Human Action Recognition
A Grand Challenge

J. K. Aggarwal

Department of Electrical and
Computer Engineering

The University of Texas at Austin

Austin, TX 78712

Overview

- Human Action Recognition
- Motion – An Important Cue
- Methodologies and Results by other Researchers
- Actions and Interaction.
  - Simple actions
  - Interaction - simple
  - Interaction - continued and recursive
  - Objects and actions
- Some Final Thoughts.

Human Action Recognition

Image Processing / Pattern Recognition / Computer Vision have matured from recognizing simple objects, textures and images to recognizing discrete actions, continuous action and sequences of actions – called events.This poses serious challenges and in the long run this may hold significant benefits.

Grand Challenges!

- A system that may detect a person having a heart attack in a hotel room or a system that may detect a person drowning at a swimming pool.
- Such systems would have universal applicability if the false alarm rate was really low.
- Research in human action recognition is at the embryonic stage, a successful system of this nature may require a mating of several disciplines including psychology and may be still far in the future.

Today's Research on Human Action Recognition

- Personal Assistants
- Surveillance
- Medical imaging
- Patient monitoring
- Analyzing sports and dance videos
- Robot/Human interaction
- Biometrics
- Kinesiology

Almost all of the above applications involve enterprise systems

System Components

- Low-level vision module – concerned with segmentation and other low-level tasks, there is no substitute for “good” segmentation!
- Modeling and recognition of the action and interactions, including motion of objects.
- Semantic description of the actions and interactions.

Motion - An Important Cue

- Detection
- Segmentation
- Tracking
- Deriving 3D information
- Structure from motion
- Point of impact
- Recognition

Study of Motion

- Long history of scientific thought spanning diverse disciplines:

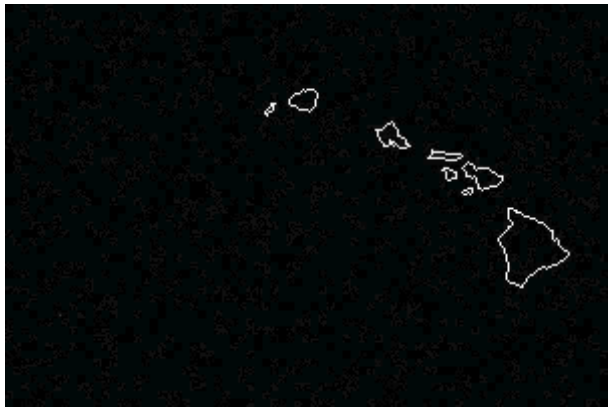
  philosophy, psychophysics, neurobiology, …and of course,

  pattern recognition, robotics, computer graphics and computer vision.

- Zeno, Aristotle, Bertrand Russell,……

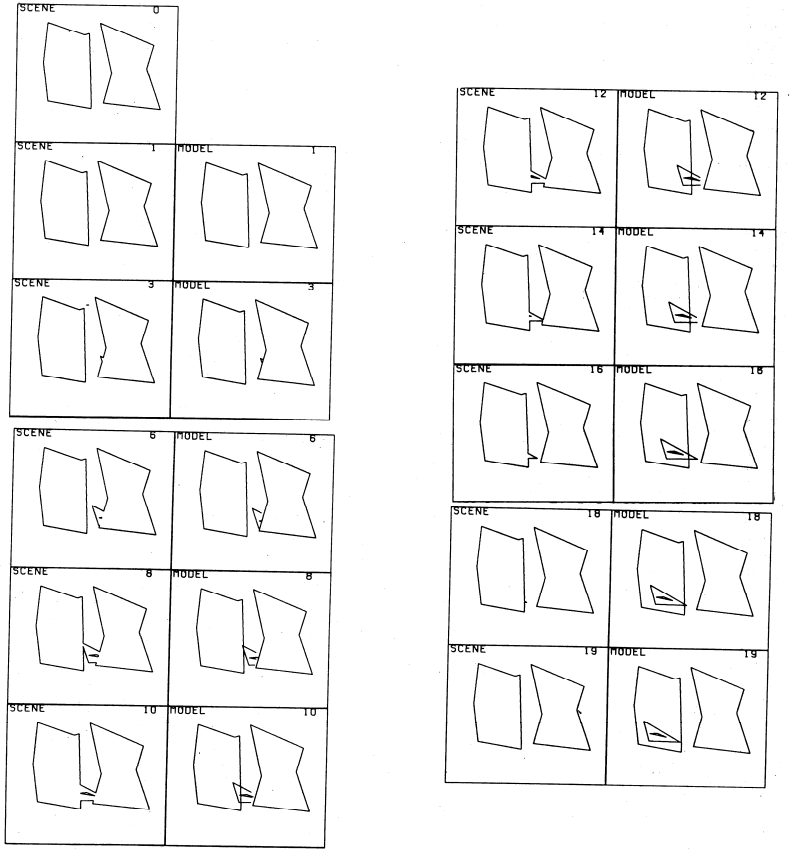
My Simple Beginnings in Motion Research (1970's)

- Cloud motions from satellite image sequences.
- Idealized clouds as rigid polygons moving in a plane (later to curvilinear figures).
- For the idealized cloud models determined linear and angular velocities.

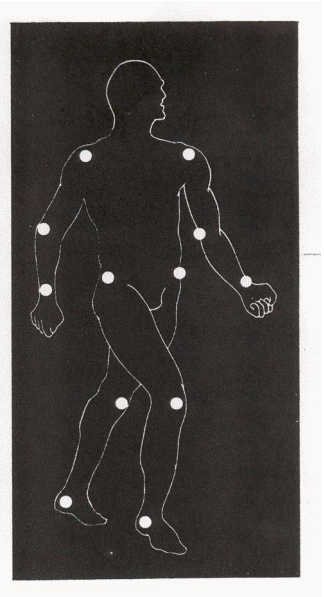


First Encounter with ‘Structure from Motion’

“Computer Analysis of Moving Polygonal Images”, IEEE Trans. on Computers Vol. C-24, no.10 October, 1975. (with R. O. Duda)



Motion of Rigid and Jointed Objects



- Johansson's experiments (1975) - lights attached to major joints of a person - inspired Jon Webb.
- Each rigid body part represented by two points.
- Webb decomposed such a motion into a rotation and a translation, where the rotation axis is fixed for short periods of time.
- Webb determined structure of jointed objects under orthographic projection.

Motion of Rigid and Jointed Objects, Contd.

- Hoffman data/MIT.
- Six points on a walking man, 0.26 sec.
- Artificial Intelligence, vol.19, 1982, 107-130.

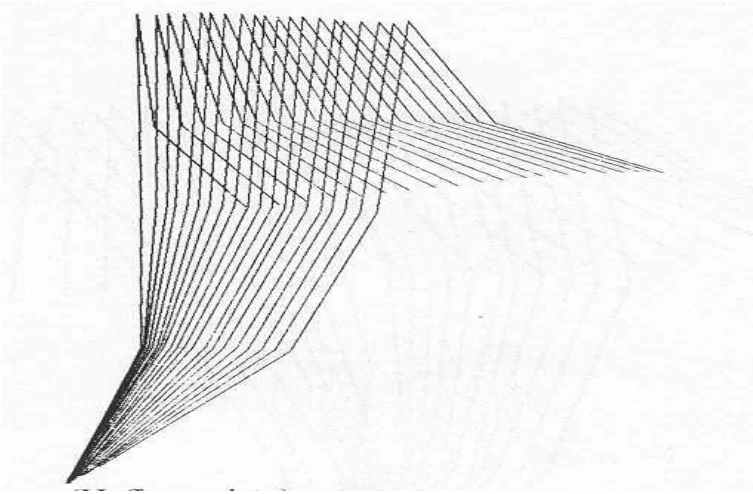
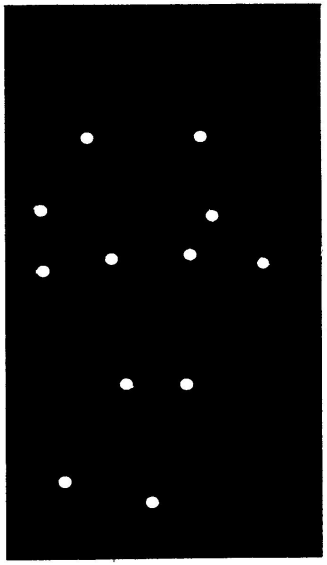


TABLE 1. Rigid part lengths (in meters); Hoffman data

From	To	Estimated length	Actual length	Relative error
shoulder	elbow	0.344	0.335	2.53%
elbow	wrist	0.283	0.274	3.35%
shoulder	hip	0.584	0.579	0.808%
hip	knee	0.438	0.437	0.175%
knee	ankle	0.435	0.437	1.90%

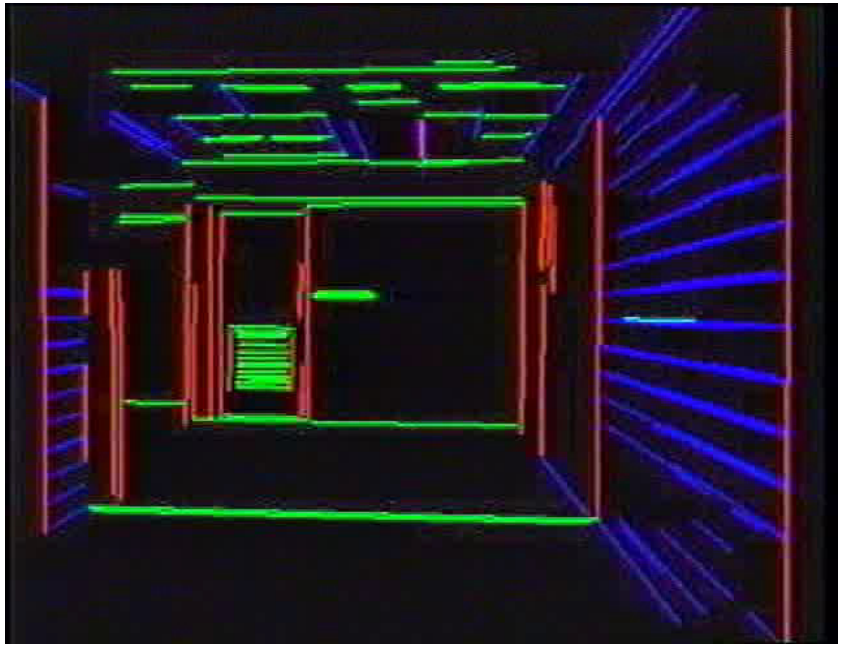
Trying to Avoid Establishing Correspondence of Points



- Emphasis on points in the computation of structure-from-motion.
- For points difficult to:
  - Find
  - Localize
  - Occlusion
  - Correspondence
- A natural question: Can one do better with lines, planes and surfaces?

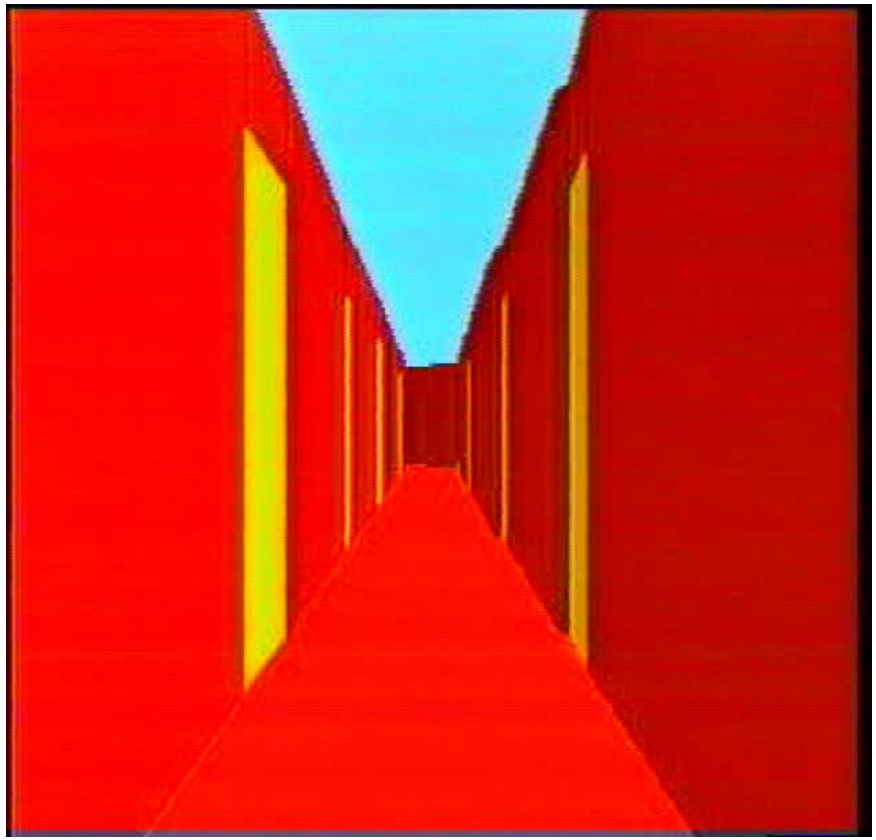
Lines

- Lebegue (1992) chose lines with particular orientation, using vanishing points.
- Tracked 3D line segments using a Kalman filter.
- Built a CAD model of a corridor by mounting a single camera on a tether-less robot.



Results

(IEEE Trans. On Robotics and Automation, vol.9 no.6, 1993, 801-815)



Recent Progress

- Today, research has progressed from the ‘structure from motion’ paradigm to tracking and recognition of events – human actions, human-human interactions and human-object interactions
- Computational speed, cheap storage and cameras.

Issues in Human Actions/ Interactions Recognition

- Diversity
  - Actions - walking, running, throwing etc
  - Interactions - ranging from hugging, kissing and shaking hands to punching and killing.
  - Interactions between objects and persons
- Occlusions, correspondence difficulties due to loose clothing and shadows, reflections and lighting conditions.
- Besides the computer vision issues, a number of contextual and philosophical issues – persons can interact by not interacting.

Human Motion/Body modeling

- Different types of motion:
  - *Rigid* vs. *non-rigid* motion
- Human motion is *articulated* motion, a subset of non-rigid motion.
  - Composed of piecewise rigid motions of individual body parts, but the overall motion of the entire human body is not rigid.
- Model-based vs. appearance-based
  - Well-defined a priori knowledge of the object shape available or not.

High-level recognition schemes with domain knowledge

- Human-involved scene understanding needs more abstract and meaningful schemes than purely physical laws for interpreting *'what is happening in the scene'*.
- Understanding very long image sequences requires another abstraction scheme: the '*event'*.
- The event is regarded as a sort of summary of the whole sequence, and the summary is closely related to real world knowledge.

Action, Interaction and Event: A Taxonomy

- Nagel (1988): Change, Event, Verb, Episode and History.
- Bobick (1997): Movement, Activity and Action.
- Park and Aggarwal (2003):Pose, Gesture, Action, Interaction.

A Bayesian Computer Vision System for Modeling Human Interactions

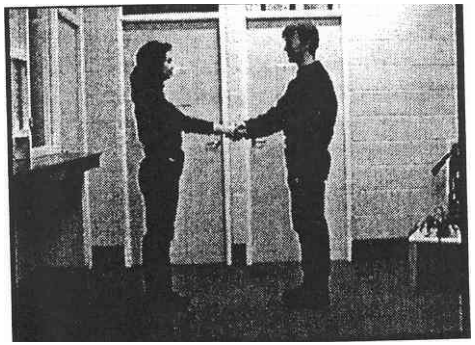
Nuria Oliver, Barbara Rosario, and Alex Pentland

- Objective: Monitor person-to-person interactions in perspective-view image sequences
  - Pedestrians walk in wide area, Humans move 2-dimensionally in image sequences
  - Blob level tracking (track a human as one blob)
  - Approach, Meet, Walk, etc.
- Method: Trajectory shape recognition
  - Coupled Hidden Markov Model
  - Recognize sequences of state change
  - Multiple state streams

Interaction Behavior Models

N. Johnson, A. Galata, and D. Hogg

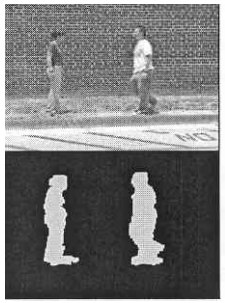
- The acquisition of interaction behaviors by observing humans and using models to simulate a plausible partner during interaction.
- A Spline type contour to specify the persons.



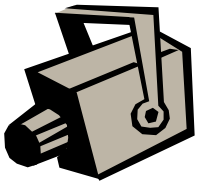
Person-on-Person Violence Detection

A. Datta, M. Shah and N. Lobo

- Detect human violence such as fighting, kicking, hitting with objects.
- Motion trajectory and orientation of person's limbs.
- AMV- Acceleration Measure Vector- direction and magnitude of motion-'jerk' temporal derivative of AMV.



Segmentation and Recognition of Continuous Human Activity



Anjum Ali

Preprocessing :

Segmentation and skeletonization

sequence frame

background

thresholded image

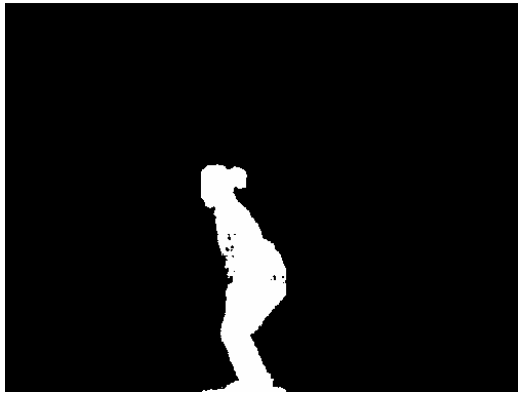
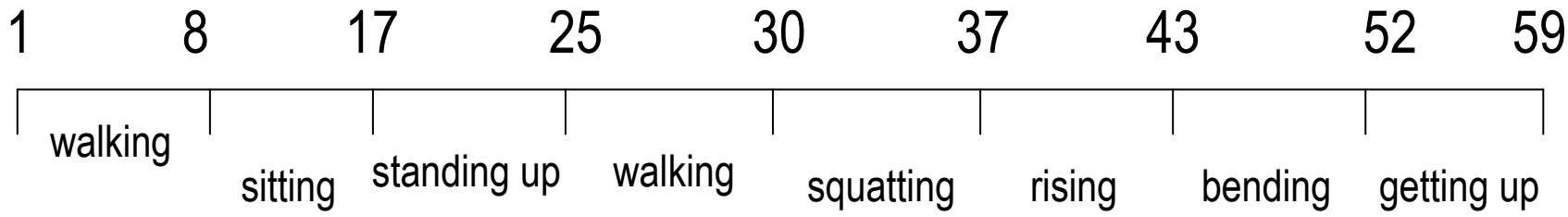


Image after filtering and connected component labeling


1
8
17
25
30
37
43
52
59
walking
sitting
standing up
walking
squatting
rising
bending
getting up

Results

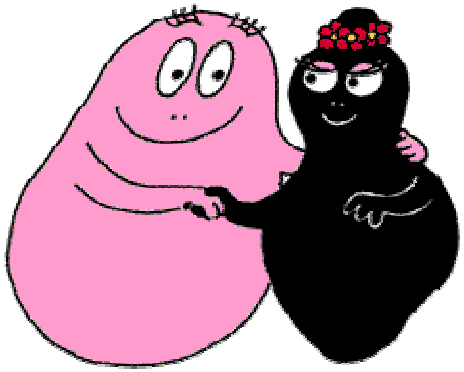
Number of	Total	Correct	Efficiency
Breakpoint frames	128	109	85
Actions	143	110	76
Walking	29	26	89
Sitting	13	10	76
Standing up	16	12	75
Bending	21	15	71
Getting up	19	14	73
Squatting	24	18	75
Rising	21	15	71

Event 2001, Vancouver, BC, pp.28-35

Blob-level Interaction of Persons

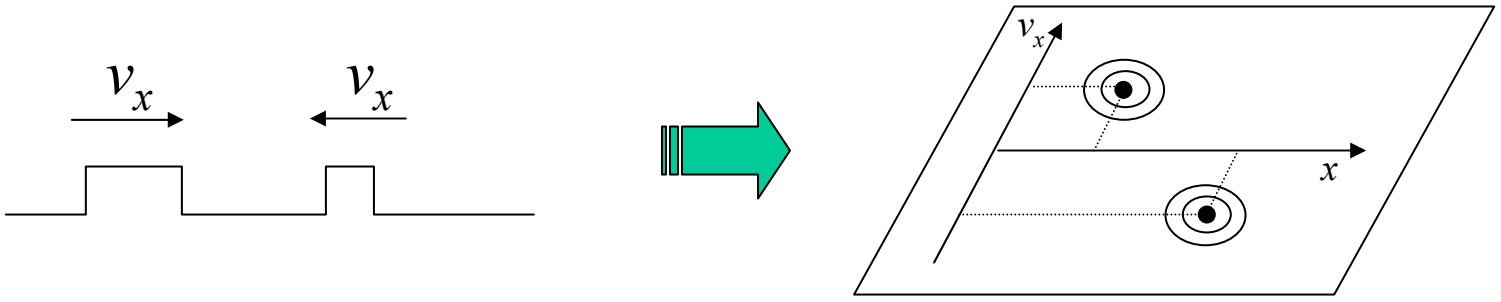
1. Temporal Spatio-Velocity (TSV) Transform
2. Tracking of blobs
3. Interaction Recognition

Koichi Sato



Temporal Spatio-velocity (TSV) Transform

- Extract pixel velocities from an image sequence
  - Remove the velocity-unstable pixels
  - Segment blobs by velocity similarity



Tracking

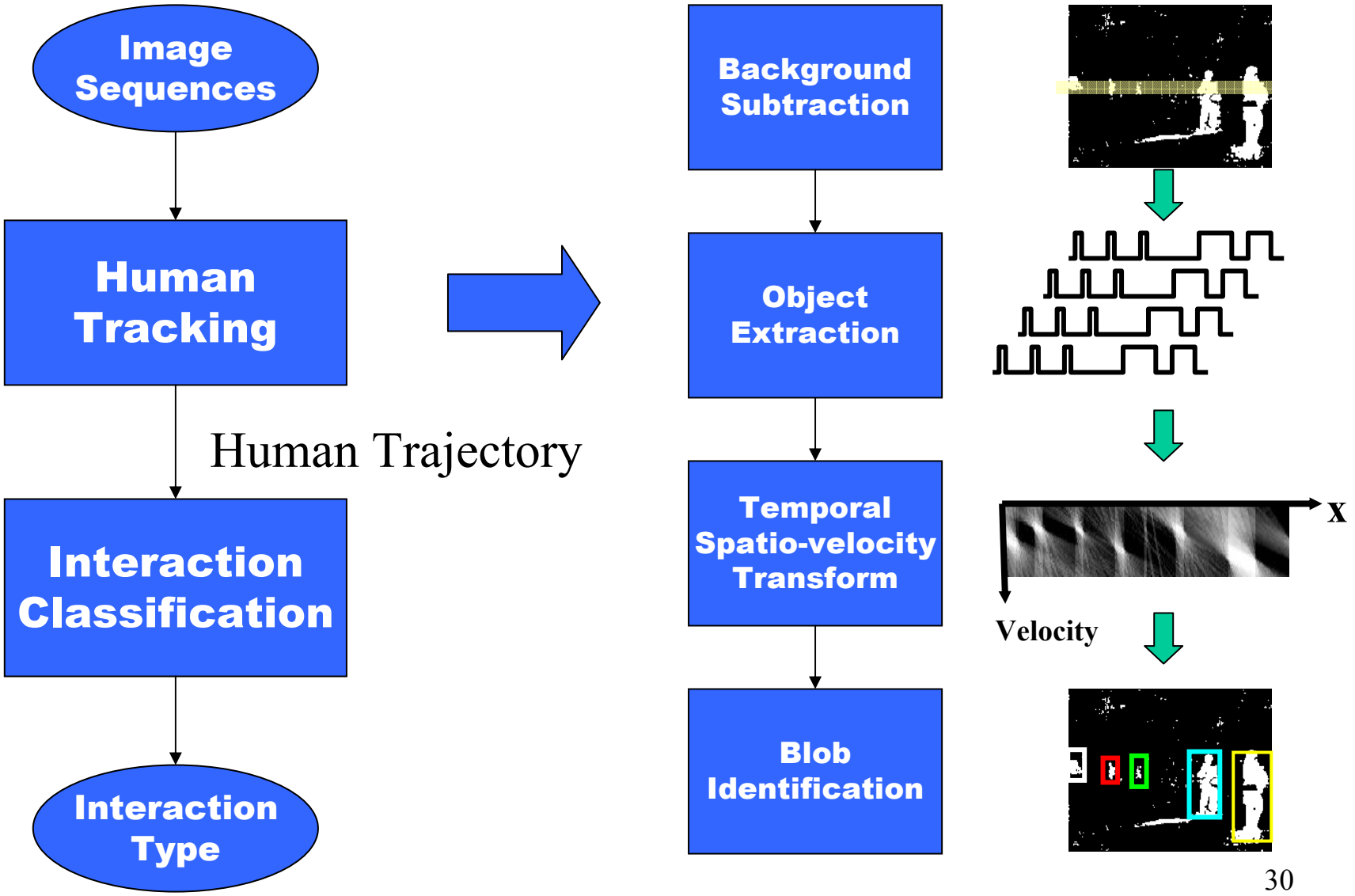
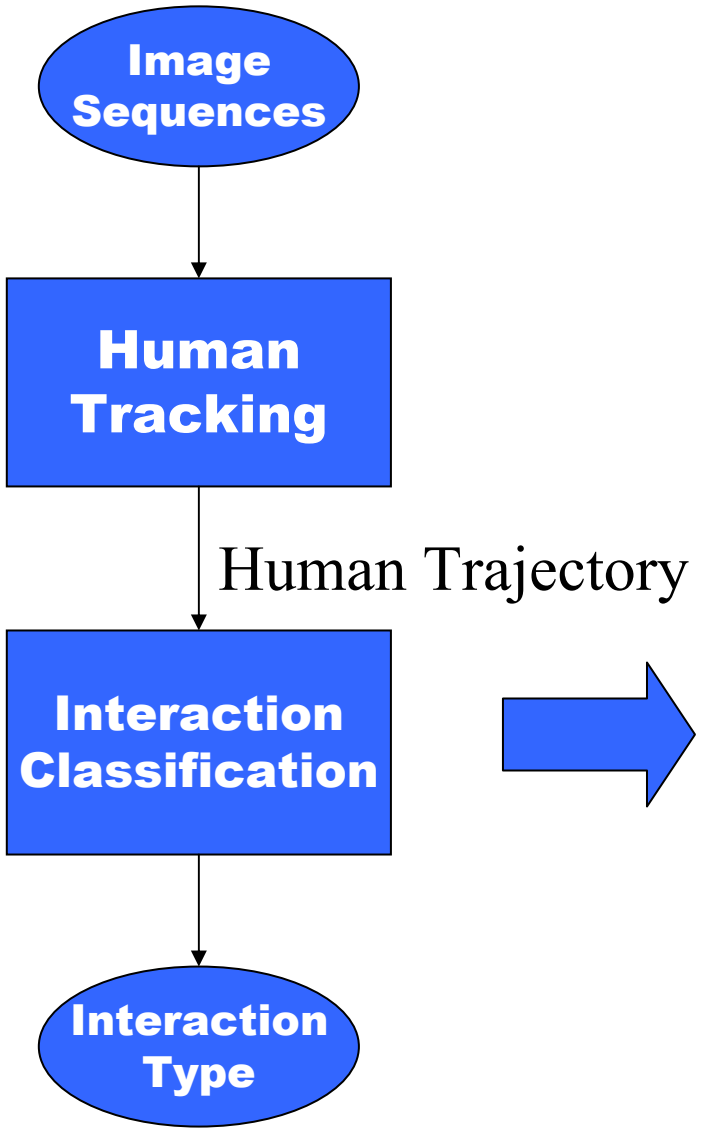
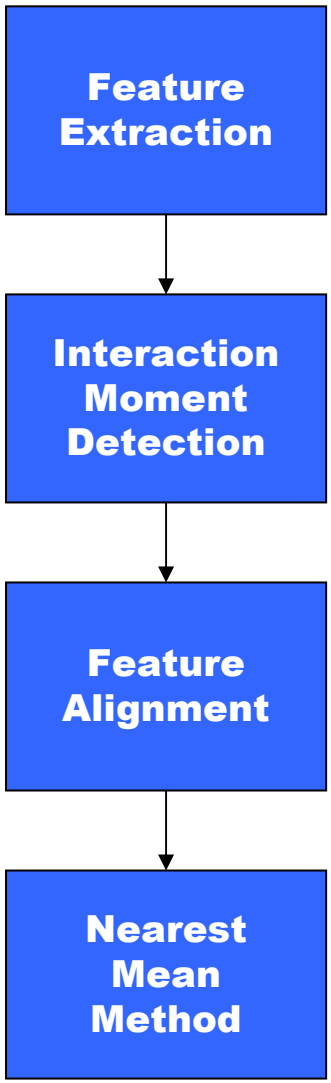


Image Sequences
Human Tracking
Interaction Classification
Interaction Type
Human Trajectory
Background Subtraction
Object Extraction
Temporal Spatio-velocity Transform
Blob Identification
x
Velocity

Interaction





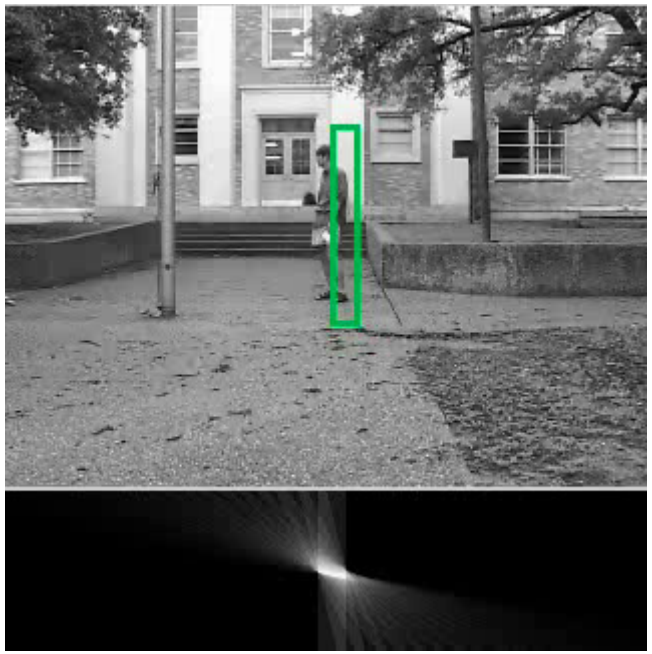
Person2 \ Person1	$S_S \rightarrow S_M$	$S_M \rightarrow S_S$	S_S	S_M
$S_S \rightarrow S_M$	APART			
$S_M \rightarrow S_S$	STOPGO	MEET		
S_S	LEAVE	STOP	STANDSTILL	
S_M	FOLLOW	WALKSTOP	PASS1	PASS2 TAKEOVER

Simple Interaction Unit (SIU) types

S_M: Moving State

S_S: State of Being Still

Follow



Meet and Leave

Results

- Human segmentation: 184 out of 194, only 10 persons lost.
- Tracked 159, correctly tracked 123, lost in tracking 5, wrongly tracked 31.
- Interaction type, training 54, test 70, correctly classified 56, classification accuracy 80%.

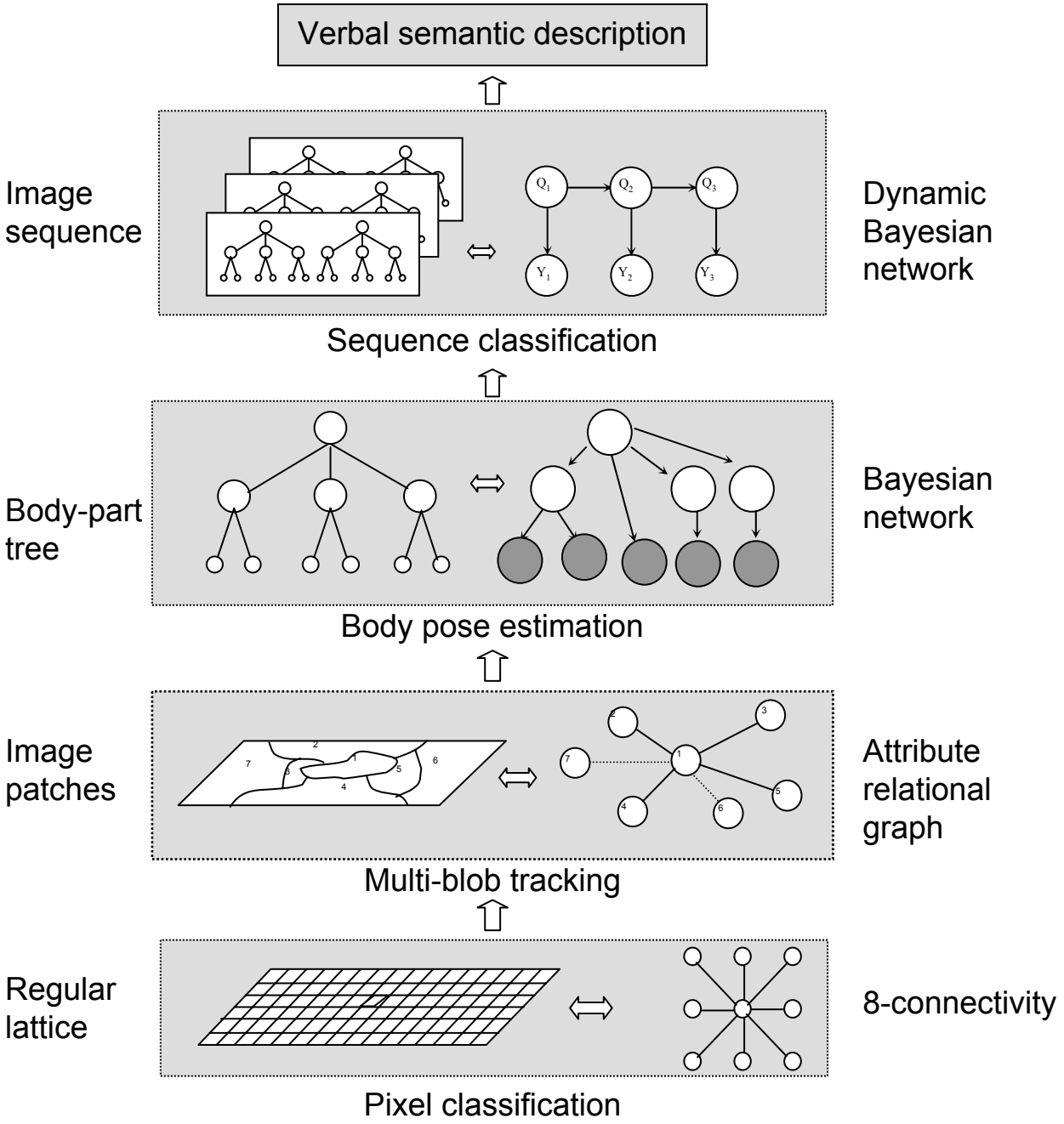
(Computer Vision and Image Understanding 96 (2004) 100-128.)

Interaction At the Detailed Level

‘Pushing’ sequence

Problems involved:

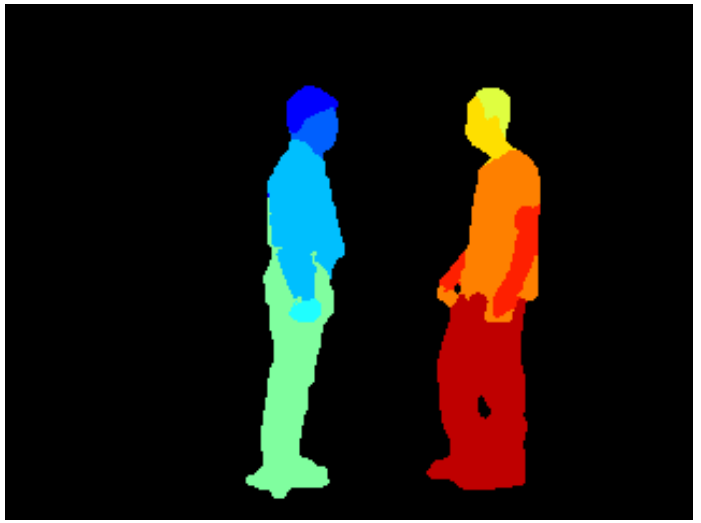
- segmenting multiple body parts
- tracking body parts
- treatment of occlusion and shadows between body parts
- recognizing human interaction type



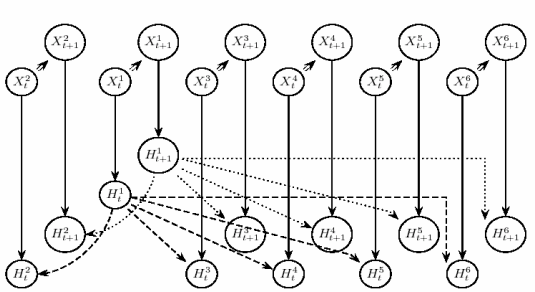
A hierarchical framework for recognizing human interaction

Results of segmentation

- Hugging

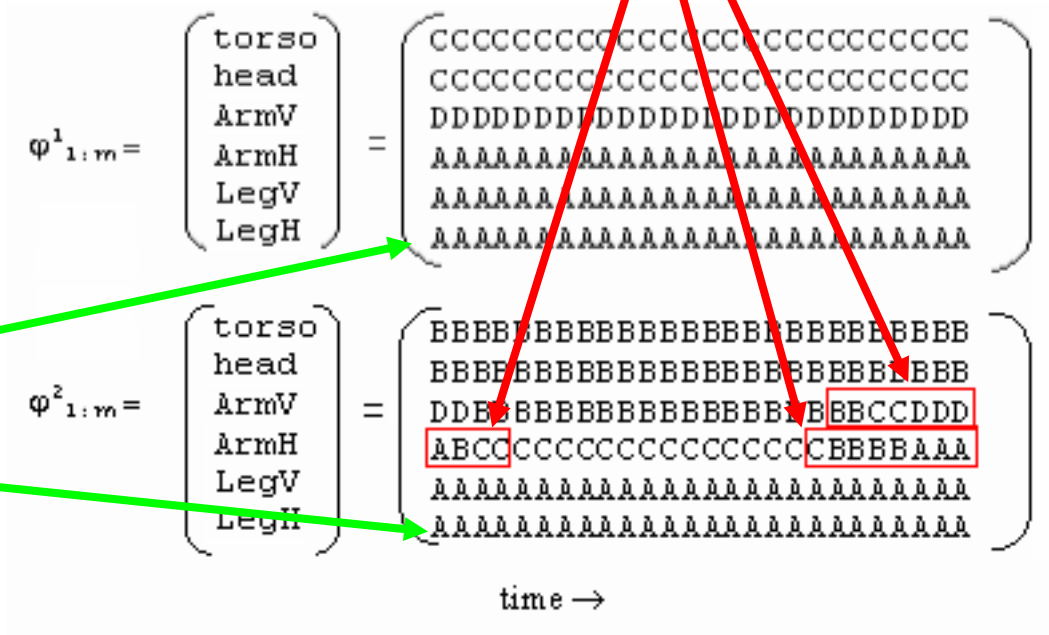


Gesture Detection using DBN

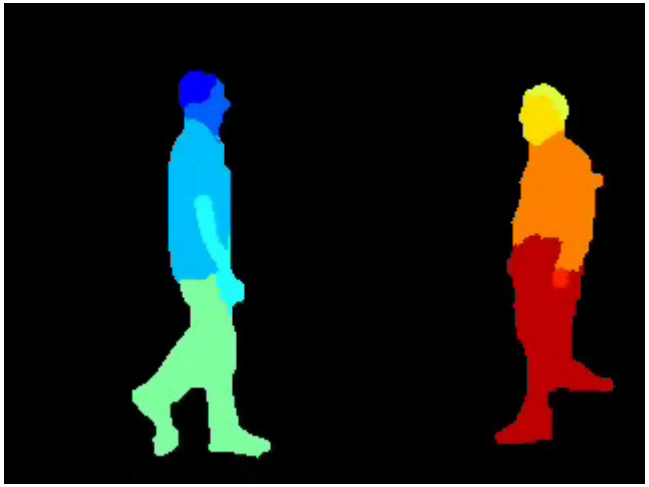


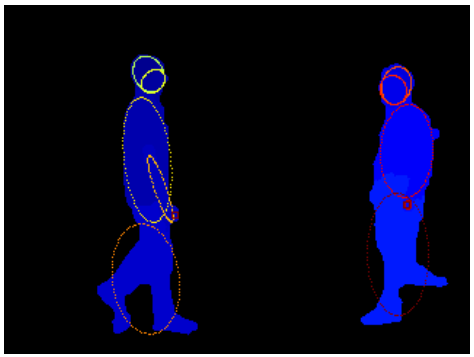
Pose Estimation using BN

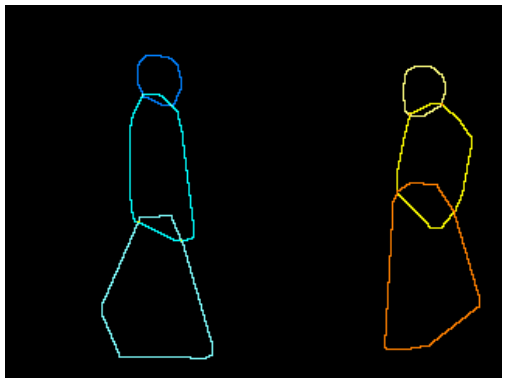
Observation $e = [V_1, \ldots, V_{11}]^T$ ➡ **BN** ➡ **Pose estimation** $\phi = [H_1, \ldots, H_6]^T$



Results: ‘Punching’ Interaction





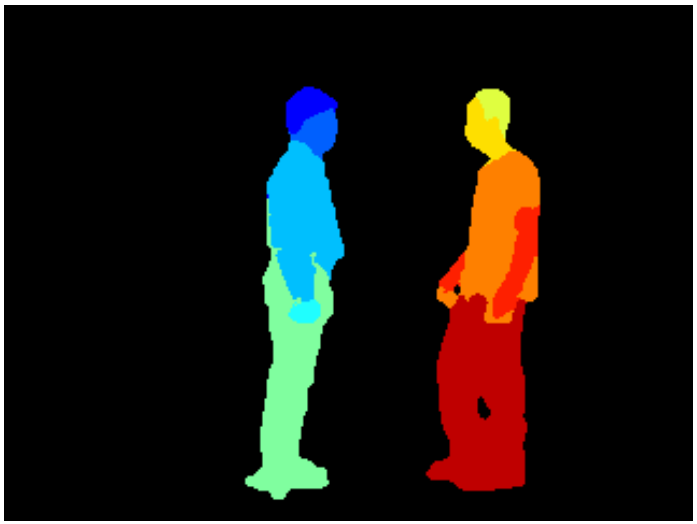


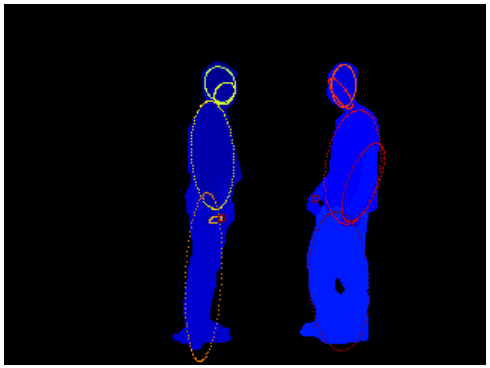
‘Punching’ sequence

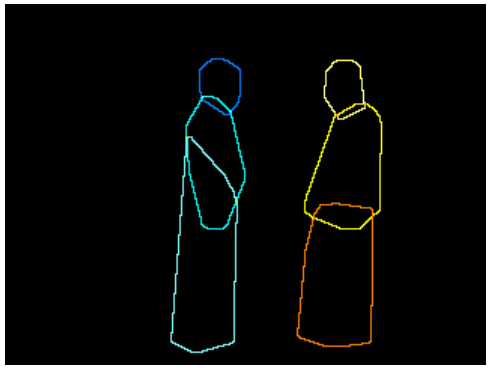
Person 1 <torso1-translate left-null> time

Person 2 <torso2-translate left-torso1> <arm2-stretch-head1> <arm2-withdraw-null> time

Results: ‘Hugging’ Interaction







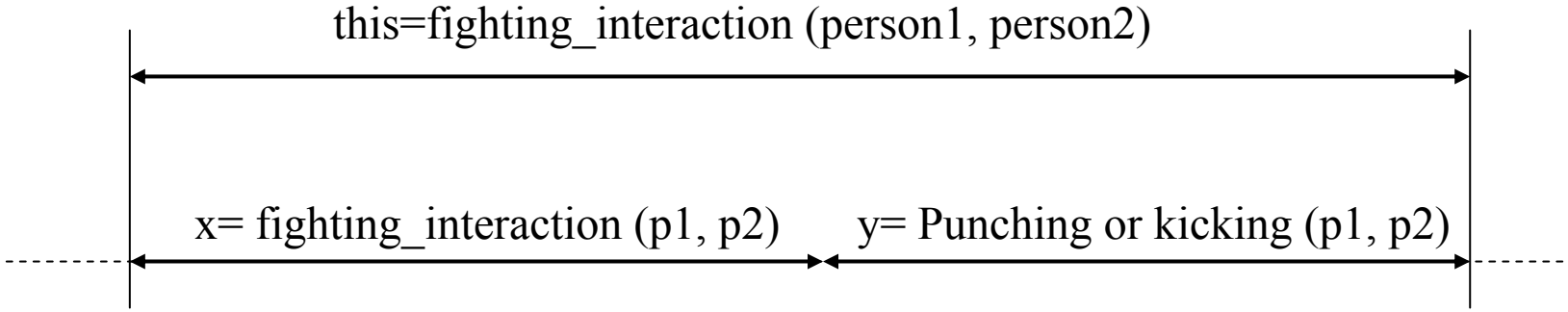
Recognition Accuracy

index	interaction	total	correct	accuracy
(a)	approach	6	6	1
(b)	depart	6	6	1
(c)	point	6	4	0.67
(d)	hand-in-hand	6	5	0.83
(e)	shake hands	6	6	1
(f)	hug	6	3	0.5
(g)	punch	6	4	0.67
(h)	kick	6	5	0.83
(i)	push	6	3	0.5
total				0.78

ACM Multimedia Systems 10(2), 164-179, 2004

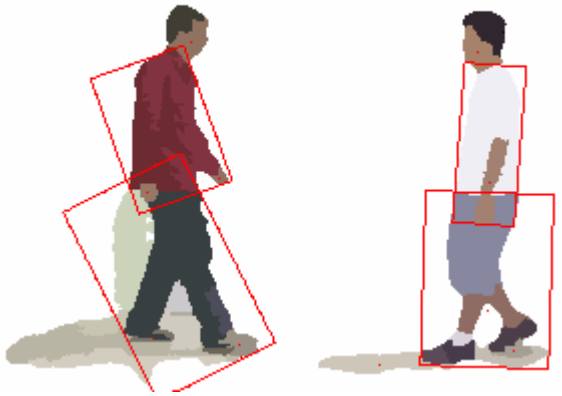
Recursive human activities

- Recursive representation will allow us to describe higher-level abstract human activities.
- For example, in ‘fighting’ interaction,
  - The number of sub-activities is not fixed. ‘punching’ or ‘kicking’ after some ‘fighting’ is also ‘fighting’.
  - ‘fighting’ can be defined in terms of smaller ‘fighting’

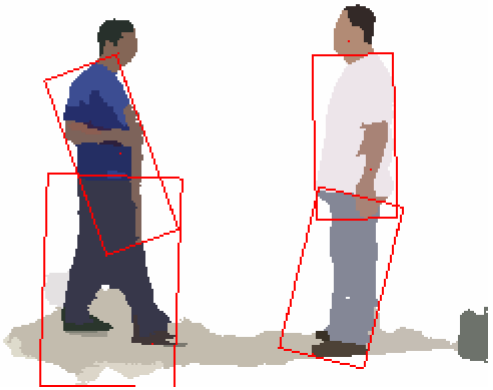


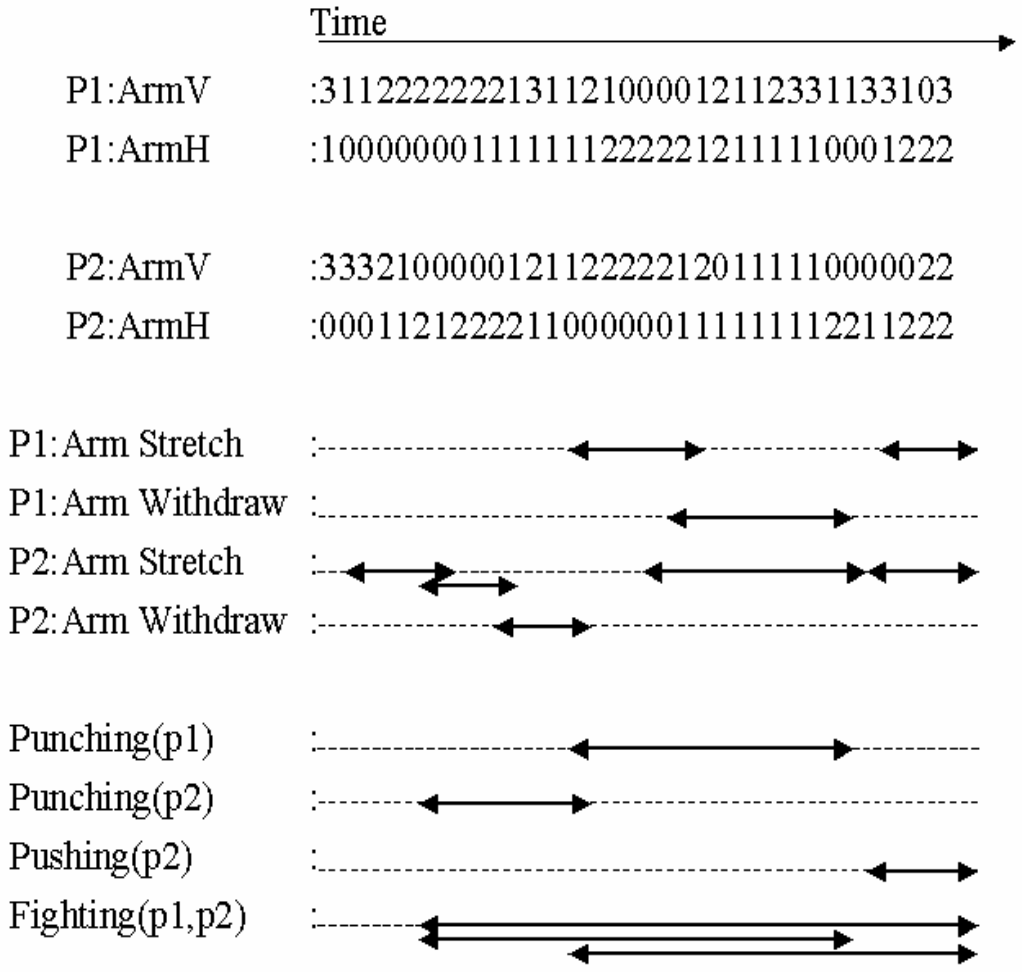
Experimental results

- Recognized following 10 types of interactions
  - (approach, depart, point, shake-hands, hug, punch, kick, and push) + (**Fighting** and **Greeting**).
- Videos of **sequences** of interactions are taken.
  - 320*240 resolution, 15 frames per sec.
  - Features are extracted for each image frame.



Experimental results (Cont'd)





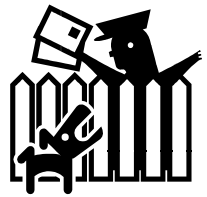
Experimental results (Cont'd)

interaction	total	Correct	Accuracy
Approach	12	12	1.000
Depart	12	12	1.000
Point	12	11	0.917
Shake-hands	12	11	0.917
Hug	12	10	0.833
Punch	12	11	0.917
Kick	12	10	0.833
Push	12	11	0.917
total	96	88	0.917

interaction	total	Correct	Accuracy
Fighting	6	4	0.667
Greeting	6	4	0.667
total	12	8	0.667

Detection of Fence Climbing from Monocular Videos

Elden Yu



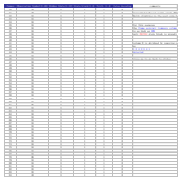
Difficulty

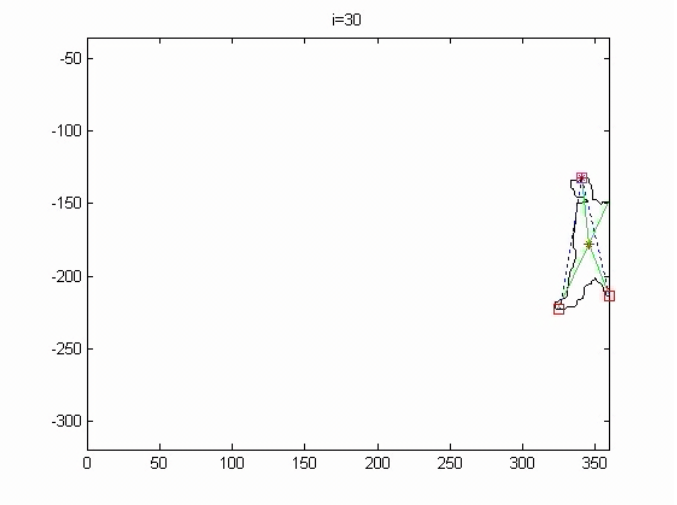
- Variability of fences
- Variability of climbing

Assumptions

- Fixed type of fences
- Single person
- Three possible views: front, back and side

Example





Some Final Thoughts

- A long way from tracking planar polygonal objects in 1975.
- Interaction of two persons is significantly more complicated than tracking persons.
- The complexities of occlusion and correspondence probably shall never leave us!!
- Does the computer “see and track” what we perceive?!

Thank You

- PICASSO
  “Figuras Al Borde Del Mar”
  “Figures By The Sea Shore”