Presentation of the Ph.D. thesis entitled

Human Pose and Activity Recognition from Stereo Images Using Probabilistic Parametric Inference

by **Nguyen Duc Thang**

Thesis advisor: Prof. Young-Koo Lee

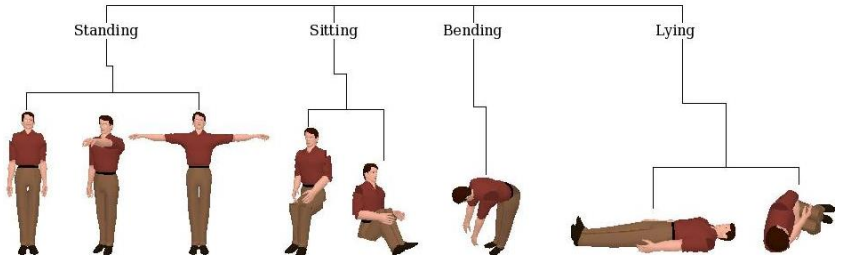
Ubiquitous Computing Laboratory
Kyung Hee University, Korea

June 2011

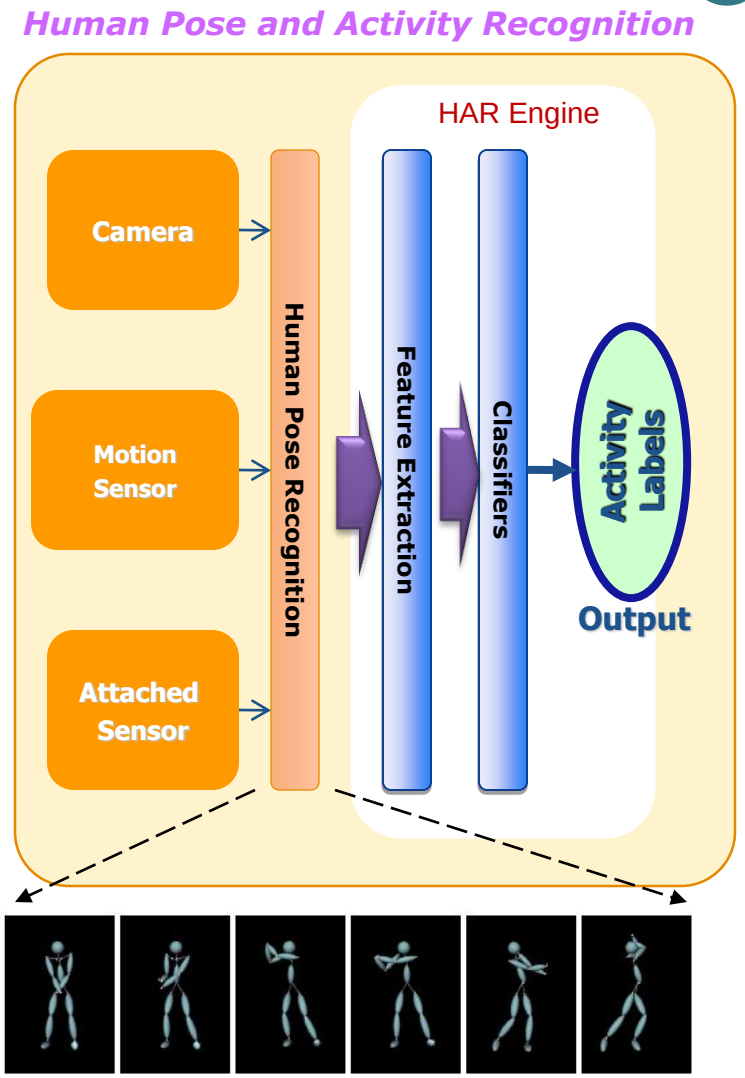
Contents

Human Pose and Activity Recognition

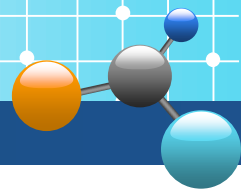
- **Human pose** presents a configuration of a human body in 3-D



- **Human pose recognition** aims to recover human body poses using data acquired by external sensors
- **Human activity recognition** (HAR) aims to recognize activities (i.e., patterns of movements of the human body) of a person
- Human pose recognition can provide info about human body configurations changing over time to distinguish different human activities



Human Pose Recognition: Wearable Sensor



Human Pose Recognition

Non-optical Based with Wearable Sensor

METAMotion

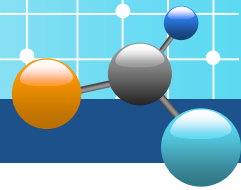
xsens

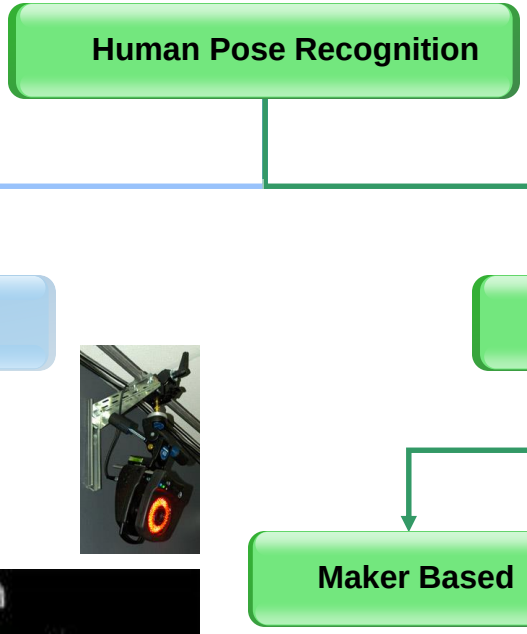
Features:

✓ Able to operate indoors or outdoors

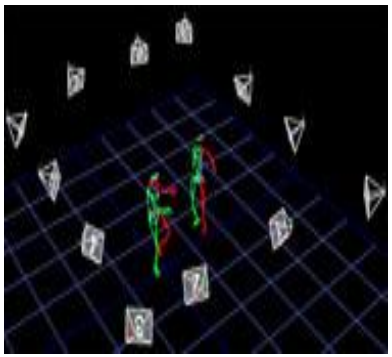
✓ Sensors are attached on an exoskeleton, a suit or a human body

Human Pose Recognition: Marker Based





VICON system uses optical markers

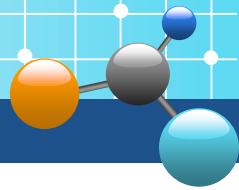


Special cameras are utilized to record motion of markers, reconstructing human body poses

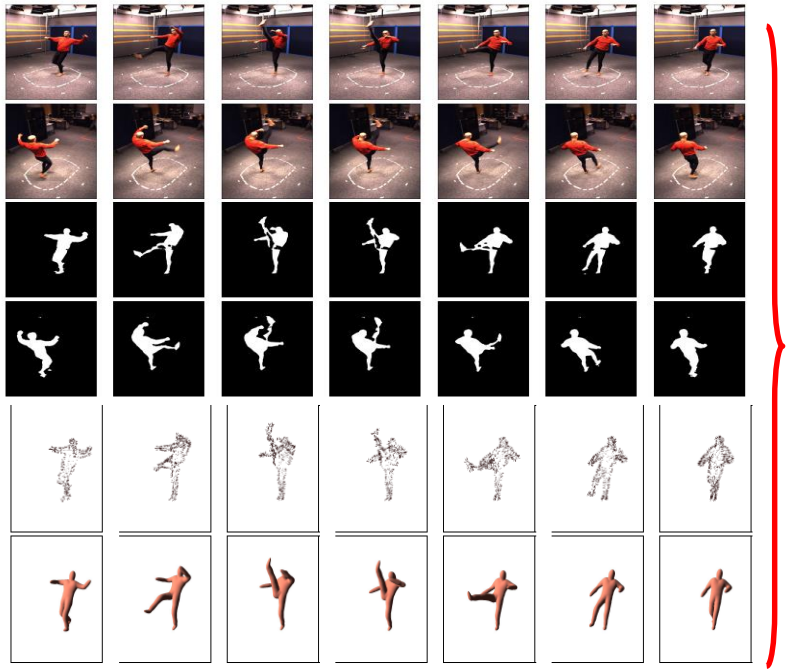
Features:

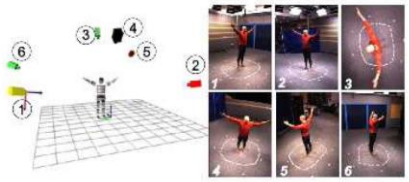
- ✓ Fast processing speed and high accuracy
- ✓ Expensive equipment (most systems)

Human Pose Recognition: Multiview Based



- Human Pose Recognition
  - Non-optical Based with Wearable Sensors
  - Optical Based (Video Based)
    - Maker Based
    - Makerless Based
      - Multi-view Based



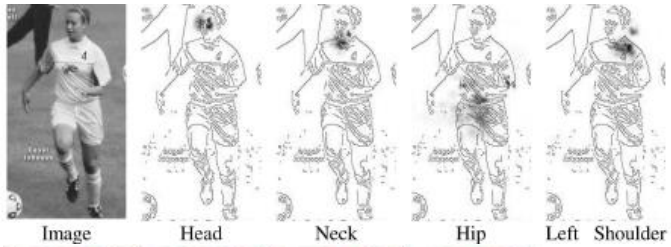


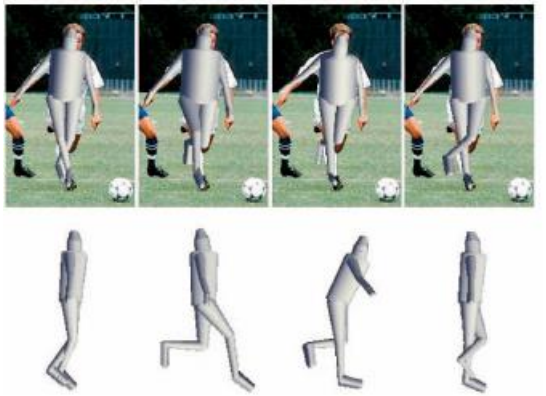
Features:

✓ Able to reconstruct both human body shapes and poses

✓ Complicated setups with calibration

Human Pose Recognition: Monocular Camera



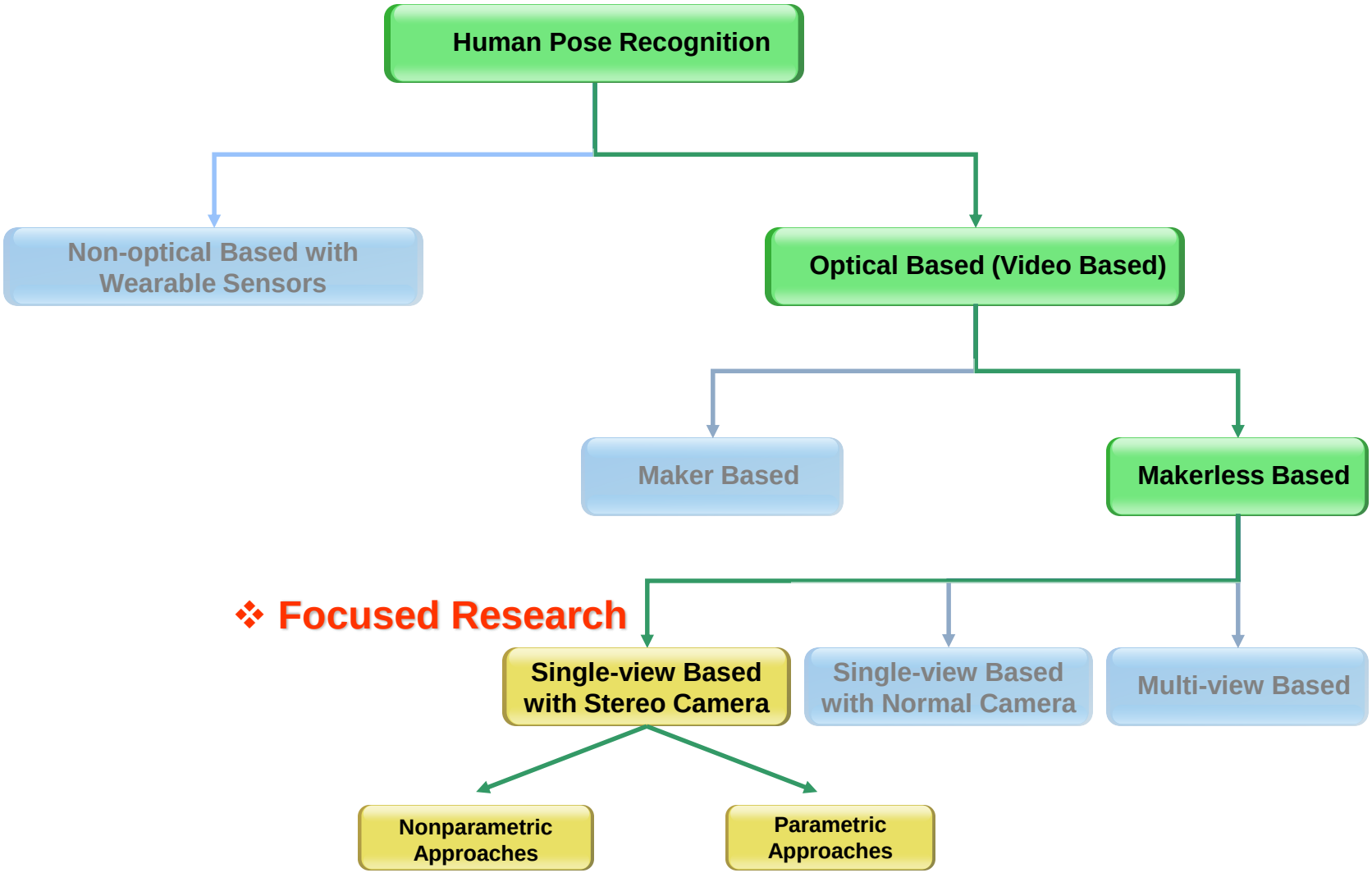


- Human Pose Recognition
  - Non-optical Based with Wearable Sensors
  - Optical Based (Video Based)
    - Maker Based
    - Makerless Based
      - Single-view Based with Normal Camera
      - Multi-view Based

Features:

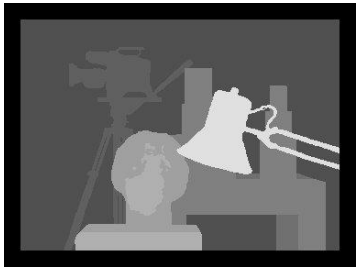
- ✓ Simple installation
- ✓ Ambiguity of 2-D information

Human Pose Recognition: Focused Research



Human Pose Recognition: Stereo Camera

- Stereo camera consists of two lenses to capture a pair of images each time

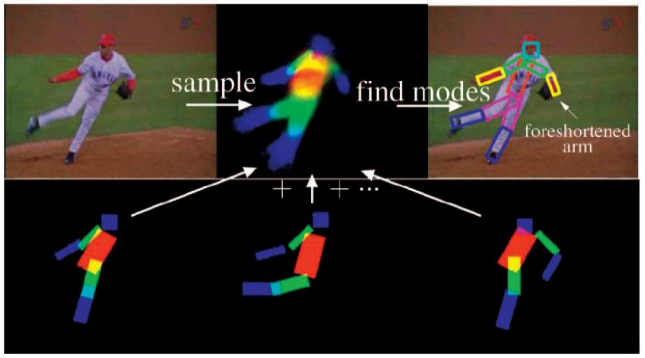


- Correspondences between two images are combined to estimate depth information, decoded in a gray scale image (depth image)

- **Features:** Capable of estimating 3-D info and flexible to be deployed, thus applicable to such areas as human-computer interface, games, surveillance, etc.
- This thesis work focuses on the uses of stereo camera for human pose and activity recognition

Human Pose Recognition Using Stereo Camera: Nonparametric Approach

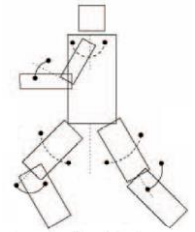
- **Nonparametric approach**: Candidate poses are generated to match with a query image
- Most previous studies proposed to estimate human poses from stereo images were based on this approach



- **Features:**
  - ✓ With candidate poses created in advanced, large pose databases and efficient pose retrieval algorithms are needed[1]
  - ✓ With candidate poses created online, human motion needs to be learned to limit the number of created poses[2]

Human Pose Recognition Using Stereo Camera: Parametric Approach

- **Parametric approach**: Establish a parametric-based formulation and apply mathematical tools for finding solutions[3]



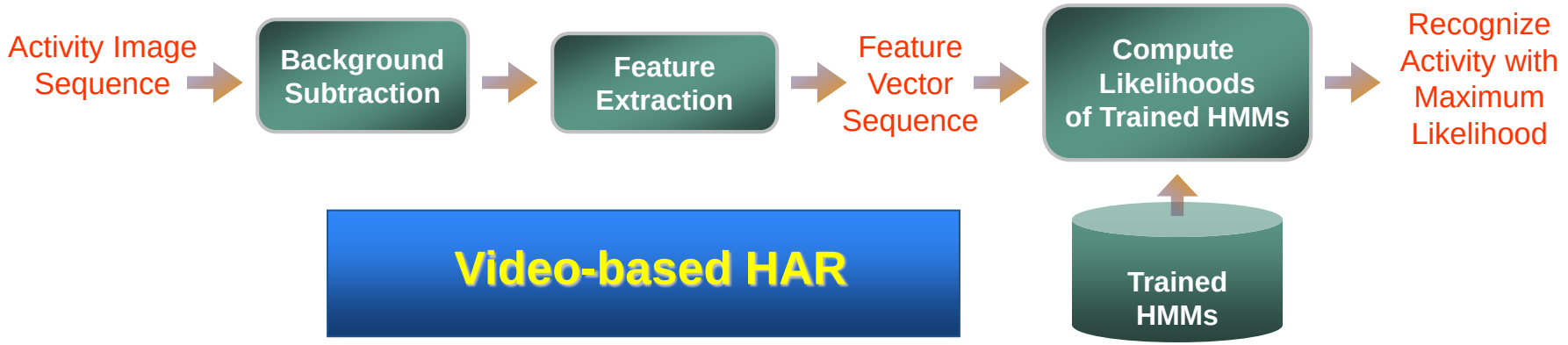
ϑ parameters of human poses

$$\vartheta^* = \max_{\vartheta} p(\vartheta|I)$$

Compatible probability

- Few attempts have concerned this approach to directly recover human poses from stereo images
- Model-data **registration using hidden variables**[4] (i.e., auxiliary variables mapping point-to-point,-to-mesh, or -to-model) is one of possible ways
- However, a generalized formulation of this method to exploit information from stereo images to estimate human poses has not been developed

HAR: Binary Silhouette Features



- Previous HAR with hidden Markov models (HMMs) uses binary silhouettes as features to distinguish different human activities
- Different 3-D human body poses may appear in the same binary silhouette, affecting the recognition rates of a HAR system

- **Exploring better features for HAR** remains an active topic and open to all researchers in fields of computer vision

Motivations

Motivations of Applications

- Wide applicable areas motivate us to develop a marker-less system to recognize human poses and activities using stereo camera

Motivations of Human Pose Recognition

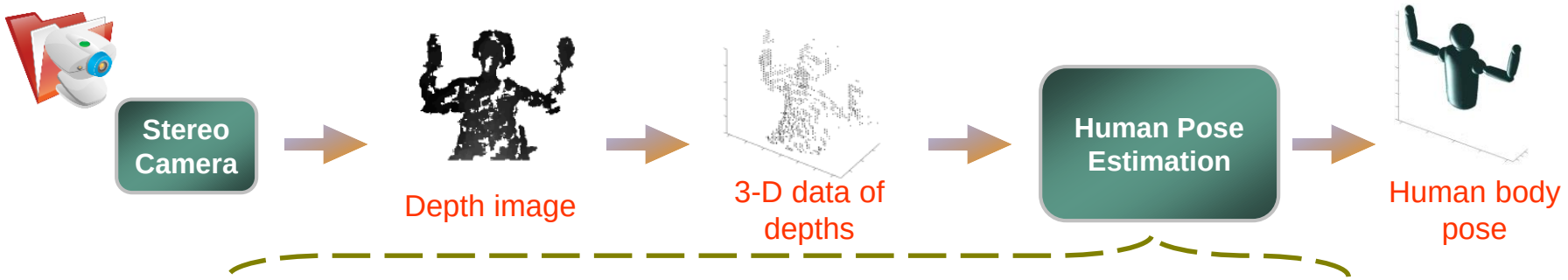
- New parametric-based method allows us to directly estimate human poses from stereo images without creating exemplar poses for matching
- Flexible and efficient model-data registration allows us to integrate more info from depths and RGB images for human pose recognition from stereo images

Motivations of HAR

- New feature with body joint angles allows us to improve accuracy of parametric-based HAR with HMMs

Proposed Human Pose Recognition

Human Pose Recognition System



- Establish a probabilistic formulation between stereo data and a human model using model-data registration with hidden variables
- Define elements of probabilistic formulation including smoothness constraints, RGB likelihoods, geodesic constraints, and reconstruction errors
- Develop co-registration algorithm to fit a human model to stereo data by apply variational expectation maximization (VEM) on the defined probabilistic formulation

Proposed HAR

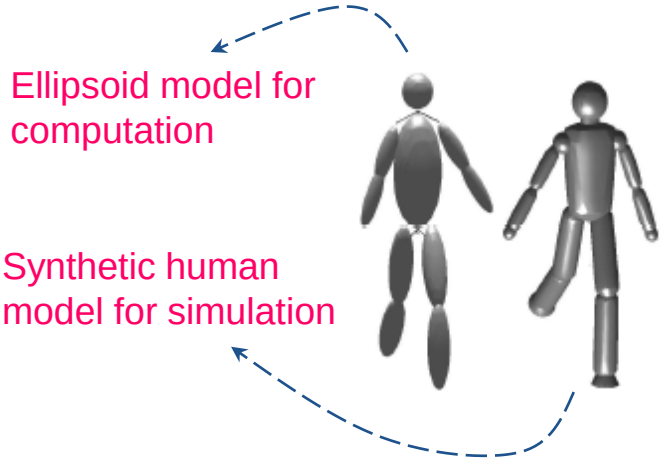
- A sequence of human poses are estimated from a sequence of stereo images
- Joint angles of recovered poses are used as new features to classify human activities
- Apply HMMs to model and recognize time sequential features

Recovering 3-D Human Body Poses from Stereo Images

Methodology

Methodology to recover human body poses from depths

- ✓ Estimate depths from stereo images
- ✓ Design human model
- ✓ Formulate probabilistic registration problem with hidden variables
- ✓ Derive a co-registration algorithm with VEM



Estimate depth image **Co-registration** **Final recovered pose**

Stereo Matching Algorithm

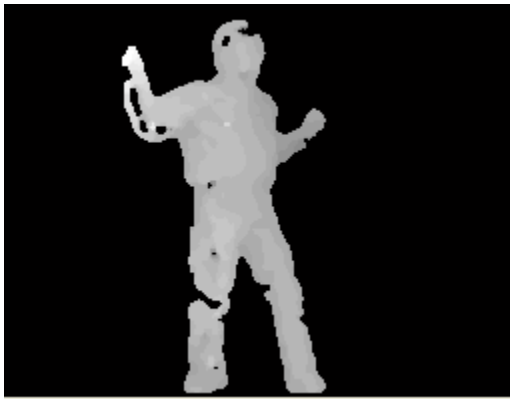
For each epipolar line

For each pixel in the left image

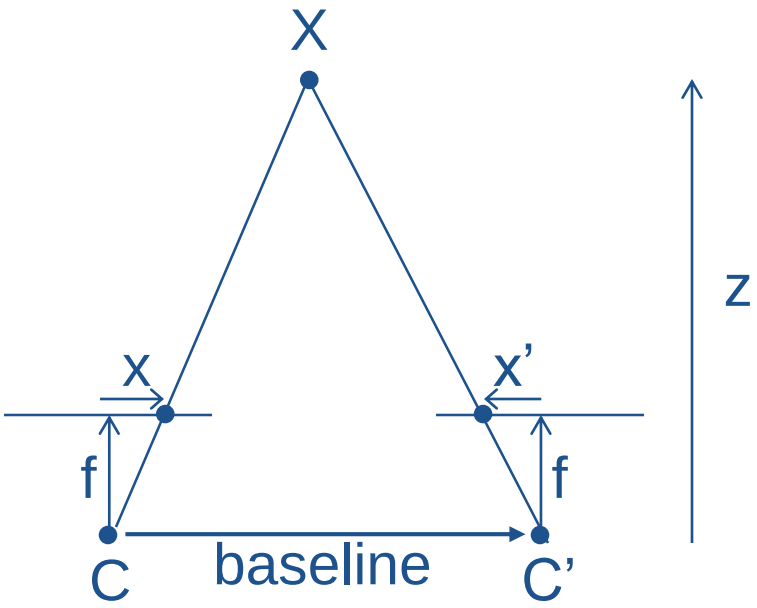
- ✓ Compare with every pixel on same epipolar line in right image
- ✓ Pick pixel with minimum match cost

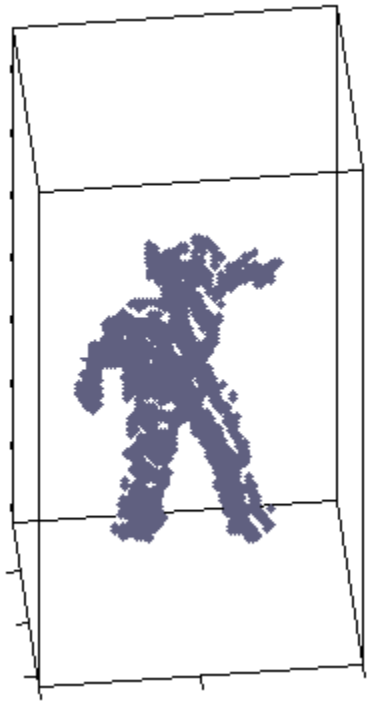
Improvement: Sparse searching with growing corresponding seed (GCS) algorithm

Compute Depth Values



Disparity (depth) image





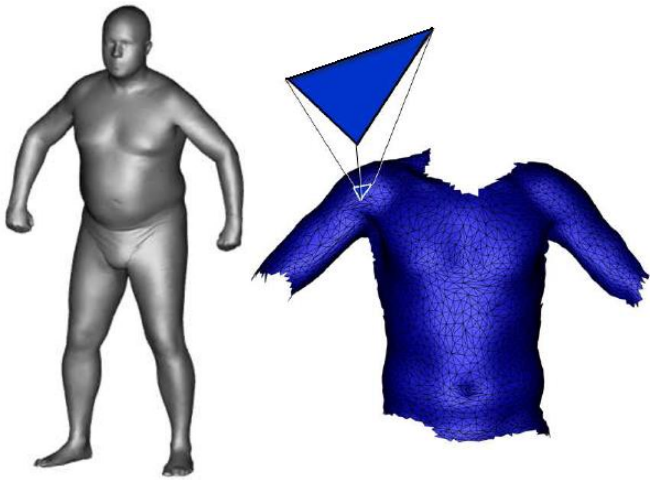
3-D data of depths

$$disparity = x - x' = \frac{baseline * f}{z}$$

3-D Human Model: Whole-body Model

- Presented by a single deformable surface for the entire body
- Originally developed in the computer graphics for animations and virtual reality applications
- Complicated to design and animate with the requirements of high accuracy of input source (multiple cameras, 3-D laser scanner, etc.)
- Slow processing (e.g, about *30s~1min* per frame)

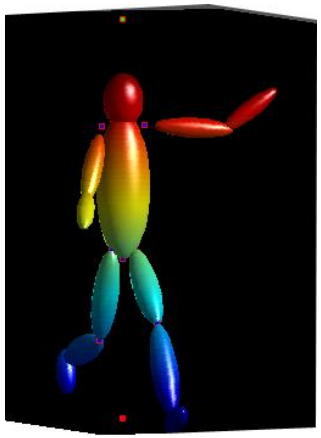
Soft object



SCAPE (Shape Completion and Animation of People)

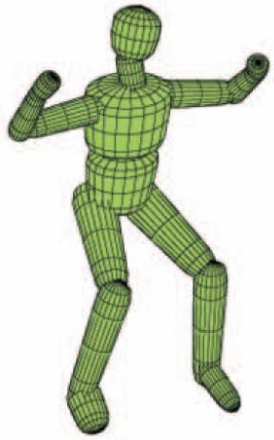
3-D Human Model: Body-part Model

- Body-part models present each part as a rigid surface attached to a joint of the kinematic tree
- These models are convenient to control and yield success for articulated human pose estimation and tracking
- Commonly used representations include geometric primitives like cylinders, truncated cones, ellipsoids, polyhedrons or superquadrics



Ellipsoids

Truncated cones

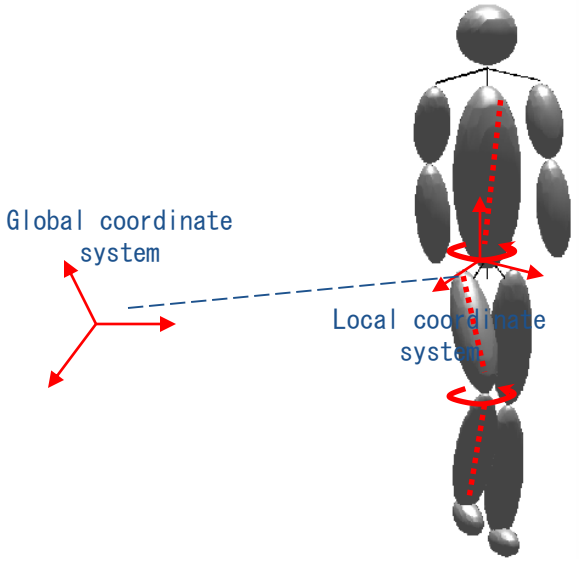


Superquadrics

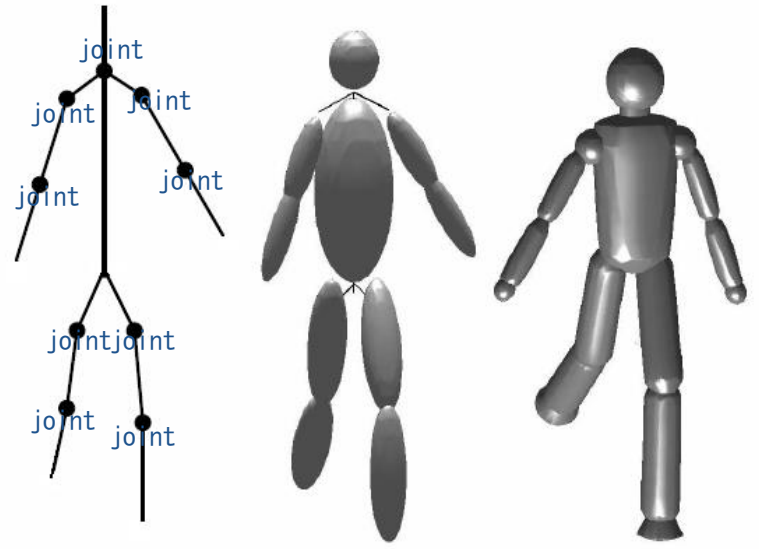
Our Body-part Model: Ellipsoid

- The 3-D human model is defined by a set of ellipsoids[7]
- Each ellipsoid is controlled by a kinematic chain of rotation angles

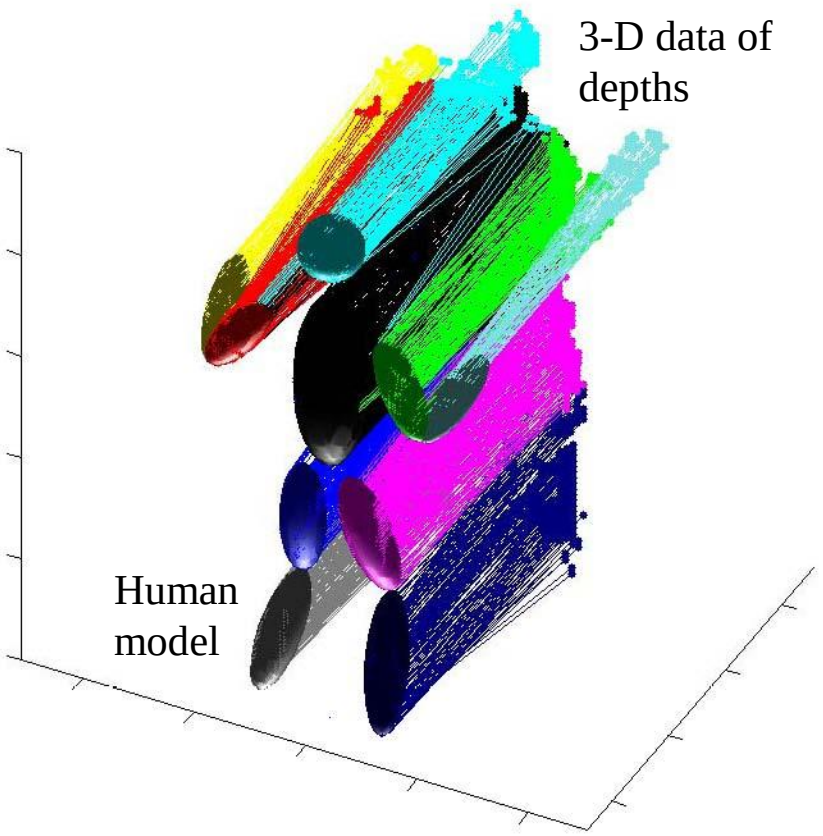
$$\mathbf{Q}_\vartheta = \mathbf{Q}_n(\vartheta_n)\mathbf{Q}_{n-1}(\vartheta_{n-1})\ldots\mathbf{Q}_1(\vartheta_1)$$



- ✓ 14 segments
- ✓ 10 ellipsoids
- ✓ 9 joints
- ✓ 2 DOF at each joint
- ✓ Total 24 DOF



Model-depth Registration with Hidden Variables



- Each 3-D point of depths should find its correspondent ellipsoid (a body part) of a human model
- Hidden variables are defined as body-part labels of each point

Probabilistic Formulation to Recover Human Poses

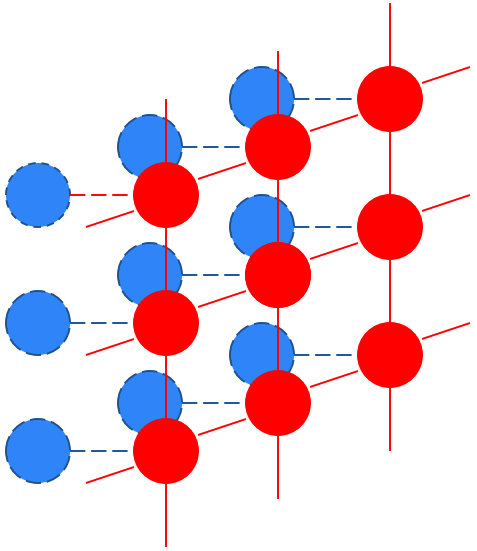
- Let $D=(X_1,X_2,\ldots,X_M)$ denote M points of 3-D data
- I for RGB images
- Let $V=(v_1,v_2,\ldots,v_M)$ be the body part assignment v_i for each point i of 3-D data (used for ellipsoid registration)
- $\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_n)$ are kinematic parameters
- Probabilistic relationship between ϑ and V given I and D

$$P(V, \vartheta | I, D) \propto P(V)P(I|V)P(D|V)P(D|V, \vartheta)$$

Elements of the Probabilistic Formulation

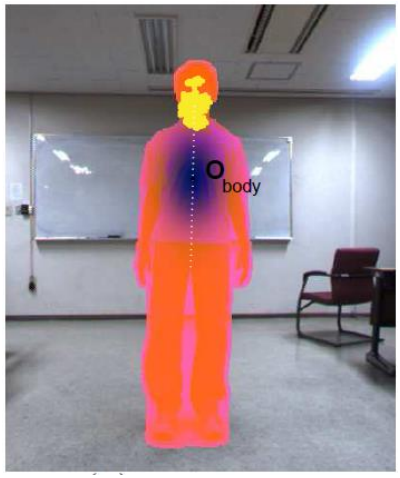
- Smoothness prior $P(V)$: Drive the label of a pixel toward the dominant label of its neighbors

$$P(v_i, v_j) = \begin{cases} e^{\gamma} & \text{if } v_i = v_j \\ 1 & \text{if } v_i \neq v_j \end{cases}$$



- Image likelihood $P(I|V)$: Use RGB cues to provide extra information of labels, e.g.

$$P(I|v_i = head) = \begin{cases} e^{c} & \text{pixel } i \text{ detected as '}\textbf{\textit{face}}\text{'} \\ 1 & \text{otherwise} \end{cases}$$

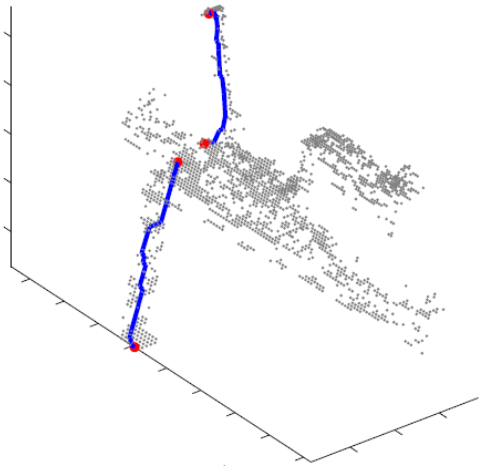


*c and γ are positive constants

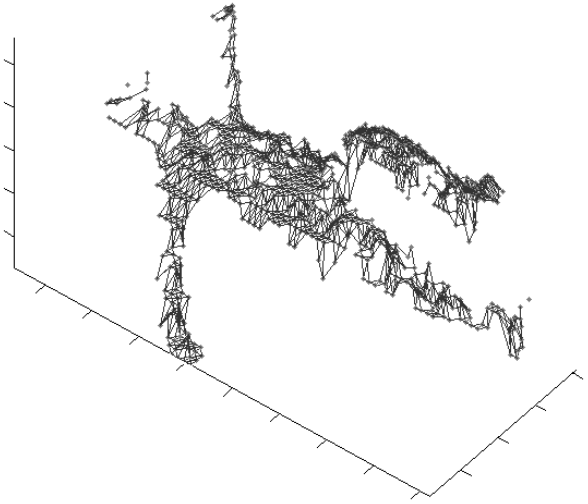
Elements of the Probabilistic Formulation (cont')

- Geodesic constraint $P(D|V)$: Two points with their corresponding label pair should not be too far or too close

$$P_{geo}(D|v_i, v_{j_c}) = \begin{cases} e^{-\alpha} & d(v_{i_c}, v_{j_c}) < d_{min}(v_{i_c}, v_{j_c}) \\ e^{-\beta} & d(v_{i_c}, v_{j_c}) > d_{max}(v_{i_c}, v_{j_c}) \end{cases}$$



- The geodesic distance is approximated by the shortest path distance in a graph

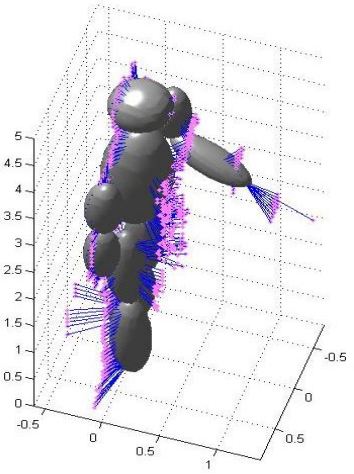


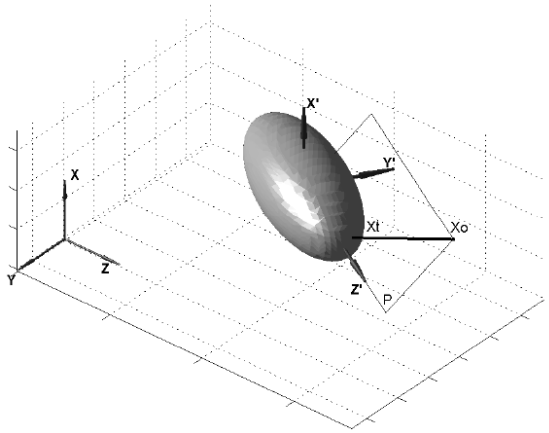
*α and β are two positive constants

Elements of the Probabilistic Formulation (cont')

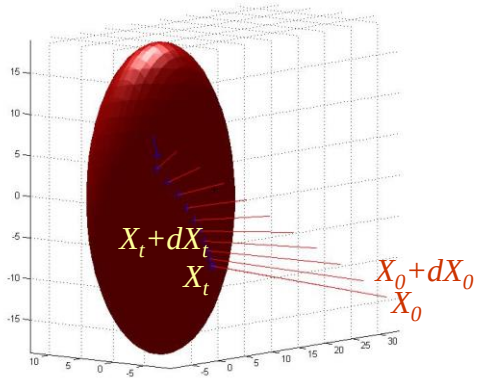
- Reconstruction error $P(D|V,\vartheta)$: Measure the errors (Euclidean distance) between ellipsoids of model and 3-D data of depths

$$P(D|V,\vartheta) = \prod_{i=1}^{M} e^{-\frac{d^2(X_i,\vartheta,v_i)}{2\sigma^2}}$$





Finding the nearest of one 3-D point requires a sixth-degree polynomial equation[8]



Finding the nearest of one 3-D point is simplified by using gradient methods

*σ is a variance and $d_i(X_i,\vartheta,v_i)$ is the distance from a point X_i to an ellipsoid v_i

Co-registration Algorithm with VEM

- The kinematic parameter ϑ of a human body posture is found to be the root of an optimal problem

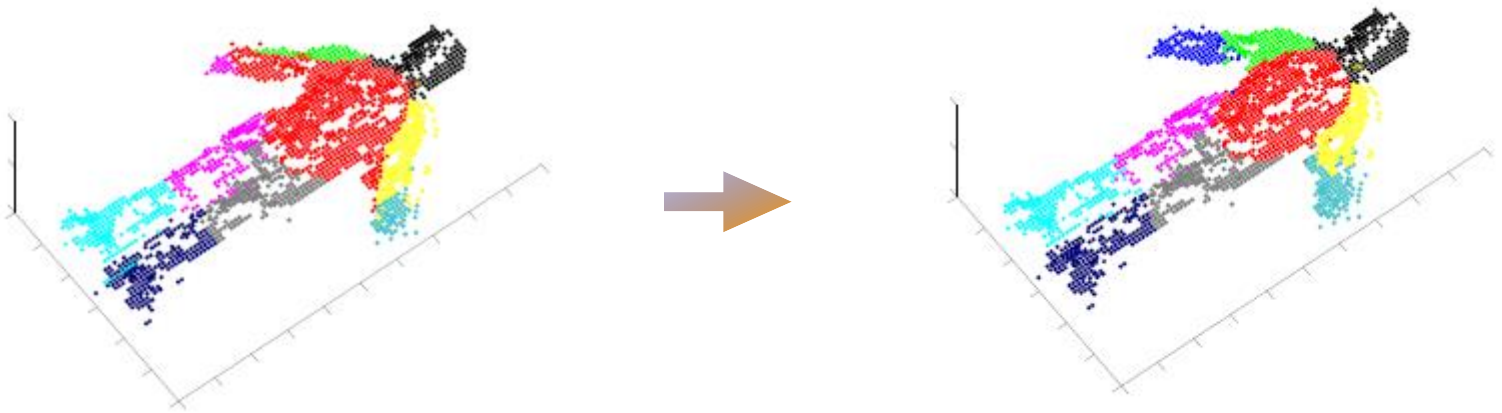
$$\vartheta^* = \text{argmax}_{\vartheta} \sum_{V} P(V, \vartheta | I, D)$$

- VEM is applied to solve this problem with two main steps
  - ✓ VE-step
  - ✓ M-step (Model fitting)
- Two steps are iterated until the algorithm converges toward an estimated pose (co-registration)

Variational E-step (VE-step)

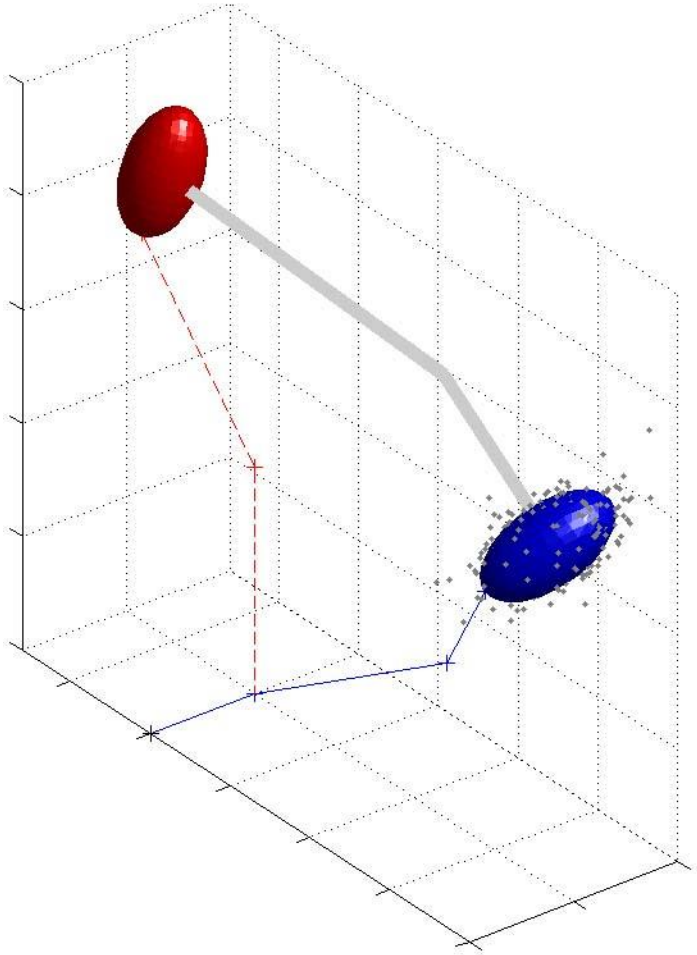
- Estimate the posterior $P(V|\vartheta, I, D)$ of hidden variables V
- The approximation of $P(V|\vartheta, I, D)$ is estimated by variational method

$$Q(V) \propto P(V|\vartheta, I, D)$$

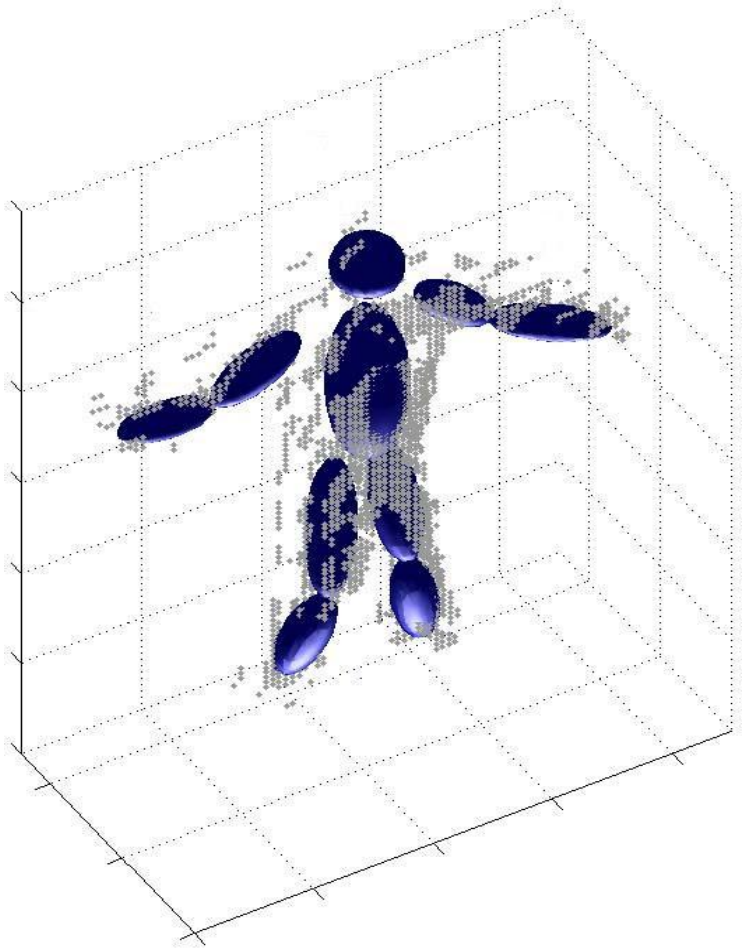


VE-step approximates the distribution of body-part labels

M-step (Model Fitting)

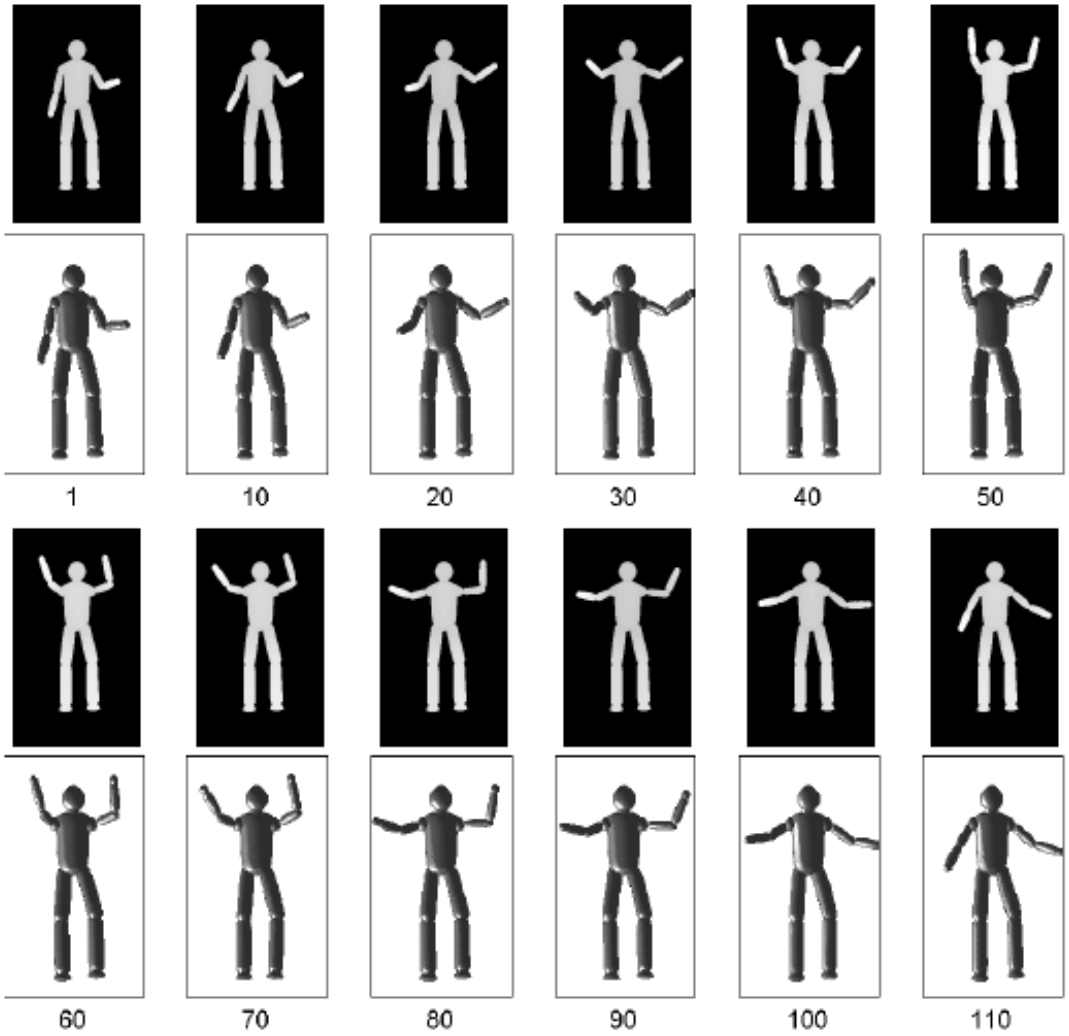


Single ellipsoid fitting

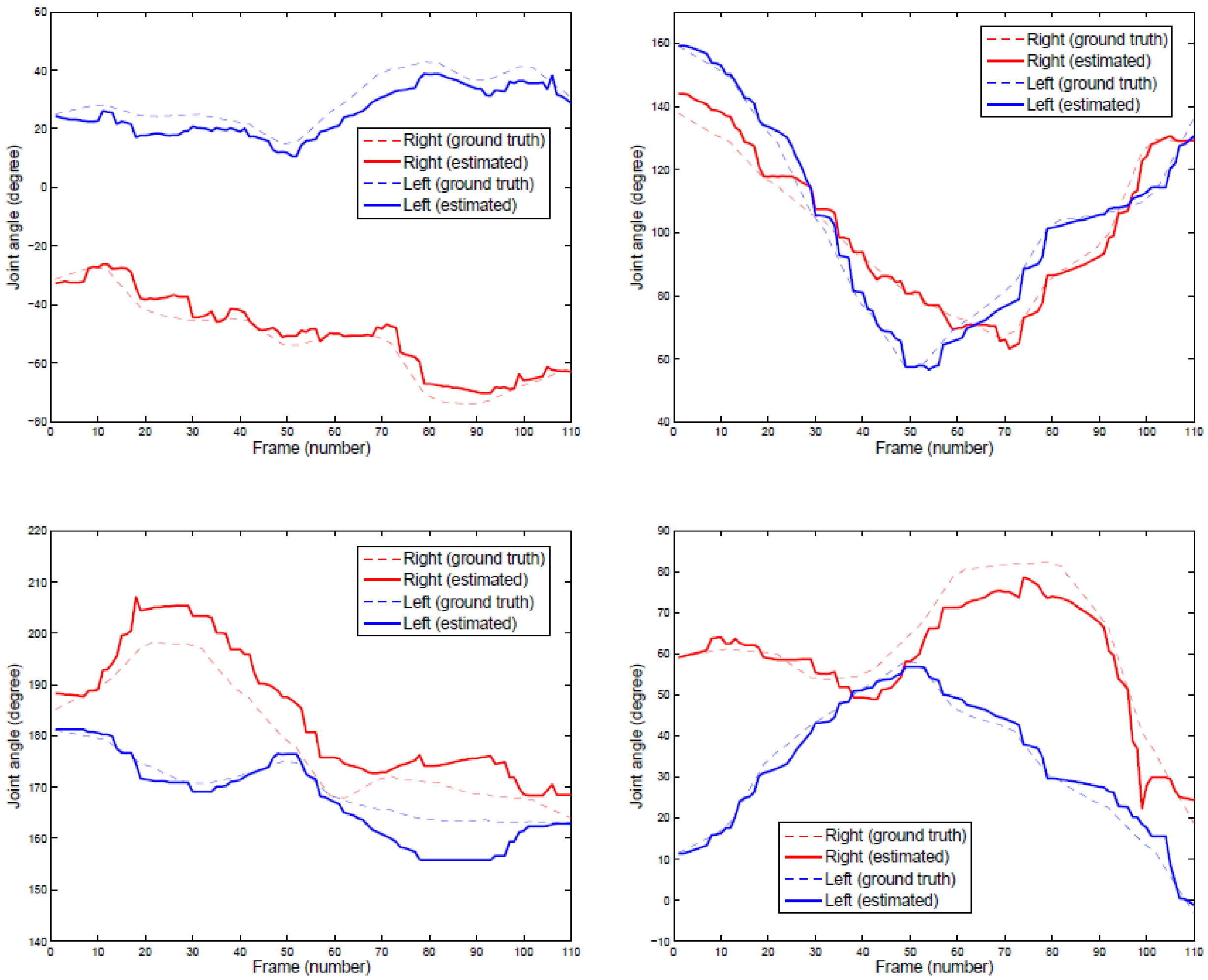


Whole human body fitting

Experimental Results with Synthetic Data



Experimental Results with Synthetic Data



Experimental Results with Real Data

Experimental Results with Real Data

The average reconstruction error (^{0}) of the joint angles of the first four experiments

Experiment	Evaluated angle	Average reconstruction error	
Elbow movement (horizontal direction)	Upper arm & lower arm	Left	8.21
		Right	7.58
Elbow movement (vertical direction)	Upper arm & lower arm	Left	6.79
		Right	7.64
Knee movement	Upper leg & lower leg	Left	8.03
		Right	13.81
Shoulder movement	Whole arm & x-axis	Left	5.66
		Right	5.72
	Whole arm & z-axis	Left	9.08
		Right	9.97

Experimental Results with Complex Motion

$$D_t = \frac{\sum_{i=1}^{M} d_t(i)}{M}$$

$d_t(i)$ is the Euclidean distance from the point i to the nearest ellipsoid, M number of points

The mean and standard derivation of the average distance D_t of two sequences

Sequences	Walking	Arbitrary activity
Mean (m)	0.062	0.037
Std. Dev. (m)	0.003	0.002

Demonstration

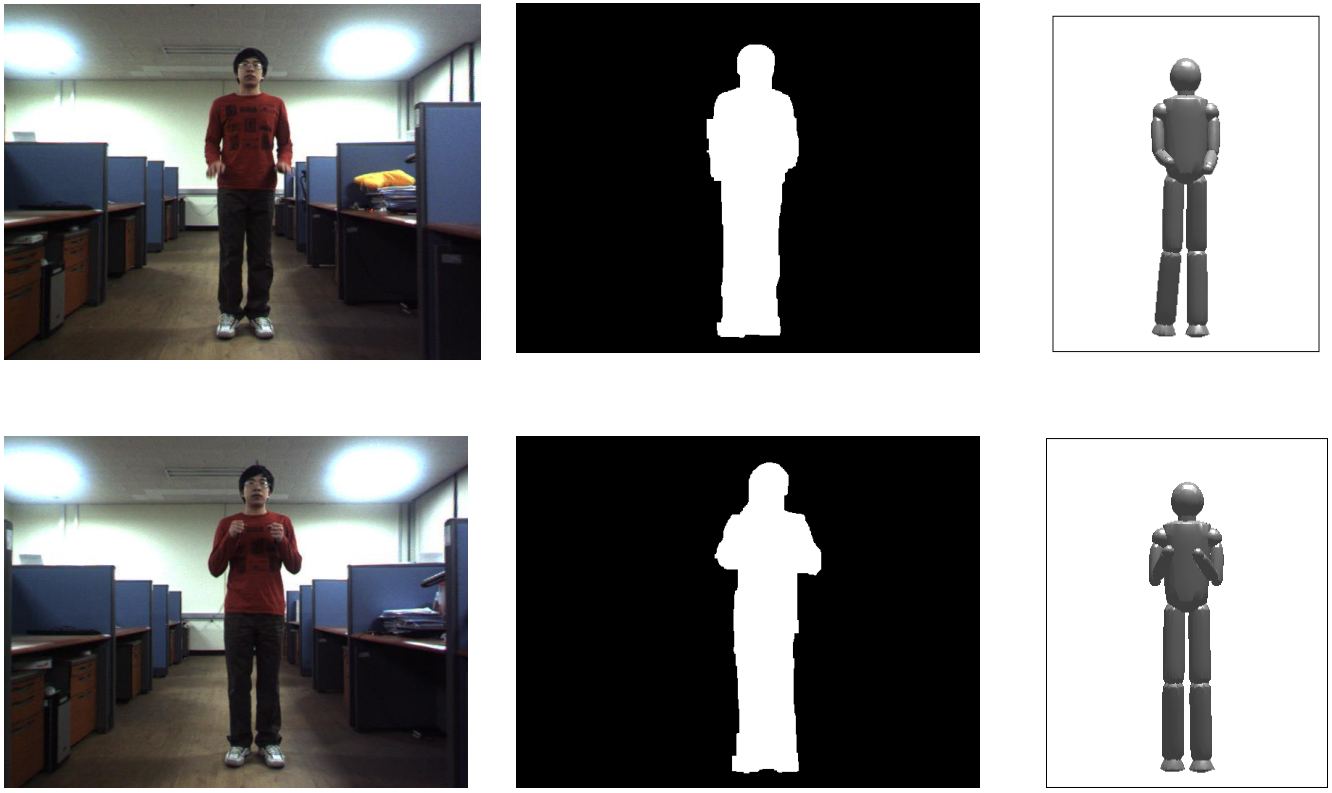
Acquired frame rates: *4~8Hz*

Conclusion

- We present a marker-less system to recover 3-D human body poses from stereo images
- Our method to recover human body poses is derived in an efficient and flexible framework using co-registration and body part detections with the cues from RGB and depth images
- Our system can reconstruct human body poses from stereo video even for complicated movements with an average error of about *6-14*^{0} of estimated kinematic angles

Human Activity Recognition Using Joint Angle Features

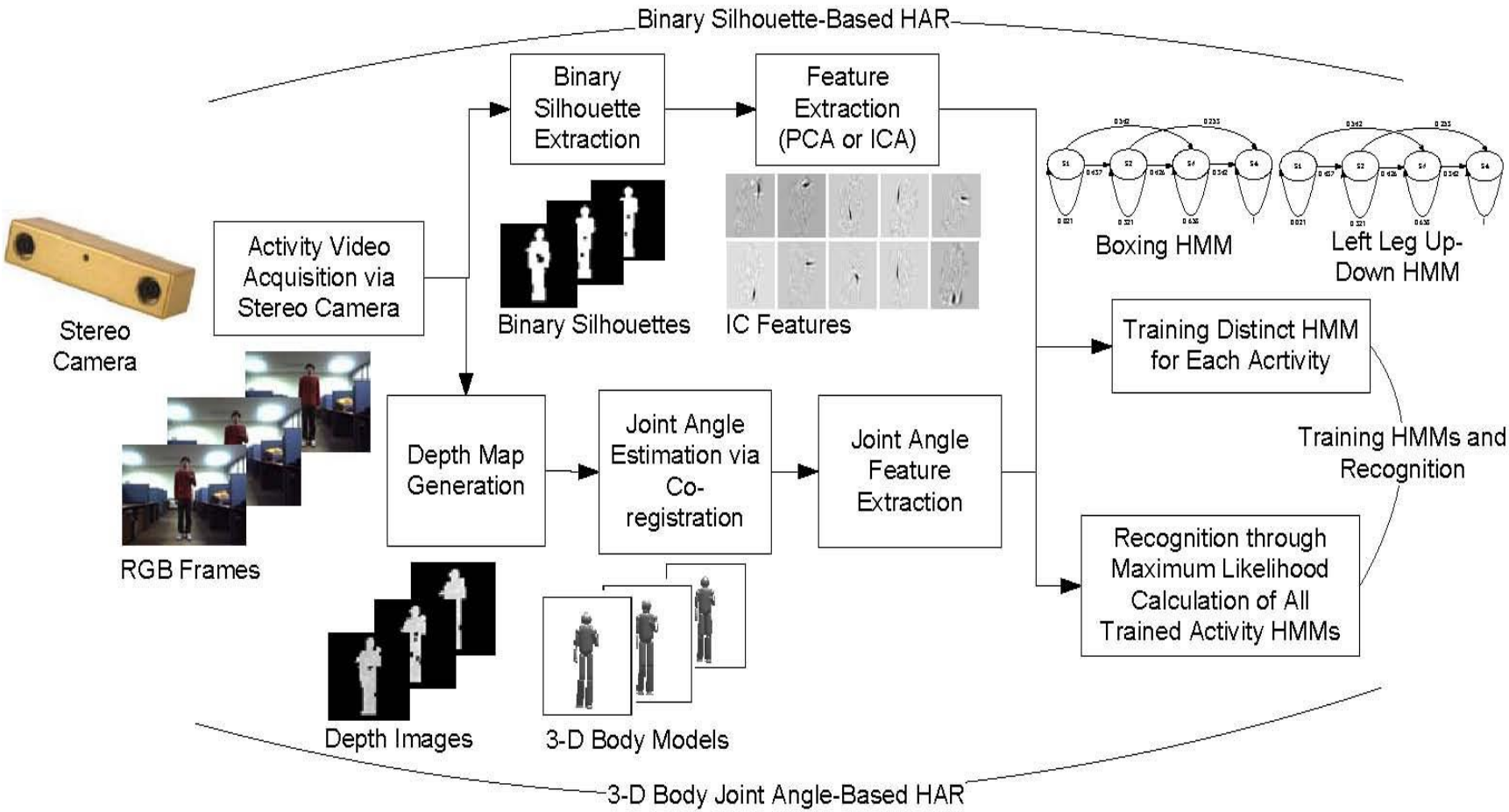
Joint Angles of Human Body Poses and HAR



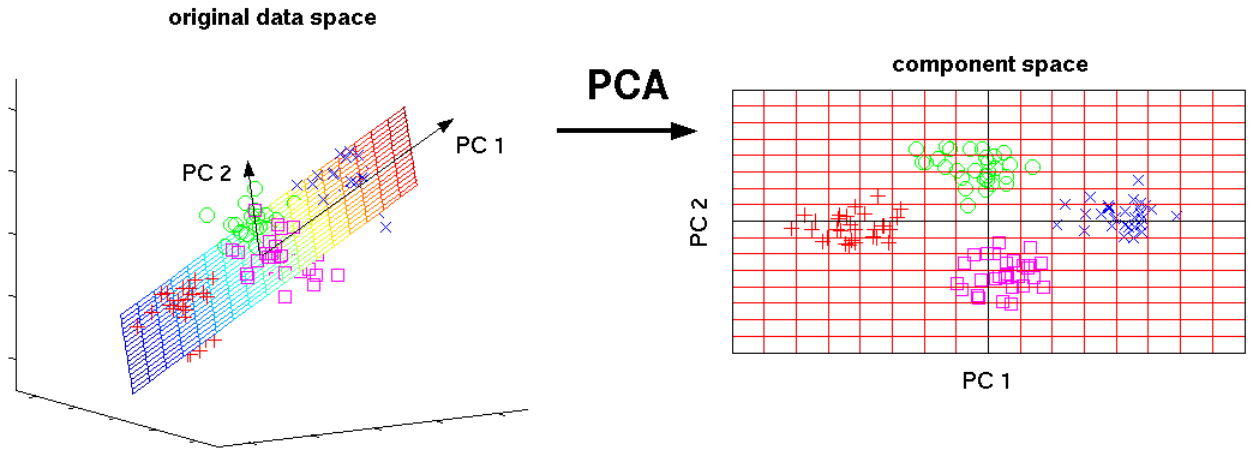
Joint angles seem to be efficient features for HAR

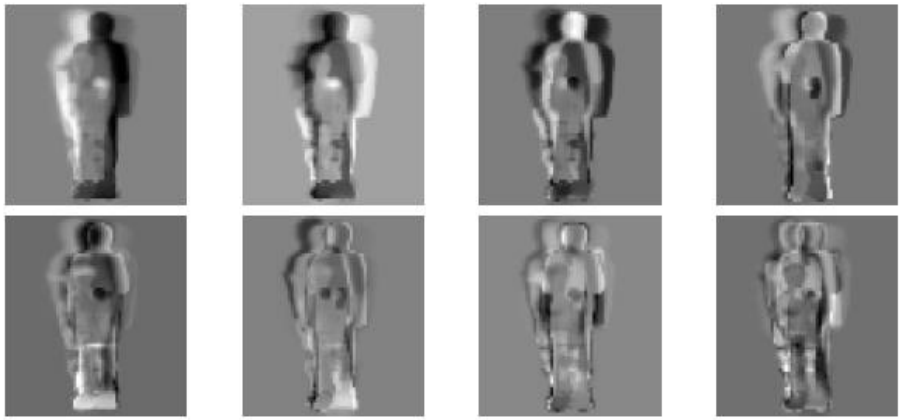
* Here are examples of binary silhouettes in an ideal case

Binary Silhouette- and Joint Angle-based HAR



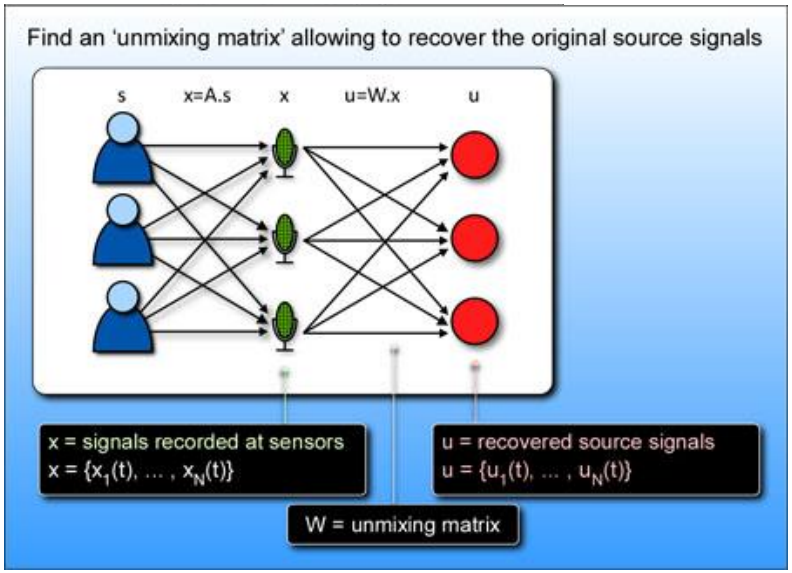
Feature Extraction: Principal Component Analysis (PCA)





- PCA extracts global principal components (PCs)
- Binary silhouette features are projected into PCs to reduce dimensions

Feature Extraction: Independent Component Analysis (ICA)

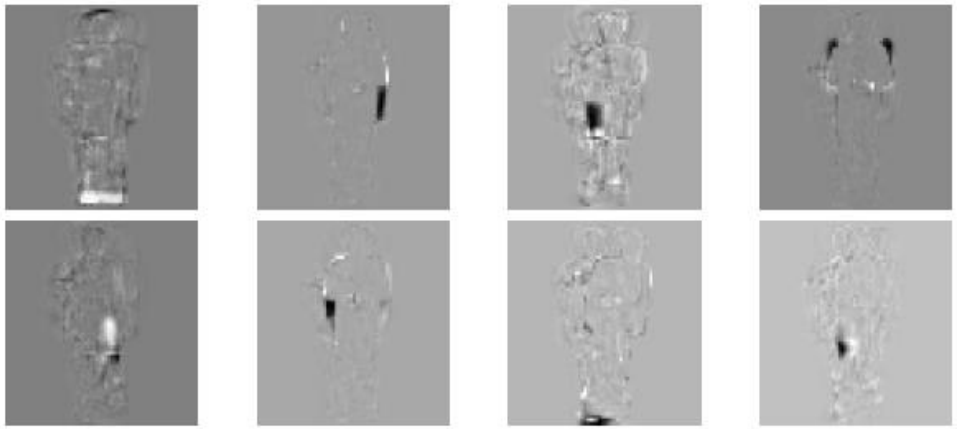


Independent outputs

$$p(u) = p(u_1, u_2, ..., u_N) = p(u_1)p(u_2)...p(u_N)$$

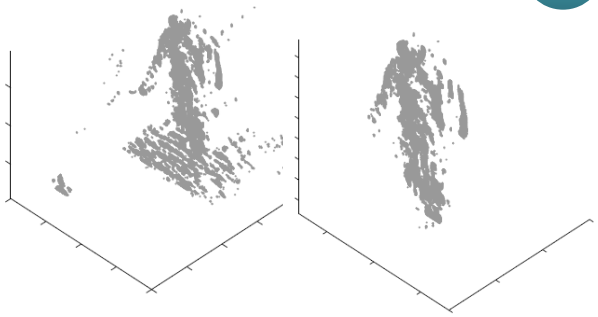
Maximum likelihood is applied to find bases

$$W^* = \text{argmax}_W p(W|u(t))$$

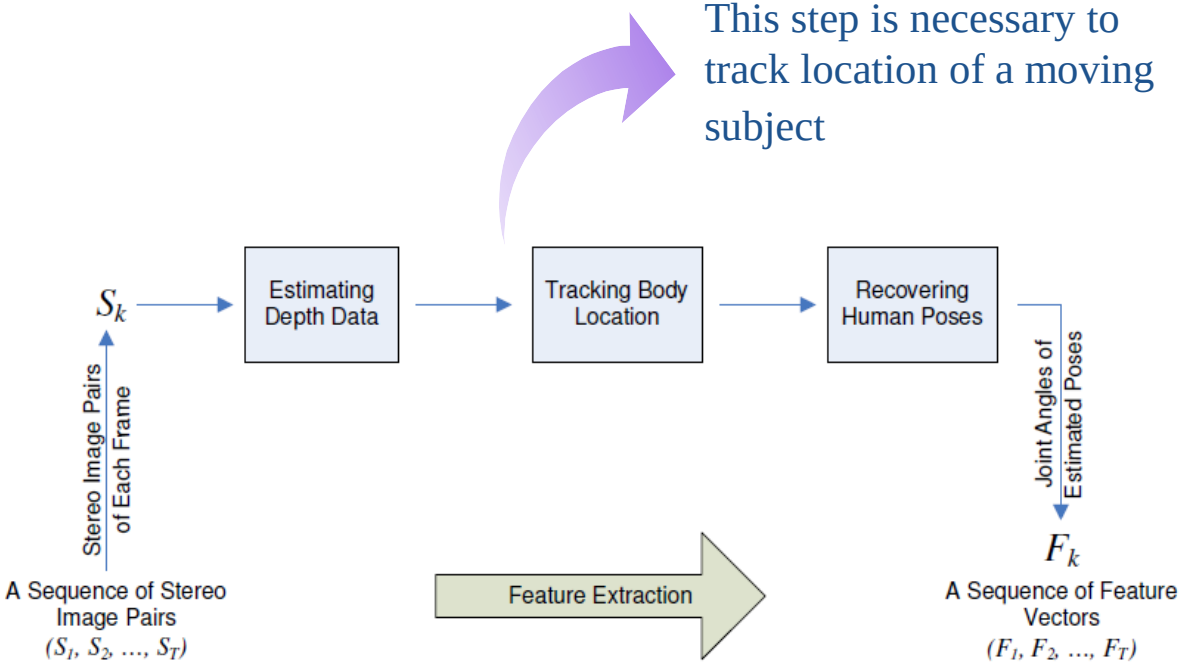


- ICA recovers independent components (ICs) focusing on local features
- Binary silhouette features are projected into ICs to acquire more discriminant features

Feature Extraction: Joint Angles



This step is necessary to track location of a moving subject

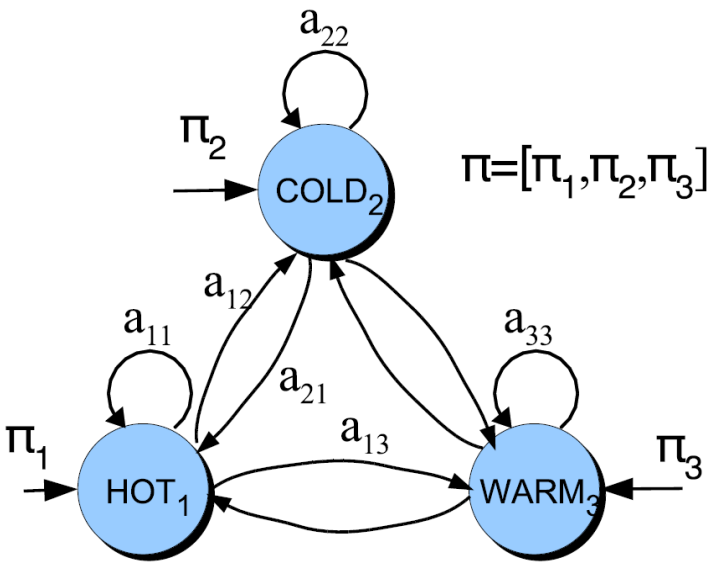


An activity in video frames is presented by a sequence $\{F_1, F_2, \ldots, F_t, \ldots, F_T\}$, where F_t contains kinematic angles of a human posture in a single frame t

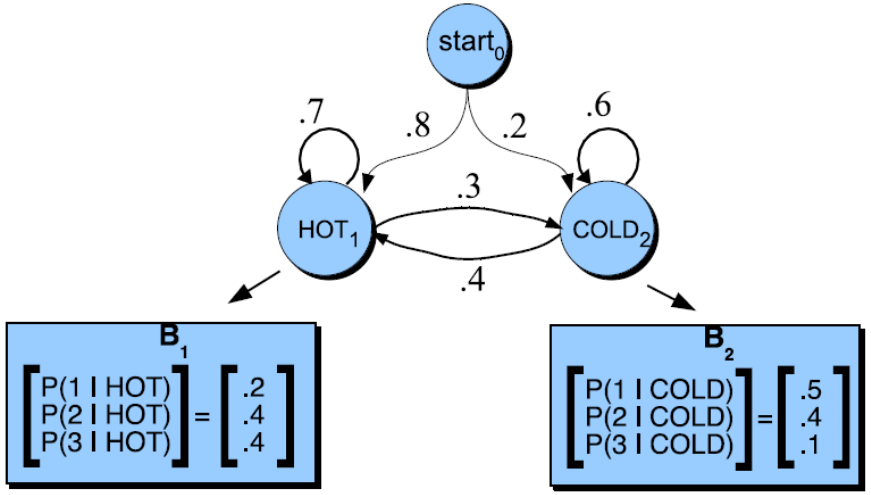
$$F_t = [\theta_{hip}, \theta_{left\_shoulder}, \theta_{right\_shoulder}, \theta_{left\_crotch}, \theta_{right\_crotch}, \theta_{neck}, \theta_{left\_elbow}, \theta_{right\_elbow}, \theta_{left\_knee}, \theta_{right\_knee}]$$

Hidden Markov Model (HMM)

- Example:

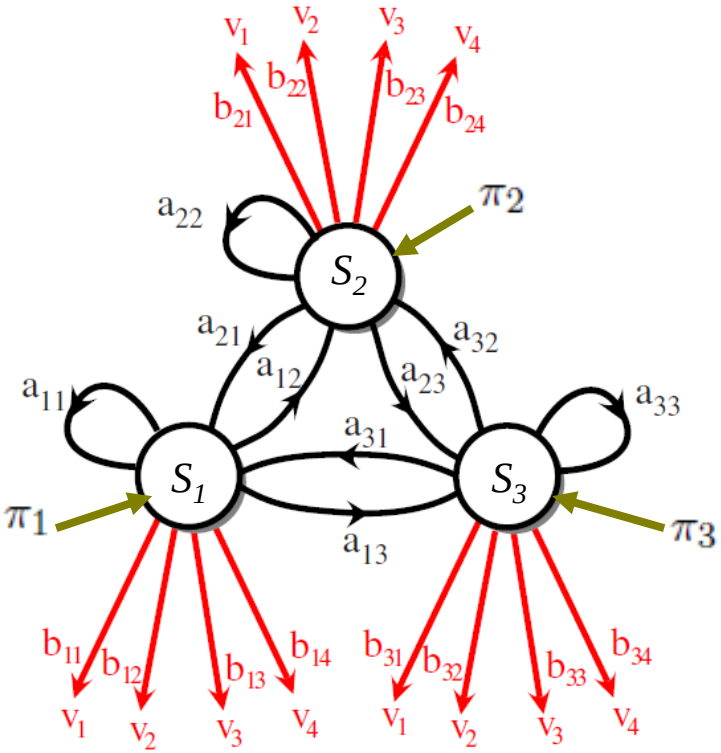


States: Real-world weather



Observations: The number of ice creams eaten by Jason

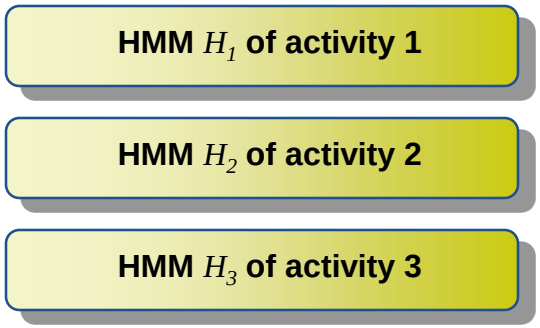
Hidden Markov Model (cont')



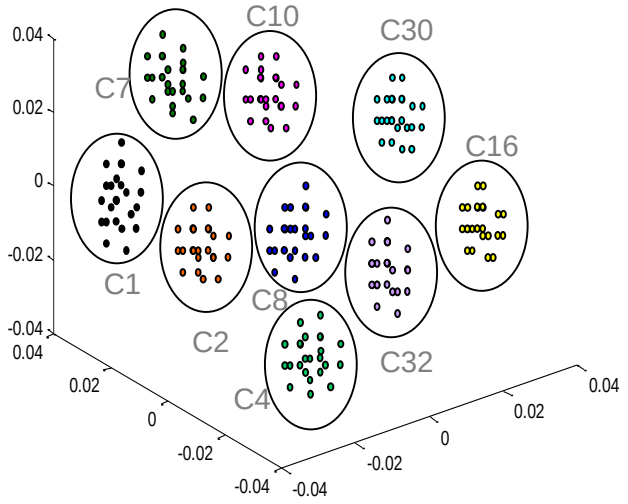
- ✓ HMMs evaluate the probability of an observed sequence $\{v(1),v(2),\ldots,v(T)\}$ using state variables S_i
- ✓ Parameters of HMMs includes
  - Transition probability $A=\{a_{ij}\}$, a_{ij} is a transition probability from state S_i to state S_j
  - Emission probability $B=\{b_{ik}\}$, b_{ik} is the probability of state S_i giving an observation $v_k(t)$
  - Initial probability $\pi=\{\pi_i\}$
- ✓ Two fundamental problems of HMMs
  - Given a sequence $V=(v(1),v(2),\ldots,v(T))$, evaluate the likelihood $P(V|A,B,\pi)$
  - Given a set of training sequences $\{V^i\}$, estimate the values of parameters $\{A,B,\pi\}$ to maximize the probability $P(A,B,\pi|\{V^i\})$

Parametric-based HAR with HMMs

- An observation $v(t)$ of PCA or ICA features and joint angle features is embedded in a continuous space
- Thus, we need to perform vector quantization to get a discrete value of $v(t)$
- Sequences of $v(t)$ from training data are used to learn the parameter $\{A_i, B_i, \pi_i\}$ of each HMM H_i
- Likelihood $P(V|A_i, B_i, \pi_i)$ are used to distinguish different activities



$$decision = \text{argmax}_{i=1,2,\ldots,M}\{P(V|A_i, B_i, \pi_i)\}$$



Vector quantization

Recognition Results of PCA-based HAR Using Binary Silhouette Features

Activity	Recognition Rate (%)	Mean	Standard Deviation
Left hand up-down	47.50	**58.12**	19.03
Right hand up-down	55		
Both hands up-down	60		
Boxing	20		
Left leg up-down	60		
Right leg up-down	67.50		
Walking	70		
Sitting	85		

Recognition Results of ICA-based HAR Using Binary Silhouette Features

Activity	Recognition Rate (%)	Mean	Standard Deviation
Left hand up-down	47.50	64.06	18.03
Right hand up-down	60		
Both hands up-down	67.50		
Boxing	30		
Left leg up-down	72.50		
Right leg up-down	72.50		
Walking	75		
Sitting	87.50		

Recognition Results of HAR Using 3-D Joint Angle Features

Activity	Recognition Rate (%)	Mean	Standard Deviation
Left hand up-down	87.50	**92.81**	3.65
Right hand up-down	97.50		
Both hands up-down	87.50		
Boxing	95		
Left leg up-down	92.50		
Right leg up-down	95		
Walking	92.50		
Sitting	95		

Conclusion

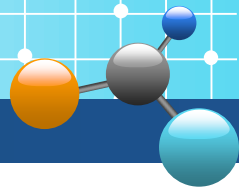
- We propose a new HAR system using joint angles of human body poses
- Our HAR system is capable of recognizing human activities with very high accuracy, about 93% in the recognition rate
- Our recognition rate is much better than that of all conventional approaches could achieve

Contributions, Applications and Future Work

Contributions

- Implement a new system to recover human body poses and to recognize human activities from stereo images without using markers or attached sensors
- Propose an efficient and flexible framework for human pose recognition utilizing cues from depths, RGB images and relationships among 3-D data
- Use variational method to derive a co-registration algorithm to fit a human model to stereo data
- Propose HAR using joint angles which is superior to conventional HAR using binary silhouettes
- Whole proposed system is well suited to many practical applications

Applications



- **Human Computer Interaction** uses motion of body limbs to distinguish human activities, which are the inputs to control external devices such as computers, games and robotics
- **Ubiquitous Healthcare** needs the tracking of human movements and activities to detect abnormal events such as dangerous falls of elderly persons
- **Security** is another application domain that requires video surveillance to monitor people in public or private areas

Future Work

- However, existing errors of recovered kinematic angles make our system face difficulty with medical applications requiring high accuracy of estimating human motion (e.g., biomecanic measurements, disease diagnosis, etc.)
- Also, improving the proposed system for real-time processing is another factor needed to be concerned
- We plan our future work to improve the reliability of our presented techniques and its robustness to handle the rapid and complex changes of human postures in a video sequence by
  - ✓ Apply more techniques to better detect human body parts
  - ✓ Integrate human motion from an exemplar database
  - ✓ Improve processing speed by hierarchy registration

Publications

Journals

1. Nguyen Duc Thang, Tahir Rasheed, Young-Koo Lee, Sungyoung Lee, and Tae-Seong Kim "*Content-based Facial Image Retrieval Using Constrained Independent Component Analysis*", Information Sciences (SCI), DOI: 10.1016/j.ins.2011.03.021, 2011
2. Md. Zia Uddin, Nguyen Duc Thang, Jeong Tai Kim, and Tae-Seong Kim, "*Human Activity Recognition Using Body Joint Angle Features and Hidden Markov Model*", ETRI Journal (SCI), 2011, (accepted)
3. Nguyen Duc Thang, Tae-Seong Kim, Young-Koo Lee, and Sungyoung Lee, "*Estimation of 3-D Human Body Posture via Co-Registration of 3-D Human Model and Sequential Stereo Information*", Applied Intelligence (SCI), DOI: 10.1007/s10489-009-0209-4, 2010

Book Chapter

1. Nguyen Duc Thang, Md. Zia Uddin, Young Koo Lee, Sung Young Lee, and Tae-Seong Kim, "*Recovering 3-D Human Body Posture from Depth Maps and Its Application for Human Activity Recognition*", Book Name: Depth Map and 3D Imaging Applications: Algorithms and Technologies (to be published)

Publications

Conferences

1. Nguyen Duc Thang, Sungyoung Lee, and Young-Koo Lee, *"Fast Constrained Independent Component Analysis for Blind Speech Separation with Multiple References",* International Conference on Computer Sciences and Convergence Information Technology (ICCIT), Seoul, Korea, November 30-December 2, 2010
2. Nguyen Duc Thang, Tae-Seong Kim, Young-Koo Lee, and Sungyoung Lee, "*Fast 3-D Human Motion Capturing from Stereo Data Using Gaussian Clusters*", International Conference on Control, Automation and Systems(ICCAS), Gyeonggi-do, Korea, October 27-30, 2010
3. Nguyen Duc Thang, Young-Koo Lee, Yoon-Hyuk Kim, and Tae Seong Kim, "*Makerless 3-D Human Motion Capturing Using a Stereo Camera*", JEGM 2010, Miami, Florida, US, May 12-15, 2010
4. Nguyen Duc Thang, P. T. H. Truc, Young-Koo Lee, Sungyoung Lee, and Tae Seong Kim, "*3D-Human Pose Estimation from 2-D Depth Images*", International Conference on Ubiquitous Healthcare (uHealthcare), Busan, Korea, November 14-17, 2008
5. M.d Zia Uddin, Nguyen Duc Thang, and Tae-Seong Kim, "*Human Activity via 3-D Joint Angle Features and Hidden Markov Models*", International Conference on Image Processing (ICIP), HongKong, China, September 26-29, 2010

Publications

Conferences

5. La The Vinh, Nguyen Duc Thang, and Young-Koo Lee, "*An Improved Maximum Relevance and Minimum Redundancy Feature Selection Algorithm Based on Normalized Mutual Information",* Annual International Symposium on Applications and the Internet (SAINT), Seoul, Korea, July 19-23, 2010
6. Ji-Hwan Kim, Nguyen Duc Thang, Hyung Sang Suh, Tahir Rasheed, and Tae-Seong Kim, *"Forearm Motion Tracking with Estimating Joint Angles from Inertial Sensor Signals*", International Conference on Biomedical Engineering and Informatics (BMEI), Tianjin, China, October 17-19, 2009
7. Ji-Hwan Kim, Nguyen Duc Thang, and Tae-Seong Kim, *"3-D Hand Motion Tracking and Gesture Recognition Using a Data Glove",* IEEE International Symposium on Industrial Electronics (ISIE), Seoul, Korea, July 5-8, 2009
8. Brian J. d'Auriol, Nguyen Thi Thanh Tuyen, Ngo Quoc Hung, Pho Duc Giang, Hassan Jameel, Le Xuan Hung, S.M.K.R. Raazi, Dao Phuong Thuy, Ngo Trong Canh, Adil Mehmood Khan, Sunghyun Kim, Shu Lei, Sakib Pathan, Tran Van Phuong, Sungyoung Lee, and Young-Koo Lee, *"Embedded Processor Security*", The 2007 International Conference on Security and Management (SAM'07), Las Vegas, Nevada, USA., June 25-28, 2007


Thank You !