Distributed and Lightweight Multi-Camera Human Activity Classification

Gaurav Srivastava, Hidekazu Iwaki, Johnny Park, Avinash C. Kak

School of Electrical and Computer Engineering, Purdue University, USA

Outline

- Motivation for multi-view analysis
- Typical multi-camera algorithms: issues with distributed implementation
- Proposed Action Classification Algorithm
- Classification results
- Why is it lightweight?
- Conclusions

Motivation for Multi-view Analysis

- Logical next step to fixed-view activity analysis
- Does not constrain the human's orientation to frontal or profile views relative to single camera
- Capturing action from multiple views $\Longrightarrow$ additional features for higher discriminative ability
- Robustness to partial occlusions

Typical Multi-camera Algorithms

- Assume that images from multiple cameras can be transmitted for central processing
- Leads to significant bandwidth requirement even for commonly used parameters:
  - frame rates of 15-30 fps,
  - image resolutions like 320x240 pixels
  - 5-10 cameras
- Computationally intensive operations: 3D visual hull construction, 3D model projection onto multiple 2D views for matching

Distributed Processing: what is desirable?

- Transmit compact representative descriptors instead of entire images
- Modular design: each camera node can independently process local sensory data
- Low memory requirements at each camera node
- Simple and fast aggregation algorithms
- Not compromise on the classification performance (compared to the centralized multi-camera approach)

Contributions of the paper

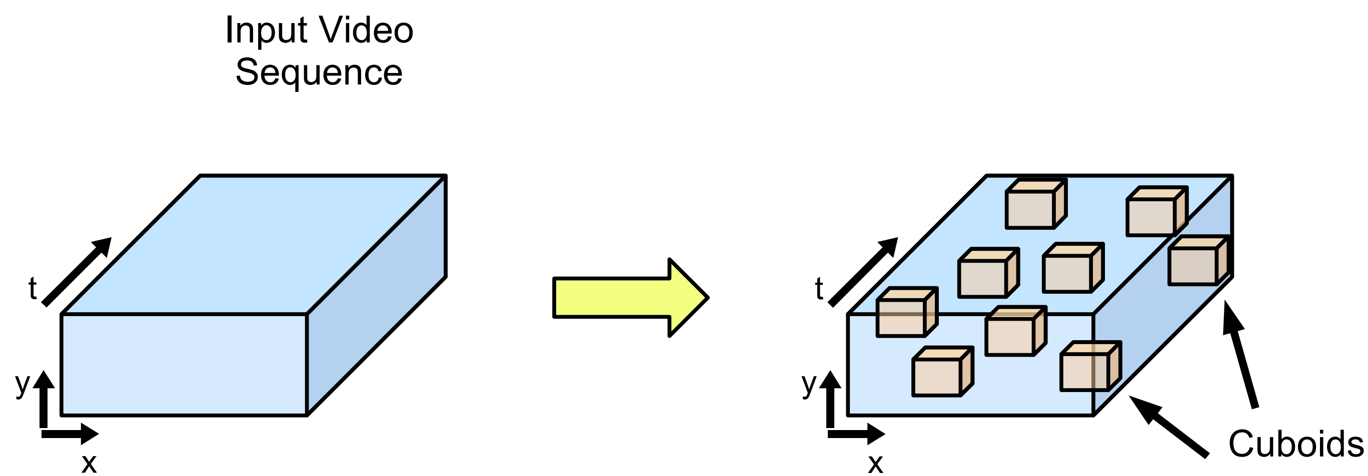
- Extend the feature histogram representation (Dollar et al. 2005) to multiple cameras and present a simple aggregation algorithm
- Demonstrate some level of invariance to actor orientation
- Demonstrate robustness to previously unseen views
- Analyze the system's superior storage and bandwidth requirements $\Rightarrow$ demonstrate suitability for a distributed implementation.

Proposed Methodology

- Represents actions using spatio-temporal features
- Achieves orientation invariance using multi-view action representation
- Suitable for distributed implementation

Spatio-temporal feature extraction

(Dollar et al. 2005)

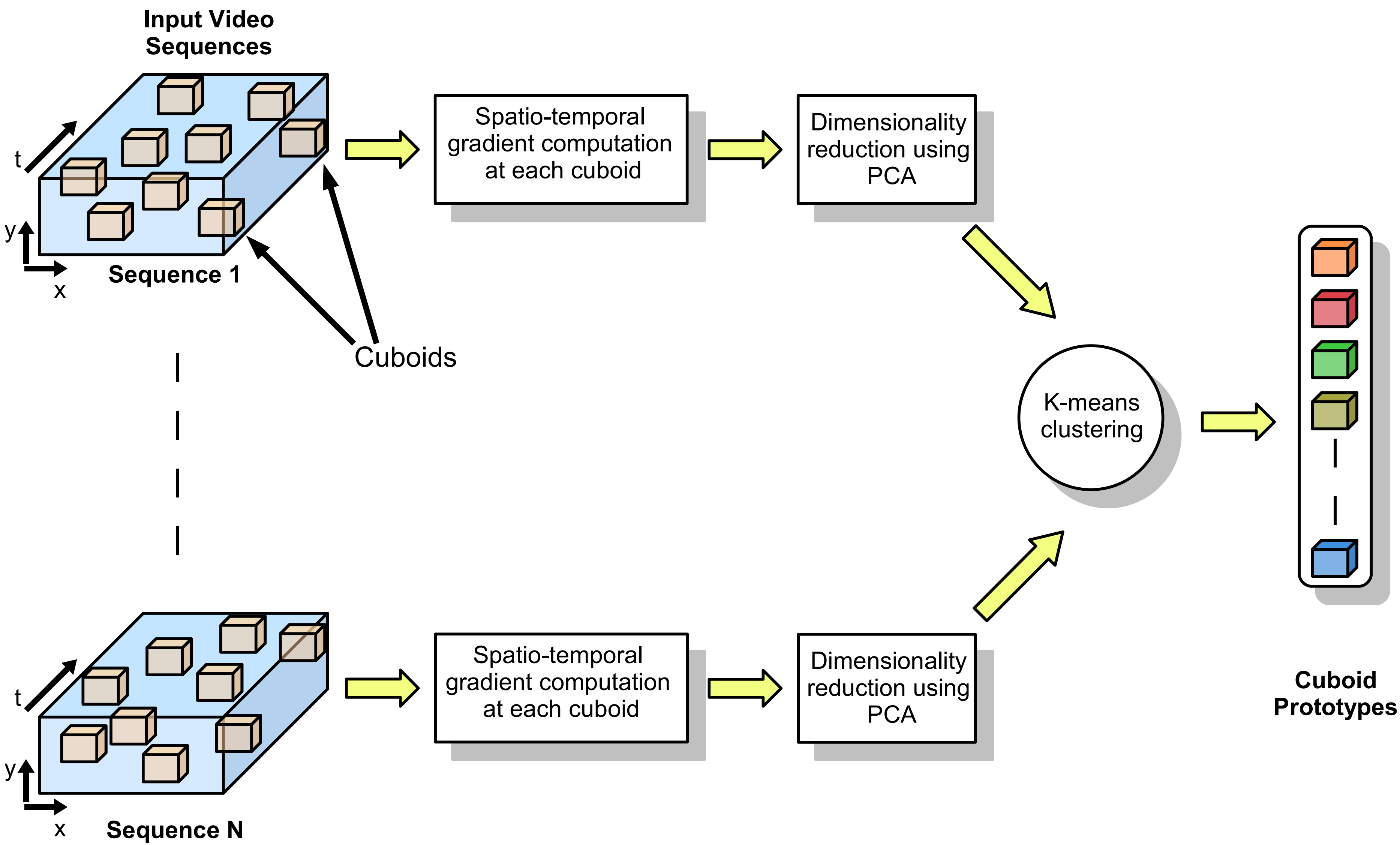


Convolve the video sequence with a spatio-temporal linear filter to obtain respose function R. Local maxima of R are locations of cuboids.

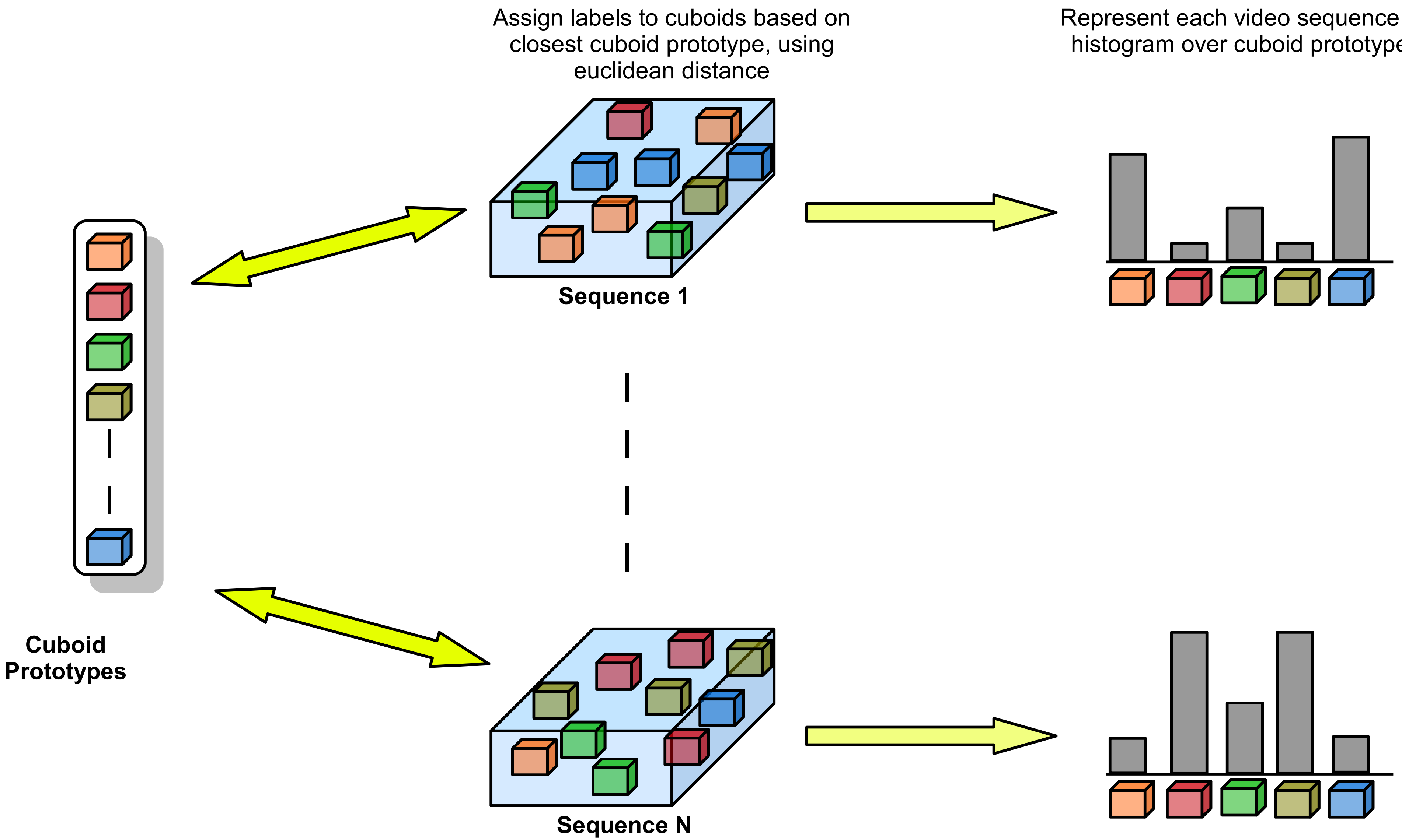
$$R = (I \star g \star h_{ev})^2 + (I \star g \star h_{od})^2$$

Spatio-temporal feature extraction

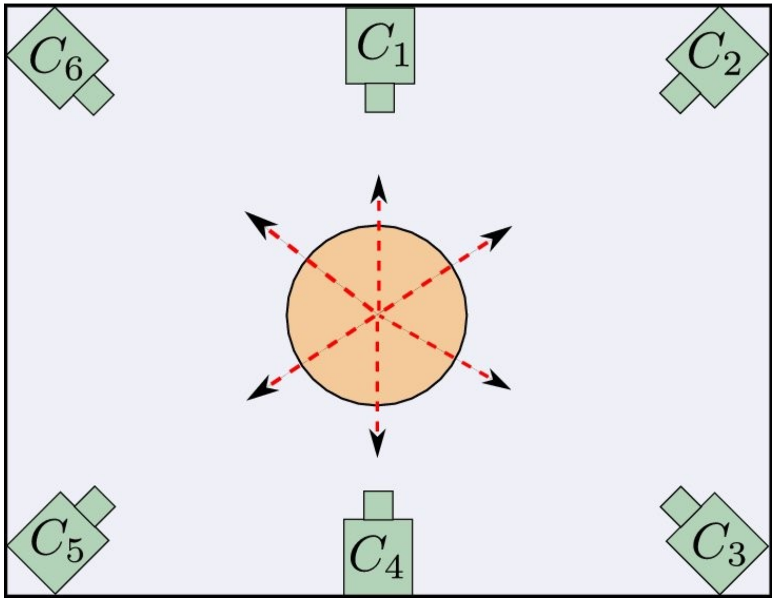
(Dollar et al. 2005)



Action Histogram Generation

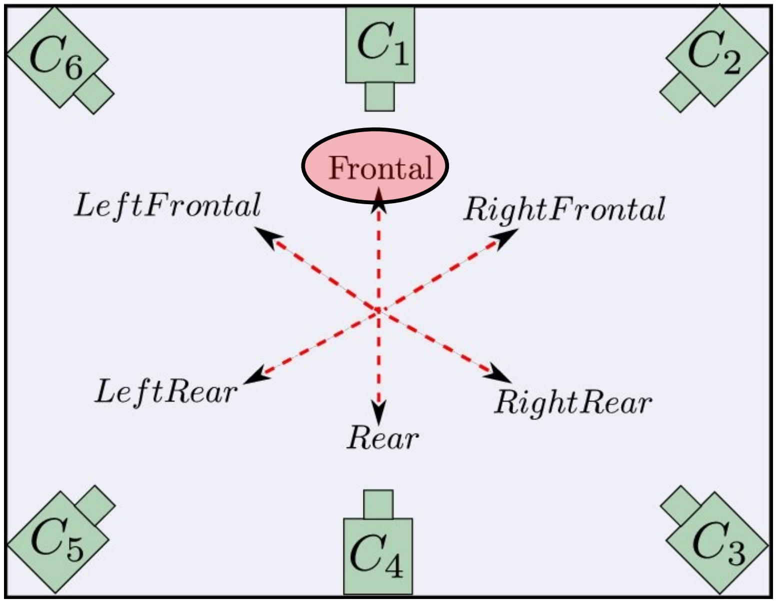


Experimental Setup

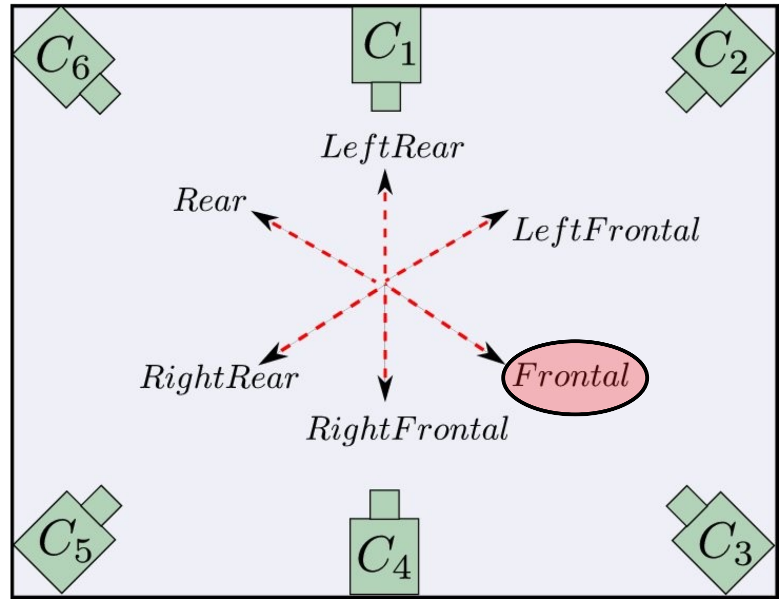


- 6 cameras, placed approximately uniformly around the room, at same height.
- Subjects can perform actions facing any of the cameras. *Discretized* orientation invariance.
- *Even if actor's orientation not along one of the cameras, still high classification performance achieved.*

Orientation Invariance



Subject facing camera C_1

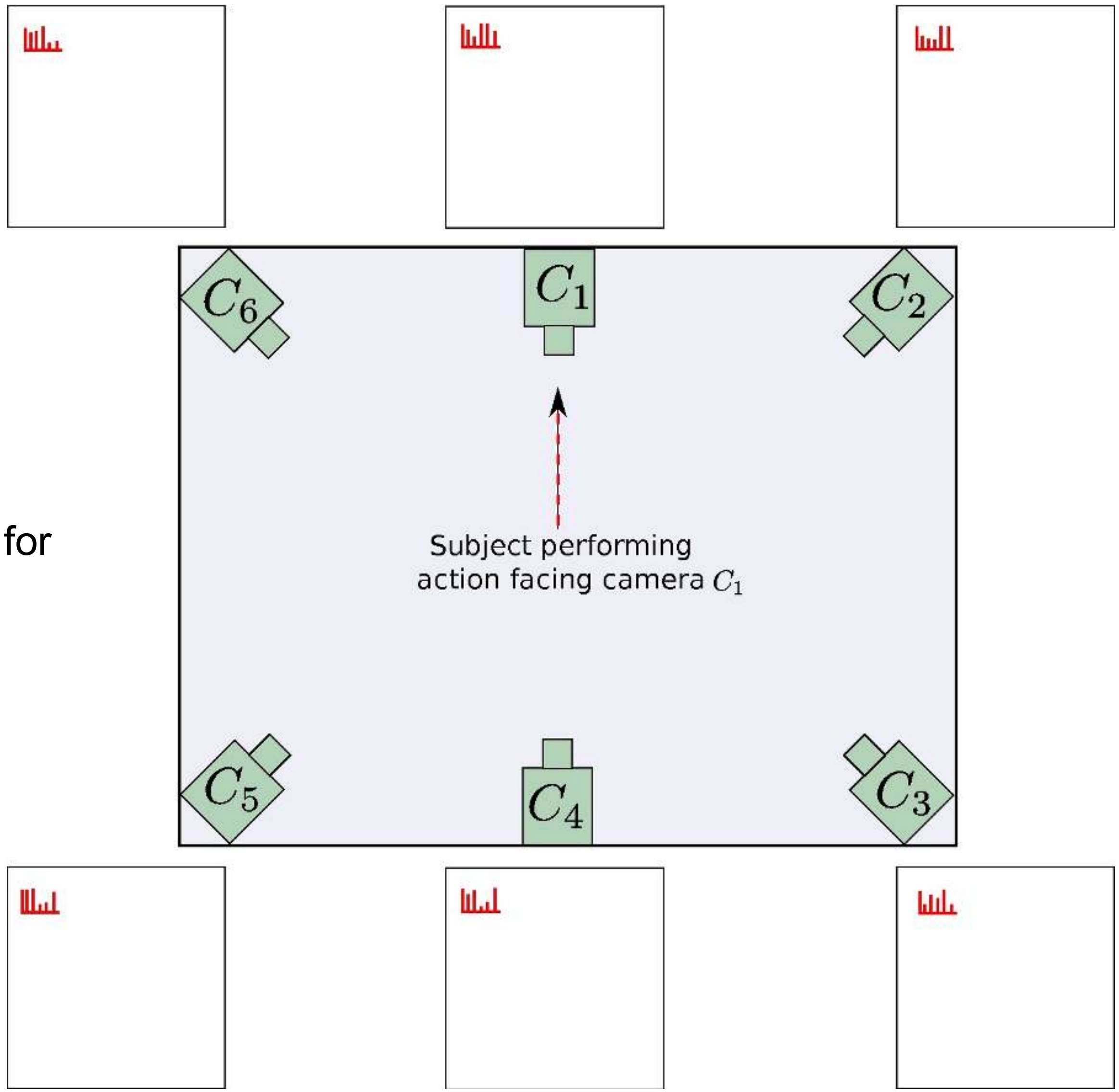


Subject facing camera C_3

	LeftFrontal	Frontal	RightFrontal	Rear	LeftRear	RightRear
Facing Camera C1	C6	C1	C2	C5	C4	C3
Facing Camera C3	C2	C3	C4	C1	C6	C5

Multi-view action representation

The 6 histograms corresponding to 6 camera views constitute the multi-view representation for any action.

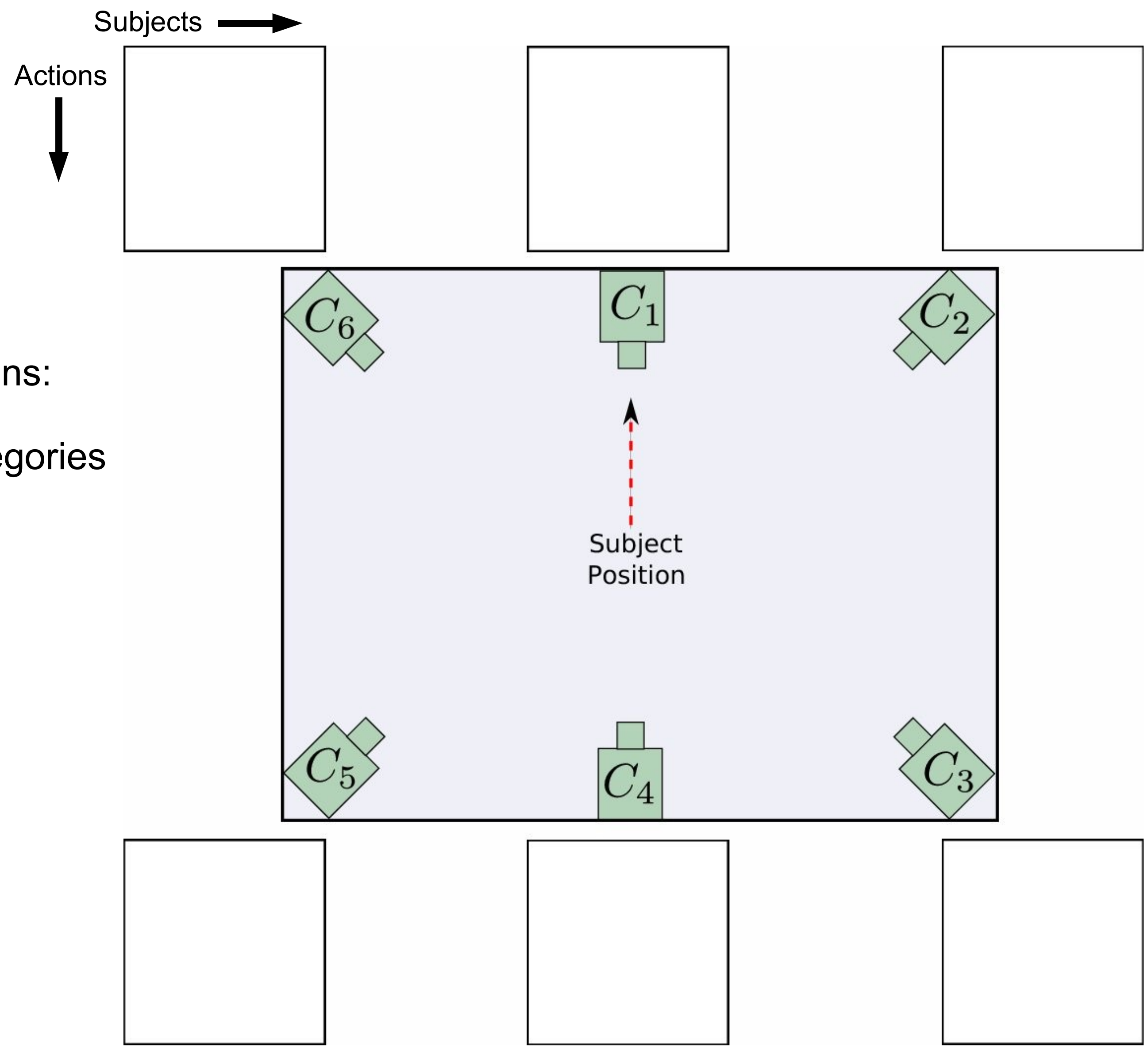


Multi-view Action Classification: Training Stage

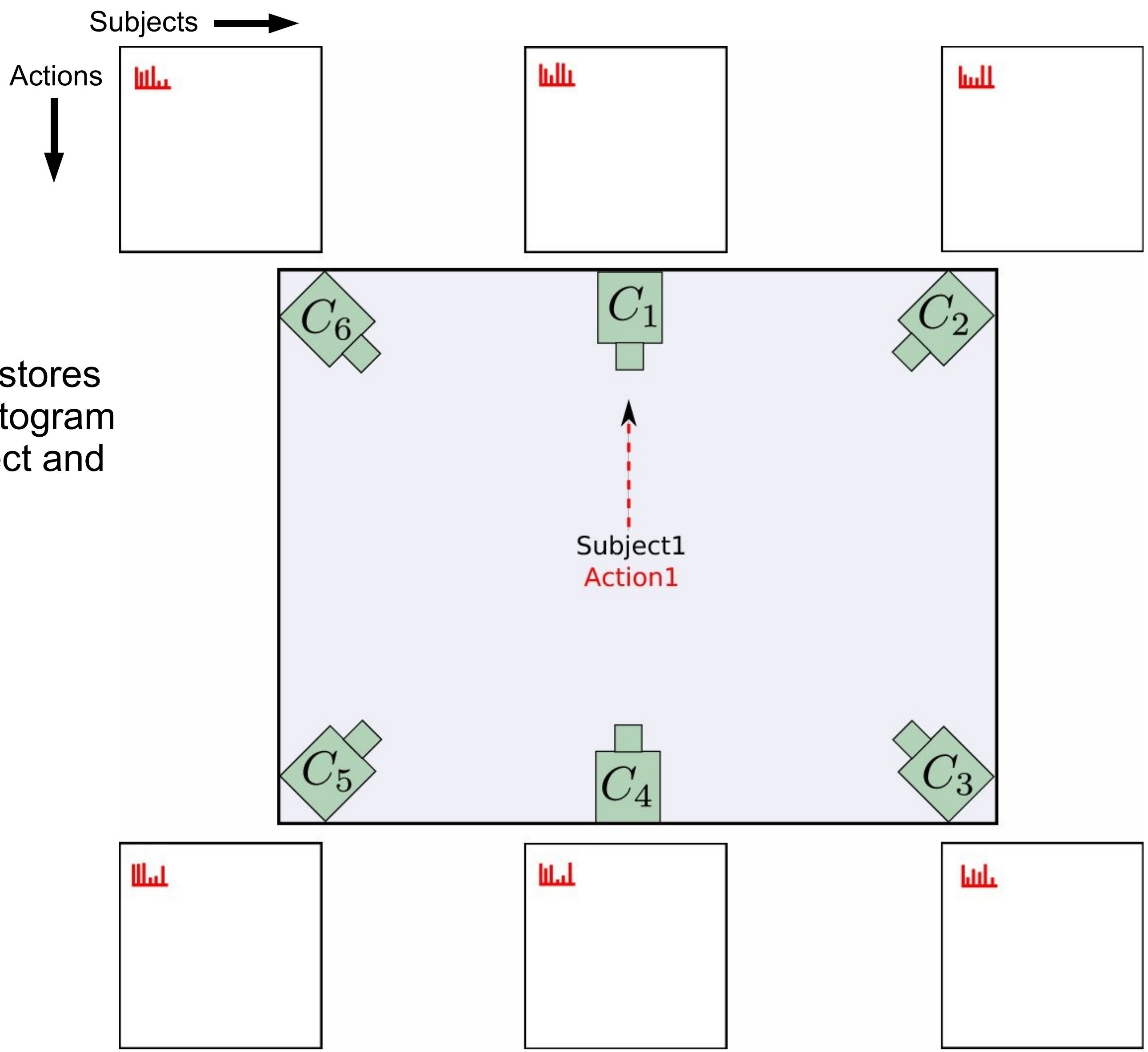
Subjects face camera C_1 while performing actions

Learning Multi-view action histograms

Training begins:
M subjects,
C action categories

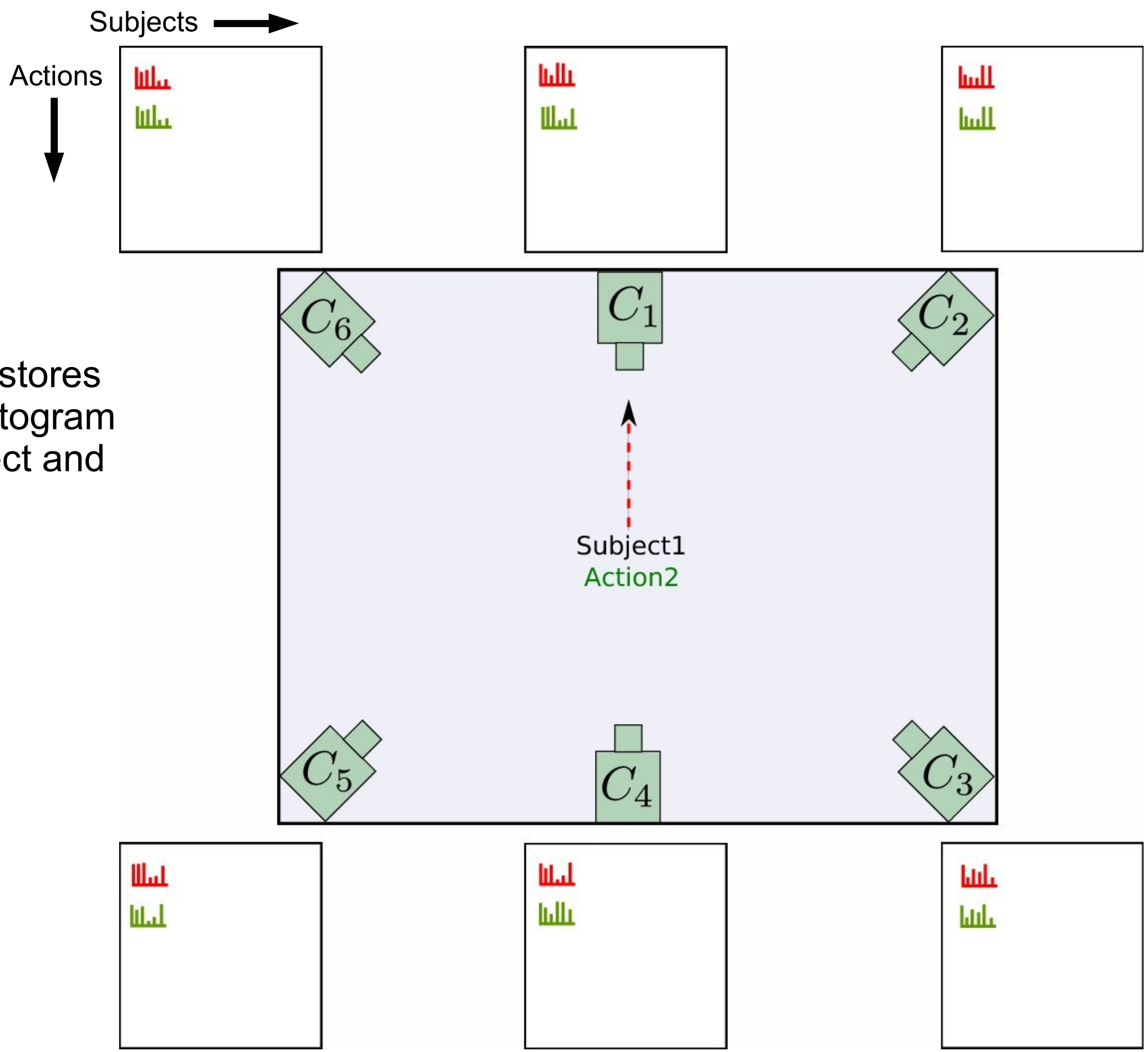


Learning Multi-view action histograms



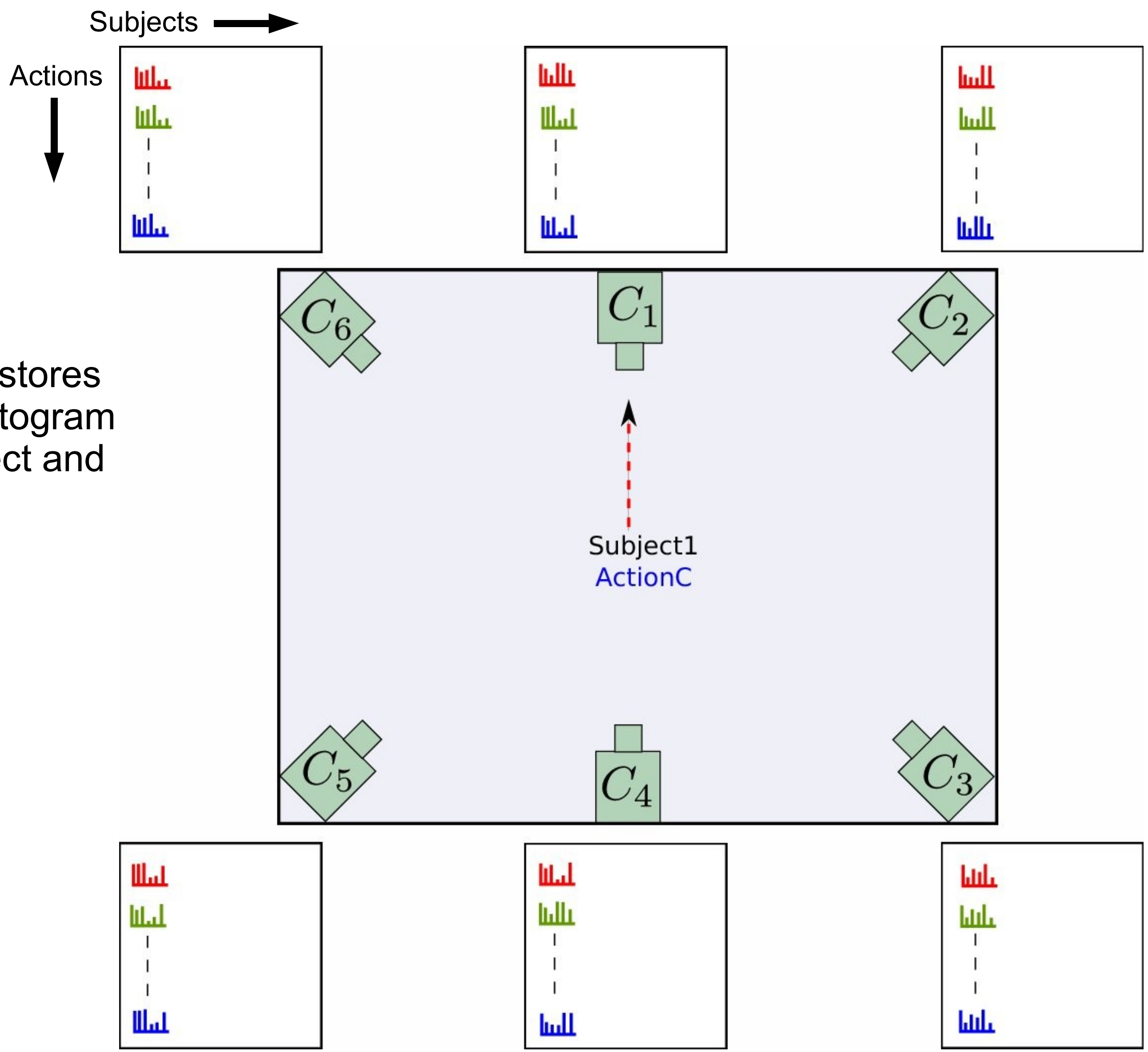
Each camera stores one action histogram for each subject and each action.

Learning Multi-view action histograms



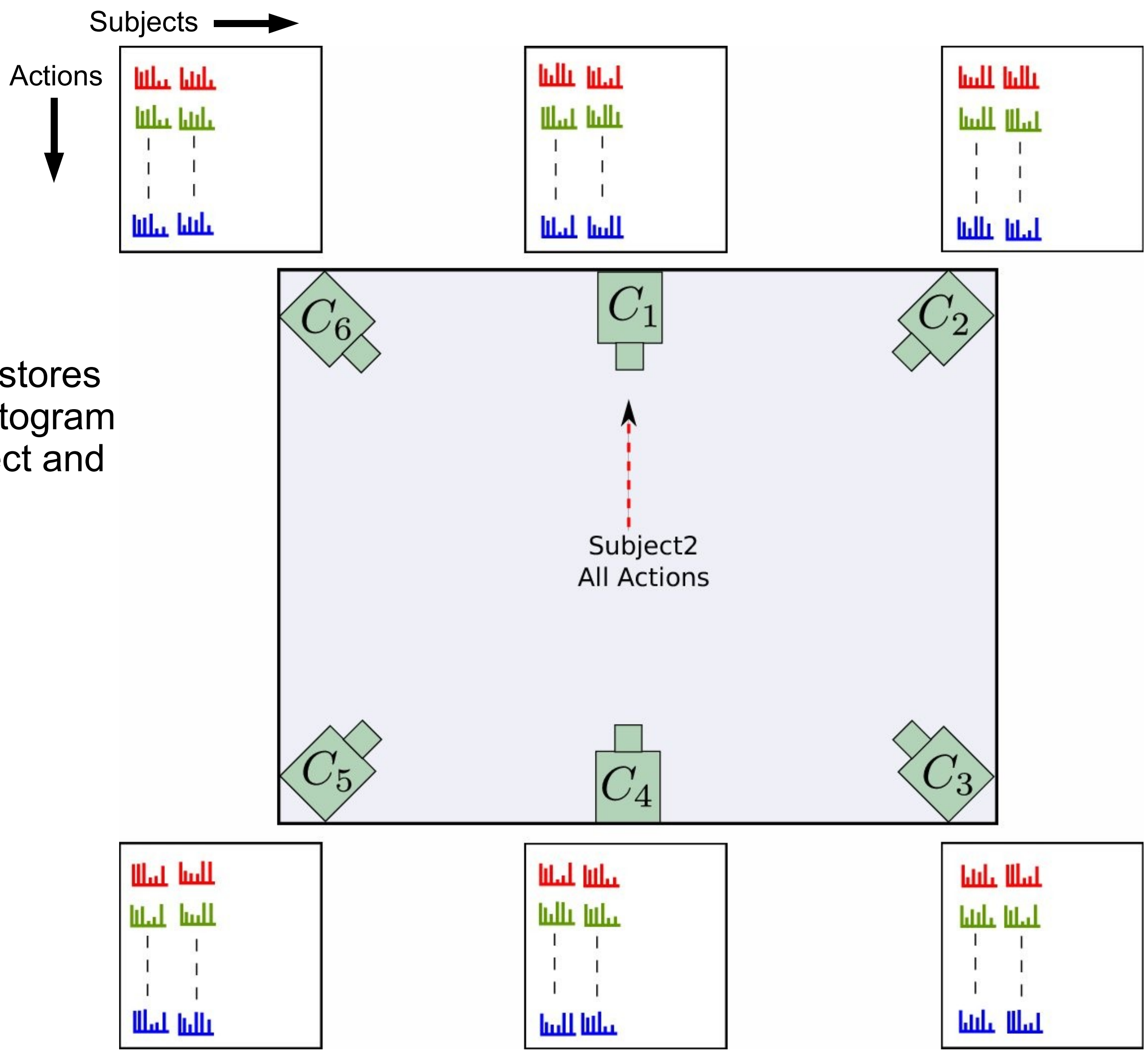
Each camera stores one action histogram for each subject and each action.

Learning Multi-view action histograms



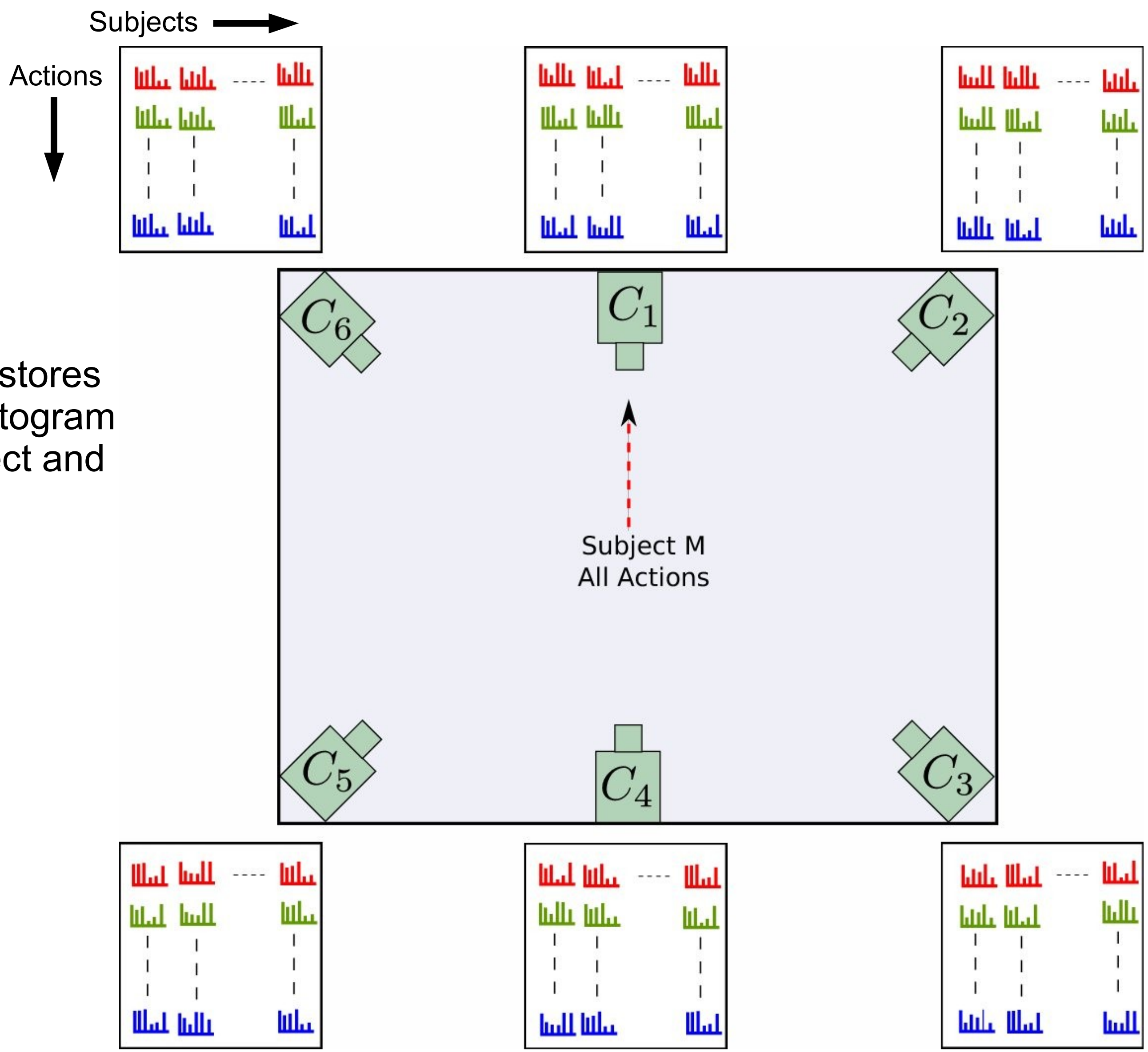
Each camera stores one action histogram for each subject and each action.

Learning Multi-view action histograms



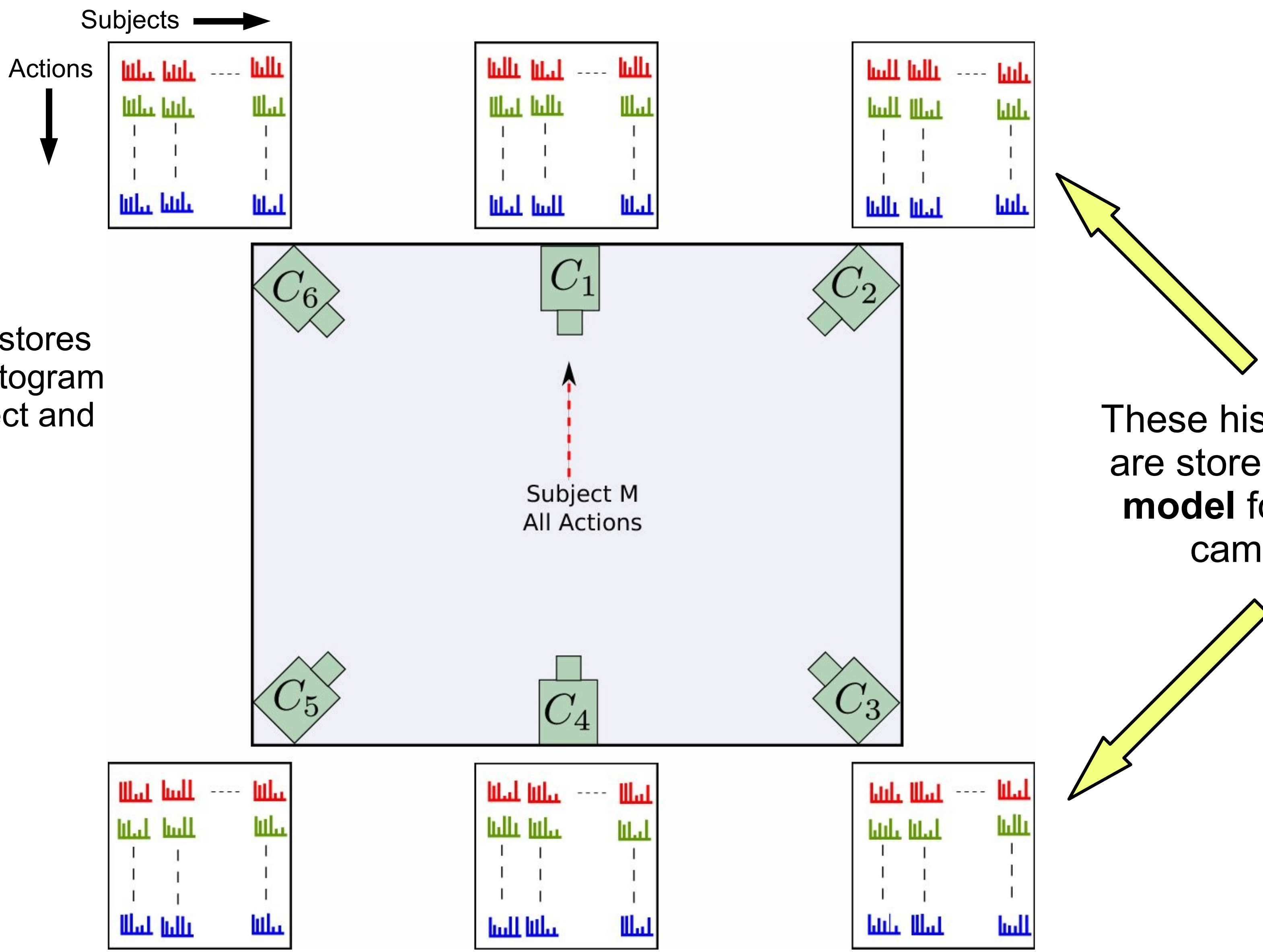
Each camera stores one action histogram for each subject and each action.

Learning Multi-view action histograms



Each camera stores one action histogram for each subject and each action.

Learning Multi-view action histograms



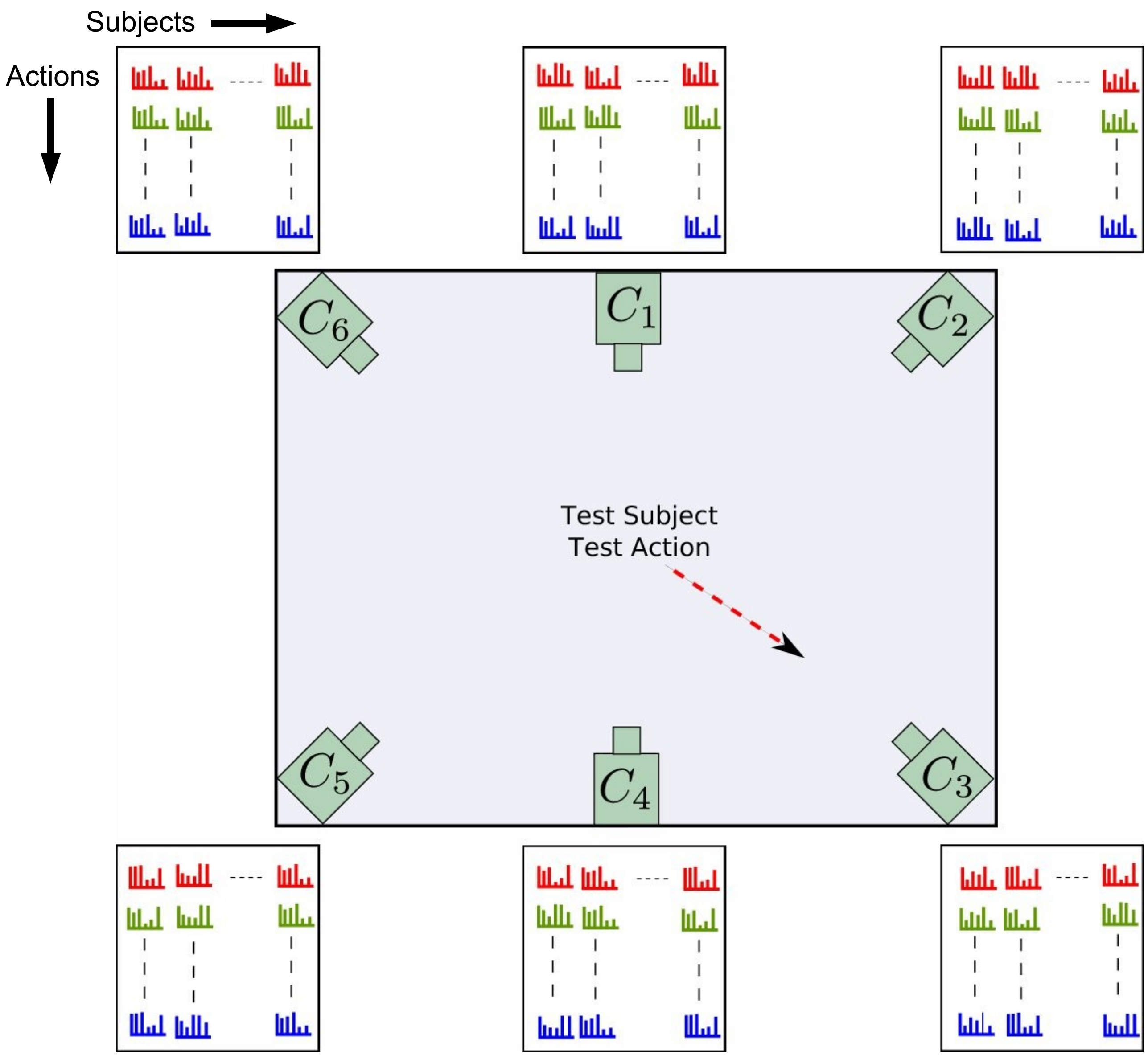
Each camera stores one action histogram for each subject and each action.

These histograms are stored as the **model** for each camera

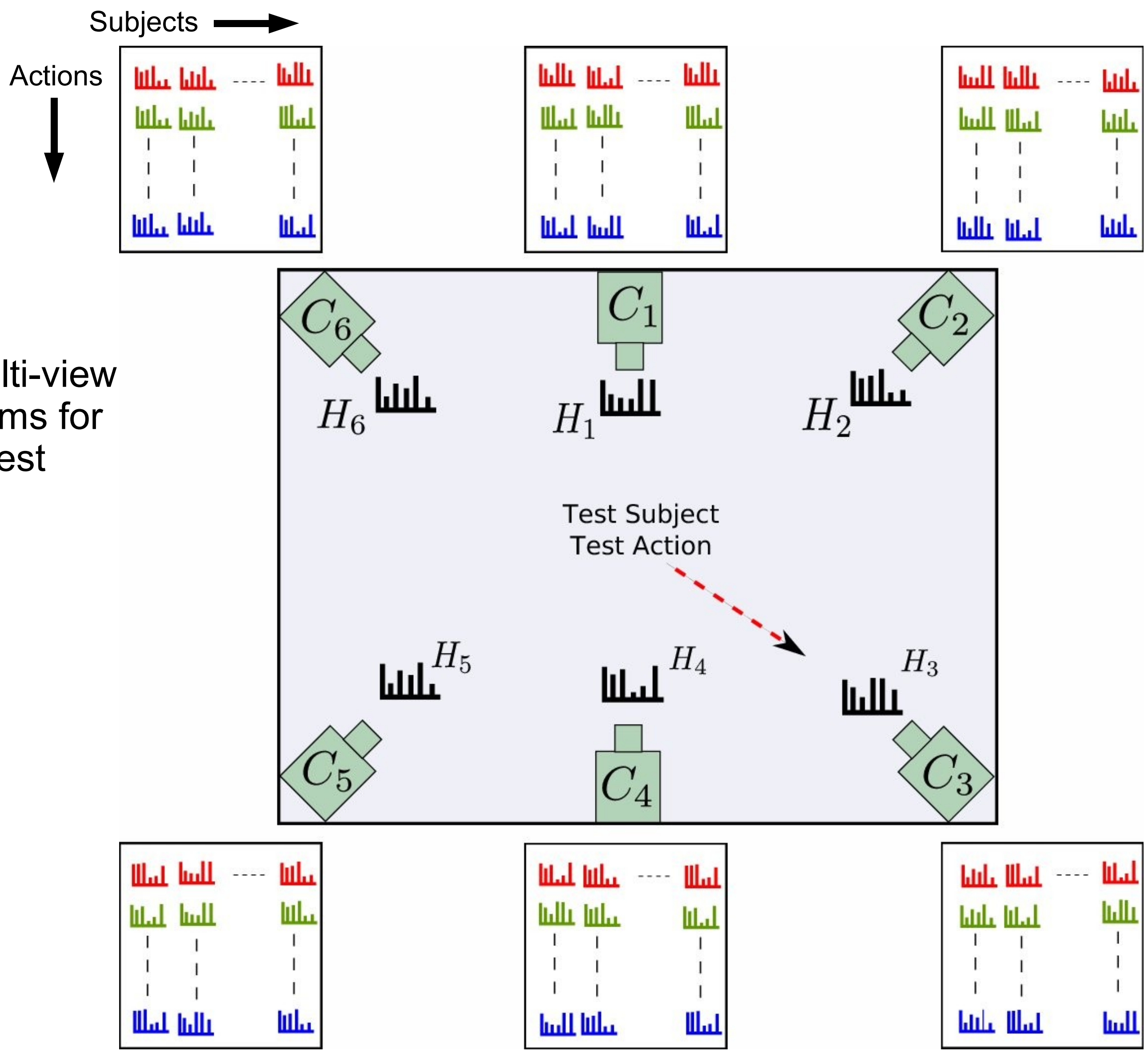
Multi-view Action Classification: Testing Stage

Subject may face any camera while performing actions. As an example, she may face camera C_3 or C_5

Test Subject facing camera C_3

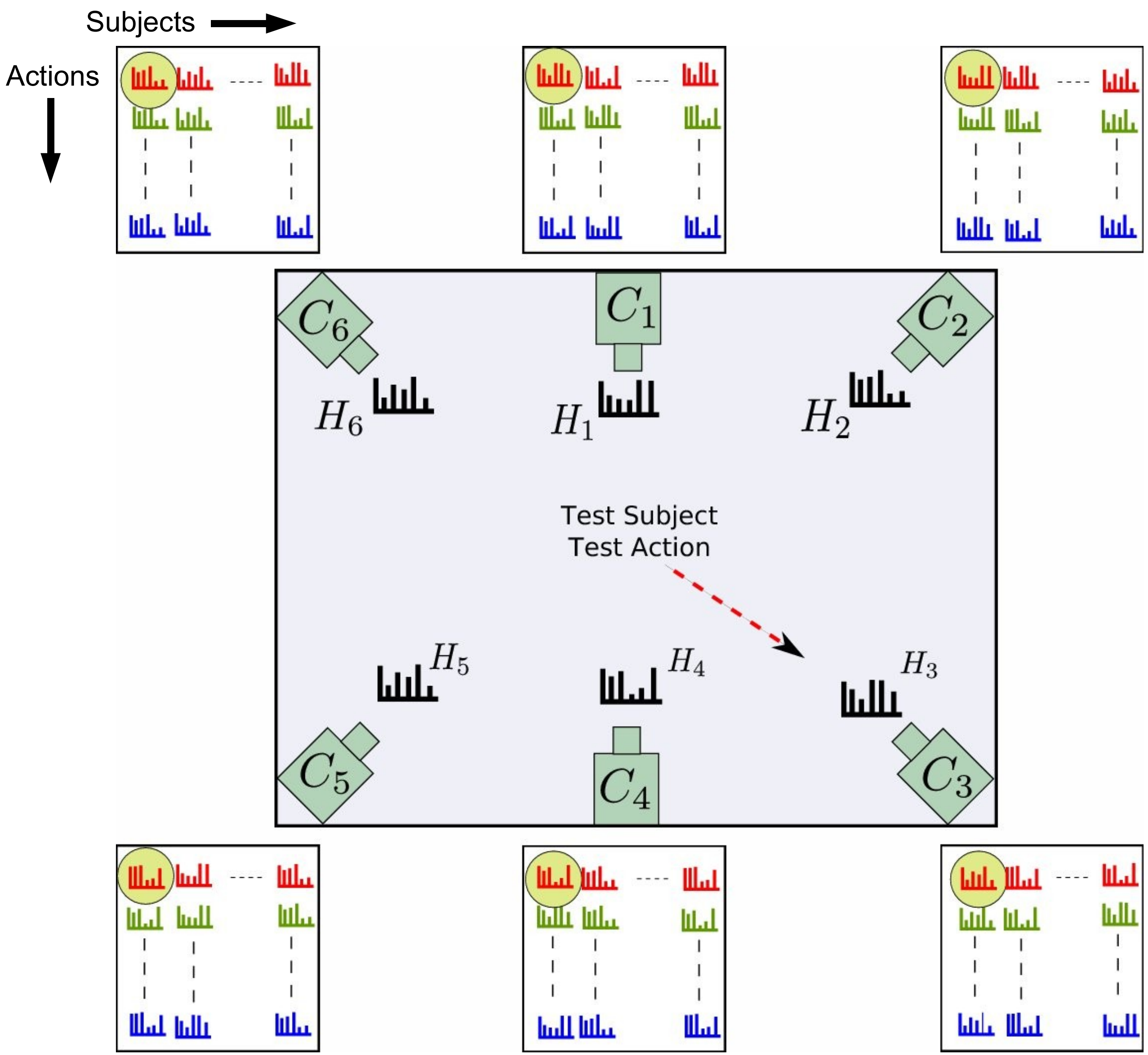


Test Subject facing camera C_3

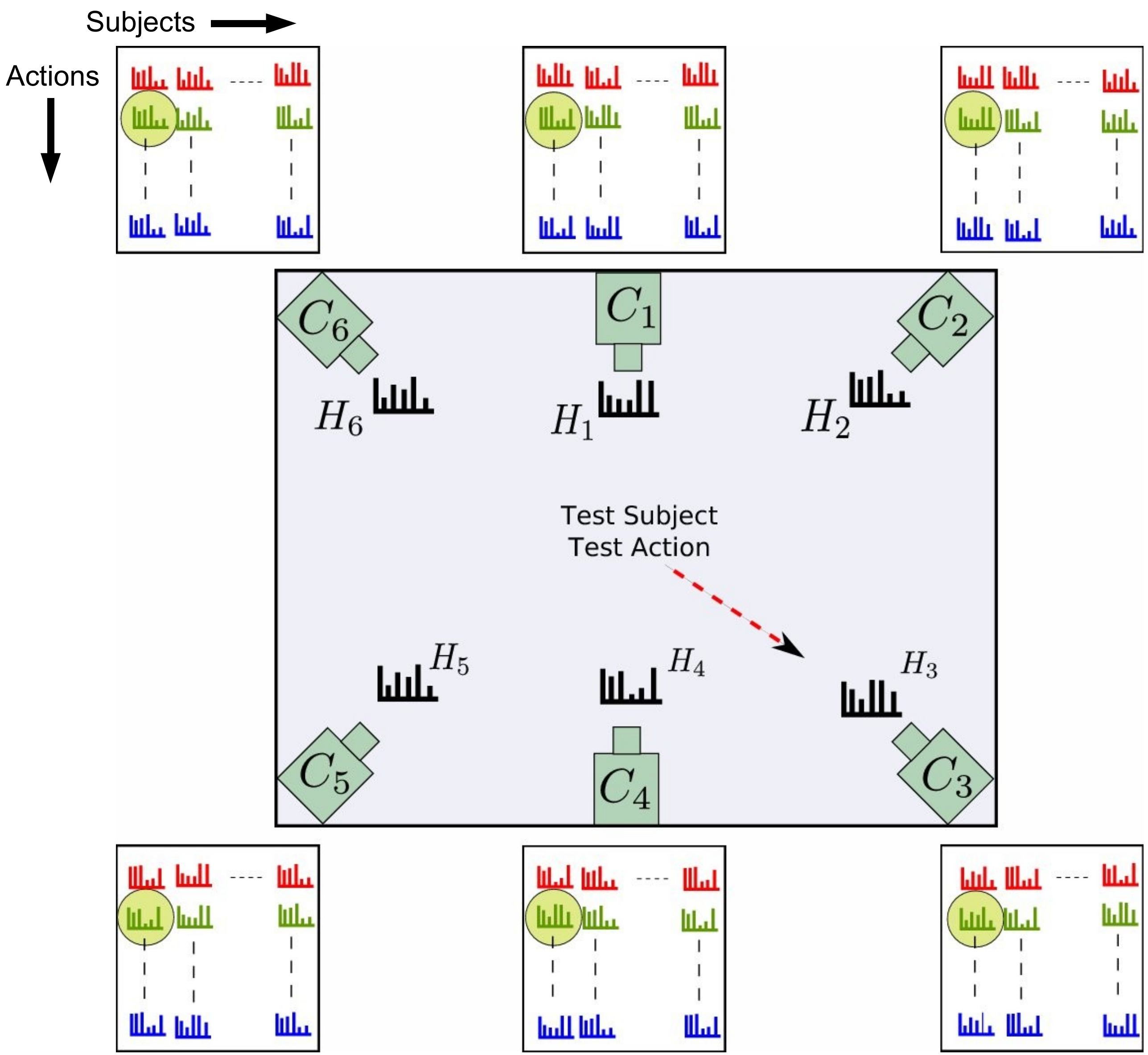


Generating multi-view action histograms for test subject – test action.

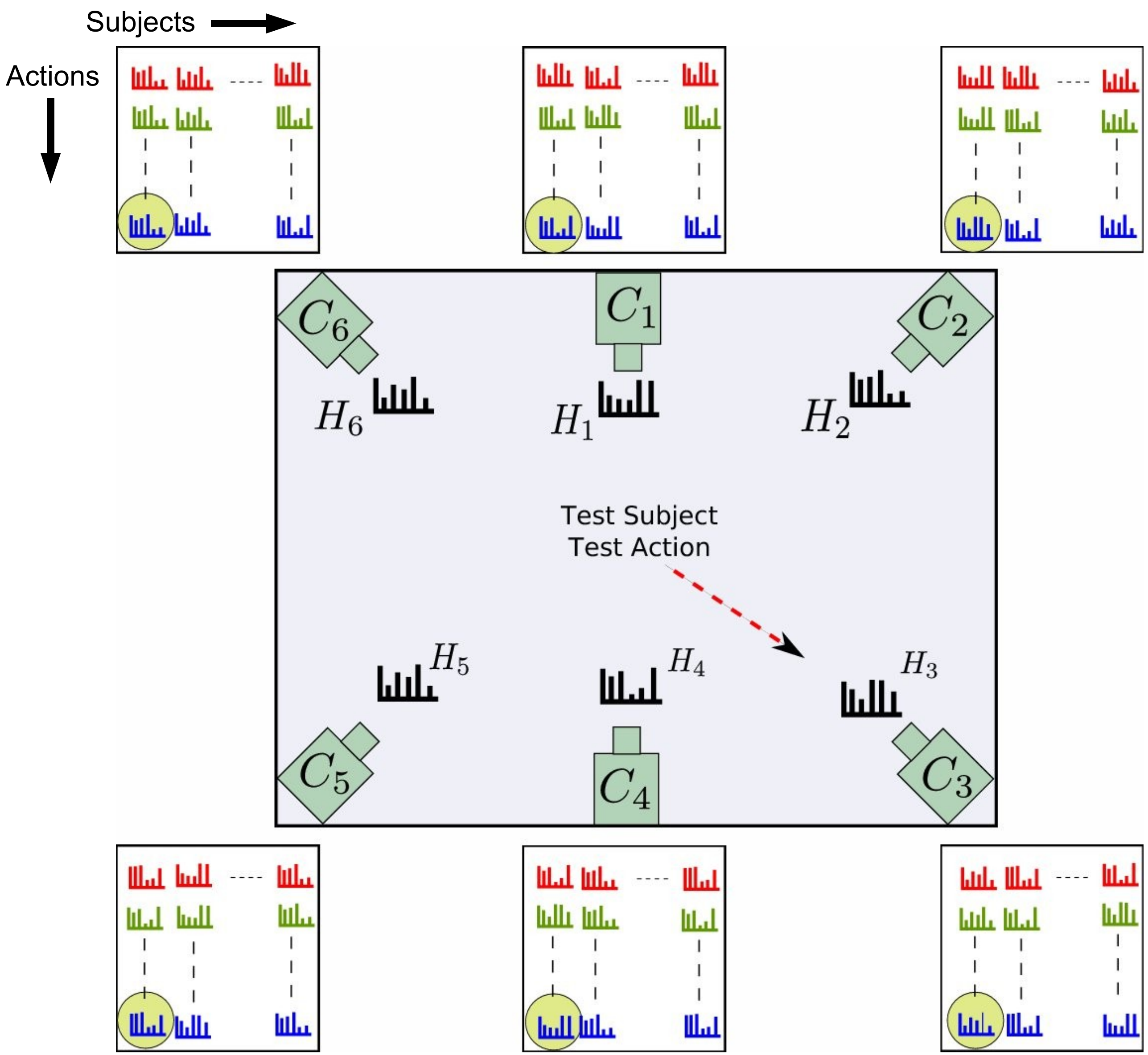
Finding the best match



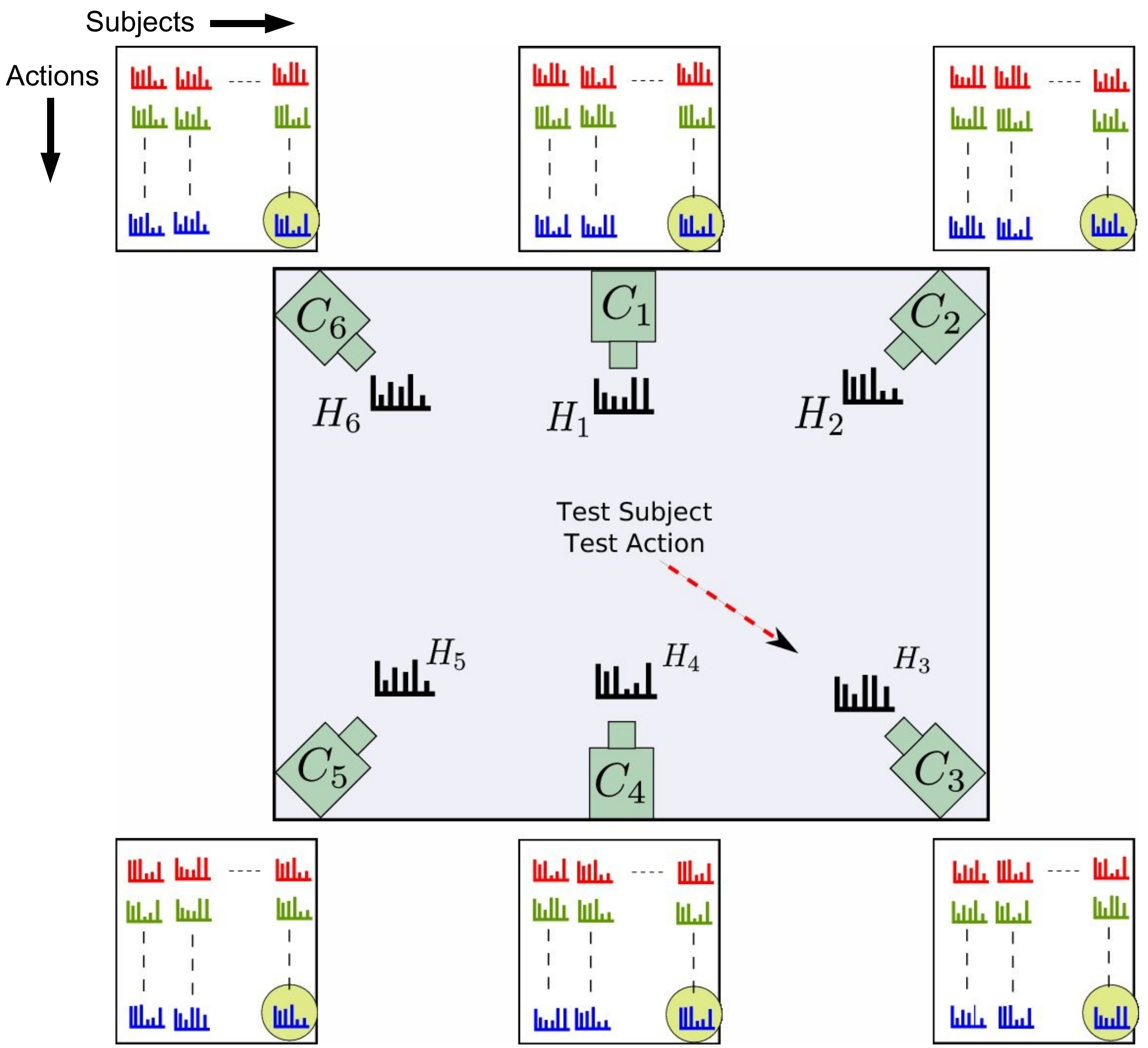
Finding the best match



Finding the best match

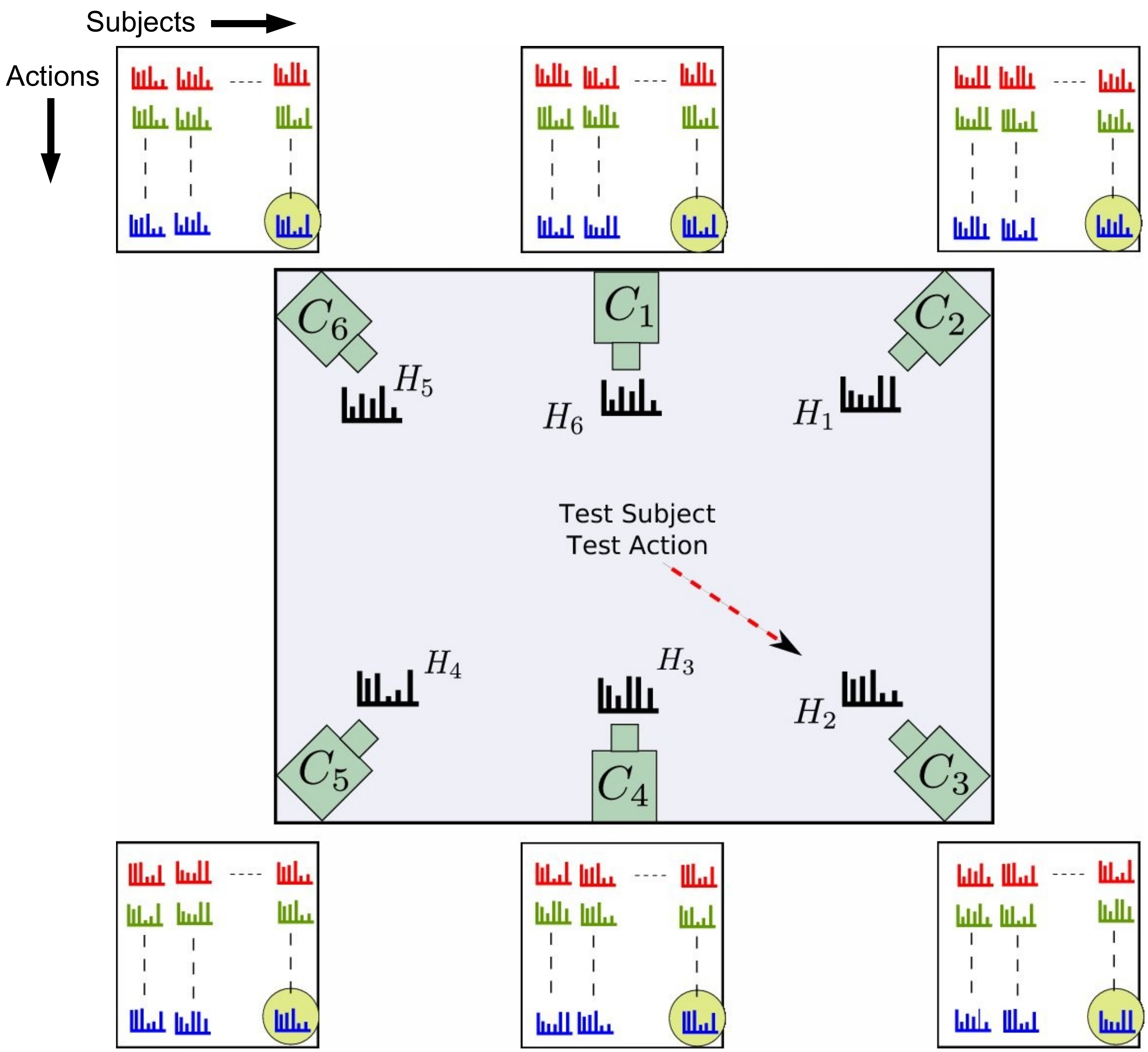


Finding the best match



d_0

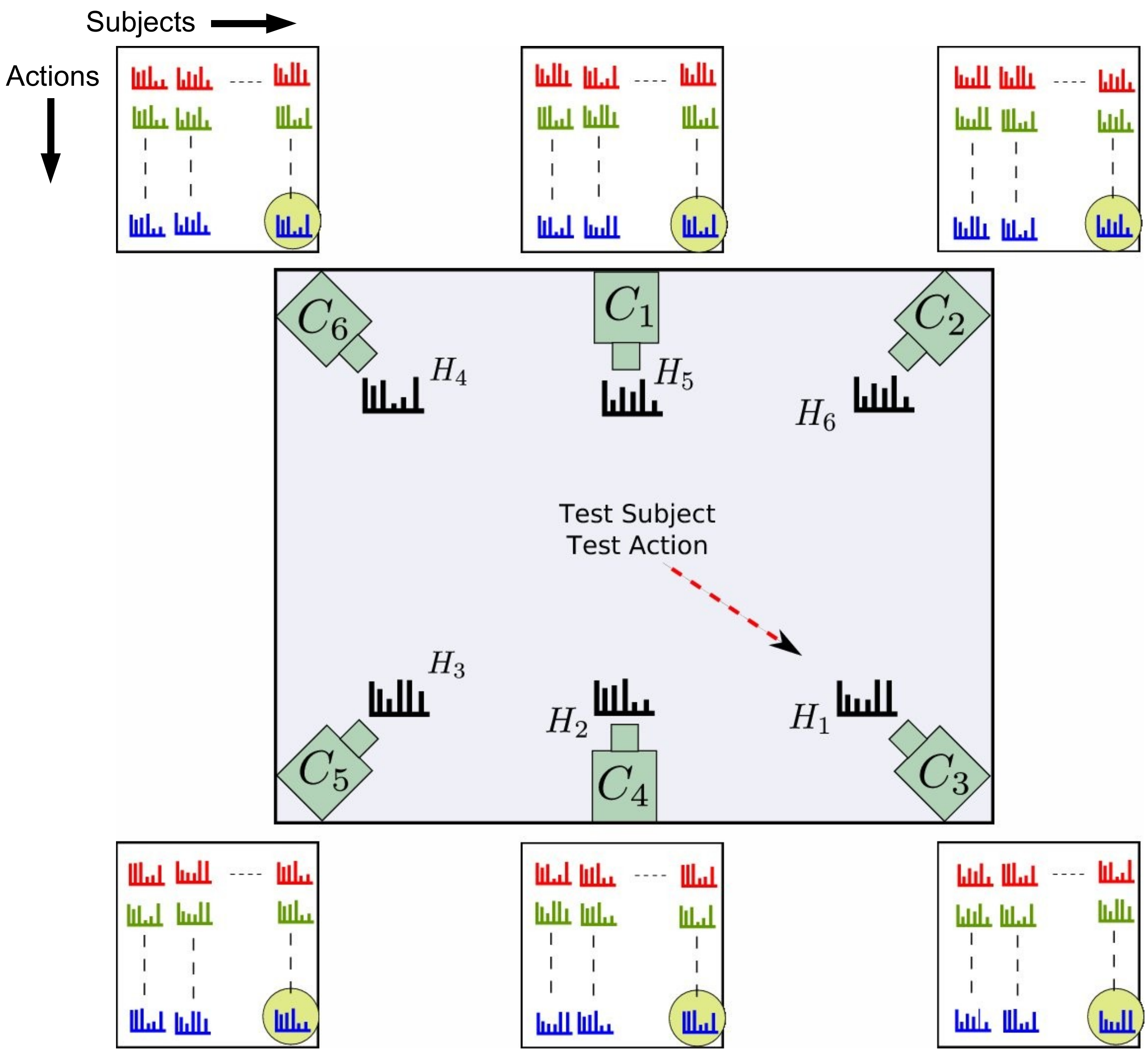
Finding the best match



d_0

d_1

Finding the best match

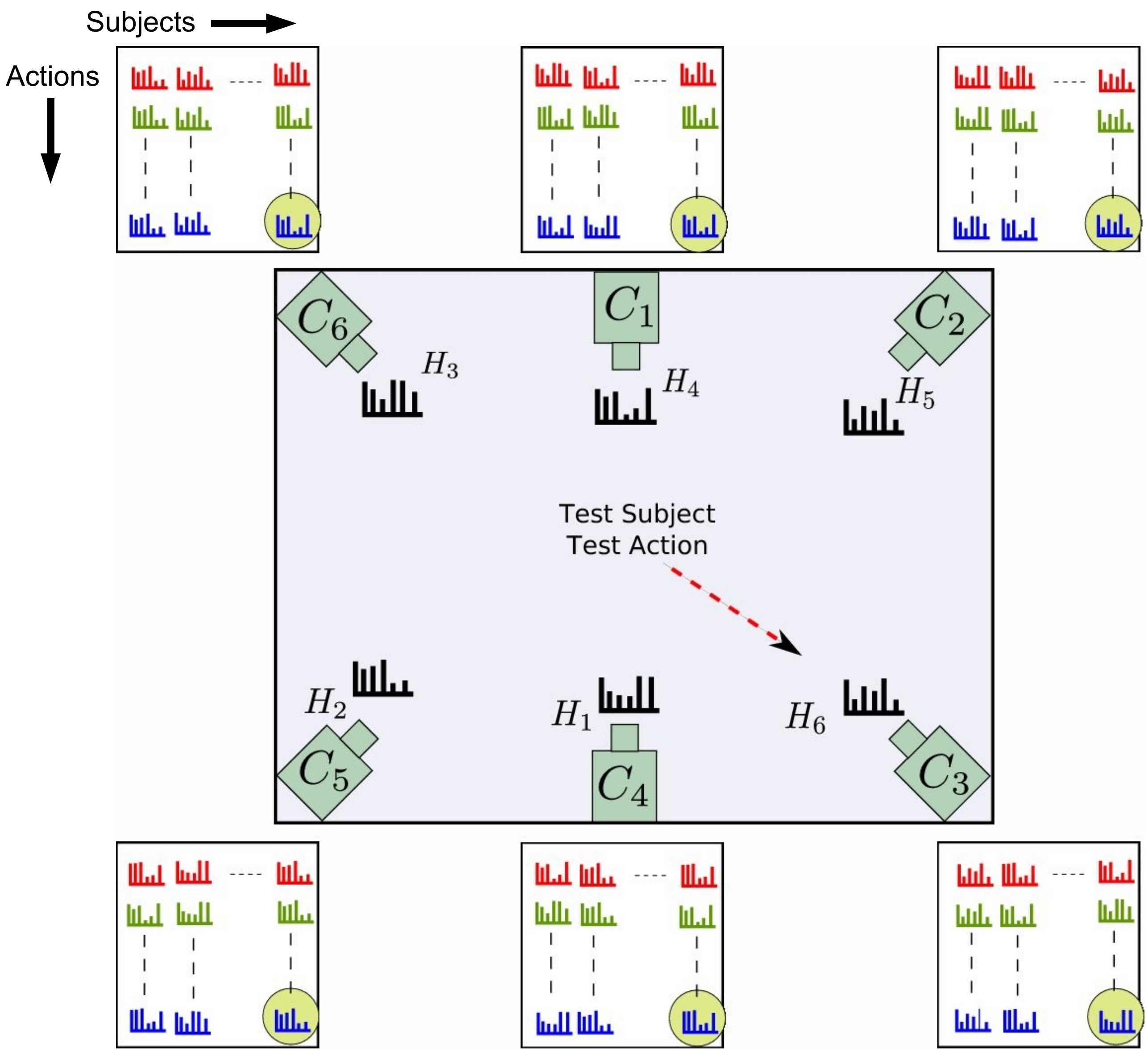


d_0

d_1

d_2

Finding the best match



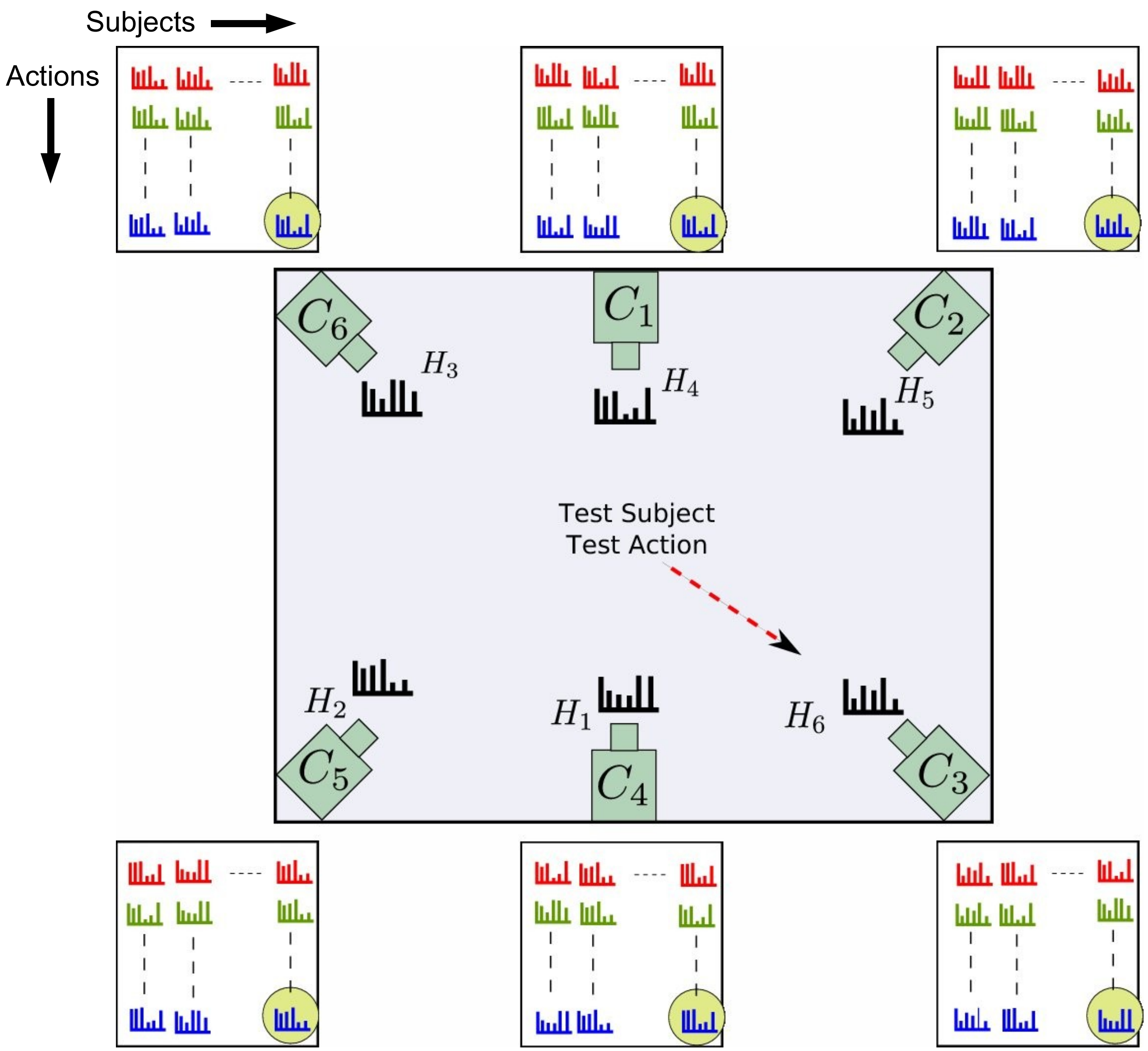
d_0

d_1

d_2

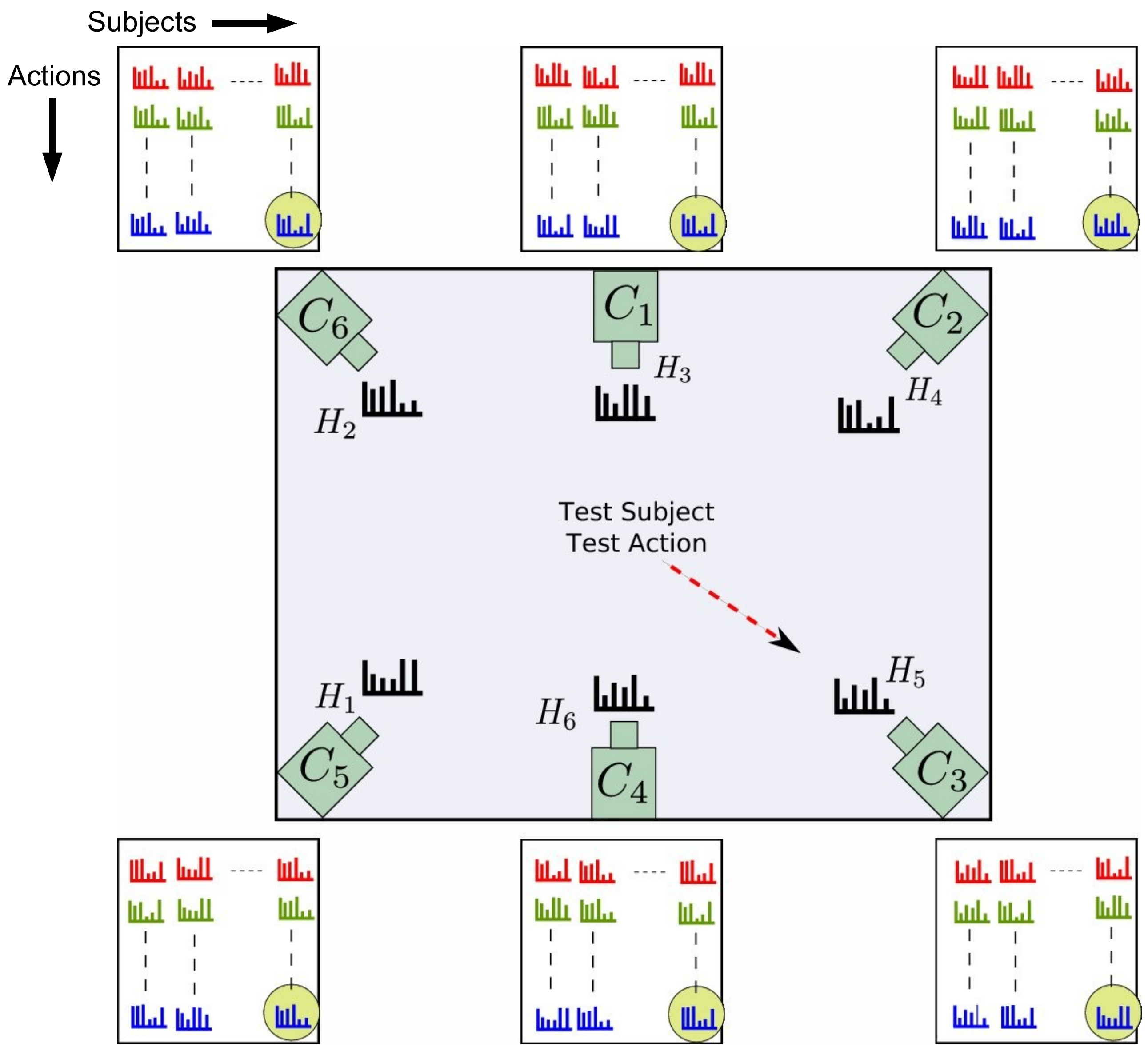
d_3

Finding the best match



d_0
d_1
d_2
d_3
d_4

Finding the best match



d_0

d_1

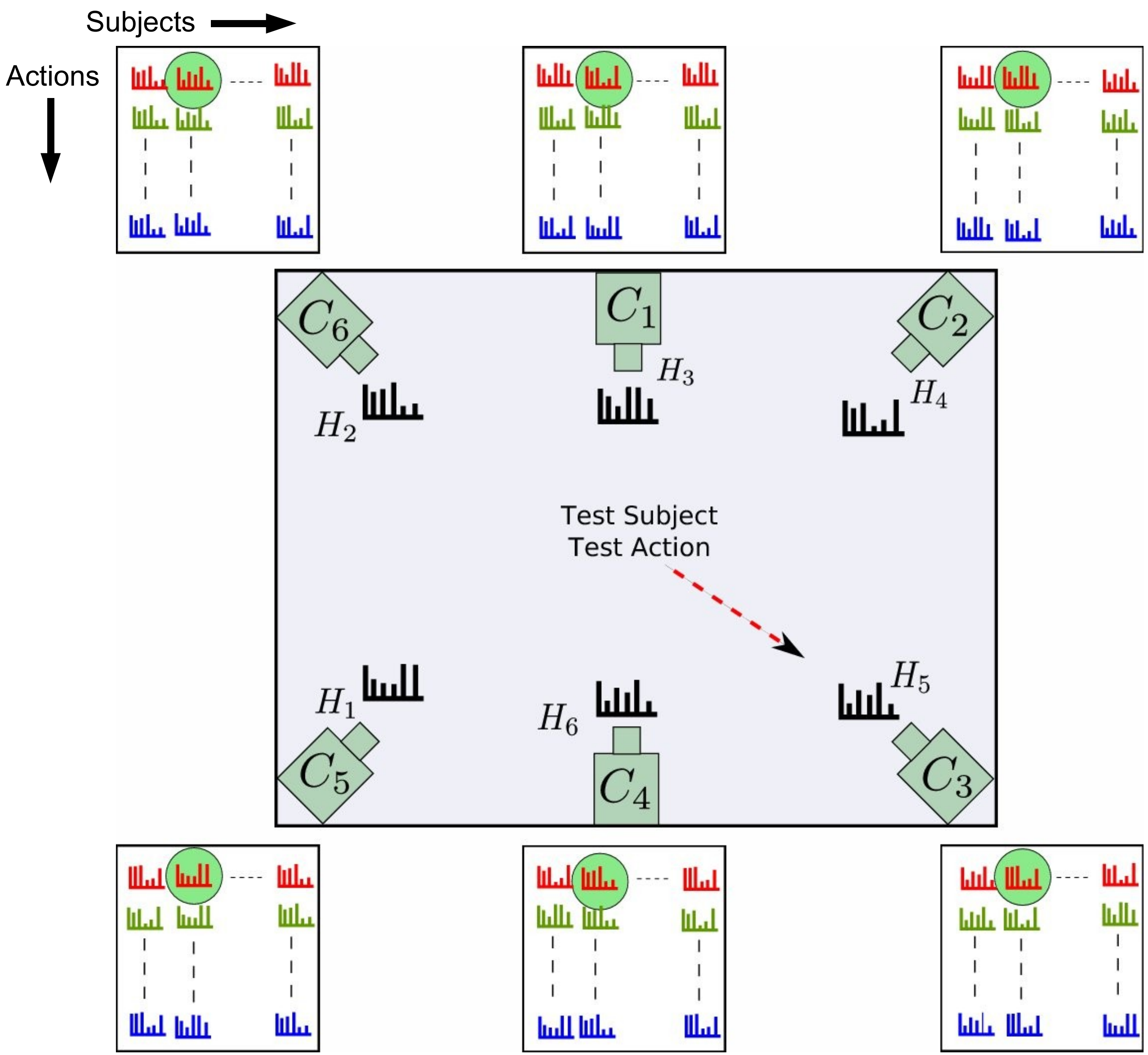
d_2

d_3

d_4

d_5

Best match found



d_0

d_1

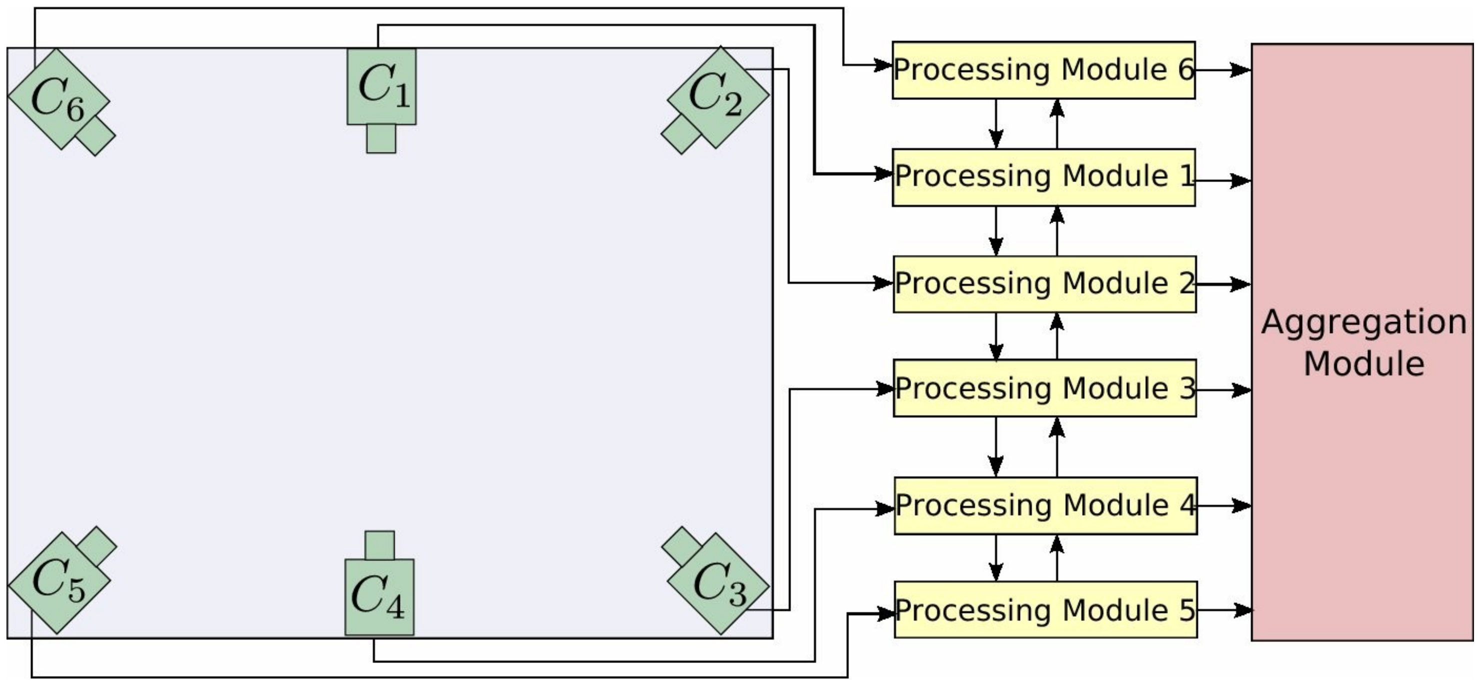
d_2

d_3

d_4

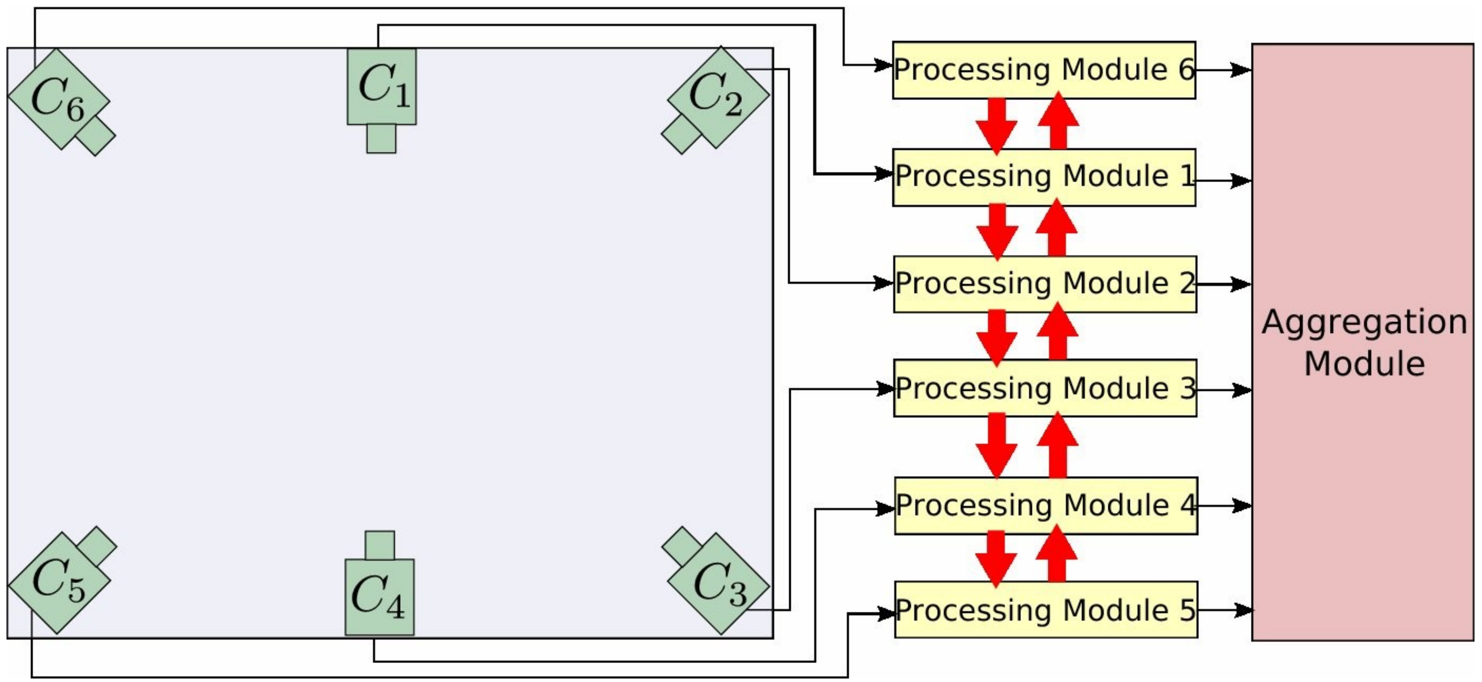
d_5

Suitability for Distributed Implementation



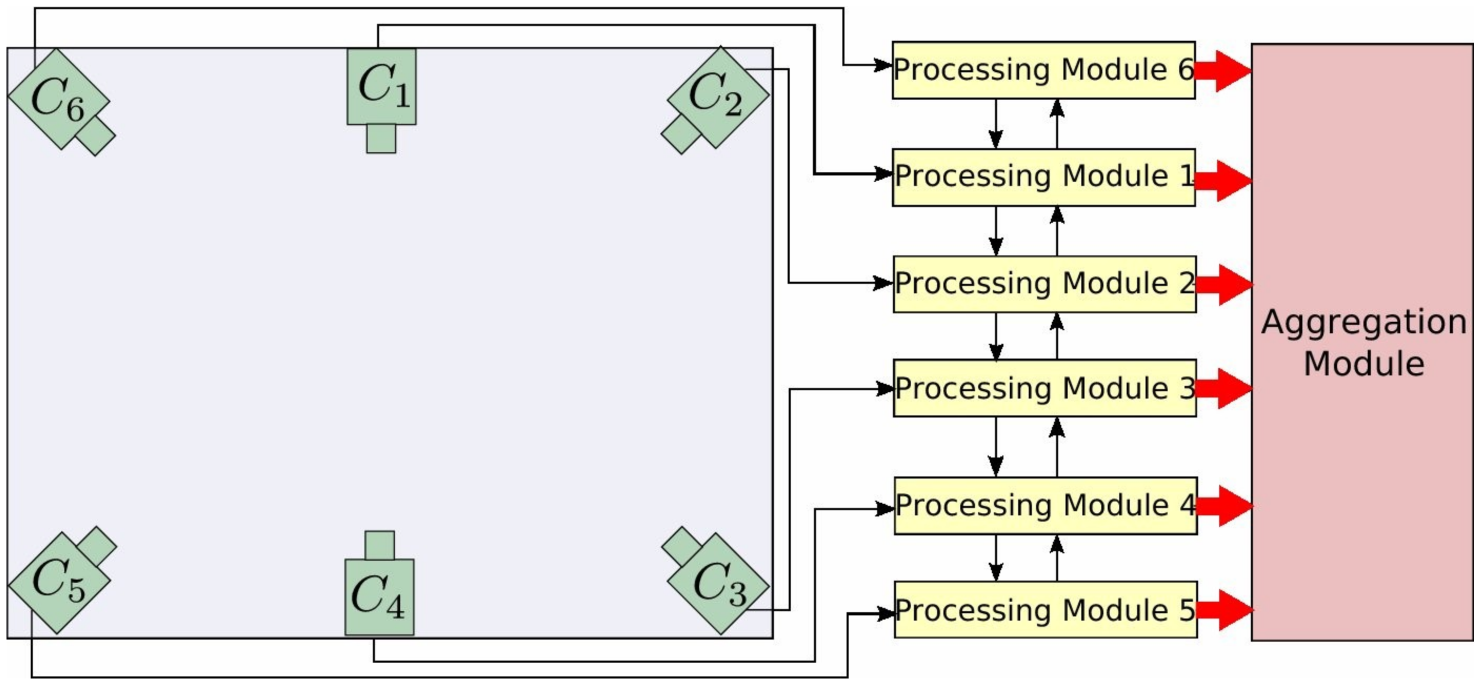
- For any particular circular shift, the histogram distances can be computed parallelly by the cameras.

Suitability for Distributed Implementation



- For any particular circular shift, the histogram distances can be computed parallely by the cameras.
- Circular shifts can be implemented by each camera broadcasting its histograms.

Suitability for Distributed Implementation



- For any particular circular shift, the histogram distances can be computed parallely by the cameras.
- Circular shifts can be implemented by each camera broadcasting its histograms.
- The histogram distances from the individual cameras can be transmitted to the aggregation module for finding the best matching training histograms corresponding to test histograms.

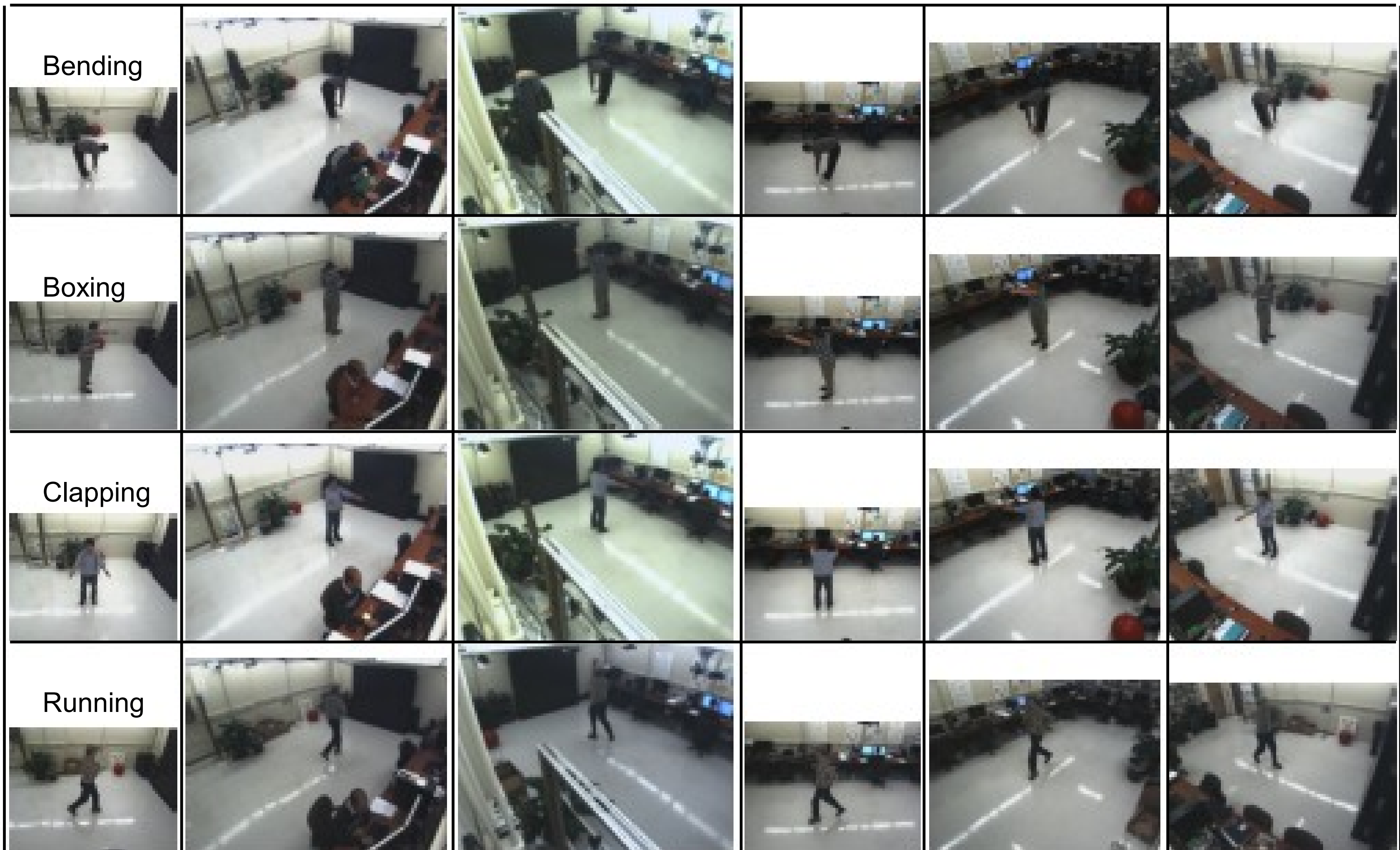
Experiments and Results

- Two multi-view multi-action datasets:
  - Purdue Dataset
    - 12 subjects, 9 action classes, 6 cameras
  - IXMAS Dataset (Weinland et al. 2007)
    - 10 subjects, 11 action classes, 4 cameras
- Action Classification using 1-NN.
- Leave-one-out cross validation

Purdue Dataset

- 3 experimental scenarios:
  - Multi camera training, multi camera testing (MM)
  - Multi camera training, single camera testing (MS)
  - Single camera training, single camera testing (SS)

Purdue Dataset



Classification Results (Purdue Dataset)

Error rate = 15.3704 %

	Walking	Running	Bending	Kicking	SitDown	StandUp	HandWaving	HandClapping	Boxing
Walking	.75	.25	.00	.00	.00	.00	.00	.00	.00
Running	.05	.95	.00	.00	.00	.00	.00	.00	.00
Bending	.00	.00	1.0	.00	.00	.00	.00	.00	.00
Kicking	.00	.00	.02	.98	.00	.00	.00	.00	.00
SitDown	.00	.00	.03	.00	.97	.00	.00	.00	.00
StandUp	.00	.00	.00	.00	.00	1.0	.00	.00	.00
HandWaving	.00	.00	.00	.00	.00	.00	.82	.18	.00
HandClapping	.00	.00	.03	.00	.00	.00	.18	.60	.18
Boxing	.00	.00	.07	.28	.00	.02	.00	.08	.55

Confusion Matrix

(Multi camera training, multi camera testing)

Classification Results (Purdue Dataset)

	Multi View Testing	Single View Testing
Multi View Training	84.6	82.96
Single View Training (Frontal)	N/A	78.89

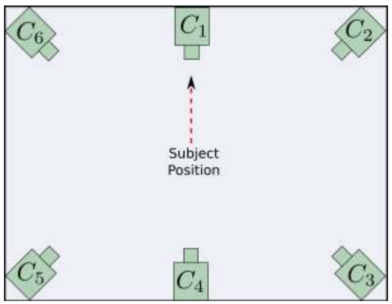
- Classifcation accuracy: MM > MS > SS

Classification Results (Purdue Dataset)

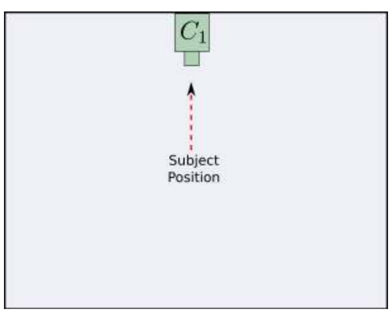
	Single View Testing		
	Left	Front	Right
Multi View Training	73.18	82.96	64.82
Single View Training (Frontal)	56.48	78.89	45.37

- Single View Testing: Front view accuracy > side view accuracy

Training

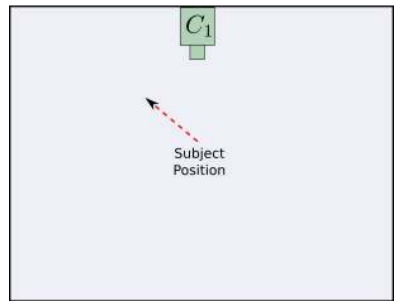


Multi-view

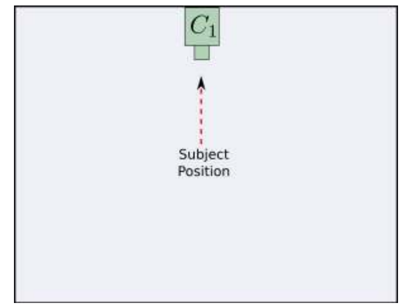


Single-view (Frontal)

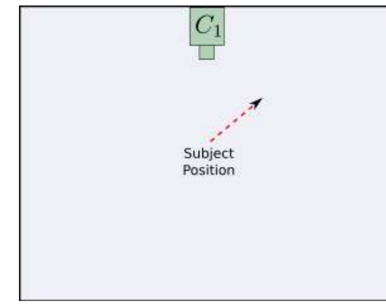
Testing (Single View)



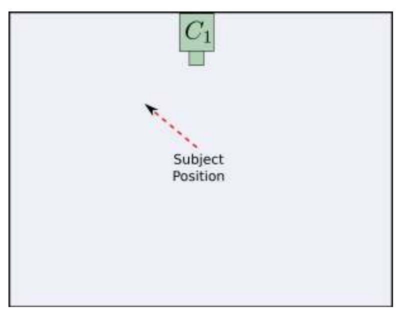
Single-view (Left)
73.18 %



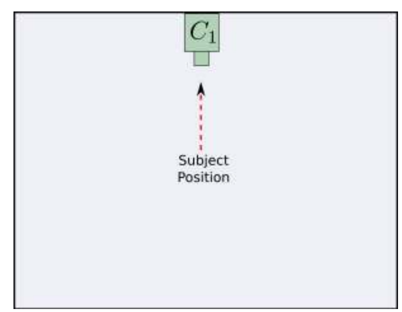
Single-view (Frontal)
82.96 %



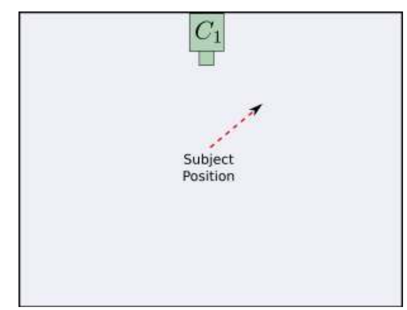
Single-view (Right)
64.82 %



Single-view (Left)
56.48 %

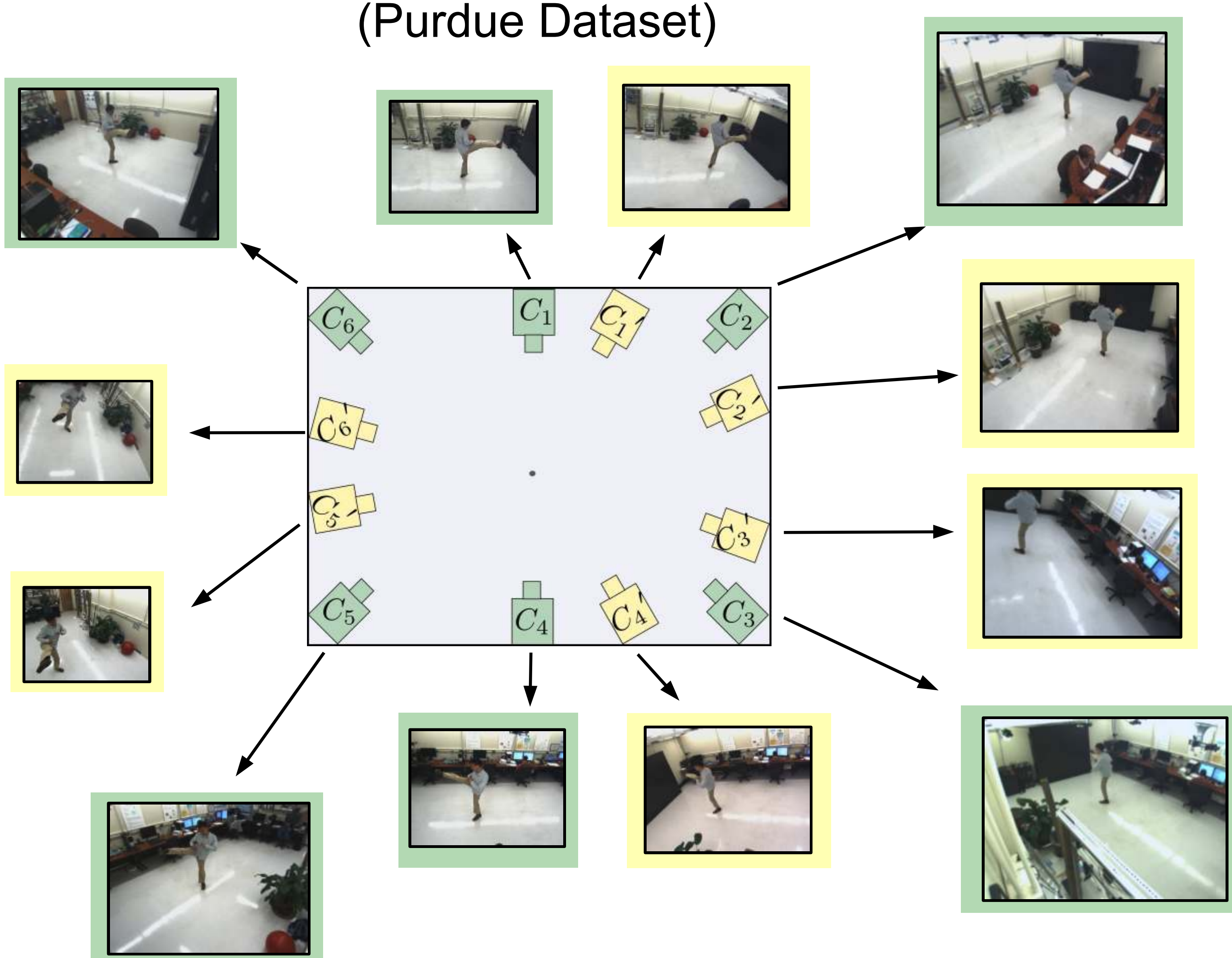


Single-view (Frontal)
78.89 %



Single-view (Right)
45.37 %

Classification for Previously Unseen Views
(Purdue Dataset)

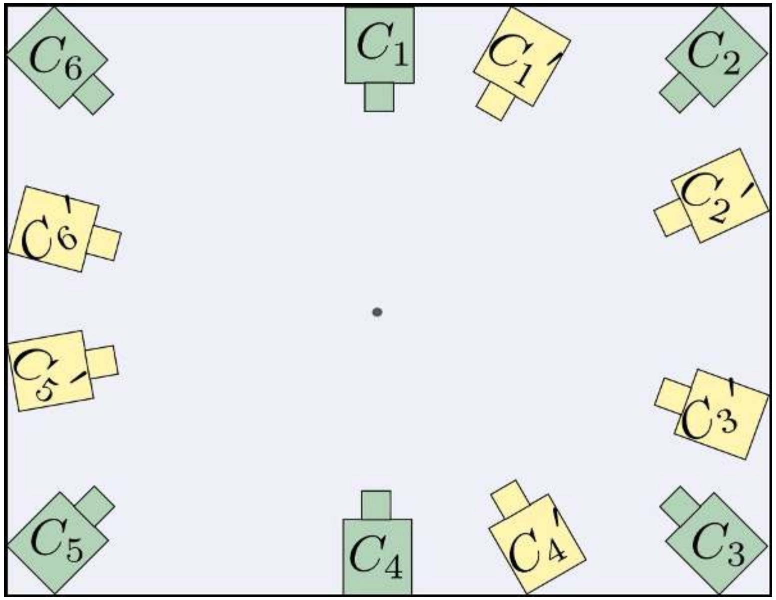


Images from different camera views were scaled differently so that the human actor has the same height in all the images.

Green cameras used during training stage

Yellow cameras used during testing stage

Classification for Previously Unseen Views
(Purdue Dataset)



# Previously unseen views	Classification Accuracy
1	83.70%
2	82.78%
3	82.22%
4	83.52%
5	78.70%
6	76.30%

IXMAS Dataset: Comparing 3 Algorithms

10 subjects, 11 action classes, 4 cameras

# Cameras in testing stage	Proposed approach	Weinland et al. (2007)	Yan et al. (2008)
4	81.40%	81.30%	78.00%
3	79.10%	70.20%	60.00%
2	75.60%	81.30%	71.00%
1	69.10%	Not reported	Not reported

Average Classification Accuracy
(as a function of number of cameras used)

Advantages for Distributed Implementation

- Comparative analysis with Weinland et al.
- Based on IXMAS dataset.
- Memory requirements:
  - Weinland et al. 1.72 Mbytes
  - Proposed approach 0.293 Mbytes / camera
- Communication Bandwidth requirements:
  - For transmitting full images (390x291 resolution, 23 fps): 30 Mbytes/s
  - Weinland et al. (transmit silhouette information): 47.1 Kbytes/s
  - Proposed approach (transmit histogram distance values): 2.7 Kbytes/s

Conclusions

- Proposed a Multi-camera orientation invariant action classification algorithm:
  - Based on simple histogram features
  - Training and testing stages are simple
  - Lightweight features $\Longrightarrow$ low memory and bandwidth requirements.
  - Algorithm suitable for distributed implementation due to simple computations, low resource requirements and modular operation.

References

- Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S., "Behavior Recognition via Sparse Spatio-Temporal Features", ICCV VS-PETS 2005, Beijing, China.
- Weinland, D., Boyer, E., Ronfard, R., "Action Recognition from Arbitrary Views using 3D Exemplars", ICCV 2007, vol. 1, no. 7, pp. 14-21, Oct. 2007.
- Yan, P., Khan, S.M., Shah, M., "Learning 4D action feature models for arbitrary view action recognition", CVPR 2008, vol. 1, no. 7, pp. 23-28, June 2008

Thank You !