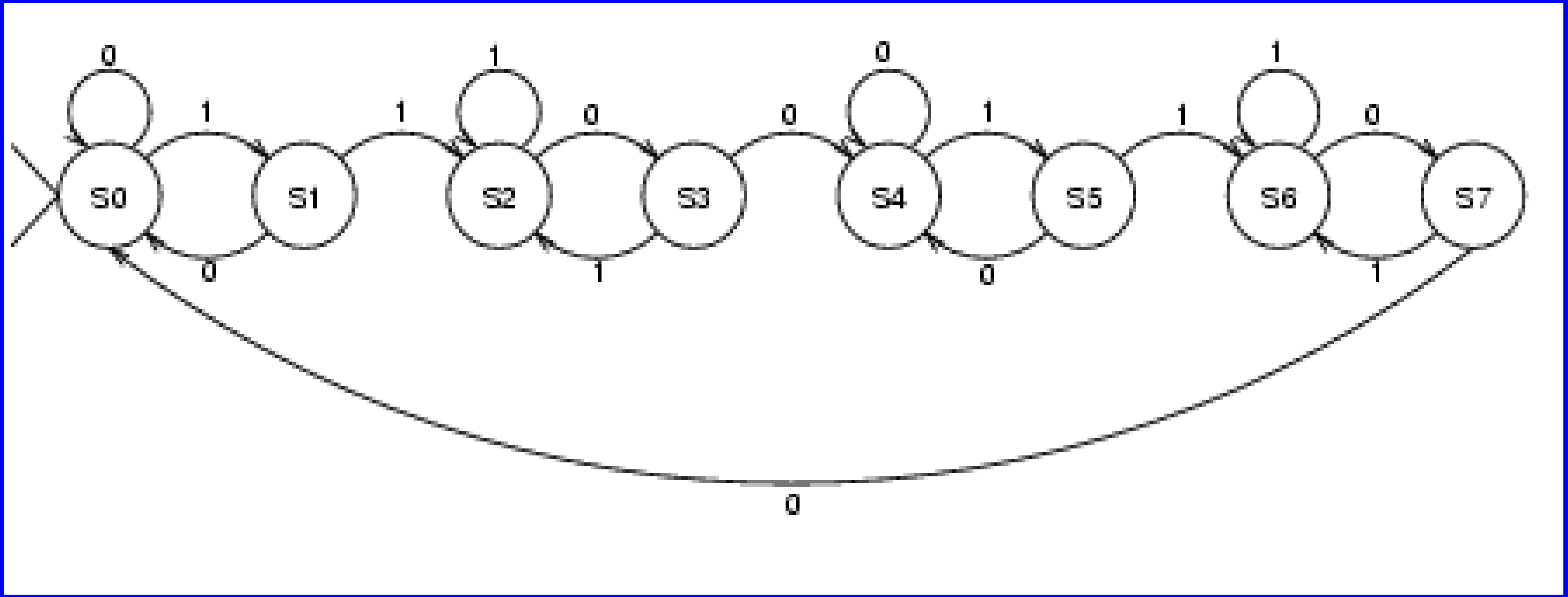
Action Recognition

Approaches

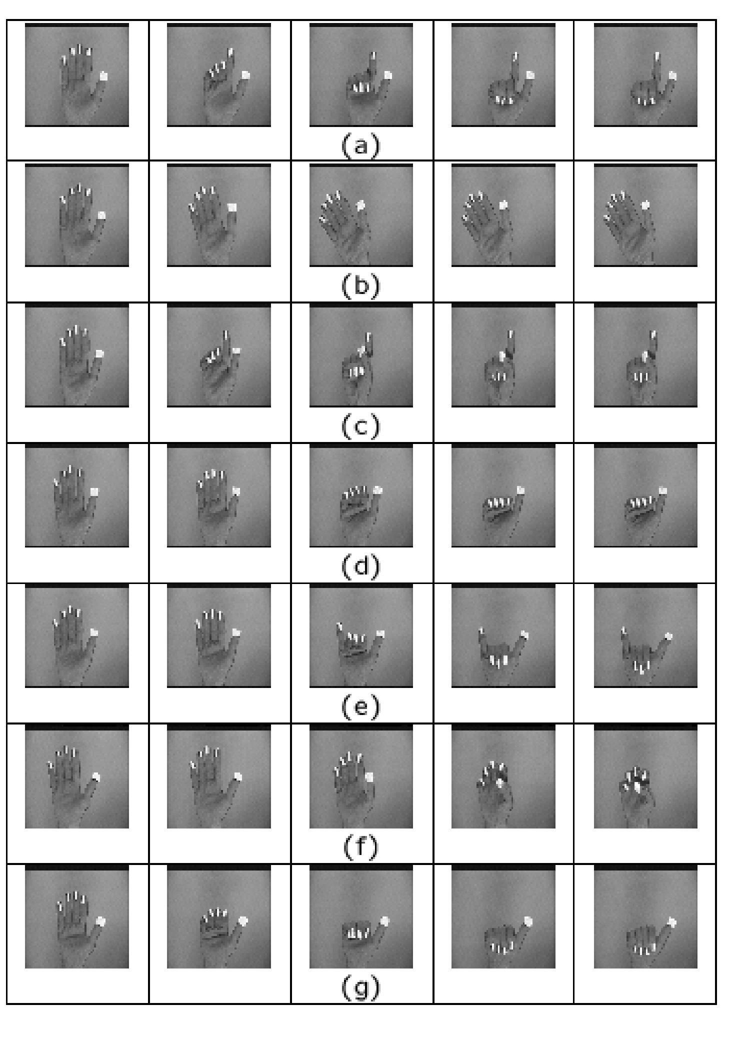
- FSA
- HMMs/NNs
- Rule-based
- ---

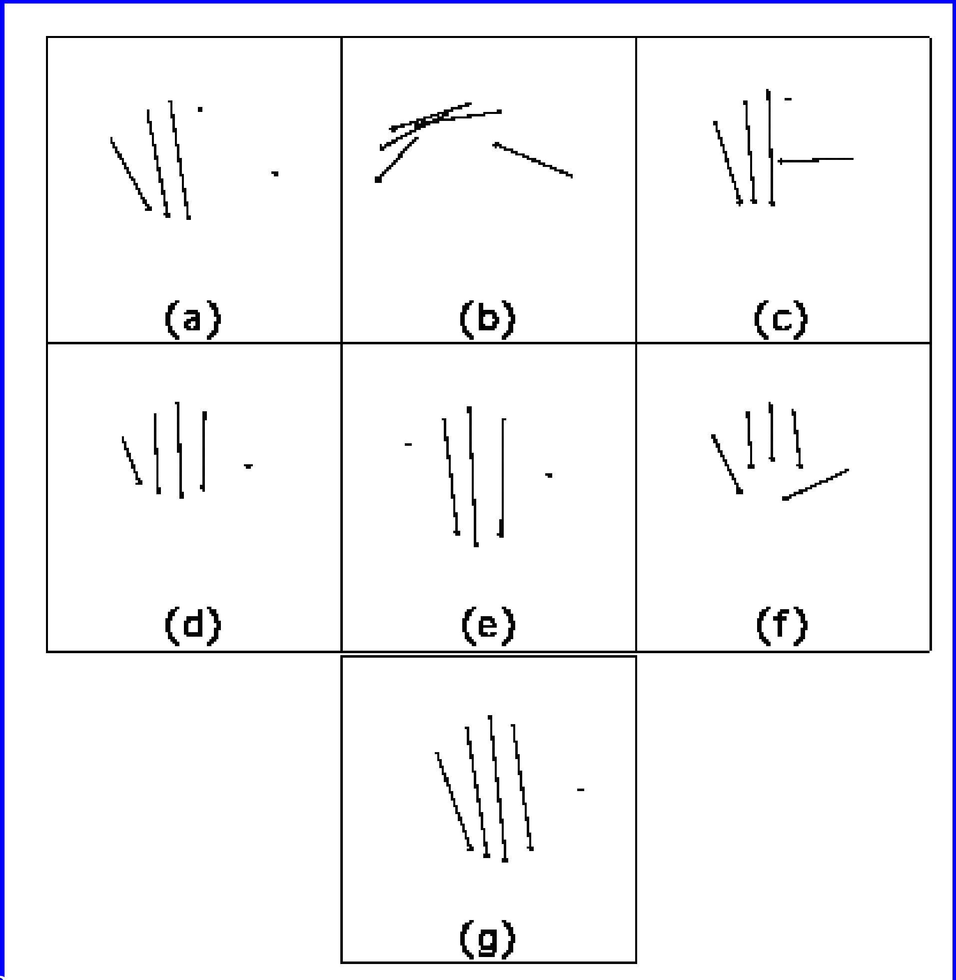
- Representation is important

FSA: Hand Gesture Recognition

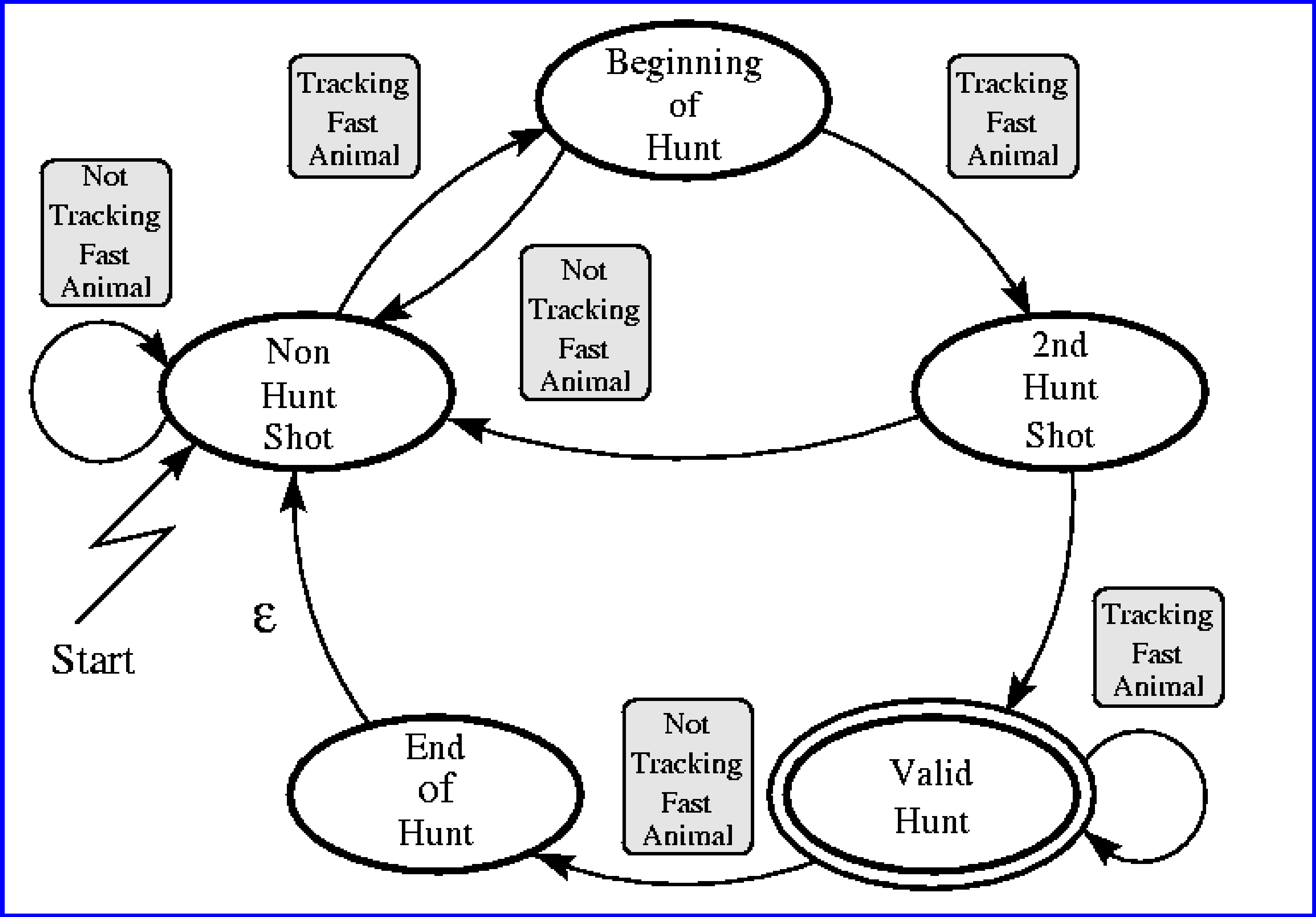


FSA: Hand Gesture Recognition



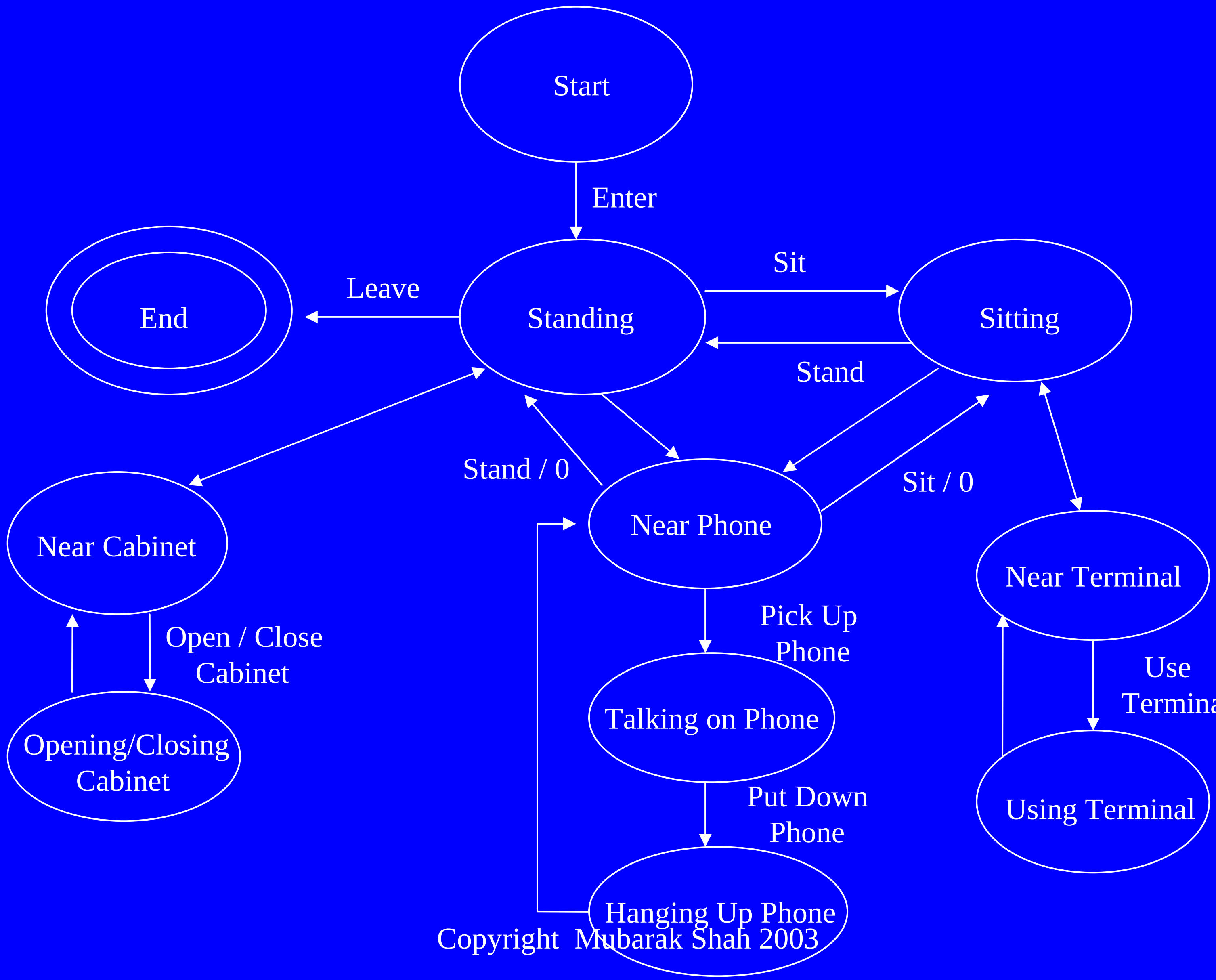


Hunt events



Beginning of Hunt
Tracking Fast Animal
Tracking Fast Animal
Not Tracking Fast Animal
Not Tracking Fast Animal
Non Hunt Shot
2nd Hunt Shot
Start
ε
Tracking Fast Animal
Not Tracking Fast Animal
End of Hunt
Valid Hunt

FSA: Recognizing Human Behavior in Office Environment



Rule-Based: Detecting Violence

Rule-Based: Detecting Violence

Rule-Based: Recognizing Outdoor Activities

Limitations

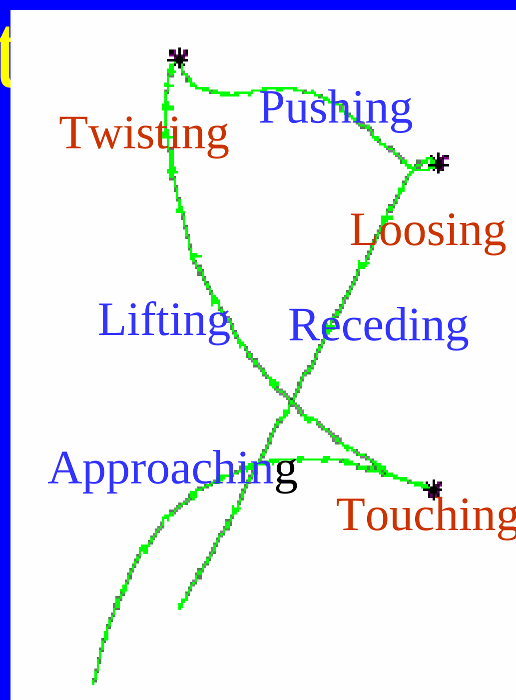
- A priori knowledge
- Extensive training
- No explanation
- No learning
- Representation
- View invariance

View-invariant Representation and Recognition of Human Action

Hand Actions Recognition

- hand generates a 3-D trajectory with respect to time.
- analyze 2-D projection of this 3-D trajectory.
- View invariance issues.

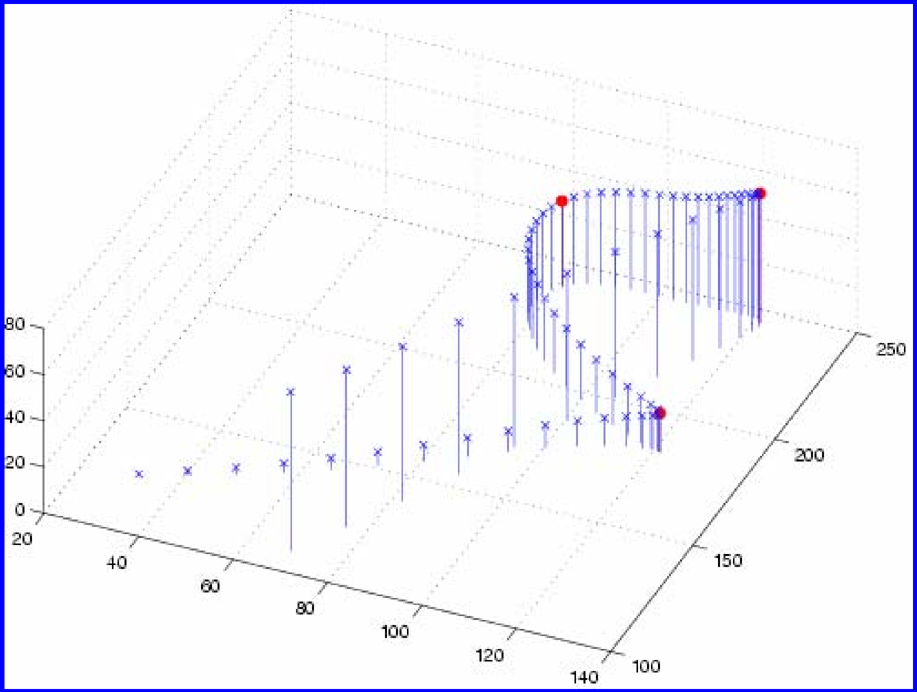
Instant epresentation

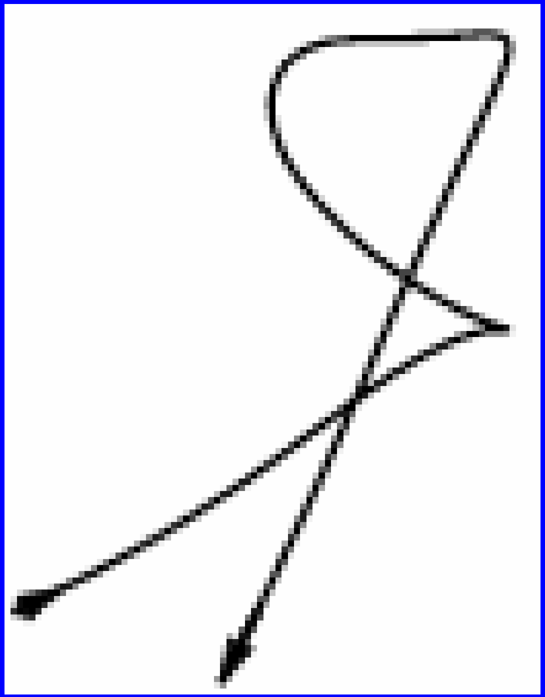


Hand Actions

Spatiotemporal Curve

$$r_{st} = [x(t) \quad y(t) \quad t]$$





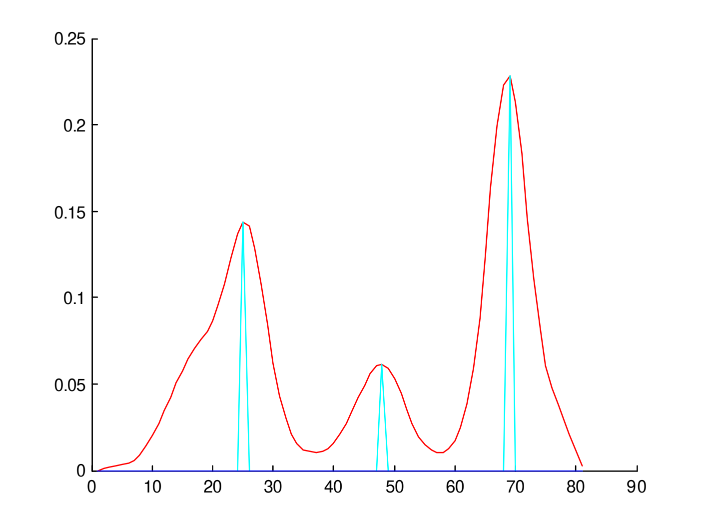
Spatiotemporal Curvature

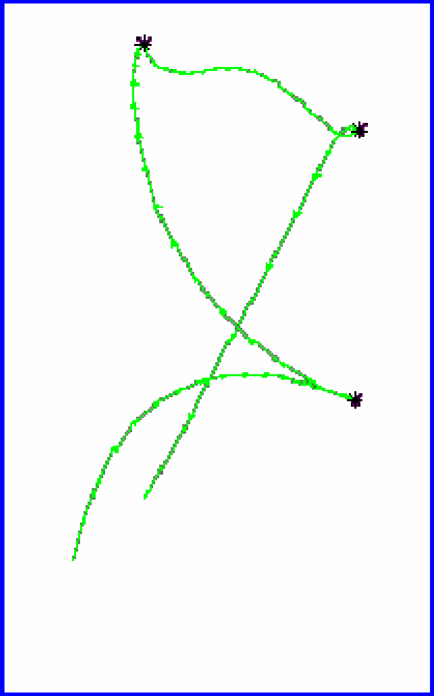
$$k = \frac{\sqrt{A^2 + B^2 + C^2}}{\left((x')^2 + (y')^2 + (t')^2\right)^{\frac{3}{2}}} \qquad A = \begin{vmatrix} y' & t' \\ y'' & t'' \end{vmatrix}, B = \begin{vmatrix} t' & x' \\ t'' & x'' \end{vmatrix}, C = \begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}$$

Spatiotemporal curvature captures both the **speed** and **direction** changes in **one quantity**.

Representation of Actions

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

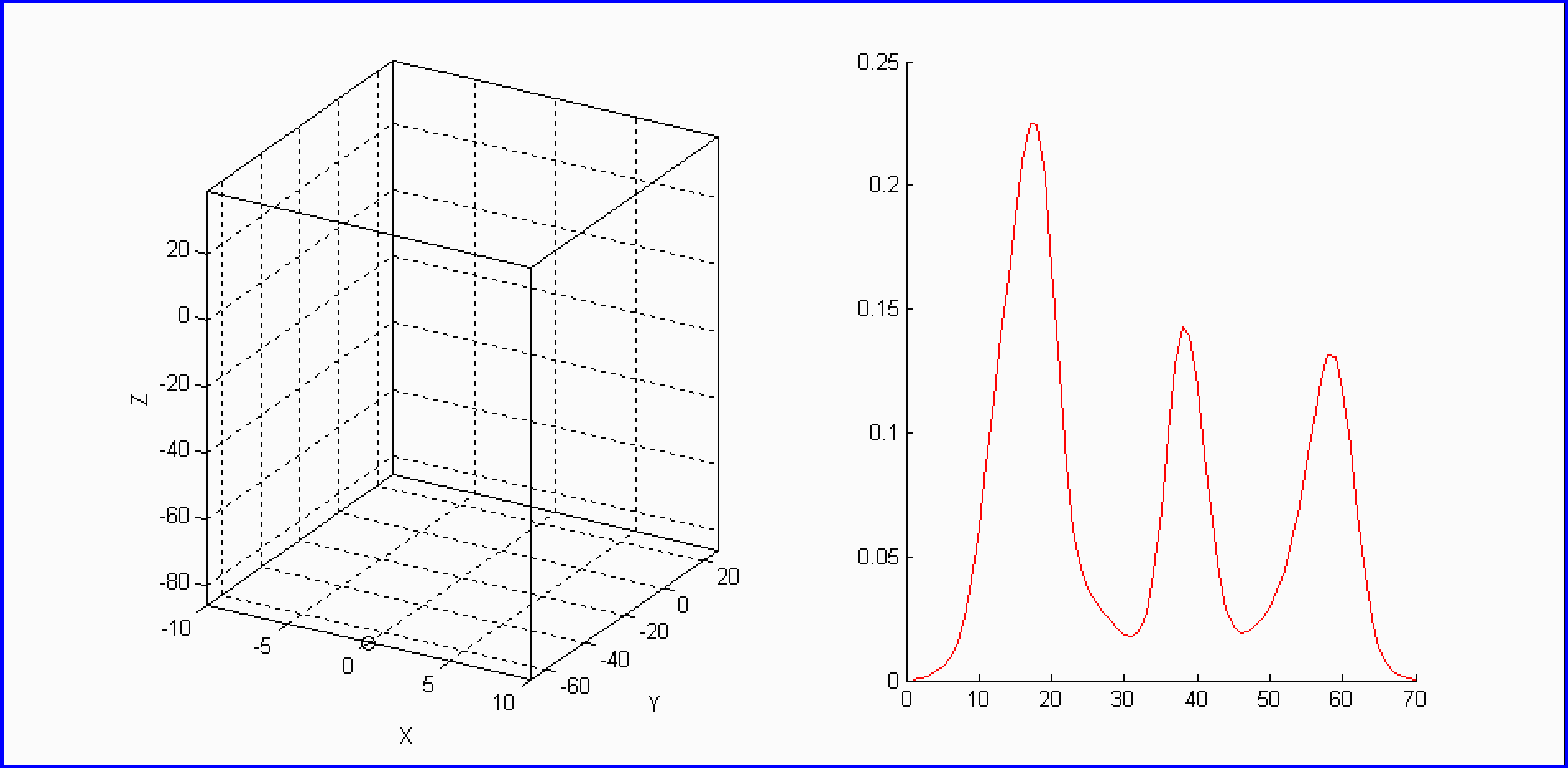




Action Units

- Psychology research shows:
  - People tend to divide an action into atomic units at the places where the motion characteristics change the most.
    - Stops, starts, pauses, dynamic instants

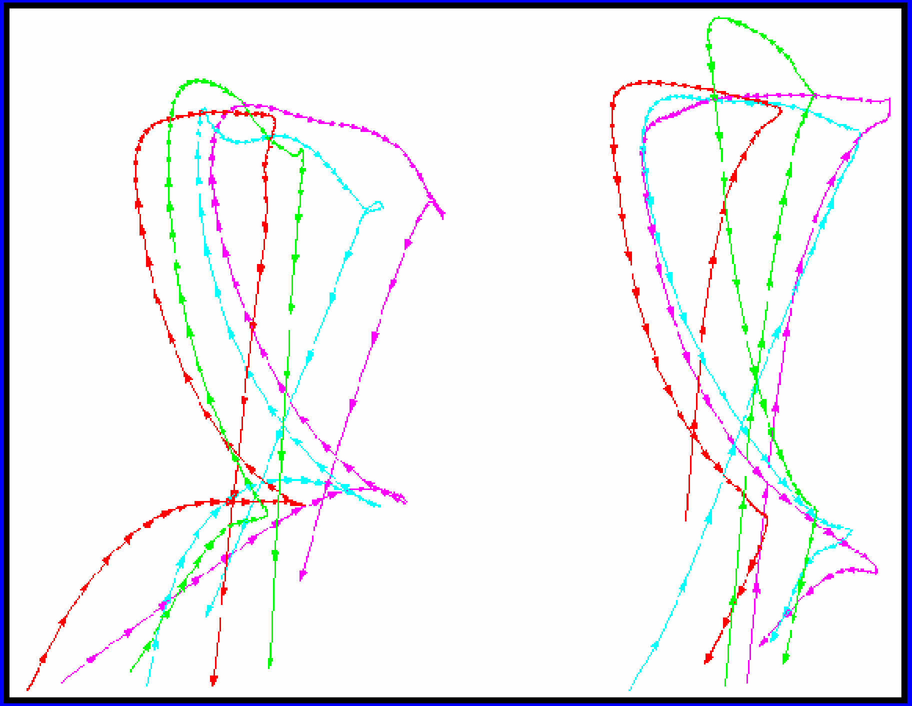
Viewing Directions and Spatiotemporal Curvature



Although the viewing directions are quite different, the peak locations are consistent.

View-invariant Action Representation

- Opening and closing an overhead cabinet

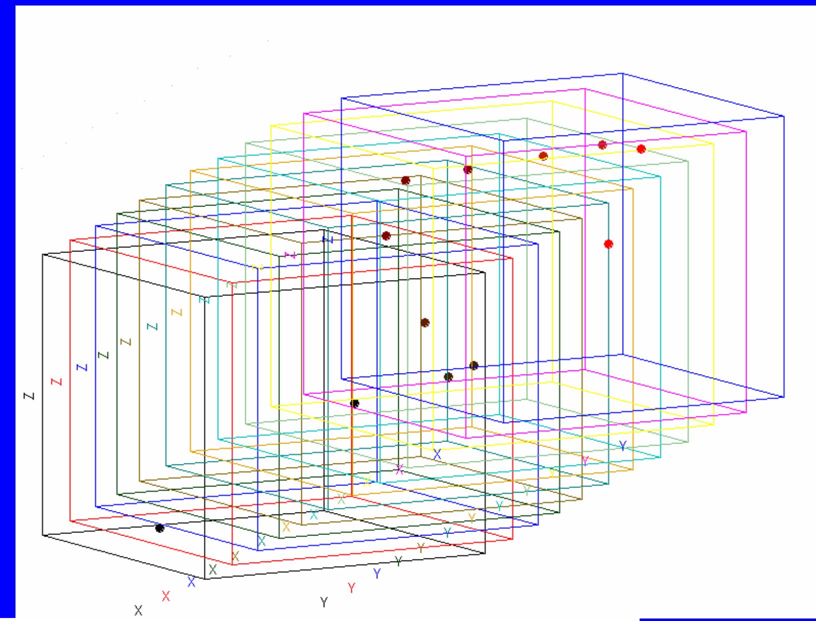


Opening Closing

View-invariant Matching

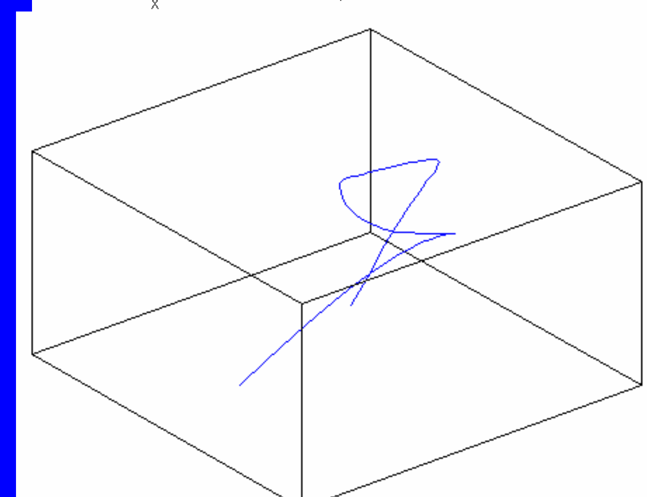
- Consider 3D trajectories as 3D objects.

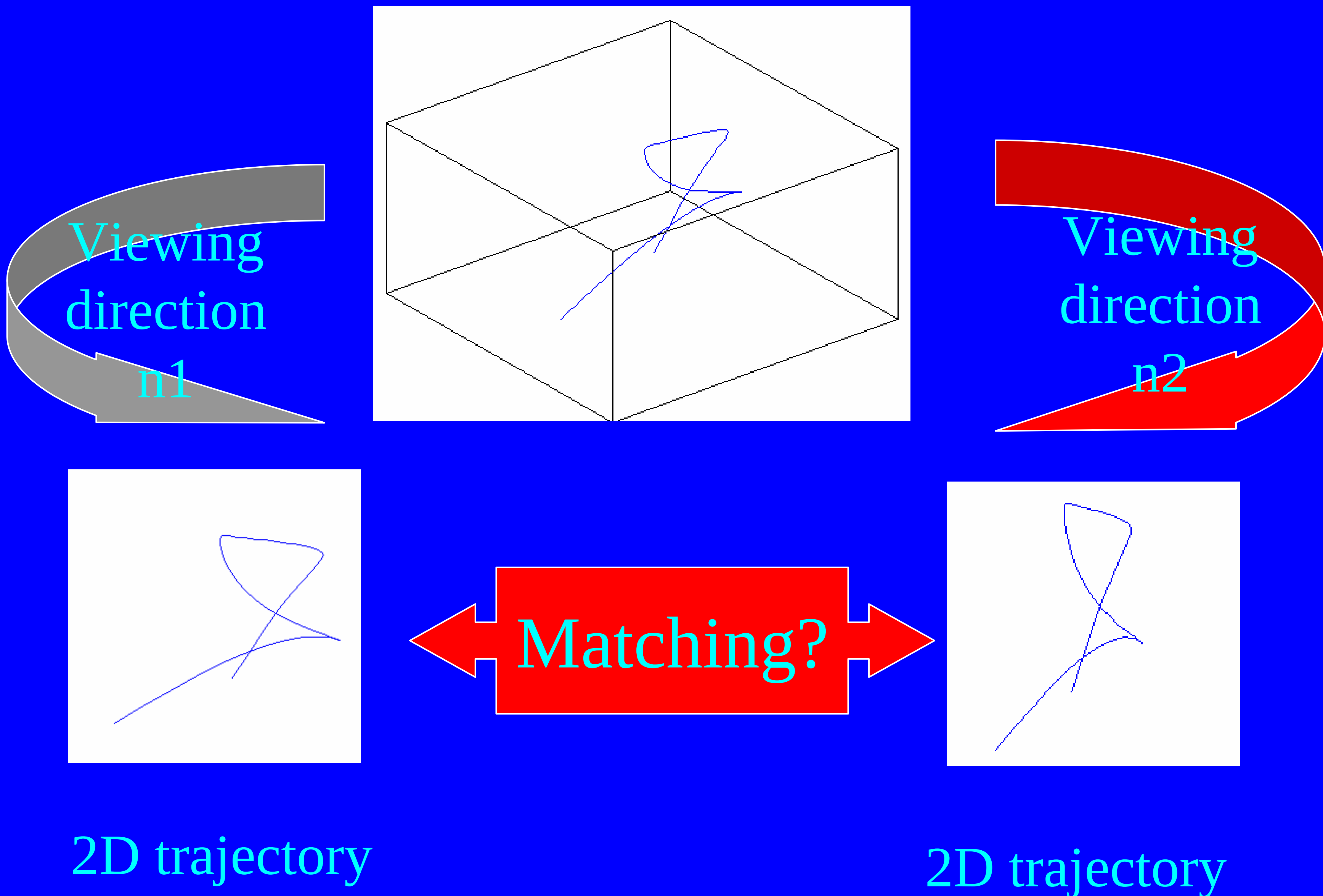
Action Trajectory in 4D



Sampling in time

Ignore the time index




Viewing direction n1
Viewing direction n2
Matching?
2D trajectory
2D trajectory

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:

$$M = \begin{bmatrix} \mu_1^{v_1} & \mu_2^{v_1} & \dots & \mu_n^{v_1} \\ \nu_1^{v_1} & \nu_2^{v_1} & \dots & \nu_n^{v_1} \\ \mu_1^{v_2} & \mu_2^{v_2} & \dots & \mu_n^{v_2} \\ \nu_1^{v_2} & \nu_2^{v_2} & \dots & \nu_n^{v_2} \end{bmatrix} = P \bullet S = \begin{bmatrix} \Pi_{v_1} \\ \Pi_{v_2} \end{bmatrix} \bullet \begin{bmatrix} X_1 & X_2 & & X_n \\ Y_1 & Y_2 & \cdots & Y_n \\ Z_1 & Z_2 & & Z_n \end{bmatrix}$$

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

M is 4 by n, P is 4×3 and S is $3 \times n$,

then the rank of M is at most 3.

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:

$$M = \begin{bmatrix} \mu_1^i & \mu_2^i & \cdots & \mu_n^i \\ \nu_1^i & \nu_2^i & \cdots & \nu_n^i \\ \mu_1^j & \mu_2^j & \cdots & \mu_n^j \\ \nu_1^j & \nu_2^j & \cdots & \nu_n^j \end{bmatrix} \qquad dist_{i,j} = |\sigma_4|$$

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:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}^T F \begin{bmatrix} u_i' \\ v_i' \\ 1 \end{bmatrix} = 0,$$

$$Mf = \begin{bmatrix} u_1'u_1 & u_1'v_1 & u_1' & v_1'u_1 & v_1'v_1 & v_1' & u_1 & v_1 & 1 \\ u_2'u_2 & u_2'v_2 & u_2' & v_2'u_2 & v_2'v_2 & v_2' & u_2 & v_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_n'u_n & u_n'v_n & u_n' & v_n'u_n & v_n'v_n & v_n' & u_n & v_n & 1 \end{bmatrix} f = 0$$

M is 9 by *n* matrix

$$f = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{21} & f_{22} & f_{23} & f_{31} & f_{32} & f_{33} \end{bmatrix}$$

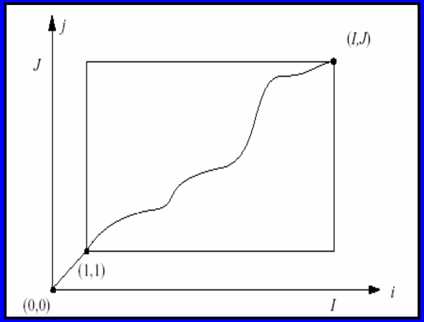
To solve the equation, the rank(M) must be 8.
The 9th singular value of M, σ_9, is the match measure.

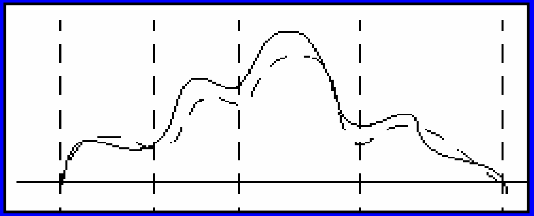
Instant-Interval Representation

- We have not used the interval information
- View Invariance
- Temporal Invariance

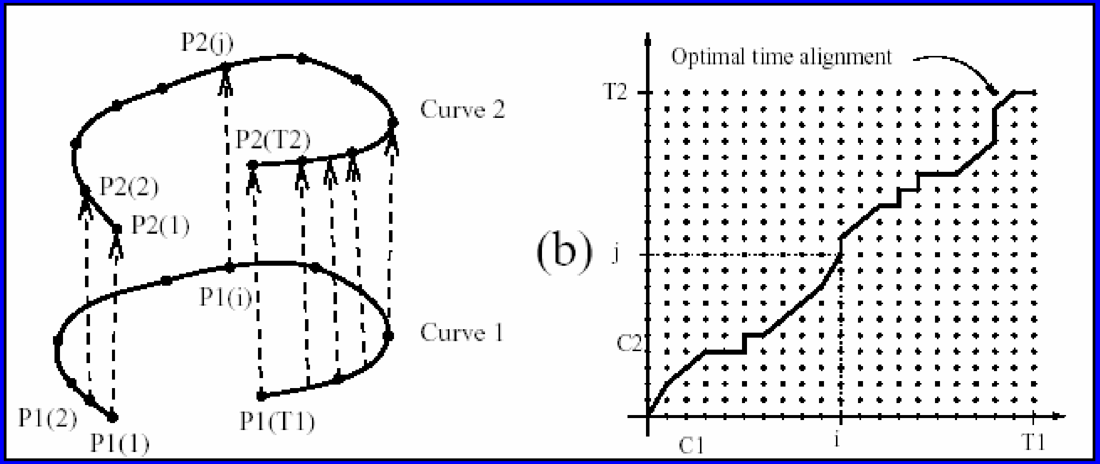
DTW and Temporal Signals

- Match two 1D temporal signals:





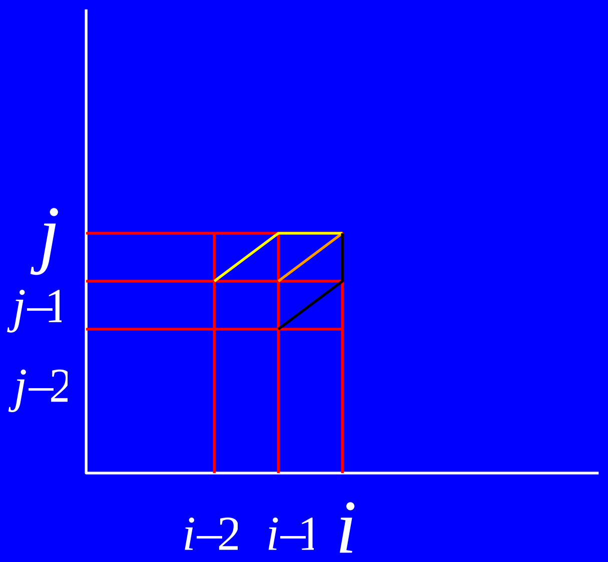
- Match two 2D temporal curves:



Warping

$$A = [a_1, a_2, \ldots, a_i, a_I]$$

$$B = [b_1, b_2, \ldots, b_j, b_J]$$



$$d_{ij} = \left| a_i - b_j \right|$$

$$g_{11} = 2d_{11}$$

$$g(i,j) = \min \begin{bmatrix} g(i-1, j-2) + 2d(i, j-1) + d(i,j) \\ g(i-1, j-1) + 2d(i,j) \\ g(i-2, j-1) + 2d(i-1, j) + d(i,j) \end{bmatrix}$$

Affine View-invariant DTW

Step 1: Pick up 4 instants from trajectories I and I', (x_1,y_1), (x_2,y_2), (x,y_3), (x_4,y_4) and (x_1',y_1'), (x_2',y_2'), (x_3',y_3'), (x_4',y_4') are image coordinates.

Step 2: Apply Dynamic Time Warping

- the similarity between i^{th} (u_i,v_i) and j^{th} (u'_j,v'_j) point is:

$$d(i,j) = |\sigma_4|$$

$$M = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & u_i \\ y_1 & y_2 & y_3 & y_4 & v_i \\ x_1' & x_2' & x_3' & x_4' & u_j' \\ y_1' & y_2' & y_3' & y_4' & v_j' \end{bmatrix}$$

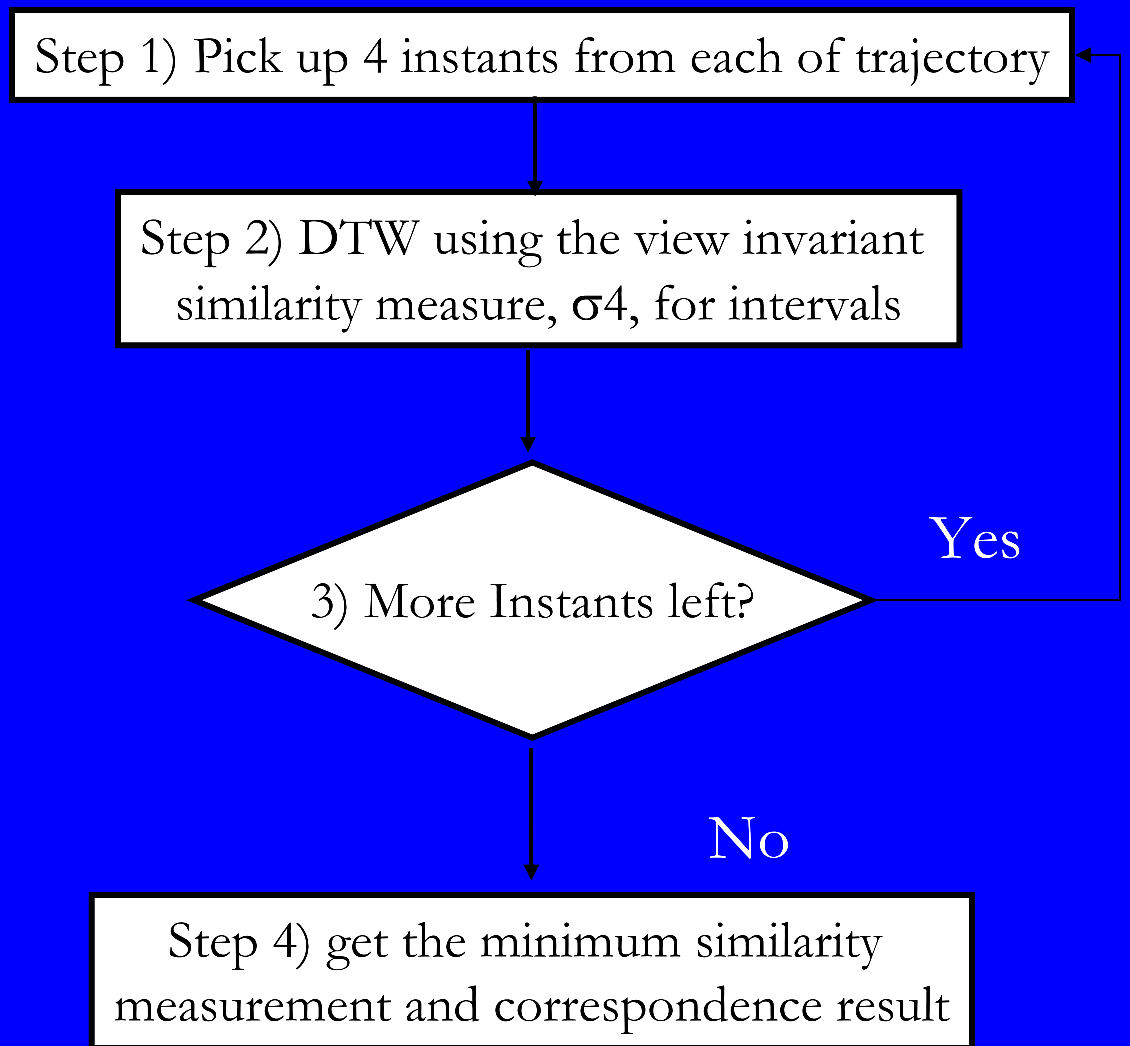
σ_4 is the 4^{th} singular value of matrix M

Affine View-invariant DTW (Con.)

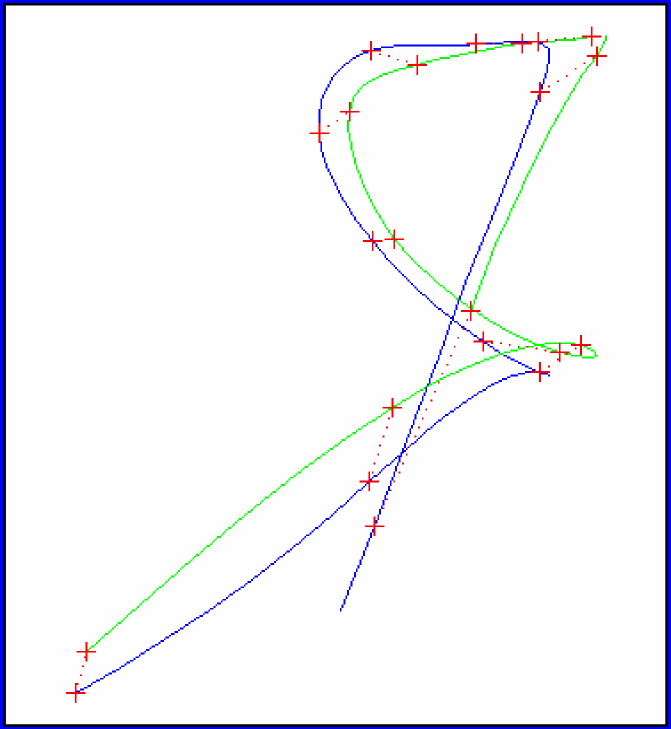
Step 3: if there are more than 4 pairs of instants in the trajectories, go back to step 1 and try other combinations of 4 instants.

Step 4: pick the minimal matching error as the similarity measurement and get the correspondence result.

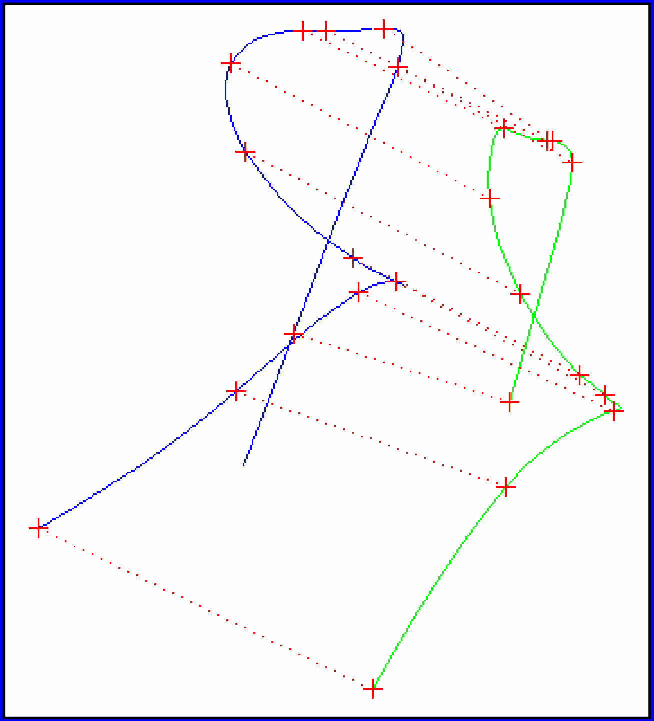
Matching using View-invariant Dynamic Time Warping



Action Recognition Results

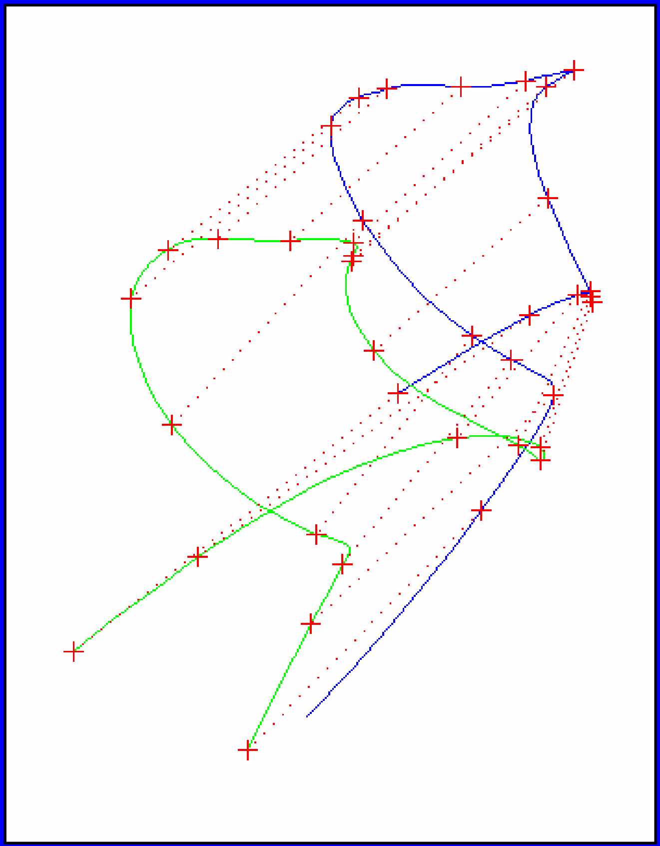


Difference = 2

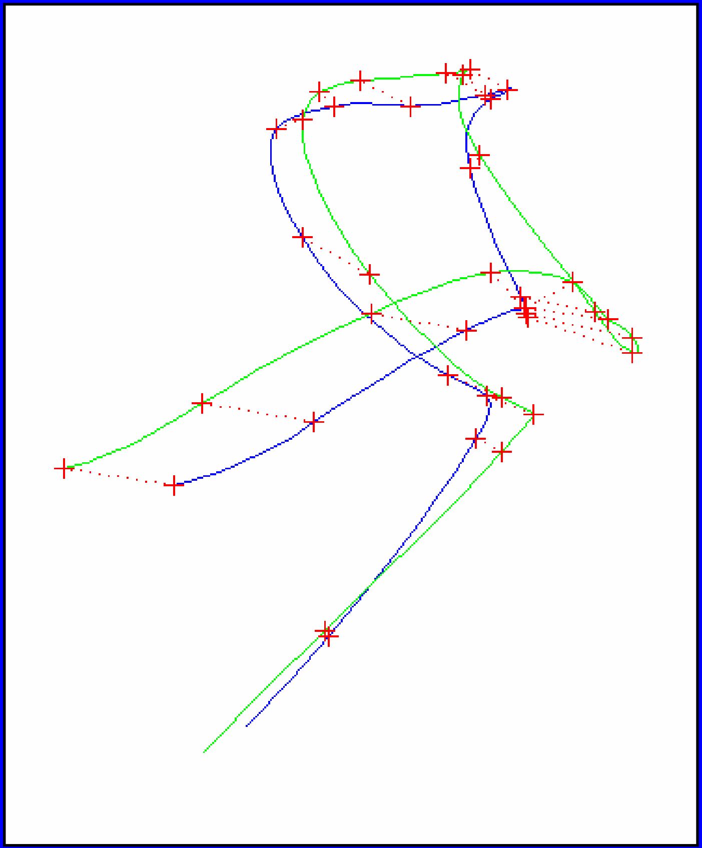


Difference = 2.3

Action Recognition Results

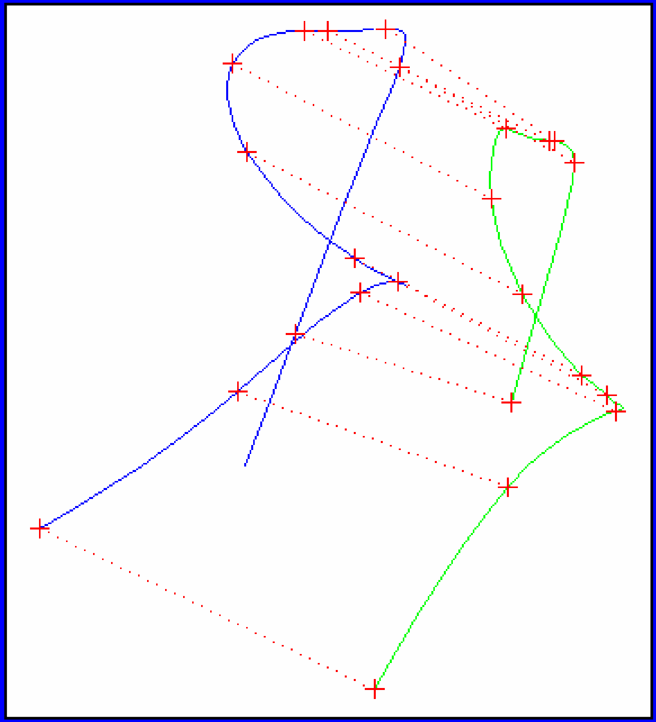


$Difference = 2.5$

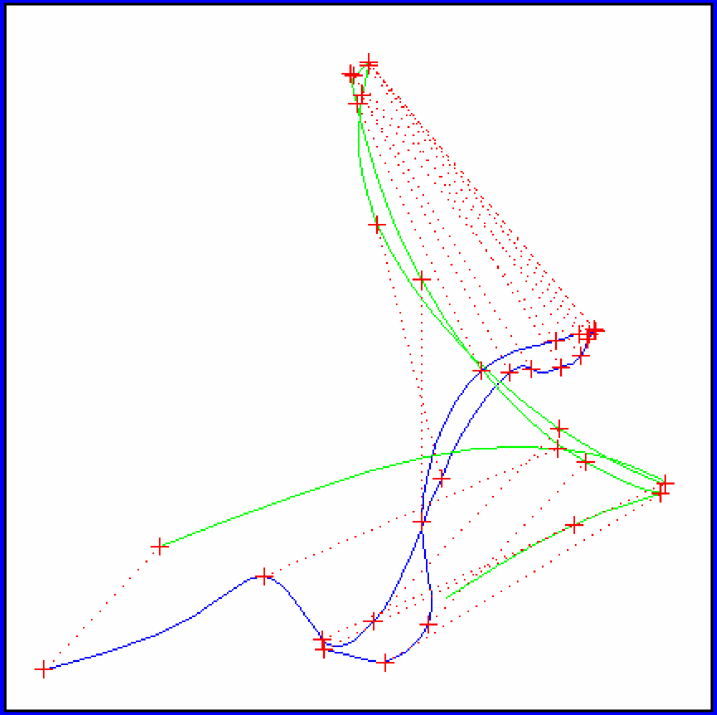


$Difference = 3.2$

View-invariant DTW Results



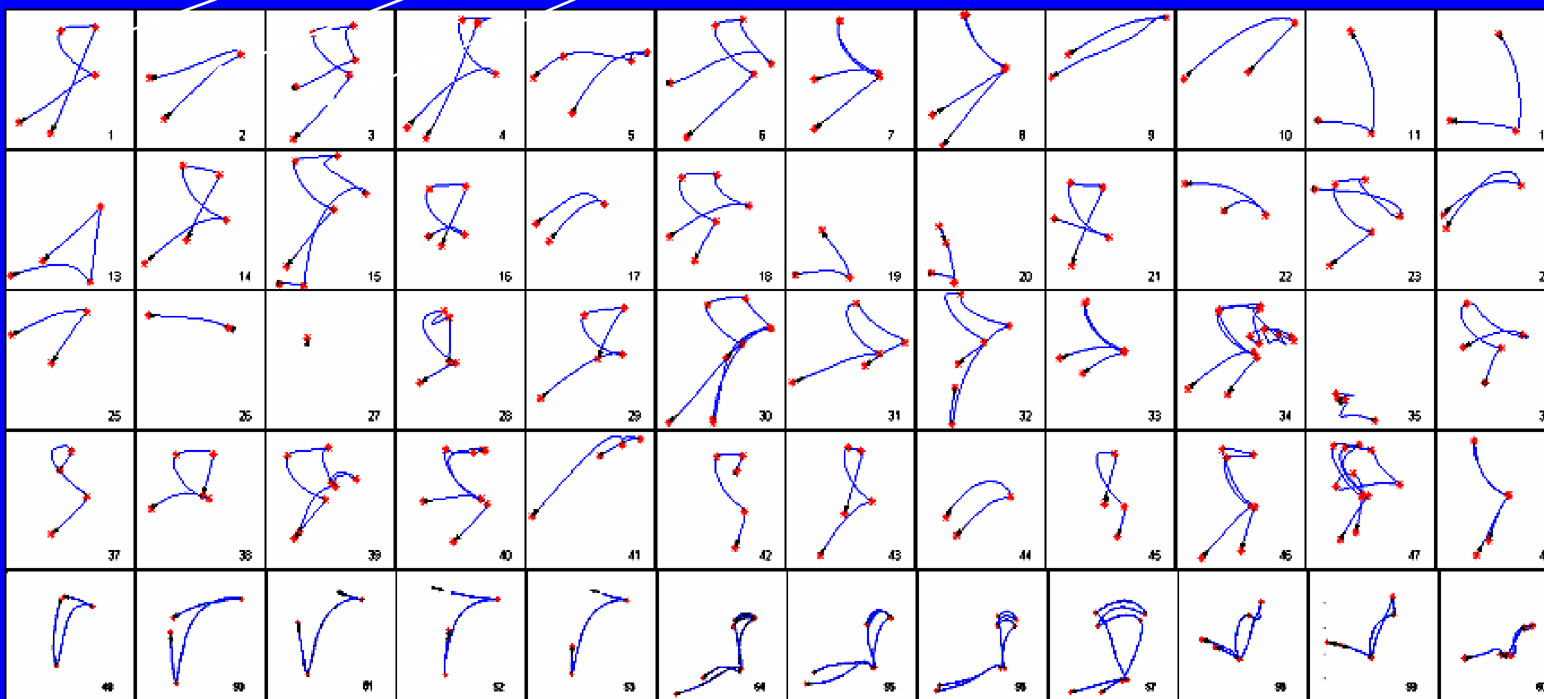
Difference = 2.3



Difference = 71

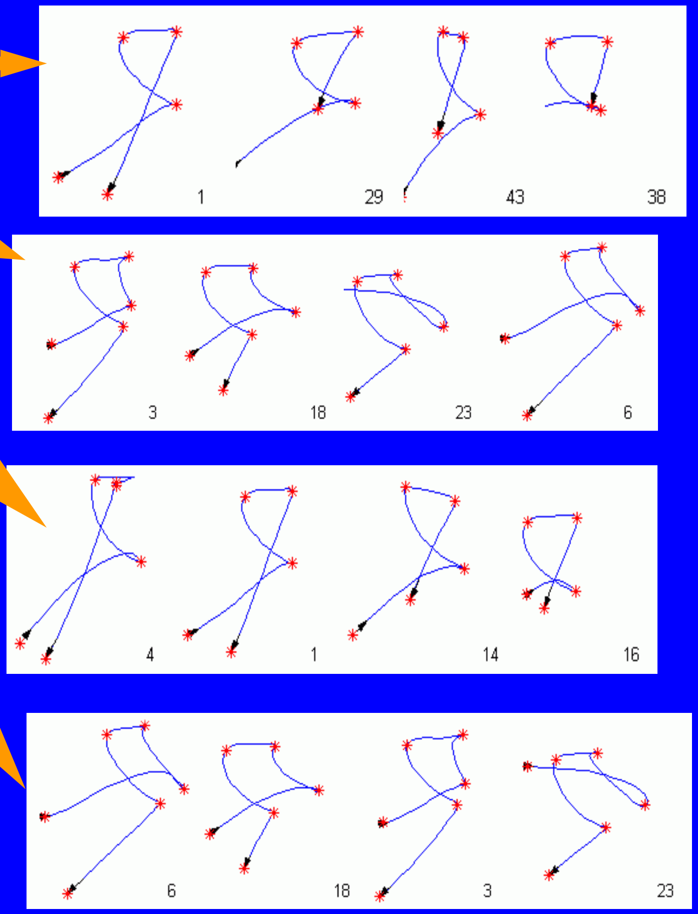
Experimental Results

60 Action Trajectories
7 People

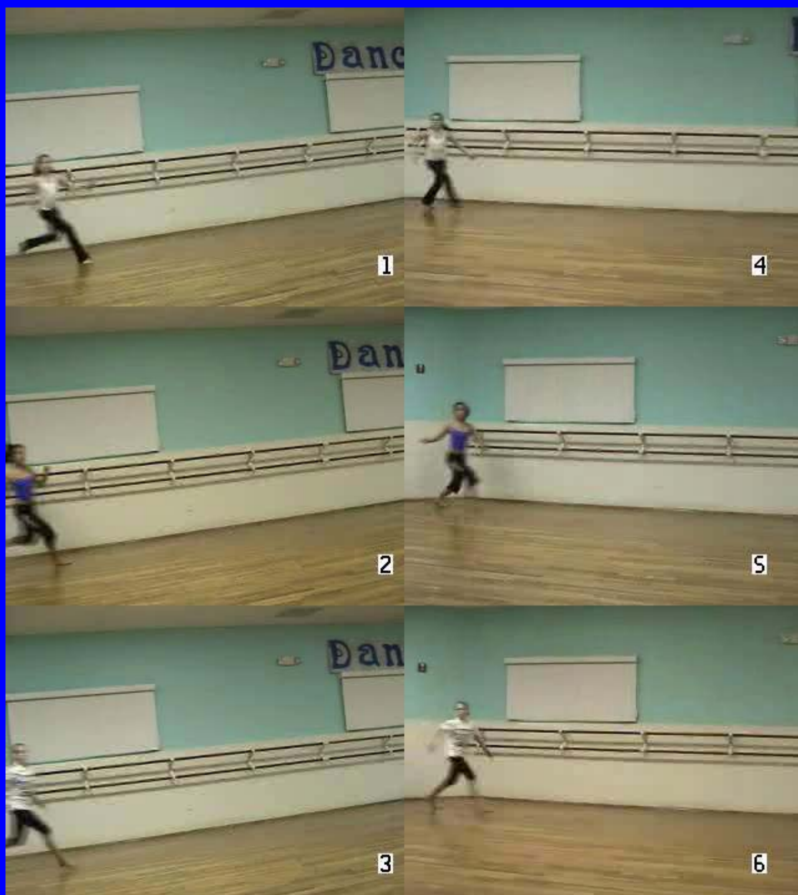


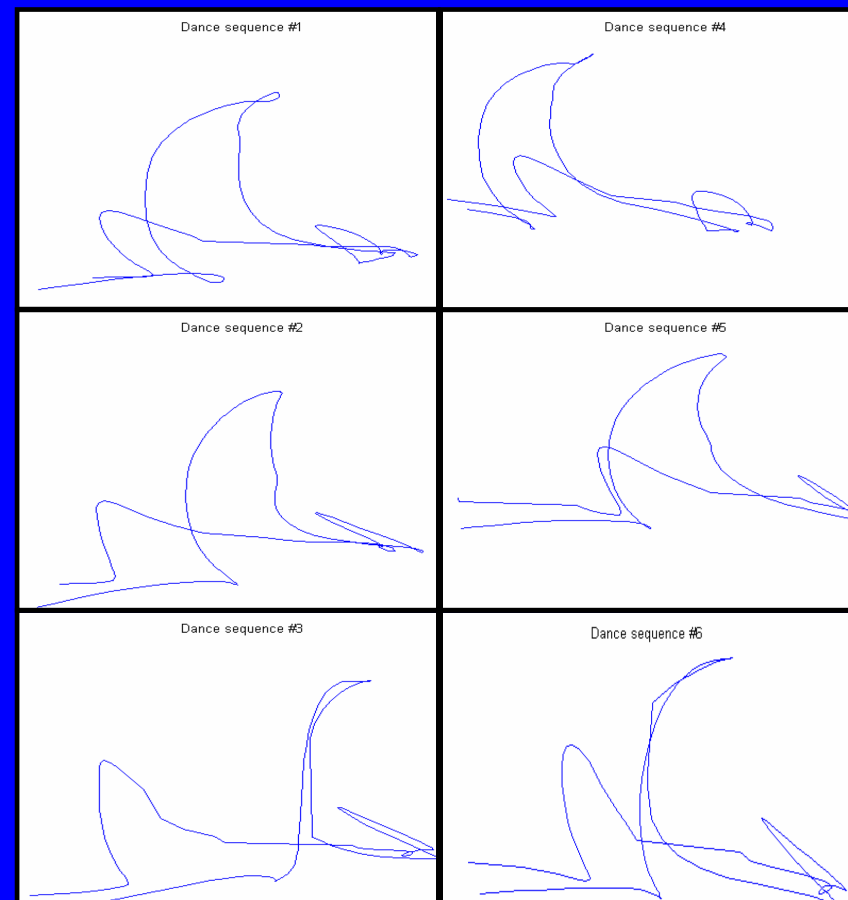
Experimental Results

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



Temporal Alignment of Videos





Trajectories of the right foot:

Input videos:

Temporal Alignment Results

Synchronized videos:

Temporal Alignment Results

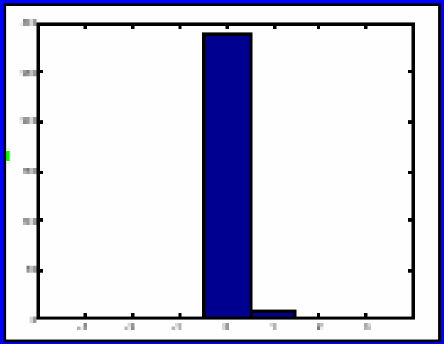
Before Temporal Alignment

After Temporal Alignment

Temporal Alignment Results

Non-overlapping video temporal alignment:



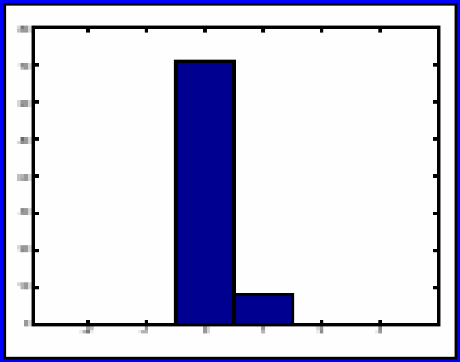


Left Seq

Right Seq

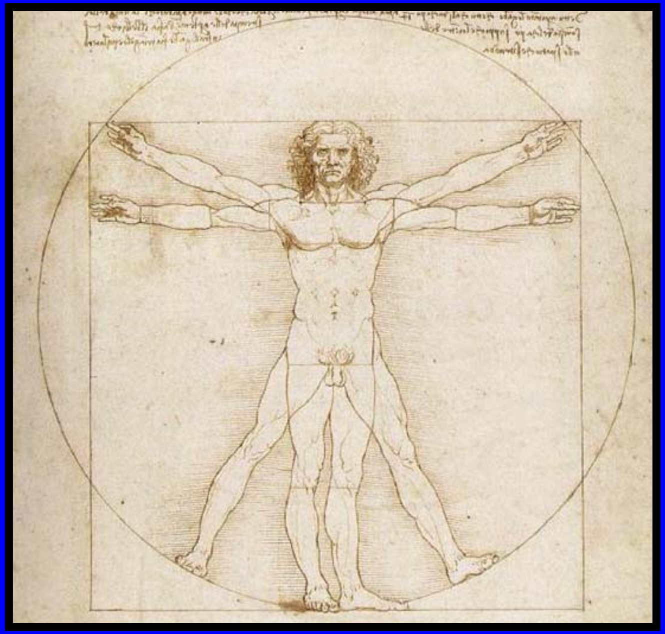
Histogram of Misalignment



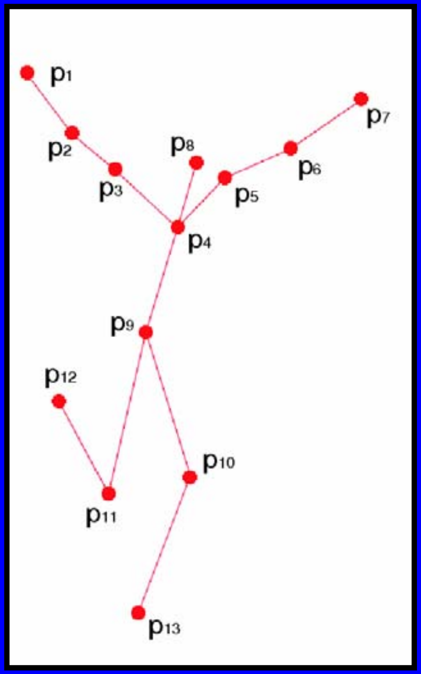


- Cen Rao, Alexei Gritai, Mubarak Shah, View-invariant Alignment and Matching of Video Sequences. The Ninth IEEE International Conference on Computer Vision, Nice, France, 2003.
  Project web page.
- Cen Rao, Alper Yilmaz, Mubarak Shah. View-Invariant Representation And Recognition of Actions, International Journal of Computer Vision, Vol. 50, Issue 2, 2002.

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

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.

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.**

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

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

Postural Constraint

- **Proposition 1:** If $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ 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.

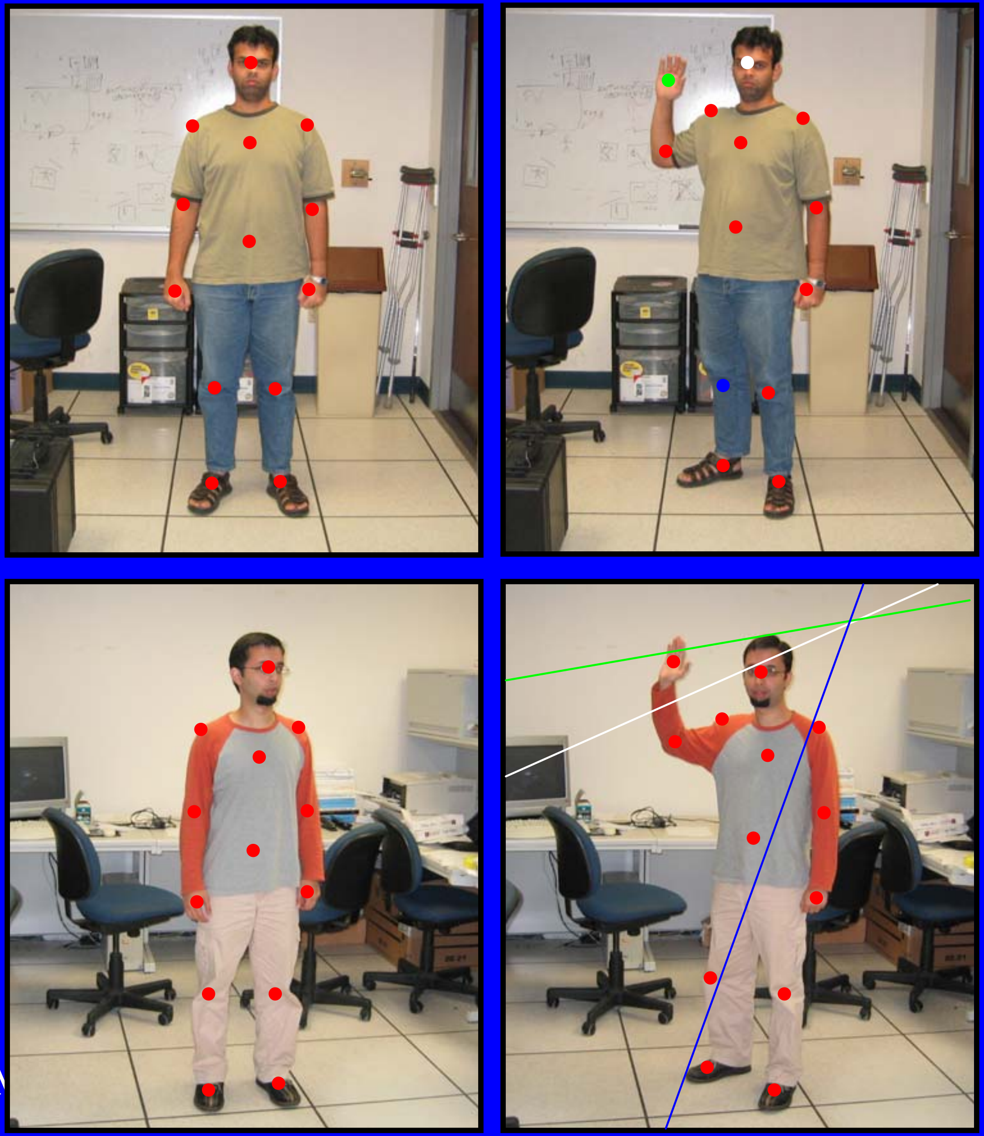
Postural Constraint

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

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

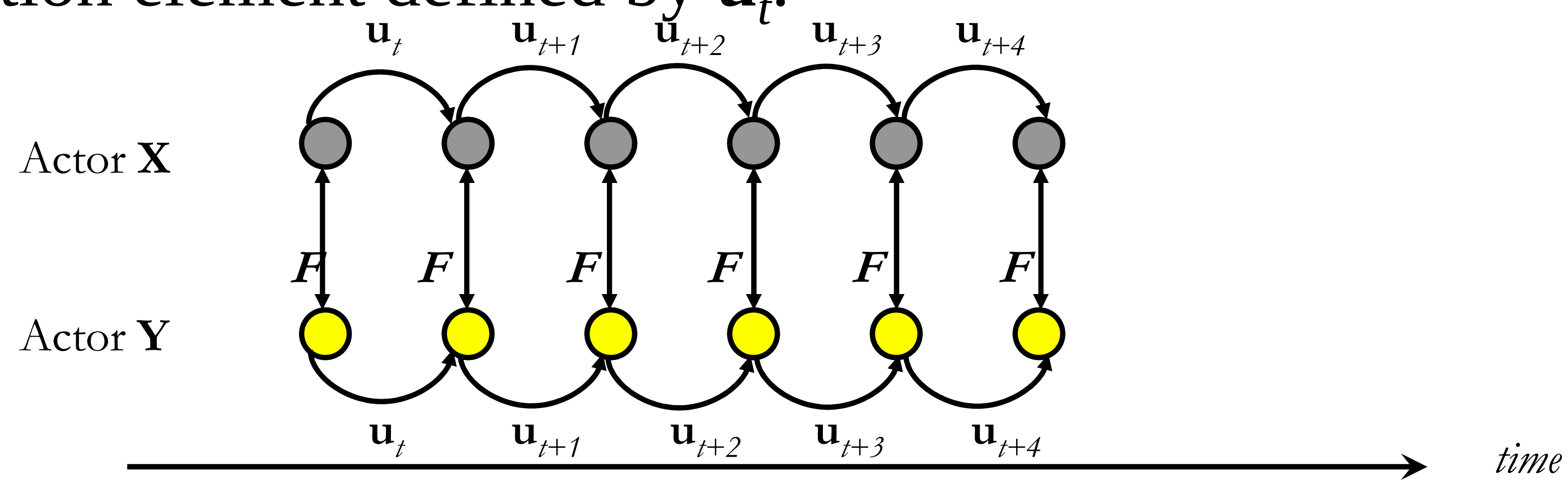
Capturing View Variance

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



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

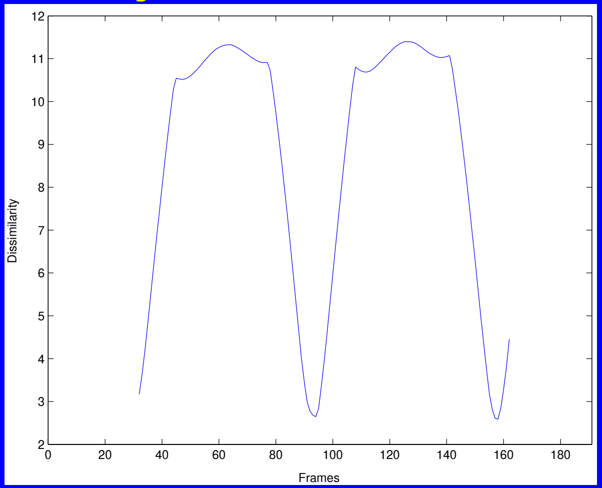
Analyzing Periodicity

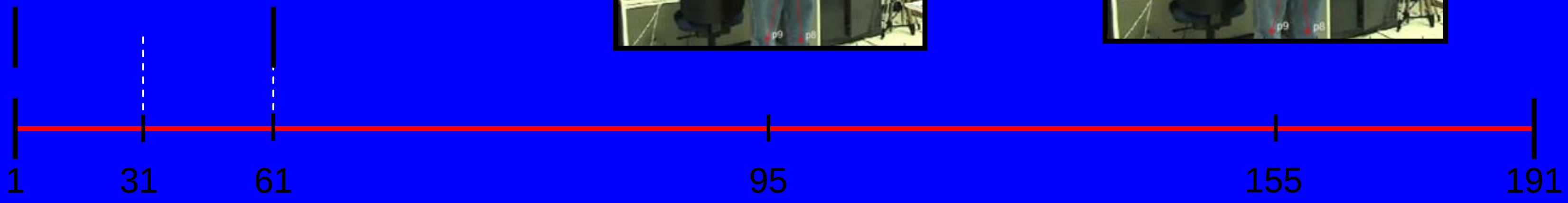
Reference Pattern

Test Sequence

Action Detection

Analyzing Periodicity



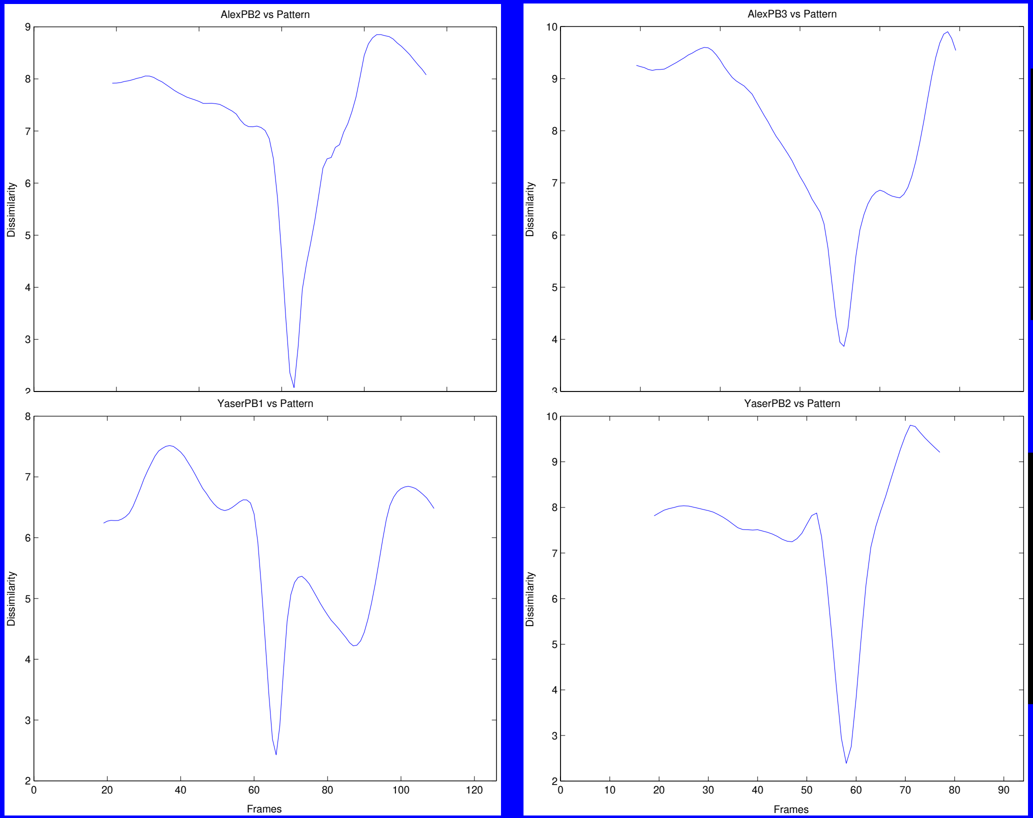


Action Detection: Different approaches, different people, the same action

ReferencePattern

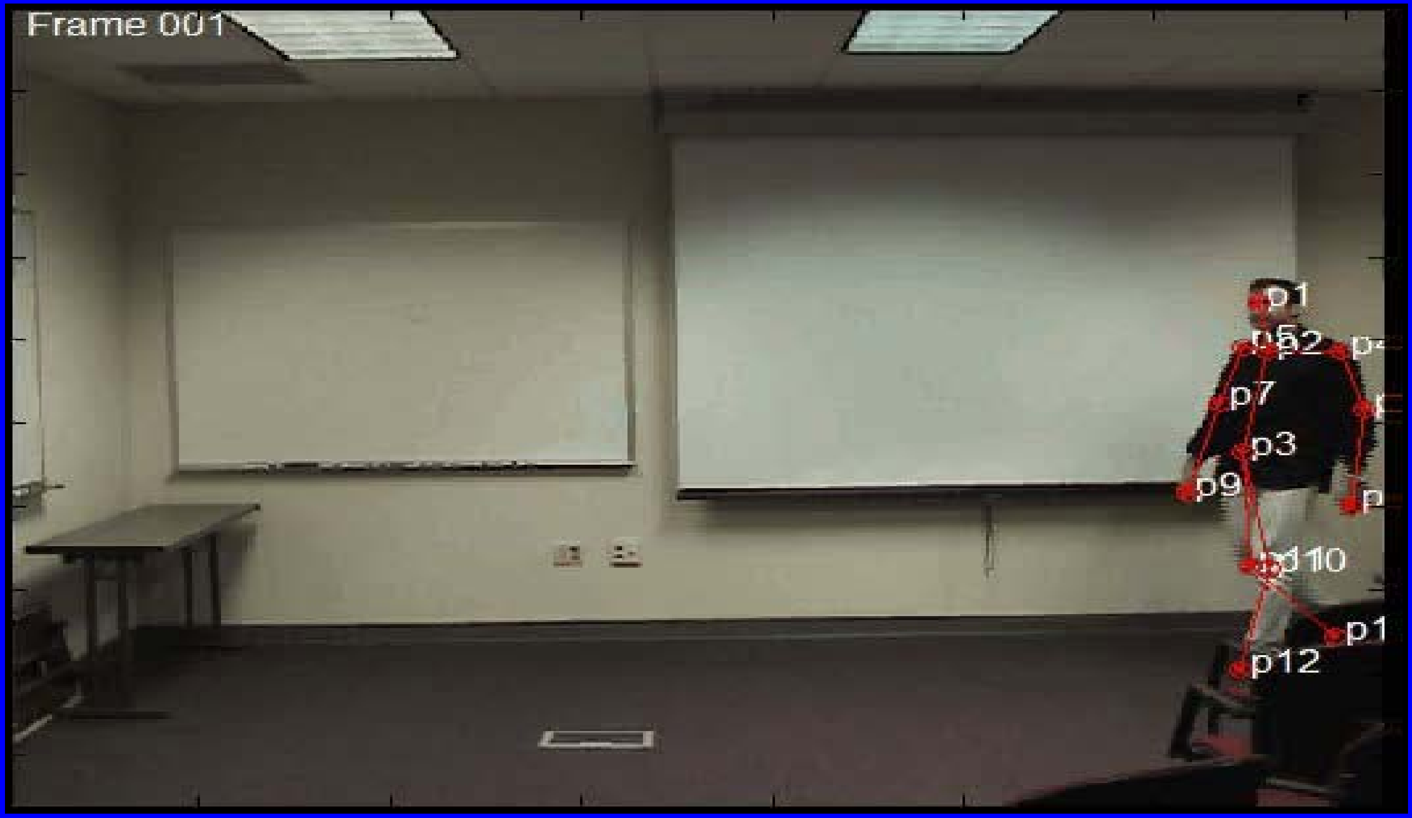
Test Sequences

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

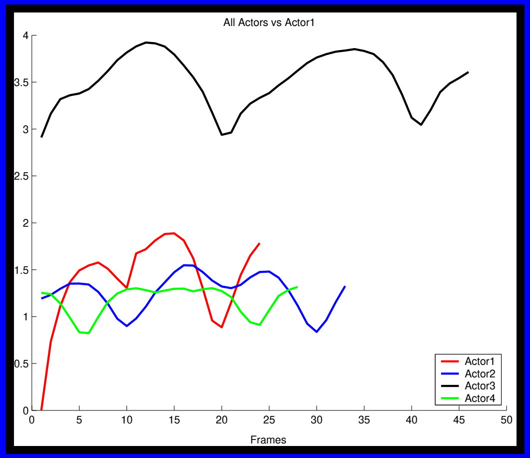


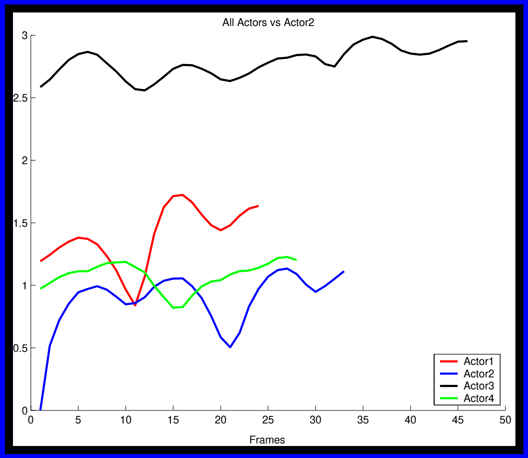
Analyzing Actions

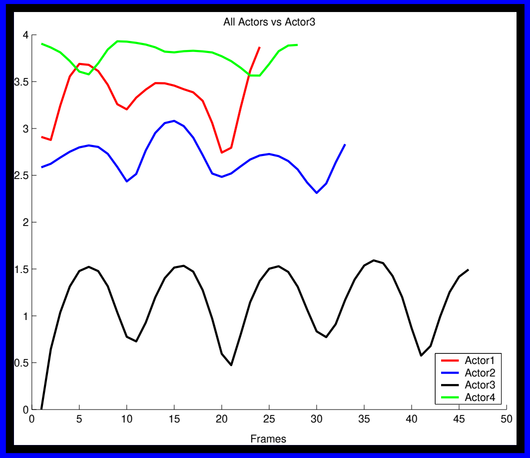
Odd One Out

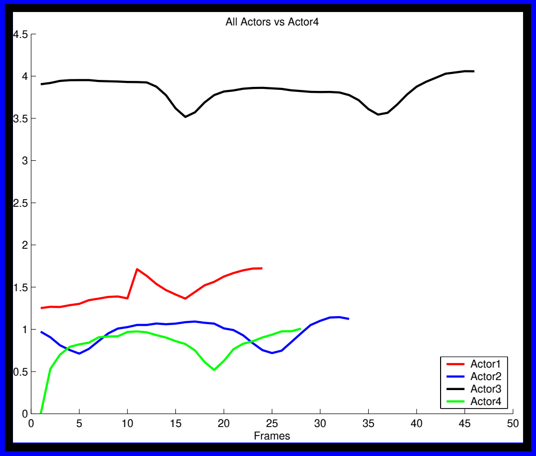


‘Odd One Out’


All Actors vs Actor1
Actor1
Actor2
Actor3
Actor4
Frames

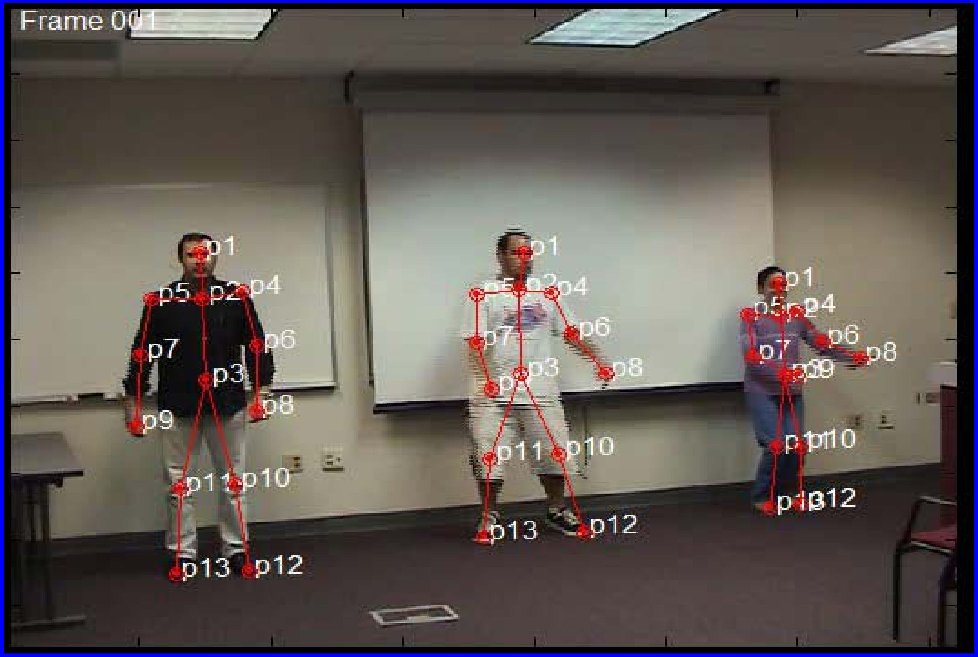

All Actors vs Actor2
Actor1
Actor2
Actor3
Actor4
Frames


All Actors vs Actor3
Actor1
Actor2
Actor3
Actor4
Frames


All Actors vs Actor4
Actor1
Actor2
Actor3
Actor4
Frames

Action Synchronization

Following the Leader



Action Synchronization

Following the Leader