Understanding Human Actions with 2D and 3D Sensors Part II

Zicheng Liu

Microsoft Research Redmond, USA

Junsong Yuan

Nanyang Technological University, Singapore

Outline

- Introduction:
  - Gesture, action, activity
  - 3D sensors
  - Depth maps
    - noises, holes, foreground/background occlusions
  - Skeleton tracking
    - Useful but has limitations
  - Datasets

Outline

- Features
  - Skeleton based features
    - Joint angle trajectory
    - EigenJoints, SMIJ, Ho3DJoints,
    - Fourier temporal pyramid
  - Depthmap based features
    - HOG, DMM-HOG
    - Spin Image
    - Bag of 3D points
    - Spacetime Occupancy Pattern, local occupancy pattern
    - Local Depth Pattern
    - Histogram of Oriented Normal Vectors (HONV), Histogram of 3D Facets
    - Histogram of Oriented 4D Normal vectors (HON4D)
  - RGB+depth

Outline

- Hand segmentation and feature extraction
- Recognition paradigms
  - Direct classification (global features)
  - Bag-of-feature framework (interest points + local descriptors)
  - Actionlet ensemble
  - Random occupancy patterns
  - Contour matching (static hand gesture)
  - Real time online action recognition
    - Temporal segmentation
    - ActionGraph
- Experiments discussed following each topic

Introduction

- Gesture, action, activity
- 3D sensors
- Depth maps
  - accuracy, holes, foreground/background occlusions
- Skeleton tracking
  - Useful but has limitations
- Datasets

Gesture, Action, Activity

- Hand gesture
  - Short, single person, focused on hands
    - American Sign Language
- Action
  - Short, single person, involving the body
    - Throw, catch, clap
- Activity
  - Longer, one or multiple people
    - Reading a book, making a phone call, eating
    - Talking to each other, hugging

Introduction

- Gesture, action, activity
- 3D sensors
- Depth maps
  - noises, holes, foreground/background occlusions
- Skeleton tracking
  - Useful but has limitations
- Datasets

3D Sensors

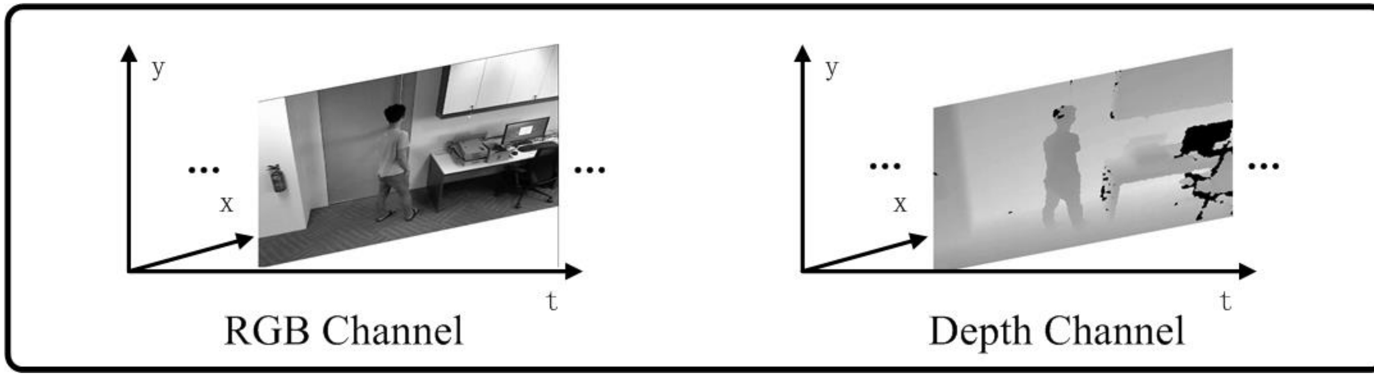
- Laser scanners:
  - Objects have to be motionless
- MoCap sensors (3D joint positions)

  - Expensive, difficult to setup, only research labs have those
- Depth cameras (RGBD)
  - Microsoft Kinect
    - Kinect for Windows driver
  - Cheap, USB, Plug-play

Introduction

- Gesture, action, activity
- 3D sensors
- Depth maps
  - noises, holes, foreground/background occlusions
- Skeleton tracking
  - Useful but has limitations
- Datasets

Depth maps



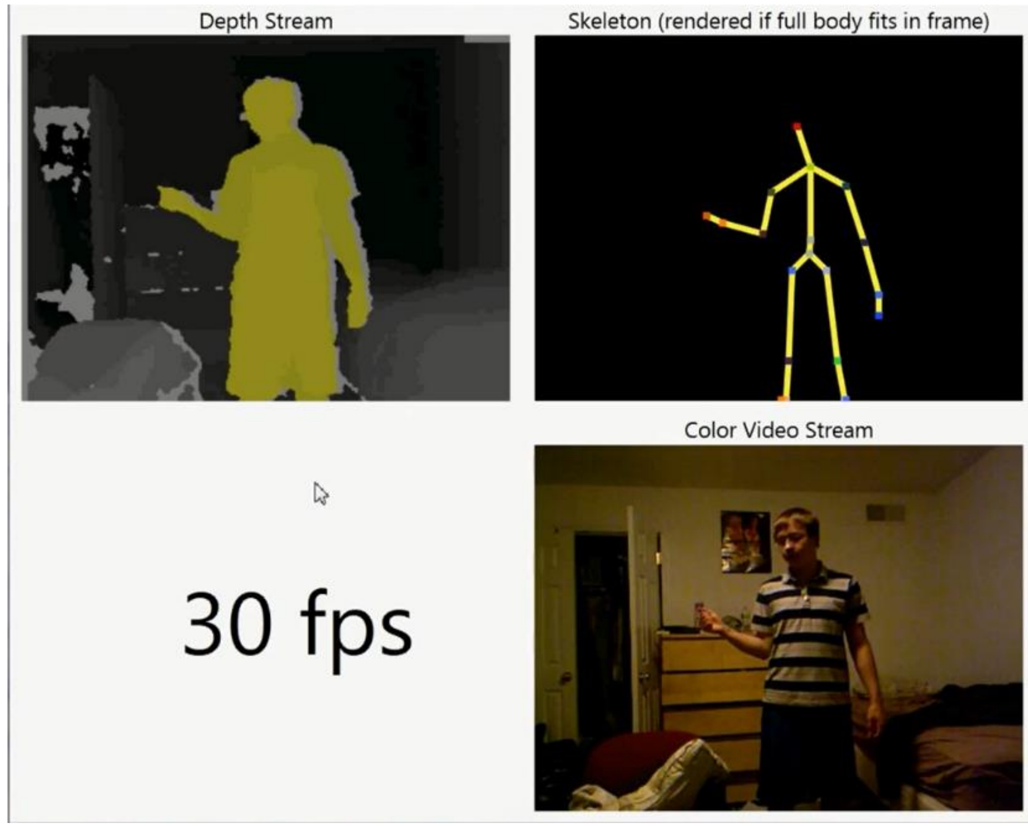
- Noises: flickering
- Accuracy: degrades with the distance to the camera
- Foreground occlusion and background occlusion
  - F/B segmentation is not always easy

Introduction

- Gesture, action, activity
- 3D sensors
- Depth maps
  - accuracy, holes, foreground/background occlusions
- Skeleton tracking
  - Useful but has limitations
- Datasets

Skeleton Tracking

- 20 joints
- Limitations
  - Side view
  - Occlusions
    - Crossing arms
    - Bending
    - Two people



Introduction

- Gesture, action, activity
- 3D sensors
- Depth maps
  - accuracy, holes, foreground/background occlusions
- Skeleton tracking
  - Useful but has limitations
- Datasets

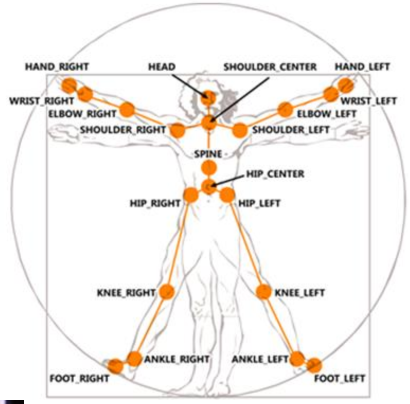
Datasets

- MSR Action3D: sports actions
- MSR Daily Activity3D: human-object interactions
- RGBD-HuDaAct (NTU): home monitoring
- MSR Action Pairs: human-object interactions
- MSR Gesture3D: dynamic ASL gestures
- NTU 10-Gesture: static, digits 0-9
- KINECT-ASL (UESTC): static, ASL digits

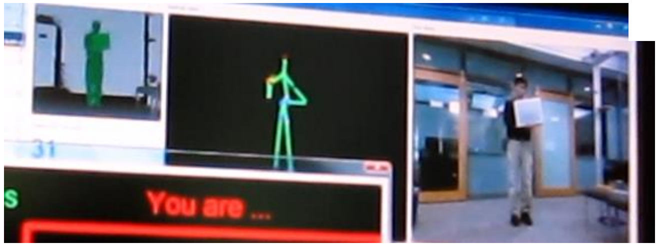
Features

- Skeleton based features
  - Joint angle trajectory
  - EigenJoints, SMIJ, Ho3DJoints,
  - Fourier temporal pyramid of pairwise joint position difference
- Depthmap based features
  - HOG, Bag of 3D points, STOP, DMM-HOG
  - Local occupancy pattern
  - Local Depth Pattern
- RGB+depth

Skeleton Based Features



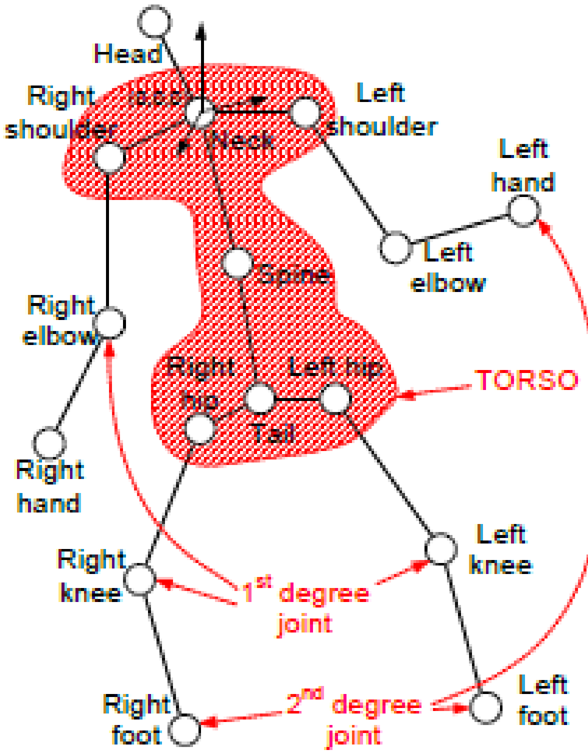
- Kinect outputs 20 joint positions
- Skeletons are noisy
  - Self-Occlusions
  - Object occlusions
  - Side view



- Directly using joint positions does not work well
  - Contrary to the MoCap data

Joint Angle Trajectory

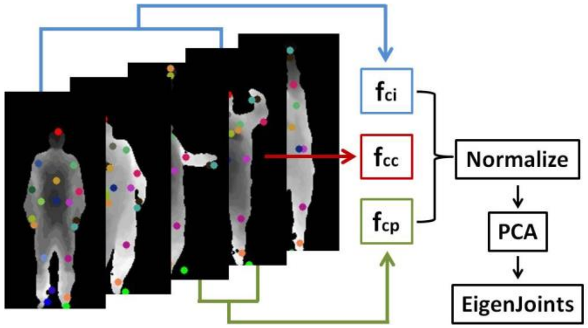
- Torso coordinate frame
  - PCA of torso points
- Joint
  - Spherical angles in torso frame
- FFT over time



M.Raptis, D. Kirovski, H. Hoppe, Real-Time Classification of Dance Gestures from Skeleton Animation, SCA 2011

EigenJoints

- Position difference between joints
  - Within frame
  - Current frame and previous frame
  - Current frame and initial frame
  - PCA: concatenated feature vector
- One concatenated feature vector per frame
- Nearest neighbor classifier
  - Frame-class distance



X. Yang, Y. Tian, EigenJoints-based Acton Recognition Using Baive-Bayes-Nearest Neighbor, HAU3D'2012

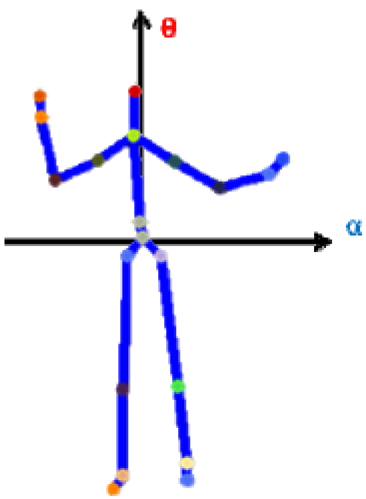
SMIJ: Sequence of Most Informative Joints

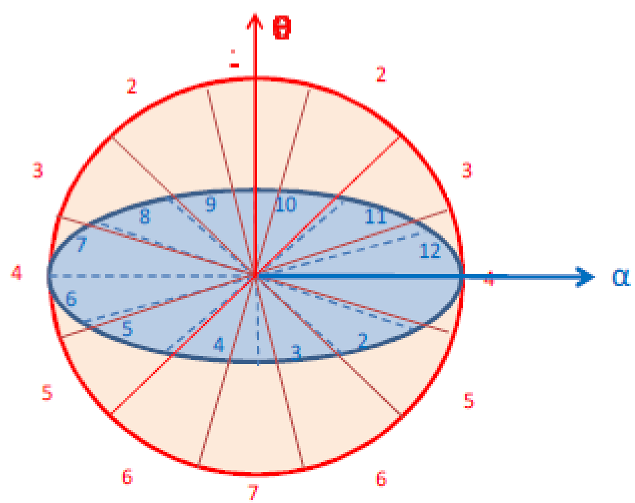
- Given a video clip, find its top 6 most informative joints: variance of joint angle, angular velocity
- The 6 indices form the feature descriptor

F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, SMIJ: A New Representation for Human Skeletal Action Recognition, HAU3D'2012

Histogram of 3D Joint locations (HOJ3D)

- Histogram of spherical coordinates of the joint positions in the HIP coordinate frame
- HIP coordinate frame is not reliable





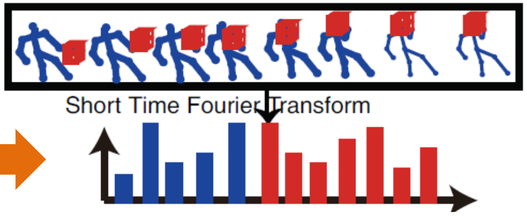
L. Xia, C.C. Chen, J. K. Aggarwal, View Invariant Human Action Recognition Using Histogram of 3D Joints, HAU3D'2012

Fourier Temporal Pyramid of Pairwise Joint Position Difference

- Let $P_i(t)$ denote the 3D position of joint i at frame t

$$P_{ij}(t) = P_i(t) - P_j(t) \qquad 1 \le i, j \le 20, 1 \le t \le T$$

$$FFT\{P_{ij}(t): t \in [1, T]\}$$



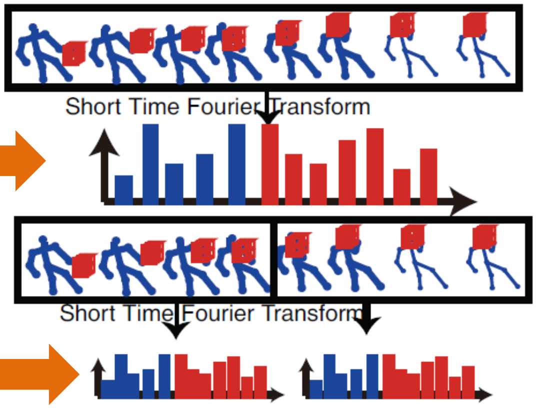
J. Wang, Z. Liu, Y, Wu, J. Yuan, Mining Actionlet Ensemble for Action Recognition with Depth Cameras, CVPR 2012

Fourier Temporal Pyramid of Pairwise Joint Position Difference

- Let $P_i(t)$ denote the 3D position of joint i at frame t

$$P_{ij}(t) = P_i(t) - P_j(t) \qquad 1 \le i, j \le 20, 1 \le t \le T$$

$$FFT\{P_{ij}(t): t \in [1, T]\}$$



  - Divide [1,T] into [1,T/2] and [T/2, T]

$$FFT\{P_{ij}(t): t \in [1, \frac{T}{2}]\} \quad FFT\{P_{ij}(t): t \in [\frac{T}{2}, T]\}$$

J. Wang, Z. Liu, Y, Wu, J. Yuan, Mining Actionlet Ensemble for Action Recognition with Depth Cameras, CVPR 2012

Fourier Temporal Pyramid of Pairwise Joint Position Difference

- Let $P_i(t)$ denote the 3D position of joint i at frame t

$$P_{ij}(t) = P_i(t) - P_j(t) \qquad 1 \le i, j \le 20, 1 \le t \le T$$

$$FFT\{P_{ij}(t): t \in [1, T]\}$$

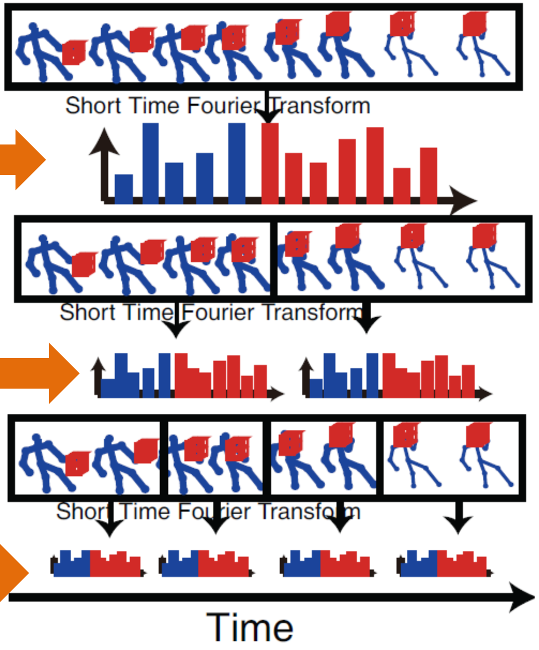
- Divide [1,T] into [1,T/2] and [T/2, T]

$$FFT\{P_{ij}(t): t \in [1, \frac{T}{2}]\} \quad FFT\{P_{ij}(t): t \in [\frac{T}{2}, T]\}$$

- Further divide [1,T] into 4 segments

$$FFT\{P_{ij}(t): t \in [1, \frac{T}{4}]\} \quad FFT\{P_{ij}(t): t \in [\frac{T}{4}, \frac{T}{2}]\}$$

$$FFT\{P_{ij}(t): t \in [\frac{T}{2}, \frac{3T}{4}]\} \quad FFT\{P_{ij}(t): t \in [\frac{3T}{4}, T]\}$$



J. Wang, Z. Liu, Y, Wu, J. Yuan, Mining Actionlet Ensemble for Action Recognition with Depth Cameras, CVPR 2012

Features

- Skeleton based features
  - Joint angle trajectory
  - EigenJoints, SMIJ, Ho3DJoints,
  - Fourier temporal pyramid of pairwise joint position difference
- Depthmap based features
  - HOG, DMM-HOG
  - Spin Image
  - Bag of 3D points
  - Spacetime Occupancy Pattern, local occupancy pattern
  - Local Depth Pattern
  - Histogram of Oriented Normal Vectors (HONV), Histogram of 3D Facets
  - Histogram of Oriented 4D Normal vectors (HON4D)
- RGB+depth

Depthmap Based Features

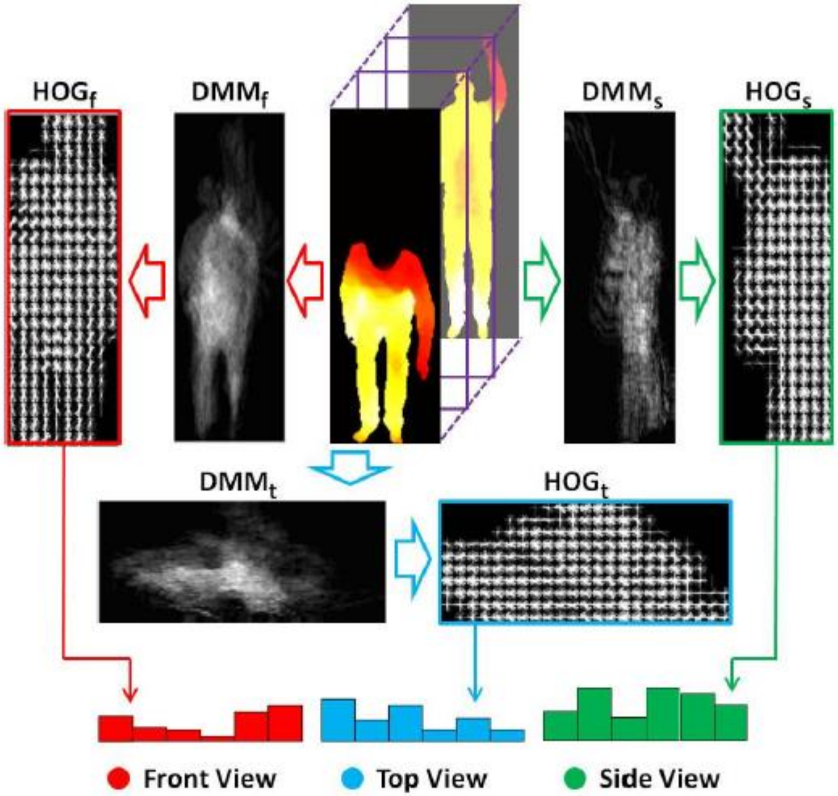
- Isn't skeleton feature sufficient?
  - No, because
    - Skeleton features are noisy, and sometimes missing
    - Cannot handle human-object interactions:
      - No info on the object that a person is holding
- Many 3D shape descriptors have been developed for shape retrieval
  - Crease Histograms
  - Shape Distributions
  - Extend Gaussian Images
  - Shape Histograms
  - Spherical Extent Functions

Treating Depth Map as Grey Image

- Features used for 2D videos
  - HoG
  - SIFT
  - STIPs + HOGHOF (Laptev et al.)
  - Kernel descriptor (Bo et al. CVPR 2011)
- Works quite well for 3D object recognition
  - RGB-D Object Dataset: http://www.cs.washington.edu/rgbd-dataset/

HOG on Depth Motion Maps (DMM-HOG)

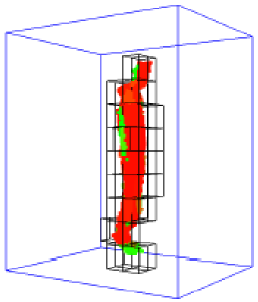
- Depth motion map (DMM)
  - Frame difference
  - Thresholding
  - Aggregation over time
- One DMM per view
  - Front
  - Top
  - Side



X. Yang, C. Zhang, and Y. Tian, Recognizing Actions Using Depth Motion Maps-based Histogram of Oriented Gradients, ACMMM12

STOP: Space-Time Occupancy Pattern

- Given a 3D point cloud and a 3D box
  - Partition the box into 3D grid with M*N*L cells
  - For cell (m,n,l), denote c(m,n,l) to be the number of points in the cell.
  - Feature $f(m,n,l) = \begin{cases} 1, if\ c(m,n,l) \geq \mu \\ \frac{c(m,n,l)}{\mu}, otherwise \end{cases}$

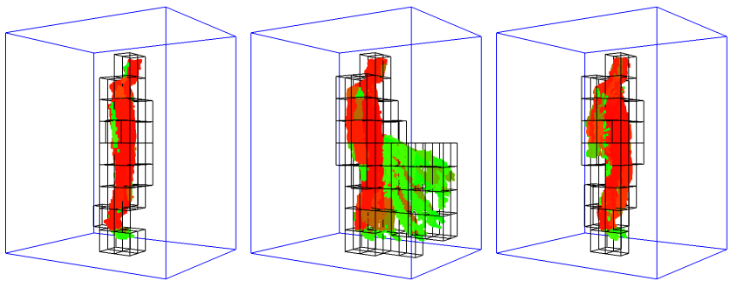


  - f(m,n,l) over all the cells forms a feature vector with dimensionality M*N*L

Vieira et al, STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences, CIARP 2012

STOP: Space-Time Occupancy Pattern

- Assuming the person is stationary
- The depthmaps over time forms a 4D spacetime volume
- Partition the 4D volume into 4D spacetime cells

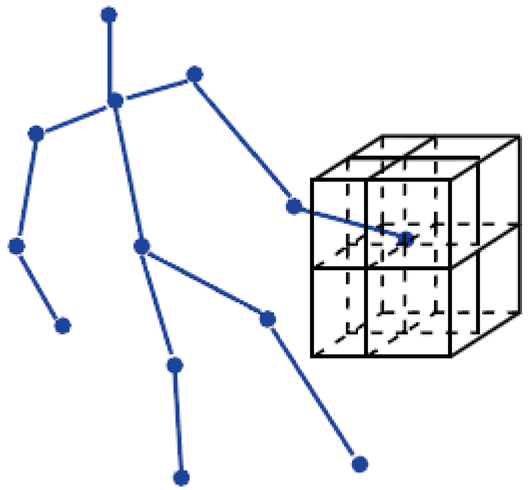


- E.g. 10x10x10x3

Vieira et al, STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences, CIARP 2012

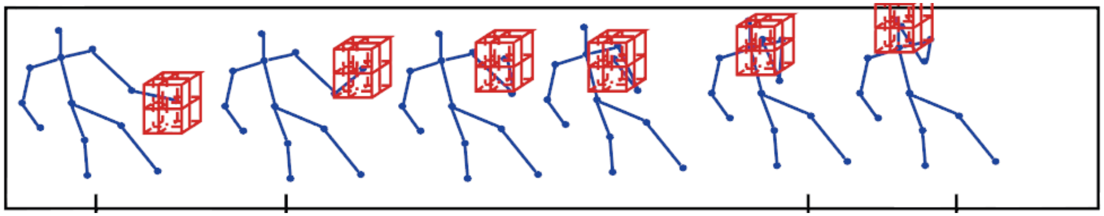
Local Occupancy Pattern (LOP)

- For each joint position
  - Create a local box centered at the point
  - Compute an occupancy pattern feature descriptor
- 20 LOPs per frame



LOP Over Time

- Given a joint j, it has a corresponding LOP feature vector per frame



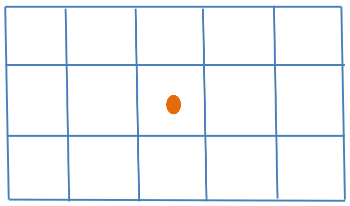
- Let $f_{j,t}(m,n,l)$ denote the occupancy value of cell (m,n,l) for joint j at frame t.
- Pyramid_FFT($f_{j,t}(m,n,l)$: $t \in [1,T]$) is the LOP feature vector of the sequence for joint j.
- Concatenation of all the joints' LOPs: overall LOP feature vector.

J. Wang et al. CVPR 2012

Local Depth Pattern (LDP)

- Form a local window (patch) centered at the interest point. The patch size is scaled inversely by the depth of the interest point
- Divide the patch into a grid
- Compute average depth value of all the valid pixels in each cell
- Difference of the average depth values for every cell pair

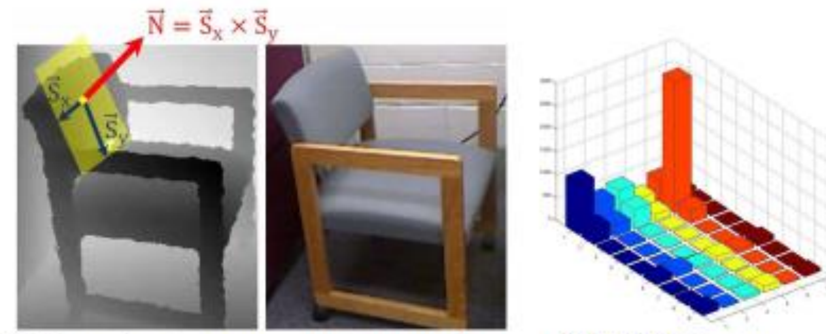
Dimension is $\binom{N_x \times N_y}{2}$



Y. Zhao, Z. Liu, L. Yang, H. Chen, Combining RGB and Depth Features for Human Activity Recognition, APSIPA2012

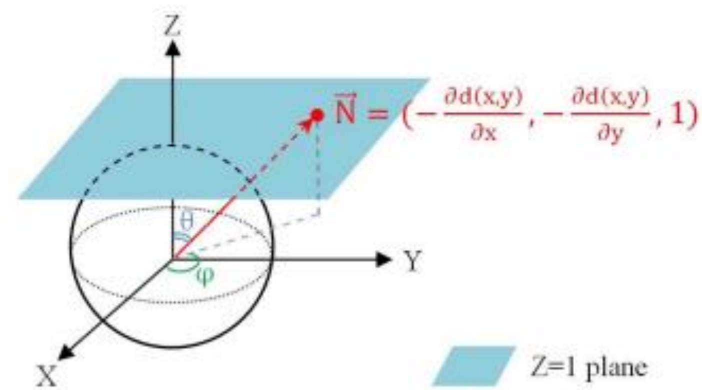
Histogram of Oriented Normal Vectors (HONV)

- Estimate a normal vector for each point
- Obtain a 2D histogram per patch



(a) Normal vector to represent tangent plane at $(x, y, d(x, y))$.

(b) HONV feature



(c) Zenith angle θ and azimuthal angle φ of a normal vector

Tang et al, Histogram of oriented normal vectors for object recognition with a depth sensor, ACCV 2012

Histogram of 3D Facets (H3DF)

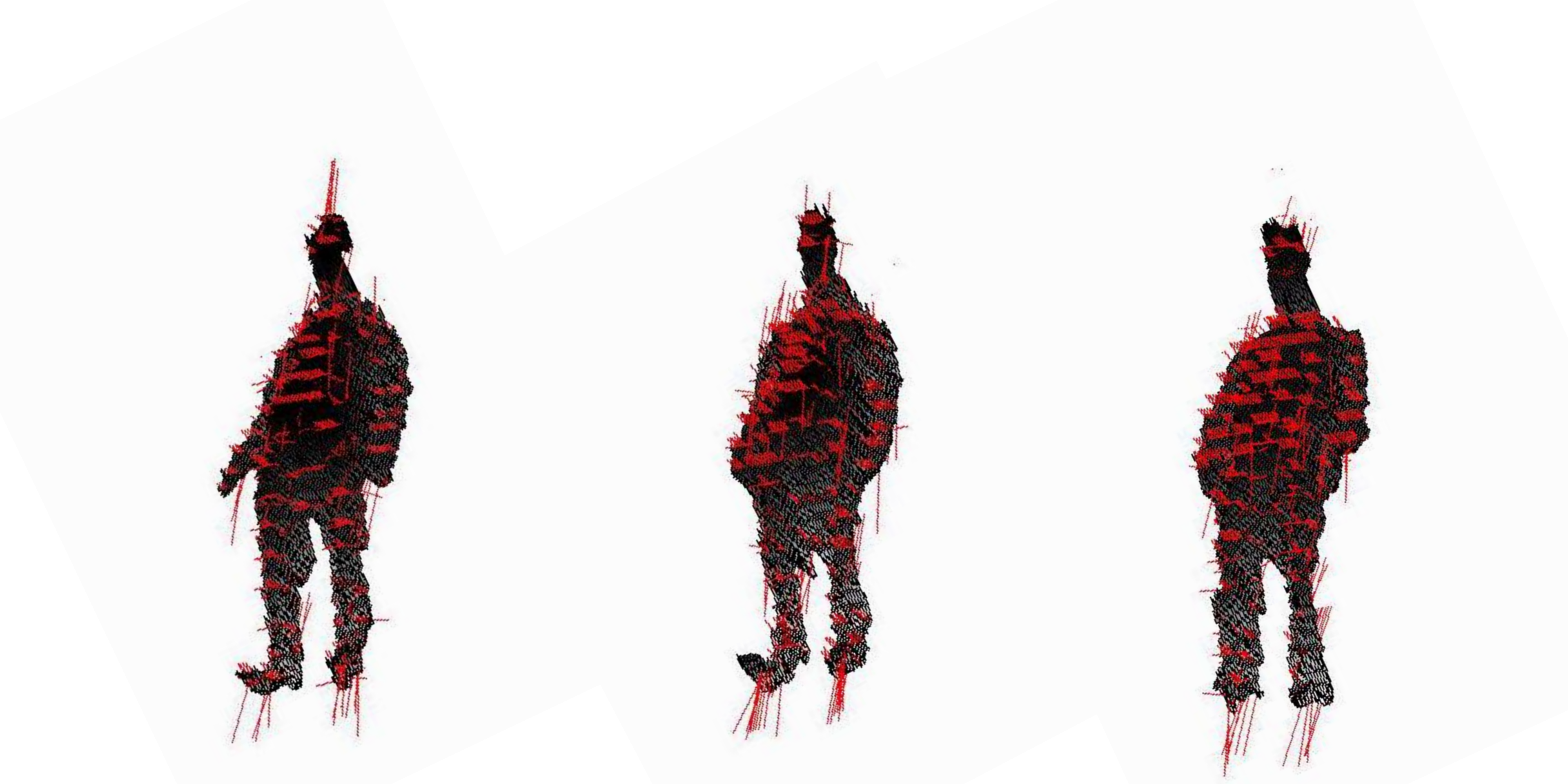
- Estimate normal vectors (similar to HONV)
- Use a different pooling scheme
- Designed for hand gesture recognition
- For details, go to Thursday’s special session on sign language

C. Zhang, X. Yang, Y. Tian, Histogram of 3D facets: a characteristic descriptor for hand gesture recognition, FG2013

Histogram of Oriented 4D Normals (HON4D)

- $\vec{n}$: Captures shape
- $\Delta\vec{n}$: Captures motion

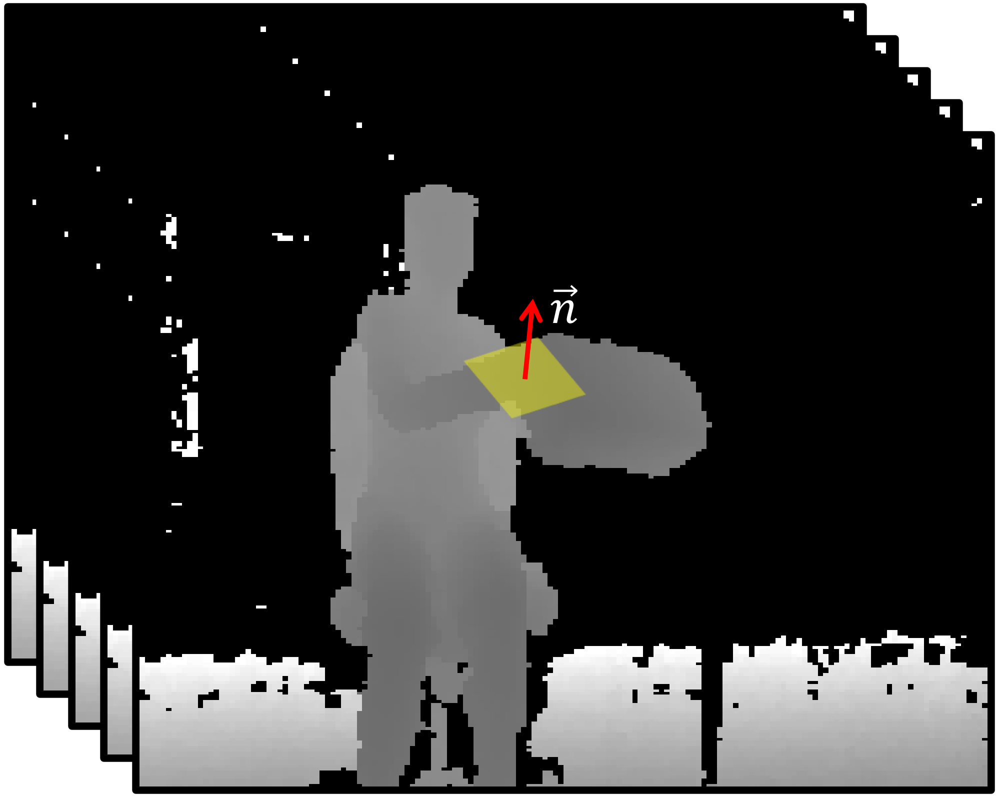
O. Oreifej, Z. Liu, HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, CVPR 2013



HON4D

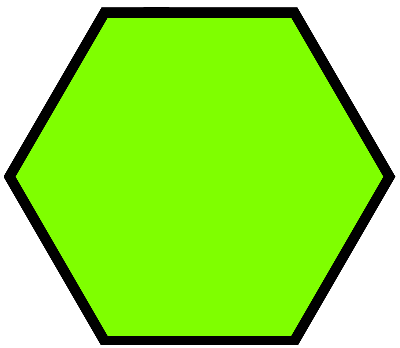
- $\vec{n} = (\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial z}{\partial t}, -1)$

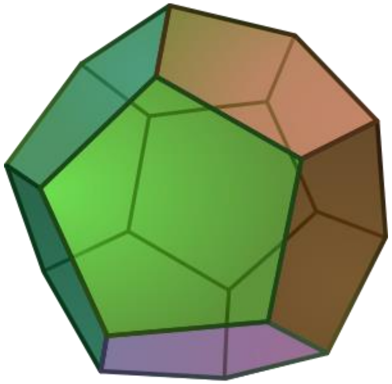
- Captures both shape and motion

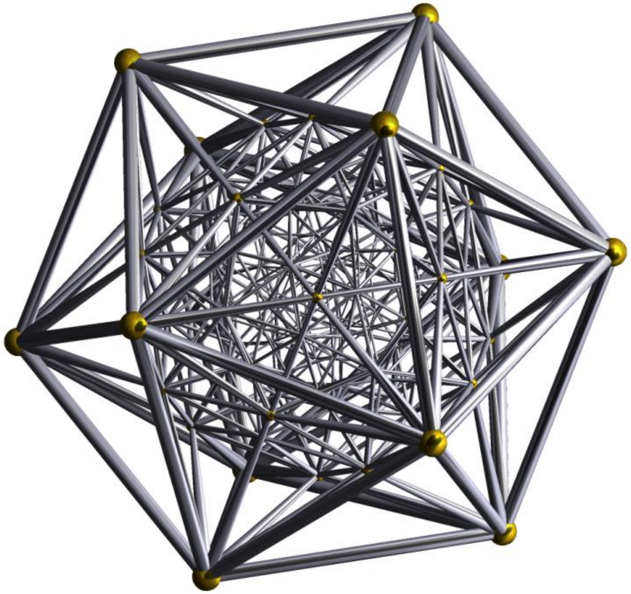


4D Space Quantization

- Polygons

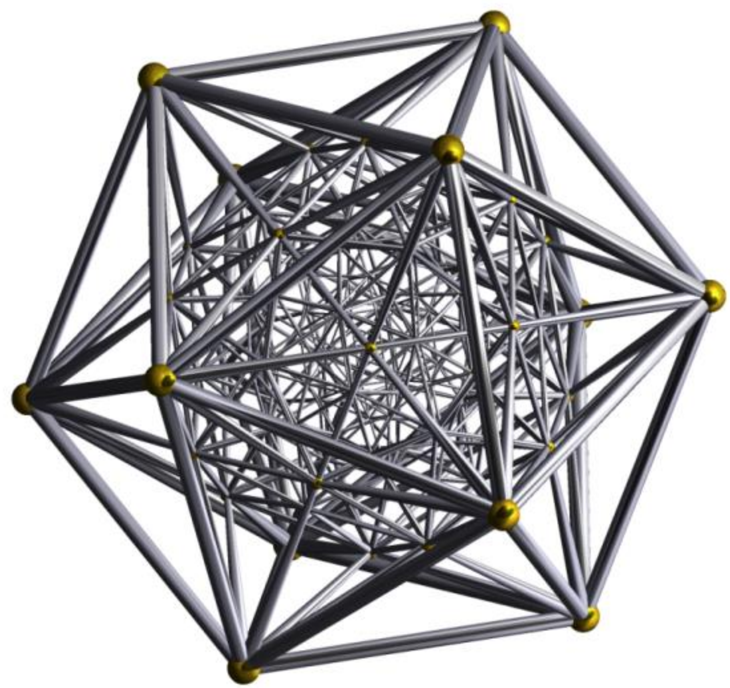

2D: Polygon


3D: Polyhedron


4D: Polychoron

600-cell

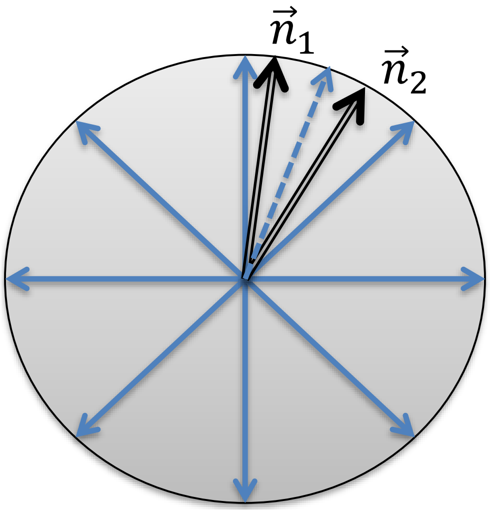
- 120 vertices
  - 16 permutations of
    $(\pm\frac{1}{2},\pm\frac{1}{2},\pm\frac{1}{2},\pm\frac{1}{2})$
  - 8 permutations of
    $(0,0,0,\pm 1)$
  - 96 even permutations of
    $\frac{1}{2}(\pm\phi,\pm 1,\pm 1/\phi,0)$
- Vertices
  - ⇨ Projectors for HONV 4D



600-cell: 120 vertices

4D Quantization

- Is the uniform 4D quantization optimal?
  - Unlikely



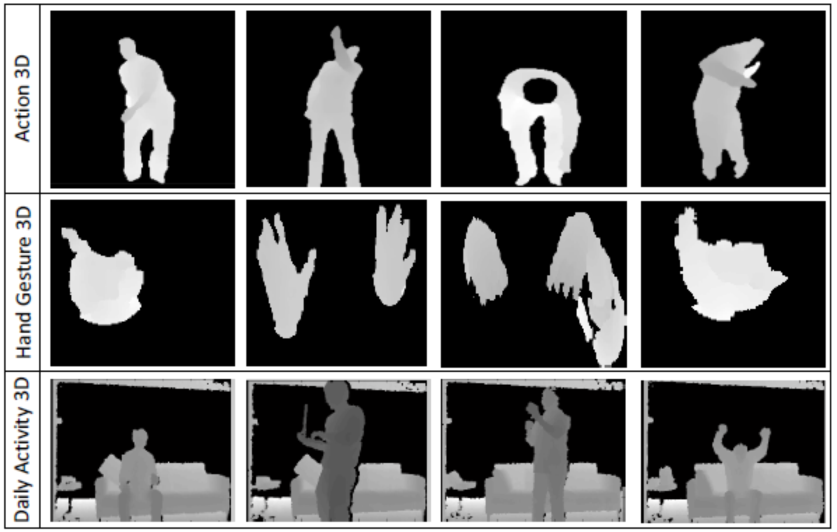
  - Non-uniform projectors

O. Oreifej, Z. Liu, HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, CVPR 2013

Experiments (SVM)

MSR Action3D

Method	Accuracy %
HON4D + D_{disc}	**88.89**
HON4D	85.85
Jiang et al. [24]	88.20
Jiang et al. [23]	86.50
Yang et al. [26]	85.52
Dollar [5] + BOW	72.40
STIP [10] + BOW	69.57
Vieira et al. [21]	78.20
Klaser et al. [9]	81.43



MSR Gesture3D

Method	Accuracy %
HON4D + D_{disc}	**92.45**
HON4D	87.29
Jiang et al. [23]	88.50
Yang et al. [26]	89.20
Klaser et al. [9]	85.23

MSR DailyActivity3D

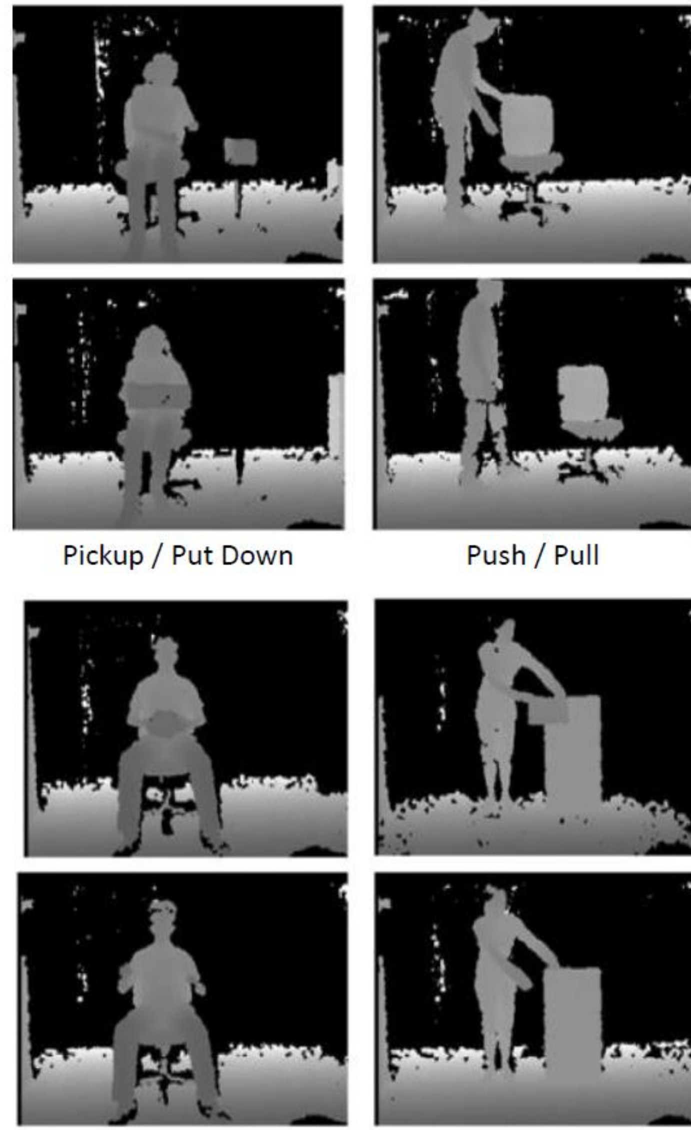
As a local descriptor per joint: 80.00%
Compared with LOP: 67.50%

MSR Action Pairs

Skeleton motions are the same for each pair

- Pick up a box – Put down a box
- Lift a box – Place a box
- Push a chair – Pull a chair
- Wear a hat – Take off a hat
- Put on a backpack – Take off a backpack
- Stick a poster – Remove a poster

Method	Accuracy %
HON4D + D_{disc}	**96.67**
HON4D	93.33
Wang et al (Skeleton + LOP)	63.33
(Skeleton + LOP + Pyramid)	82.22
Yang et al. DMM-HOG	66.11



Pickup / Put Down

Push / Pull

Wear /Take off

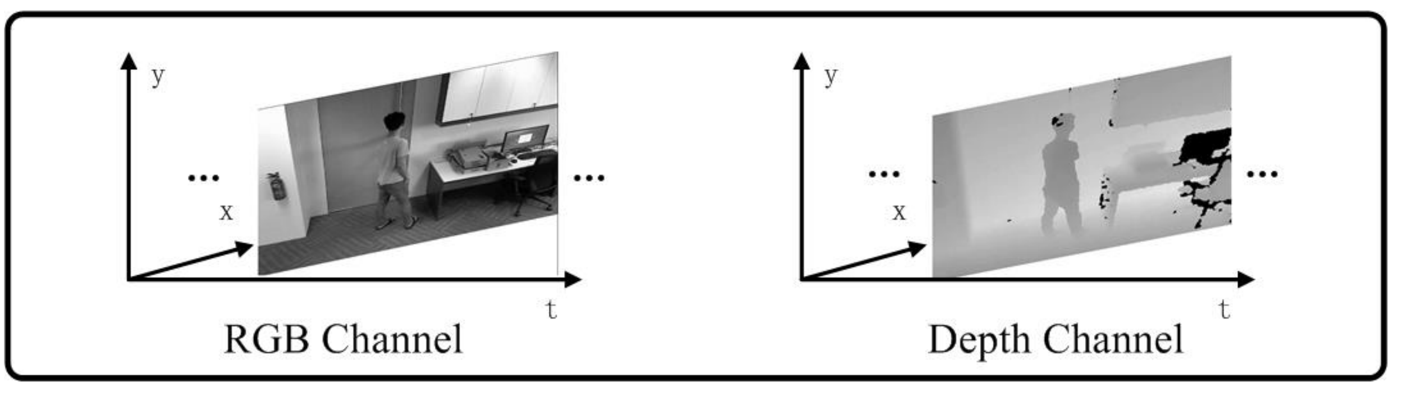
Stick / Remove

Features

- Skeleton based features
  - Joint angle trajectory
  - EigenJoints, SMIJ, Ho3DJoints,
  - Fourier temporal pyramid of pairwise joint position difference
- Depthmap based features
  - HOG, DMM-HOG
  - Spin Image
  - Bag of 3D points
  - Spacetime Occupancy Pattern, local occupancy pattern
  - Local Depth Pattern
  - Histogram of Oriented Normal Vectors (HONV), Histogram of 3D Facets
  - Histogram of Oriented 4D Normal vectors (HON4D)
- RGB+depth

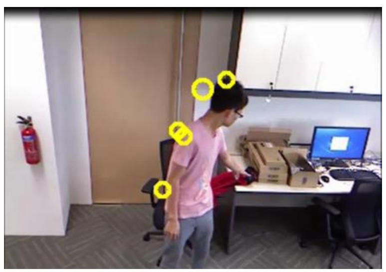
RGB + Depth

- Global feature – human tracking
  - One descriptor for the RGB channel
  - One descriptor for the depth channel
  - Concatenate RGB descriptor and depth descriptor

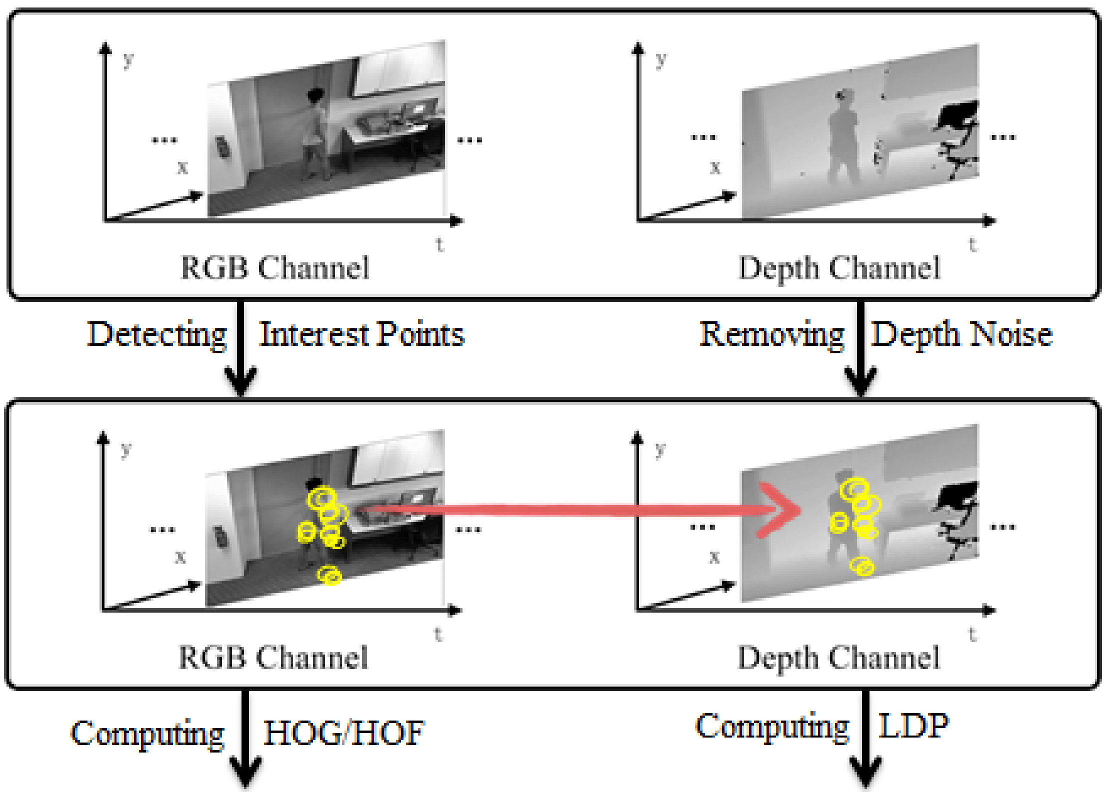


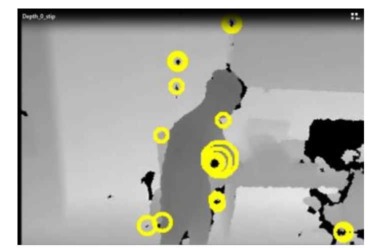
RGB + Depth

- Local feature
  - Detecting interest points from which channel?

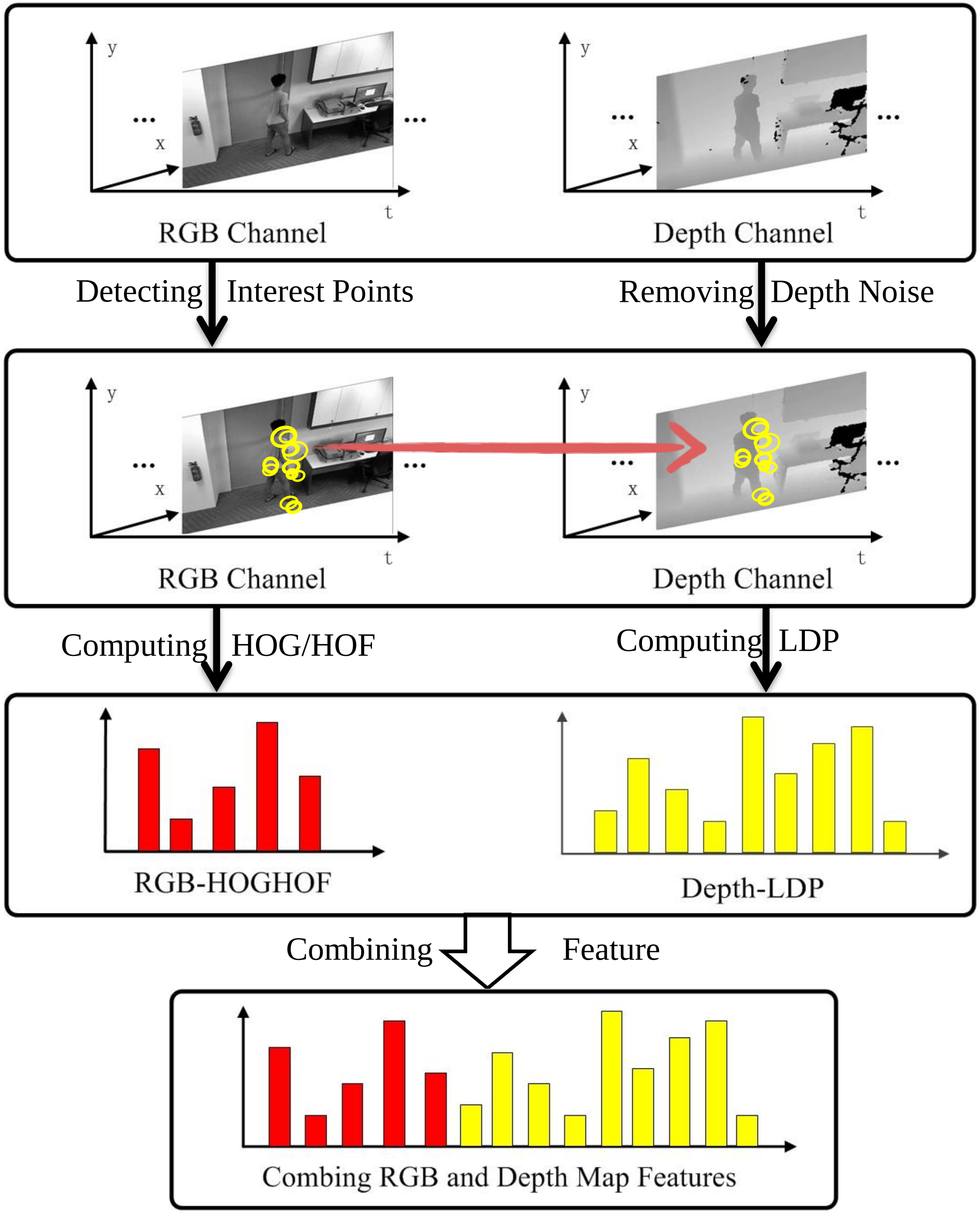


RGB-STIP

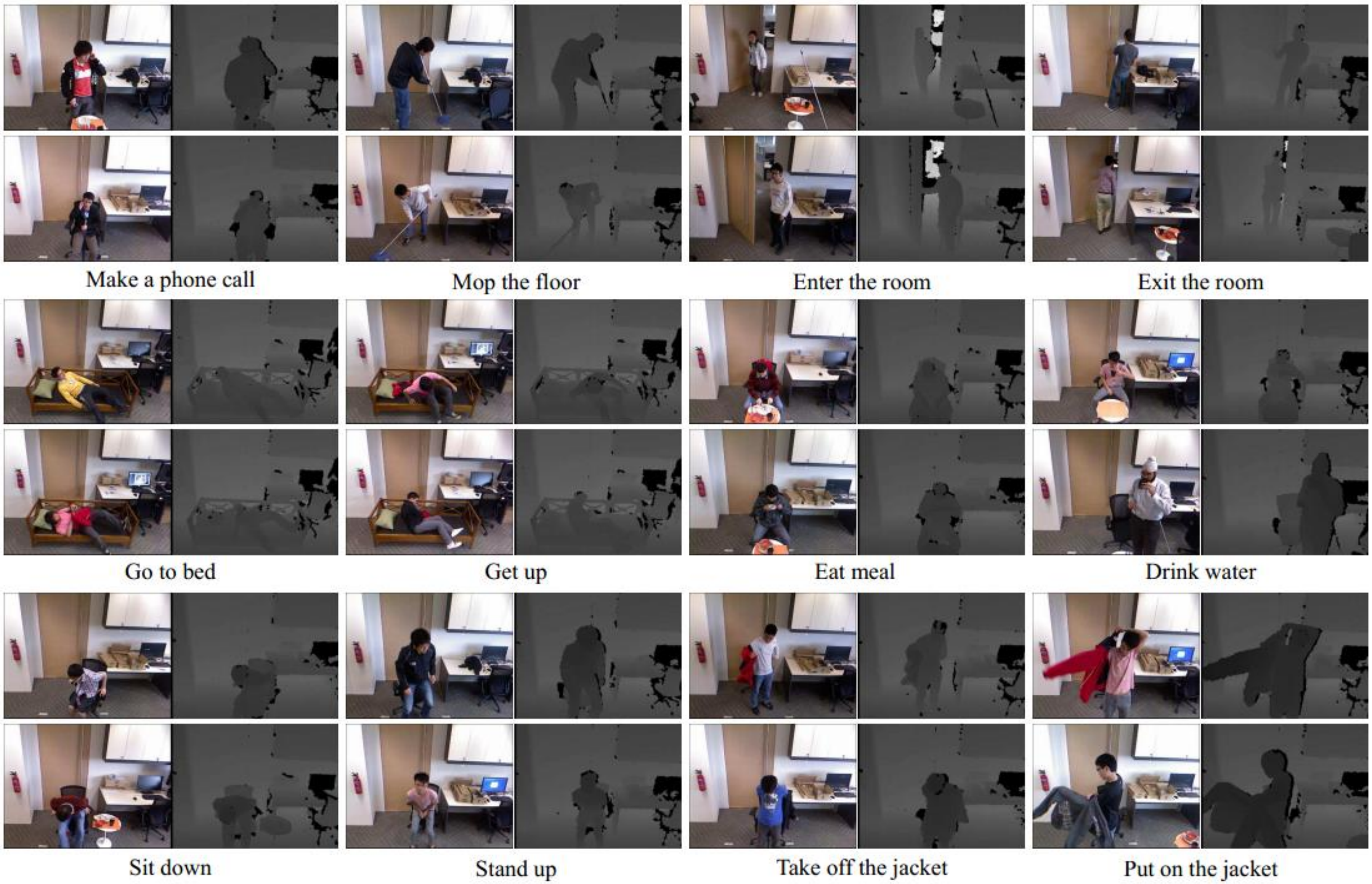




Depth-STIP


RGB Channel
Depth Channel
Detecting Interest Points
Removing Depth Noise
RGB Channel
Depth Channel
Computing HOG/HOF
Computing LDP
RGB-HOGHOF
Depth-LDP
Combining Feature
Combing RGB and Depth Map Features

RGBD-HuDaAct



B.Ni, G.Wang, P.Moulin, ICCV Workshop 2011

Results

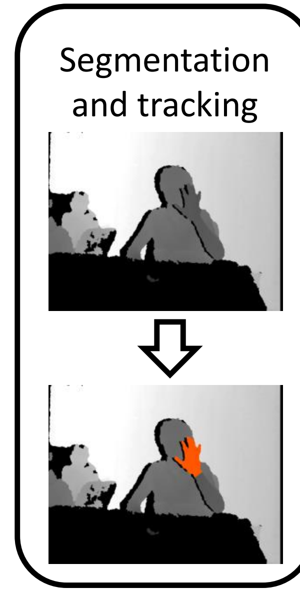
Method	Accuracy(%)
DLMC-STIPs[14]	81.5
3D-MHIs[14]	70.5
Zhao et al	**89.1**

DLMC: Depth-Layered Multi-Channel

[14]: *B.Ni, G.Wang, P.Moulin, ICCV Workshop 2011*

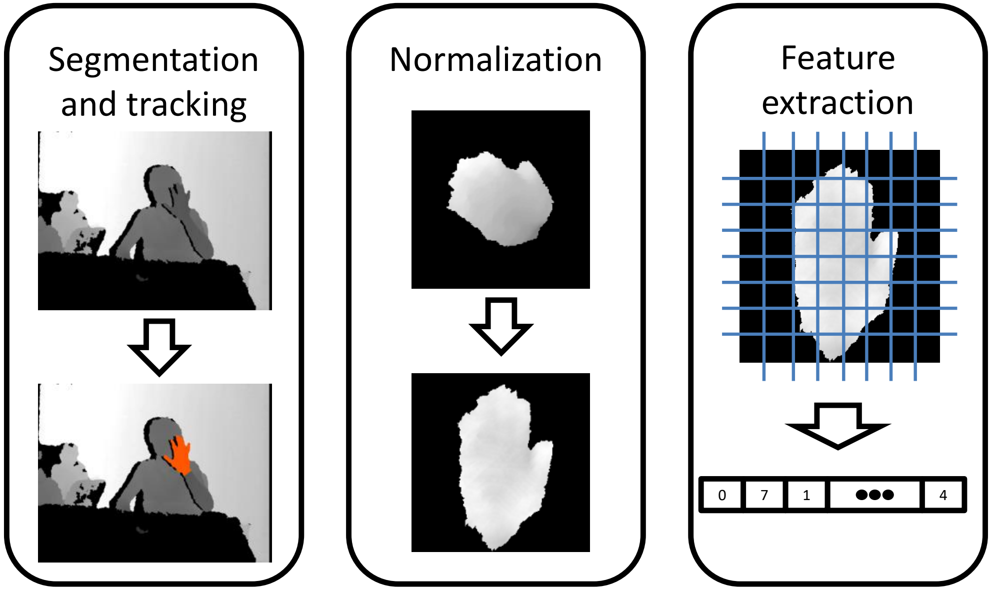
Hand Segmentation and Feature Extraction

- Hand gesture recognition
  - Info at the finger level
- Hand segmentation
  - Depth thresholding
  - Detect wrist and segment the hand
- Feature extraction
  - Depthmap based descriptor
  - Time-series curve (hand contour)



Depthmap Based Descriptor in Hand Region

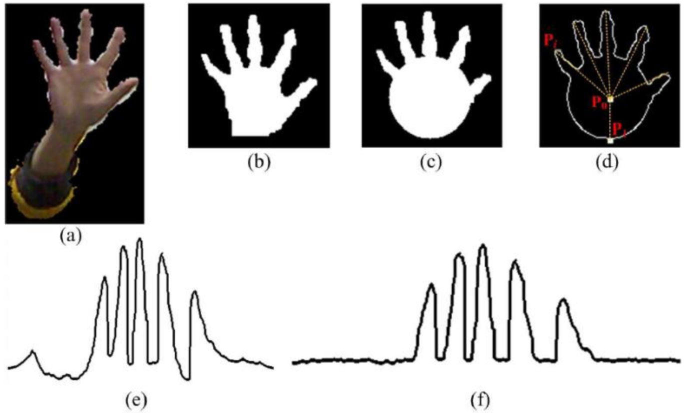
- Find the hand plane
- 2D projection
- 2D Occupancy Pattern



A. Kurakin, Z. Zhang, Z. Liu, A real-time system for dynamic hand gesture recognition with a depth sensor, EUSIPCO 2012

Time-Series Curve (Contour)

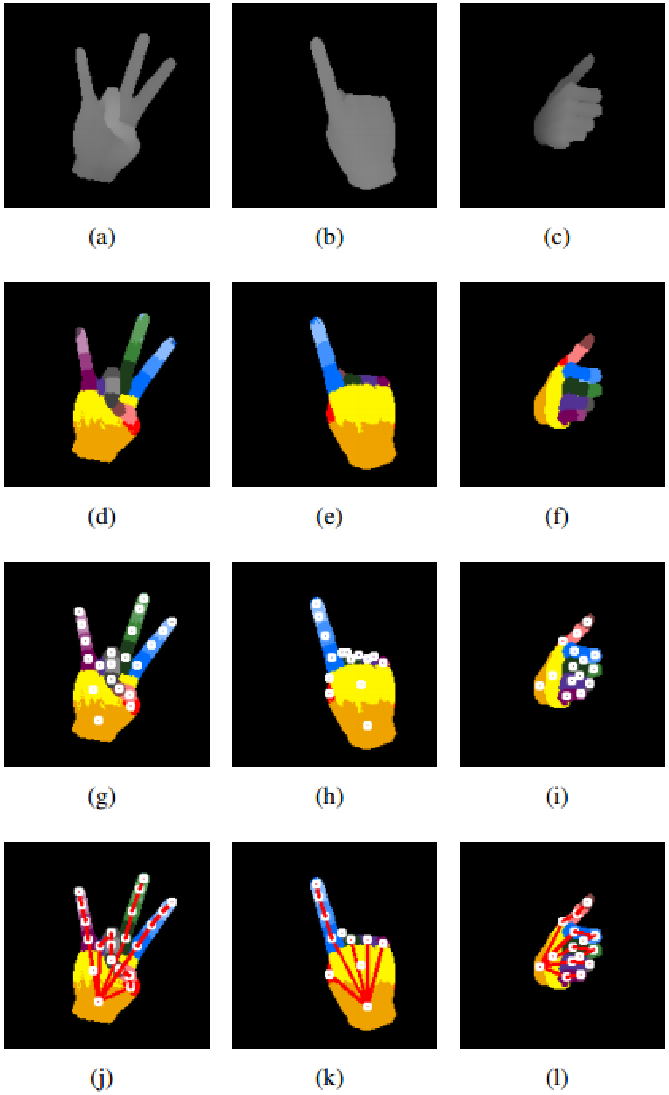
- Requires more accurate wrist segmentation
  - (a) Depth thresholding
  - (b) Detect wrist and segment the hand
  - (c) Remove palm
  - (d) Find contour by edge detection
  - (f) Contour curve with time-series representation



Z, Dai, H. Cheng, Z. Liu, Image-to-class dynamic time warping for 3D hand gesture recognition, ICME2013

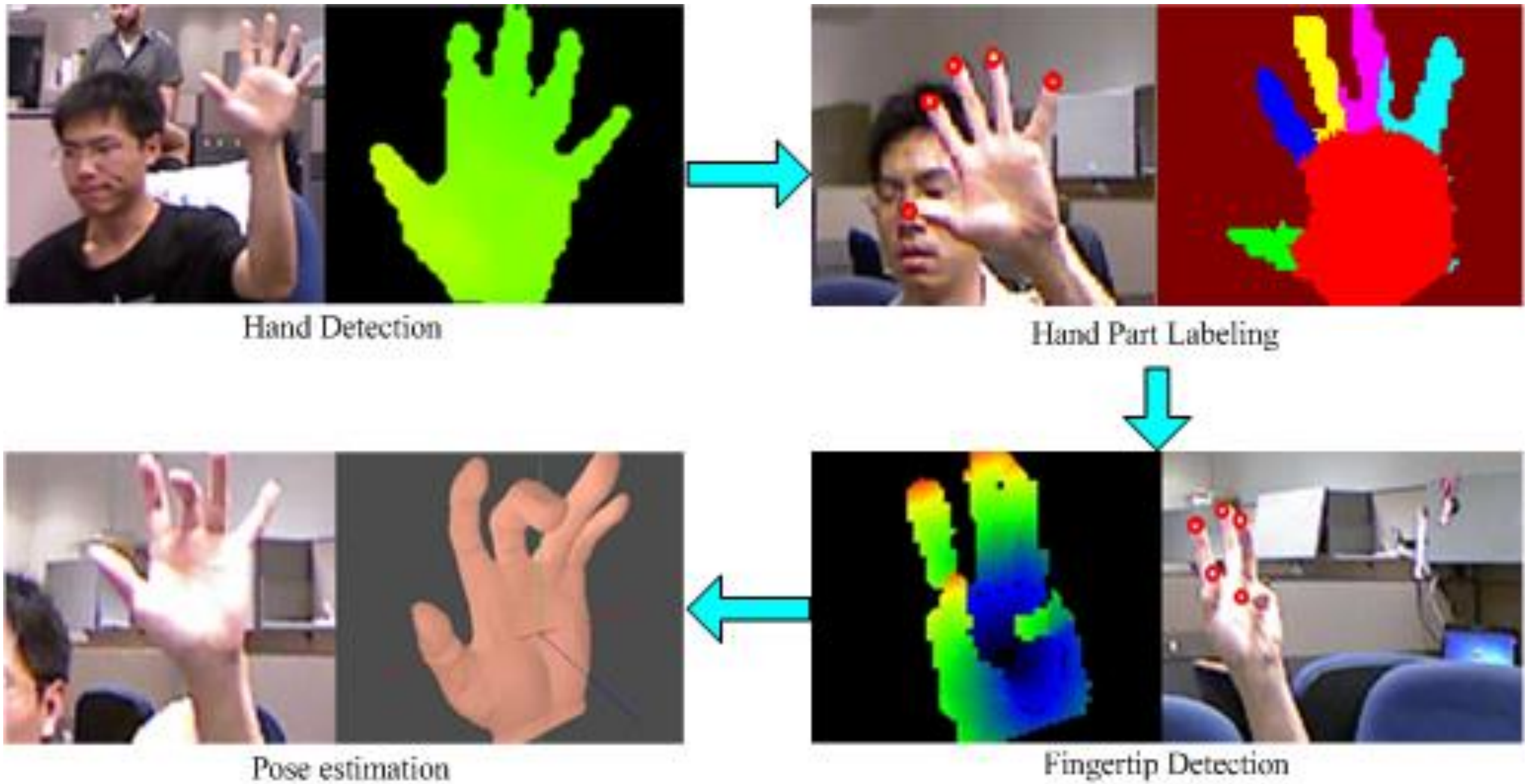
Hand Skeletonization

- Obtain the hand "skeleton"
  - Per pixel classification
  - Similar to Shotton et al's body skeleton detection method
  - Requires lots of training data
    - *Row#1: input*
    - *Row#2: pixel classification*
    - *Row#3: detected joints*
    - *Row#4: detected skeleton*



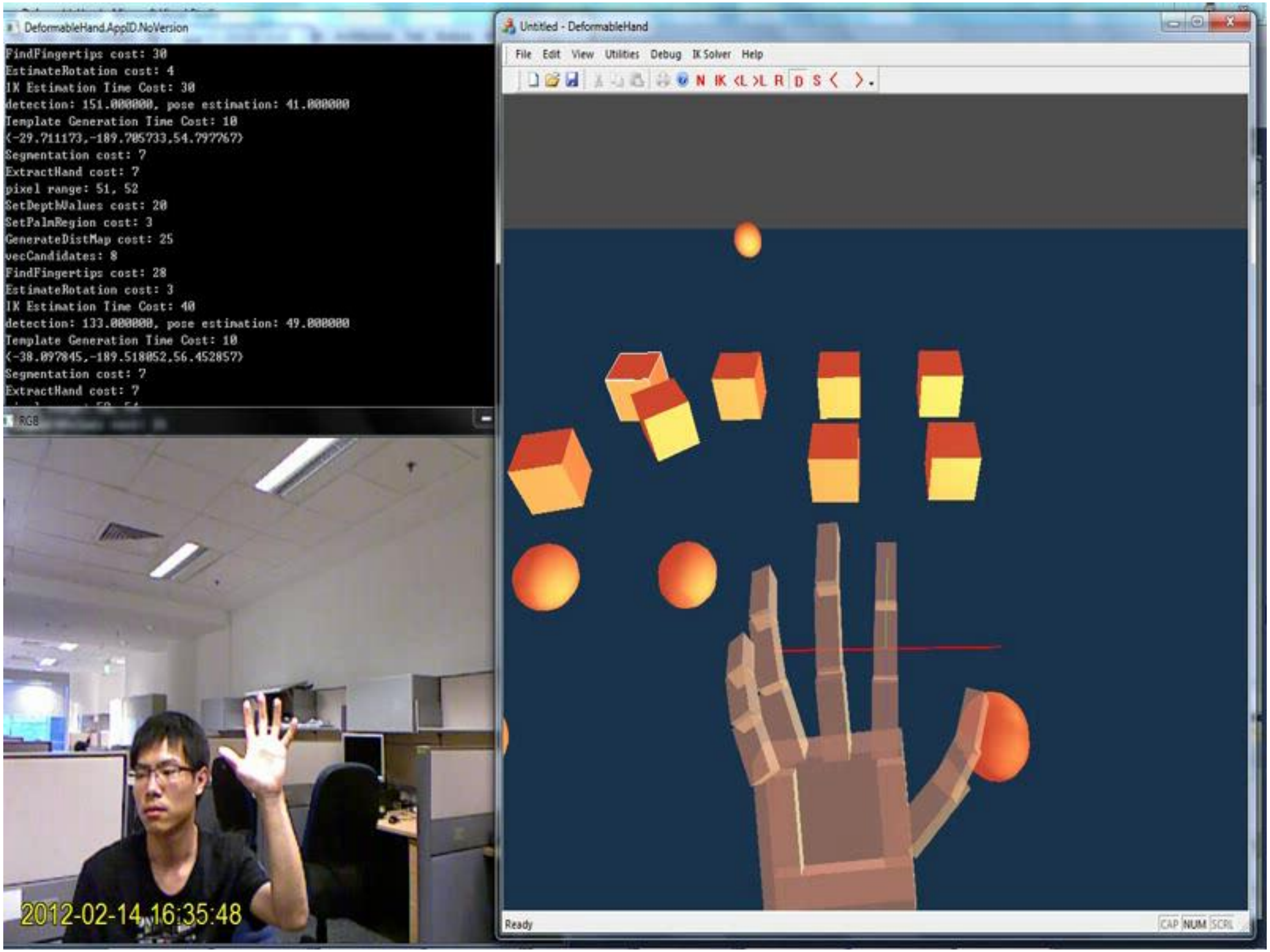
Keskin et al, Real Time Hand Pose Estimation using Depth Sensors, ICCV Workshop on Gesture Rec. 2011

Hand Skeletonization



Hui Liang, Junsong Yuan and Daniel Thalmann, 3D Fingertip and Palm Tracking in Depth Image Sequences, in ACM Int'l Conf. on Multimedia, 2012

Virtual Object Manipulation:



Hui Liang, Junsong Yuan and Daniel Thalmann, Hand pose estimation by combining fingertip tracking and articulated ICP, in SIGGRAPH VRCAI, 2012

Recognition Paradigms

- Direct classification
  - Global feature descriptor: one vector per clip
  - SVM, RF, etc.
- Bag of Words framework
  - Interest points + local feature descriptor
- Actionlet Ensemble
  - J. Wang, Z. Liu, Y, Wu, J. Yuan, CVPR2012
- Random Occupancy Pattern
  - J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, ECCV2012
- Contour Matching (static hand gesture)
- Online recognition
  - Temporal segmentation
  - Action graph, Li et al, TCSVT 2008

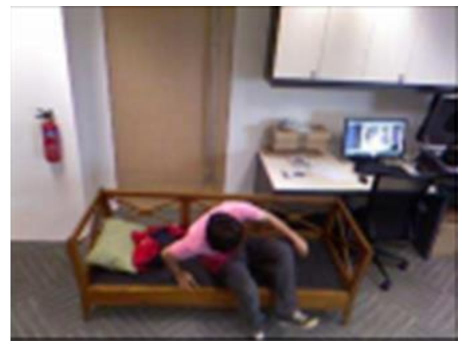
Direct Classification

- Global feature descriptors:
  - One feature vector per video clip
    - SVM, RF, etc.
  - Easier to obtain global feature descriptor for depth sequences than for conventional videos
  - Feasible as long as skeleton tracking works

Recognition Paradigms

- Direct classification
  - Global feature descriptor: one vector per clip
  - SVM, RF, etc.
- Bag of Words framework
  - Interest points + local feature descriptor
- Actionlet Ensemble
  - J. Wang, Z. Liu, Y, Wu, J. Yuan, CVPR2012
- Random Occupancy Pattern
  - J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, ECCV2012
- Contour Matching (static hand gesture)
- Online recognition
  - Temporal segmentation
  - Action graph, Li et al, TCSVT 2008

Bag-of-Feature Framework



- If skeleton tracking is not available
  - Camera looking down
    - RGBD-HuDaAct
  - BoW scheme
    - Detect interest points
    - Obtain a local descriptor per interest point
    - Build a codebook
    - Obtain a word histogram vector per clip
    - Word histogram vectors are used for classification
  - Nearest neighbor: instance-class distance
    - No need to build codebook

Recognition Paradigms

- Direct classification
  - Global feature descriptor: one vector per clip
  - SVM, RF, etc.
- Bag of Words framework
  - Interest points + local feature descriptor
- Actionlet Ensemble
  - J. Wang, Z. Liu, Y. Wu, J. Yuan, CVPR2012
- Random Occupancy Pattern
  - J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, ECCV2012
- Contour Matching (static hand gesture)
- Online recognition
  - Temporal segmentation
  - Action graph, Li et al, TCSVT 2008

Actionlet Ensemble

- Actionlet: a conjunctive (AND) structure on the base features (a subset of joints):
  - base feature: Fourier Pyramid of a joint
  - Joint *i*, overall feature vector G_i:

$$Pyramid\_FFT\{P_{ij}(t): t \in [1, T]\}\ for\ all\ j \neq i,$$

$$Pyramid\_FFT\left\{f_{i,t}(m, n, l): t \in [1, T]\right\}$$

J. Wang, Z. Liu, Y, Wu, J. Yuan, Mining Actionlet Ensemble for Action Recognition with Depth Cameras, CVPR 2012

Measuring the Discriminativity of a Joint

- Given class c, joint i, train a SVM using feature G_i
- Probability that its predicted label is equal to true label (pairwise coupling):

$$P_i(y^{(j)} = c|\boldsymbol{x}^{(j)})$$

- Let S denote a subset of joints->actionlet
- Probably that S predicts the correct label is:

$$P_S(y^{(j)} = c|\boldsymbol{x}^{(j)}) = \prod_{i \in S} P_i(y^{(j)} = c|\boldsymbol{x}^{(j)})$$

- Denote $\mathcal{X}_c$ as $\{j : t^{(j)} = c\}$
  - Data samples with label c
- In order for S to be discriminative for class c
  - $P_S(y^{(j)} = c|\boldsymbol{x}^{(j)})$ should be large for some of the data in $\mathcal{X}_c$
  - And small for other data which does not belong to $\mathcal{X}_c$

Confidence score: $\mathrm{Conf}_S = \max_{j \in \mathcal{X}_c} \log P_S(y^{(j)} = c|\boldsymbol{x}^{(j)})$

Ambiguity score: $\mathrm{Amb}_S = \sum_{j \notin \mathcal{X}_c} \log P_S(y^{(j)} = c|\boldsymbol{x}^{(j)})$

Discriminative Actionlet Mining

Look for actionlets with large confidence score and small ambiguity score

$$\mathrm{Conf}_S = \max_{j \in \mathcal{X}_c} \log P_S(y^{(j)} = c | \boldsymbol{x}^{(j)})$$

$$\mathrm{Amb}_S = \sum_{j \notin \mathcal{X}_c} \log P_S(y^{(j)} = c | \boldsymbol{x}^{(j)})$$

X_c : data items with label c

T_{conf} : *confidence threshold*

T_{amb} : *ambiguity threshold*

Aprior mining process:

Take the set of joints, the feature $\boldsymbol{G}_i$ on each joint i, the number of the classes C, thresholds T_{conf} and T_{amb}.
Train the base classifier on the features $\boldsymbol{G}_i$ of each joint i.
for *Class* $c = 1$ *to* C **do**
  Set P_c, the discriminative actionlet pool for class c to be empty : $P_c = \{\}$. Set $l = 1$.
  repeat
    Generate the l-actionlets by adding one joint into each $(l-1)$-actionlet in the discriminative actionlet pool P_c.
    Add the l-actionlets whose confidences are larger than T_{conf} to the pool P_c.
    $l = l + 1$
  until *no discriminative actionlet is added to* P_c *in this iteration*;
  remove the actionlets whose ambiguities are larger than T_{amb} in the pool P_c.
end
return the discriminative actionlet pool for all the classes

Learning Actionlet Ensemble

- Multiclass-MKL
- Assume there are *p* actionlets, each corresponding to a kernel

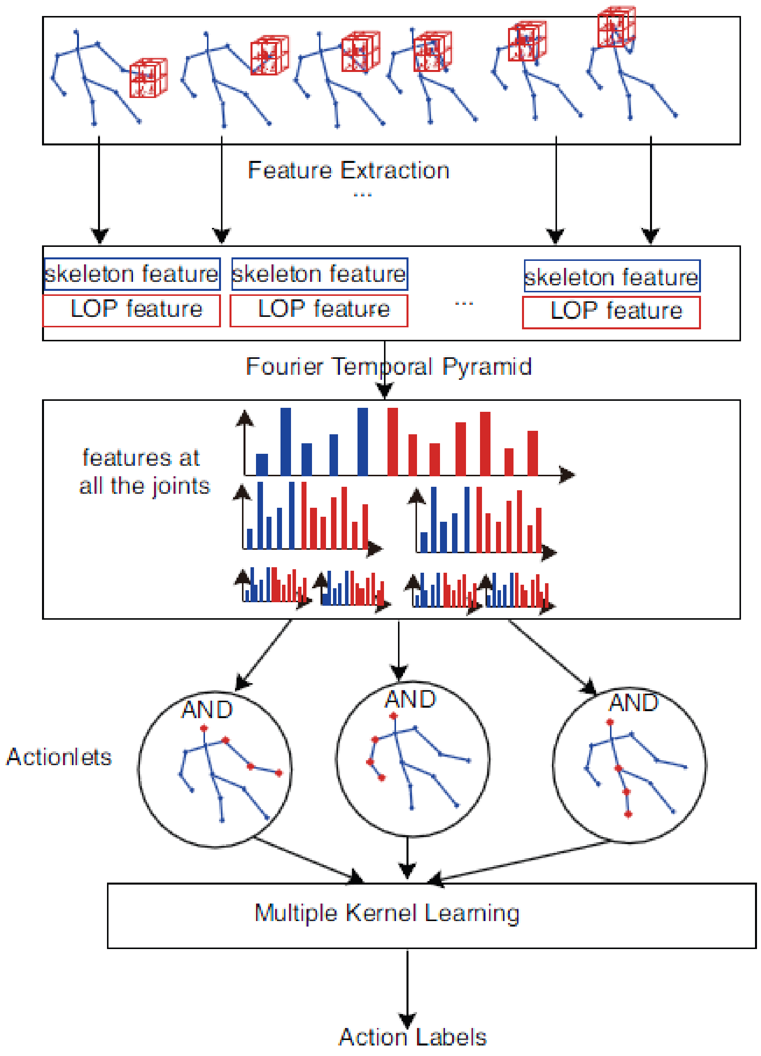
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \beta_k K_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$f_{\text{final}}(\boldsymbol{x}, y) = \sum_{k=1}^{p} \left[\beta_k \langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}, y)\rangle + b_k\right]$$

$$\min_{\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{b}, \xi} \frac{1}{2}\|\boldsymbol{\beta}\|_1^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } \forall i : \xi_i = \max_{u \neq y_i} l(f_{\text{final}}(\boldsymbol{x}^{(i)}, y^{(i)}) - f_{\text{final}}(\boldsymbol{x}^{(i)}, u))$$

Overall Framework



Datasets

- MSR Action3D

  - Sports actions
  - 20 classes, 10 subjects
  - Each subject performing each action 1-3 times
  - 567 depth sequences in total
- MSR Daily Activity

  - Daily activities
    - Eat, drink, real book, call, use laptop,etc
    - Human-object interactions
  - 16 classes, 10 subjects, each performing 2 times

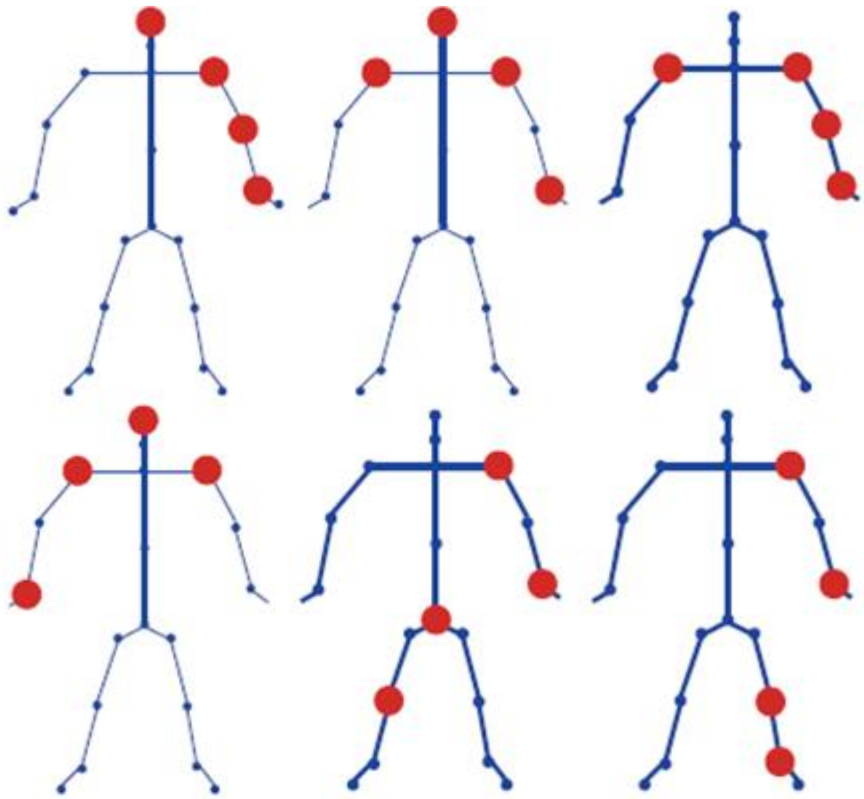
MSR Action3D

Method	Accuracy
Action graph + bag of 3D points (Li et al, CVPR4HB’10)	74.7%
Recurrent Neural Network (Martens&Sutskever’11)	42.5%
Dynamic Time Warping	54%
STOP (Vieira et al, CIARP’12)	84.8%
Actionlet Ensemble (Wang et al, CVPR’12)	88.2%
Joint Angle Trajectory (Raptis’al SCA11, Miranda’al SIBGRAPI12)	80.3%
EigenJoints (Yang&Tian, HAU3D’12)	81.4%
SMIJ (Ofli et al, HAU3D’12)	33.33%
Ho3DJoints(Xia et al, HAU3D’12)	78.97%
DMM-HOG (Yang et all, ACMMM’12)	85.52%
HON4D (Oreifej&Liu, CVPR’13)	88.89%

MSR Daily Activity

Method	Accuracy
Dynamic time warping	54%
LOP feature only	42.5%
Joint feature only	68%
SVM on both features (no actionlets)	78%
Actionlet Ensemble	85.75%
SVM on skeleton + local HON4D (no actionlets)	80.00%

Example Actionlets



Learned from MSR Daily Activity Dataset

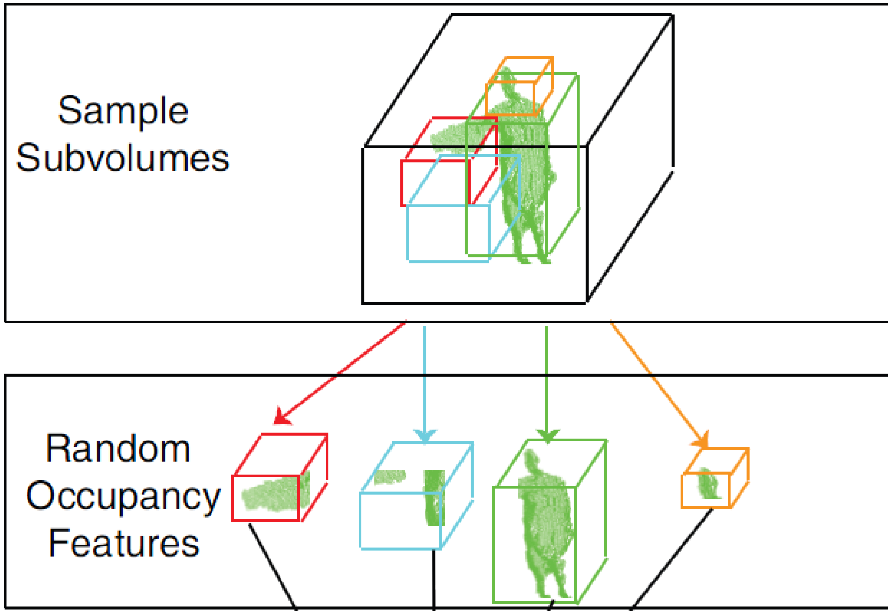
Recognition Paradigms

- Direct classification
  - Global feature descriptor: one vector per clip
  - SVM, RF, etc.
- Bag of Words framework
  - Interest points + local feature descriptor
- Actionlet Ensemble
  - J. Wang, Z. Liu, Y, Wu, J. Yuan, CVPR2012
- Random Occupancy Pattern
  - J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, ECCV2012
- Contour Matching (static hand gesture)
- Online recognition
  - Temporal segmentation
  - Action graph, Li et al, TCSVT 2008

Randomized Occupancy Pattern

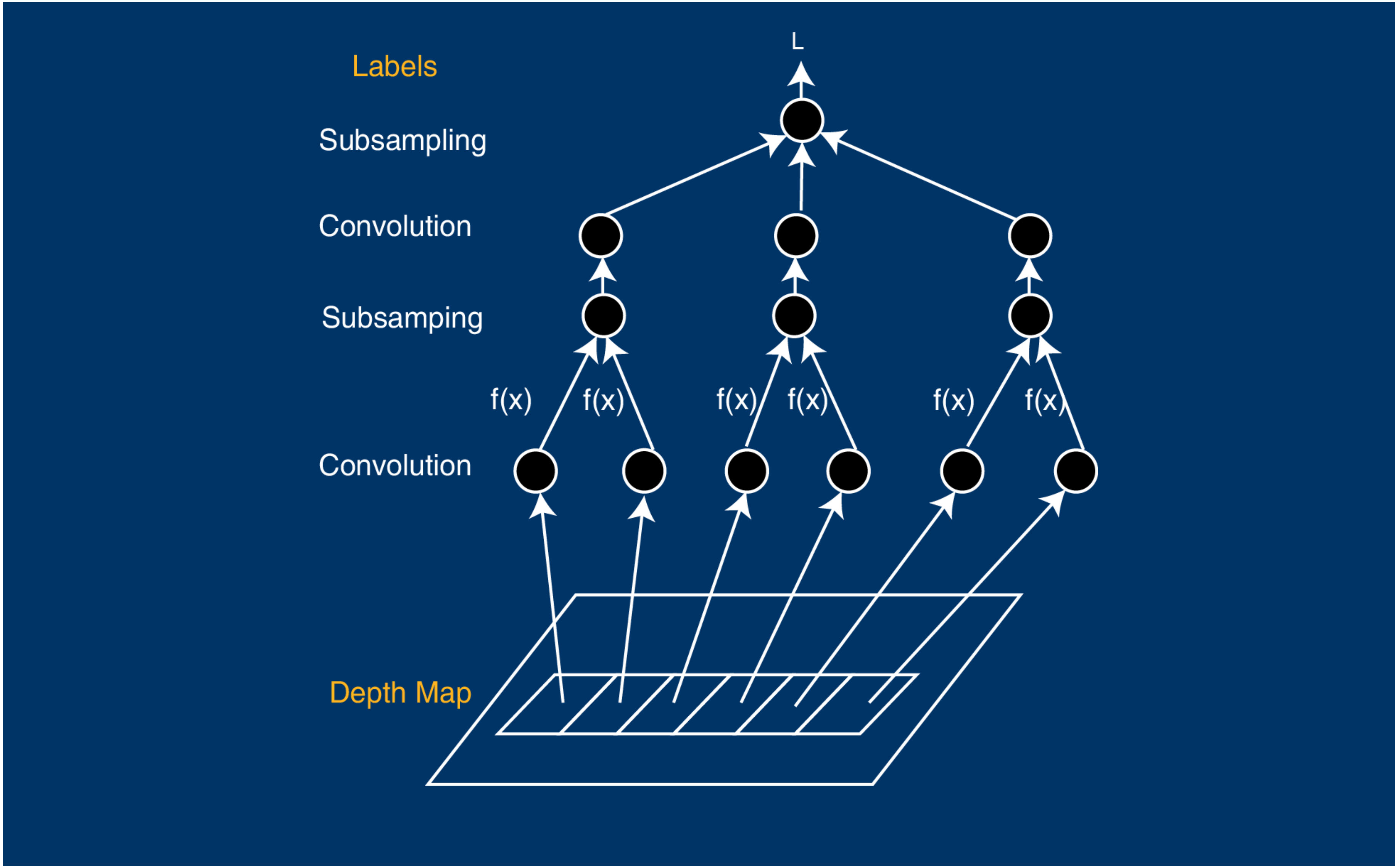
- Randomly sampling a large number of subvolumes at different positions with different sizes
  - 4D: depthmap sequence
  - 3D: single depthmap
- One occupancy value

  per subvolume



J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3D Action Recognition with Random Occupancy PatternsECCV2012

Relationship with Convolutional Neural Network



Problems of Convolutional Neural Network

- Too many parameters (weights at each layer, kernel size, etc.)
  - Difficult to train
- Empirical experiments showed
  - Kernel size (structure) more important than kernel coefficients

Weighted Sampling

- Down-sample the 4D volume of a depth sequence into resolution: $W_x * W_y * W_z * W_t$
- Total number of possible subvolumes is
$$\binom{W_x}{2} * \binom{W_y}{2} * \binom{W_z}{2} * \binom{W_T}{2}$$
- Sampling a subvolume with a probability that is proportional to the discriminativity of the subvolume.

Class Separability Score

- Given a pixel p, create a box centered at p
- For each video sequence in the training data, extract an 8-dimensional Haar feature vector from the box
- h_{ij}:feature vector from sequence j of class i.
- Within scatter matrix: $$\boldsymbol{S}_W = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(\boldsymbol{h}_{i,j} - \boldsymbol{m}_i)(\boldsymbol{h}_{i,j} - \boldsymbol{m}_i)^T$$
- Between class scatter: $$\boldsymbol{S}_B = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$
- Total scatter matrix: $$\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B$$

Class Separability Score

- The pixel's class separability score $J = \dfrac{\mathrm{tr}(\boldsymbol{S}_{\mathrm{W}})}{\mathrm{tr}\boldsymbol{S}_{\mathrm{B}}}$
- Given a subvolume, its separability score is the average separability score of all the pixels inside the subvolume
- The probability that a subvolume is sampled is proportional to its separability score

$$P_{R\ \mathrm{sampled}} \propto J_R = \frac{1}{N_R} \sum_{p \in R} J_p$$

Sampling Strategy

- Uniformly draw a subvolume
- Accept with probability

$$P_{R\ \mathrm{accept}} = \frac{W_x^2 W_y^2 W_z^2 W_t^2}{\sum_{p \in V} J_p} J_R$$

- Speed up computation:
  - 4-dimensional integral image

Feature Selection

- Elastic-Net regularization
  - Effective if feature dimension >> training data

Training data: $(x_i, t_i), i = 1, \ldots, n$

Extracting ROP feature vector: $x_i : \rightarrow h_i$

$$\min_{w} \sum_{i=1}^{n} (t_i - w \cdot h_i - b) + \lambda_1 \left| |w| \right|_1 + \lambda_2 \left| |w| \right|_2^2$$

- Discarding those h_i^j for which w^j is small

$h_i : \rightarrow y_i$ Dim(y_i)<<Dim(h_i)

$y_i^j = h_i^j * w^j$

Sparse Coding

- Handling occlusions: some boxes are occluded
- Using all the training data as the dictionary

$$A = (f_1, f_2, \ldots, f_n)$$

- Given a test data feature vector f

$$\min \frac{1}{2} \left|\left|f - A\alpha\right|\right|_2^2 + \lambda \left|\left|\alpha\right|\right|_1$$

- $\alpha(f)$ is the final feature vector to feed into a SVM classifier.

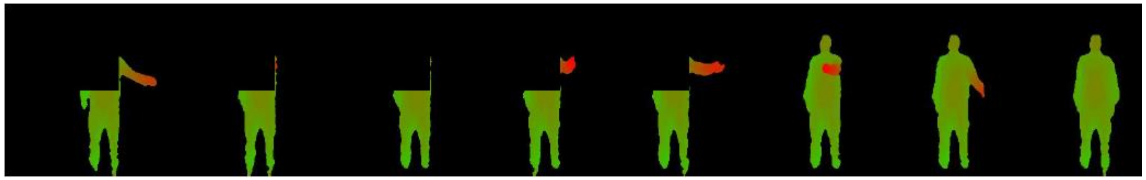
Experiments

- MSR Action3D
  - All sequences are resized to the same size 80x80x80x10

STIP	42.3%
Action Graph on Bag of 3D Points (Li et al’10)	74.7%
4D Convolutional Network (Ji et al’10)	72.5%
SVM on raw occupancy features	79%
Actionlet Emsemble	88.2%
HON4D	88.89%
ROP (no sparse coding)	85.92%
ROP(with sparse coding)	86.20%

Occlusion Handling

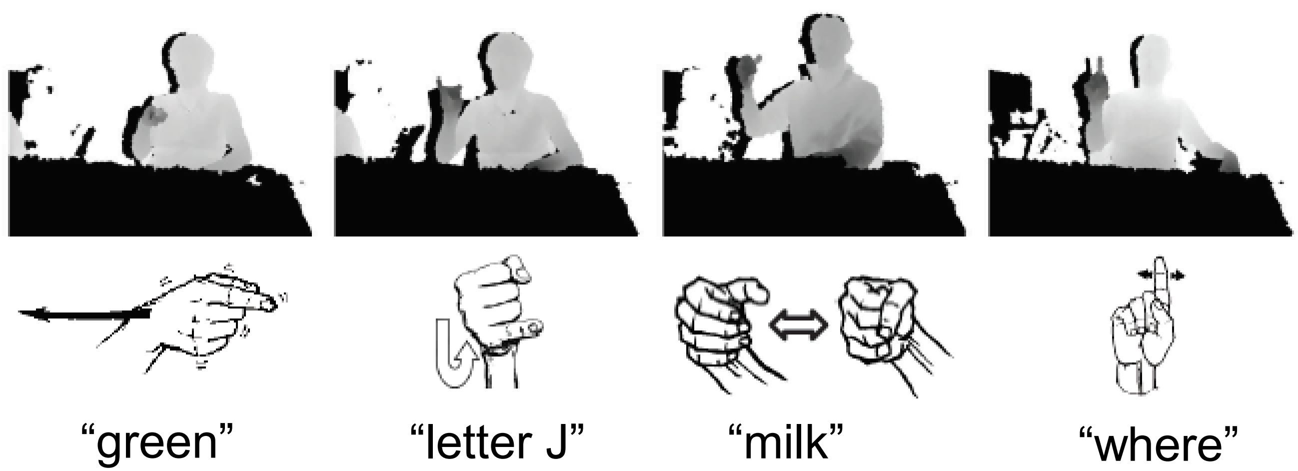
Simulated occlusions: a depth sequence partitioned into 2x2x1x2 subvolumes, removing one of the subvolumes



Occluded region	No sparse coding	With sparse coding
1	83.047	86.165
2	84.18	86.5
3	78.76	80.09
4	82.12	85.49
5	84.48	87.51
6	82.46	87.50
7	80.10	83.80
8	85.83	86.83

Hand Gesture

- MSR Gesture3D
  - 12 dynamic gestures
    - ASL
  - 10 subjects
  - Each subject performs each gesture 3 times



MSR Gesture3D

Method	Accuracy
Action graph + (2D) occupancy feature (Kurakin et al)	83.3%
4D Convolutional Network (Ji et al)	69%
HON4D (Oreifej&Liu 2013)	92.45%
ROP	86.8%
ROP + sparse coding	88.5%

Object Recognition

- RGB-D dataset (Ren et al)

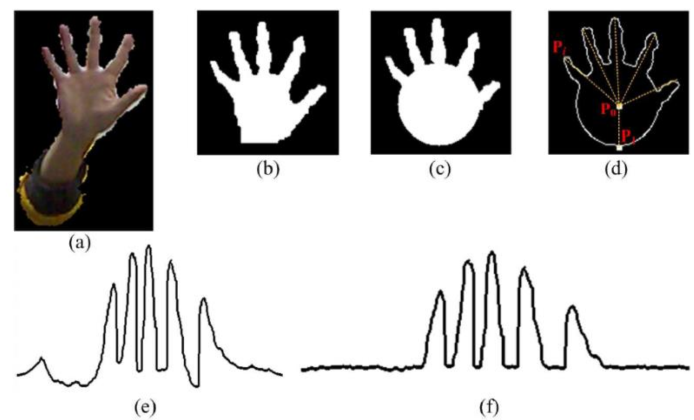
Method	Accuracy
3D SIFT (Lai et al)	66.8%
Hierarchical Kernel Descriptor on depth (Bo et al)	75.7%
ROP	80%
HONV (Tang et al)	91.25%
HOG on depth	85.00%

Recognition Paradigms

- Direct classification
  - Global feature descriptor: one vector per clip
  - SVM, RF, etc.
- Bag of Words framework
  - Interest points + local feature descriptor
- Actionlet Ensemble
  - J. Wang, Z. Liu, Y, Wu, J. Yuan, CVPR2012
- Random Occupancy Pattern
  - J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, ECCV2012
- Contour Matching (static hand gesture)
- Online recognition
  - Temporal segmentation
  - Action graph, Li et al, TCSVT 2008

Contour Matching

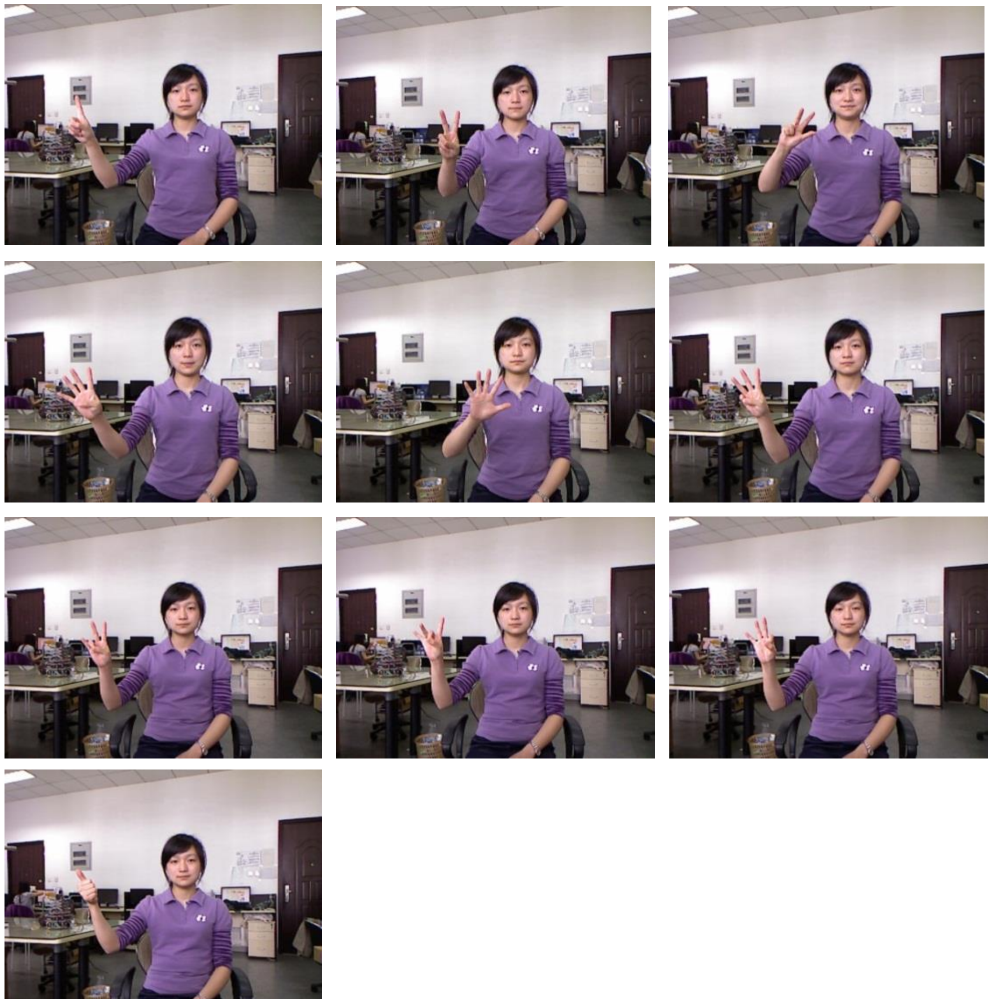
- Finger-Earth mover's distance (**FEMD**)
  - Ren et al, ACMMM2011
- Image-to-class dynamic time warping (**I2C-DTW**)
  - Dai et al, ICME2013



NTU 10-Gesture Dataset

- Digits 0-9

KINECT-ASL (UESTC)



Hands up! - Hand Gesture Based Human-Computer-Interaction

Zhou Ren, Jingjing Meng and Junsong Yuan
School of EEE, Nanyang Technological University

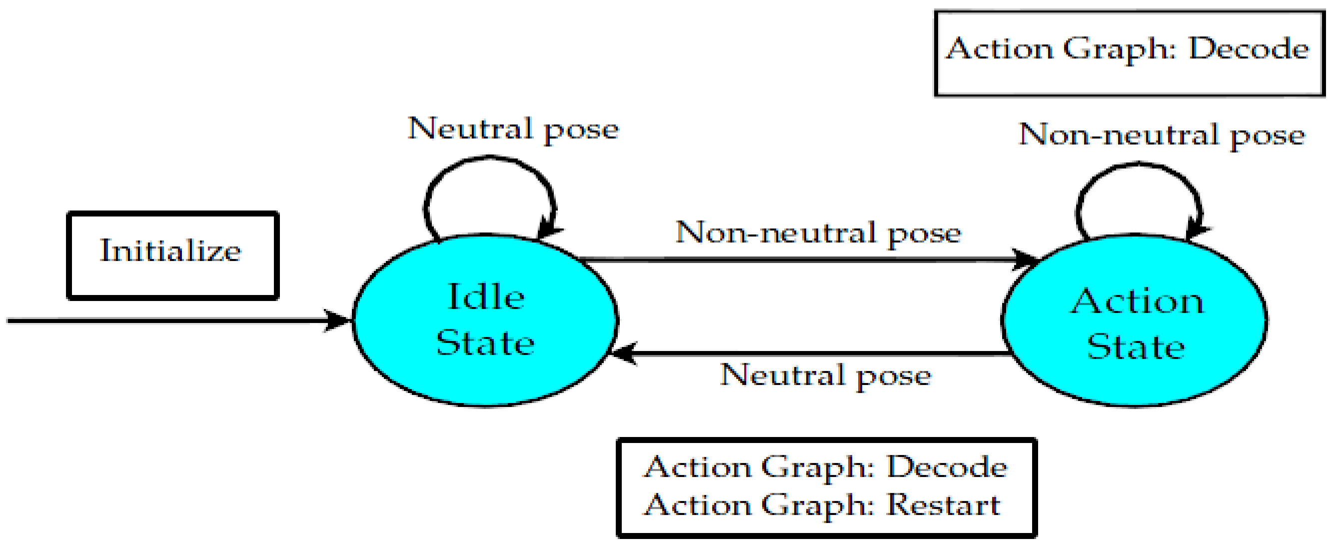
Innovative Technology Showcase 2011, Singapore

Recognition Paradigms

- Direct classification
  - Global feature descriptor: one vector per clip
  - SVM, RF, etc.
- Bag of Words framework
  - Interest points + local feature descriptor
- Actionlet Ensemble
  - J. Wang, Z. Liu, Y, Wu, J. Yuan, CVPR2012
- Random Occupancy Pattern
  - J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, ECCV2012
- Contour Matching (static hand gesture)
- Online recognition
  - Temporal segmentation
  - Action graph, Li et al, TCSVT 2008

Online (Real-time) Action Recognition

- Temporal segmentation
  - Short-time feature vector (e.g. every 5 frames)
  - Idle pose classifier



Back-end Classifier

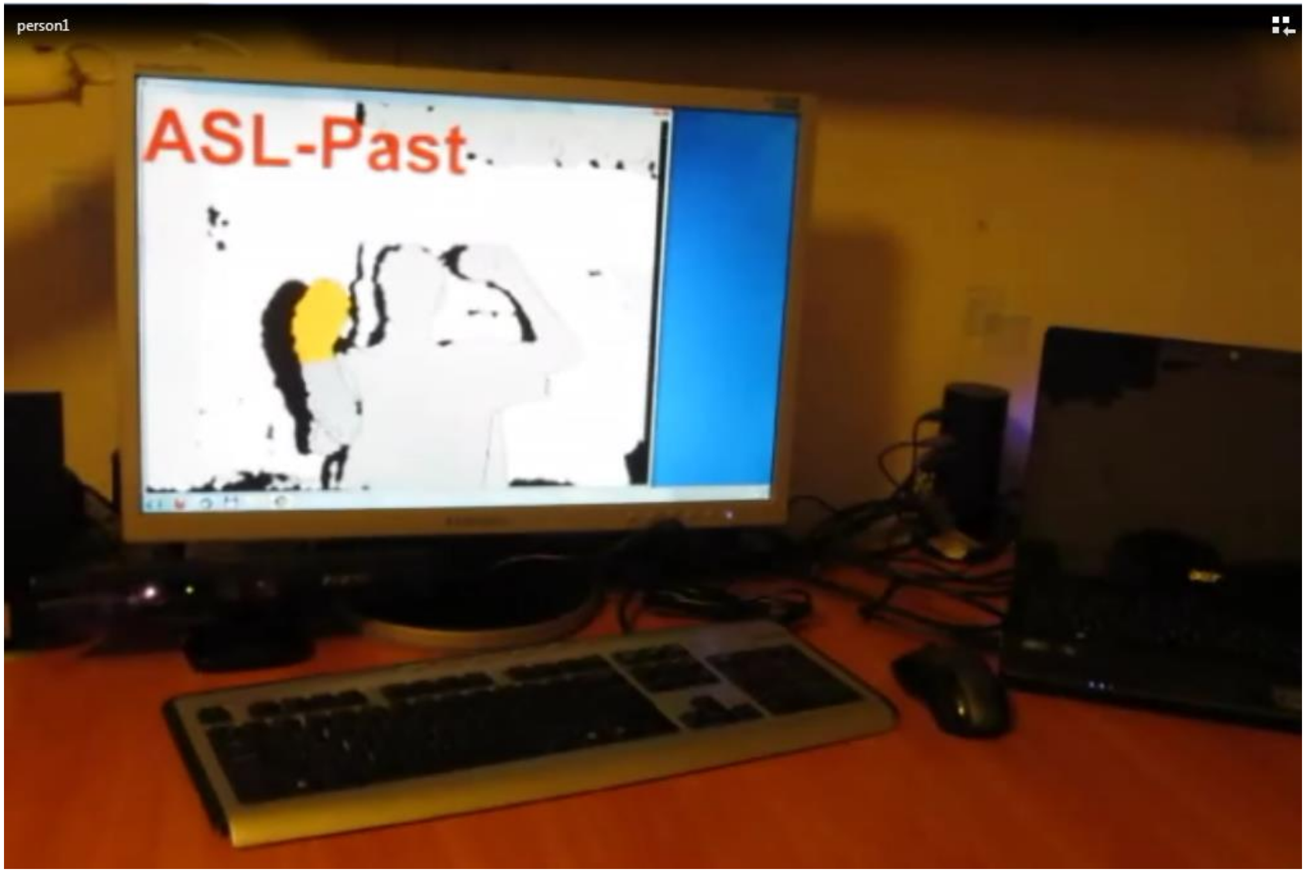
- Batch-mode classifier applied to the accumulated frames between last idle state and current idle state
- Action graph (Li et al, TCSVT2008)
  - Better handling temporal alignment
  - Outputs recognition results without having to wait until the action is finished

Video

- Daily activity recognition

Video

- Hand Gesture Recognition



A. Kurakin, Z. Zhang, Z. Liu, EUSIPCO 2012

Summary

- Action/gesture recognition from 3D sensors
  - Lots of new problems to work on
  - Exciting application scenarios
  - Robotics, HCI, Medical, VR/AR, etc
- Many new features
  - From skeleton: Fourier Pyramid
  - From depth data: HON4D
- Actionlet ensemble
  - Combining skeleton + local shape features
  - Discriminative actionlet mining

Summary

- Random occupancy patterns
  - Not relying on skeletons
  - Useful for action, hand gesture, and object recognition
- Hand gesture recognition
  - Hand segmentation and feature extraction
  - Hand skeletonization
- Datasets and codes

Future Directions

- Bag of feature scheme
  – Better interest point detection from depth maps
- Handling realistic occlusions
  – Don't know whether there is an occlusion and where
- Continuous activity recognition
    - Without clear separation boundaries over time
- Human-object interactions
  – Many interesting problems.
  – Combining object recognition with activity recognition
  – Stochastic grammar for complex activities

Future Directions

- Hand gesture recognition
  - Exciting applications in user interface
- Attention and intention recognition
  - Understanding user's interests
  - Javier et al: Measuring the Engagement Level of TV Viewers, FG2013

Thanks!

Contacts:

Zicheng Liu
zliu@microsoft.com
http://research.microsoft.com/~zliu

Junsong Yuan
jsyuan@ntu.edu.sg