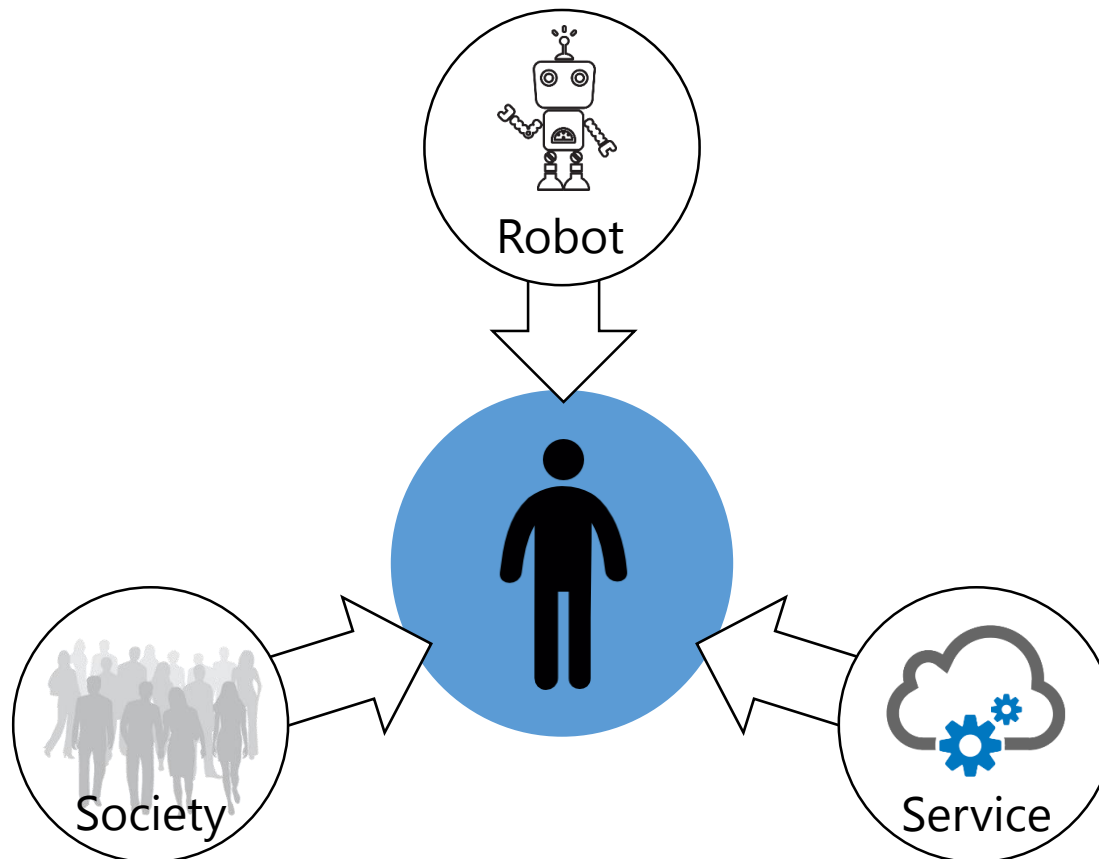


# **Understanding humans:** identity, communication, state, and more

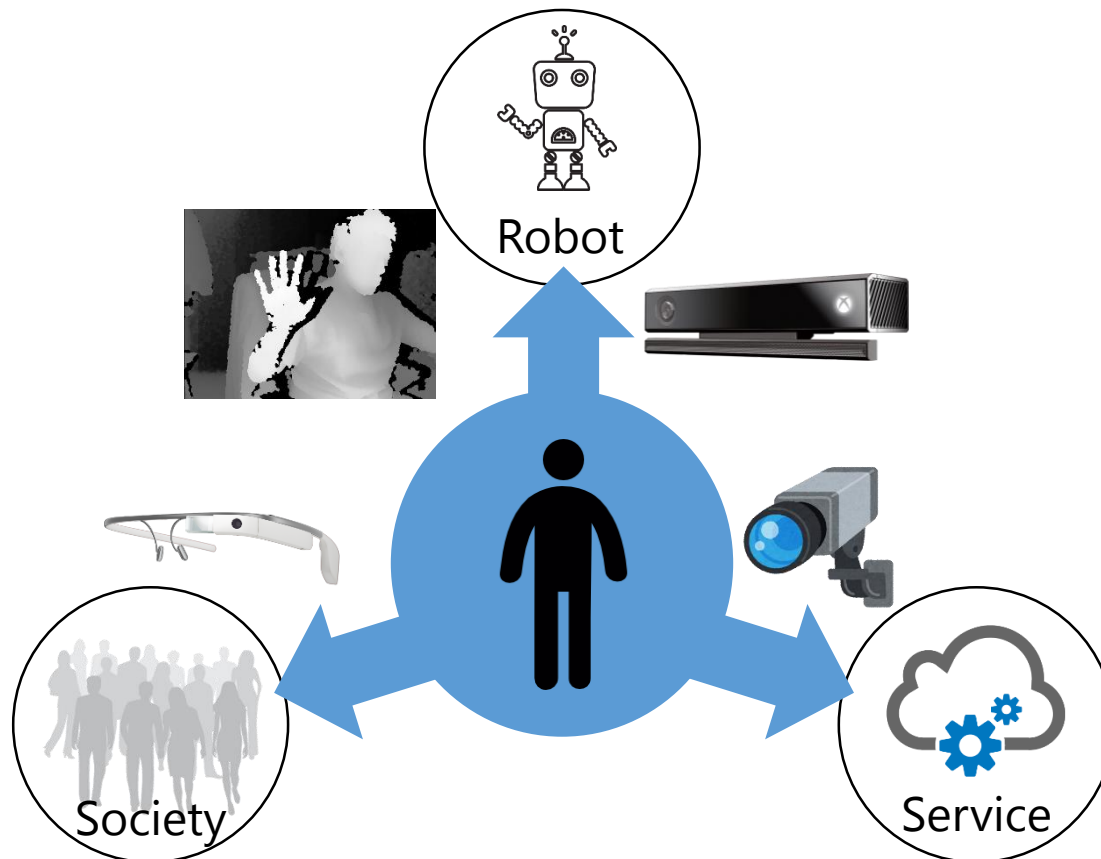
**Yang Wu**  
**(伍洋)**

Nara Institute of Science and Technology  
奈良先端科学技术大学院大学

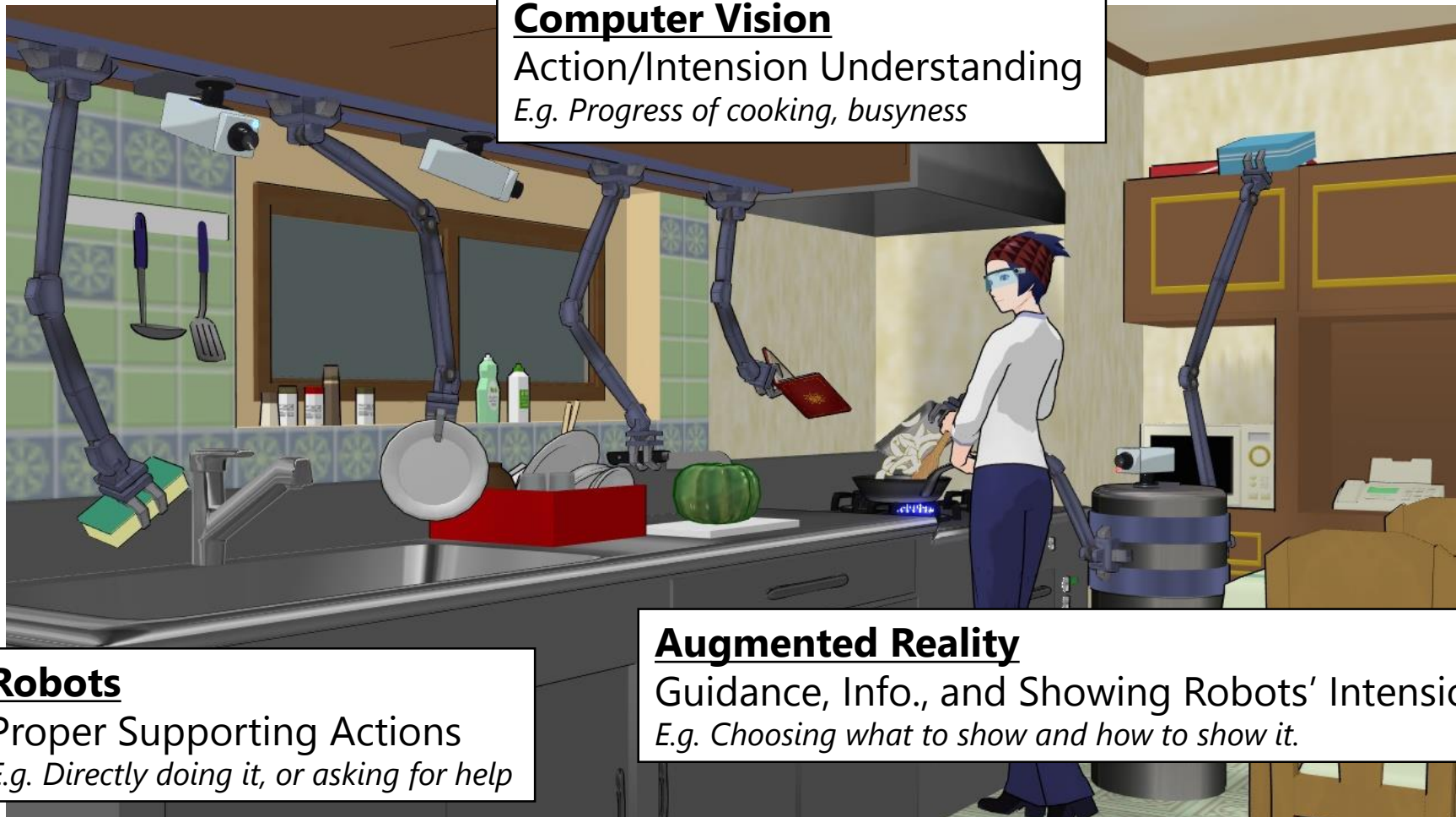
## For helping a person



the system needs to  
**understand** the person



## A possible application scenario



### Computer Vision

Action/Intension Understanding  
*E.g. Progress of cooking, busyness*

### Robots

Proper Supporting Actions  
*E.g. Directly doing it, or asking for help*

### Augmented Reality

Guidance, Info., and Showing Robots' Intension  
*E.g. Choosing what to show and how to show it.*



# Identity

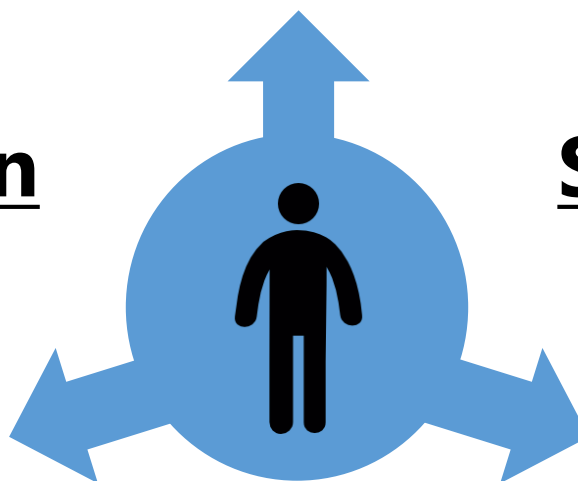
*(Who?)*

# Communication

*(What [does he/she want]?)*

*How [does he/she feel]?)*

**Explicit expression**



# State, Action, ...

*(What [is he/she doing]?)*

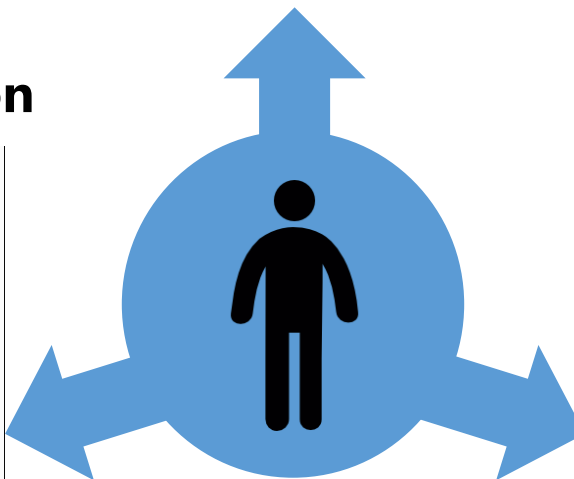
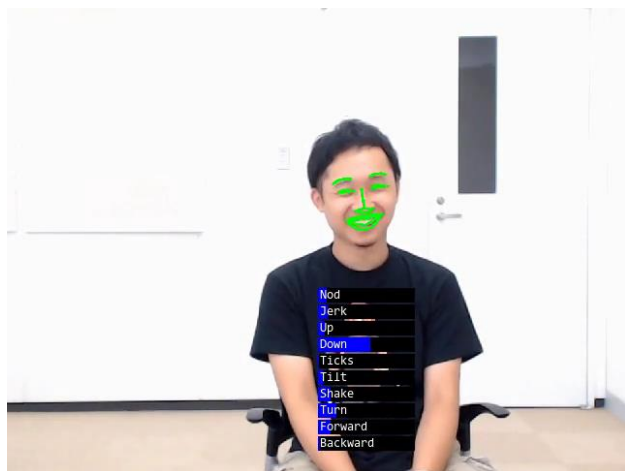
*How [does he/she do it]?)*

**Implicit expression**

## Across-camera Person Re-identification



## Head Gesture Recognition

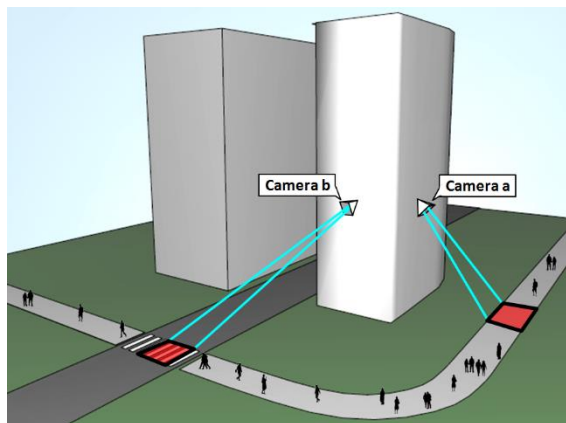


## 3D Hand Tracking



Identity: in-a-distance and unobtrusive

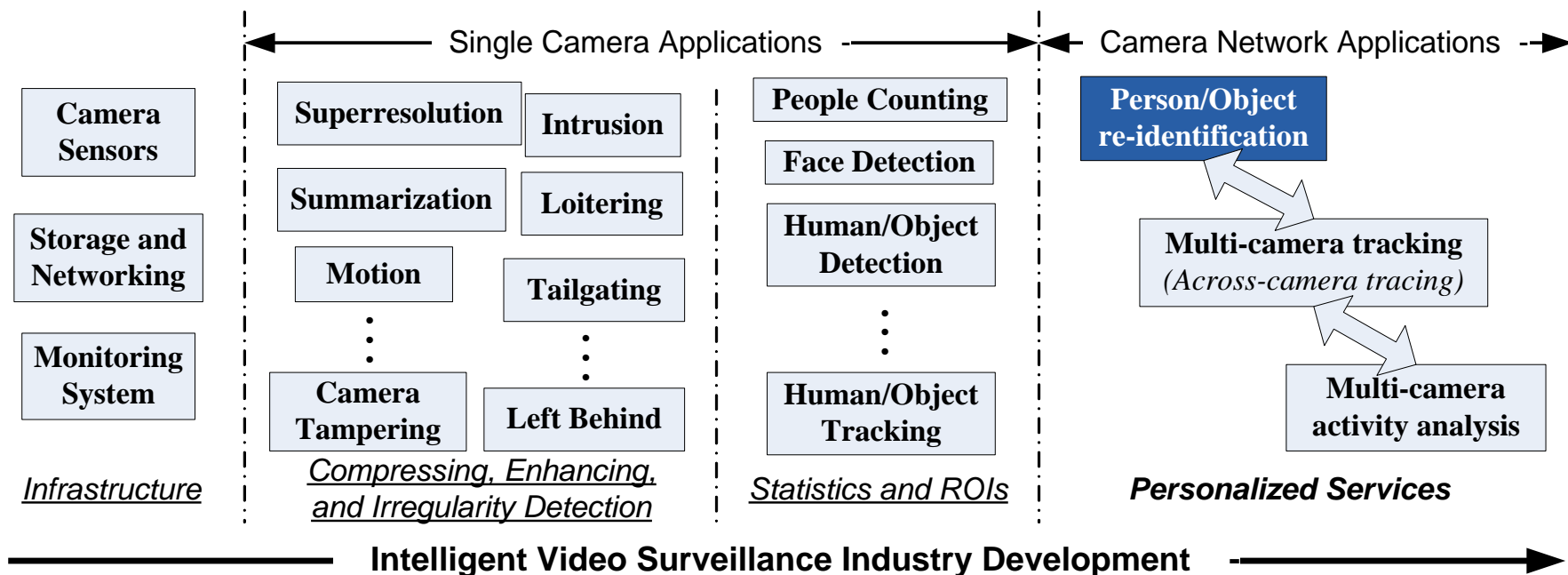
## Person re-identification (Re-ID)



To look for a specific person  
in a camera network



# Re-ID in the Context of Video Surveillance



## Problem Introduction: *Subtypes and Our Focus*

Single-shot



**“multiple-shot”  
is more generic  
and useful**

Multiple-shot



(a) Two camera views

**Our main  
interest!**

(b) Images of sampled individual persons

# Single-shot: Looking at the “Pose”

- Pose Normalization

[ECCV 2018] Qian et. al., “Pose-Normalized Image Generation for Person Re-identification”.

- Pose Adaptation

[submitted to AAAI 2019] Qiu et. al., “Pose-adaptive Image Generation for Person Re-identification”.

# Challenges

Body movements

Camera viewpoints

Occlusions

## **Key challenges**

Background

Illumination

## **Environmental challenges**

Clothes

Accessories

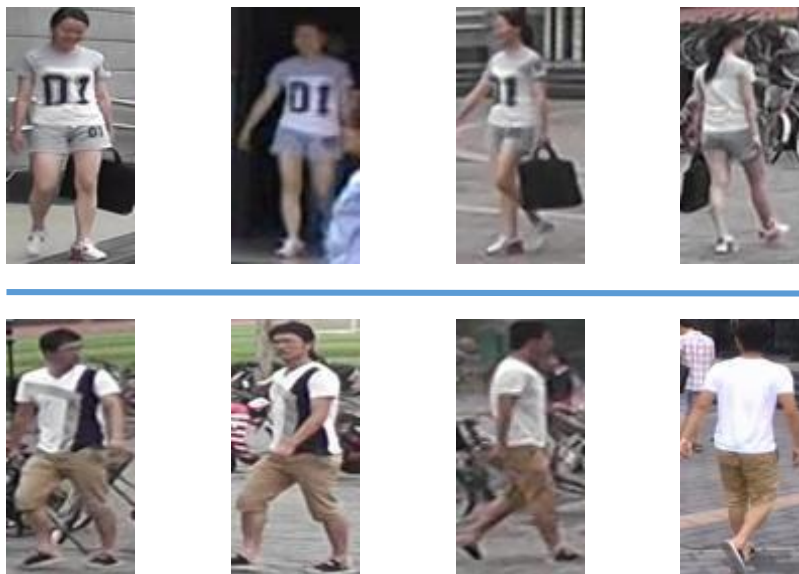
## **Others**

**pose**  
**variations**



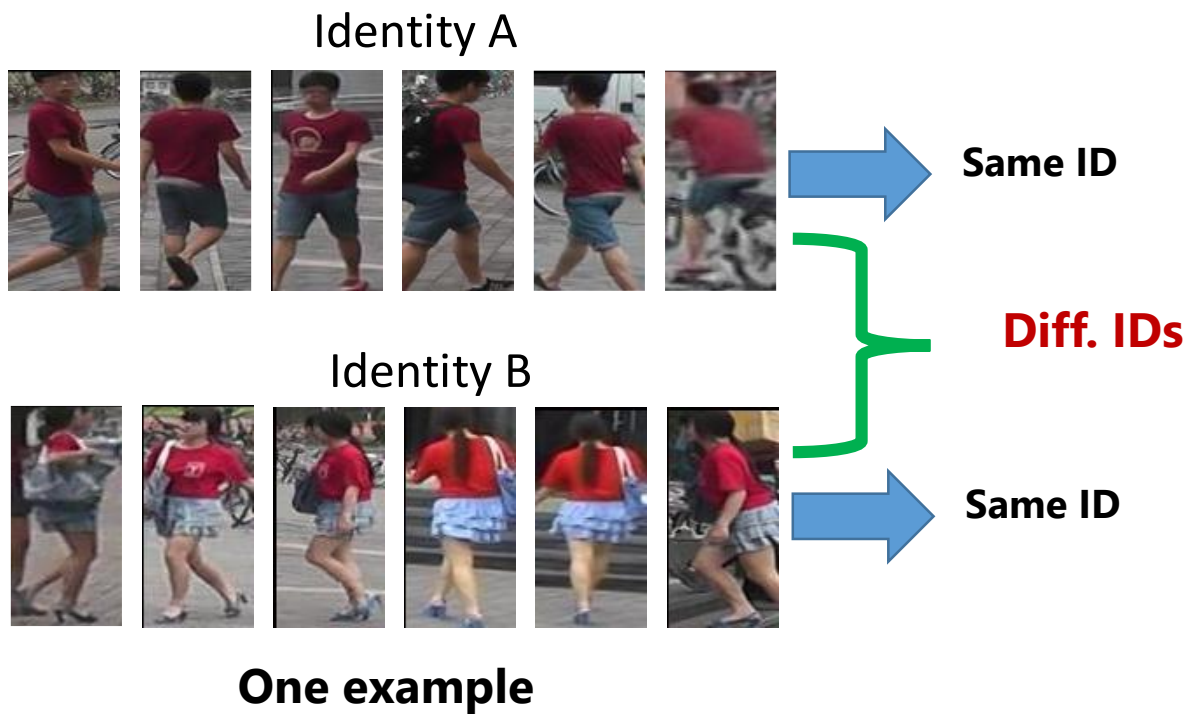
# Motivation

## 1. Lack of cross-view paired **training data**



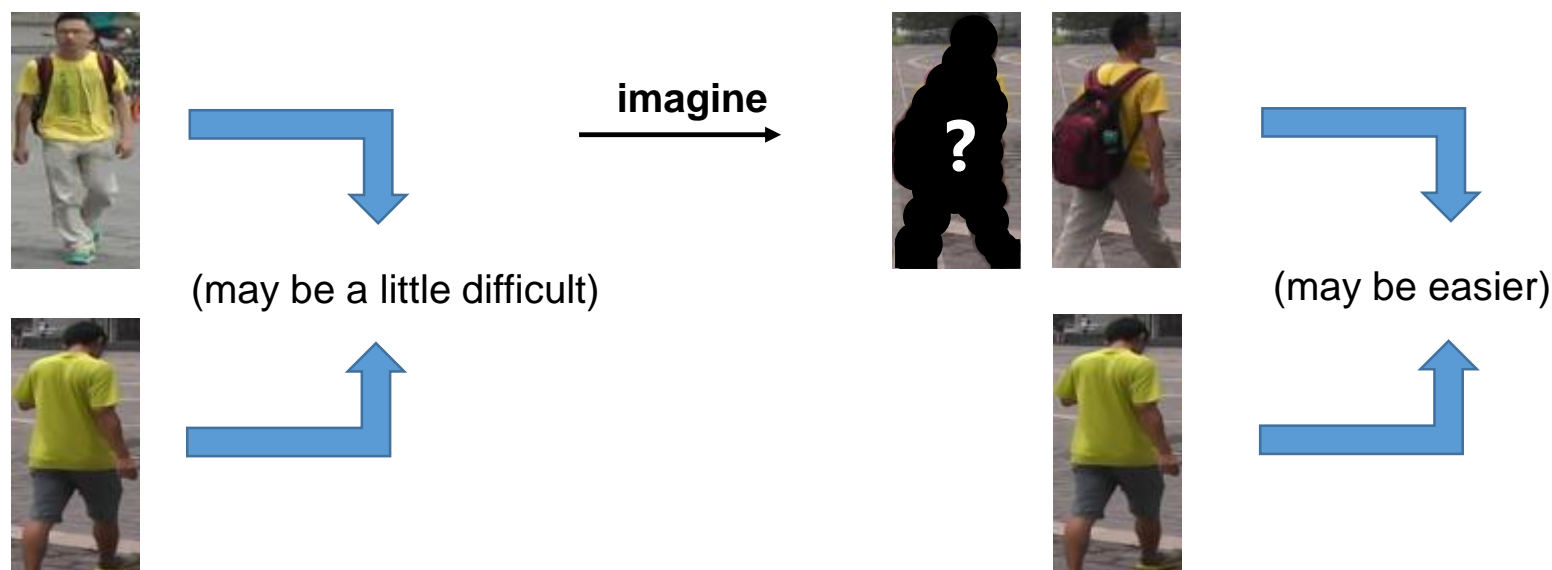
# Motivation

## 2. **Identity-sensitive** and **View-invariant** representation

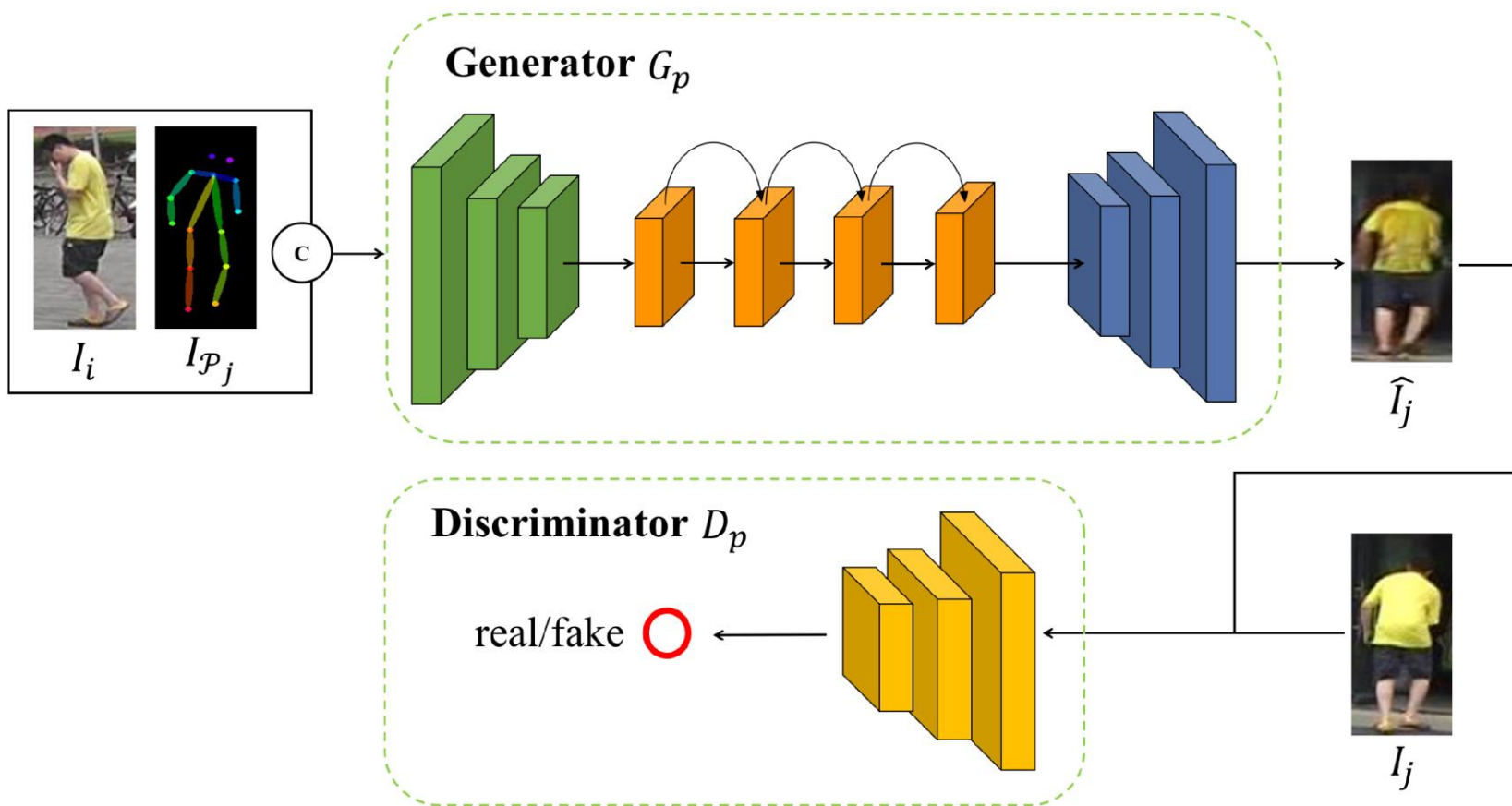


# Proposal

**Key idea:** Eliminating the pose differences



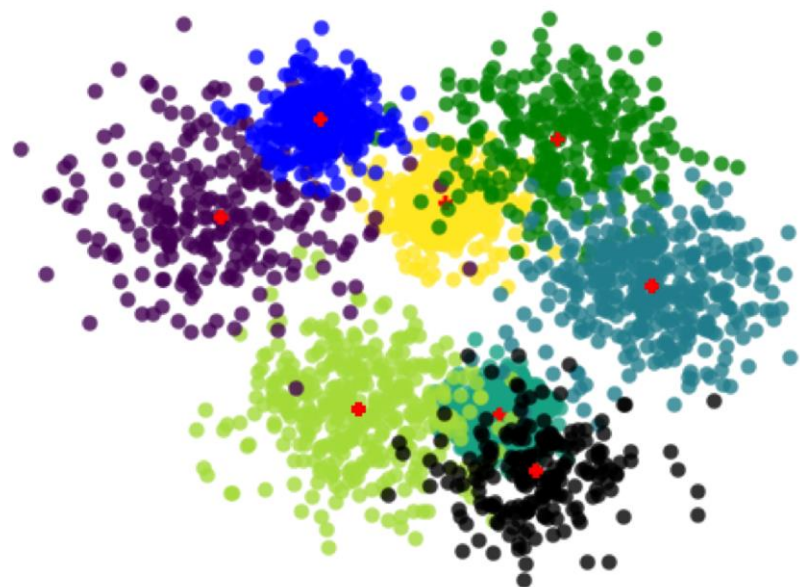
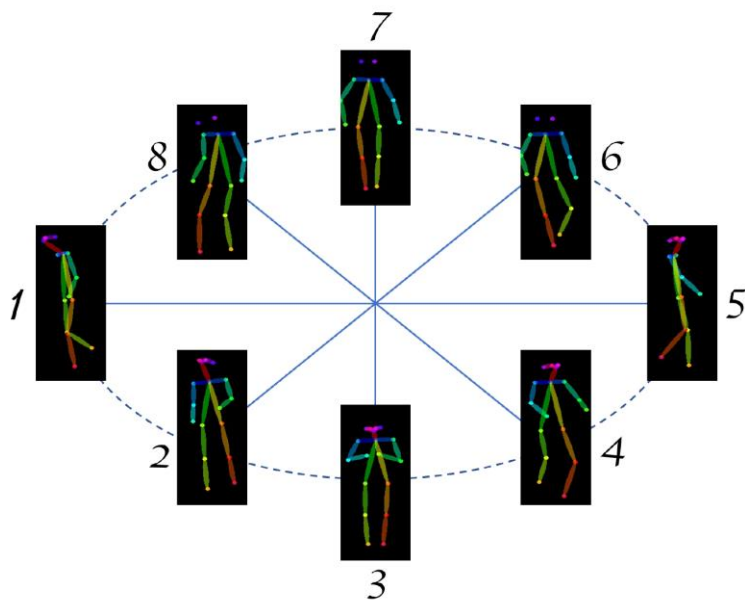
# Network (PN-GAN)



## Network (eight canonical poses)

1. **Pose estimation** – OpenPose [1];
3. **Pose clustering** – K-means.

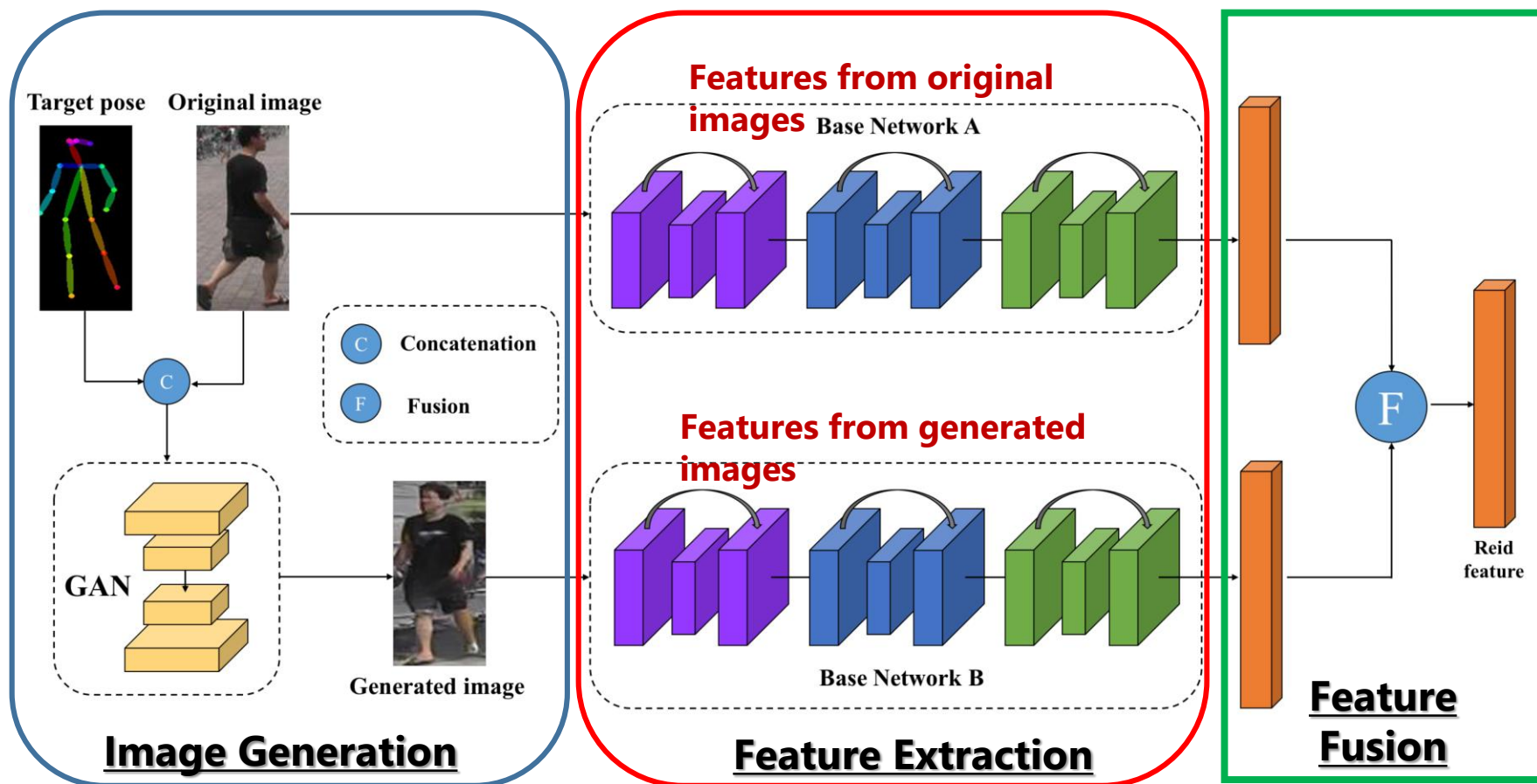
2. **Feature extraction** – ResNet-50;



(a) Eight canonical poses on Market-1501      (b) t-SNE visualization of different poses.

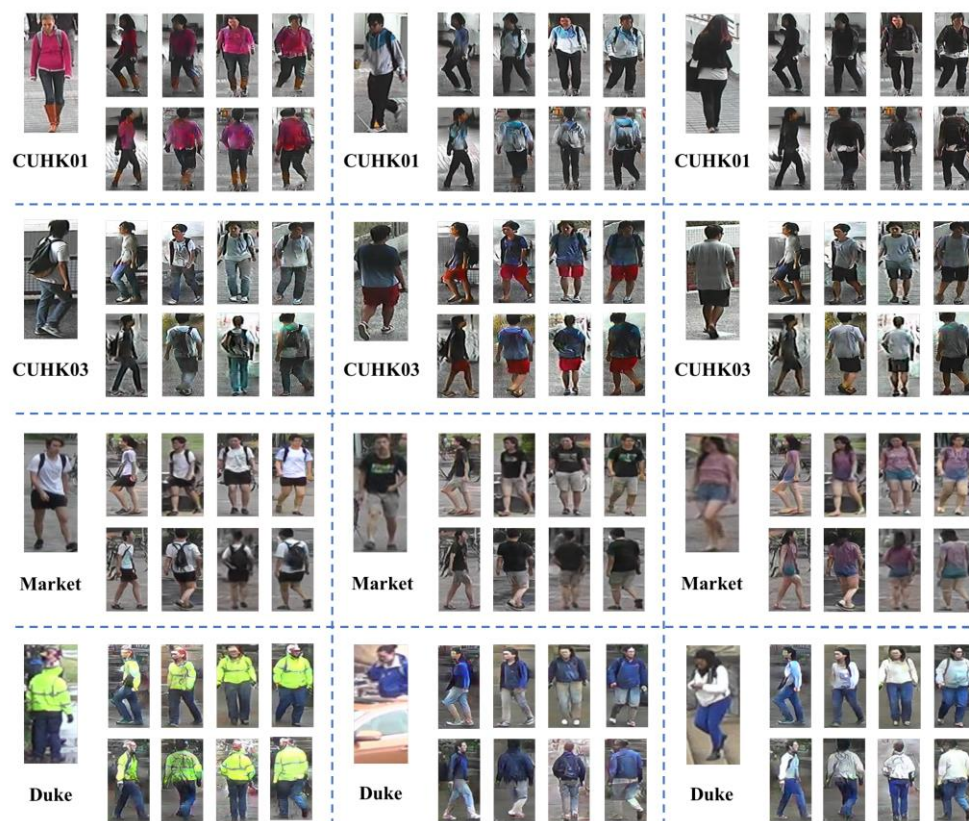
[1] Cao,Z.,Simon,T.,Wei,S.E.,Sheikh,Y.:Realtimemulti-person2dposeestimation using part affinity fields. In: CVPR (2017)

# Network (framework)





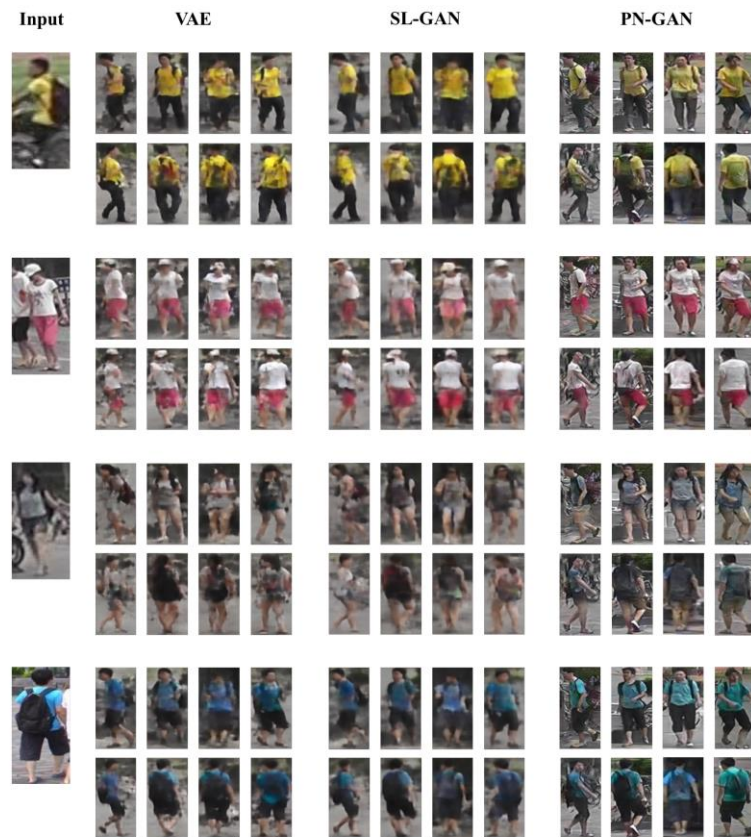
# Visualization



**Fig. 2.** The visualization of generated images on Market-1501, DukeMTMC-reID, CUHK03 and CUHK01.



# Visualization



**Fig. 3.** Comparing images with the eight canonical poses synthesised using our PN-GAN and a number of alternative models.

## Code

**[https://github.com/naiq/PN\\_GAN](https://github.com/naiq/PN_GAN)**



# Another Strategy: Pose Adaptation

- Generating data with an **arbitrary pose** for **any specific person**.
  - Enhancing the generation with **Re-ID specific losses**
- Forcing the ReID model to be **pose invariant**.

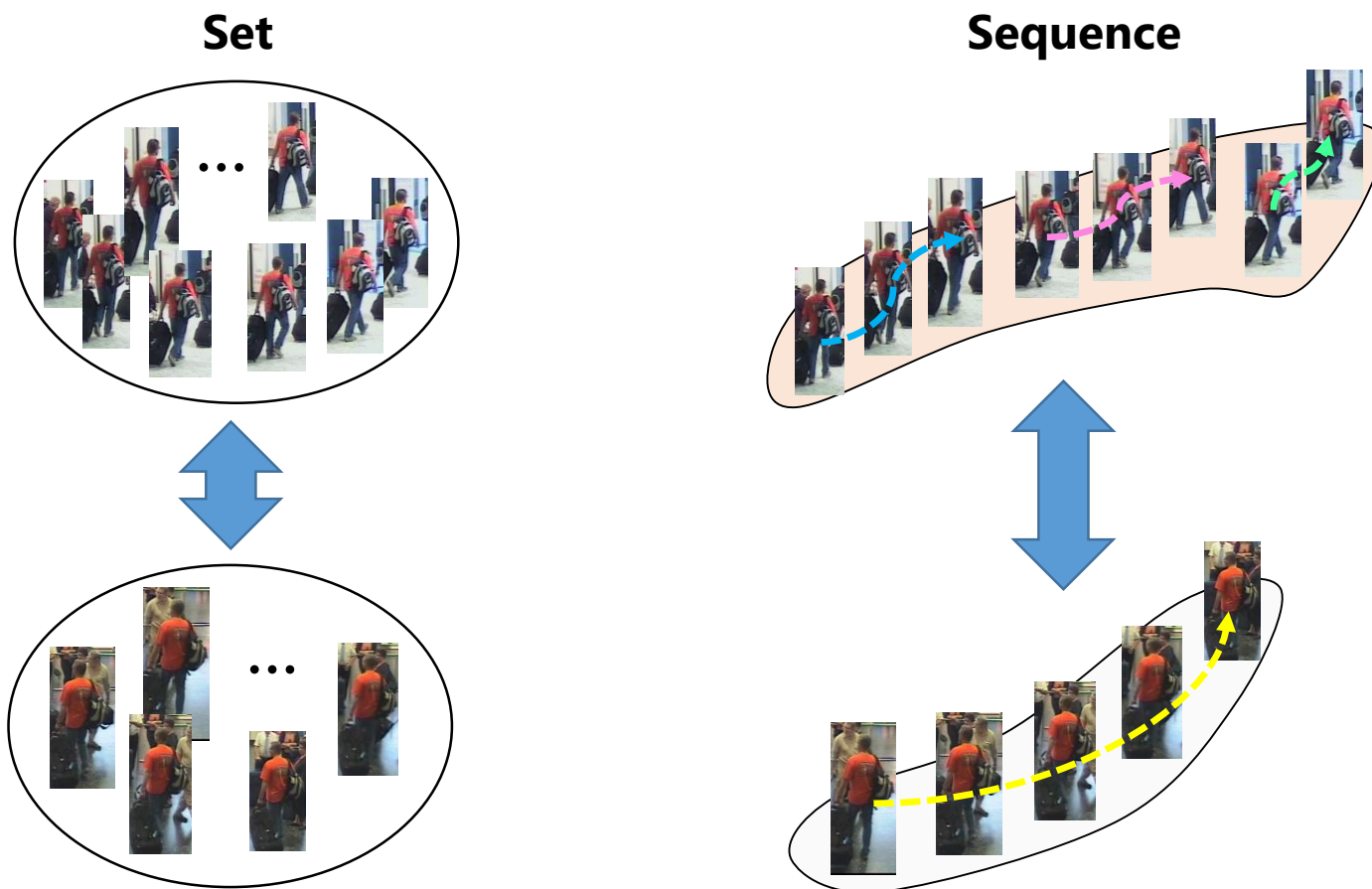


[submitted to AAAI 2019]

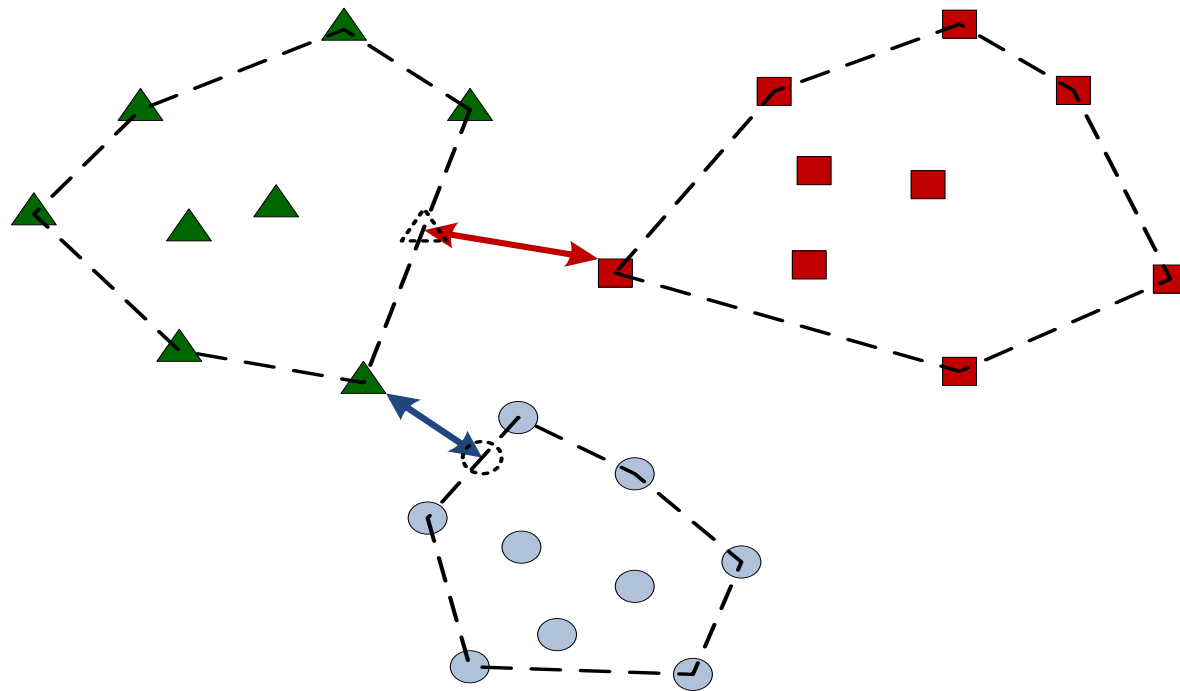
Qiu et. al., "Pose-adaptive Image Generation for Person Re-identification".

# Video-based ReID:

## *Perspectives of Set and Sequence*

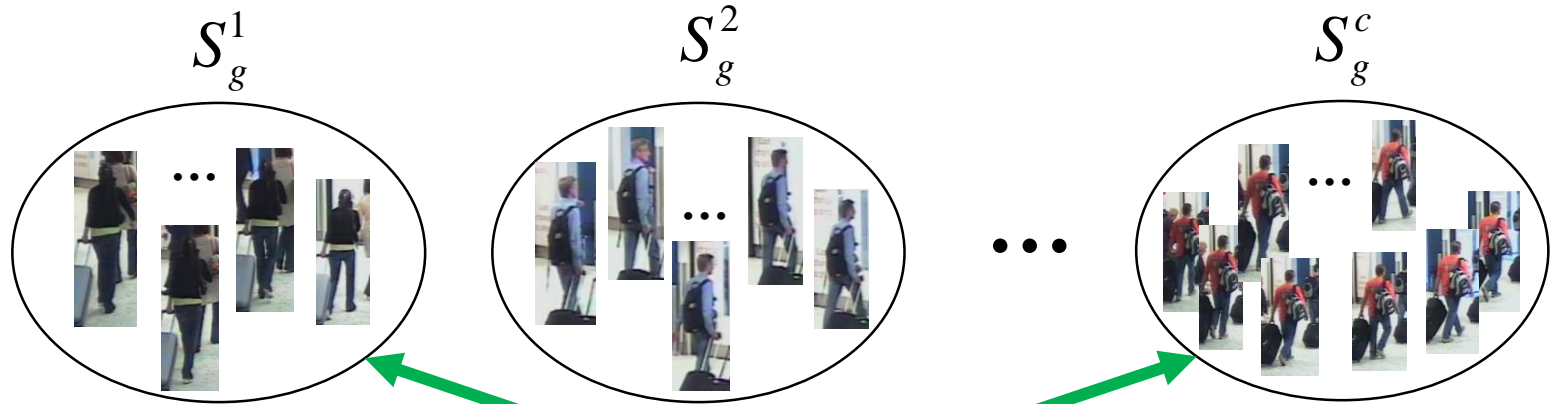


# Set: Robustness and Flexibility of Geometry

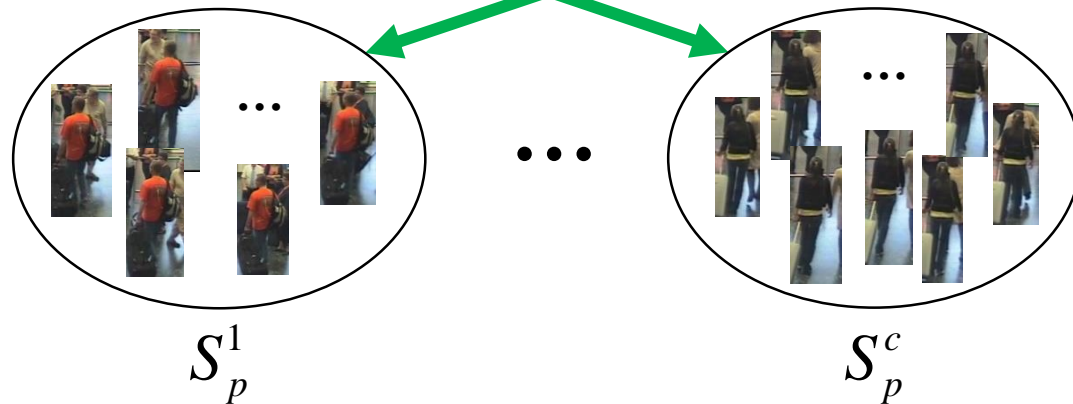


# Multiple-shot Re-ID: A Set-based Perspective

Gallery

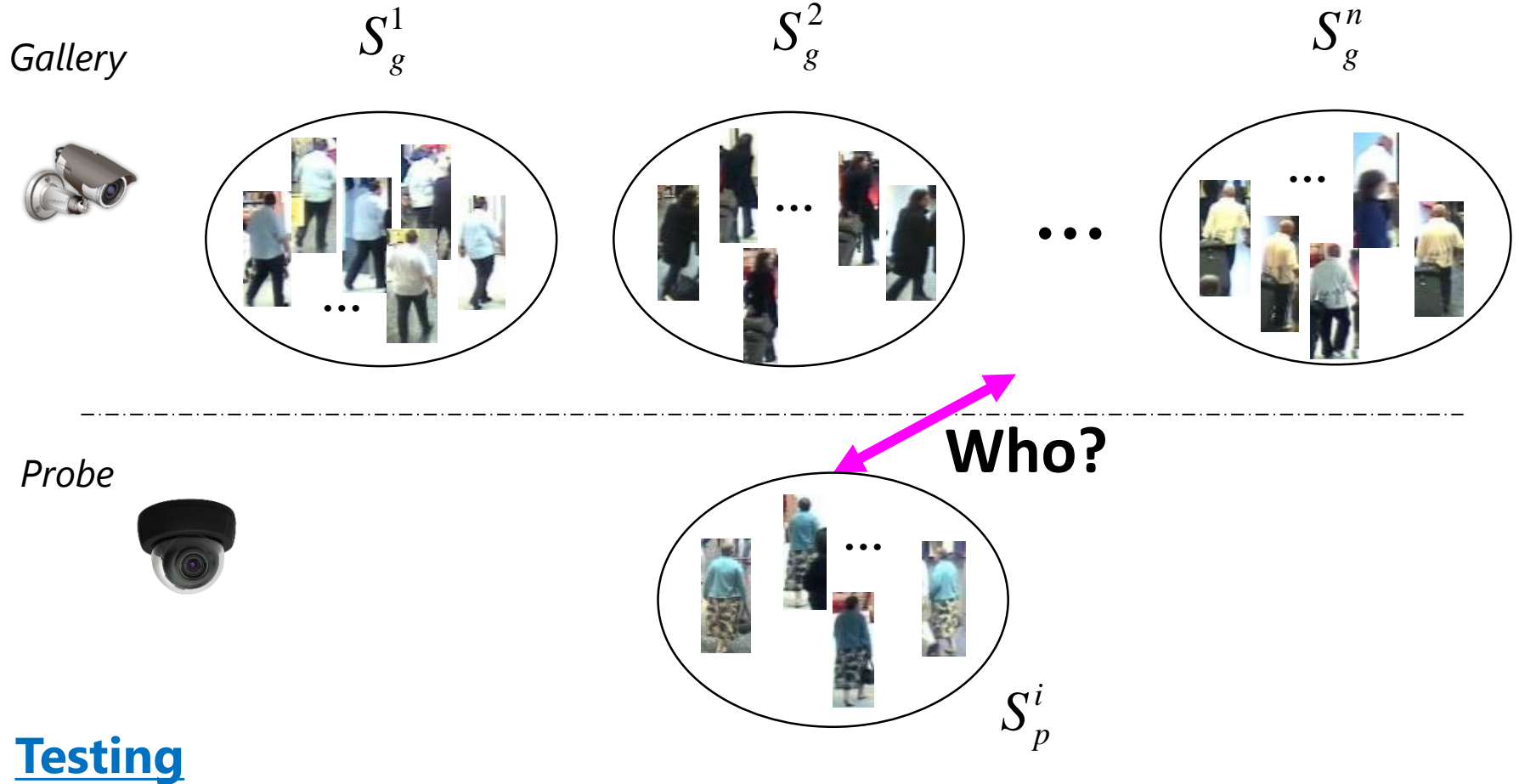


Probe



**Training**

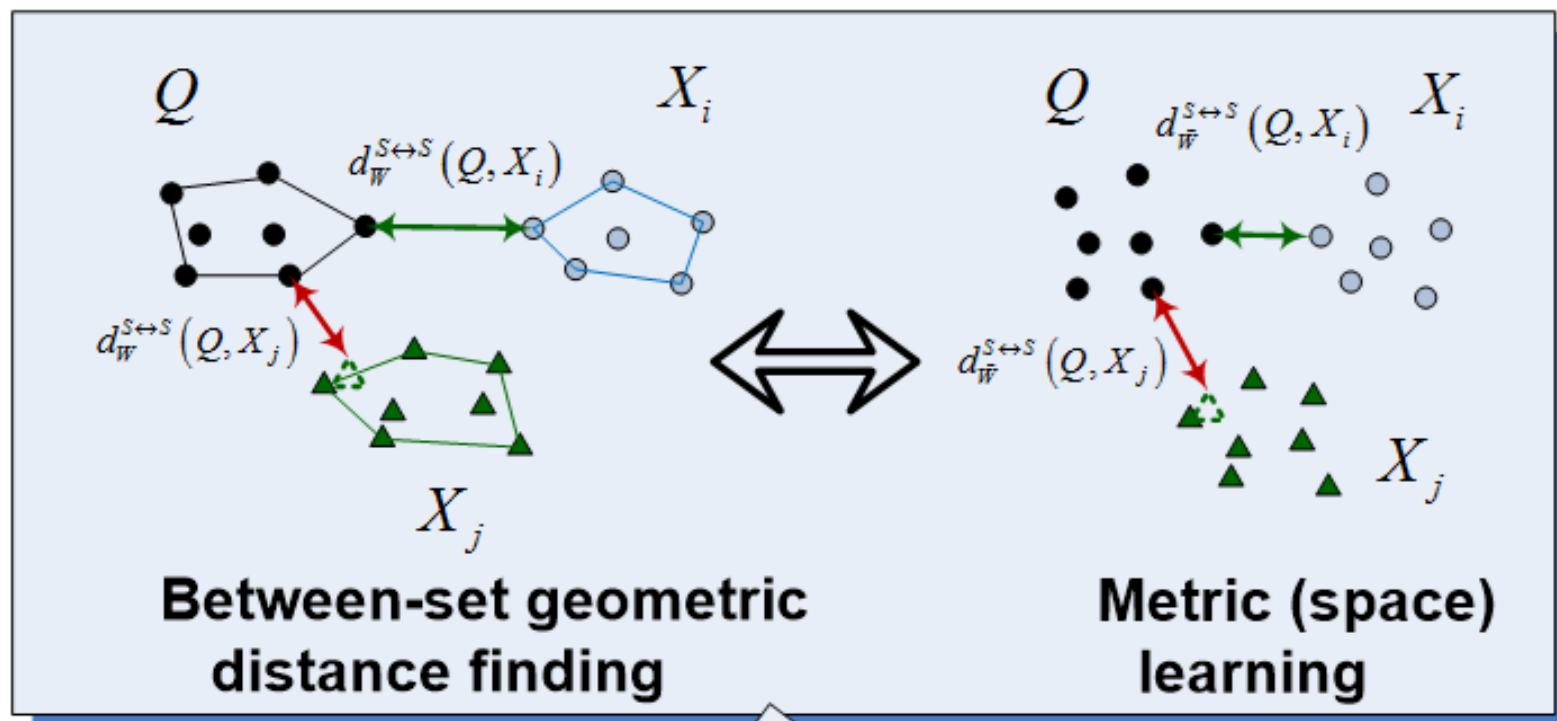
# Multiple-shot Re-ID: A Set-based Perspective





One direction → **Parametric** methods

## Set-to-set distance + metric learning

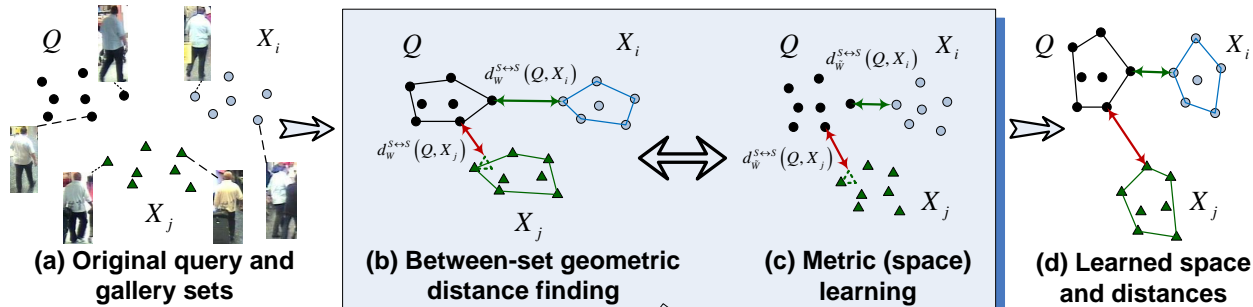


[ECCV 2012] Yang Wu, et al., "**Set based discriminative ranking for recognition**".

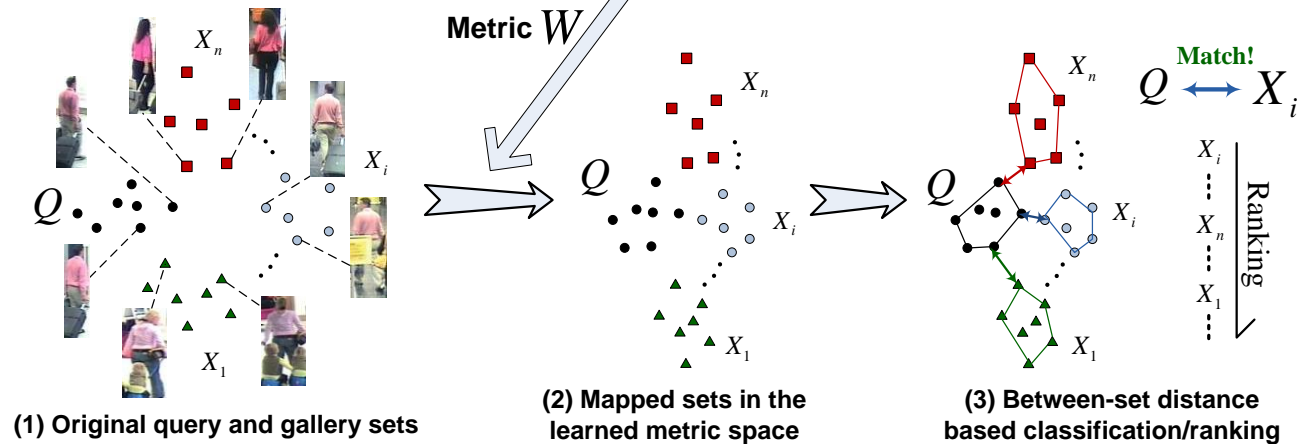
One direction → Parametric methods

# Set-to-set distance + metric learning

## Training Stage:



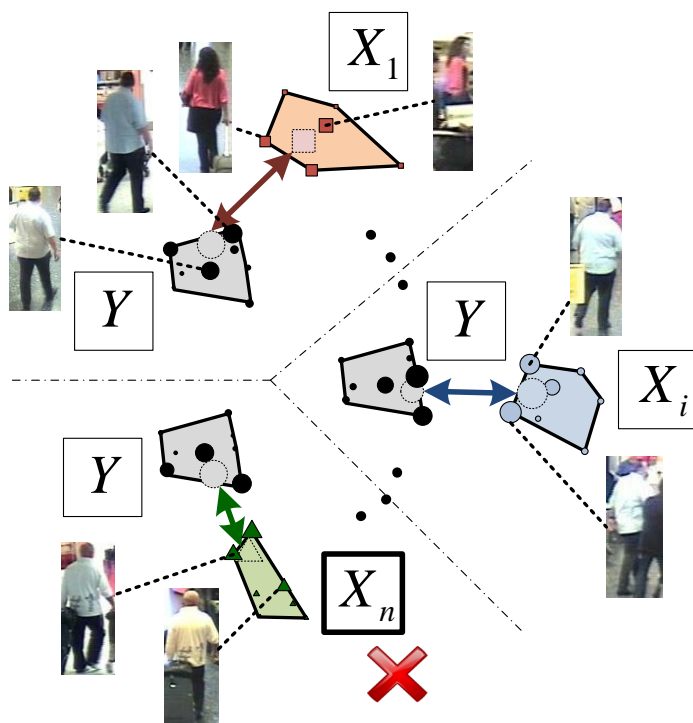
## Testing Stage:



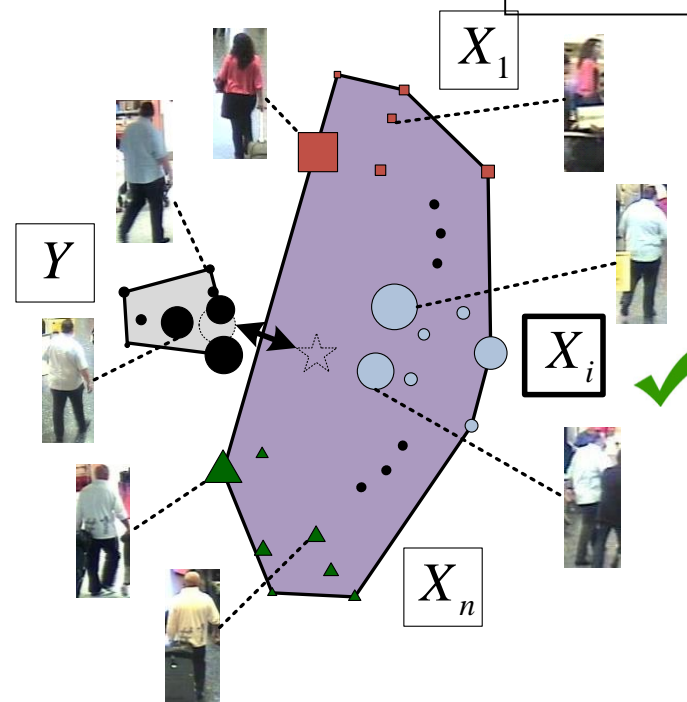
[ECCV 2012] Yang Wu, et al., "**Set based discriminative ranking for recognition**".

# Collaborative representation

$Y$  -- Query/Probe Set  
 $\mathbf{X}_i, i \in \{1, \dots, n\}$   
-- Gallery Sets



(a) Set-to-set distances



(b) Set-to-sets distance

(MPD, AHISD/CHISD, SANP/KSANP, RNP)

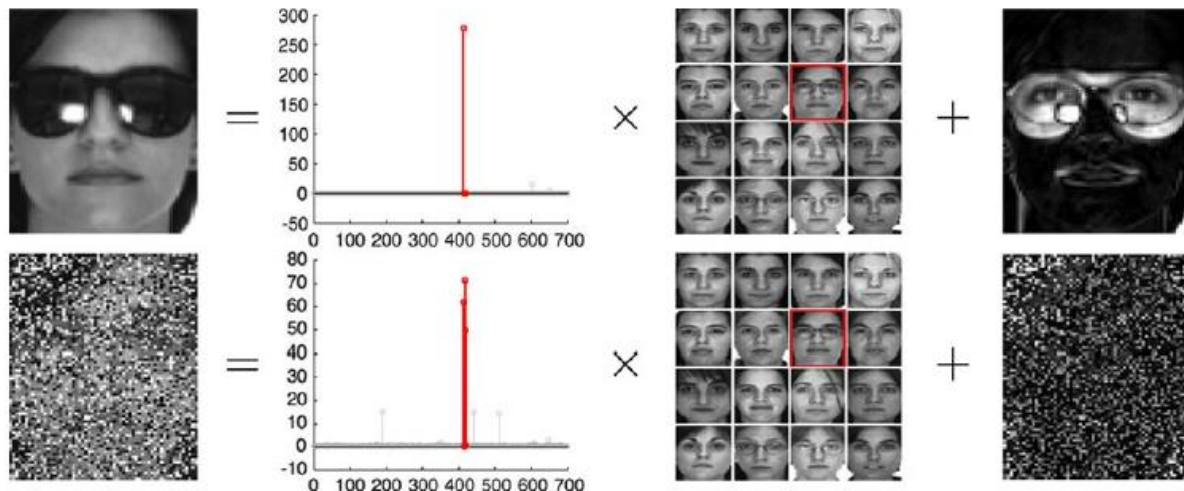
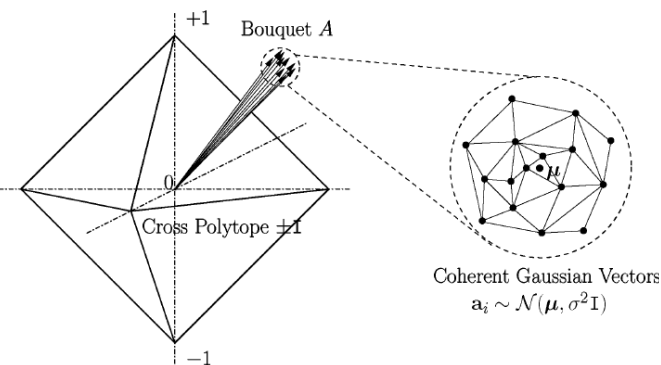
(CSA, CRNP, LCSA, LCRNP, CMA)

## Sparse representation based classification

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{y} - \mathbf{X}\mathbf{a}\|_2^2 + \lambda_1 \|\mathbf{a}\|_1.$$

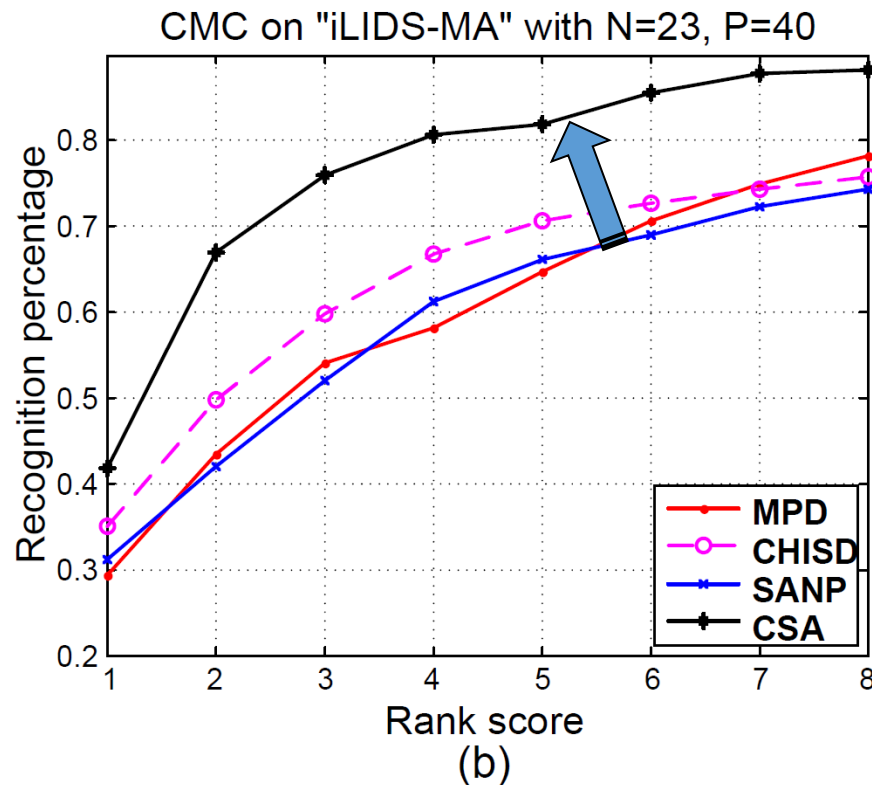
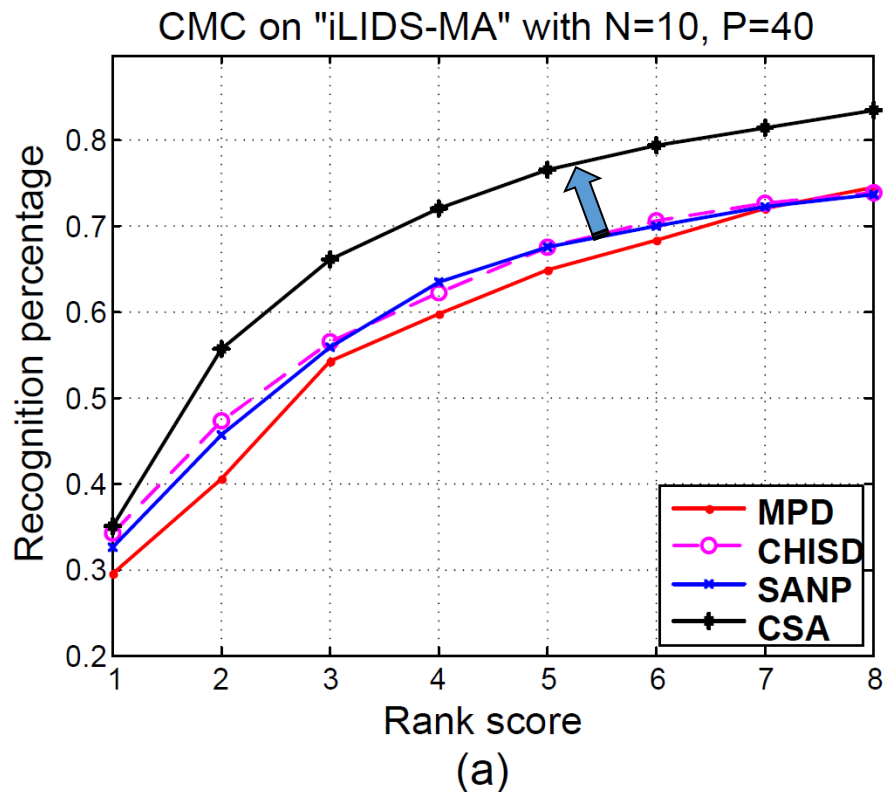
$$r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}_i \hat{\mathbf{a}}_i\|_2^2, \forall i,$$

$$C(\mathbf{y}) = \arg \min_i r_i(\mathbf{y}).$$



**J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma., Robust Face Recognition via Sparse Representation. IEEE TPAMI, 31(2):210–227, 2009.**

## Results (Sparse model)



## Results (**Non-sparse** model)

**Face recognition** accuracy (%) comparison on the **Honda/UCSD** dataset.

| Method     | MPD[4] | SRC[8] | CRC[14] | CHISD[2]     | SANP[13]     | KSANP[6] | SBDR[10] | CSA[9] | RNP[12]      | CRNP         |
|------------|--------|--------|---------|--------------|--------------|----------|----------|--------|--------------|--------------|
| 50 frames  | 79.49  | 76.92  | 76.92   | 79.49/82.05* | 84.62/84.62* | 87.18*   | 87.69*   | 84.62  | 66.67/87.18* | <b>89.74</b> |
| 100 frames | 87.18  | 94.87  | 82.05   | 79.49/84.62* | 89.74/92.31* | 94.87*   | 89.23*   | 92.31  | 92.31/94.87* | <b>97.44</b> |

**Face recognition** accuracy (%) comparison on the **CMU MoBo** dataset.

| Method     | MPD[4] | SRC[8] | CRC[14] | CHISD[2] | SANP[13] | SBDR[10] | CSA[9] | RNP[12]            | CRNP         |
|------------|--------|--------|---------|----------|----------|----------|--------|--------------------|--------------|
| 50 frames  | 92.22  | 88.89  | 89.72   | 90.83    | 90.14    | 95.00*   | 86.25  | 91.81/91.9*        | <b>93.33</b> |
| 100 frames | 94.31  | 92.36  | 93.06   | 94.17    | 93.61    | 96.11*   | 94.44  | <b>94.58/94.7*</b> | 94.44        |

Performance comparison for **person re-identification** on three benchmark datasets.

| Dataset     | MPD[4]     | SRC[8]            | CRC[14]    | CHISD[2]   | SANP[13]   | CSA[9]            | RNP[12]    | CRNP                |
|-------------|------------|-------------------|------------|------------|------------|-------------------|------------|---------------------|
| iLIDS-MA    | 50.0(75.0) | 57.3(74.8)        | 28.5(50.0) | 52.5(72.8) | 46.8(74.8) | <b>59.0(71.3)</b> | 53.3(76.0) | <b>59.0(78.3)</b>   |
| iLIDS-AA    | 23.8(60.4) | <b>36.0(68.9)</b> | 24.7(54.1) | 24.6(58.2) | 19.2(57.3) | 22.5(59.6)        | 25.5(59.9) | 35.4( <b>71.6</b> ) |
| CAVIAR4REID | 19.0(47.2) | 25.4(50.8)        | 16.6(37.6) | 25.4(51.2) | 25.2(52.4) | 24.6(48.8)        | 24.0(50.2) | <b>26.8(63.6)</b>   |

# Results (**Non-sparse** model)

- Computational cost

For those methods which can have (parts of) their models pre-computed using the training data, the total pre-computation time (in seconds) is listed for comparison.

| Dataset  | Honda/UCSD         |                    | CMU MoBo           |                    | iLIDS-MA | iLIDS-AA | CAVIAR4REID |
|----------|--------------------|--------------------|--------------------|--------------------|----------|----------|-------------|
|          | 50 frames          | 100 frames         | 50 frames          | 100 frames         |          |          |             |
| SBDR[10] | $9.23 \times 10^3$ | $1.46 \times 10^4$ | $1.23 \times 10^4$ | $3.14 \times 10^4$ | N/A      | N/A      | N/A         |
| CSA[9]   | 0.59               | 0.74               | 28.7               | 50.2               | 0.39     | 0.62     | 0.26        |
| RNP[12]  | 0.06               | 0.20               | 0.17               | 0.64               | 0.02     | 0.05     | 0.02        |
| CRNP     | 0.22               | 0.87               | 0.64               | 2.66               | 0.04     | 0.22     | 0.02        |

Computational cost comparison with all the related methods on all of the recognition tasks (in the "**milliseconds per sample**" manner, excluding the time for feature extraction).

| Dataset          | MPD[4]     | SRC[8]            | CRC[14] | CHISD[2] | SANP[13] | SBDR[10] | CSA[9] | RNP[12] | CRNP        |
|------------------|------------|-------------------|---------|----------|----------|----------|--------|---------|-------------|
| Honda/UCSD (50)  | 3.2        | $1.2 \times 10^3$ | 0.28    | 77.7     | 19.6     | 259      | 17.4   | 11.5    | <b>0.32</b> |
| Honda/UCSD (100) | 6.4        | $4.1 \times 10^3$ | 0.55    | 330      | 17.3     | 97.8     | 32.6   | 14.5    | <b>0.46</b> |
| CMU MoBo (50)    | 12.4       | $7.6 \times 10^3$ | 0.94    | 89.0     | 47.2     | 85.0     | 29.0   | 3.5     | <b>2.1</b>  |
| CMU MoBo (100)   | 71.4       | $2.7 \times 10^4$ | 1.8     | 394      | 53.0     | 79.3     | 39.1   | 5.9     | <b>2.5</b>  |
| iLIDS-MA         | 3.9        | 741               | 0.51    | 58.7     | 121      | N/A      | 9.6    | 24.5    | <b>3.3</b>  |
| iLIDS-AA         | 9.9        | 2337              | 1.2     | 150      | 344      | N/A      | 36.8   | 83.4    | <b>7.2</b>  |
| CAVIAR4REID      | <b>3.8</b> | 214               | 0.35    | 55.3     | 249      | N/A      | 15.8   | 30.8    | 8.0         |



# Collaboratively Regularized Nearest Points

- Distance finding optimization

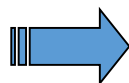
$$\min_{\alpha, \beta} \left\{ \left\| \mathbf{z} - \mathbf{Q}\alpha - \mathbf{X}\beta \right\|_2^2 + \lambda_1 \left\| \alpha \right\|_2^2 + \lambda_2 \left\| \beta \right\|_2^2 \right\},$$

One-step closed-form solution?

**Yes!**

**But,**

-- it is expensive,  
-- the whole optimization is needed for each query/probe set.



Iterative Optimization:

**Fix  $\beta$ ,** and optimize  $\alpha$  :

$$\alpha^* = \mathbf{P}_q (\mathbf{z} - \mathbf{X}\beta), \text{ with } \mathbf{P}_q = (\mathbf{Q}^T \mathbf{Q} + \lambda_1 \mathbf{I})^{-1} \mathbf{Q}^T.$$

**Fix  $\alpha$ ,** and optimize  $\beta$  :

$$\beta^* = \mathbf{P}_x (\mathbf{z} - \mathbf{Q}\alpha), \text{ with } \mathbf{P}_x = (\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I})^{-1} \mathbf{X}^T.$$

# Collaboratively Regularized Nearest Points

- Classification

Like sparse/collaborative representation models for single-instance based recognition, here the set-specific coefficients  $\boldsymbol{\beta}^* = [\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_n^*]$  is implicitly made to have some discrimination power.

Therefore, we design our classification model as follows.

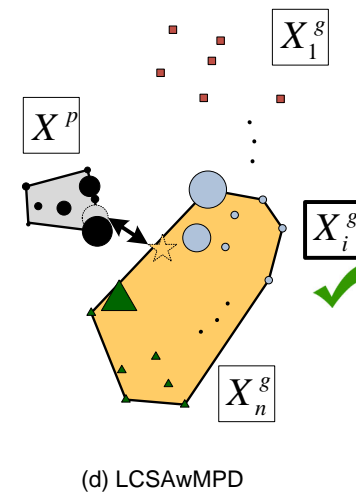
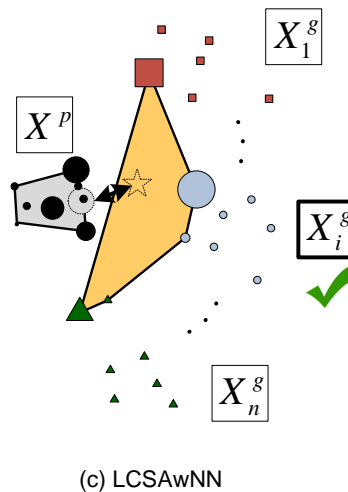
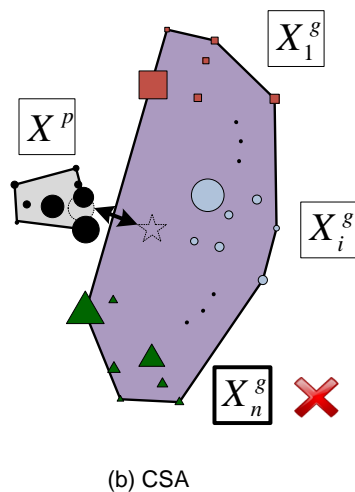
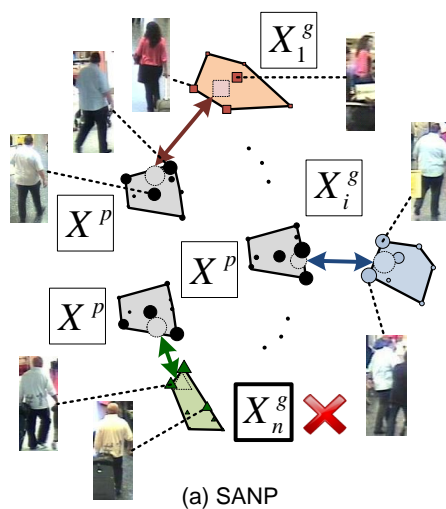
$$C(\mathbf{Q}) = \arg \min_i \{d_{CRNP}^i\},$$

where 
$$d_{CRNP}^i = \left( \|\mathbf{Q}\|_* + \|\mathbf{X}_i\|_* \right) \cdot \left\| \mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}_i^* \right\|_2^2 / \left\| \boldsymbol{\beta}_i^* \right\|_2^2.$$

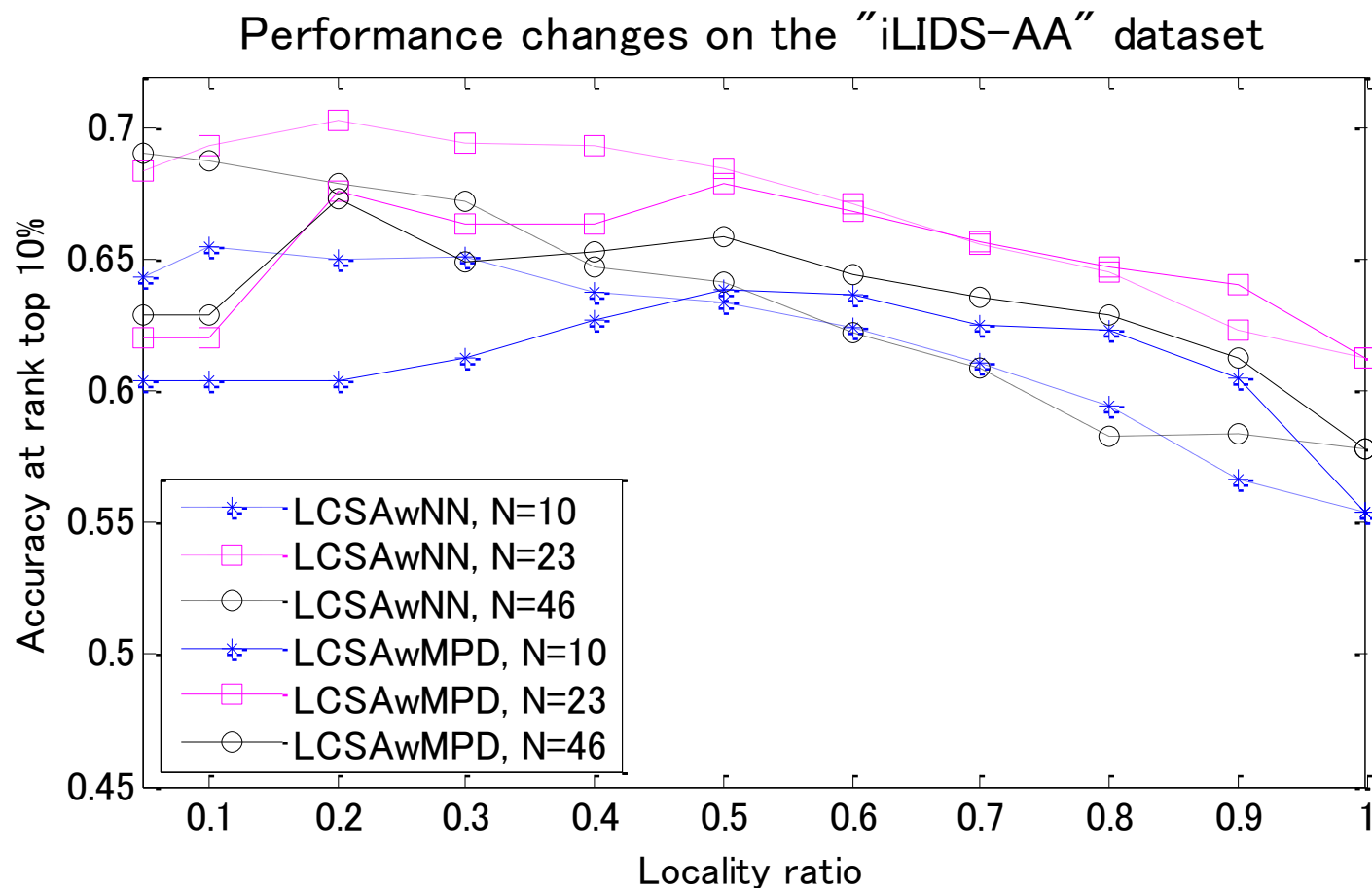
Recall that RNP doesn't directly use the coefficients themselves which are actually also discriminative.

$$d_{RNP}^i = \left( \|\mathbf{Q}\|_* + \|\mathbf{X}_i\|_* \right) \cdot \left\| \mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}^* \right\|_2^2,$$

# LCSA (Locality-constrained Collaborative Sparse Approximation)

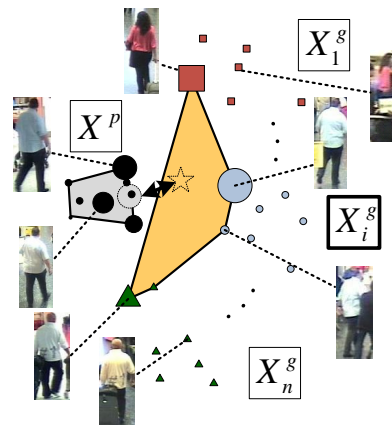


# Experimental Results

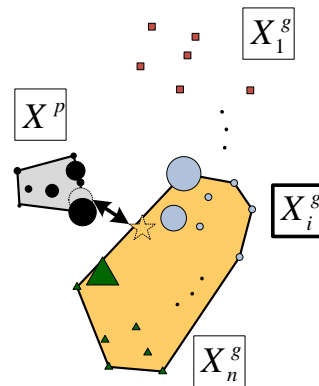


## Collaborative Representation for Re-ID → Non-sparse CR

### LCRNP (Locality-constrained Collaboratively Regularized Nearest Points)

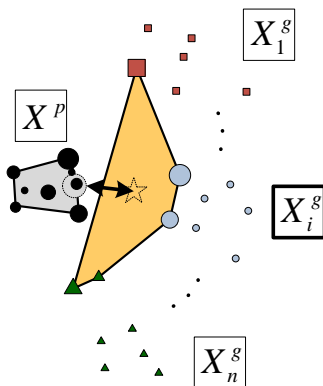


(a) LCSaWNN

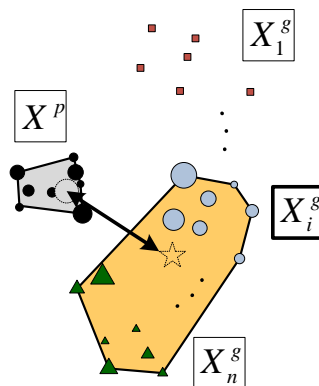


(b) LCSaWMPD

Sparse



(c) LCRNPwNN

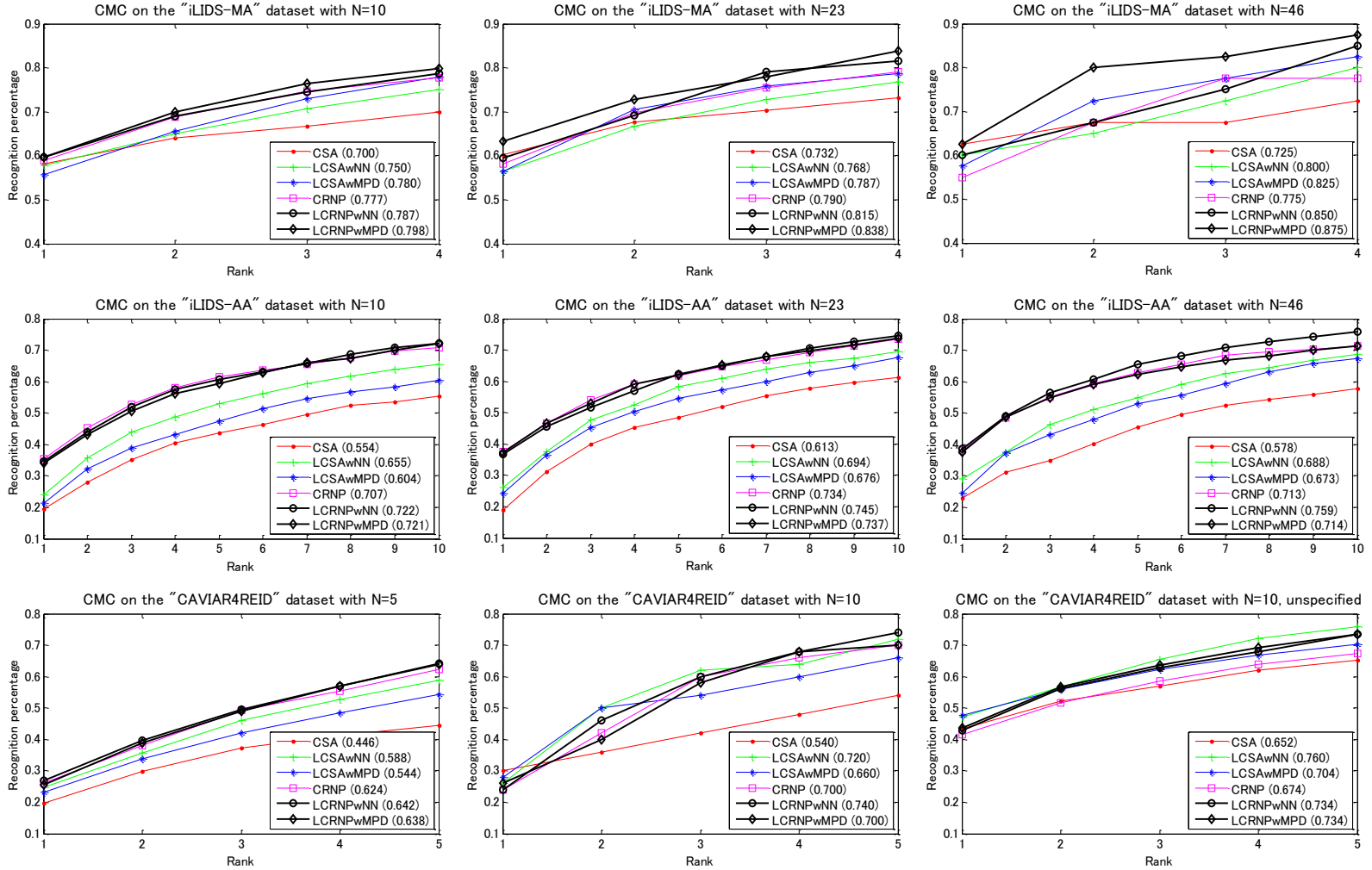


(d) LCRNPwMPD

Non-sparse

# Collaborative Representation for Re-ID → Non-sparse CR

## Experimental results for LCRNP , in comparison with the others

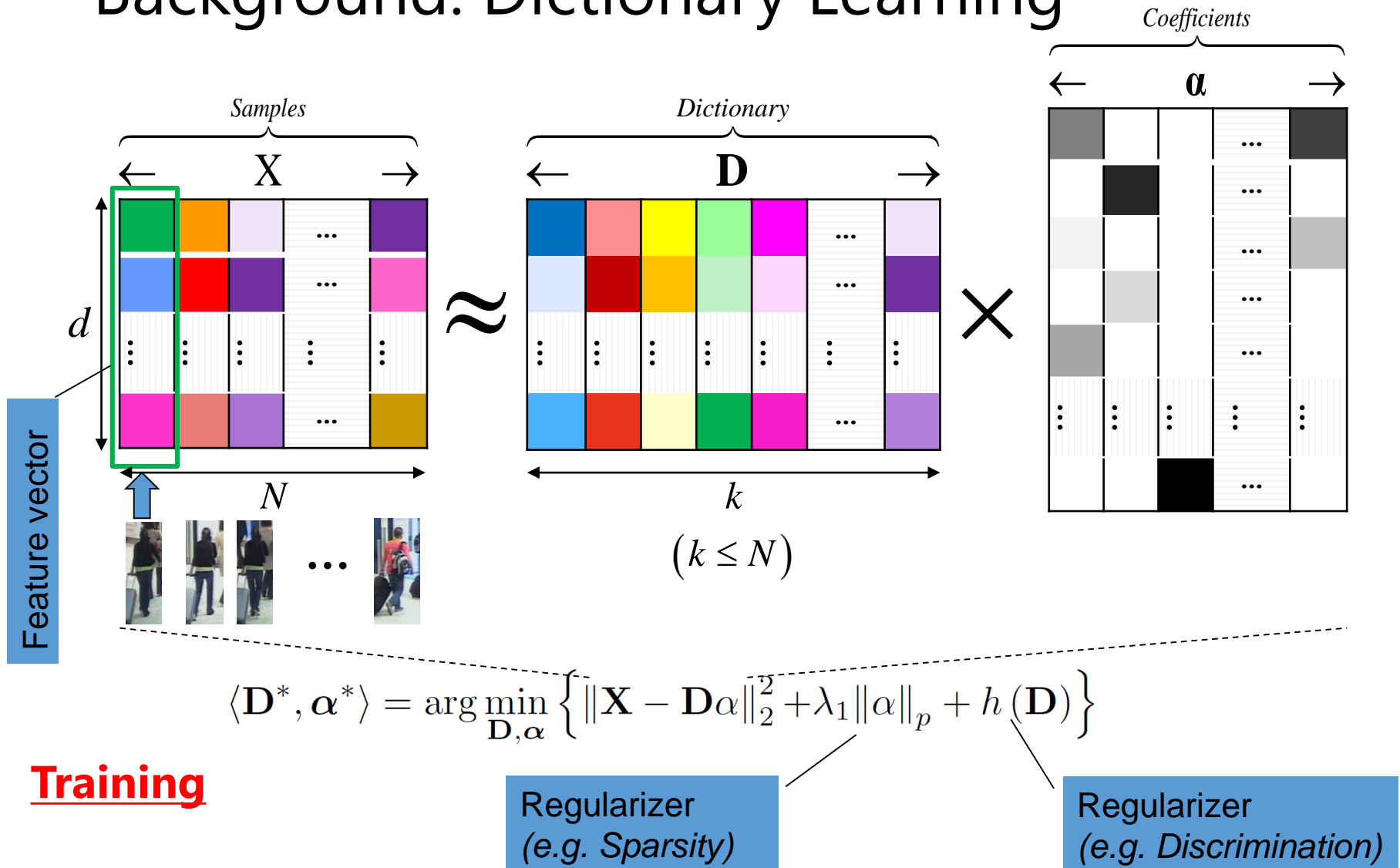


1. **Parametric** (Set-to-set distance + metric learning)
2. **Non-parametric** (Collaborative representation)

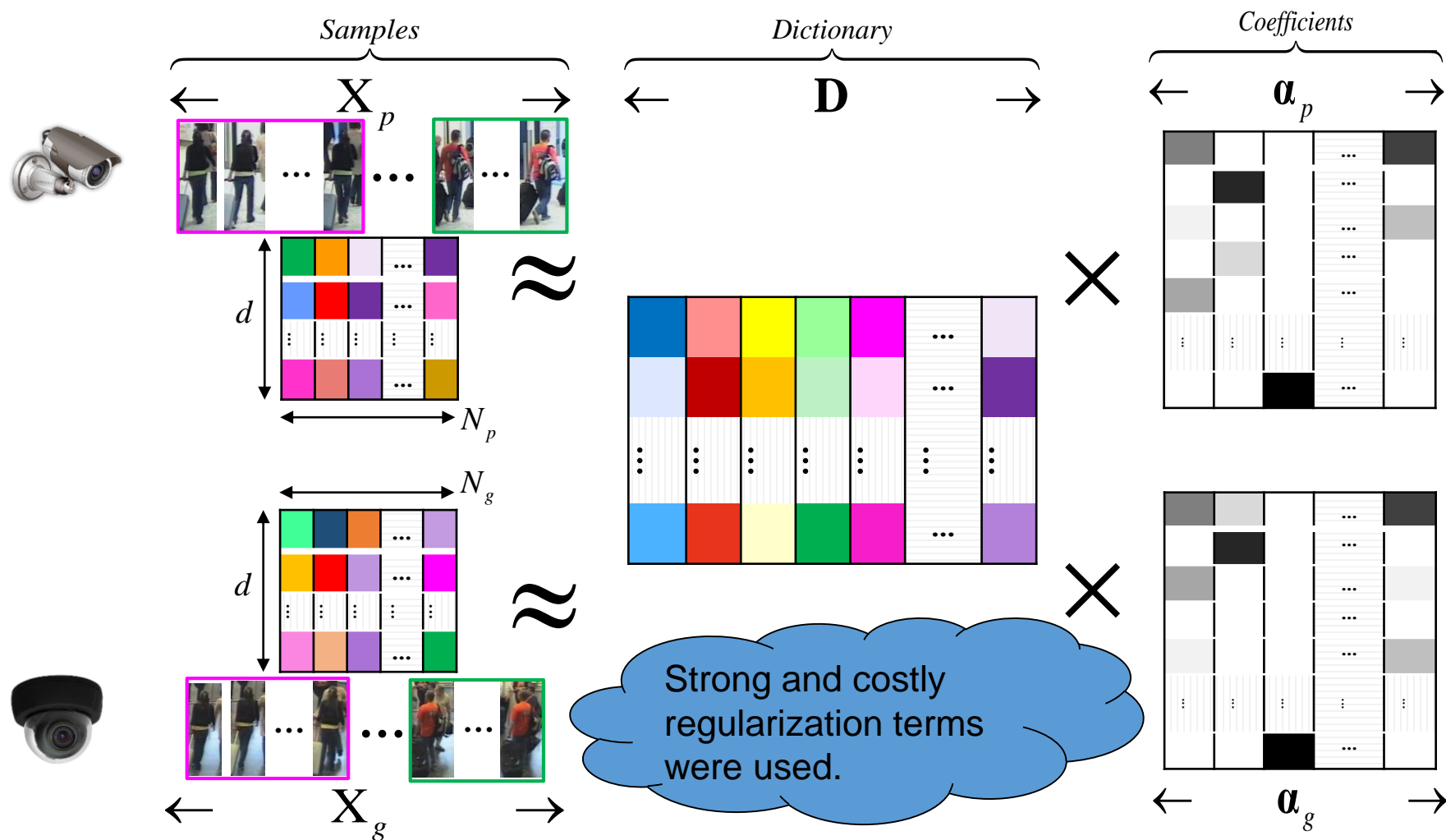
**How about combining them?**



# Background: Dictionary Learning

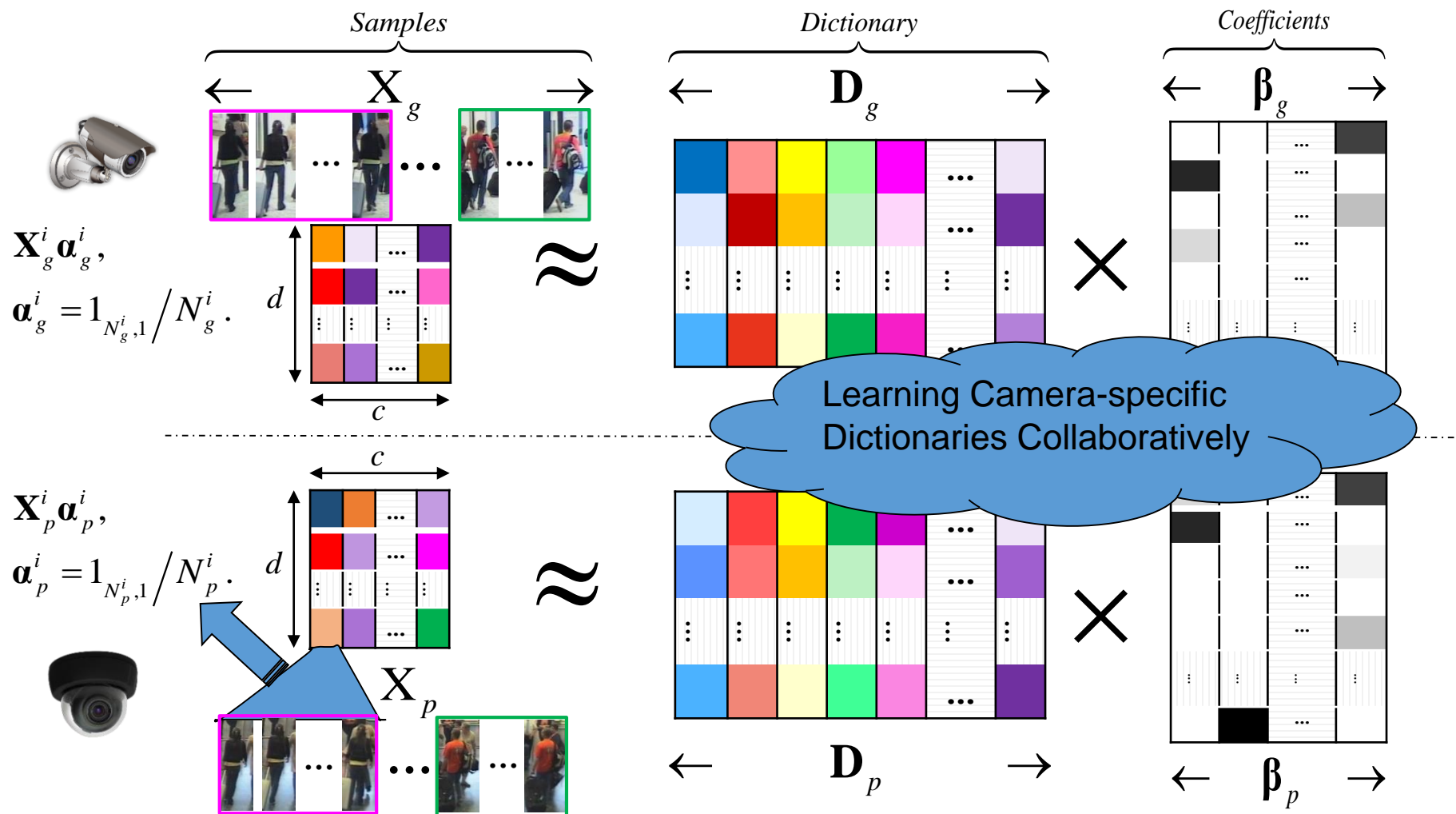


# Discriminative Collaborative Representation (DCR)



## New proposal: dictionary co-learning

# Dictionary Collaborative Learning (DCL)



# Experimental results: Effectiveness

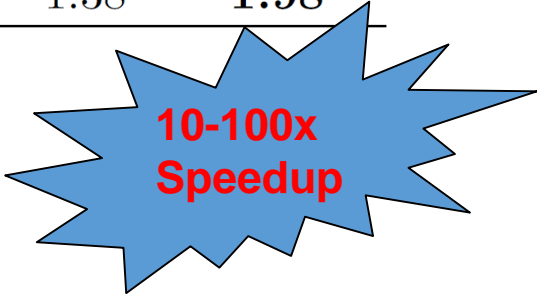
*Rank 1 accuracy*

|               | Experiment | iLIDS-MA    |             |             | CAVIAR4REID |             |                | Performance Change |
|---------------|------------|-------------|-------------|-------------|-------------|-------------|----------------|--------------------|
|               |            | $s = 10$    | $s = 23$    | $s = 46$    | $s = 5$     | $s = 10$    | $s = 10$ mixed |                    |
| Nonparametric | MPD[2]     | 52.0        | 56.5        | 55.0        | 25.2        | 28.0        | 48.8           | 74.3%              |
|               | SRC[16]    | <u>60.0</u> | 58.5        | <u>65.0</u> | 28.0        | 28.0        | 48.0           | 71.4%              |
|               | CRC[17]    | 25.5        | 30.5        | 25.0        | 16.8        | 8.0         | 42.8           | 435%               |
|               | CHISD[3]   | 49.0        | 51.5        | 55.0        | <u>31.6</u> | 32.0        | 53.6           | 67.5%              |
|               | SANP[4]    | 48.0        | 48.5        | 45.0        | 30.8        | <u>36.0</u> | 50.0           | 38.9%              |
|               | CSA[6]     | 50.0        | 52.0        | 50.0        | 24.8        | 28.0        | 50.0           | 78.6%              |
|               | RNP[5]     | 50.0        | 50.5        | 55.0        | 28.0        | <u>36.0</u> | 47.6           | <u>32.2%</u>       |
|               | CRNP[1]    | 59.5        | 57.0        | 55.0        | 29.6        | 28.0        | 48.8           | 74.3%              |
|               | CMA[8]     | 55.5        | 52.0        | 50.0        | 31.2        | 32.0        | <u>54.8</u>    | 71.3%              |
|               | LCRNP[7]   | 54.0        | 55.5        | 60.0        | 30.0        | 32.0        | 48.0           | 50.0%              |
| Parametric    | SBDR[9]    | 49.0        | 51.5        | 55.0        | <u>31.6</u> | 32.0        | 53.6           | 67.5%              |
|               | DCR[10]    | 55.0        | <u>60.0</u> | <u>65.0</u> | 25.2        | 28.0        | 47.2           | 70.7%              |
|               | DCL (ours) | <b>65.5</b> | <b>65.0</b> | <b>70.0</b> | <b>32.8</b> | <b>48.0</b> | <b>60.8</b>    | <b>26.7%</b>       |
|               | Gain       | 9.2%        | 8.3%        | 7.7%        | 3.8%        | 33.3%       | 11.0%          |                    |

# Experimental results: Efficiency

- Running time in *milliseconds/person*, using *matlab* with a normal CPU.

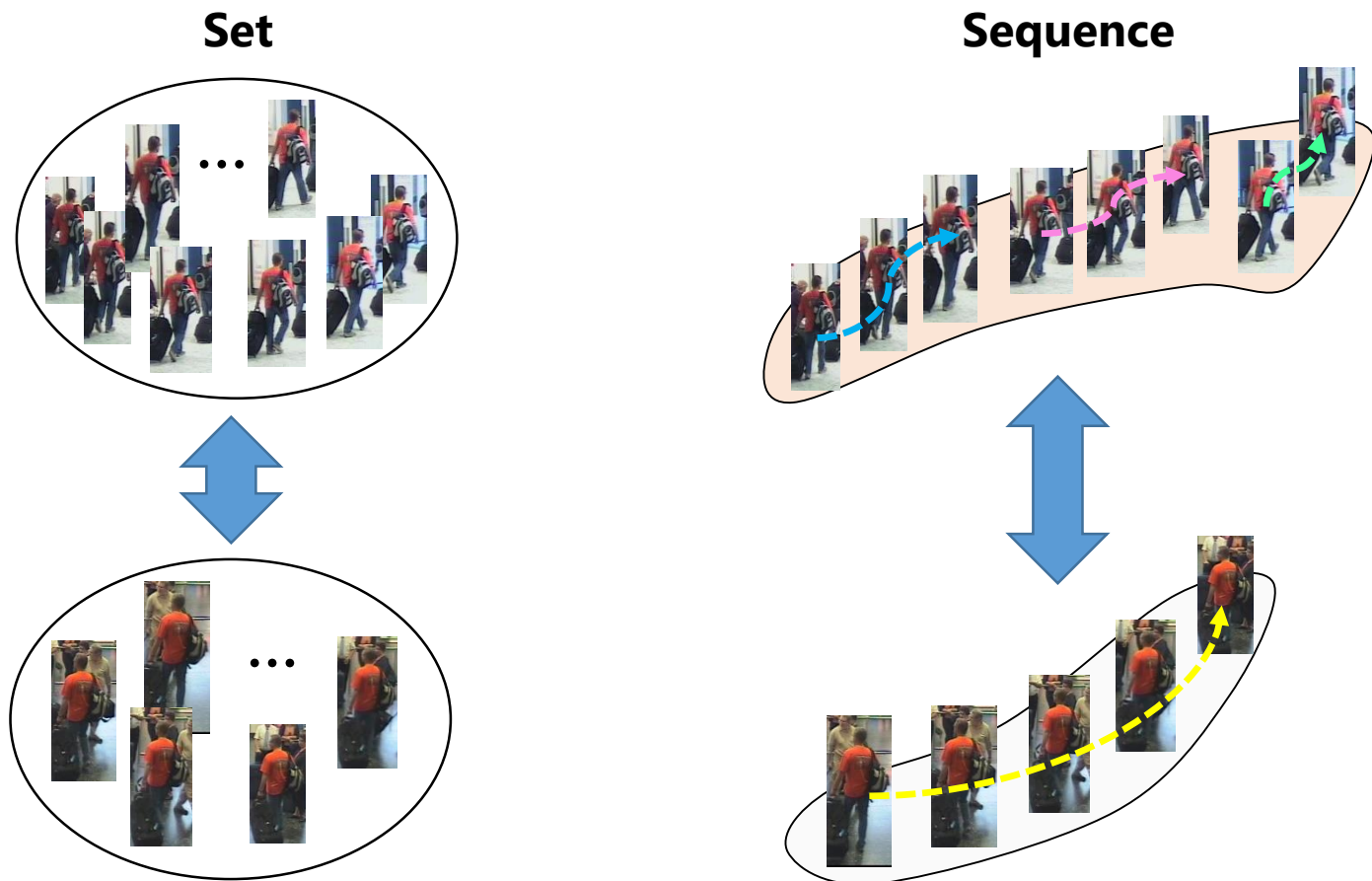
|               | Experiment | iLIDS-MA    |             |             | CAVIAR4REID |             |
|---------------|------------|-------------|-------------|-------------|-------------|-------------|
|               |            | $s = 10$    | $s = 23$    | $s = 46$    | $s = 5$     | $s = 10$    |
| Nonparametric | CRNP[1]    | <b>1.44</b> | 2.80        | 5.16        | <b>1.12</b> | 2.12        |
|               | CMA[8]     | <b>1.44</b> | 3.72        | 7.56        | 1.30        | 2.18        |
| Parametric    | SBDR[9]    | 76.3        | 116         | 328         | 95.4        | 85.6        |
|               | DCR[10]    | 16.4        | 66.8        | 291         | 10.8        | 22.9        |
|               | DCL (ours) | 1.70        | <b>2.37</b> | <b>2.68</b> | 1.38        | <b>1.98</b> |



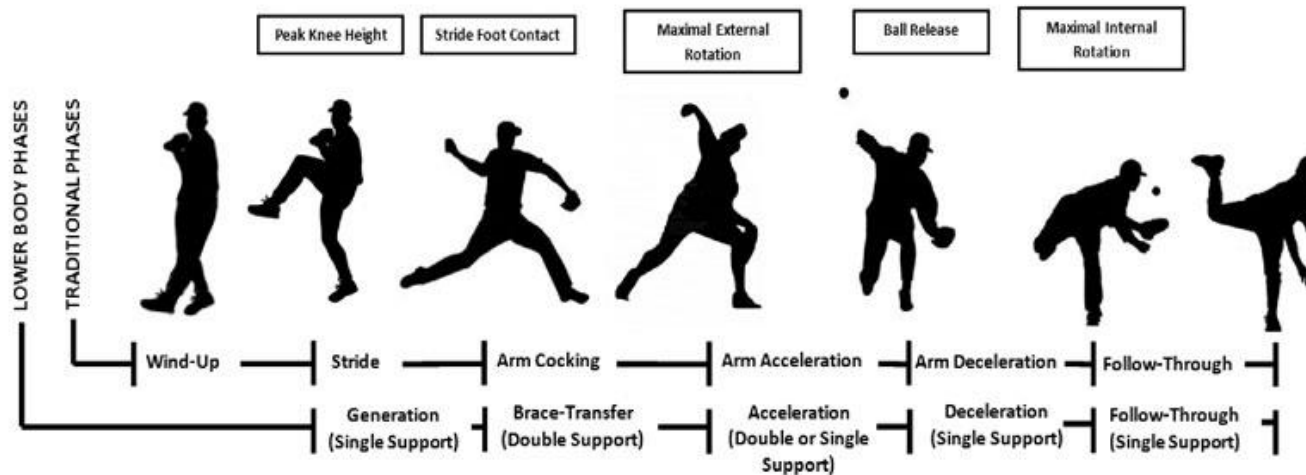
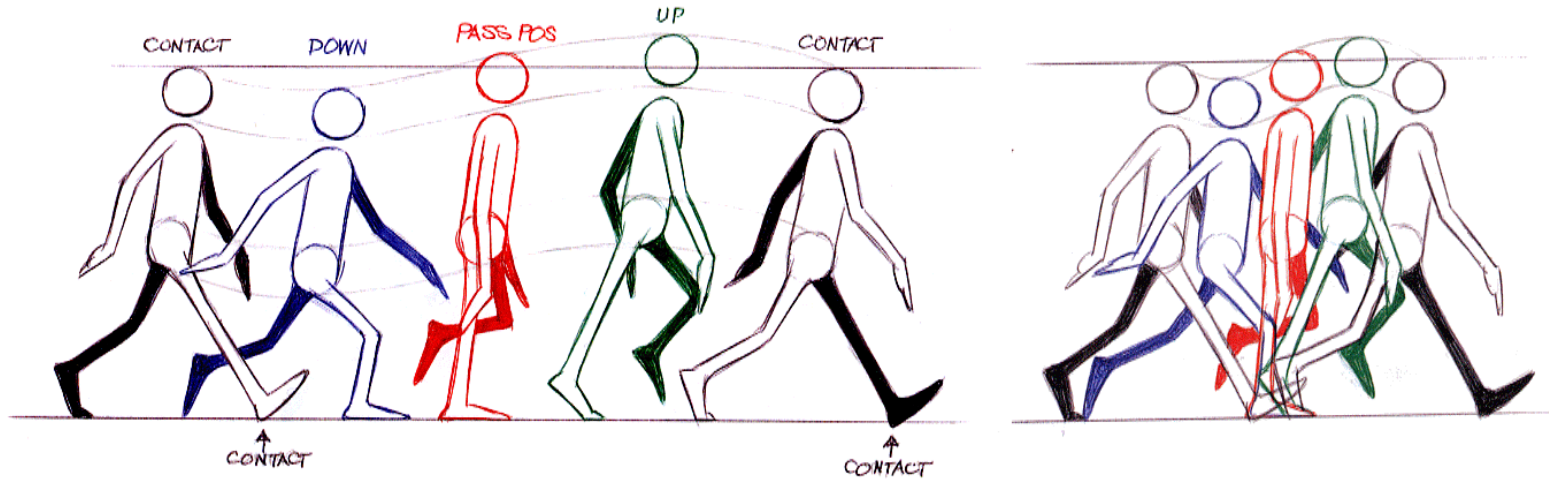
**10-100x  
Speedup**

# Video-based ReID:

## *Perspectives of Set and Sequence*

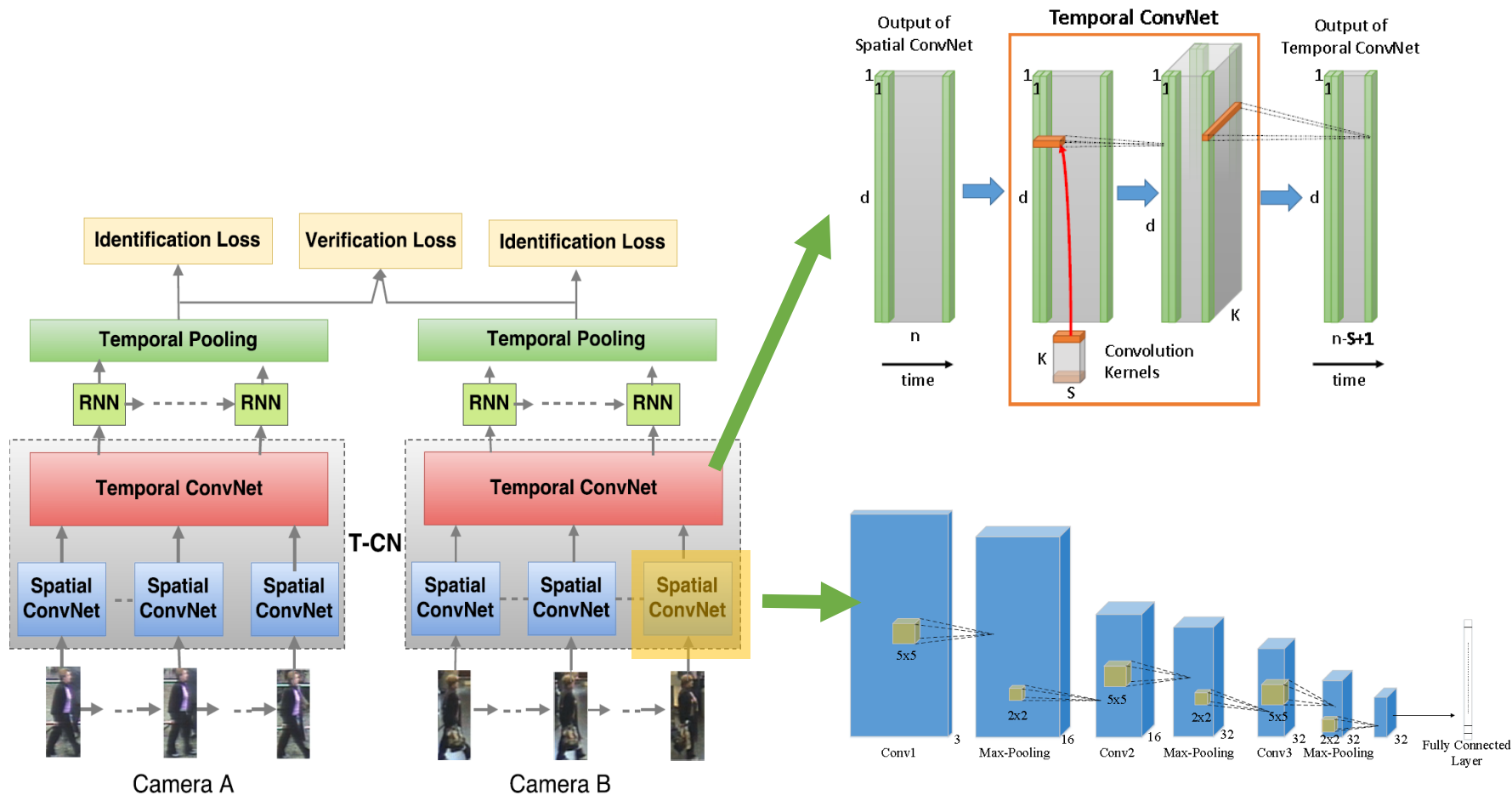


# Sequence: the order matters!





# Proposal: Temporal Convolution



## Identity

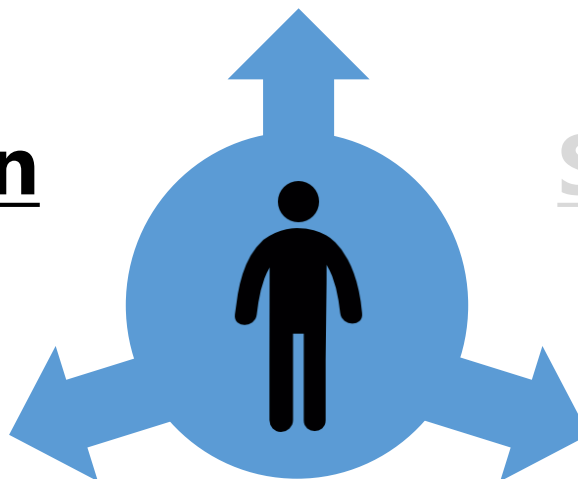
*(Who?)*

## Communication

*(What [does he/she want]?)*

*How [does he/she feel]?)*

**Explicit expression**



## State, Action, ...

*(What [is he/she doing]?)*

*How [does he/she do it]?)*

**Implicit expression**

People communicate to understand each other



What if machines understand them?

Our goal: automatic recognition of spontaneous head gestures



# Targeted head gestures



Nod



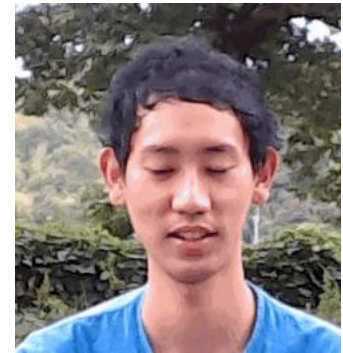
Ticks



Jerk



Up



Down



Tilt



Shake



Turn



Forward

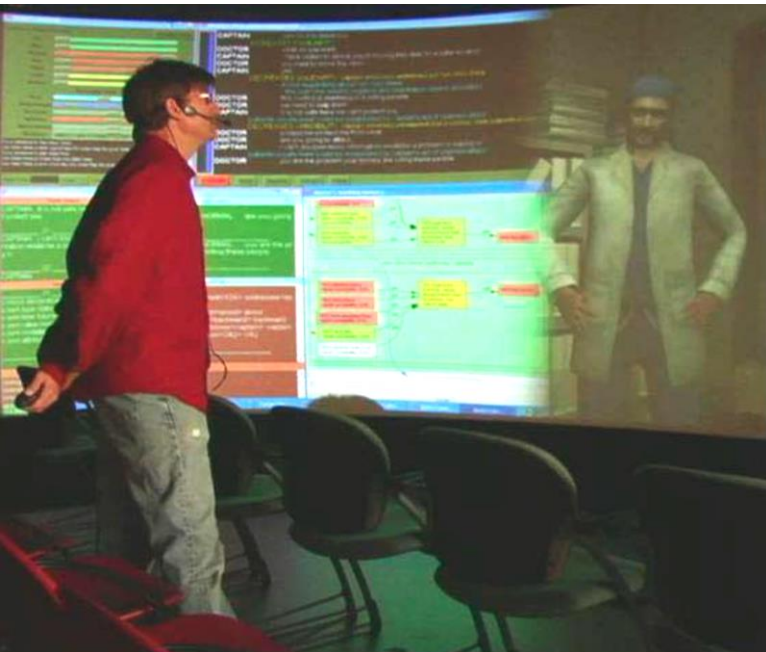


Backward



# Benefits of understanding communication

## Human-robot interaction



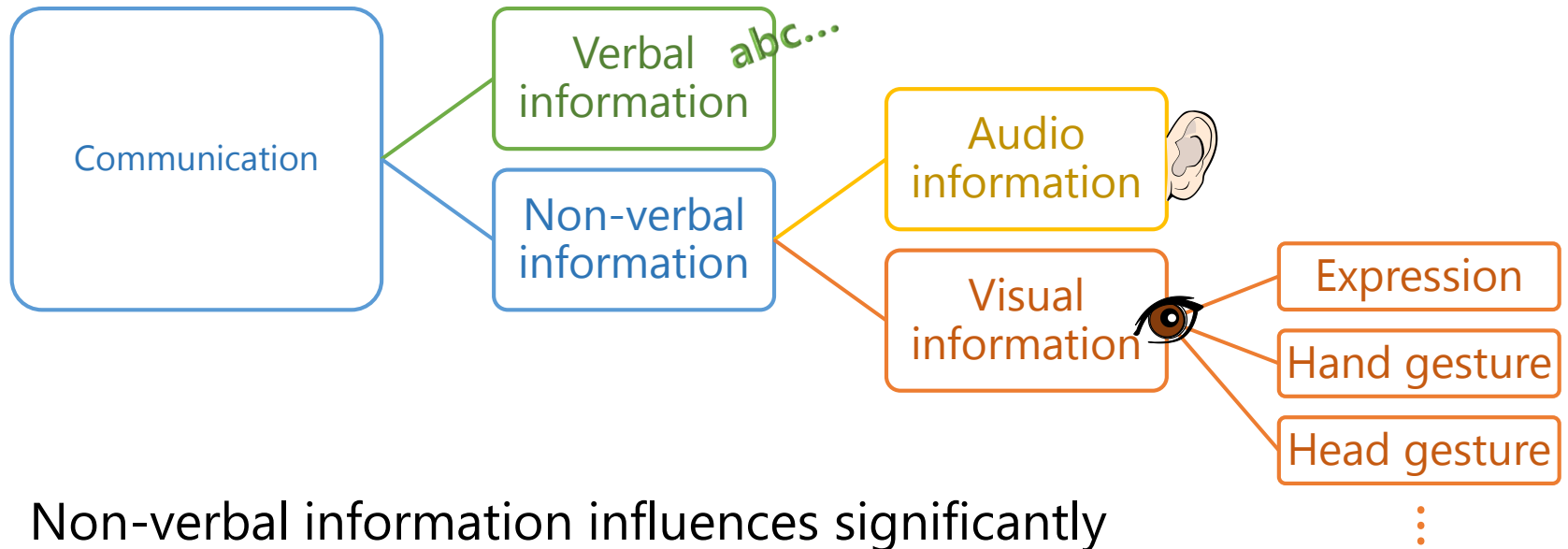
[Maatman *et al.* 2005]

## Communication assistance



[Asakawa 2015]

# Importance of non-verbal information



Non-verbal information influences significantly

e.g.) Mehrabian's rule (Rule of 7%-38%-55%)

Verbal Audio Visual

We focus on **head gesture detection**

- Appears frequently [Hadar *et al.* 1983]
- Takes important role [Kousidis *et al.* 2013, McClave 2000]



# Our contributions and novelties

## □ Contributions

- ✓ Built a **novel dataset**
- ✓ Evaluated **representative automatic recognition models**

## ➤ Novelties (*with comparison to existing work*)

### ✓ **Dataset:**

- closer to real applications
- better for deeper and further researches

### ✓ **Solution:**

- *a general hand-crafted feature*
- *a comparative study of representative recognition algorithms*

# Previous studies on head gesture detection

Recognized  
head gestures

|                  |                                    |  |
|------------------|------------------------------------|--|
| Previous studies | <b>Nod</b>                         | [Morency <i>et al.</i> 2007]<br>[Nakamura <i>et al.</i> 2013]<br>[Chen <i>et al.</i> 2015]   |
|                  | <b>Nod, Shake</b>                  | [Kawato <i>et al.</i> 2000]<br>[Kapoor <i>et al.</i> 2001]<br>[Tan <i>et al.</i> 2003]<br>[Morency <i>et al.</i> 2005]<br>[Wei <i>et al.</i> 2013] |
|                  | <b>Nod, Shake,<br/>Turn</b>        | [Saiga <i>et al.</i> 2010]   |
|                  | <b>Nod, Shake,<br/>Tilt, Still</b> | [Fujie <i>et al.</i> 2004]   |

Only **Nod** and **Shake** are widely handled gestures.

**Nod** is commonly concerned.

# Previous studies on head gesture detection

Recording  
conditions

|                  |                        |  |
|------------------|------------------------|--|
| Previous studies | No interlocutors       | [Kawato <i>et al.</i> 2000]<br>[Kapoor <i>et al.</i> 2001]<br>[Tan <i>et al.</i> 2003]<br>[Wei <i>et al.</i> 2013] |
|                  | Against a robot        | [Fujie <i>et al.</i> 2004]<br>[Morency <i>et al.</i> 2005]<br>[Morency <i>et al.</i> 2007]                         |
|                  | Speaker-listener style | [Nakamura <i>et al.</i> 2013]  |
|                  | Mutual conversations   | [Chen <i>et al.</i> 2015]<br>[Saiga <i>et al.</i> 2010]  |

Few people have worked on spontaneous head gestures in human conversations

# **Dataset Construction**

# Recording

- 30 sequences of approx. 10 min. from 15 participant
- Includes familiar/unfamiliar pairs, indoor/outdoor records
- Conversations with topics chosen beforehand
- Purpose of the recording is announced

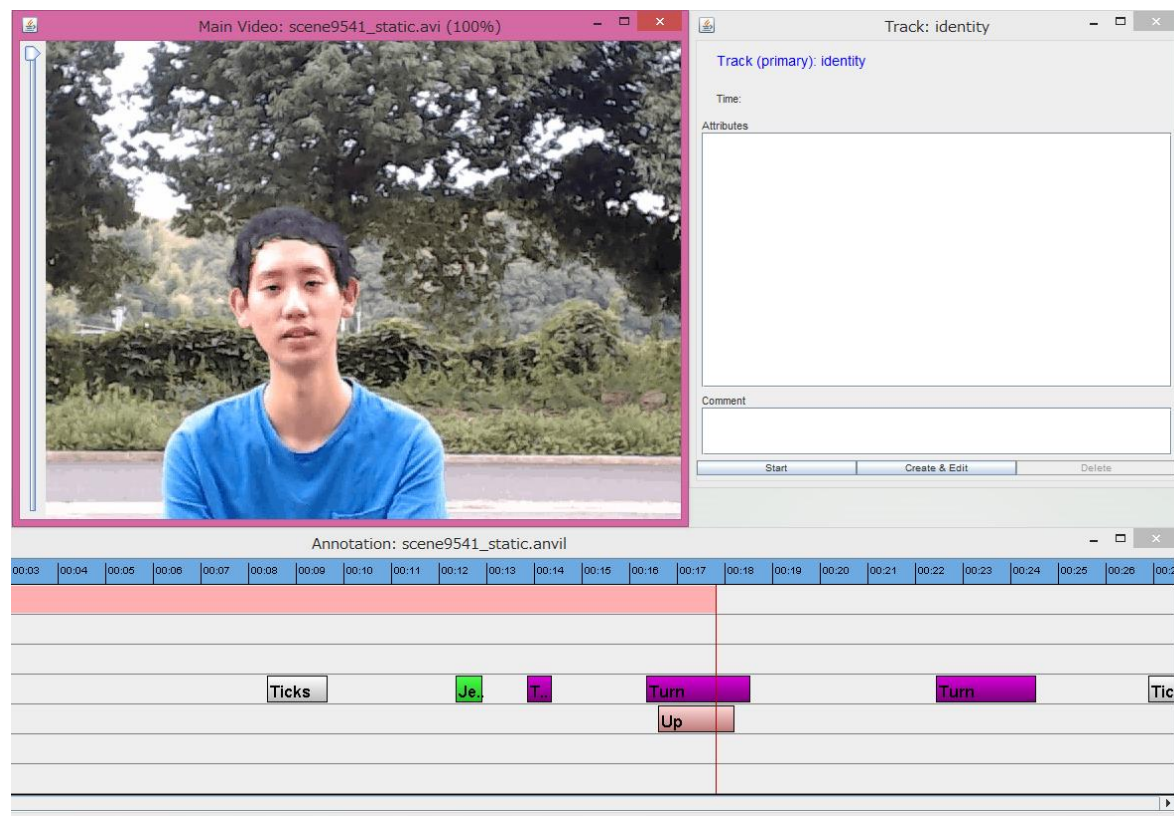


# Annotation

A freeware Anvil5 [Kipp 2014] was used for manual annotation.

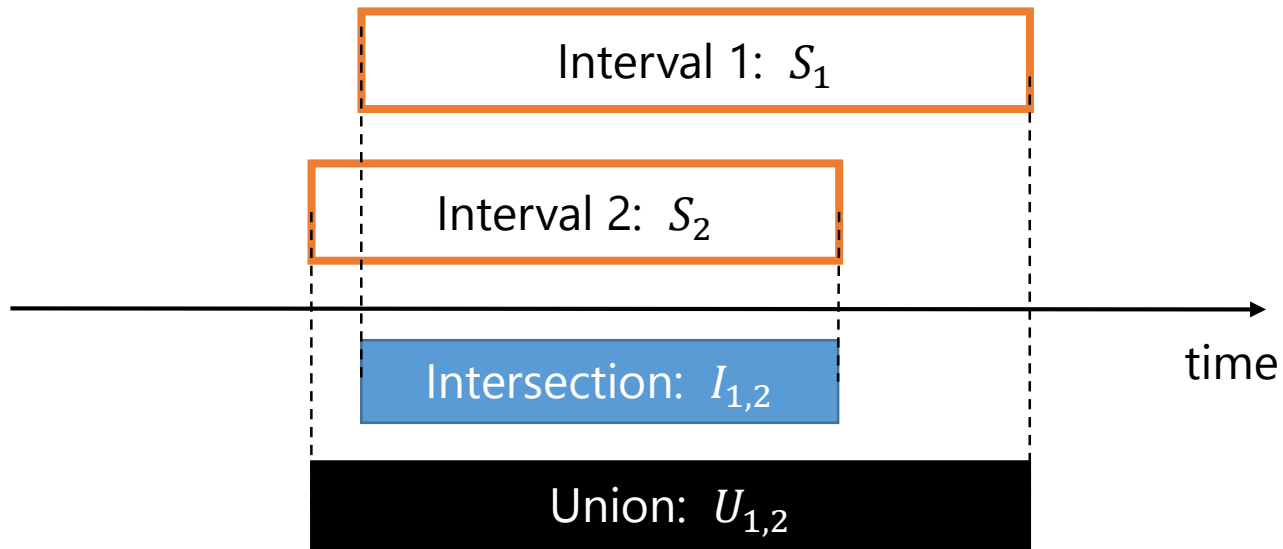
(up to 3 overlapping gestures were allowed)

**3 naive annotators** annotated all the data independently, after a quick training with guideline and examples.



# Ground-Truth Inference

- **IoU**: Interaction over Union

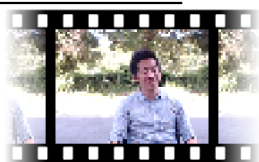


$$IoU(S_1, S_2) = \frac{length(I_{1,2})}{length(U_{1,2})}$$



# Ground-Truth Inference

Annotator A:



Nod, 2

Nod, 3

Shake, 3

Turn, 1

Annotator B:



Nod, 2

Tilt, 1

Shake, 2

Annotator C:



Down, 3

Up, 2

Shake, 3

**Inferred:**



**Nod, 2.5**

**Shake, 3**

Suppose  $\text{IoU}_{\text{th}} = 0.5$

A&B:  $\text{IoU}=0.6 > \text{IoU}_{\text{th}}$

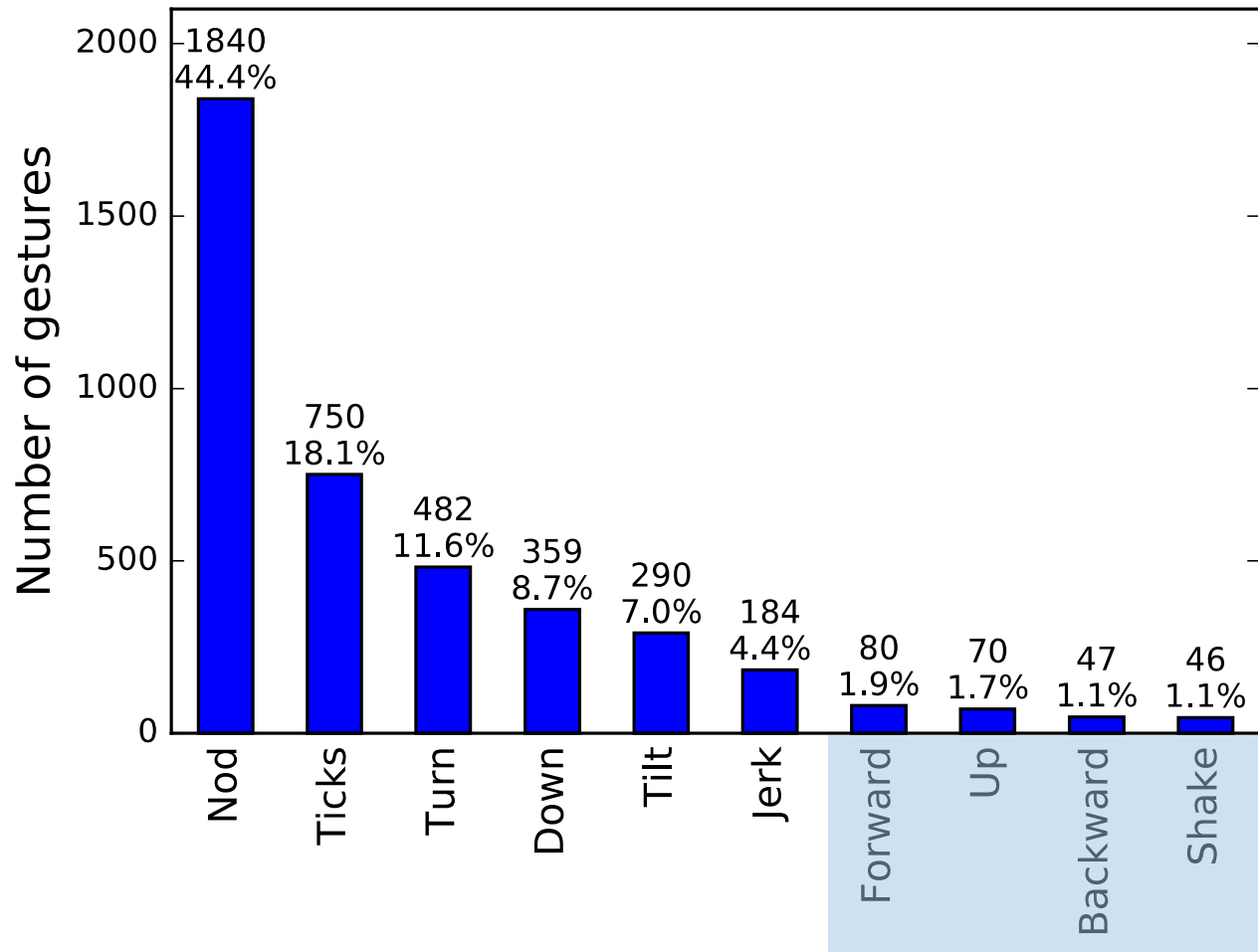
A&B:  $\text{IoU}=0.65 > \text{IoU}_{\text{th}}$

A&C:  $\text{IoU}=0.8 > \text{IoU}_{\text{th}}$

B&C:  $\text{IoU}=0.6 > \text{IoU}_{\text{th}}$

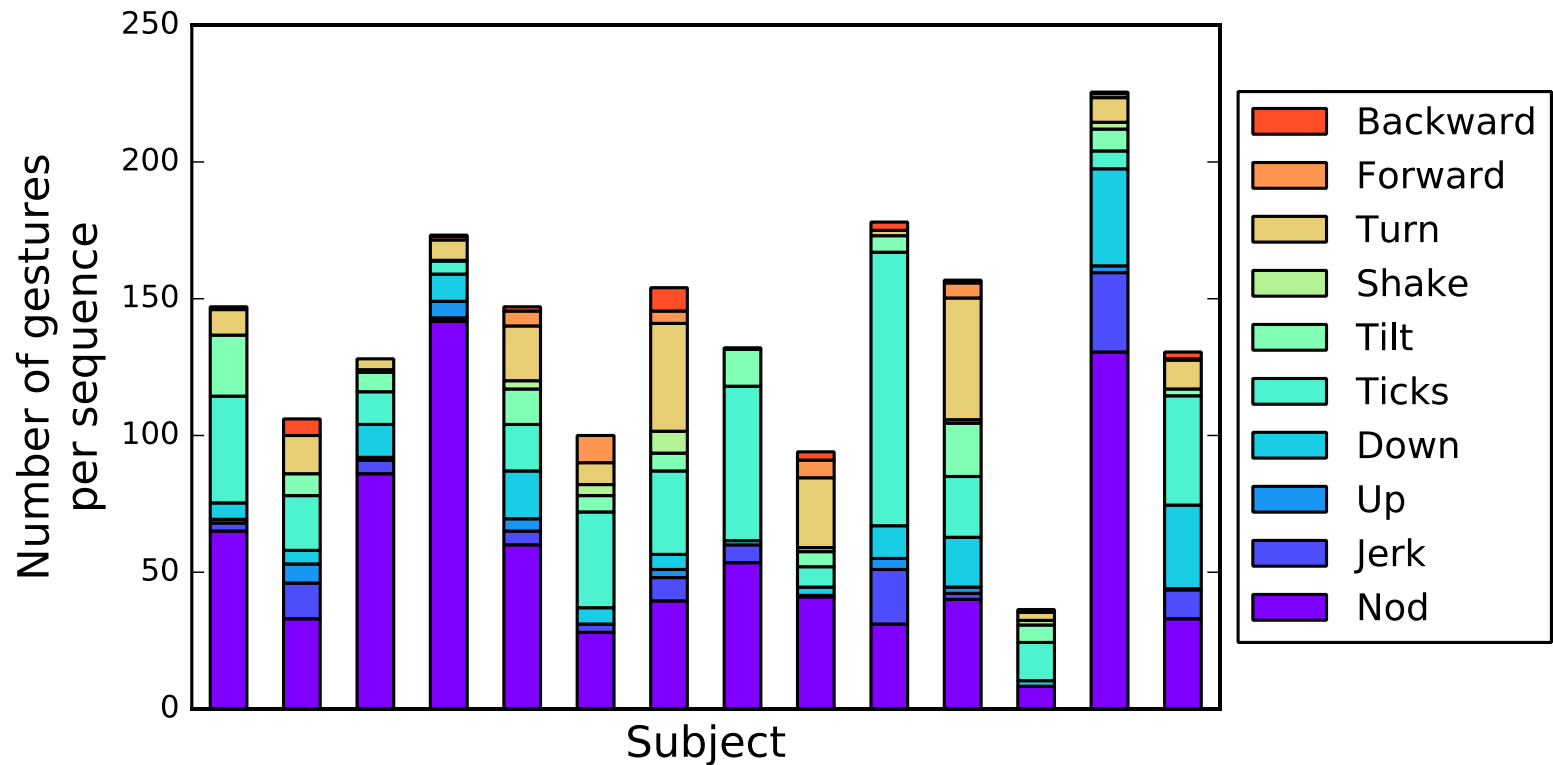
Non-maximum  
Suppression

# Statistics (Inferred Ground-truth with IoU=0.5)

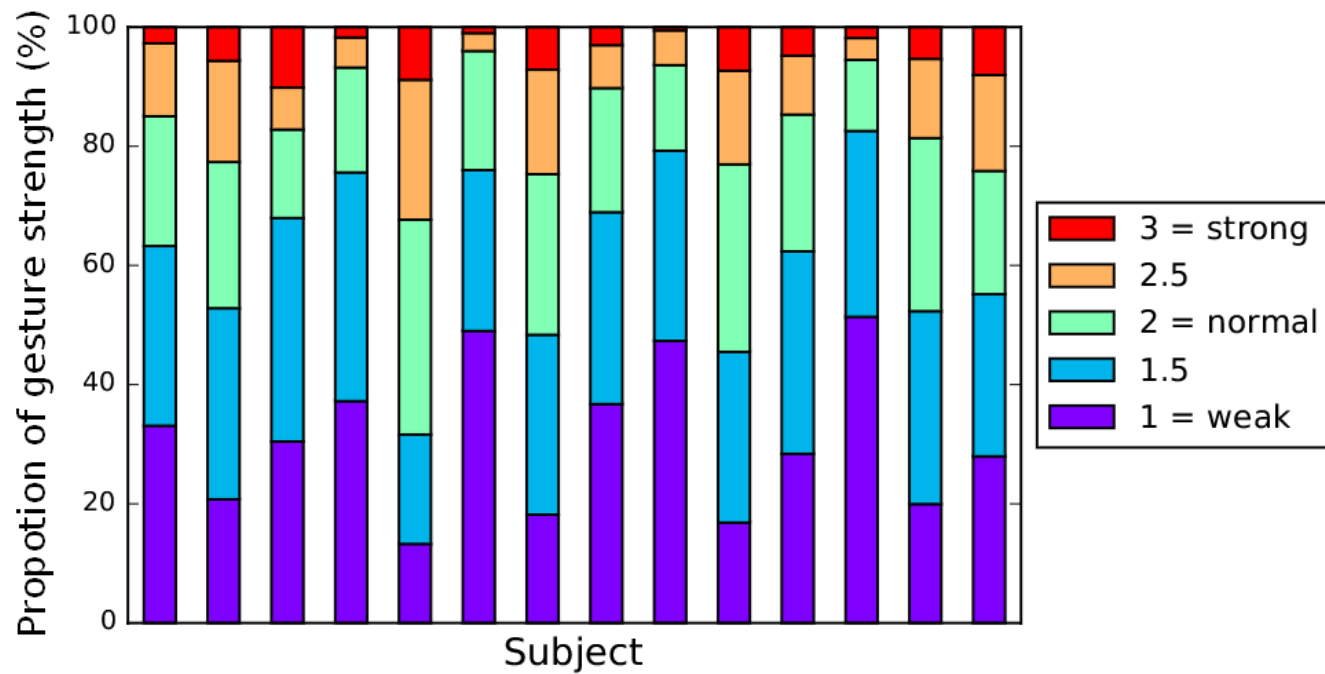


**Total No. of Samples: 4147**

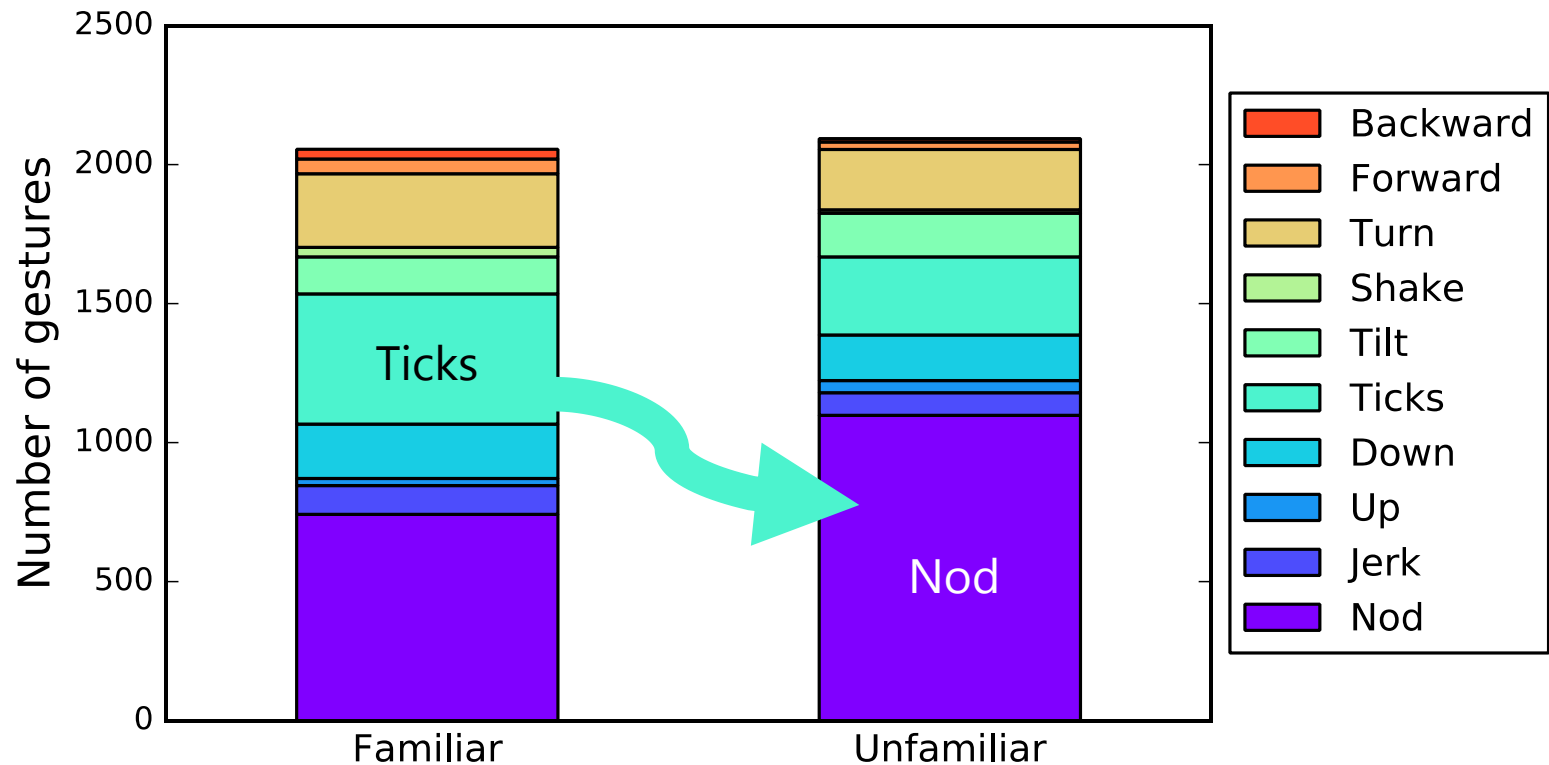
# Type Distribution per Subject



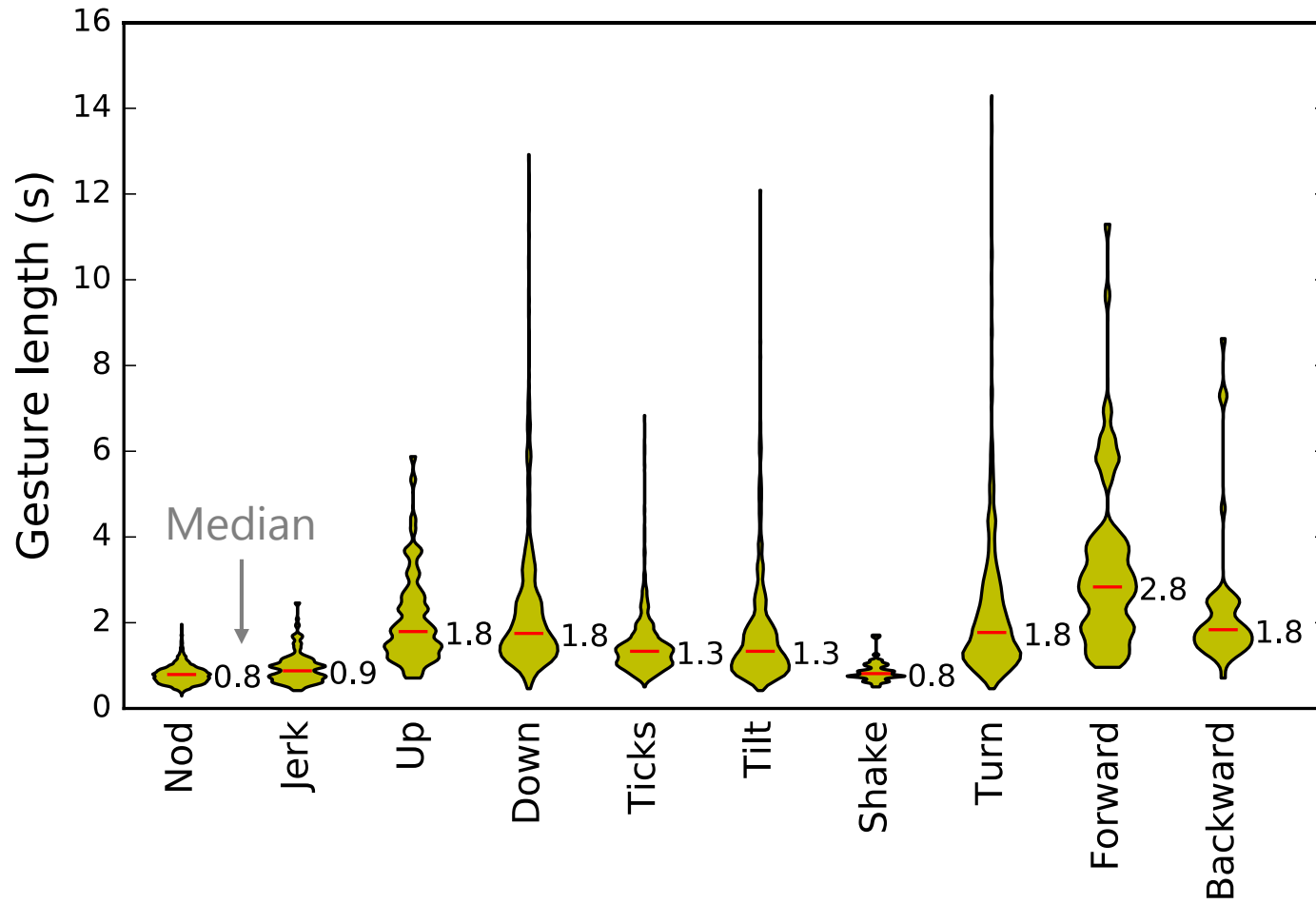
# Strength Distribution per Subject



# Familiar vs. Unfamiliar



# Length Distribution



# Recognition tasks

To **detect** varied head gestures from spontaneous conversations

**Detection** : Given a **sequence**, to infer **when** and **which** gestures appear.



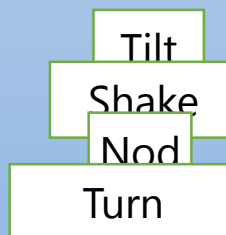
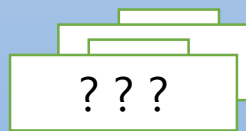
Nod

Nod

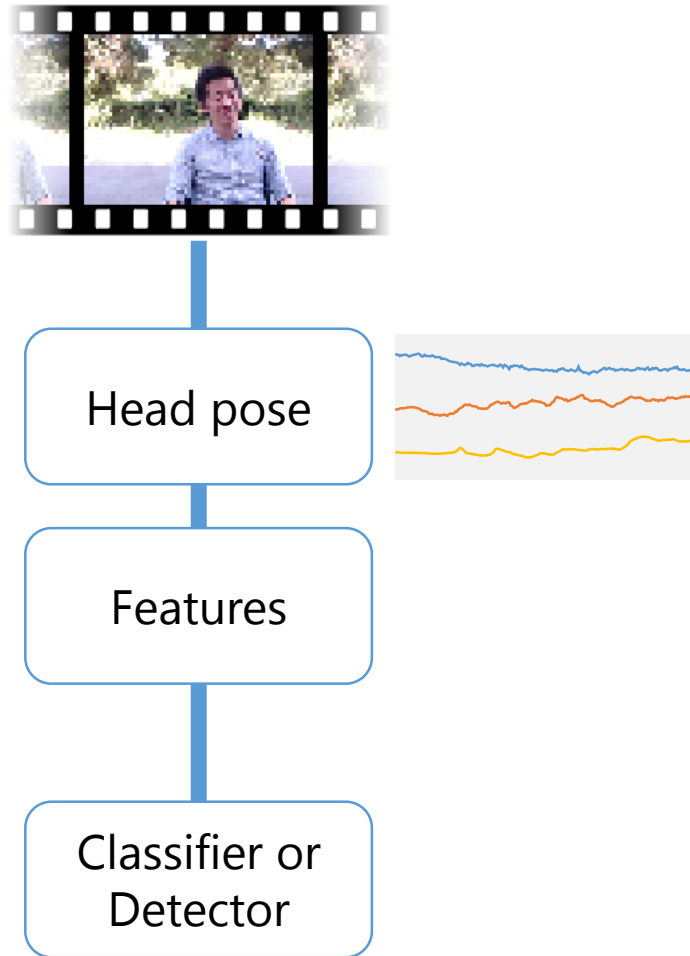
Shake

To understand the problem better, we also work on the task of

**Classification** : Given a **segmented gesture clip**, to infer **which** type it belongs to.



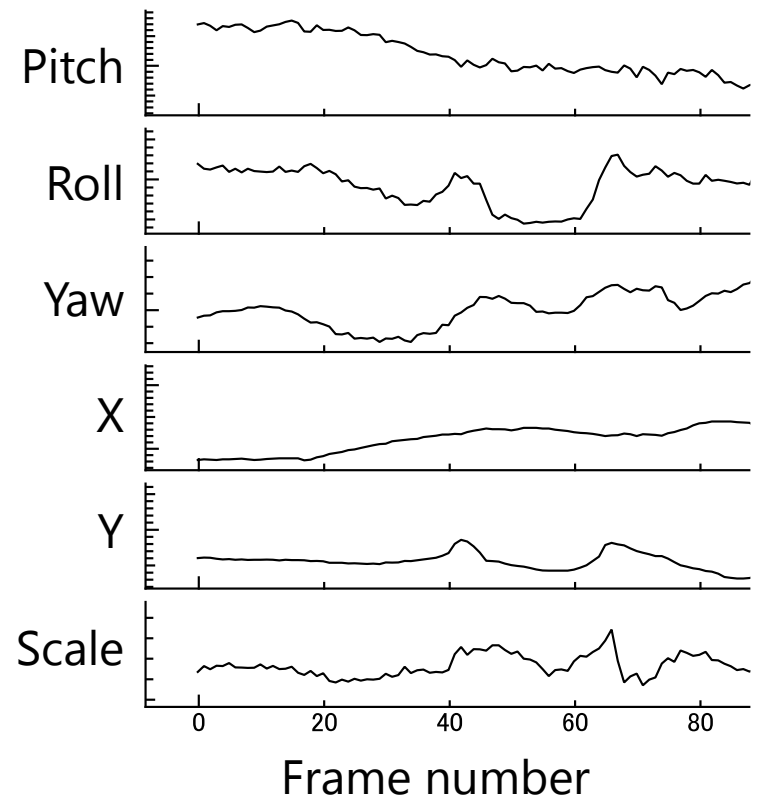
# General framework





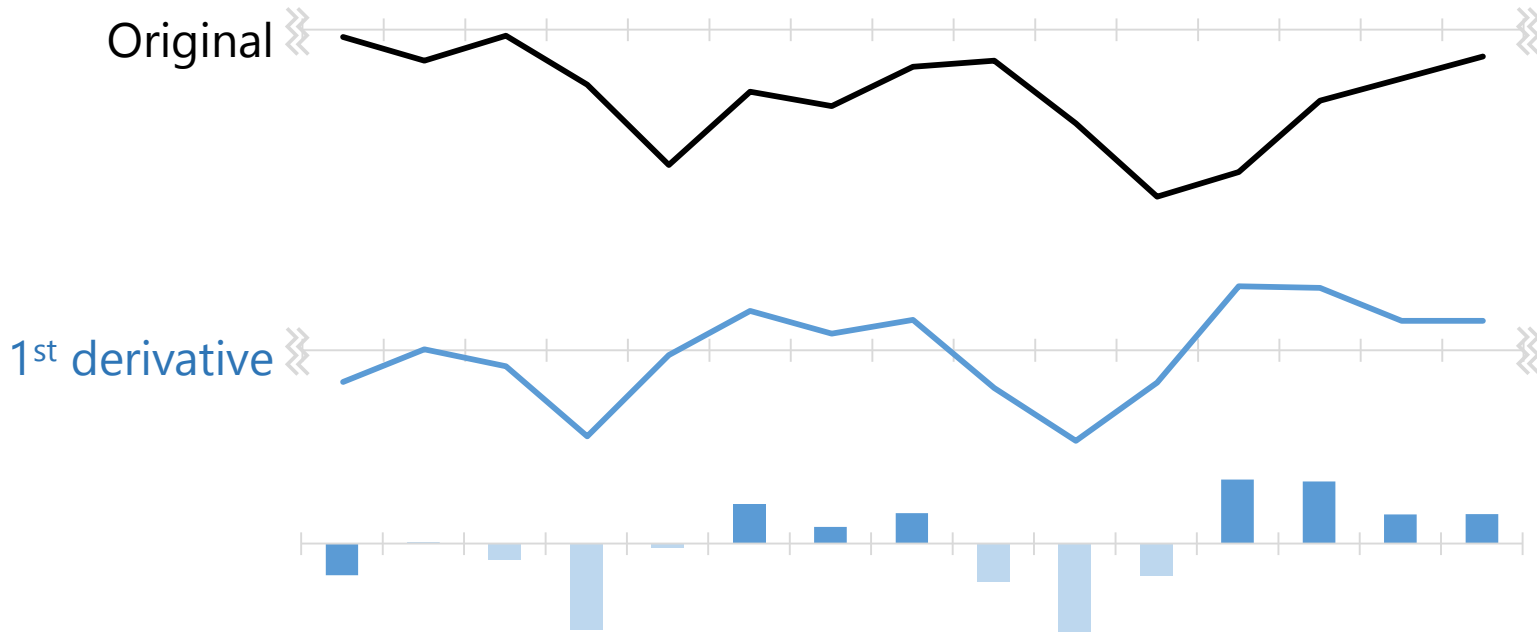
# Head pose estimation

Head pose (and position) were estimated with ZFace [Jeni *et al.* 2015]

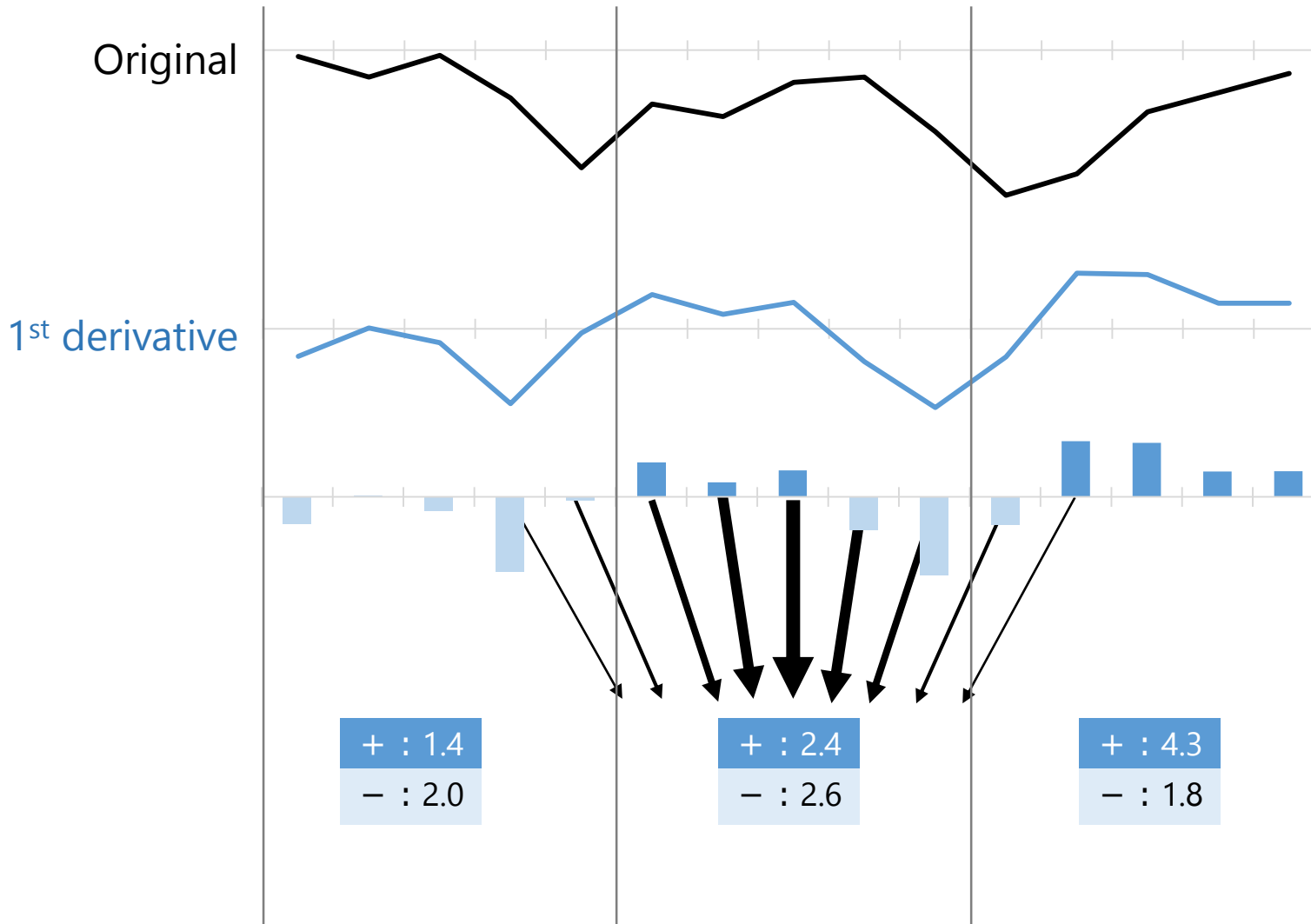


# A general hand-crafted feature

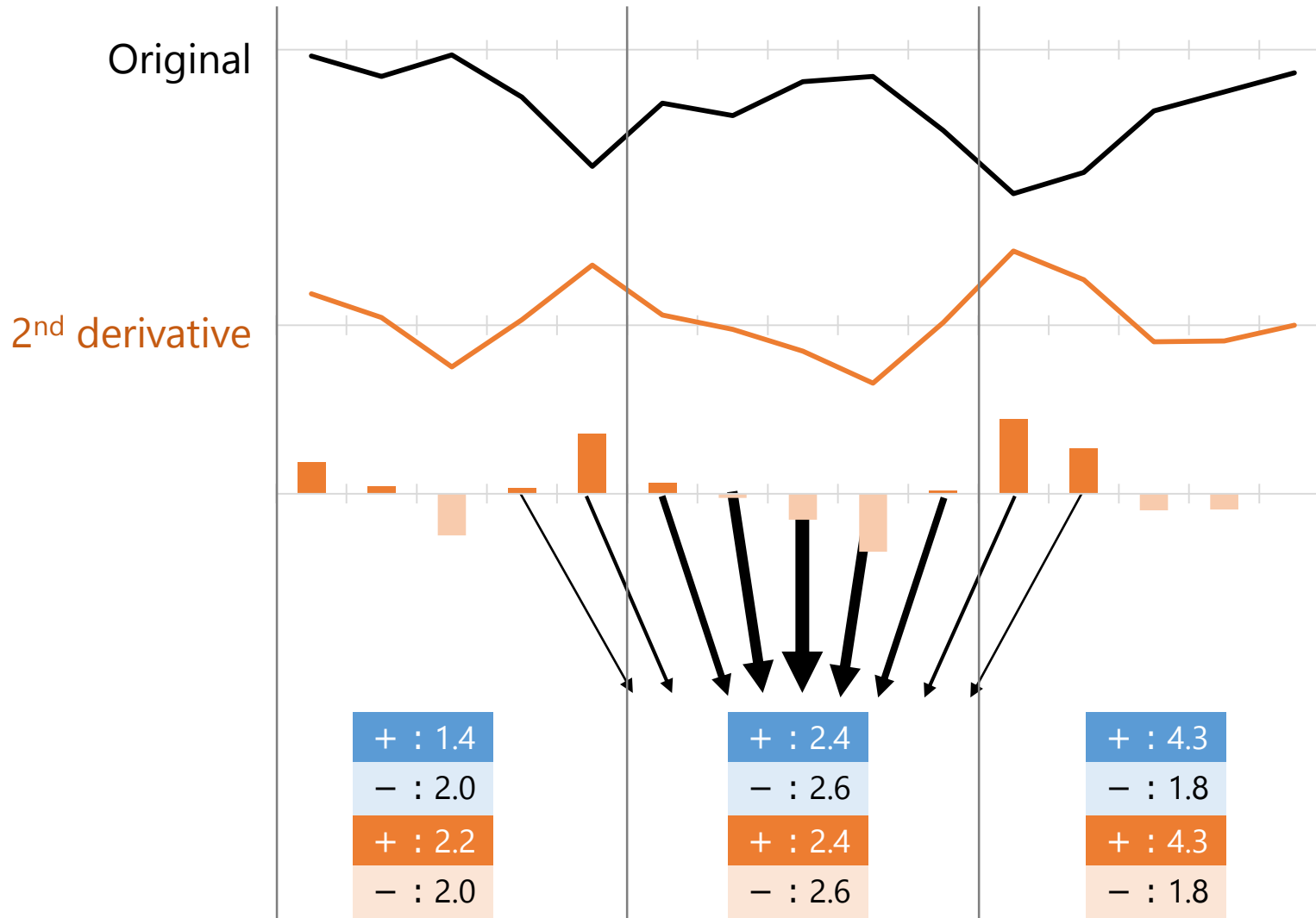
## Histogram of Velocity and Acceleration (HoVA)



# Histogram of Velocity and Acceleration (HoVA)



# Histogram of Velocity and Acceleration (HoVA)



# Existing classification models

|                  |             |   |
|------------------|-------------|---|
| Learning model   |             |   |
| Previous studies | (rule-base) | [Kawato <i>et al.</i> 2000]<br>[Saiga <i>et al.</i> 2010]<br>[Nakamura <i>et al.</i> 2013]                        |
|                  | SVM         | [Morency <i>et al.</i> 2005]<br>[Chen <i>et al.</i> 2015]   |
|                  | HMM         | [Kapoor <i>et al.</i> 2001]<br>[Tan <i>et al.</i> 2003]<br>[Fujie <i>et al.</i> 2004]<br>[Wei <i>et al.</i> 2013] |
|                  | LDCRF       | [Morency <i>et al.</i> 2007]  |

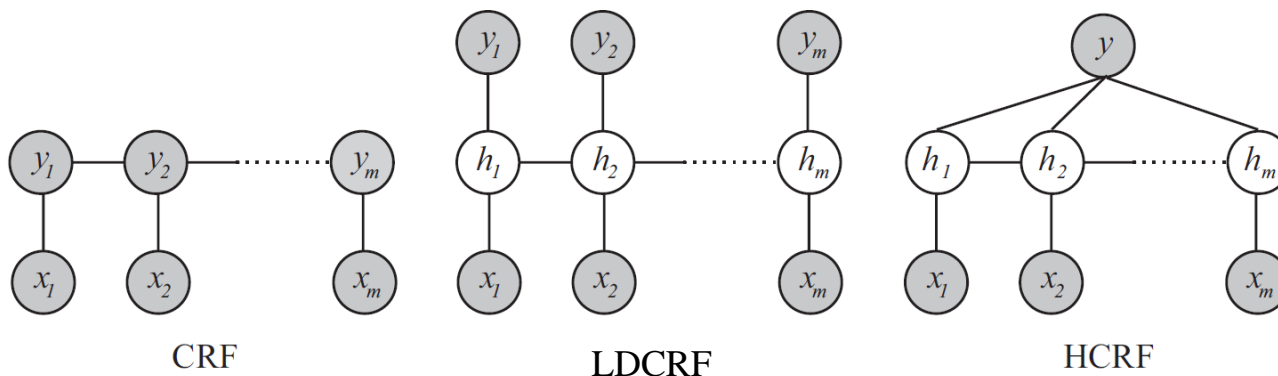
# We evaluate the following models

## ❑ Non-graphical

- SVM

## ❑ Graphical

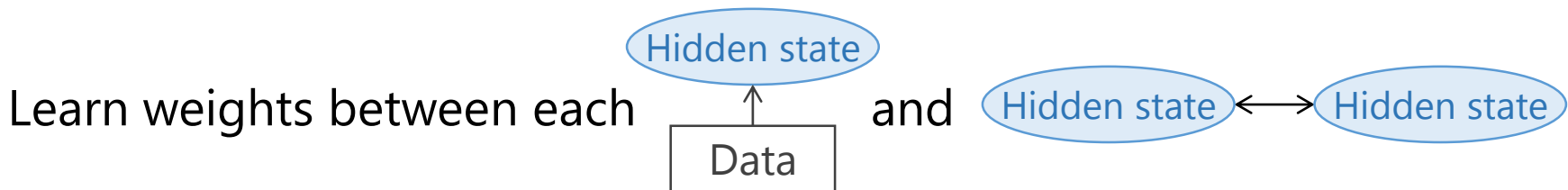
- Hidden-state Conditional Random Field (HCRF) *for classification*
- Latent-Dynamic Conditional Random Field (LDCRF) *for detection*
- Long-Short Term Memory (LSTM)



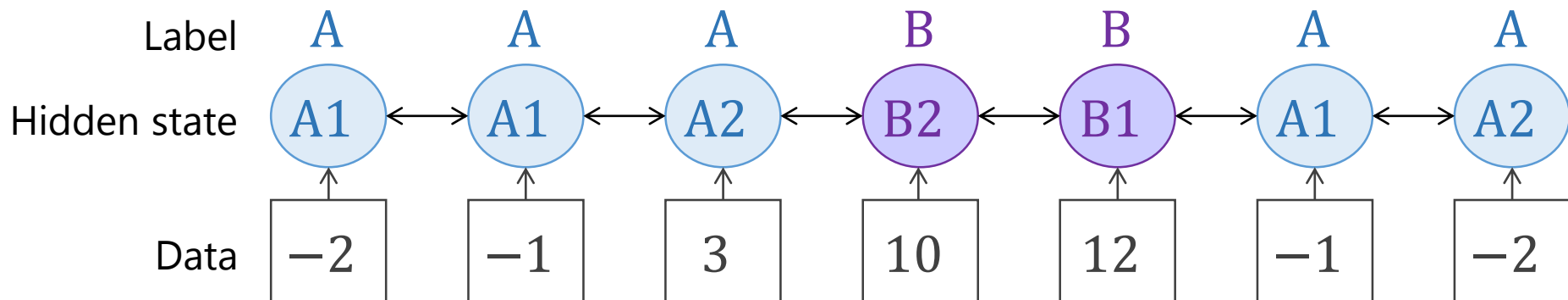
# LDCRF (Latent-Dynamic Conditional Random Field)

[Morency et al. 2007]

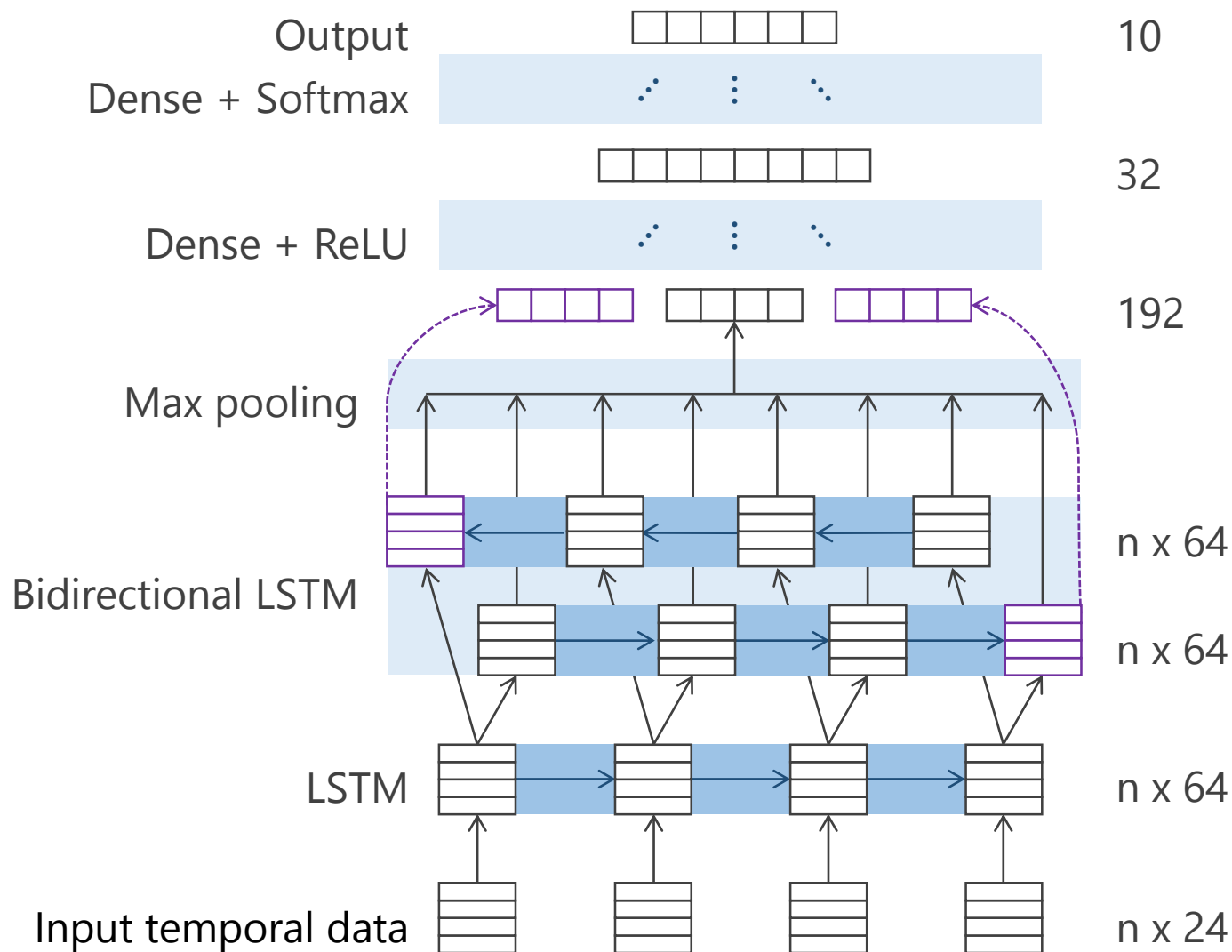
Conditional Random Field enhanced for action detection



Optimize the order of hidden states throughout temporal data



# LSTMs (Long-Short Term Memory)





# Results – Classification (Accuracy, F-score)

- **Accuracy** (Averaged)

| Method       | Training Set | Training-Val Set | Validation Set | Test Set  |
|--------------|--------------|------------------|----------------|-----------|
| SVM          | 0.68±0.02    | 0.74±0.04        | 0.62±0.11      | 0.60±0.12 |
| SVM_weighted | 0.65±0.02    | 0.76±0.01        | 0.59±0.11      | 0.57±0.13 |
| HCRF         | 0.88±0.04    | 0.83±0.03        | 0.66±0.14      | 0.64±0.10 |
| LSTMs        | 0.79±0.02    | 0.84±0.06        | 0.63±0.14      | 0.61±0.15 |

- **F-score** (Averaged)

| Method       | Training Set | Training-Val Set | Validation Set | Test Set |
|--------------|--------------|------------------|----------------|----------|
| SVM          | 0.483        | 0.318            | 0.387          | 0.307    |
| SVM_weighted | 0.493        | 0.324            | 0.408          | 0.388    |
| HCRF         | 0.799        | 0.386            | 0.433          | 0.382    |
| LSTMs        | 0.600        | 0.394            | 0.386          | 0.391    |

# Results – Classification (Confusion Matrix)

- Test set only, overall accumulation

## SVM

| Ground Truth | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |
|--------------|------|------|------|------|-------|------|-------|------|---------|----------|
|              | .900 | .005 | .001 | .040 | .043  | .003 | .000  | .008 | .000    | .000     |
|              | .177 | .242 | .016 | .048 | .468  | .000 | .000  | .032 | .000    | .016     |
|              | .171 | .200 | .086 | .029 | .257  | .029 | .000  | .229 | .000    | .000     |
|              | .212 | .000 | .000 | .561 | .106  | .032 | .000  | .090 | .000    | .000     |
|              | .237 | .022 | .006 | .047 | .623  | .008 | .000  | .053 | .000    | .003     |
|              | .277 | .020 | .014 | .142 | .216  | .142 | .000  | .169 | .014    | .007     |
|              | .333 | .056 | .000 | .278 | .222  | .056 | .000  | .056 | .000    | .000     |
|              | .230 | .017 | .004 | .119 | .243  | .021 | .000  | .353 | .000    | .013     |
|              | .419 | .000 | .000 | .032 | .323  | .097 | .000  | .097 | .032    | .000     |
|              | .087 | .000 | .130 | .043 | .478  | .000 | .000  | .217 | .000    | .043     |
| Prediction   |      |      |      |      |       |      |       |      |         |          |

## SVM\_weighted

| Ground Truth | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |
|--------------|------|------|------|------|-------|------|-------|------|---------|----------|
|              | .828 | .011 | .005 | .057 | .032  | .011 | .017  | .009 | .014    | .016     |
|              | .097 | .613 | .081 | .048 | .097  | .000 | .000  | .016 | .016    | .032     |
|              | .086 | .171 | .486 | .029 | .029  | .000 | .000  | .086 | .029    | .086     |
|              | .106 | .000 | .011 | .598 | .053  | .048 | .074  | .058 | .042    | .011     |
|              | .126 | .103 | .020 | .056 | .469  | .050 | .045  | .050 | .045    | .036     |
|              | .135 | .054 | .027 | .169 | .054  | .196 | .061  | .162 | .088    | .054     |
|              | .167 | .222 | .000 | .278 | .000  | .056 | .222  | .056 | .000    | .000     |
|              | .089 | .081 | .064 | .128 | .102  | .055 | .051  | .345 | .047    | .038     |
|              | .226 | .065 | .000 | .097 | .097  | .097 | .065  | .065 | .290    | .000     |
|              | .043 | .087 | .087 | .000 | .000  | .043 | .000  | .043 | .174    | .522     |
| Prediction   |      |      |      |      |       |      |       |      |         |          |

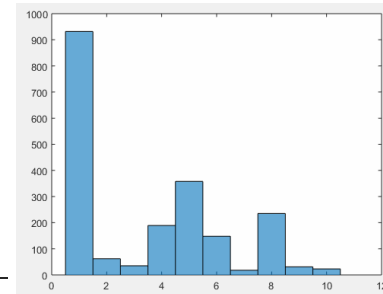
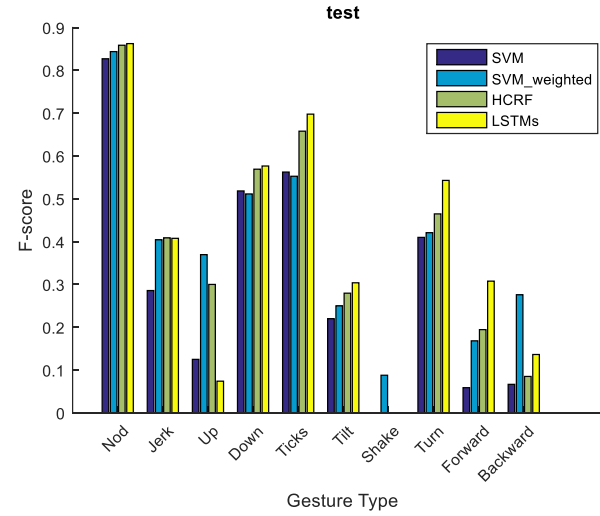
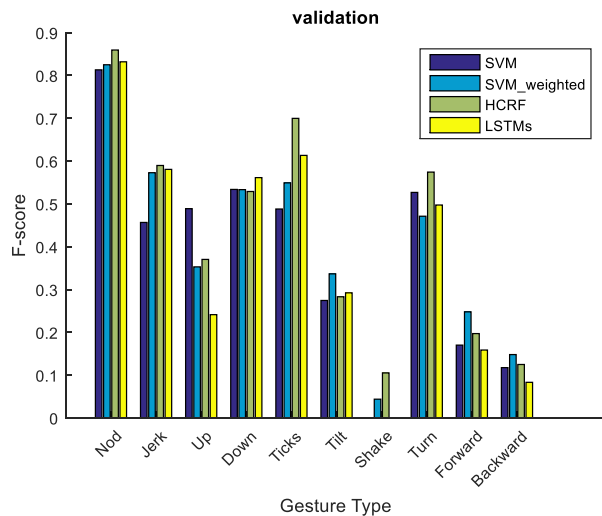
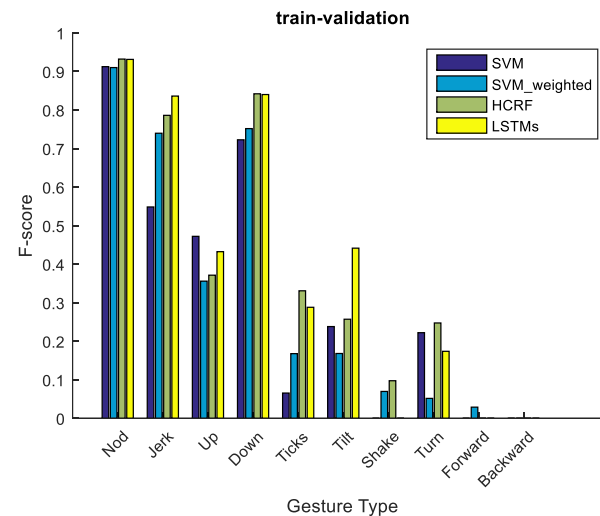
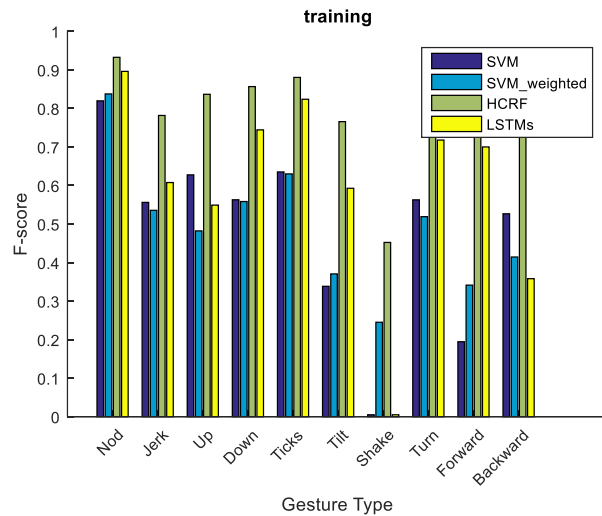
## HCRF

| Ground Truth | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |
|--------------|------|------|------|------|-------|------|-------|------|---------|----------|
|              | .889 | .015 | .000 | .010 | .050  | .013 | .006  | .012 | .004    | .000     |
|              | .371 | .452 | .000 | .000 | .048  | .032 | .016  | .032 | .016    | .032     |
|              | .000 | .200 | .257 | .029 | .114  | .000 | .000  | .229 | .029    | .143     |
|              | .132 | .005 | .000 | .566 | .037  | .079 | .005  | .132 | .037    | .005     |
|              | .170 | .025 | .014 | .036 | .628  | .047 | .014  | .039 | .017    | .008     |
|              | .216 | .014 | .020 | .135 | .061  | .264 | .027  | .203 | .047    | .014     |
|              | .444 | .167 | .000 | .167 | .056  | .000 | .000  | .167 | .000    | .000     |
|              | .064 | .030 | .026 | .111 | .106  | .149 | .009  | .434 | .034    | .038     |
|              | .161 | .000 | .000 | .194 | .000  | .226 | .000  | .194 | .226    | .000     |
|              | .043 | .174 | .087 | .087 | .217  | .174 | .000  | .130 | .000    | .087     |
| Prediction   |      |      |      |      |       |      |       |      |         |          |

## LSTMs

| Ground Truth | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |
|--------------|------|------|------|------|-------|------|-------|------|---------|----------|
|              | .878 | .019 | .000 | .011 | .047  | .040 | .000  | .003 | .002    | .000     |
|              | .355 | .516 | .016 | .000 | .048  | .032 | .000  | .000 | .000    | .032     |
|              | .086 | .229 | .057 | .029 | .029  | .143 | .000  | .114 | .029    | .286     |
|              | .127 | .005 | .011 | .598 | .069  | .058 | .000  | .090 | .037    | .005     |
|              | .134 | .034 | .008 | .045 | .670  | .028 | .000  | .067 | .014    | .000     |
|              | .135 | .088 | .014 | .196 | .061  | .324 | .000  | .162 | .020    | .000     |
|              | .500 | .222 | .000 | .000 | .056  | .222 | .000  | .000 | .000    | .000     |
|              | .064 | .017 | .017 | .123 | .081  | .162 | .000  | .498 | .017    | .021     |
|              | .161 | .000 | .032 | .065 | .000  | .355 | .000  | .065 | .323    | .000     |
|              | .043 | .130 | .174 | .130 | .000  | .087 | .000  | .217 | .087    | .130     |
| Prediction   |      |      |      |      |       |      |       |      |         |          |

# Results – Classification (Class-specific)



# Simulated Human Performance -- Classification

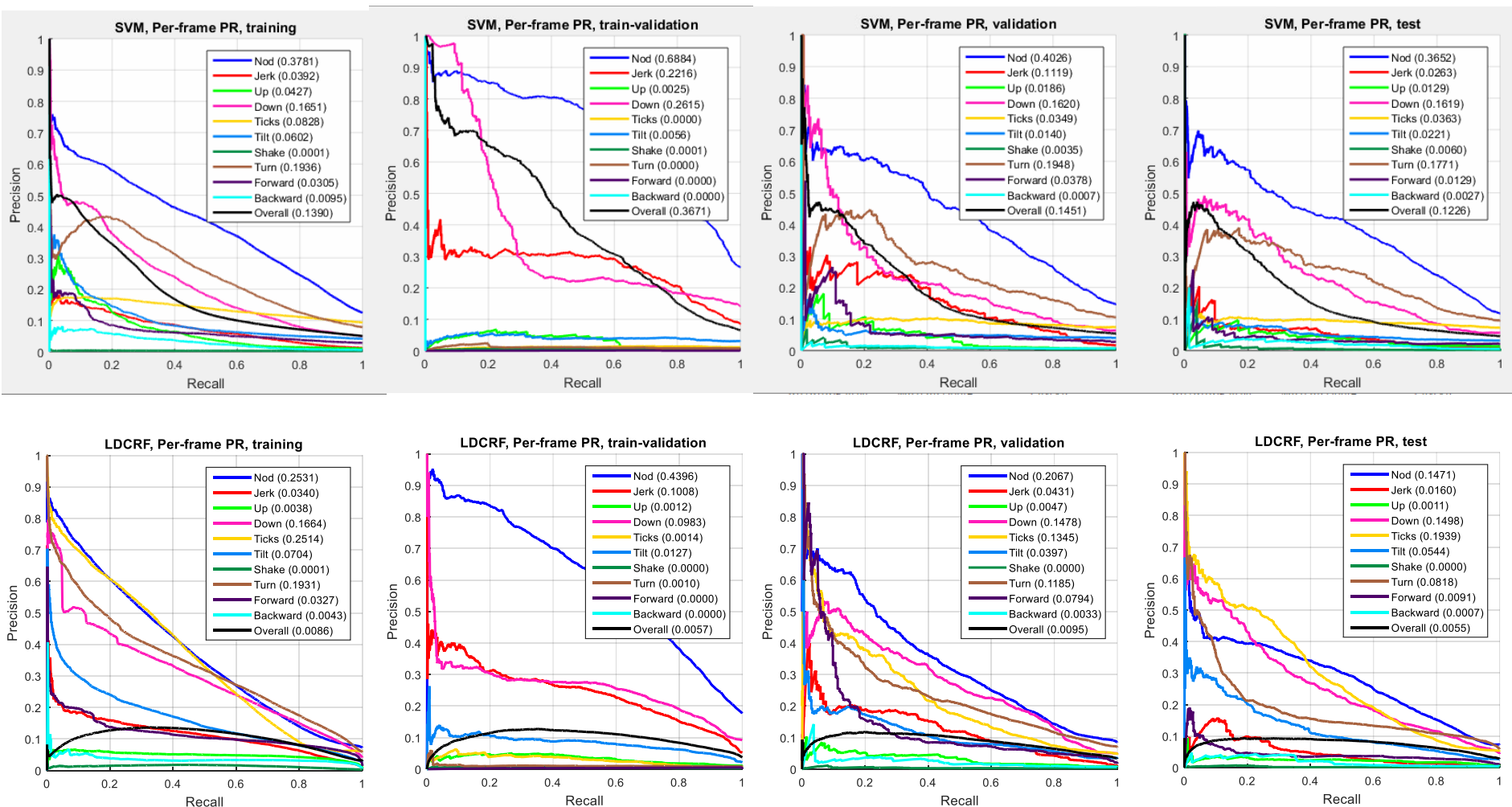
Frame-wise confusion matrix  
(with "None" class)

| Ground Truth | None | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |
|--------------|------|------|------|------|------|-------|------|-------|------|---------|----------|
| None         | .000 | .000 | .000 | .000 | .000 | .000  | .000 | .000  | .000 | .000    | .000     |
| Nod          | .406 | .462 | .012 | .003 | .020 | .037  | .015 | .001  | .034 | .007    | .002     |
| Jerk         | .372 | .074 | .412 | .039 | .007 | .067  | .005 | .000  | .018 | .004    | .002     |
| Up           | .320 | .053 | .031 | .373 | .022 | .013  | .026 | .000  | .088 | .022    | .053     |
| Down         | .341 | .070 | .011 | .009 | .448 | .022  | .026 | .002  | .055 | .015    | .002     |
| Ticks        | .397 | .062 | .038 | .004 | .009 | .445  | .015 | .000  | .023 | .002    | .004     |
| Tilt         | .344 | .029 | .004 | .017 | .053 | .016  | .448 | .007  | .074 | .007    | .001     |
| Shake        | .389 | .028 | .007 | .000 | .021 | .000  | .042 | .458  | .056 | .000    | .000     |
| Turn         | .349 | .072 | .014 | .014 | .032 | .037  | .035 | .001  | .425 | .014    | .006     |
| Forward      | .326 | .071 | .019 | .004 | .052 | .026  | .026 | .000  | .064 | .393    | .019     |
| Backward     | .395 | .048 | .007 | .020 | .020 | .014  | .020 | .000  | .020 | .020    | .435     |
| Prediction   | None | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |

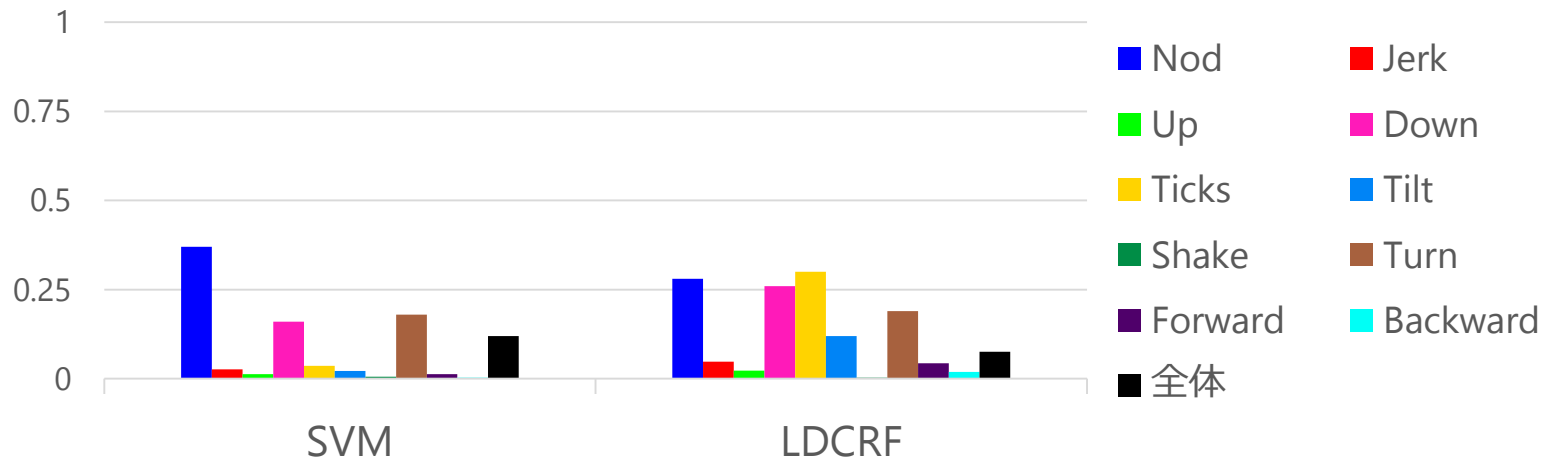
Frame-wise confusion matrix  
(without "None" class)

| Ground Truth | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |
|--------------|------|------|------|------|-------|------|-------|------|---------|----------|
| Nod          | .778 | .021 | .005 | .034 | .062  | .026 | .002  | .057 | .012    | .004     |
| Jerk         | .118 | .656 | .062 | .011 | .107  | .008 | .000  | .028 | .006    | .003     |
| Up           | .077 | .045 | .548 | .032 | .019  | .039 | .000  | .129 | .032    | .077     |
| Down         | .107 | .016 | .013 | .679 | .033  | .040 | .003  | .084 | .023    | .003     |
| Ticks        | .103 | .063 | .007 | .015 | .738  | .024 | .000  | .037 | .004    | .007     |
| Tilt         | .043 | .006 | .026 | .081 | .024  | .683 | .011  | .113 | .011    | .002     |
| Shake        | .045 | .011 | .000 | .034 | .000  | .068 | .750  | .091 | .000    | .000     |
| Turn         | .111 | .022 | .022 | .049 | .057  | .054 | .001  | .652 | .022    | .010     |
| Forward      | .106 | .028 | .006 | .078 | .039  | .039 | .000  | .094 | .583    | .028     |
| Backward     | .079 | .011 | .034 | .034 | .022  | .034 | .000  | .034 | .034    | .719     |
| Prediction   | Nod  | Jerk | Up   | Down | Ticks | Tilt | Shake | Turn | Forward | Backward |

# Results – Detection (PR-curve, AP)



# Results – Detection (AP)



- Poorer results when fewer samples are available
- LDCRF can better model classes with more diversities, e.g. Ticks.

# Conclusions and discussions

- ❑ Spontaneous head gesture recognition is a hard problem
  - Hard for humans, but even harder for automatic recognition
- ❑ Gestures types are not equally hard for automatic recognition
- ❑ Larger model is stronger
- ❑ Deep learning is more promising, but more data is needed.

Identity

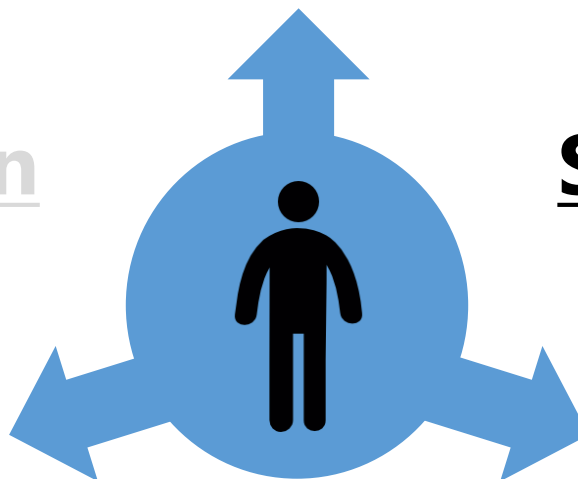
*(Who?)*

Communication

*(What [does he/she want]?)*

*How [does he/she feel]?)*

Explicit expression



State, Action, ...

*(What [is he/she doing]?)*

*How [does he/she do it]?)*

Implicit expression



# Proposal of a Wrist-mounted Depth Camera for Finger Gesture Recognition

Kai Akiyama, Yang Wu  
*Nara Institute of Science and Technology*

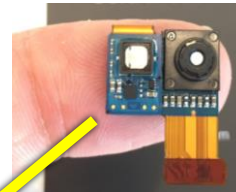


AR/VR controller

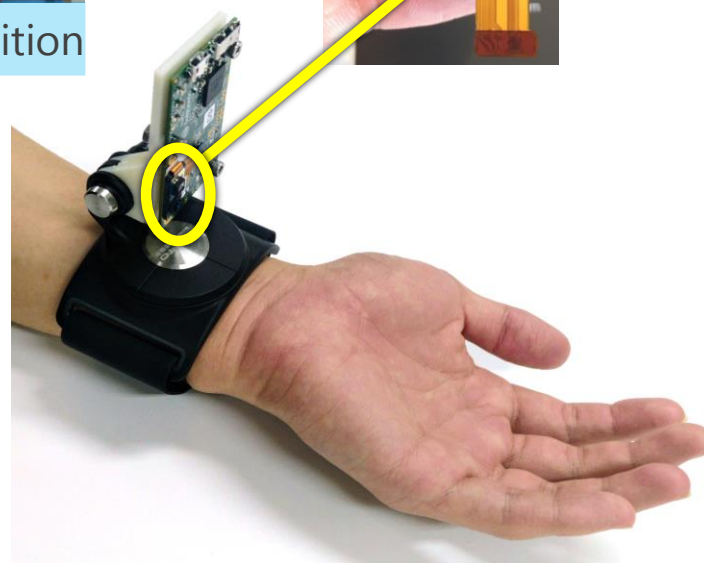


Daily activity recognition

Time-of-Flight camera

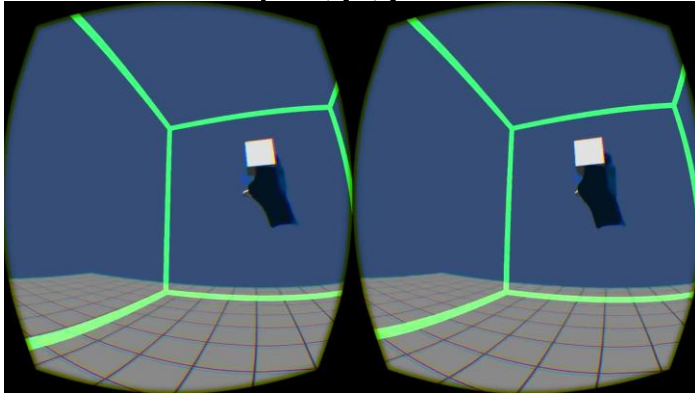


Retrieved depth images



# Hand pose estimation - Applications

Playing games



Driving assistant



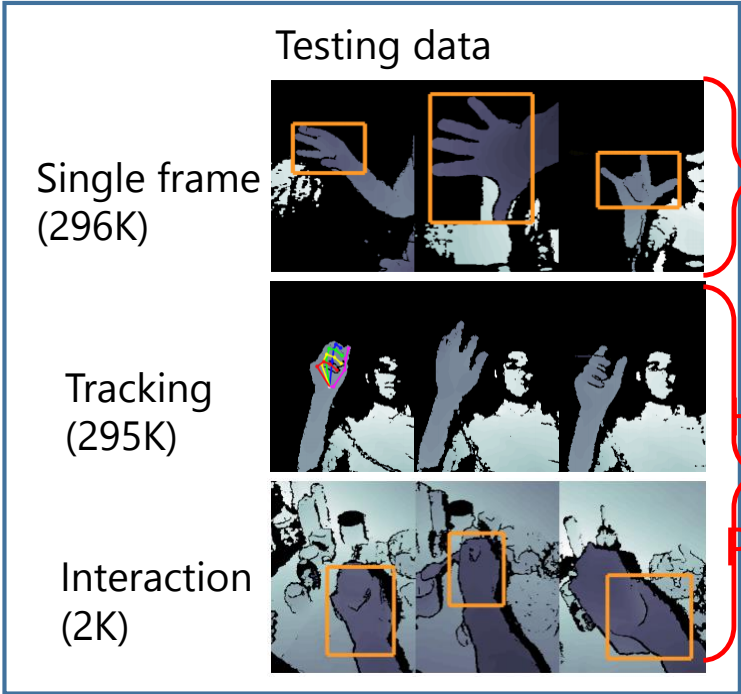
Surgery assistant



etc.

# Background – Depth-based 3D hand pose estimation benchmark

## Hands In the Million Challenge (HIM2017)

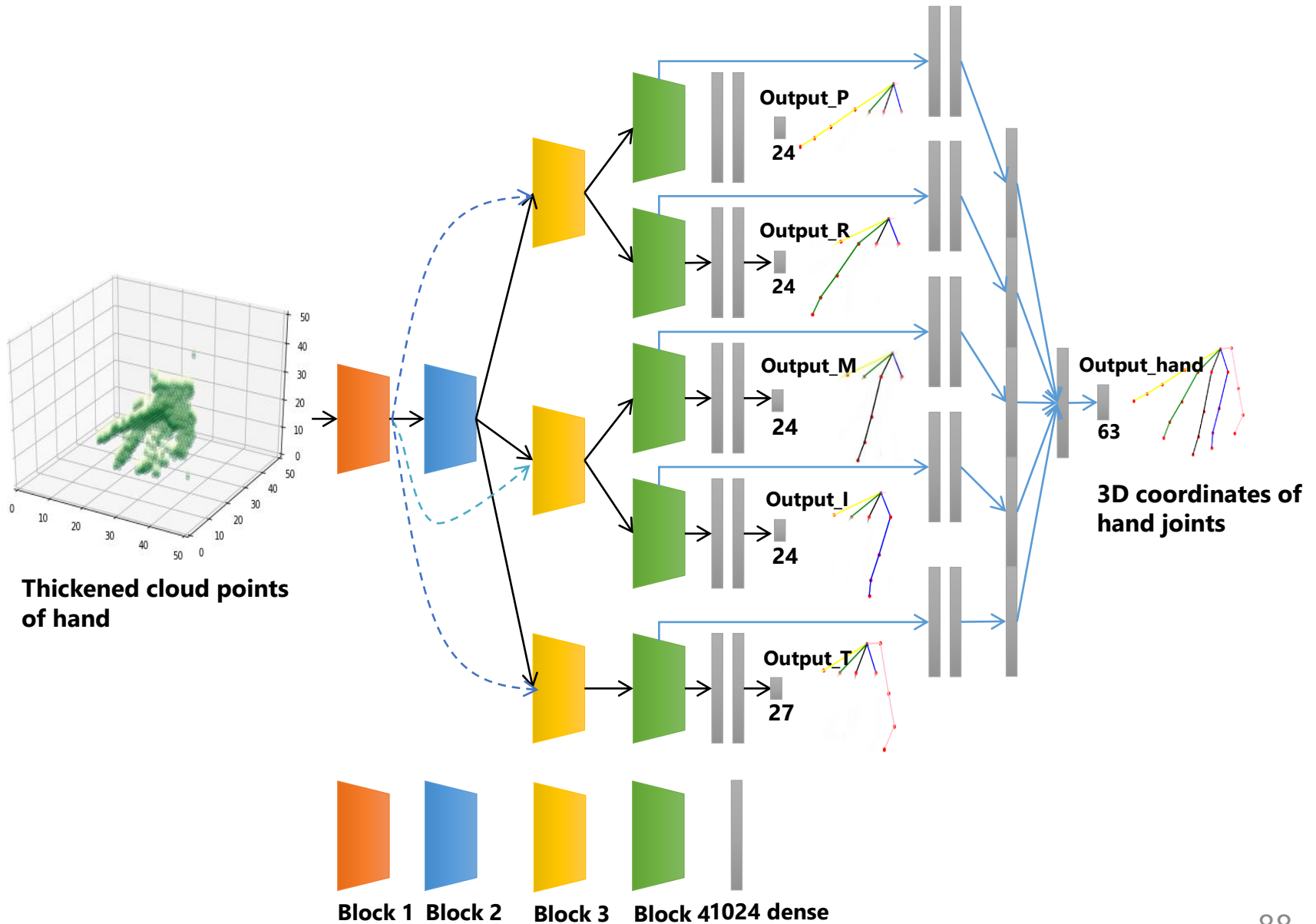


Pose estimator

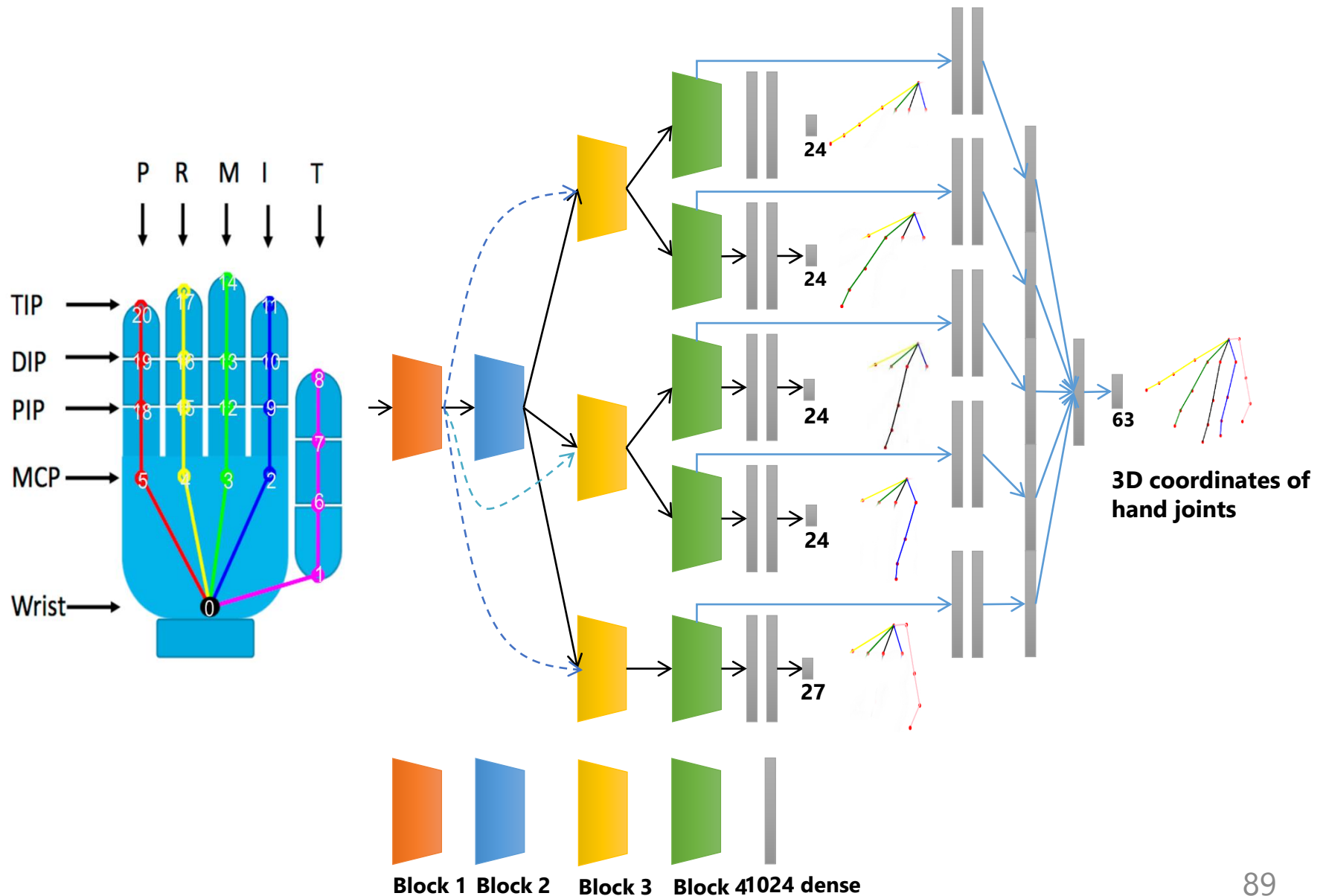
Hand detector + Pose estimator

(S. Yuan, et al. 2017)

# Proposed 3D hand pose estimator architecture (1)



## Proposed 3D hand pose estimator architecture (2)



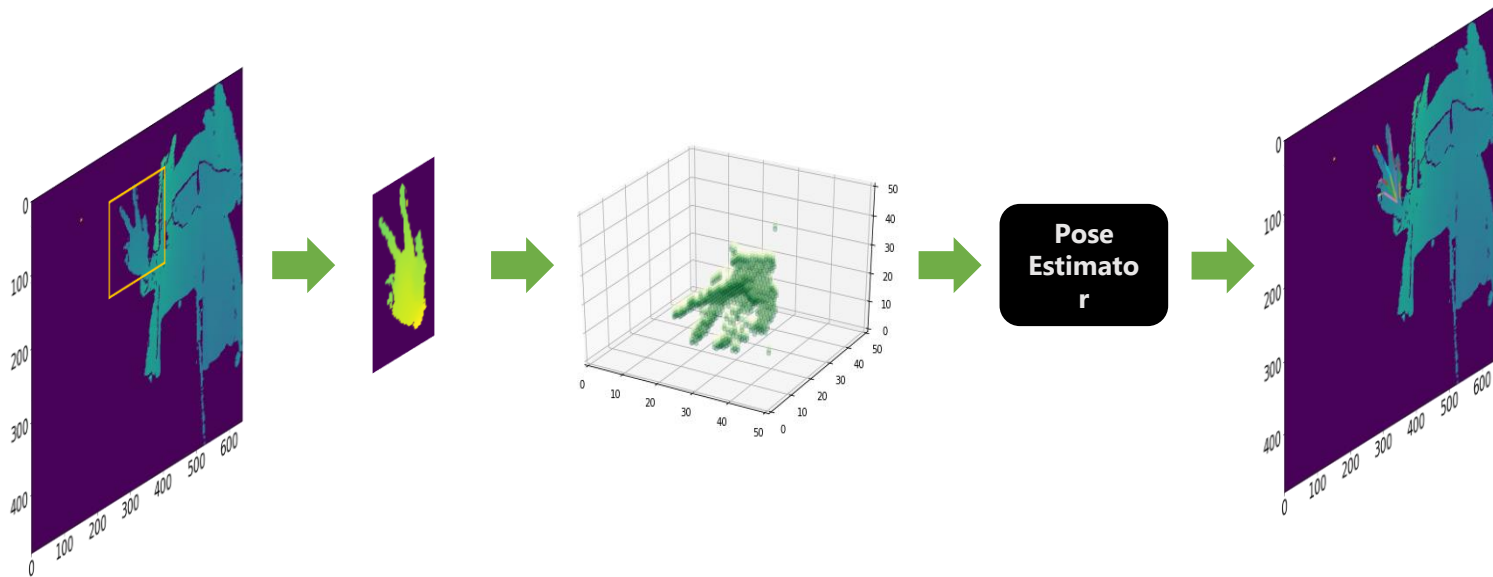
# Pipeline of Pose estimator

## Single frame pose estimation

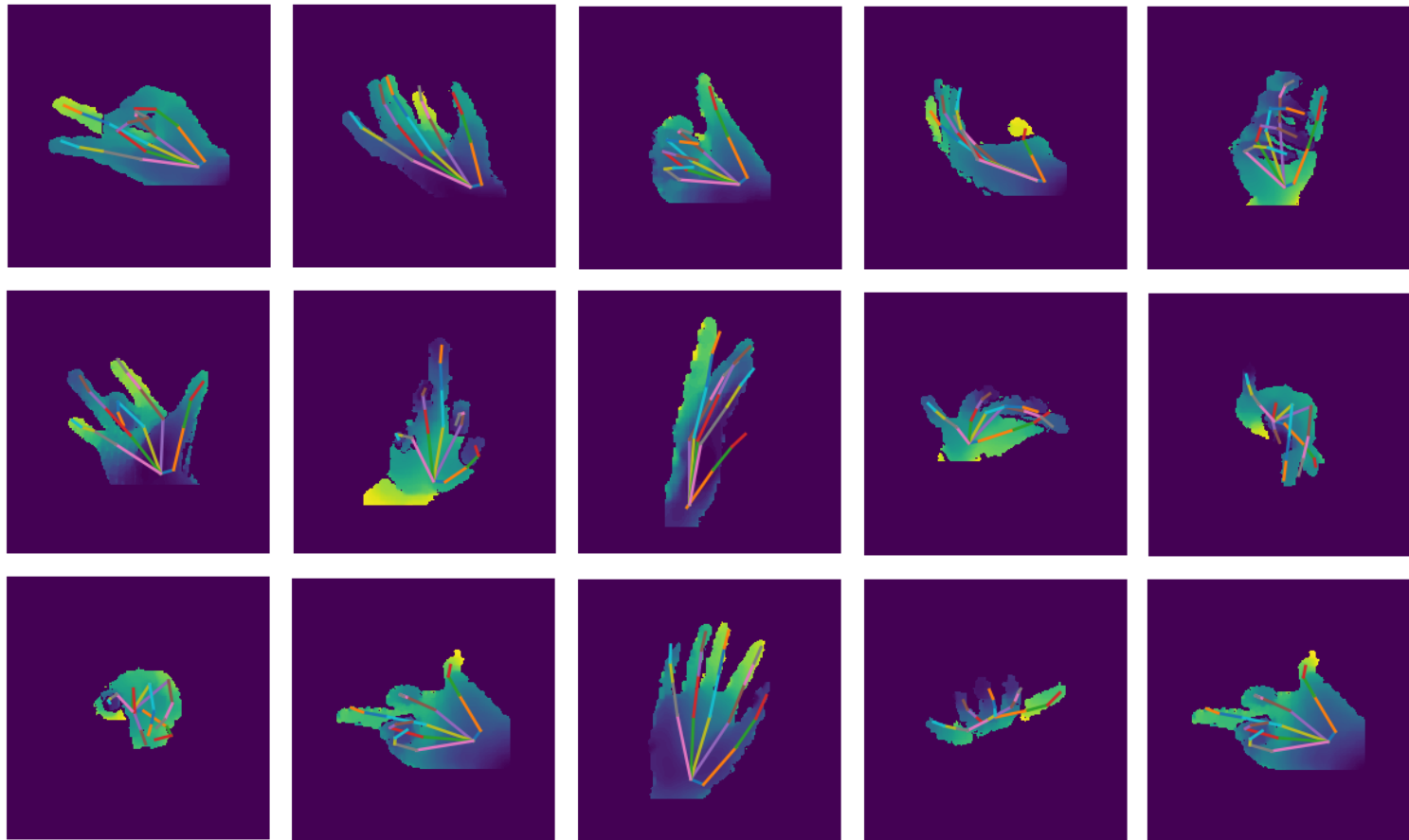
Extracting hand based by given bounding box

Represent data by 50x50x50 volume

Estimate 3D hand pose and transform back to original coordinates



# Qualitative results of 3D hand pose estimator

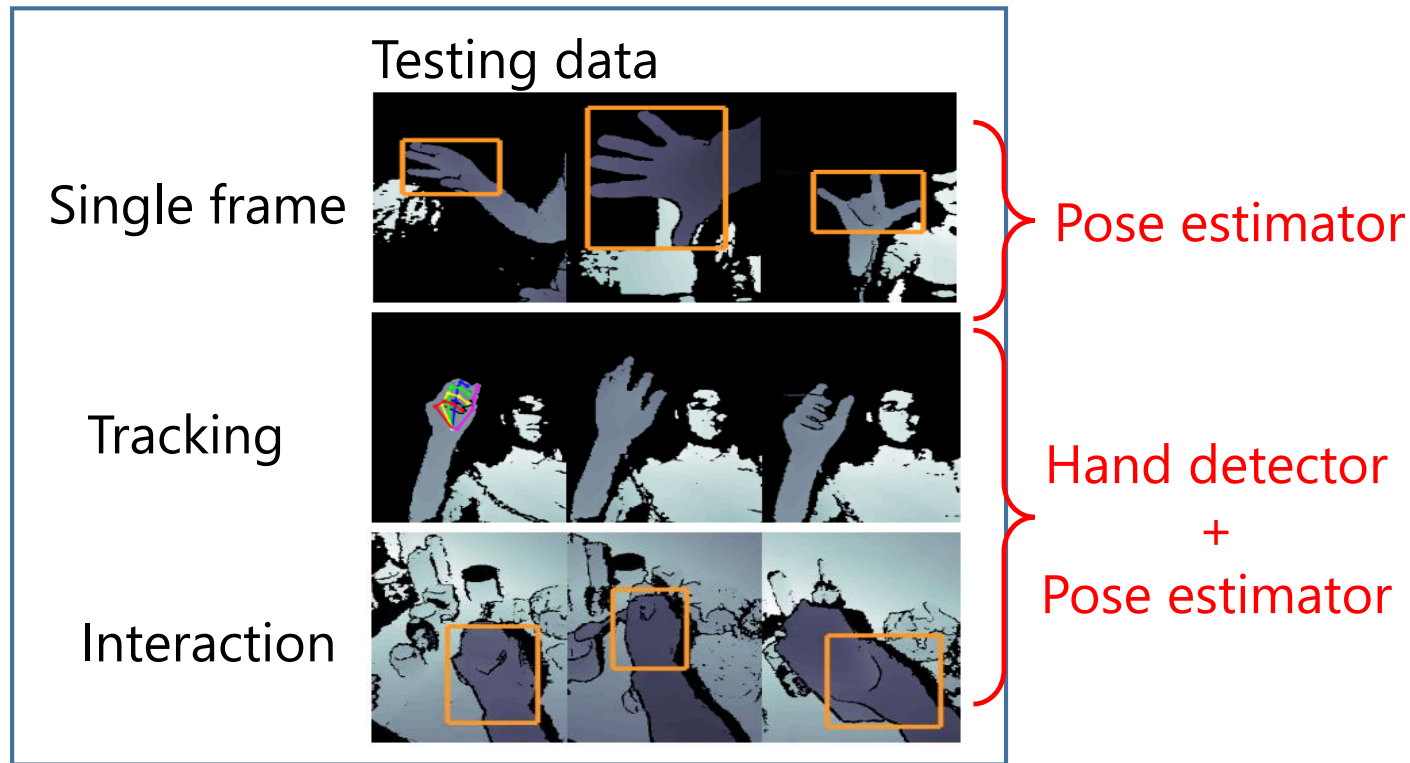


# Evaluation on the 3D hand pose estimation task of HIM2017 benchmark

| Team name                     | AVG ERROR (mm) | SEEN ERROR (mm) | UNSEEN ERROR (mm) |
|-------------------------------|----------------|-----------------|-------------------|
| SNU CVLAB                     | 9.95           | 6.97            | 12.43             |
| NVIDIA Research and UMontreal | 9.97           | 7.55            | 12.00             |
| NTU                           | 11.30          | 8.86            | 13.33             |
| THU VCLab                     | 11.70          | 9.15            | 13.83             |
| NAIST RVLab                   | 11.90          | 9.34            | 14.04             |
| Baseline                      | 19.71          | 14.58           | 23.98             |



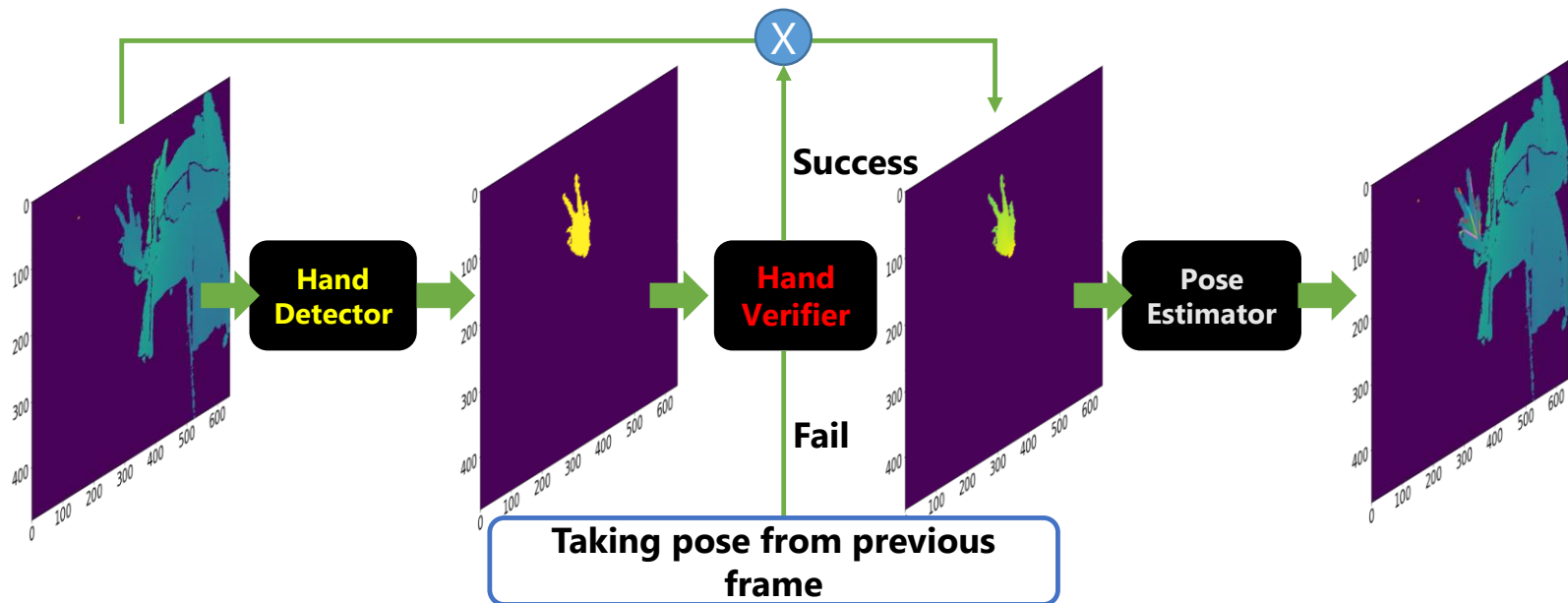
# Utilizing a hand detector for tracking and interaction task



We need a hand detector to find where is the hand in real application

# Architecture of the 3D hand pose tracking system

## Hand detector + Hand verifier + Pose estimator



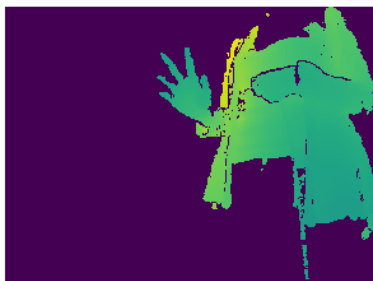
Hand verifier:

1. Comparing with the previous frame, whether the center of detected hand area shift more than 150 mm;
2. Whether the number of pixels for detected hand area is more than 1000.

# Qualitative results of 3D hand pose tracking

Sequential Frames

Depth  
image



...



...

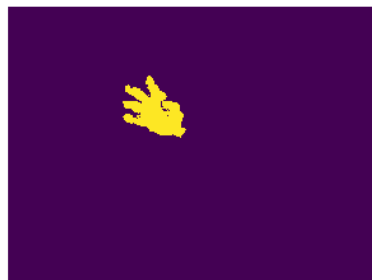


...

Hand  
mask



...

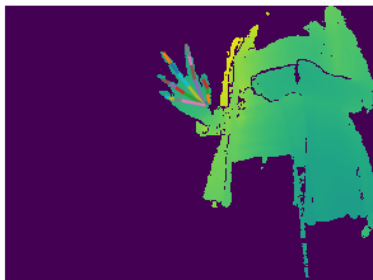


...



...

Estimated  
hand pose  
and depth  
image



...



...

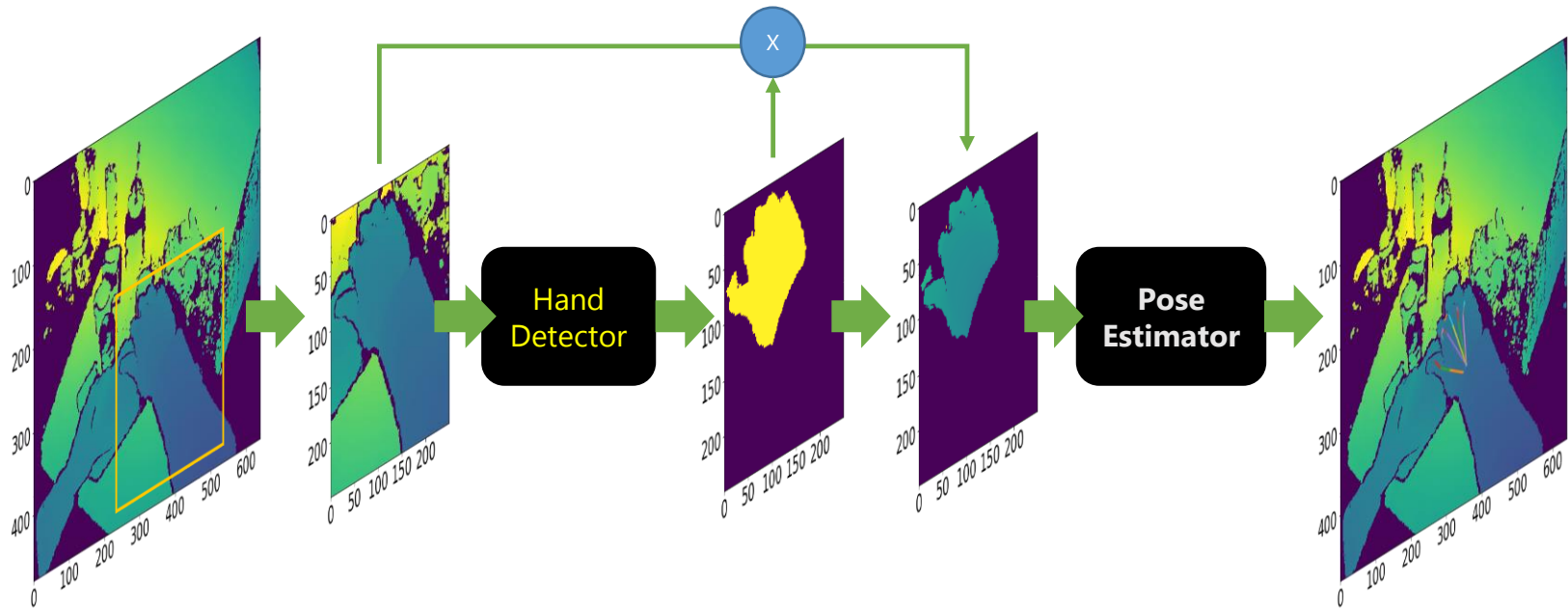


...

# Evaluation on the 3D hand tracking task of HIM2017 benchmark

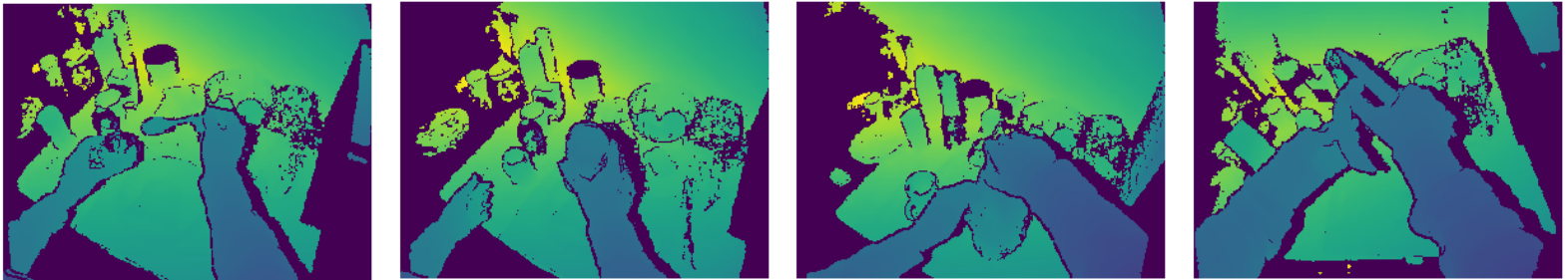
| Team name                     | AVG ERROR (mm) | SEEN ERROR (mm) | UNSEEN ERROR (mm) |
|-------------------------------|----------------|-----------------|-------------------|
| NVIDIA Research and UMontreal | 10.51          | 8.21            | 12.37             |
| NAIST RVLab                   | 12.64          | 10.20           | 14.62             |
| THU VCLab                     | 13.65          | 11.02           | 15.70             |
| Baseline                      | 20.63          | 16.04           | 24.36             |

# Applying modified tracking system on Hand object interaction



# Qualitative results of 3D hand-object interaction pose estimation

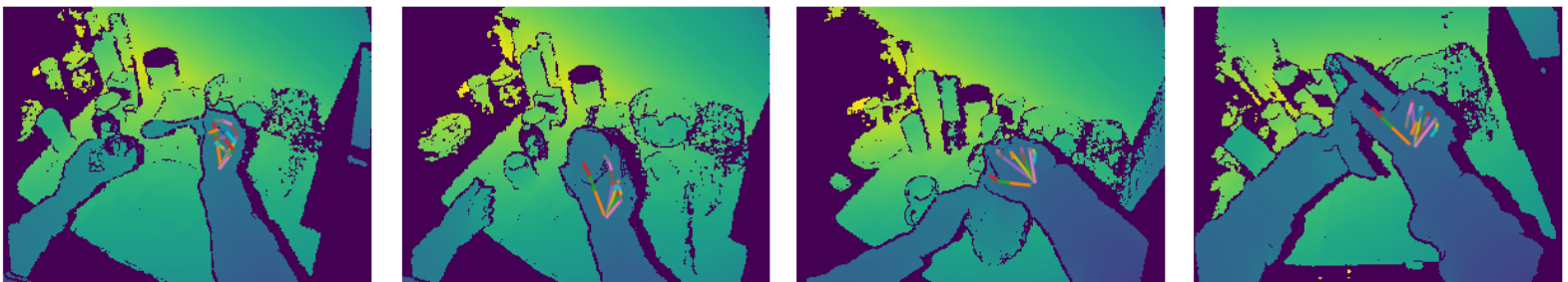
Depth image



Hand mask



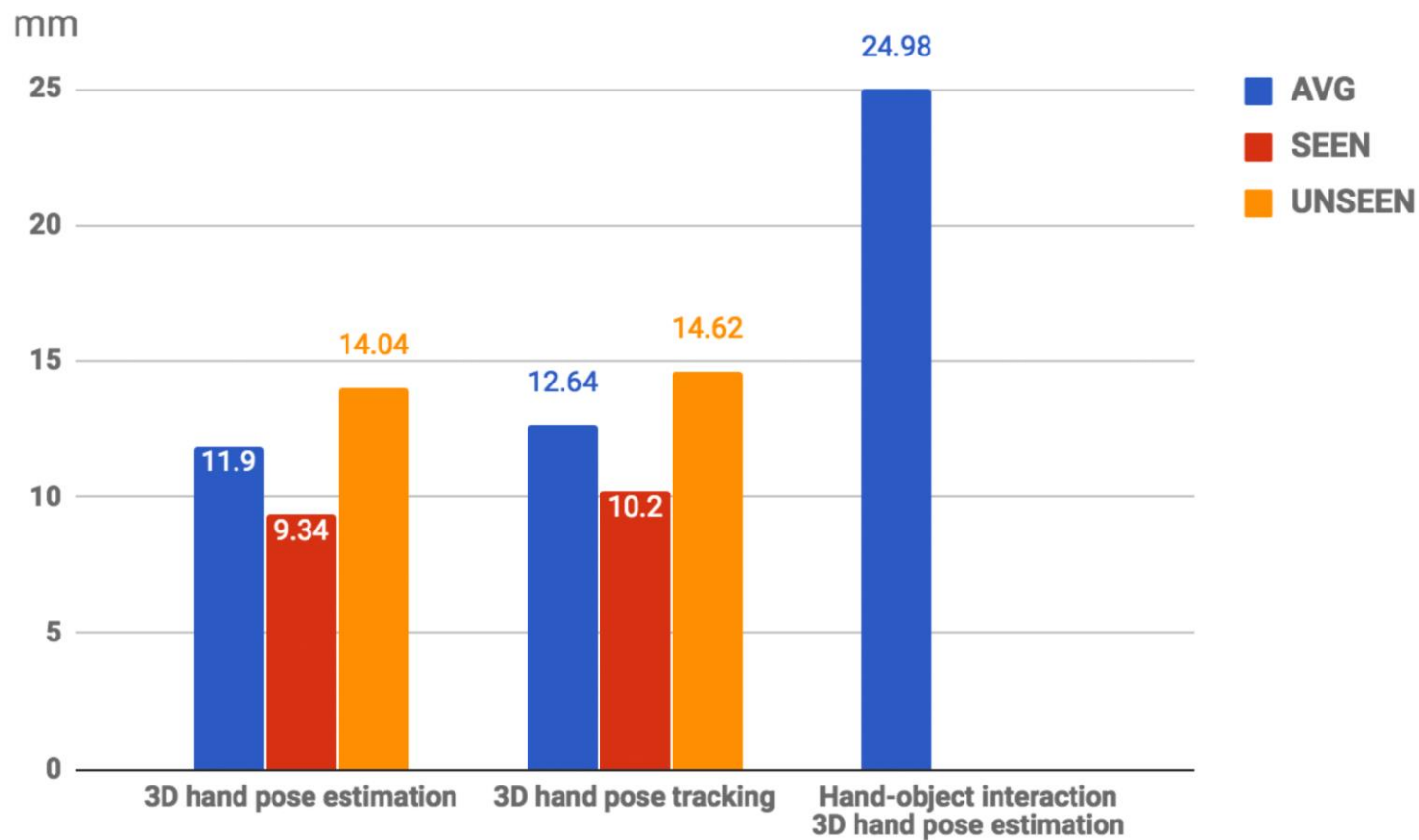
Estimated  
hand pose  
and depth  
image



# Evaluation on the hand object interaction task of HIM2017 benchmark

| Team name                     | AVG ERROR (mm) |
|-------------------------------|----------------|
| NAIST RVLab                   | 24.98          |
| THU VCLab                     | 29.19          |
| NVIDIA Research and UMontreal | 32.44          |
| Baseline                      | 46.10          |

# Evaluation results on all tasks of HIM2017 benchmark





# Who is Doing What in Drone-recorded WAMI

Pred: Carrying, Standing Pred: Carrying, Standing Pred: Reading, Sitting Pred: Reading, Standing  
True: Carrying, Standing True: Carrying, Standing True: Reading, Sitting True: Reading, Standing



Pred: Pushing/Pulling, Walking  
True: Pushing/Pulling, Walking



Pred: **Sitting**  
True: **Sitting**



Pred: Reading, Sitting  
True: Carrying, Sitting



Pred: Pushing/Pulling, Walking  
True: Pushing/Pulling, Walking



Pred: **Walking**  
True: Calling, Walking



Pred: **Standing**  
True: Carrying, Standing



Pred: Carrying, Walking  
True: Carrying, Walking



Pred: Walking  
True: Carrying, Walking



Pred: Lying  
True: Lying



Pred: **Sitting**  
True: Reading, Sitting

# Results – Region Proposal



# Results – Tracking





# Results – Action Recognition

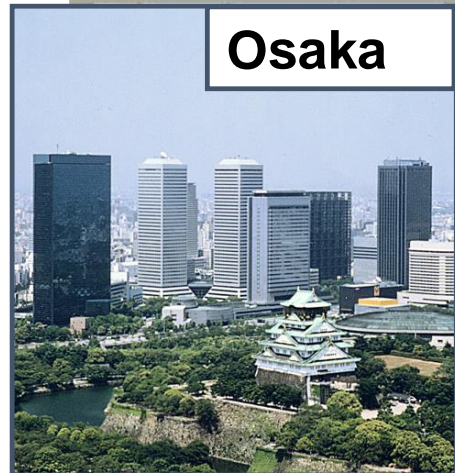


# **About NAIST**

# NAIST Location



**Kyoto**



**Osaka**



**Nara**



# Kansai Science City (Keihanna)

Research park in the Kansai Hills area, extending to three prefectures, Kyoto, Osaka and Nara, and covering about 150 km<sup>2</sup>. More than 110 companies and institutes such as:

Kyocera



Panasonic



ATR (Advanced  
Telecommunications  
Research Institute International)



NICT (National Institute of  
Information and Communications  
Technology)

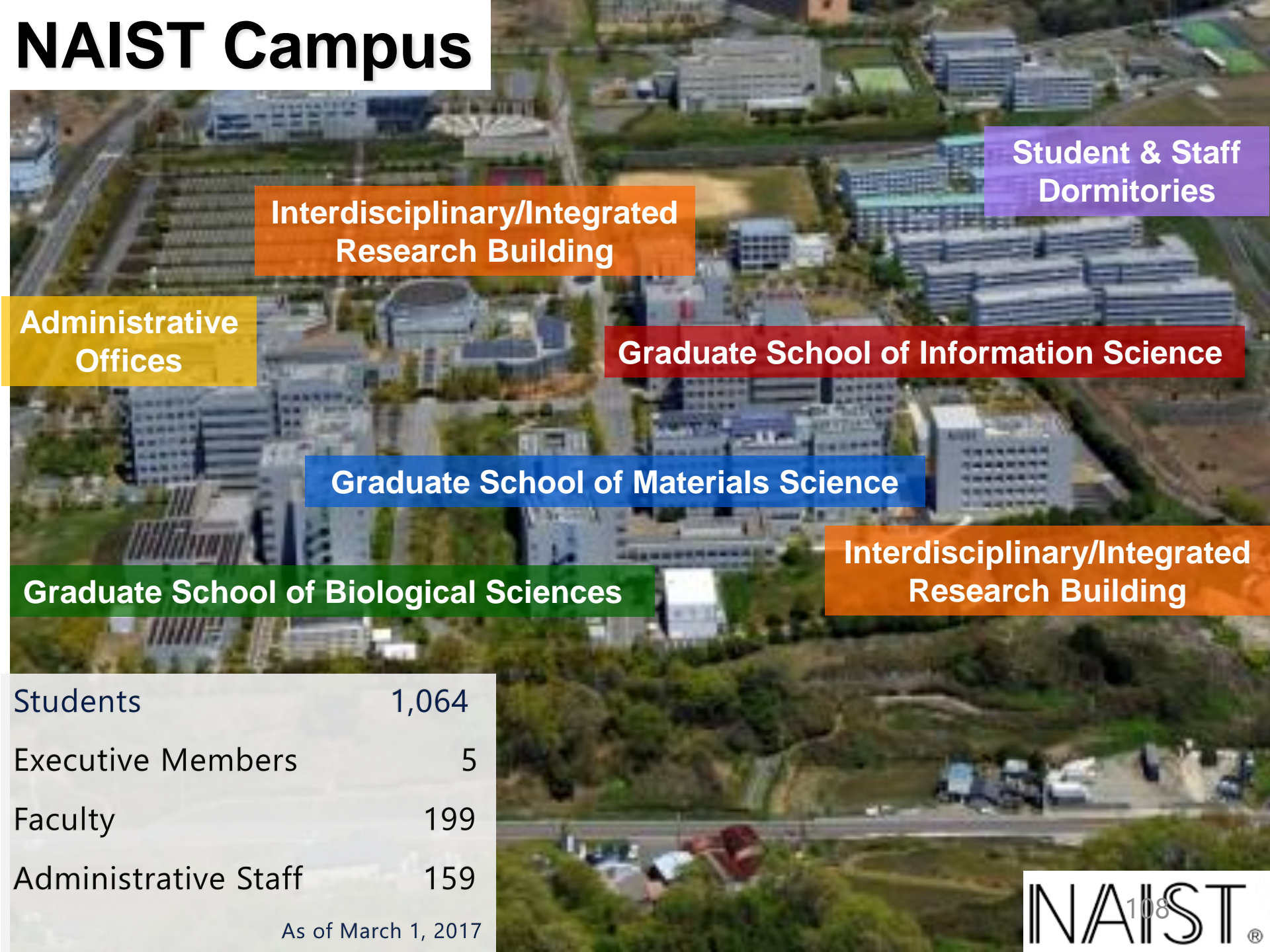


RITE (Research Institute of  
Innovative Technology for the  
Earth)





# NAIST Campus



Student & Staff  
Dormitories

Interdisciplinary/Integrated  
Research Building

Administrative  
Offices

Graduate School of Information Science

Graduate School of Materials Science

Graduate School of Biological Sciences

Interdisciplinary/Integrated  
Research Building

|                      |       |
|----------------------|-------|
| Students             | 1,064 |
| Executive Members    | 5     |
| Faculty              | 199   |
| Administrative Staff | 159   |

As of March 1, 2017



# GSIS: Core Laboratories

## *Media Informatics*

Computational Linguistics  
Augmented Human Communication  
Network Systems  
Vision and Media Computing  
Interactive Media Design  
Optical Media Interface  
Ambient Intelligence

## *Computer Science*

Computing Architecture  
Dependable System  
Ubiquitous Computing System  
Mobile Computing  
Software Engineering  
Software Design and Analysis  
Internet Engineering  
Internet Architecture and Systems

## *Applied Informatics*

Robotics  
Intelligent System Control  
Large-Scale Systems Management  
Mathematical Informatics  
Imaging-based Computational Biomedicine  
Computational Systems Biology  
Robotics Vision

# NAIST External Evaluation

## The 87<sup>th</sup> Session of the Council for Science and Technology Policy

|                                    |   |
|------------------------------------|---|
| Ranked 1 <sup>st</sup><br>in Japan | Revenue for research expenses (per faculty member)                      |
|                                    | Number of Grants-in-Aid for scientific research (per faculty member)    |
|                                    | Allotment of Grants-in-Aid for Scientific Research (per faculty member) |
|                                    | Revenue from patent implementation (per faculty member)                 |
|                                    | Number of university business ventures (per faculty member)             |
|                                    | Percentage of Young Faculty (Younger than 37 years old)                 |

## Ranking 2013 by Asahi Shimbun

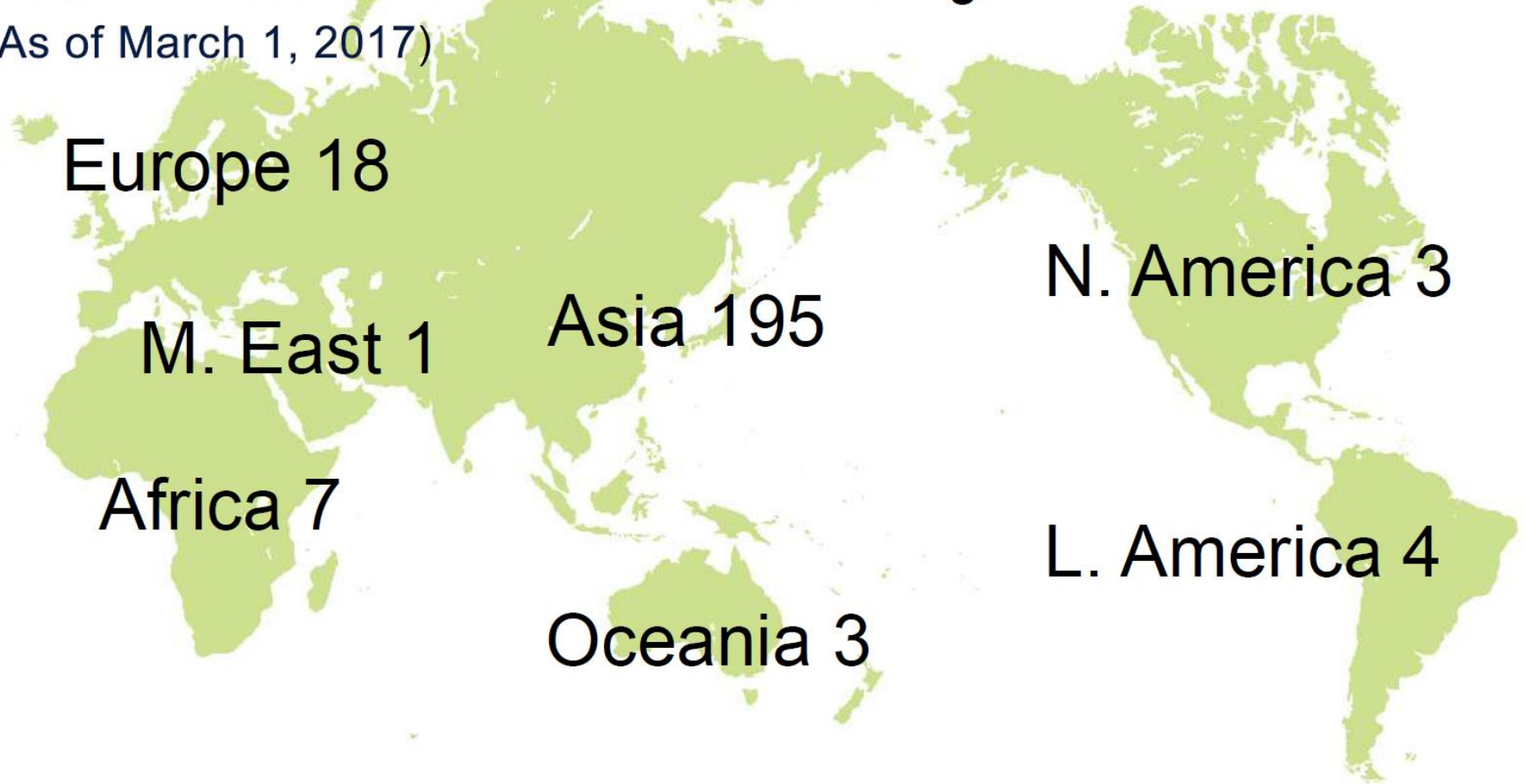
|                        |  |
|------------------------|--|
| Ranked 1 <sup>st</sup> | Citation Index of ISI (overall) among Japanese National Universities |
|------------------------|--|

# International Students @ NAIST

International students comprise about 22%

Total: 231 students from 31 countries/regions

(As of March 1, 2017)



# **NAIST elected for major university programs by MEXT**

- ✓ 2014 Top Global University Project
- ✓ 2013 The Program for Promoting the Enhancement of Research Universities

# **About My Lab**

## ABOUT THE LAB

**NAIST International Collaborative Laboratory for Robotics Vision** is focusing on enhancing robotics and improving people's life quality by exploring the most of computer vision technologies. Established in Dec 2014 by NAIST and CMU (Carnegie Mellon University), this laboratory provides a unique platform for conducting leading research (both basic and applied) via close collaboration among international talents.

As one of the first two international collaborative laboratories built by NAIST, this laboratory is different from any other research labs and collaborative labs that you can find in the Graduate School of Science and Technology.



### NAIST-CMU

Lead by Prof. Takeo Kanade and closely collaborated with Carnegie Mellon University.



### International

Gathering active and talented researchers with very diverse nationalities.



### Research

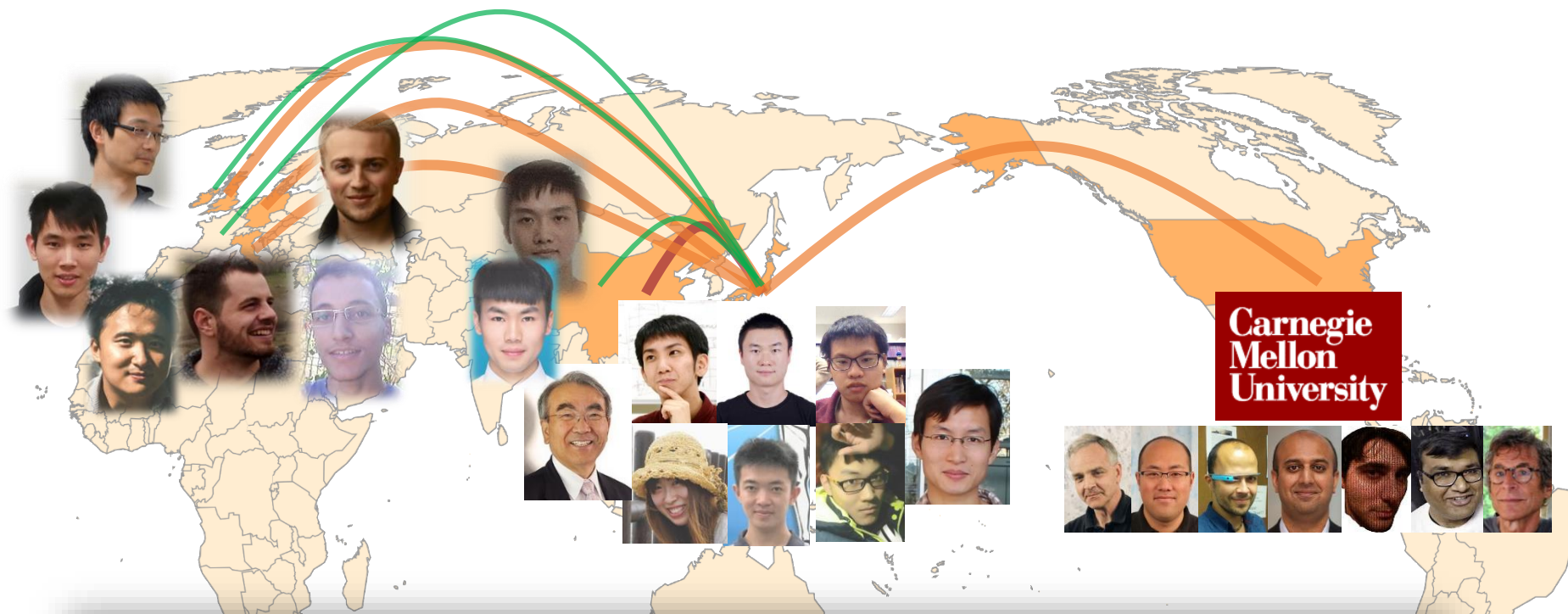
Targeting at innovative research with global and long-lasting impact.



### Bridge

Better bridging NAIST researchers and international peers.

# NAIST International Collaborative Laboratory for Robotics Vision





## We won

- **Best Student Paper Award:** The Piero Zamperoni Best Student Paper Award of ICPR 2018 (Global)
- **Best Paper Award:** The AutoML 2018 workshop @ ICML/IJCAI-ECAI 2018 (Global)
- **Winner:** The 2017 Hands in the Million Challenge (Hand-Object Interaction Task) (Global)
- **Winner:** ISMAR 2015 Tracking Competition (Off-Site Category: Level 1) (Global)
- **Excellent Demo:** IPSJ Distributed Processing System Workshop 2016 (Japan)
- **Excellent Award:** Creative and International Competitiveness Project 2017 (NAIST)
- **Excellent Student Award:** 2018 Excellent Student Award (NAIST)
- **Excellent Student Award:** 2017 Excellent Student Award (NAIST)



The Best International Collaborative Lab of NAIST, 2017

