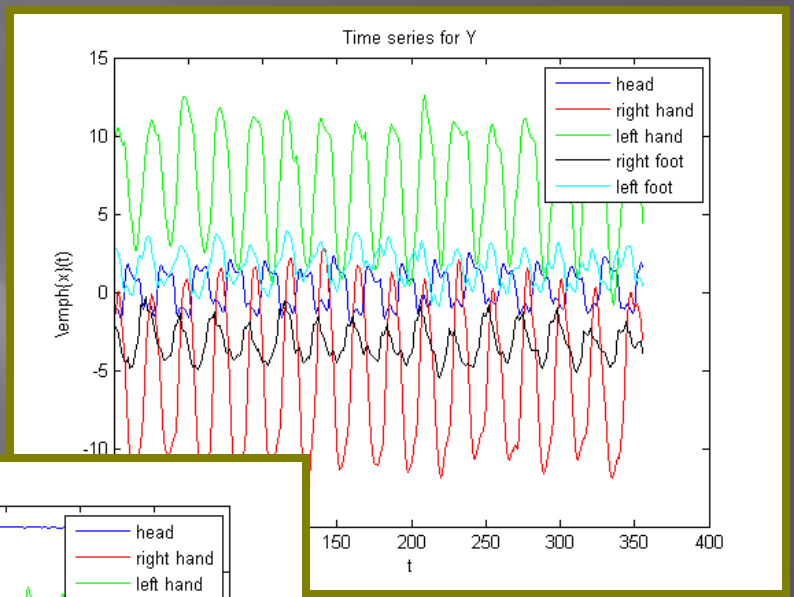
- Six Body Joints
  - Two Hands, Two Feet, Head, Belly.
- Normalized with respect to the belly point.
- Results in 5 trajectories per action.

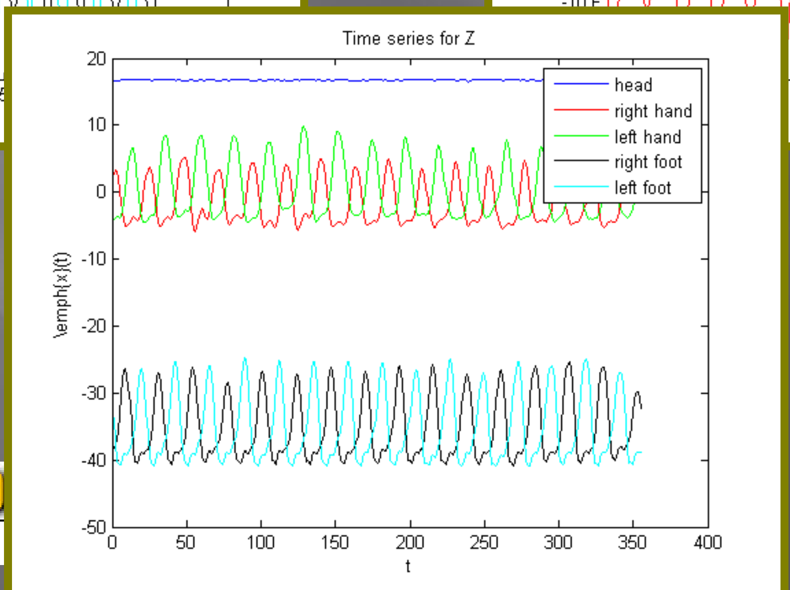
University of Central Florida VISION

Action Representation

- Each dimension of the trajectory is considered as a univariate time series







University of Central Florida

Phase Space Embedding

- Underlying Idea: All the variables of the dynamical system influence each other.

$$= z_1, z_2, z_3, \ldots, z_t$$

τ

Every point z_i of the series results from the intricate combination of influences of all the true state variables.

$$(\theta_1, \theta_2, \ldots, \theta_N)$$

Therefore, $z_{i+\tau}$ can be considered as a second substitute variable which carries the influence of all the systems variables during time interval τ.

$$z_{i+2\tau}, z_{i+3\tau}, \ldots, z_{i+m\tau}$$

Using this reasoning, introduce a series of substitute variables and obtain the whole m-dimensional space.

University of Central Florida VISION

Reconstructed Phase Space

$m = 3$

$\tau = 2$

$z_1, z_2, z_3, z_4, z_5, z_6, z_7, \ldots\ldots, z_t$

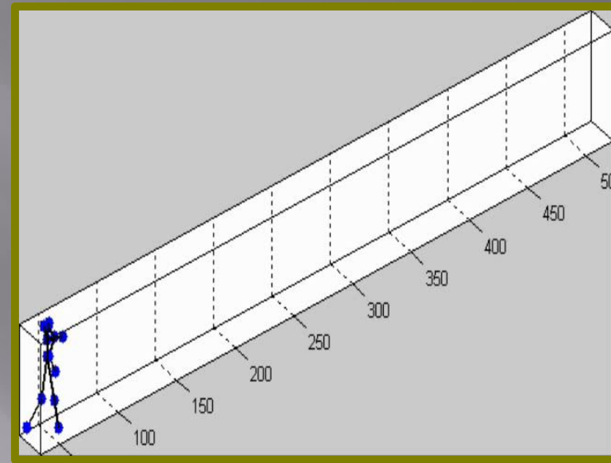
$\tau \qquad \tau$

$$X = \begin{bmatrix} z_1 & z_3 & z_5 \\ z_2 & z_4 & z_6 \\ . & . & . \end{bmatrix}$$

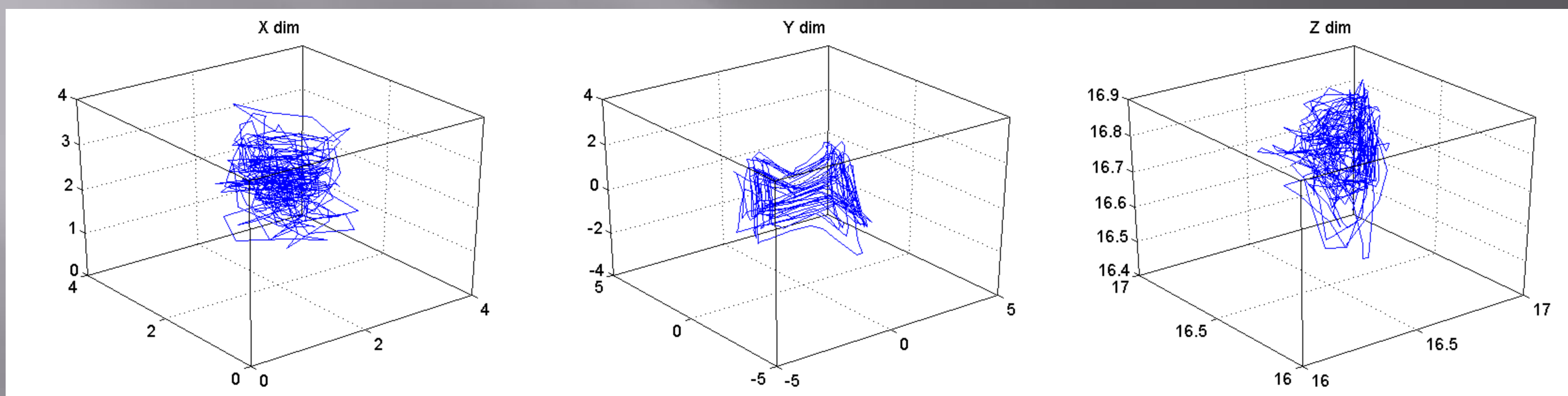
Each row is a point in a m-dimensional phase space.

m-dimensional reconstructed phase space

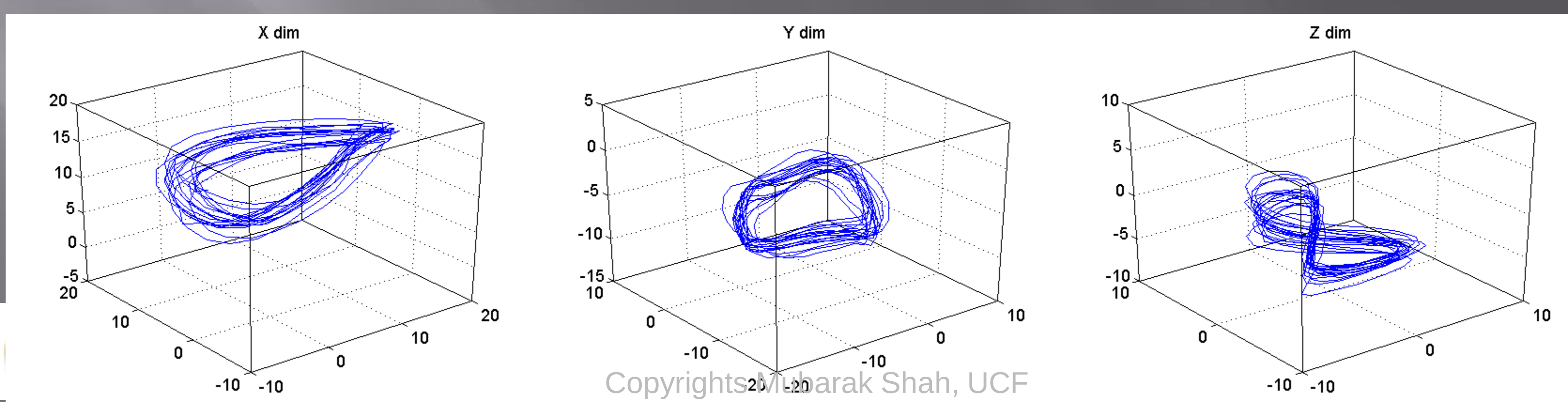
University of Central Florida VISION



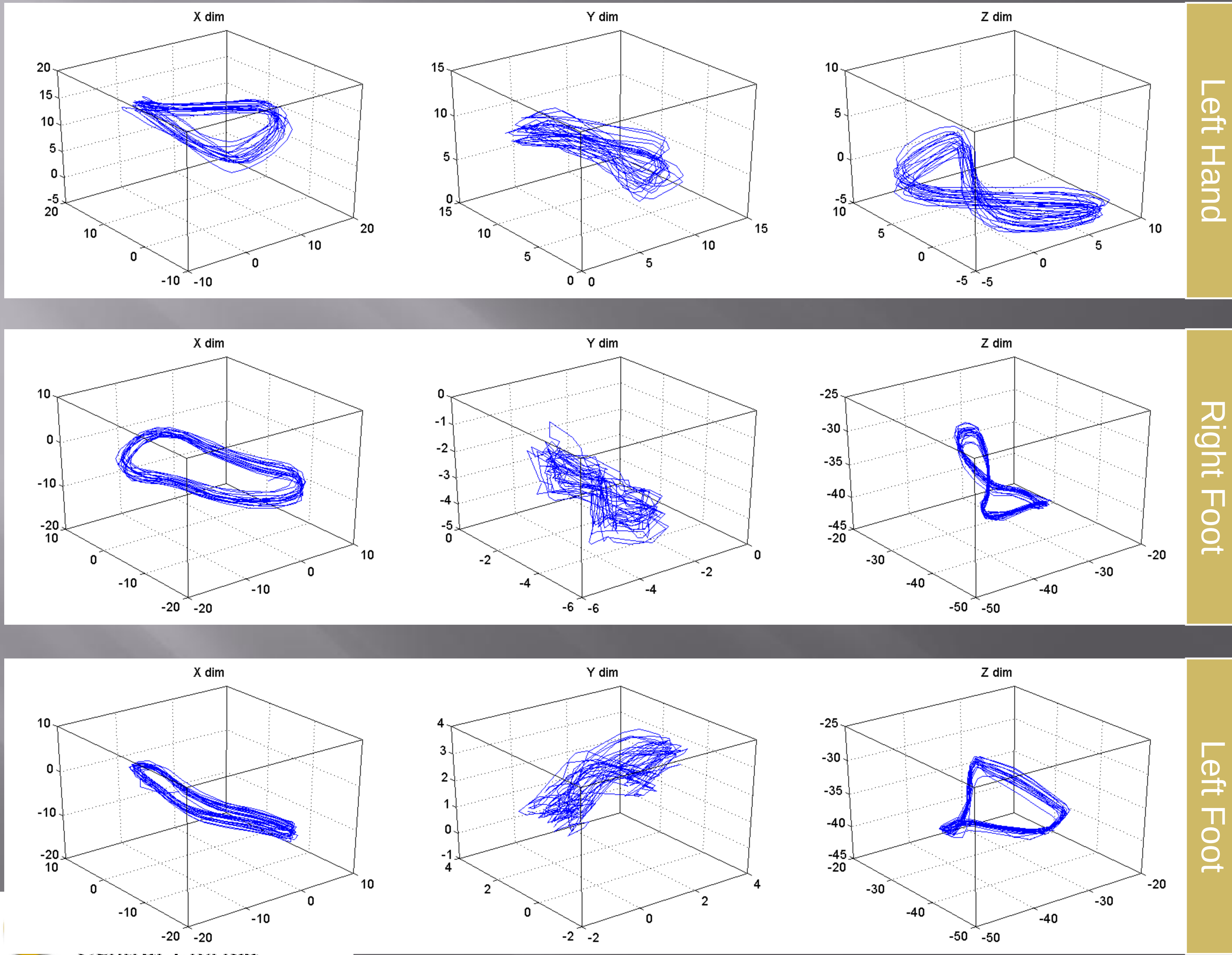
Phase Spaces



Head



Right Hand


X dim
Y dim
Z dim
Left Hand
X dim
Y dim
Z dim
Right Foot
X dim
Y dim
Z dim
Left Foot

Invariant Features

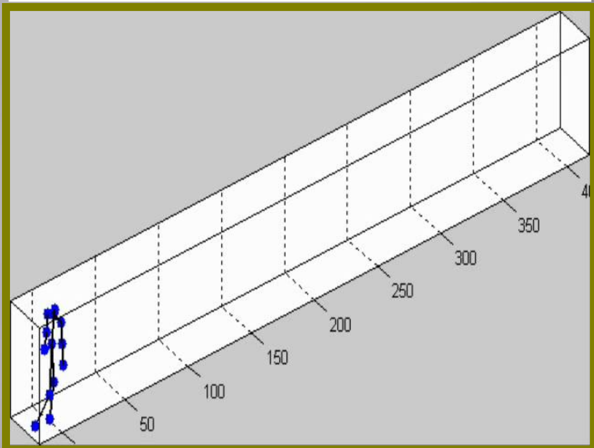
- Maximal Lyapunov Exponent
- Correlation Integral
- Correlation Dimension

Experiment-I

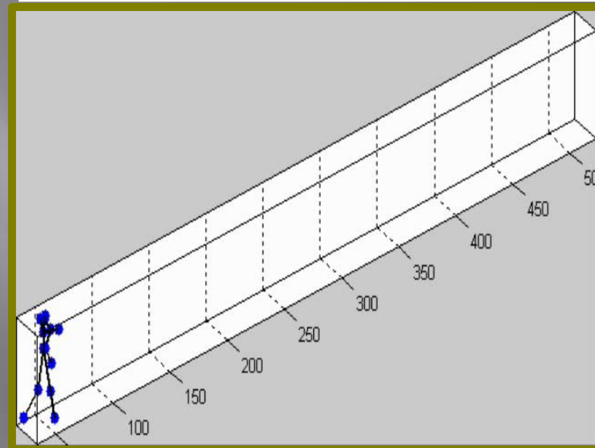
- Motion capture data
- Dataset size
  - Dance : 19
  - Run : 26
  - Walk : 46
  - Sit: 14
  - Jump: 33
- Leave-One-Out Cross validation using K-means classifier.

Experiments Motion Capture Data

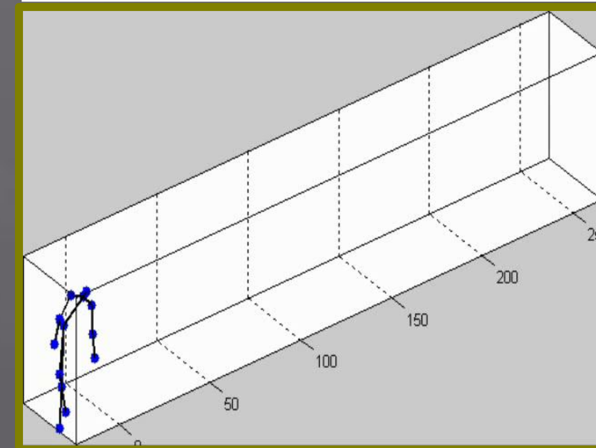
Walking



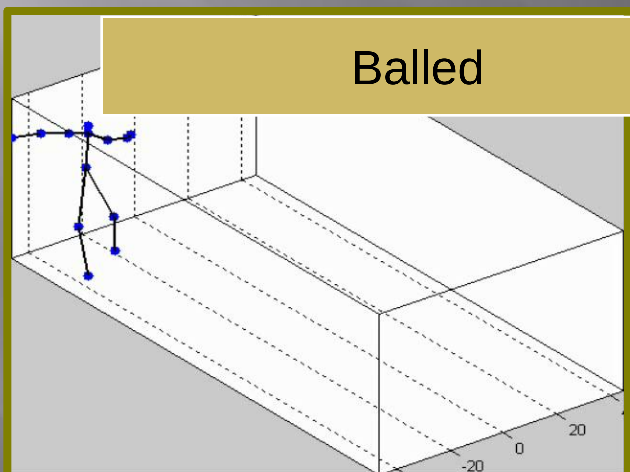
Running



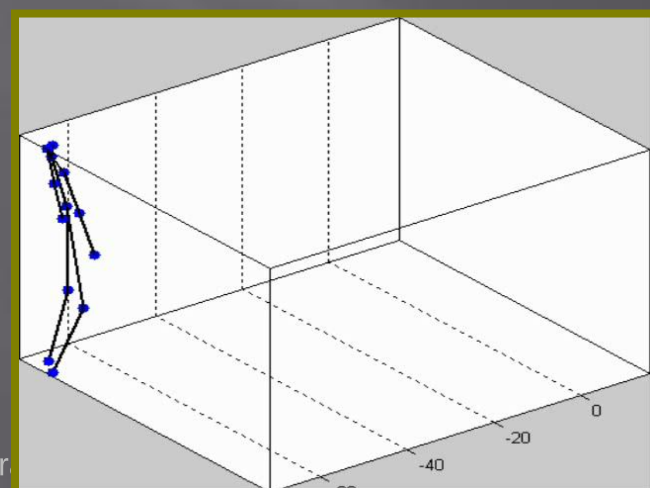
Jumping



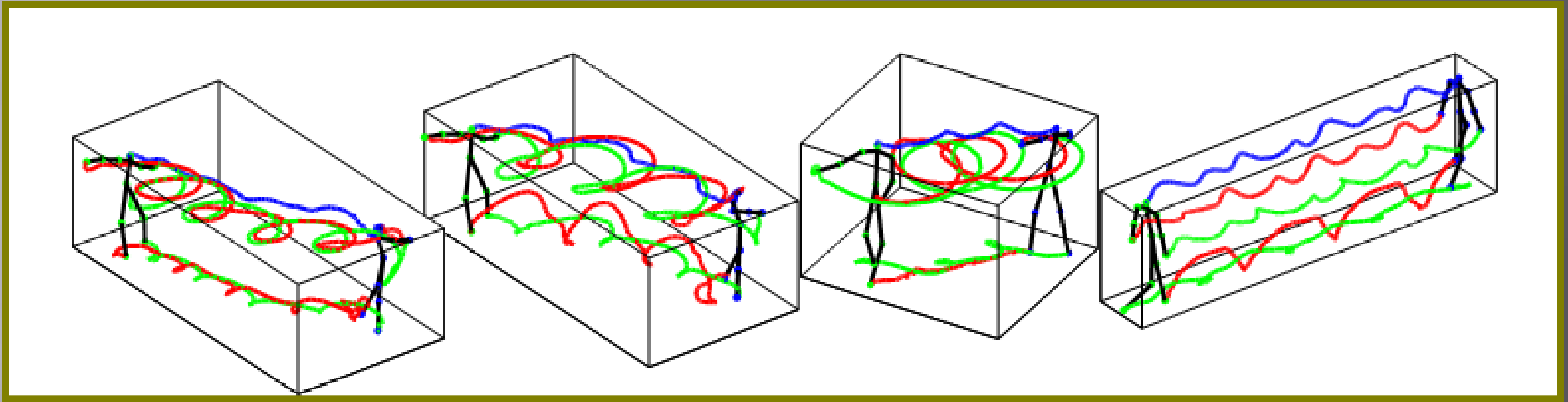
Balled



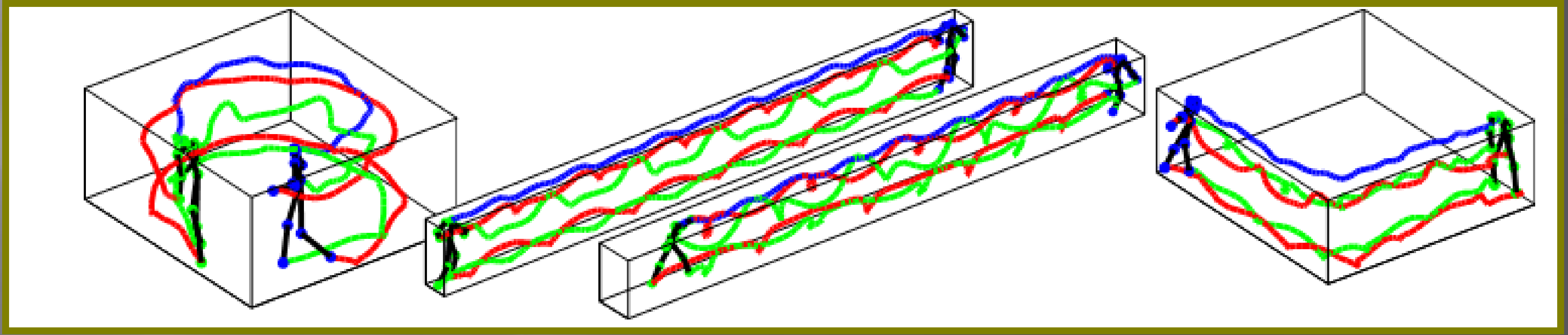
Sitting



Experiments Motion Capture Data



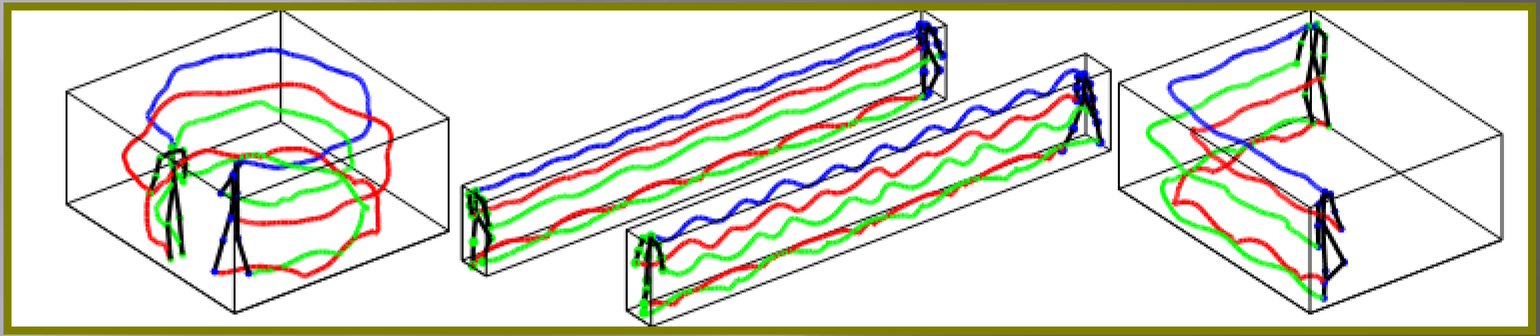
Dance



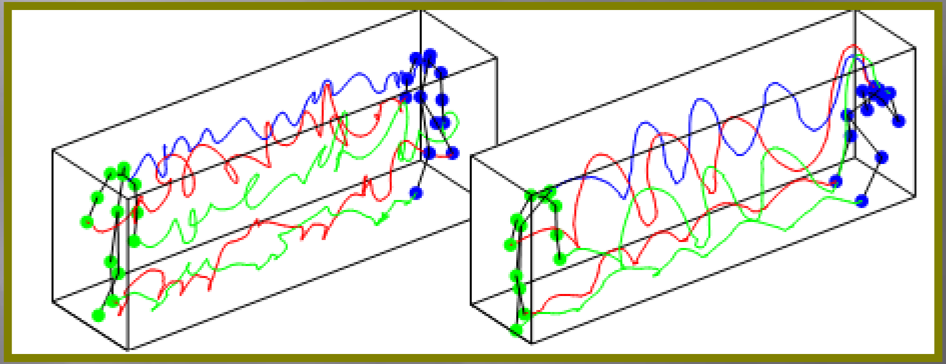
Run

University of Central Florida VISION

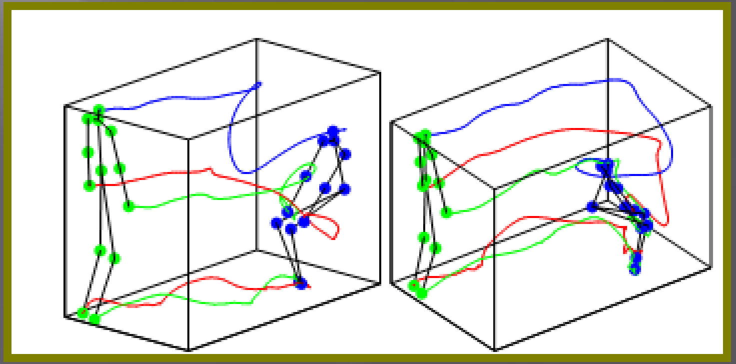
Experiments Motion Capture Data



Walk



Jump



Sit

University of Central Florida VISION

Experiments Motion Capture Data

	Dance	Jump	Run	Sit	Walk
Dance	28				2
Jump		13			1
Run	2	1	22	1	4
Sit				33	
Walk	3		2		43

Mean Accuracy: 89.7%

University of Central Florida VISION

Experiment -II

- Wizemann Action Data Set
- Nine actions performed by nine different actors:
  - Bend, Jumping Jack, Jump Forward, Jump in Place, Run, Side Gallop, Walk, Wave One Hand, Wave Two Hands
- 81 videos

Experiment-II

Experiments Wizemann Action Data Set

	Bend	Jumping Jack	Jumping Forward	Jumping In Place	Run	Side Gallop	Walk	Wave1	Wave2
Bend	9								
Jumping Jack		9							
Jump Forward			5	2	2				
Jump In Place				9					
Run					8		1		
Side Gallop					1	8			
Walk							9		
Wave1								9	
Wave2									9

University of Central Florida VISION

Papers

- Cen Rao, Alper Yilmaz, and Mubarak Shah, View-Invariant Representation And Recognition of Actions, International Journal of Computer Vision, Vol.50, Issue 2, 2002. (http://www.cs.ucf.edu/~vision/papers/ijcv2002.pdf).
- Alper Yilmaz and Mubarak Shah, Actions As Objects: A Novel Action Representation, IEEE CVPR 2005, San Diego, June 20-26. (http://www.cs.ucf.edu/~vision/papers/yilmaz_cvpr_2005.pdf)
- Alexei Gritai, Yaser Sheikh, and Mubarak Shah, On the Invariant Analysis of Human Actions, 17th conference of the International Conference on Pattern Recognition, 2004. (http://www.cs.ucf.edu/~vision/papers/gritai_icpr_2004.pdf)

Papers

- Yaser Sheikh, Alexei Gritai, and Mubarak Shah On the Spacetime Geometry of Galilean Cameras, IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA 2007. (http://server.cs.ucf.edu/~vision/papers/758.pdf)
- Saad Ali, Arslan Basharat, and Mubarak Shah, Chaotic Invariants for Human Action Recognition, ICCV 2007, Rio de Janeiro, Brazil. (http://server.cs.ucf.edu/~vision/papers/SaadICCV07.pdf)