Feature Improvement and Analysis with Dense Trajectories for Multi-purpose Activity Recognition

Hirokatsu KATAOKA

The Univ. of Tokyo, Postdoctoral Fellow
Keio Univ., Visiting Researcher
Technische Universität München (TUM), Visiting Scientist

http://www.hirokatsukataoka.net/

Research Introduction in 2008 – 2014

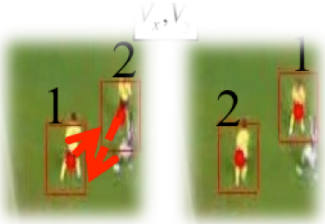
Computer vision theory and its application, mainly interested in human sensing

SIT '08 – Keio Univ. '09 - '14 – UC Riverside '11 - '12 – AIST '12 - '13 – TUM '13 - Univ. of Tokyo '14

Human Activity Recognition

Football Video Analysis

- Multi-player tracking in crowded situations
- Real-time (30fps) tracking and analysis
- Papers：ACCVW'09, IEEJournal'10, DIA'10Award, SICE'10Award, IIEEJornal'12

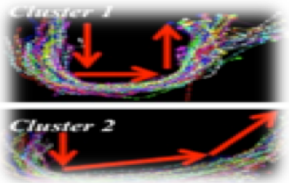


Pedestrian Detection System

- Improved feature description for complicated scenes
- Real-time tracking
- Papers：IEEJournal'12Award, IEEE IECON'12 Oral Award, Int. J. of Vehicle Safety '12, IEICE Trans. '13

Clustering for Trajectory Analysis

- Trajectory clustering
- Indoor scene analysis
- Papers：ViEW'12, IAPR MVA'13



Improved Feature for Fine-grained Recognition

- Dense sampling and detailed feature description in activity modeling
- Feature vector representation with subset mining
- Papers：MIRU'12, SSII'13, ACCV'14

Activity Prediction through Recognition & Data Analysis

- Pattern analysis in activity history DB
- Activity prediction by means of Bayesian framework
- Papers：MIRU'12, SSII'13, Int. conf. preparing

Collaboration

Co-research with company, university and institute

【Institutes】
- AIST
- NTSEL

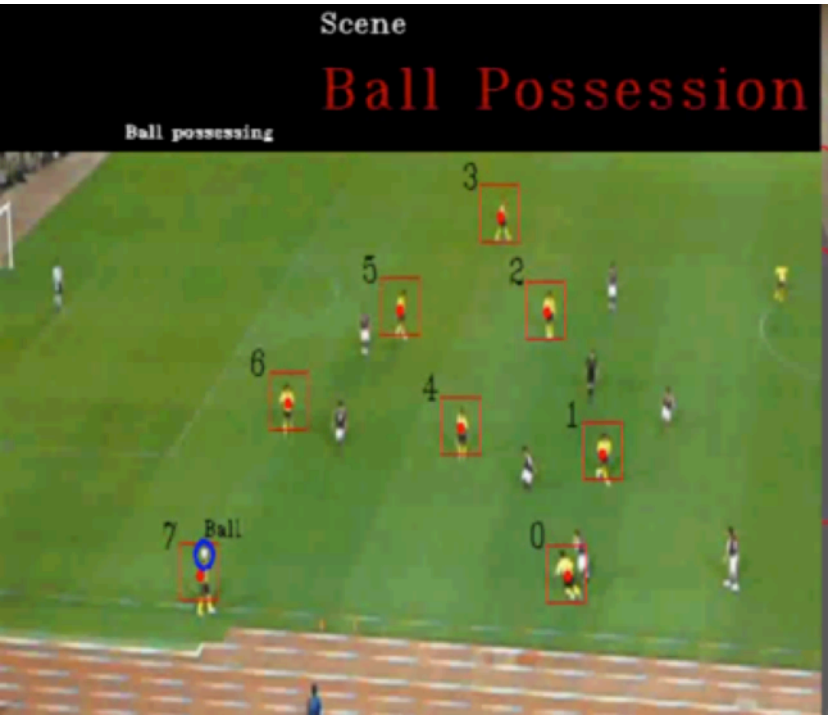
【Company】
- Panasonic
- Denso IT Lab.
- OKI
- TOYOTA
- Toshiba

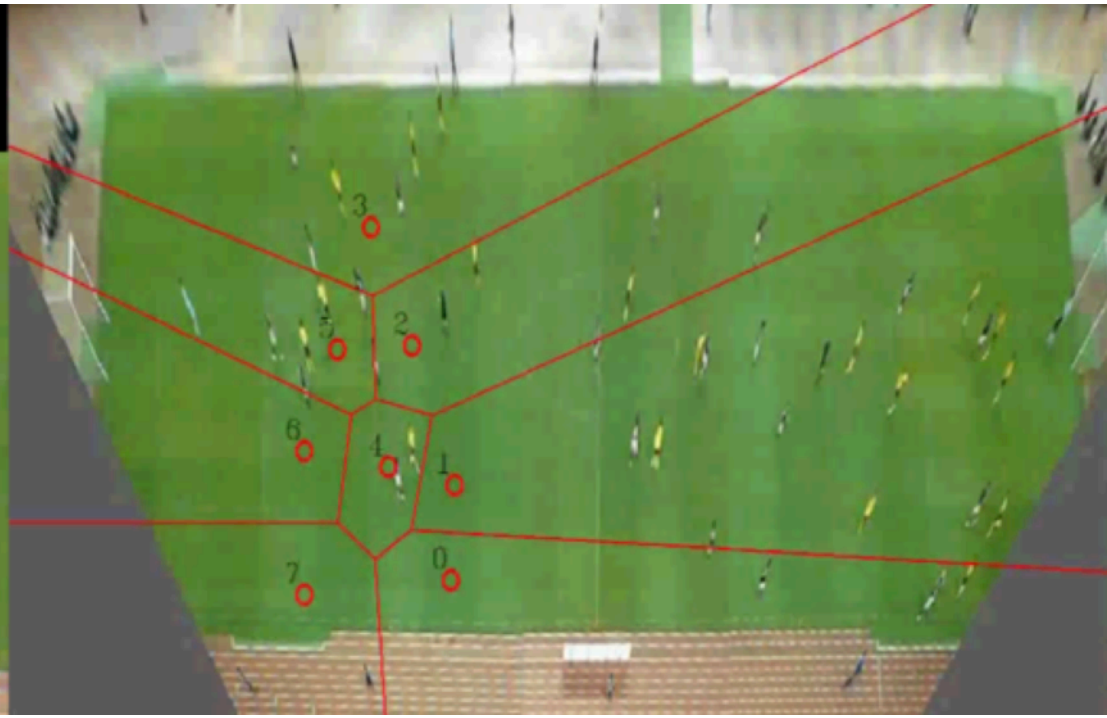
【University】
- TUM
- UC Riverside
- Hokkaido Univ.
- Tokyo Denki Univ.

※ SIT: Shibaura Institute of Technology, AIST: National Institute of Advanced Industrial Science and Technology

Football Video Analysis ‘08 – ‘11

- Football video analysis for strategy analysis
  - Multi-player positioning with color-based particle filtering
  - Detection with Adaboost classifier in crowded area and resampling
  - Resampling is considering player’s velocity





Pedestrian Detection ‘09 – ‘11

- Pedestrian detection & tracking with improved feature
  - Co-occurrence feature description
  - Classifier in tracking-by-detection framework

[ACCV'14] Extended Co-occurrence HOG with Dense Trajectories for Fine-grained Activity Recognition

H. Kataoka, K. Hashimoto, K. Iwata, Y. Satoh,
N. Navab, S. Ilic, Y. Aoki

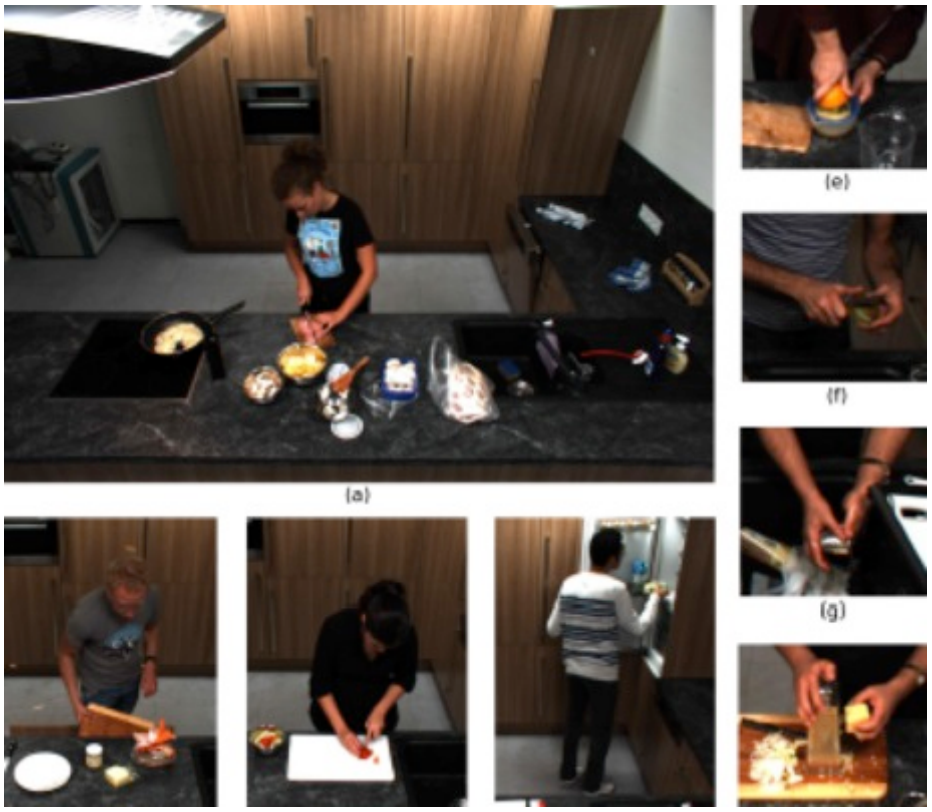
Now available here!
http://www.hirokatsukataoka.net/

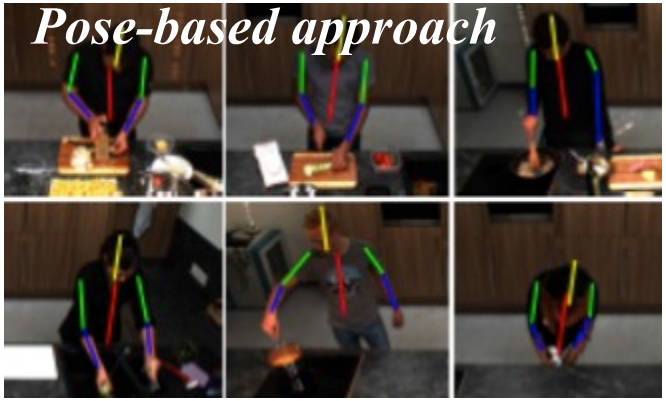
Fine-grained Activity Recognition

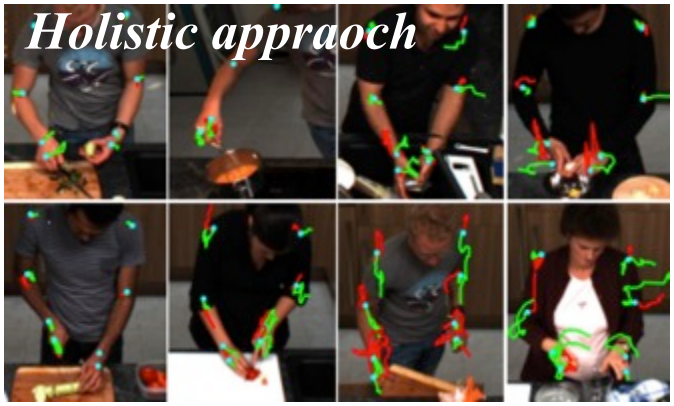
Fine-grained recognition

https://sites.google.com/site/cvprfgvctwo/home

- The visual distinctions between similar categories are often quite subtle
- It's difficult to address with general-purpose object recognition machinery



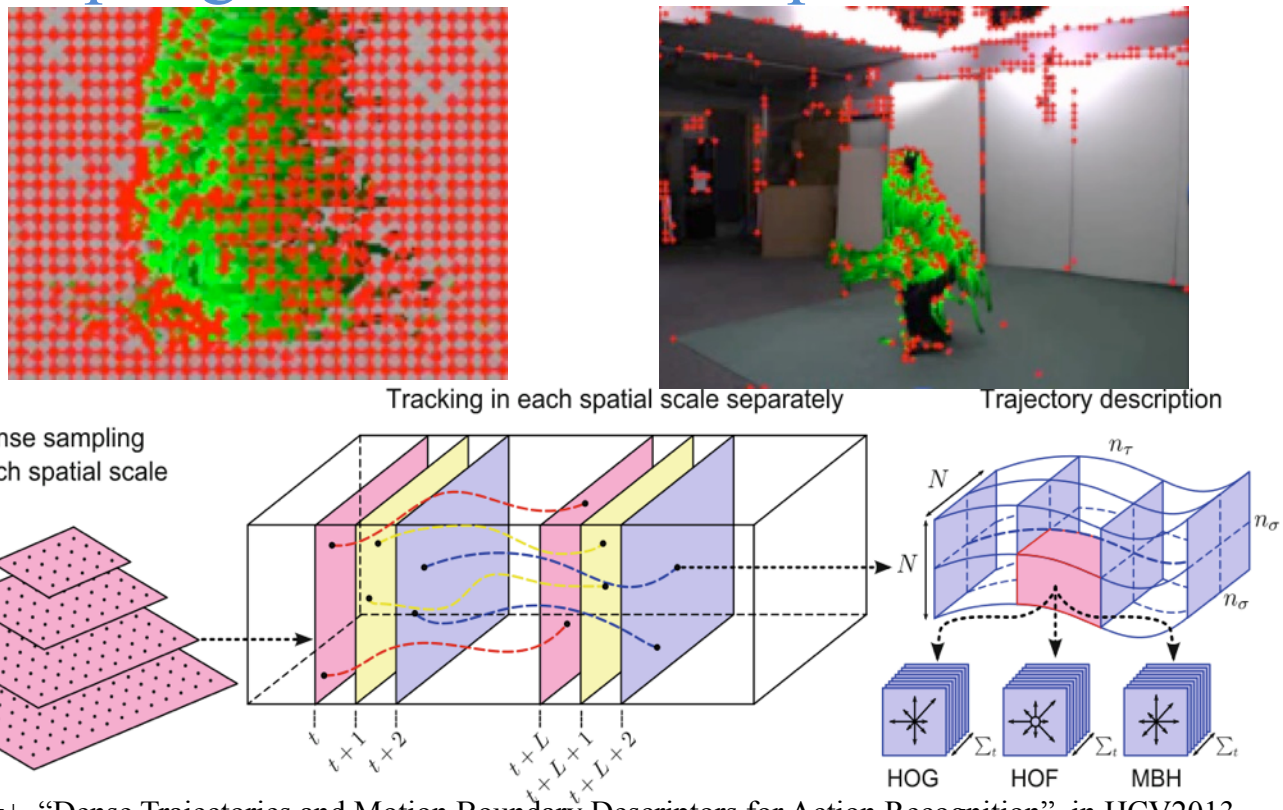




M. Rohrbach+, "A Database for Fine Grained Activity Detection of Cooking Activities", in CVPR2012.

Dense Trajectories

- Dense sampling and feature description on tracks



H. Wang+, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", in IJCV2013.

- Pyramidal image division, and dense tracks with optical flow
- Feature description (HOG, HOF, MBHx, MBHy)
- Visual words representation

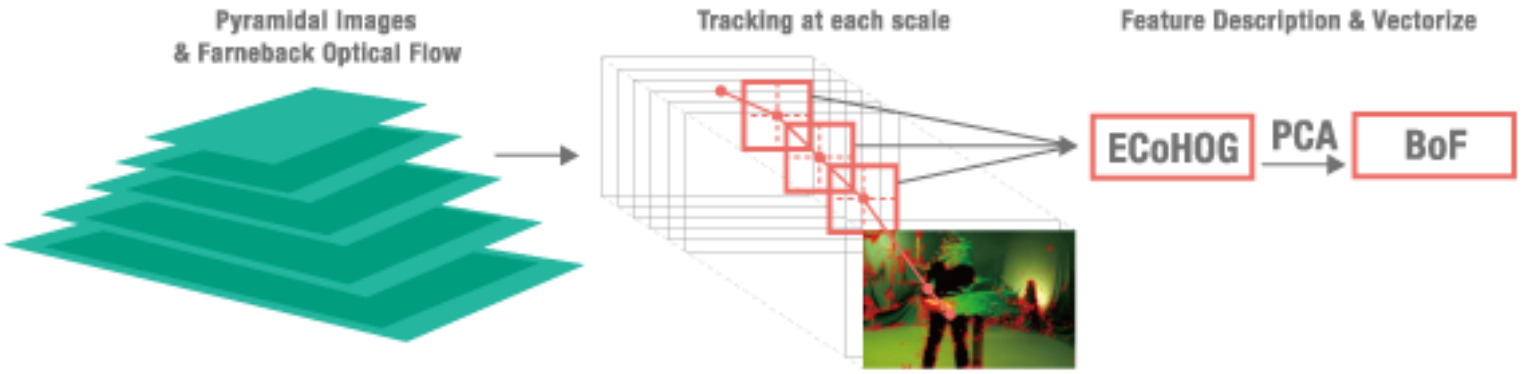
Motivation

Improved Co-occurrence feature in Dense Trajectories

- Contribution
  - Detailed feature description with co-occurrence
  - Effective feature descriptor in fine-grained recognition
- Approach
  - Co-occurrence HOG with edge-magnitude

  (to enhance edge-boundary)
  - Dimension reduction with PCA
  - MPII cooking activities dataset & TUM surgery dataset

Dense Trajectories (DT)
On KTH & IXMAS dataset



Co-occurrence feature：CoHOG&ECoHOG

“Edge-pair counting” & “Edge-magnitude accumulation(proposal)”

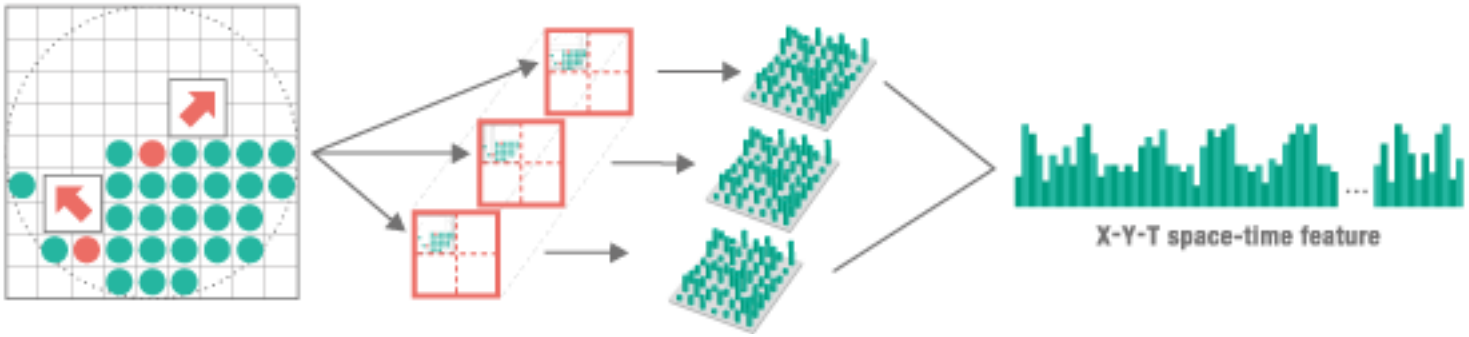
CoHOG: edge-pair counting to corresponding histogram position

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} 1, & \text{if } d(p,q) = i \text{ and } d(x+p, y+q) = j \\ 0 & \text{otherwise} \end{cases}$$

Extended CoHOG(ECoHOG): edge-magnitude accumulation

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} \|g_1(p,q)\| + \|g_2(p+x, q+y)\| \\ \text{if } d(p,q) = i \text{ and } d(p+x, q+y) = j \\ 0 & \text{otherwise} \end{cases}$$

– PCA dim. reduction: 10^3 - 10^4 dims into 10^1-10^2 ,easy to divide in feature space



Parameter decision

"# of PCA dims" & "size of edge extraction window"

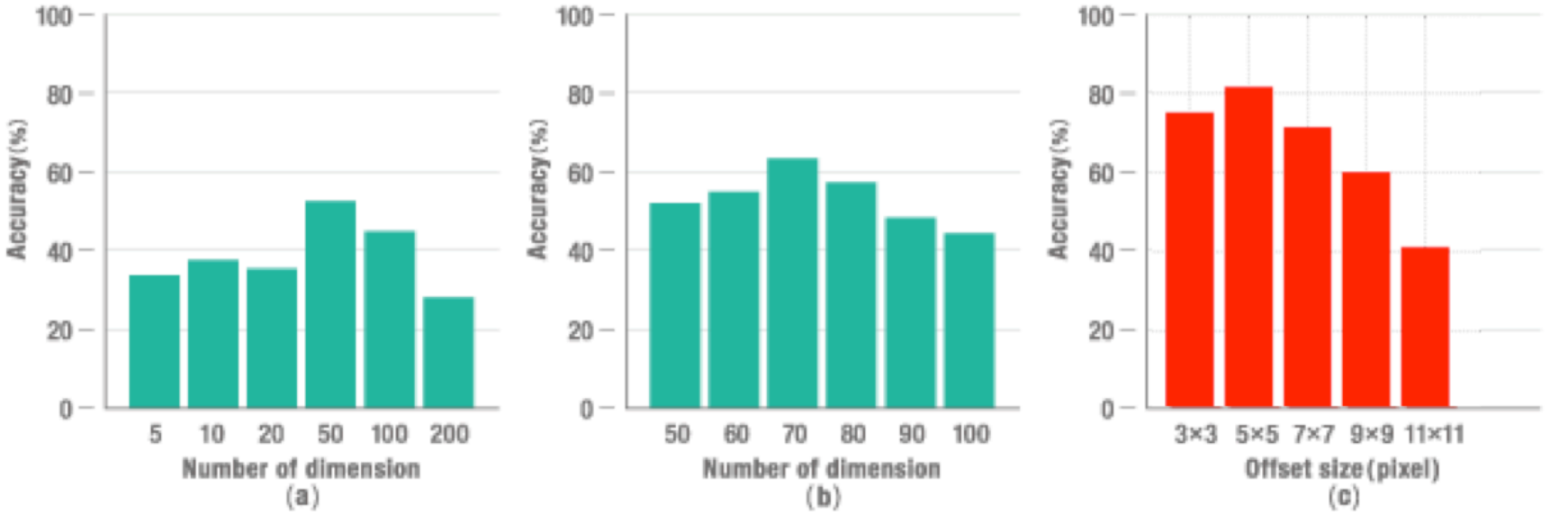
(a) PCA [dimensions] – 5, 10, 20, 50, 100, 200

(b) PCA [dimensions] – 50, 60, 70, 80, 90, 100

(c) Size of edge extraction window [pixels] – 3x3, 5x5, 7x7, 9x9, 11x11

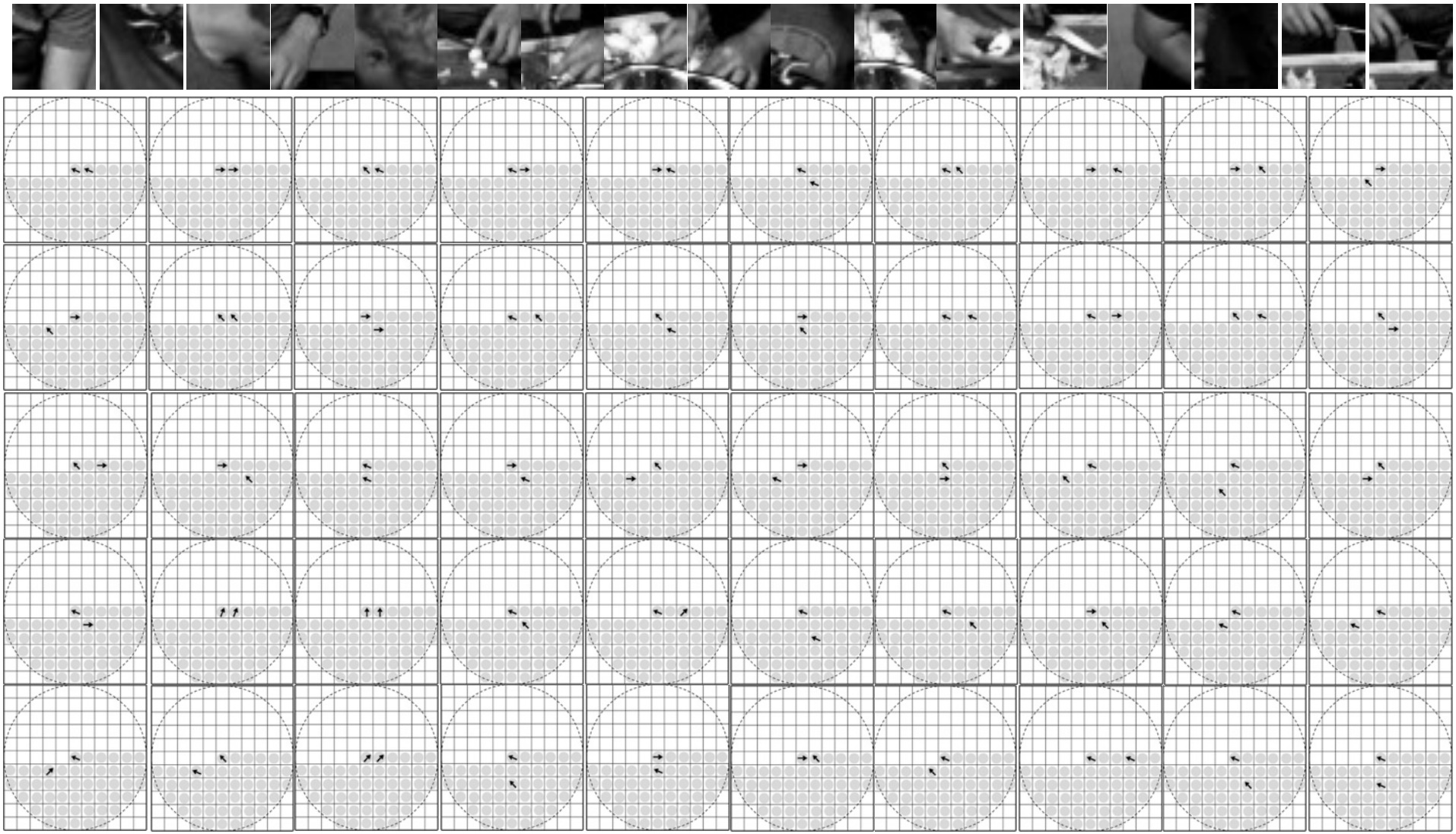
– PCA dim: 70 dimensions => consider "contribution ratio" & "feature size"

– Window size: 5x5 pixels => pixel similarity



Top 50 frequently used features

- Neighbor pixels are used in classification



Experiment on TUM & MPII dataset

INRIA surgery dataset

Approach	Accuracy (%)
HOG + Tracking	40.16
Original DT (HOG, HOF, MBH, Trajectory)	93.58
CoHOG in DT	81.05
ECoHOG in DT	96.36
All model **(ECoHOG, HOF, MBH, Trajectory)**	**97.31**

MPII cooking activities dataset

Approach	Accuracy (%)
HOG + Tracking	18.2
Original DT (HOG, HOF, MBH, Trajectory)	44.2
CoHOG in DT	46.2
ECoHOG in DT	46.6
All model **(ECoHOG, HOF, MBH, Trajectory)**	**49.1**

Consideration

- Improved co-occurrence feature into dense traj.
  - Dense sampling and detailed description in human activity
  - To grab small feature changing

- Parameter optimization
  - PCA dim: 70 dimensions => consider “contribution ratio” & “feature size”
  - Window size: 5x5 pixels => pixel similarity

Topic Modeling Based Feature Mining for Activity Recognition

H. Kataoka, M. Hayashi K. Iwata,
Y. Satoh, N. Navab, S. Ilic, Y. Aoki

Feature Mining Research [in 2013]

Association analysis in dense traj.[1] (DT) based BoF

- Bag-of-features (BoF) almost has zero values
- Association rules make BoF effective vector

$$support = \frac{(X \cup Y).count}{n} \qquad confidence = \frac{(X \cup Y).count}{X.count}$$

Bin No. →

	VW1	VW2	…			VW6			VW9	VW10
vec1	0.00	0.00	0.00	0.00	0.00	0.41	0.00	0.00	0.10	0.00
vec2	0.00	0.00	0.00	0.00	0.00	0.41	0.00	0.00	0.20	0.00
vec3	0.00	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.10	0.00
vec4	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.30	0.00
vec5	0.00	0.00	0.00	0.10	0.00	0.80	0.00	0.00	0.10	0.00

We compress 2 dimensions out of 10 in this situation, and accuracy is better

Non-zero values are stationary

→ Effective vector on the side of “accuracy” and “memory”

Topic Modeling based Feature Mining in AR

We propose topic modeling based feature mining

- Approach
  - Subset vector extraction with latent Dirichlet allocation (LDA)
  - Input vectors are captured by means of dense traj.

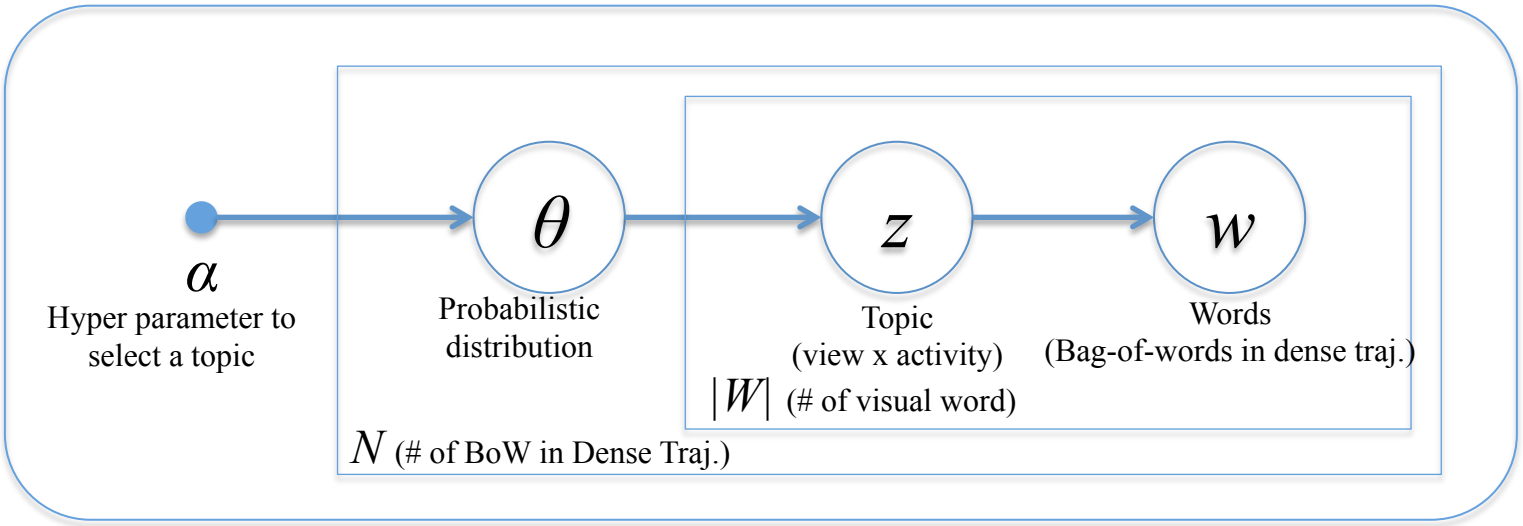
- Contributions
  - “Recognition accuracy”, “processing speed” and “Memory”
  - Common framework in single/multi-view

- Experiments
  - INRIA surgery(multi-view & fine-grained)
  - Common framework in single/multi-view

Topic Modeling based Feature Mining

Analyzing BoW vectors with topic model such as LDA[3]

- Inducing latent topics by sampling Dirichlet distribution
- A topic includes frequent words in articles (bottom-left figure)
- Unsupervised learning for frequency mining



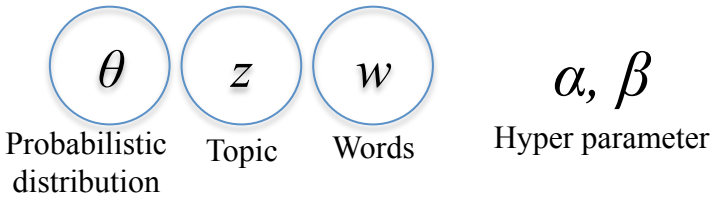
[3] D. Blei, et al., "Latent Dirichlet Allocation", Journal of Machine Learning Research, 2003.

Here, we give # of topic and words
- # of topic: view x activity or activity at each view

Analyzing BoW from Dense Traj. to create effective vectors

Latent Dirichlet Allocation (LDA) Settings

Sampling, and number of topic



- Probabilistic distribution

$$p(\theta, z_n|w, \alpha, \beta) = \frac{p(\theta, z_n, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

- Replace simple distribution, and Gibbs sampler allows us to estimate

$$p(\theta, z_n|\gamma, \phi) = p(\theta|\gamma) \prod_{n=1}^{N} p(z_n|\phi_n)$$

- In number of topic is lead as

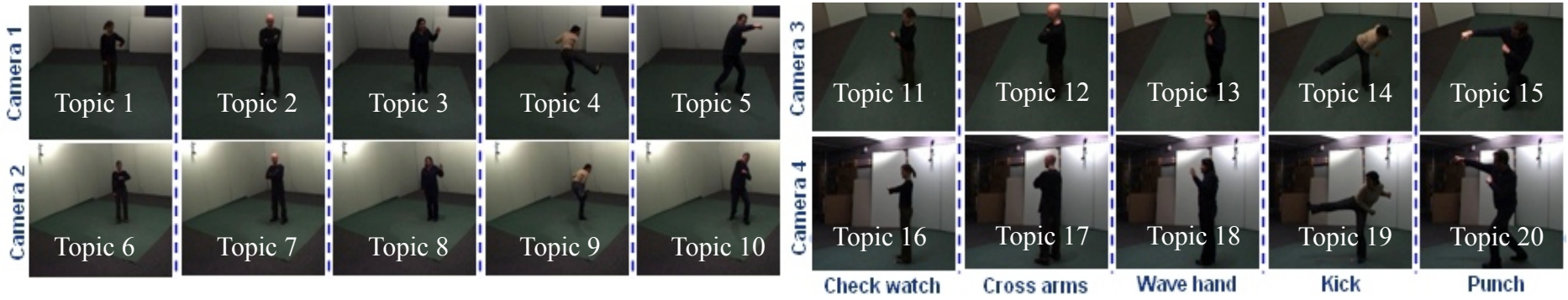
$$N = N_{action} \times N_{view} \quad \text{or} \quad N = N_{view}$$

T. Griffiths+, “Finding Scientific Topics”, in Proc. of the National Academy of Sciences, 101, pp.5228-5235, 2004.

Feature Mining

The feature mining framework with LDA

Example: dataset includes “4 views” and “5 actions”



Input vector: 10 dimensions (VW 1 – 10)

Feature analysis with topic modeling

Topic 1: VW3 VW7 VW5 | Topic 2: VW3 VW7 VW4 | . . . | Topic 20: VW1 VW7 VW4

=> Visual Word(VW): 1, 3, 4, 5, 7

OR operation

Output vector: 5 dimensions as subset of input vector

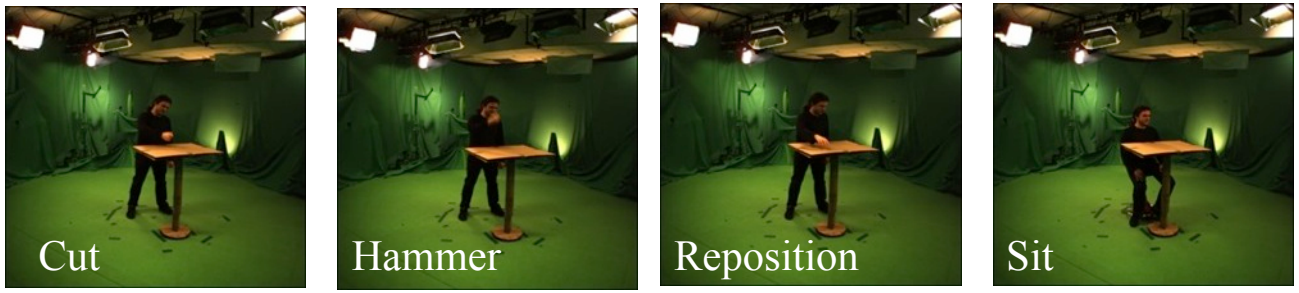
VW 1, 3, 4, 5, 7

INRIA surgery dataset

Original Dense Traj. vs LDA Subset

- Dense traj.: 4,000 dims, LDA subset: less than 4,000 dims
- 2-fold-cross-validation

INRIA surgery dataset: 4classes, 4 views and 8 persons



Descriptor	DT %	LDA allview %	LDA eachview %
HOG	73.52±0.44	70.41±1.62	74.40±0.00
HOF	75.88±1.17	75.73±0.88	76.19±0.44
MBHx	76.25±2.58	75.96±0.36	76.71±1.86
MBHy	65.97±0.74	68.78±0.29	71.27±0.74
Feature Integration	**75.51±0.36**	**75.07±0.07**	**80.43±1.71**

Results (Original Vector)

Dense trajectories (total: **75.51±0.36%, 16,000 dims**)

	View00	View01	View02	View03	ViewIntegration
HOG	68.12±0.37	73.00±1.26	71.29±1.92	69.44±0.07	73.52±0.44
HOF	73.44±3.62	72.55±1.55	70.7±5.02	71.37±5.25	75.88±1.17
MBHX	69.67±3.11	71.51±0.66	71.22±1.85	68.56±4.36	76.25±2.58
MBHY	61.83±0.59	66.27±1.48	61.38±2.08	63.6±2.07	65.97±0.74
FeatureIntegration	71.67±1.10	71.96±0.36	73.52±0.29	71.37±1.54	**75.51±0.36**

Results (Subset in All-view and Each-view)

Feature Mining in All-view (total: **75.07±0.07%, 12,885 dims**)

	View00	View01	View02	View03	ViewIntegration
HOG (3015dim)	68.26±2.44	72.18±2.81	71.22±3.47	69.74±1.25	70.41±1.62
HOF (3430dim)	70.93±0.96	74.11±0.00	70.26±6.06	73.59±0.67	75.73±0.88
MBHx(3114dim)	72.11±3.62	71.67±0.96	76.25±1.7	64.78±1.92	75.96±0.36
MBHy(3326dim)	65.82±3.99	69.74±0.07	66.05±0.37	62.79±1.55	68.78±0.29
Feature Integration	71.81±2.14	73.74±0.96	73.22±0.44	69.30±2.58	**75.07±0.07**

Feature Mining at Each-view (total: **80.43±1.71%, v.0: 6371dims v.1: 8157dims v.2: 4936dims v.3: 9447dims**)

	View00	View01	View02	View03	ViewIntegration
HOG	70.33±0.07	73.74±3.92	69.96±1.62	74.92±1.26	74.40±0.00
HOF	72.85±1.85	71.89±0.89	72.40±2.14	73.14±2.14	76.19±0.44
MBHX	68.71±0.96	73.37±2.81	76.54±0.07	71.89±2.07	76.71±1.86
MBHY	65.82±1.18	65.08±0.44	70.26±0.15	72.18±2.36	71.27±0.74
Feature Integration	73.06±1.93	75.81±0.37	74.33±3.05	77.52±0.74	**80.43±1.71**

Consideration

Subset vector at each view is better

- Dealing with small space
- Easy to divide in feature space
- Noise canceling (e.g. wrong tracks in bg and motion boundary)

Future work

- Experiments on IXMAS and MPII data
  - IXMAS: multi-view domain
  - MPII: fine-grained activity recognition
- Feature analysis
  - Single/multi-view, fine-grained recognition, # of dimension, each descriptor (HOG, HOF, MBHx, MBHy), dimension reduction, processing speed

Activity Prediction in a Bayesian Framework

H. Kataoka, K. Iwata, Y. Satoh, N. Navab, S. Ilic, Y. Aoki

At the beginning…

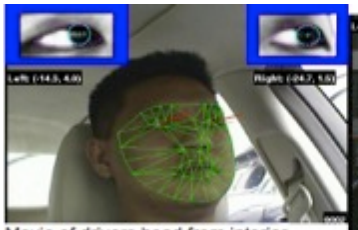
Human motion analysis

- Detection, tracking and trajectory analysis
- Face recognition and gaze estimation
- Posture estimation, activity recognitiion

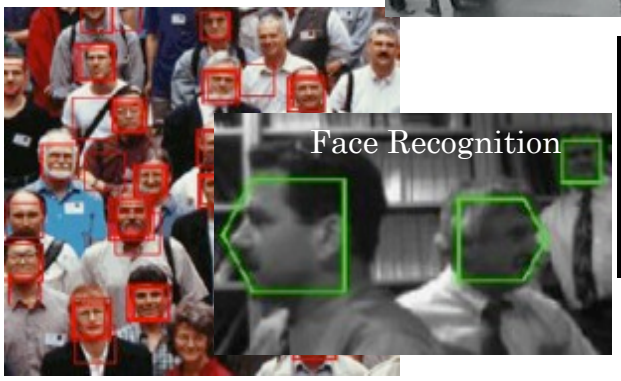
Look at people

Trajectory extraction

Attention



Mental state



Body Situation

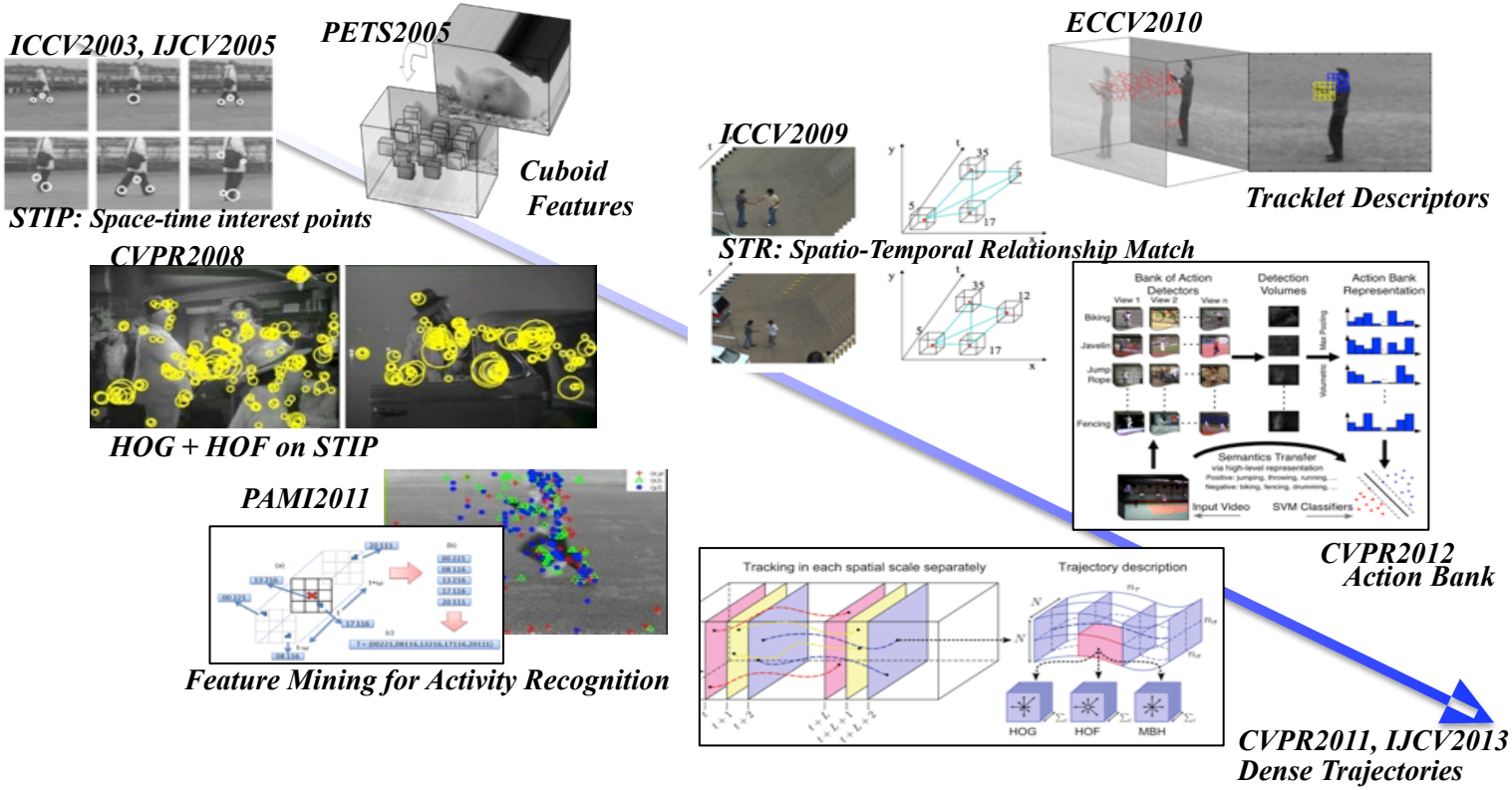
Activity Analysis

Activity Recognition

Understanding what people are (a person is) doing

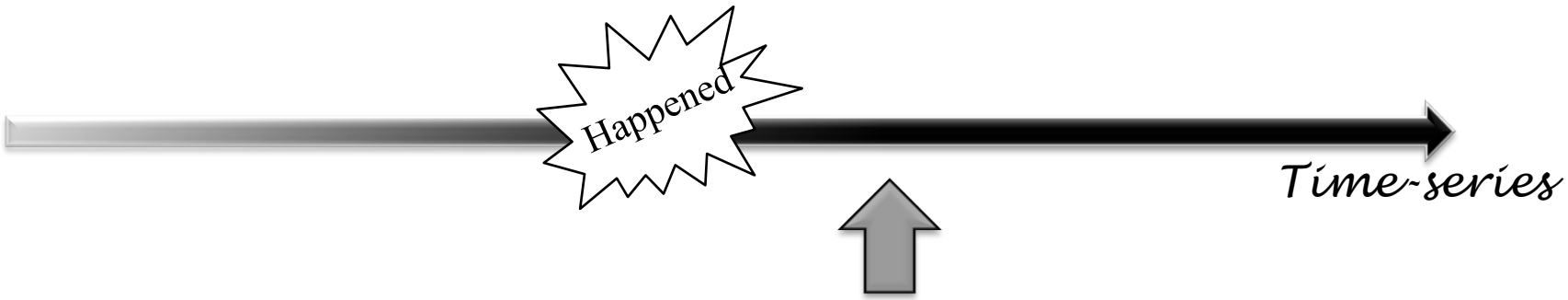
- Local-feature based recognition
- Posture based recognition
- Trajectory based recognition

…



Conventional Activity Analysis

Post-event Analysis : conventional works

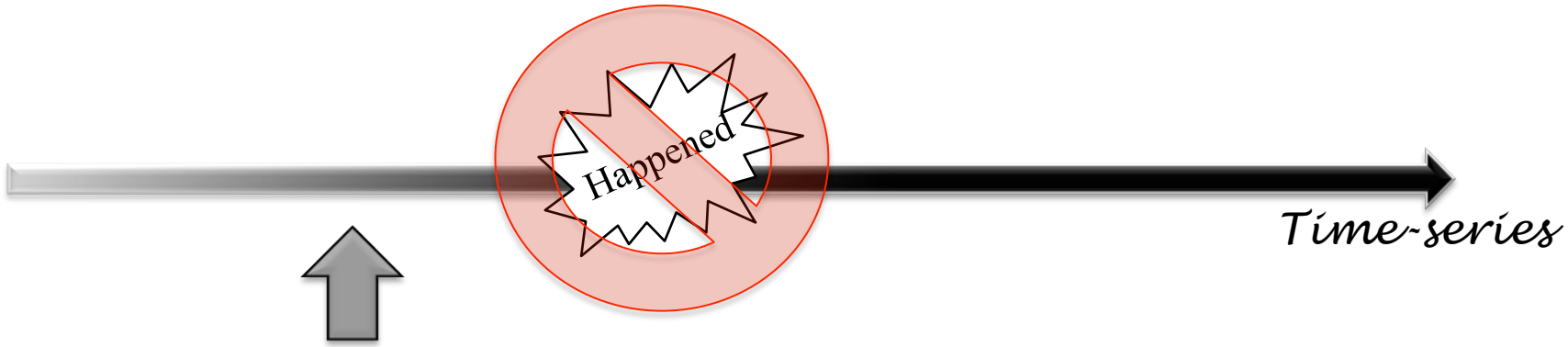


Event detection
(Action tag : A_i)

It's too late!!

Activity Prediction

Pre-event Analysis : new concept



Event prediction
(Prediction tag : P_i)

Activity prediction and it's prevented

Two Types of Activity Prediction

- Early activity recognition
  - M.S.Ryoo, "Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos", in ICCV2011.

  => Bag-of-integral-histogram for temporal tag analysis
  - G.Yu et al., "Predicting Human Activities using Spatio-Temporal Structure of Interest Points", in ACM MM2012.

  => Space-time implicit shape model & random forests
- Activity prediction
  - K. Li+, "Prediction of Human Activity by Discovering Temporal Sequence Patterns", in TPAMI2014.

  => Tree-based spatio-temporal matching

Our Proposal

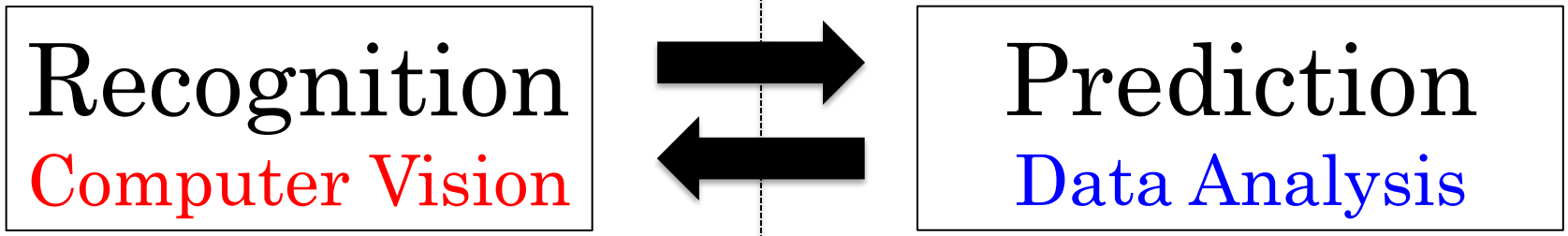
- Prediction through recognition and data analysis

Activity Recognition

- Dense traj. & postural feature
- Adaptive selection in Bayes

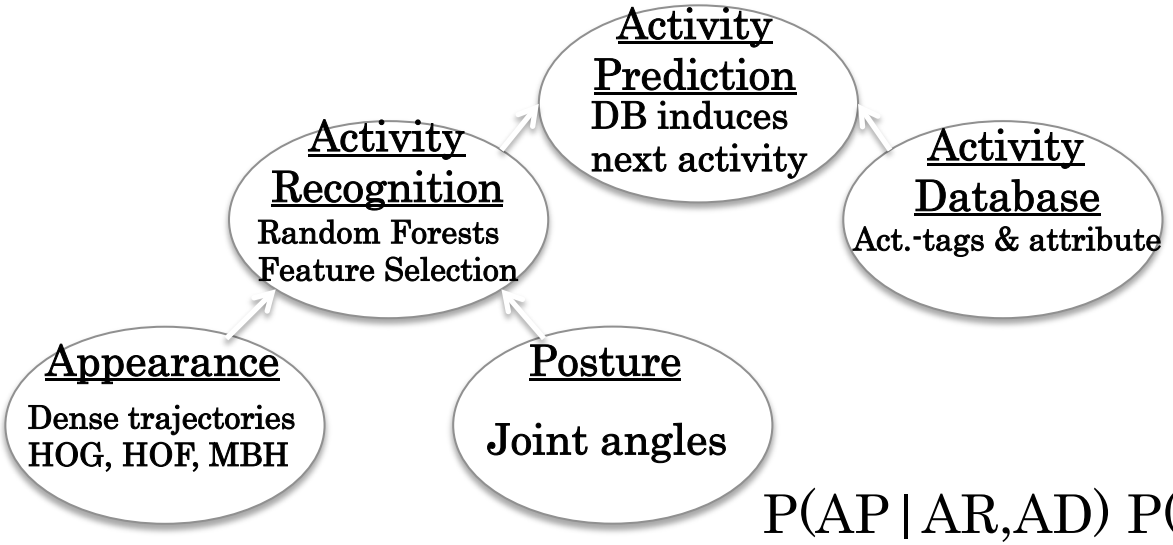
Activity Prediction

- Activity accumulation
- Effective info extraction by Mining



Framework for Prediction

Activity Prediction



$$P(AP \mid AR,AD)\ P(AR \mid A,P)\ P(AD)\ P(A)\ P(P)$$

Recognition

- Dense trajectories & joint angles with random forests integration

Prediction (data analysis)

- DB includes prev., curr., and next activity
- Naïve Bayes approach

Example of Activity Sequence

- Action sequence
  Walk -> Bend -> Sit -> Have a Meal
- Time zone
  Night

Activities in Daily Life

Activities from ICF [WHO, 2001]

d166 reading

d4103 sitting

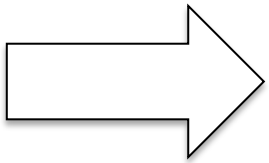
d4104 standing

d4105 bending

d4452 reaching

d450 walking

d550/d560 meal (eating & drinking)



Important!

d166 reading

d4452 reaching

d550/d560 meal

: Objective Activities

Activity History Database

Input activities and attributes

1	morning	walk	sit	book
2	morning	walk	stand	book
3	day	walk	sit	pc
4	night	bend	sit	meal
5	night	walk	sit	pc
6	night	walk	sit	pc
7	night	bend	sit	meal
8	night	walk	stand	book
9	night	walk	sit	pc
10	morning	walk	sit	book
11	morning	bend	sit	meal
12	morning	walk	sit	book
13	day	walk	sit	pc
14	day	walk	stand	pc
15	day	walk	sit	pc

► Time Zone
- From Time --:--:-- to Time Zone
- “morning”, “day”, “night”

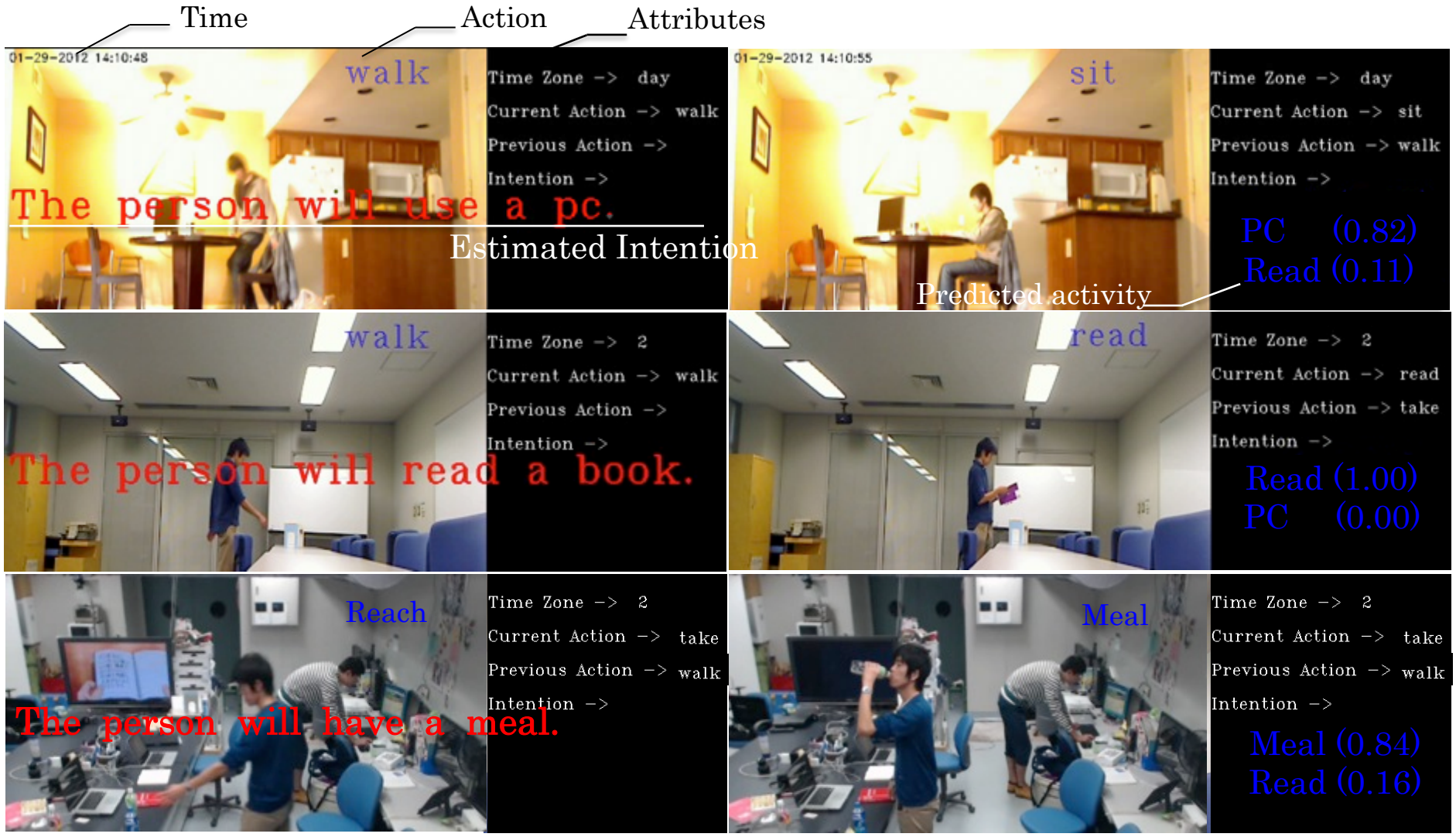
► Previous & Current Action
- “walk”, “bend”, “stand”, “sit”

► Next Activity
- d166: reading
- d4452: reaching
- d550/d560: meal (eating & drinking)

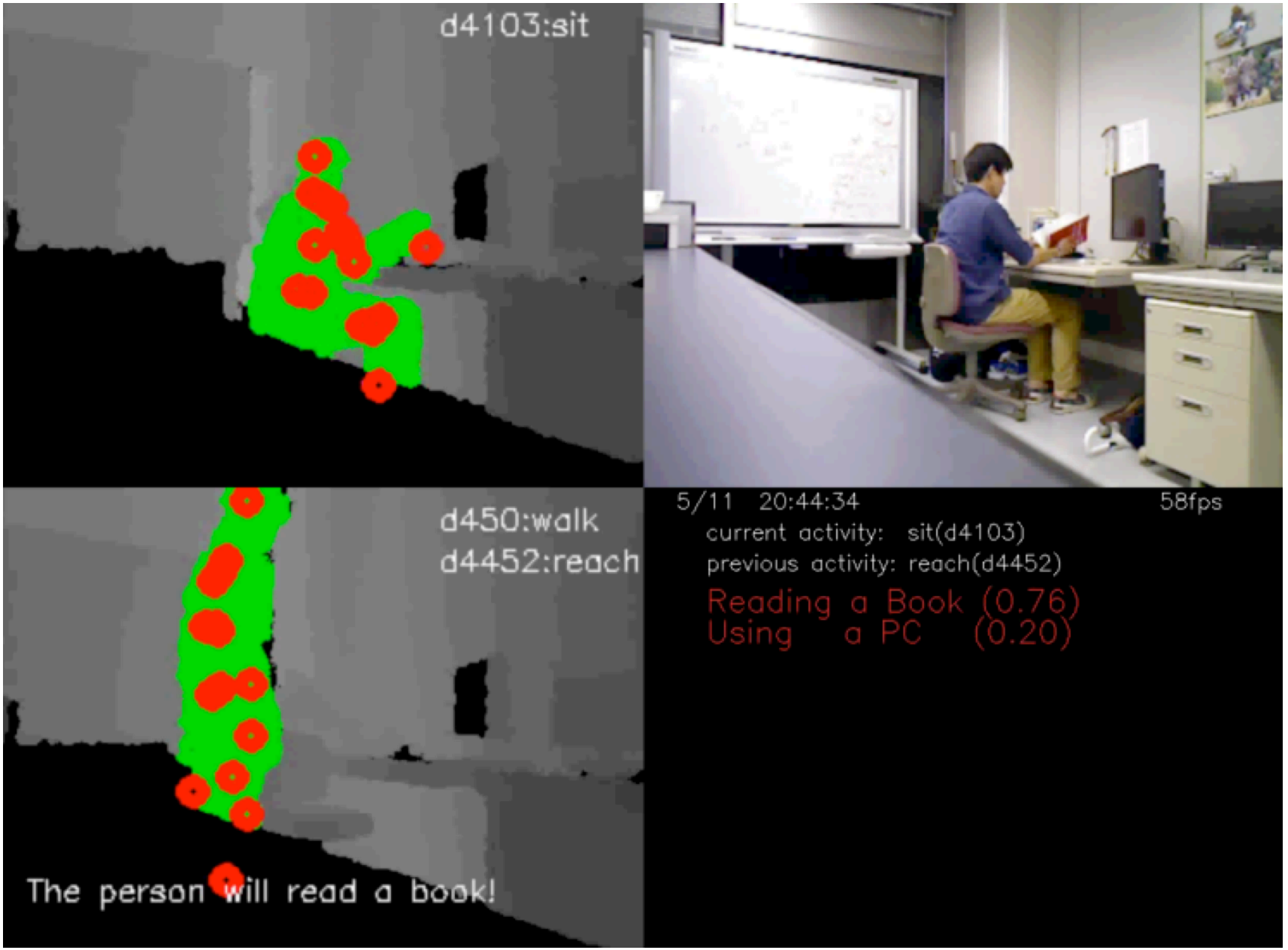
The next activity is predicted from attributes

$$\hat{\theta} = argmax_{\theta} \frac{\prod_i P(x_i|\theta)P(\theta)}{\int \prod_i P(x_i|\theta)P(\theta)d\theta}$$

Results


Time
Action
Attributes
01-29-2012 14:10:48
walk
Time Zone -> day
Current Action -> walk
Previous Action ->
Intention ->
The person will use a pc.
Estimated Intention
01-29-2012 14:10:55
sit
Time Zone -> day
Current Action -> sit
Previous Action -> walk
Intention ->
PC (0.82)
Read (0.11)
Predicted activity
walk
Time Zone -> 2
Current Action -> walk
Previous Action ->
Intention ->
The person will read a book.
read
Time Zone -> 2
Current Action -> read
Previous Action -> take
Intention ->
Read (1.00)
PC (0.00)
Reach
Time Zone -> 2
Current Action -> take
Previous Action -> walk
Intention ->
The person will have a meal.
Meal
Time Zone -> 2
Current Action -> take
Previous Action -> walk
Intention ->
Meal (0.84)
Read (0.16)

Video


d4103:sit
5/11 20:44:34
58fps
current activity: sit(d4103)
previous activity: reach(d4452)
Reading a Book (0.76)
Using a PC (0.20)
d450:walk
d4452:reach
The person will read a book!

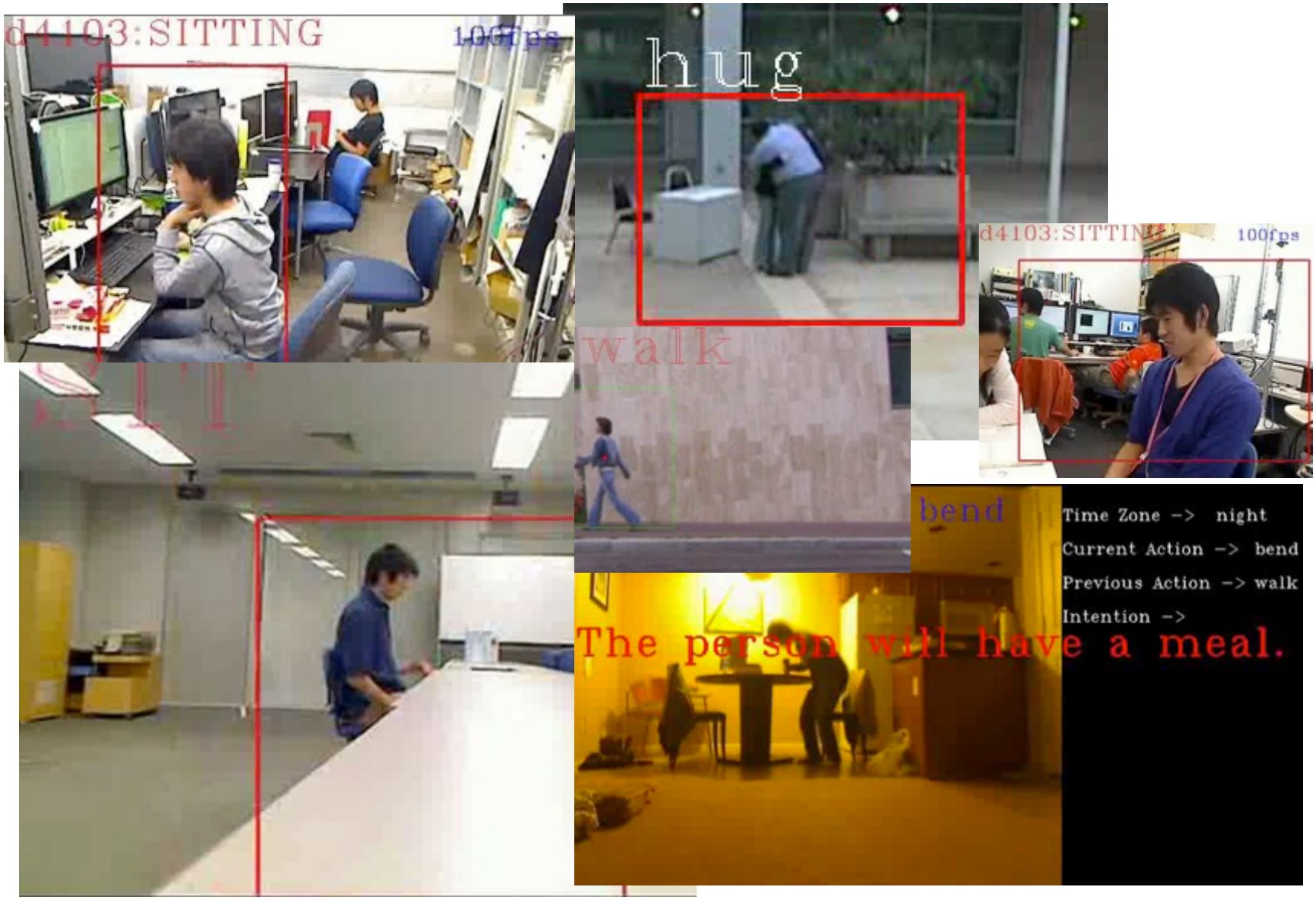
Prediction Results

- We have tried to understand in two different scenes
  - Daily living scene
  - Laboratory scene

Predicted activity	Accuracy (%)
d166: reading	73.4
d4452: reaching (using a PC)	82.0
d550 & d560: having a meal	88.5
Total	81.0

Totally, we understood 81.0% in all scenes

The Examples of Activity Prediction



Conclusion

Activity recognition & prediction projects

- Improved co-occurrence feature in fine-grained recognition
  - Extended Co-occurrence HOG
  - Parameter optimization

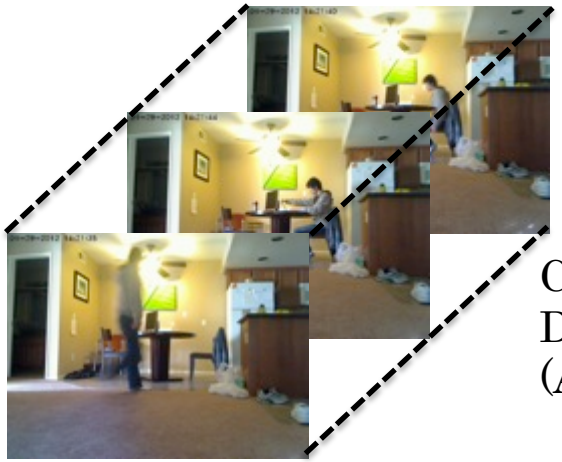
- Topic modeling based feature mining in activity recognition
  - Subset mining with latent Dirichlet allocation

- Activity recognition through recognition and data analysis
  - In Bayesian approach
  - Dense traj. and postural feature is implemented

Activity Recognition + Activity Mining

Early Mental-disorder Detection

Modeling constancy activities



Daily scenes

Creating DB,
Data Mining
(Apriori algorithm)

Activities 1 :
walking - sitting - … - eating

Activities 2 :
standing – using ** - … - cooking

…

Activities N :

Detecting abnormal behavior

Activities
(sitting, walking, running, cooking…)

Constancy activities
(from constructed data)

Anomaly detection : 67%
Anomaly is calculated by using
"one action" or "set of activity"

Confusion matrix

DT & LDA DT

Dense Trajectories (total: **73.63%, 16,000 dims**)

0.64	0.05	0.31	0.00
0.01	0.75	0.20	0.04
0.02	0.06	0.88	0.05
0.01	0.03	0.35	0.61

Latent Dirichlet Allocation (total: **84.56%, 4,161 dims**)

0.82	0.04	0.11	0.03
0.03	0.84	0.11	0.02
0.04	0.04	0.84	0.08
0.03	0.02	0.05	0.90

Multi-view implementation

Dense Trajectories (4-view in TUM data)

Descriptor	DT %(dims)
HOG	74.70(4,000)
HOF	73.07(4,000)
MBHx	73.66(4,000)
MBHy	66.71(4,000)
Integration	75.14(16,000)

Dense Trajectories (total: **75.14%, 16,000 dims**)

0.68	0.20	0.06	0.06
0.23	0.65	0.03	0.09
0.13	0.08	0.74	0.05
0.01	0.00	0.05	0.94

Topic Modeling for BoF Analysis

LDA adopts "frequent subset" in DT based BoF

- Topic word analysis converts visual computing with topic modeling
  - Word in article = Visual word in dense trajectories
  - Topic = Feature vector of activity at each view

- "An activity at each view" has "a topic"
  - Number of topic in LDA => "numof activity" x "numof view"
  - e.g. IXMAS dataset[4] : 55 topics (11 activities in 5 views)

[4] D. Weinland, R. Ronfard, E. Boyer, "Free Viewpoint Action Recognition using Motion History Volumes", CVIU, 2006.

Process Flow of Association Analysis[2]

vec no.	Item
vec1	B6, B9
vec2	B6, B9
vec3	B6, B9
vec4	B6, B9
vec5	B4, B6, B9

Cutting Bin4

Item	support
{B4}	1 (20%)
{B6}	5 (100%)
{B9}	5 (100%)

Item	support
{B6}	5
{B9}	5

Calculating support value with combination

Item	support
{B6, B9}	5 (100%)

("B6", "B9") is selected vector

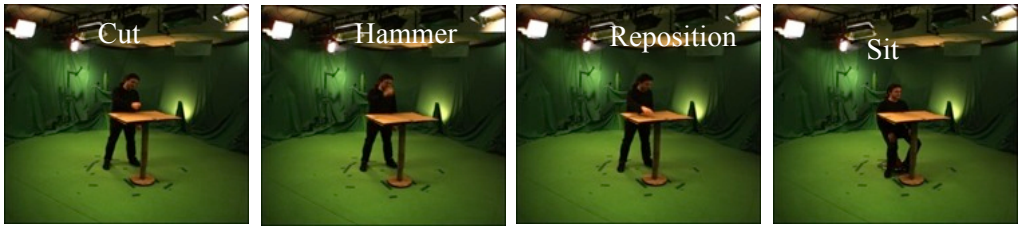
- The vector is not suitable for high-dimensional vector
  (only effective for a couple of activity classes)
- Feature combination tends to be a large space

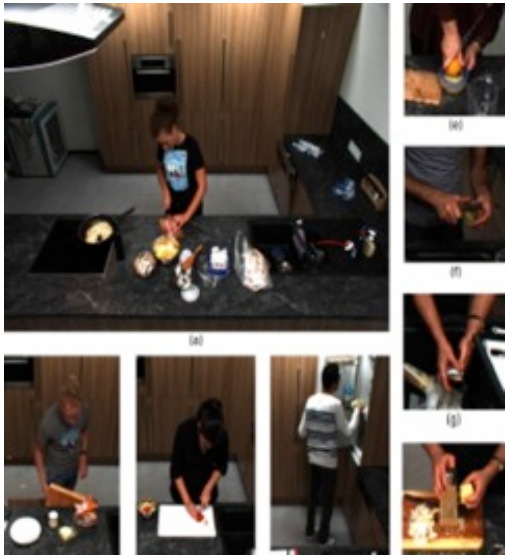
[2] R. Agrawal et al., "Mining Association Rules between Sets of Items in Large Databases", in SIGMOD1993.

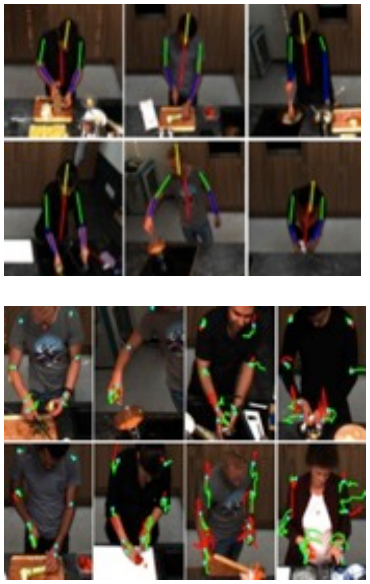
Experiment

Settings of the experiment

- Data: IXMAS(multi-view), INRIA(multi-view), and MPII cooking dataset(fine-grained)
- Dense trajectories + BoF
- Feature mining with LDA
- Numof topic: “action” x “view”







"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

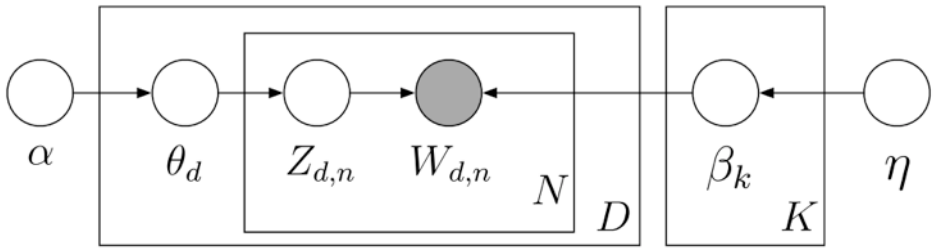


Figure 4: **The graphical model for latent Dirichlet allocation.** Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes–the topic proportions, assignments and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The N plate denotes the collection words within documents; the D plate denotes the collection of documents within the collection.