BOSTON UNIVERSITY

A RANDOM FOREST APPROACH TO SEGMENTING AND CLASSIFYING GESTURES

Ajjen Joshi, Camille Monnier, Margrit Betke, Stan Sclaroff

1. INTRODUCTION

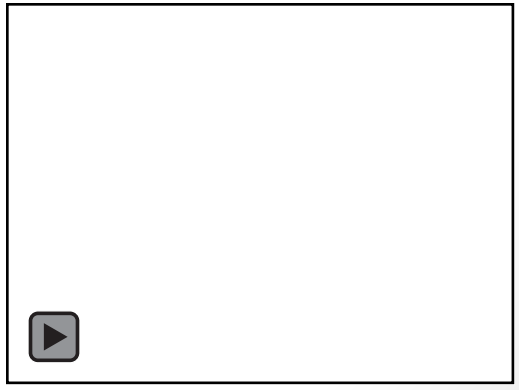
- Gesture: “A movement of part of the body, especially a hand or the head, to express an idea or meaning.”

Image reprinted from Google Image Search

Examples of specific applications

- Aircraft communication gesture recognition

- Recognition of socio-cultural gestures



Gesture spotting and recognition

- Gesture spotting

- Gesture recognition

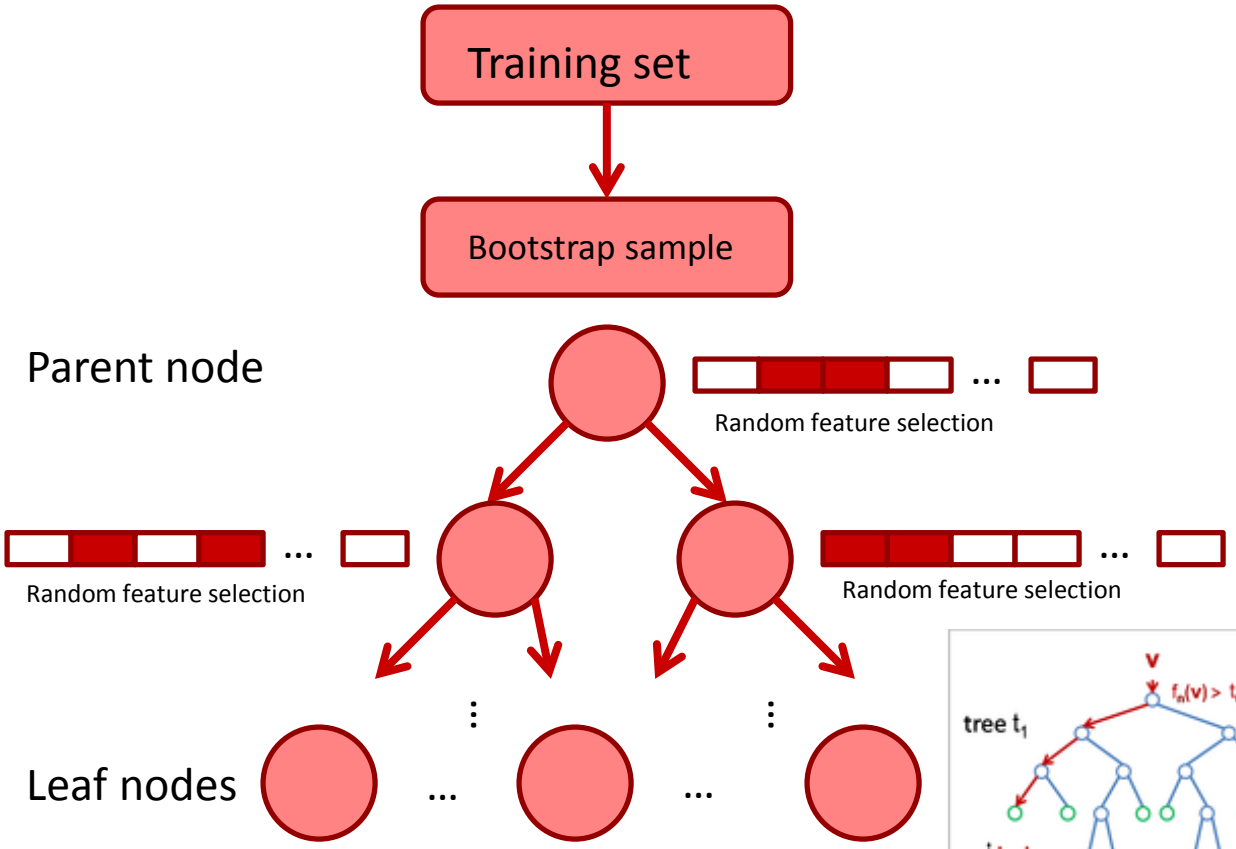
Problem Definition

- Problem:
  - Given a training set of multi-modal videos with multiple examples of all gestures in a gesture vocabulary, design a framework capable of automatically and accurately spotting and classifying gestures present in a set of test videos

Related work

- Generative Graphical Models
  - Hidden Markov Model
    - Starner et al., 1997
- Discriminative Graphical Models
  - Hidden Conditional Random Fields
    - Song et al., 2011
- Other Discriminative Models
  - Support Vector Machines
    - Huang et al., 2009
  - Tree Ensembles
    - (ours)

2. RANDOM FORESTS



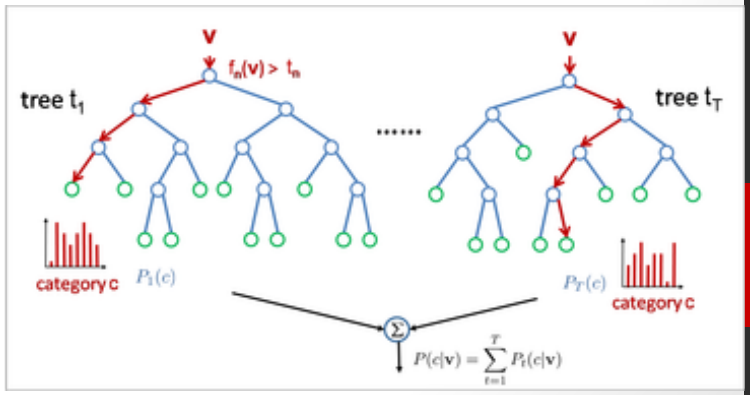


Image reprinted from:
Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." *Communications of the ACM* 56.1 (2013): 116-124.

Information Gain

- Given a training set *S* of data points and their labels, trees are built to optimize a certain function, e.g. Information gain (*I*)

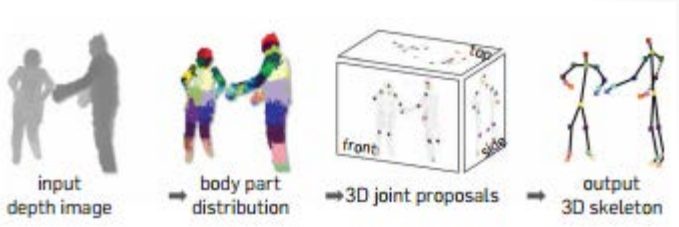
$$I_j = H(S_j) - \sum_{k \in (L,R)} \frac{|S_j^k|}{|S|} H(S_j^k)$$

S_j: set of training points at node j
$H(S_j)$: Shannon entropy at node j before split
S_j^R: sets of points at right child of parent node *j* after split
S_j^L: sets of points at left child of parent node *j* after split

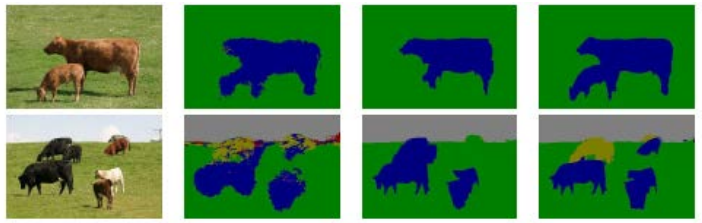
$$H(S) = -\sum_{c \in C} p_c \log(p_c)$$

S: set of training points
p_c: probability of a sample being in class *c*

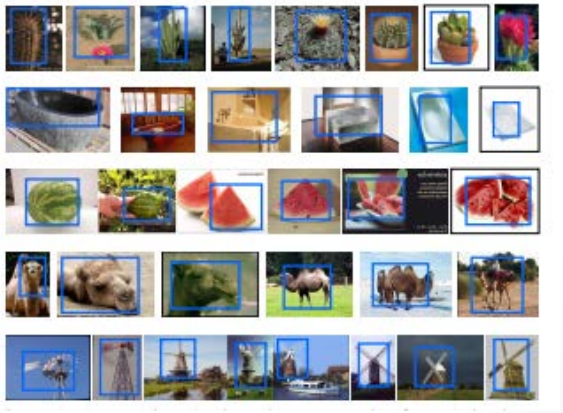
- Random forests have been used to good effect in:
  - human pose recognition
    - Shotton et al., 2013



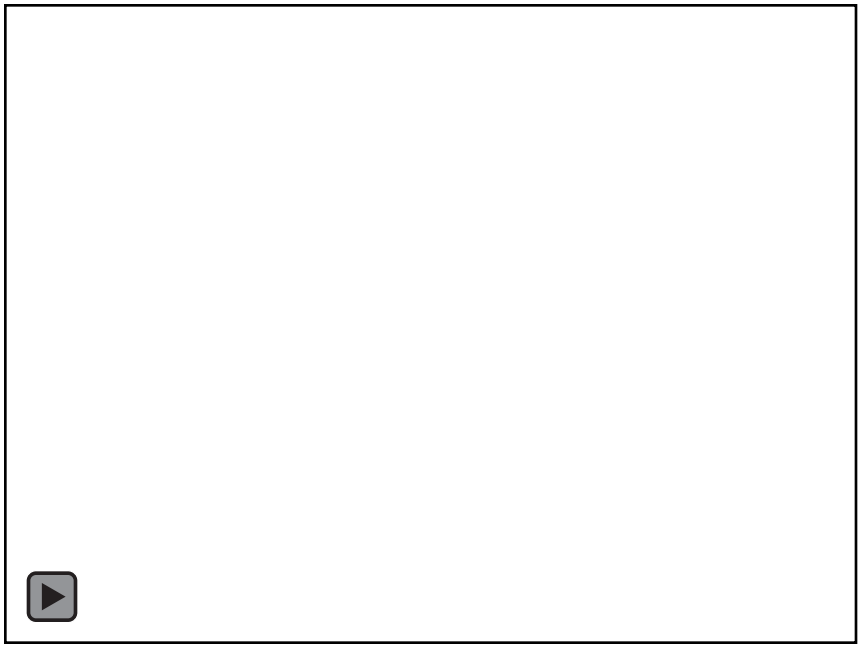
  - object segmentation
    - Schroff et al., 2008



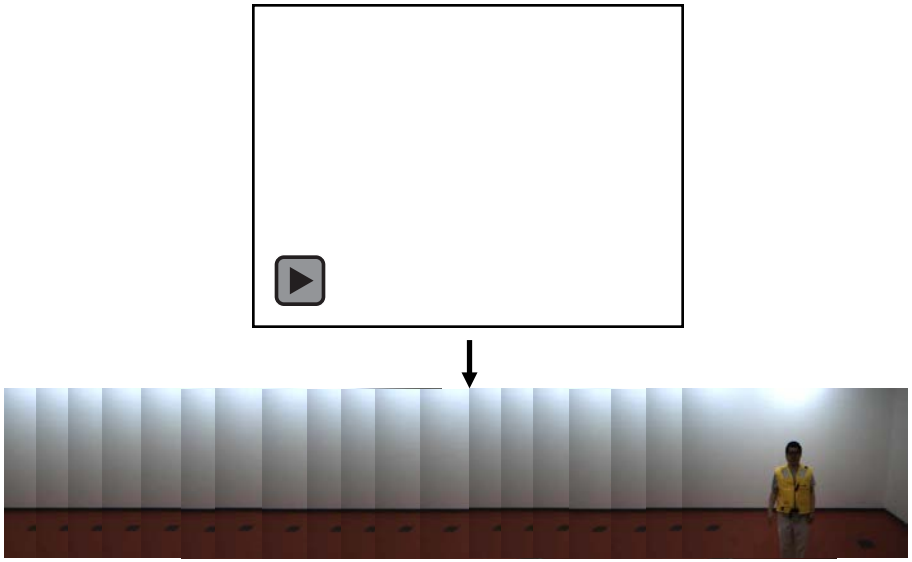
  - image classification
    - Bosch et al., 2007



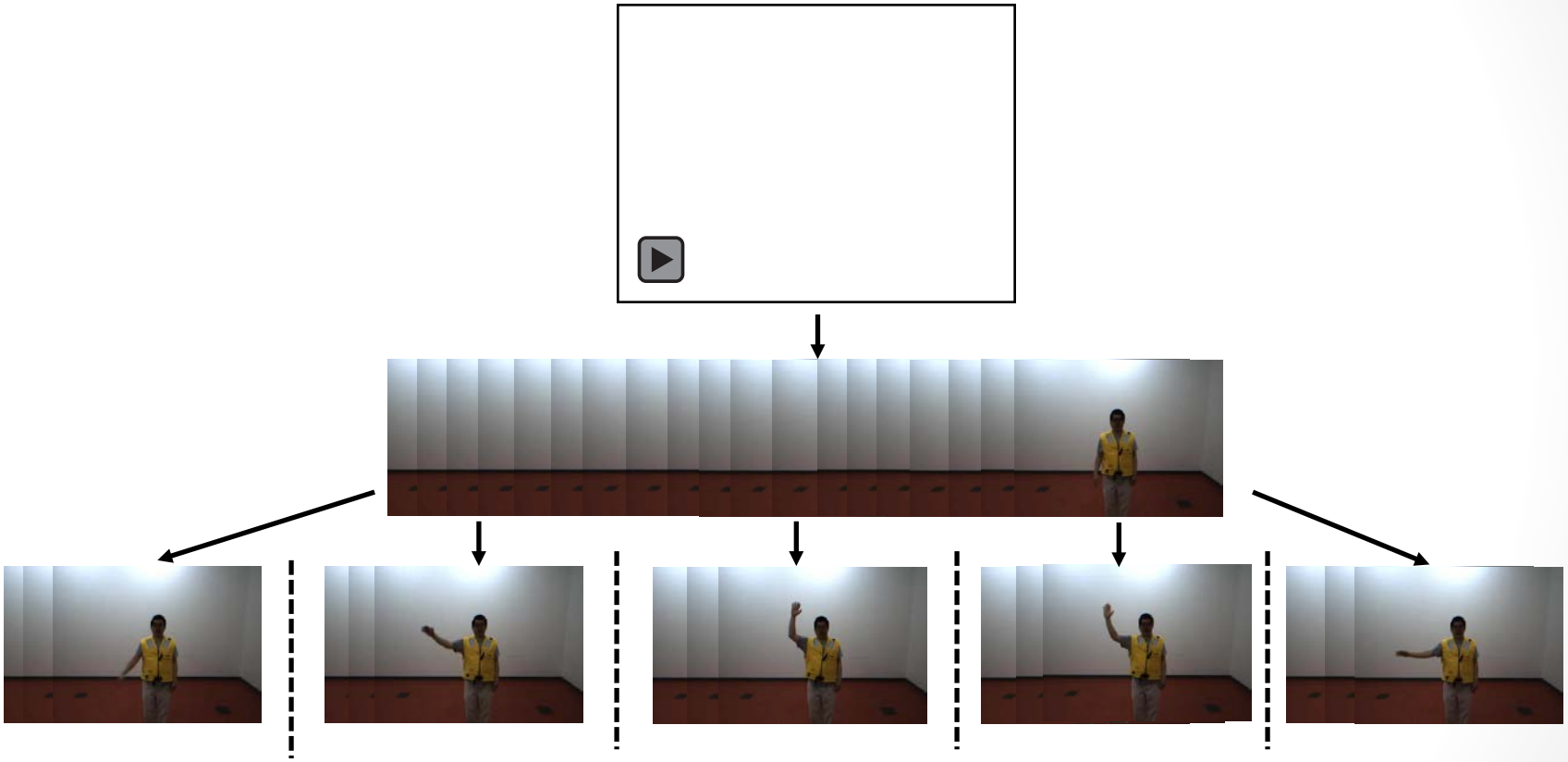
3. FRAMEWORK



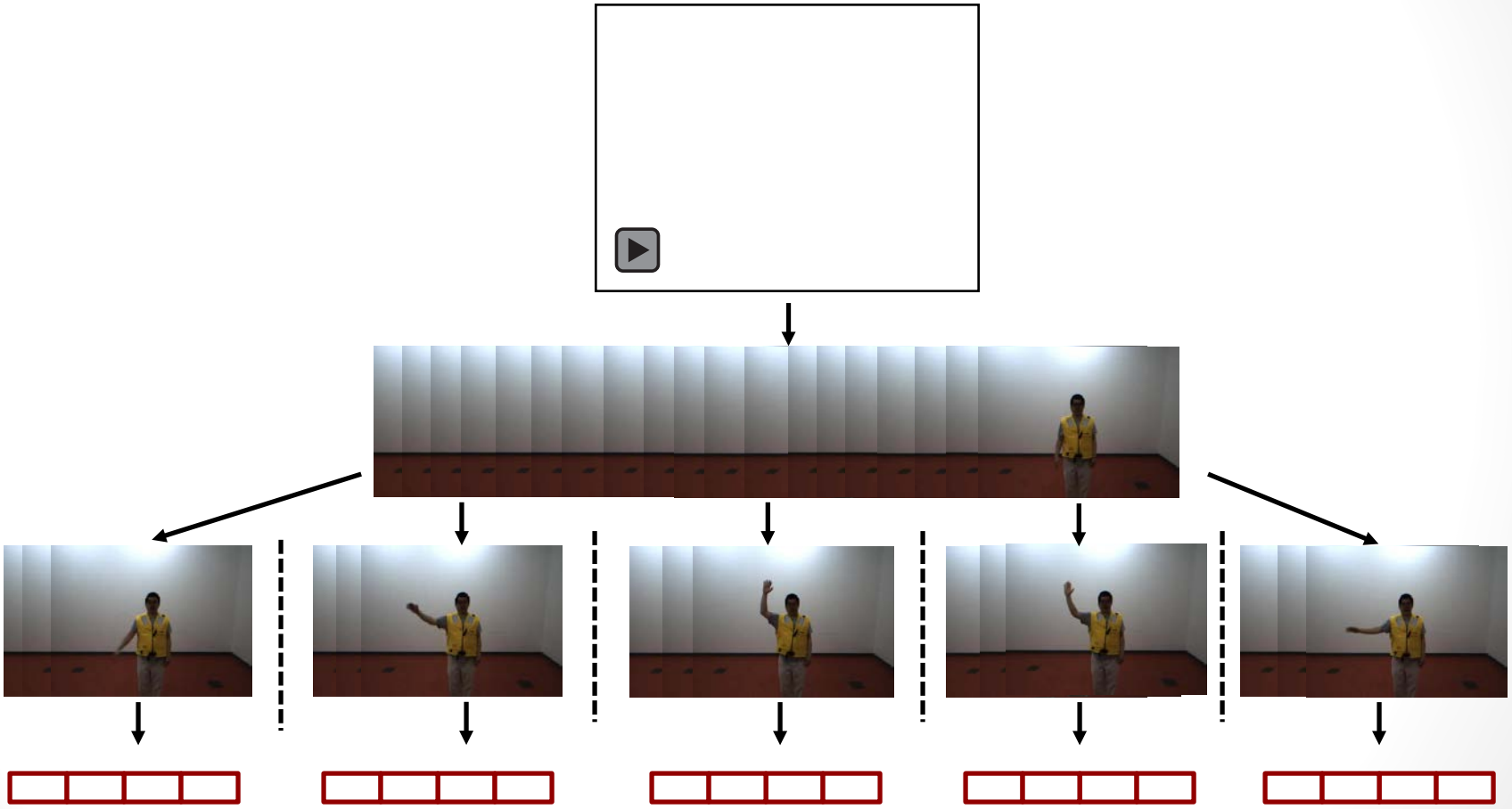
3. FRAMEWORK



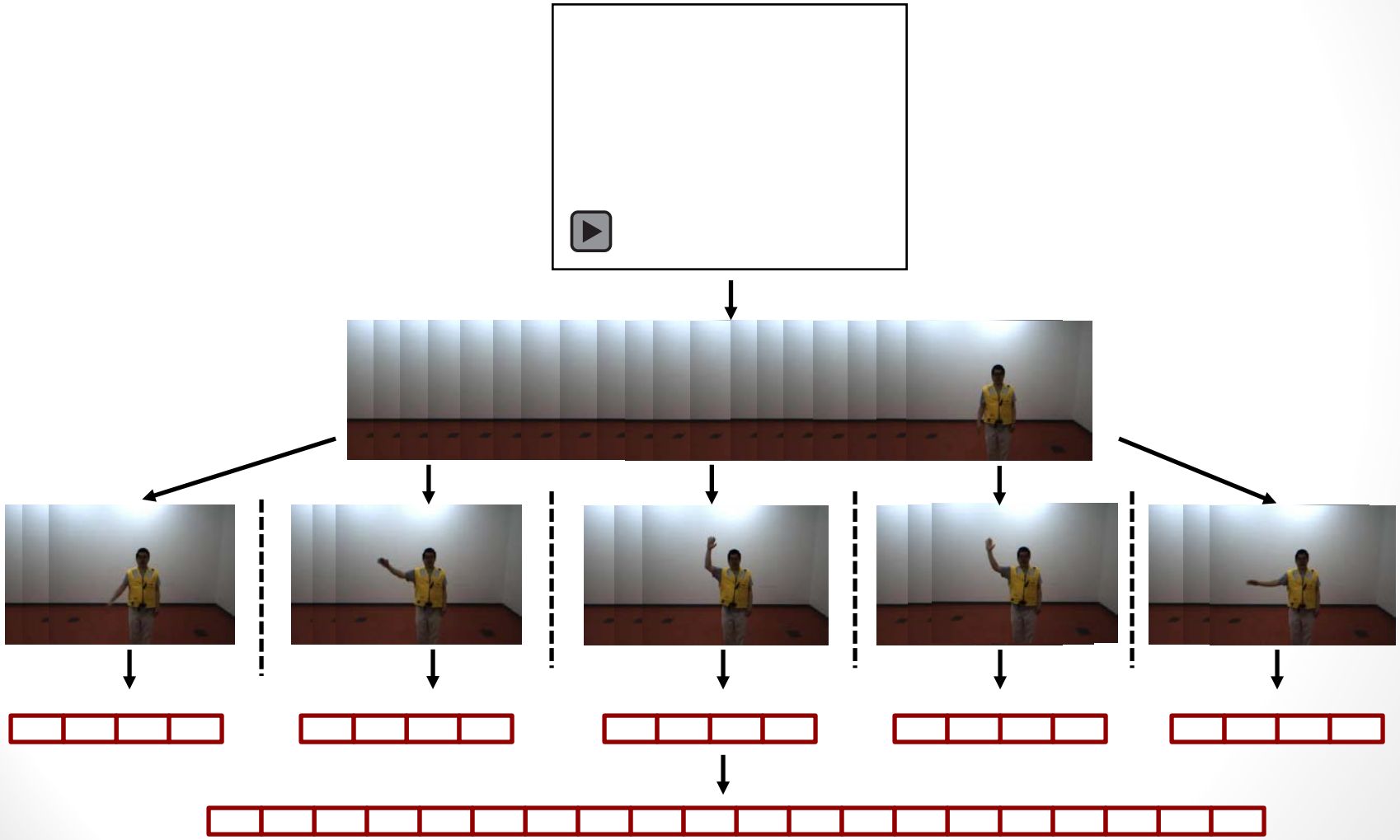
3. FRAMEWORK



3. FRAMEWORK



3. FRAMEWORK



Training: Input

Input

Training: Feature Extraction

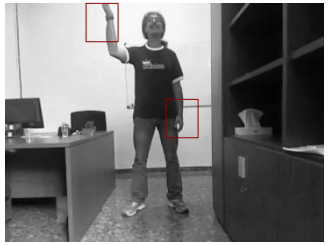
Input




Feature Extraction

- Normalized positional coordinates of joints
- Rotation angles of joints
- Positional and angular velocity of joints

Joint-based features



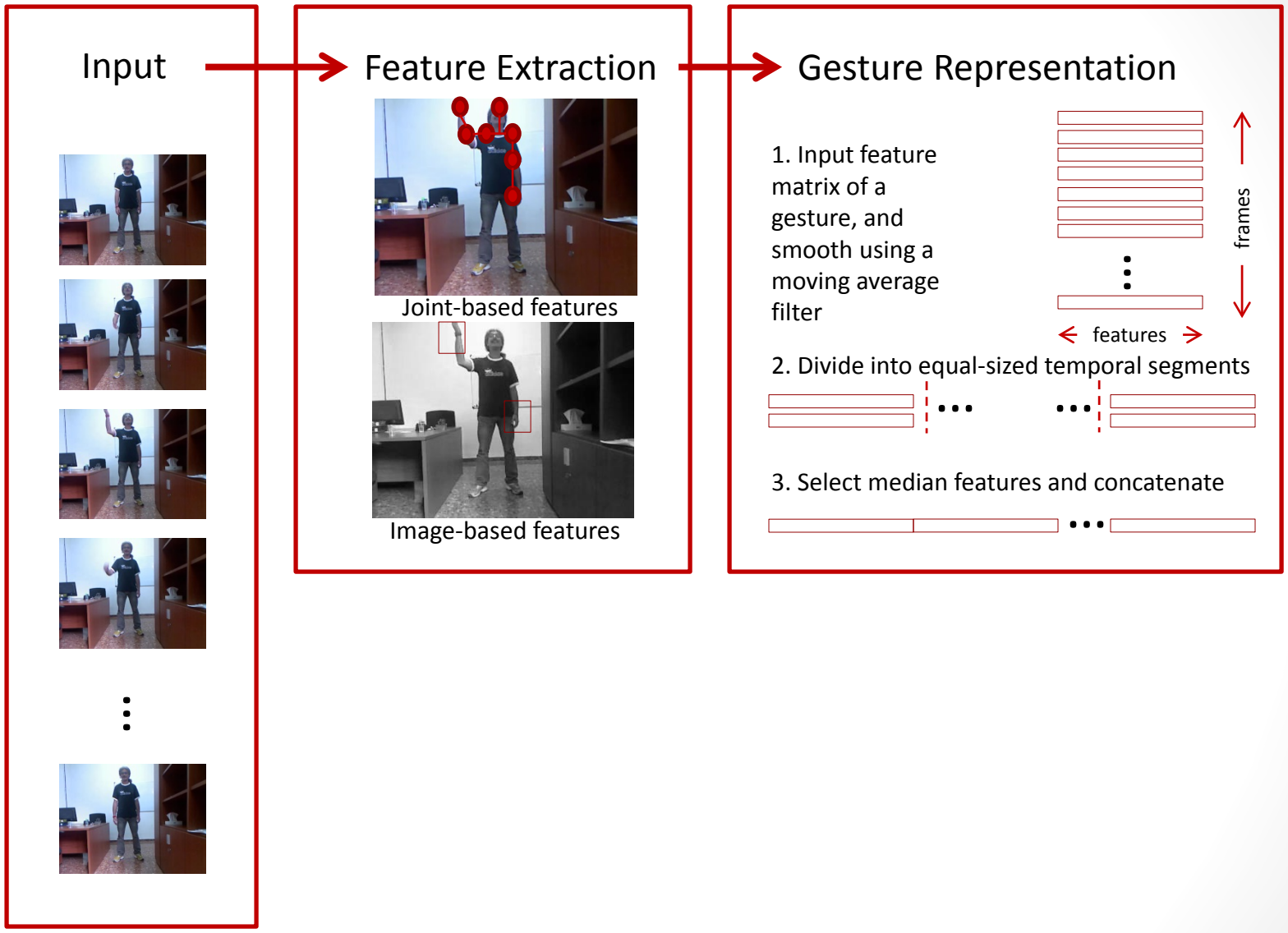
- Histogram of Oriented Gradients computed on boxes centered around the left and right hands

Image-based features

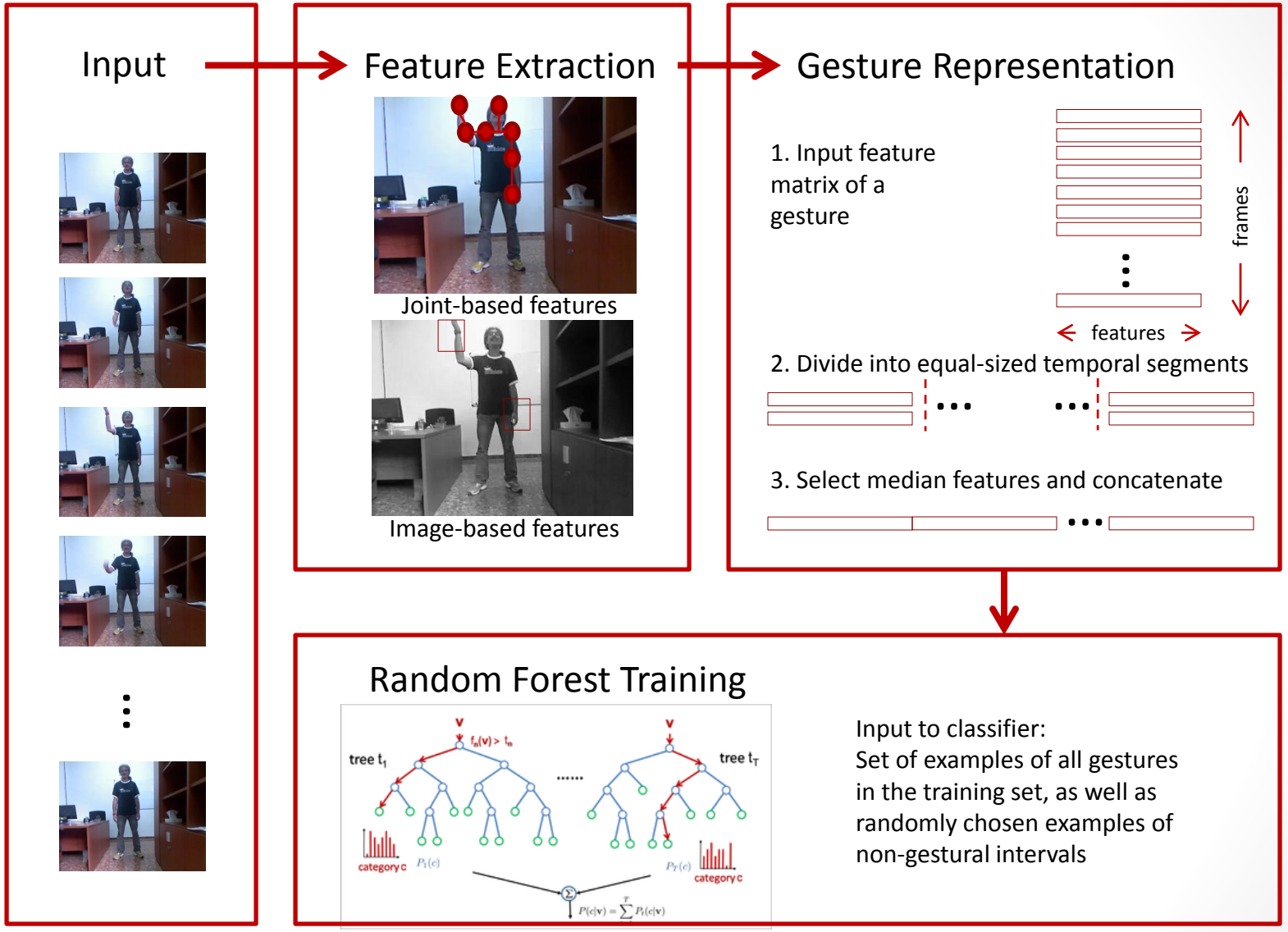
Feature Fusion

- Feature are fused to create a combined feature descriptor

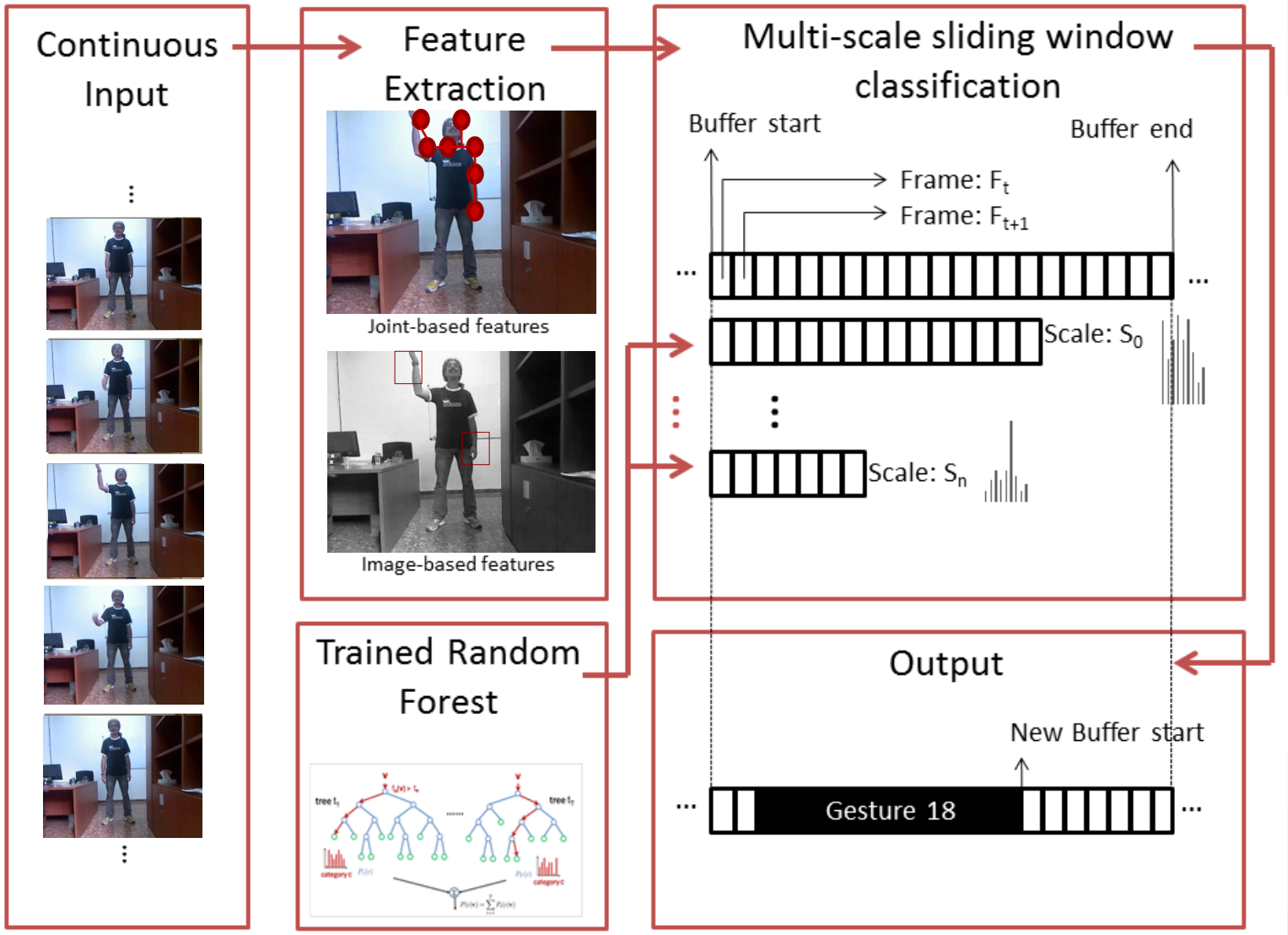
Training: Gesture Representation



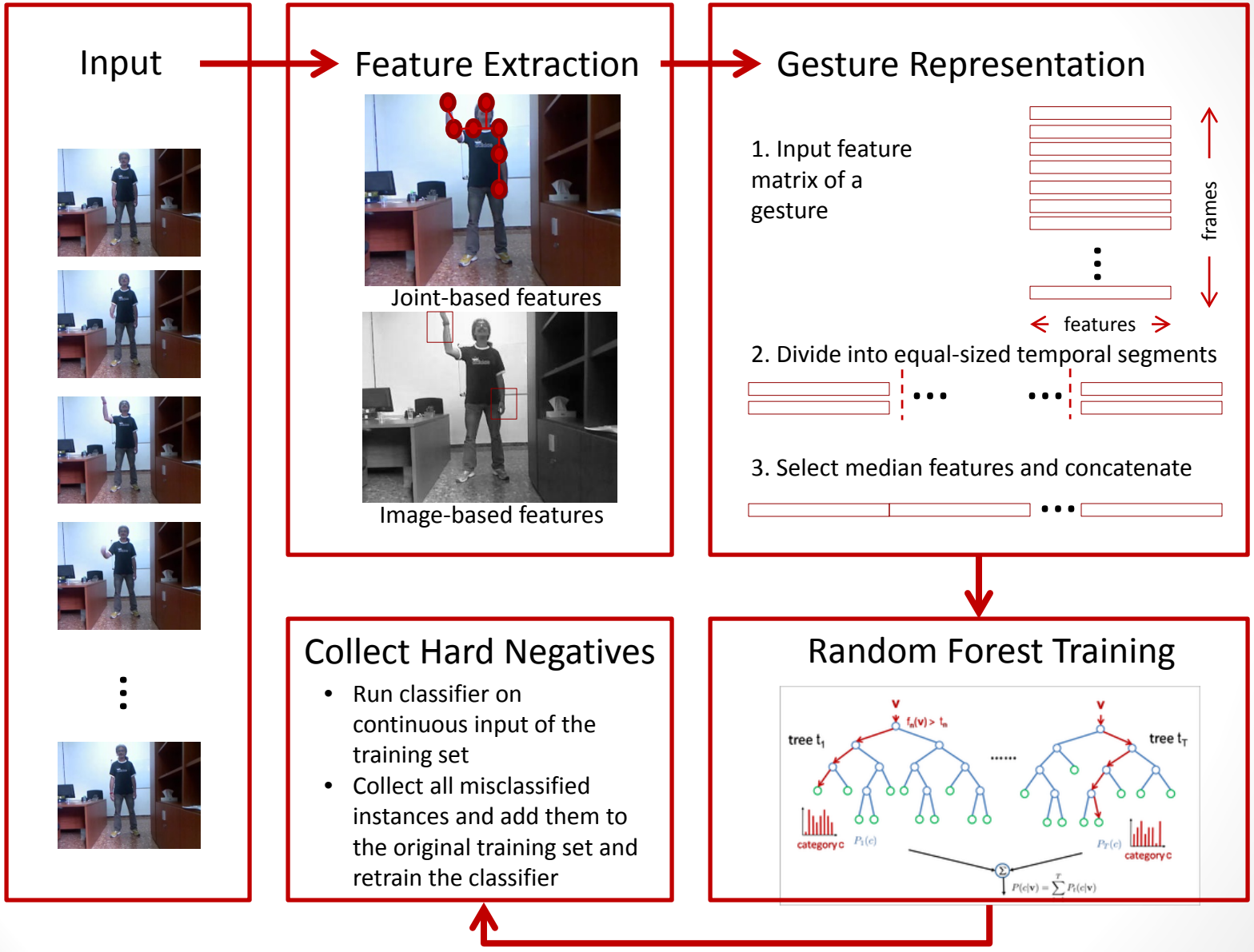
Training: Initial Random Forest Training



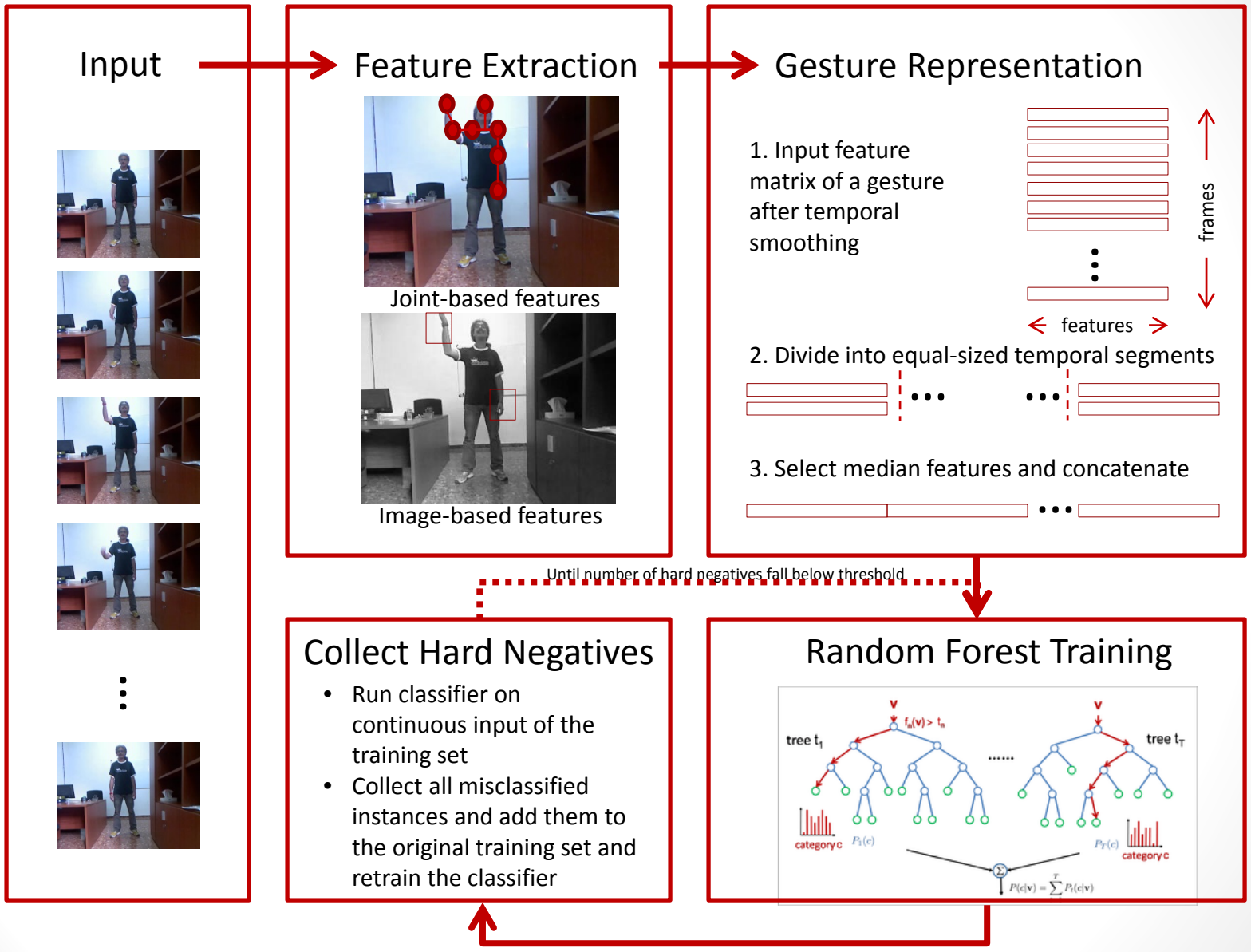
Training: Test on continuous input of training set



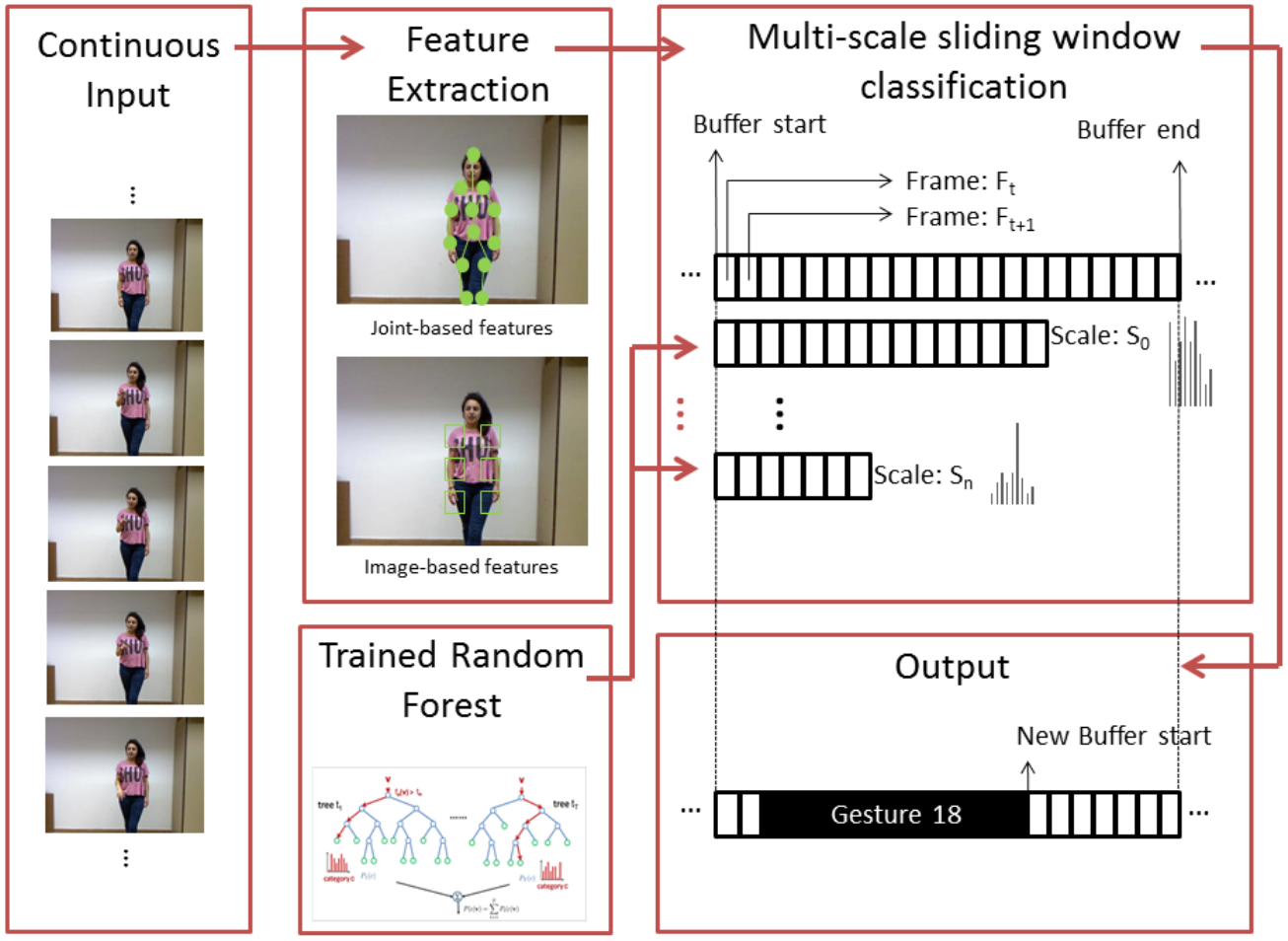
Training: Collection of hard negatives



Training: Iterative Random Forest Training

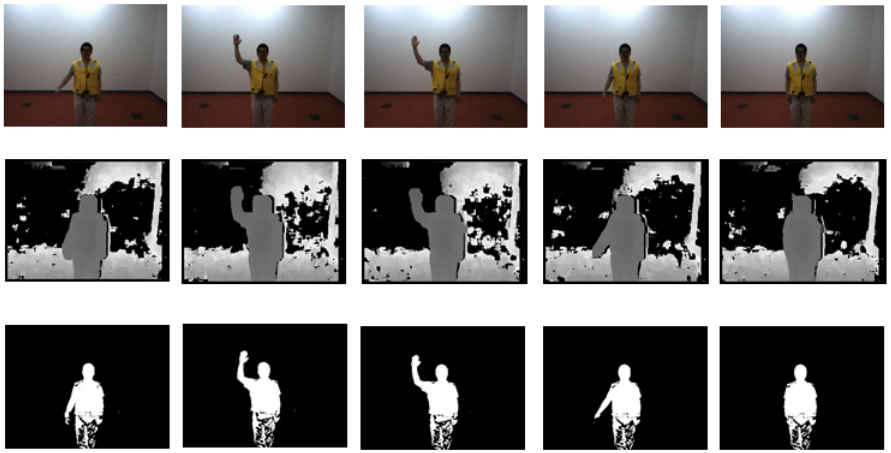


Testing


Continuous Input
Feature Extraction
Joint-based features
Image-based features
Trained Random Forest
Multi-scale sliding window classification
Buffer start
Buffer end
Frame: F_t
Frame: F_{t+1}
Scale: S_0
Scale: S_n
Output
New Buffer start
Gesture 18

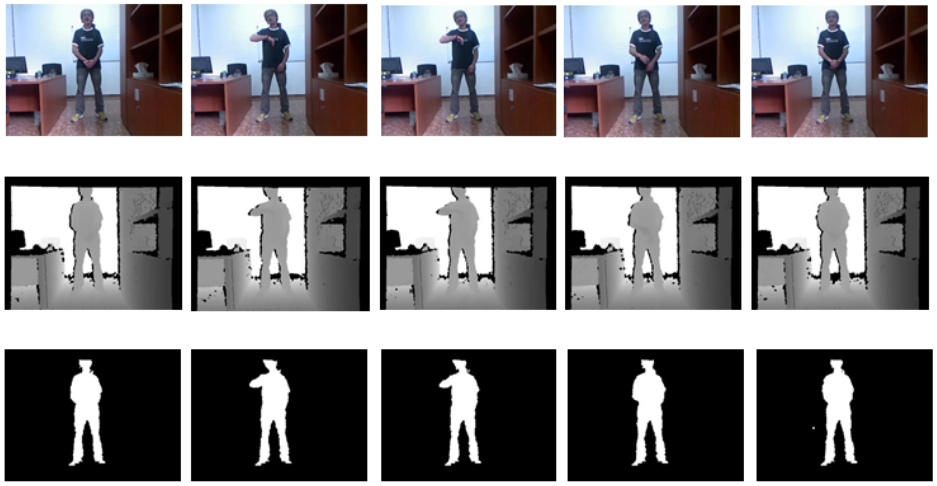
4. DATASETS

NATOPS



- Naval Air Training and Operating Procedures Standardization gestures
- 24 aircraft handling signals, performed by 20 subjects, 20 times
- Dataset includes RGB color images, depth maps, mask images and skeletal information

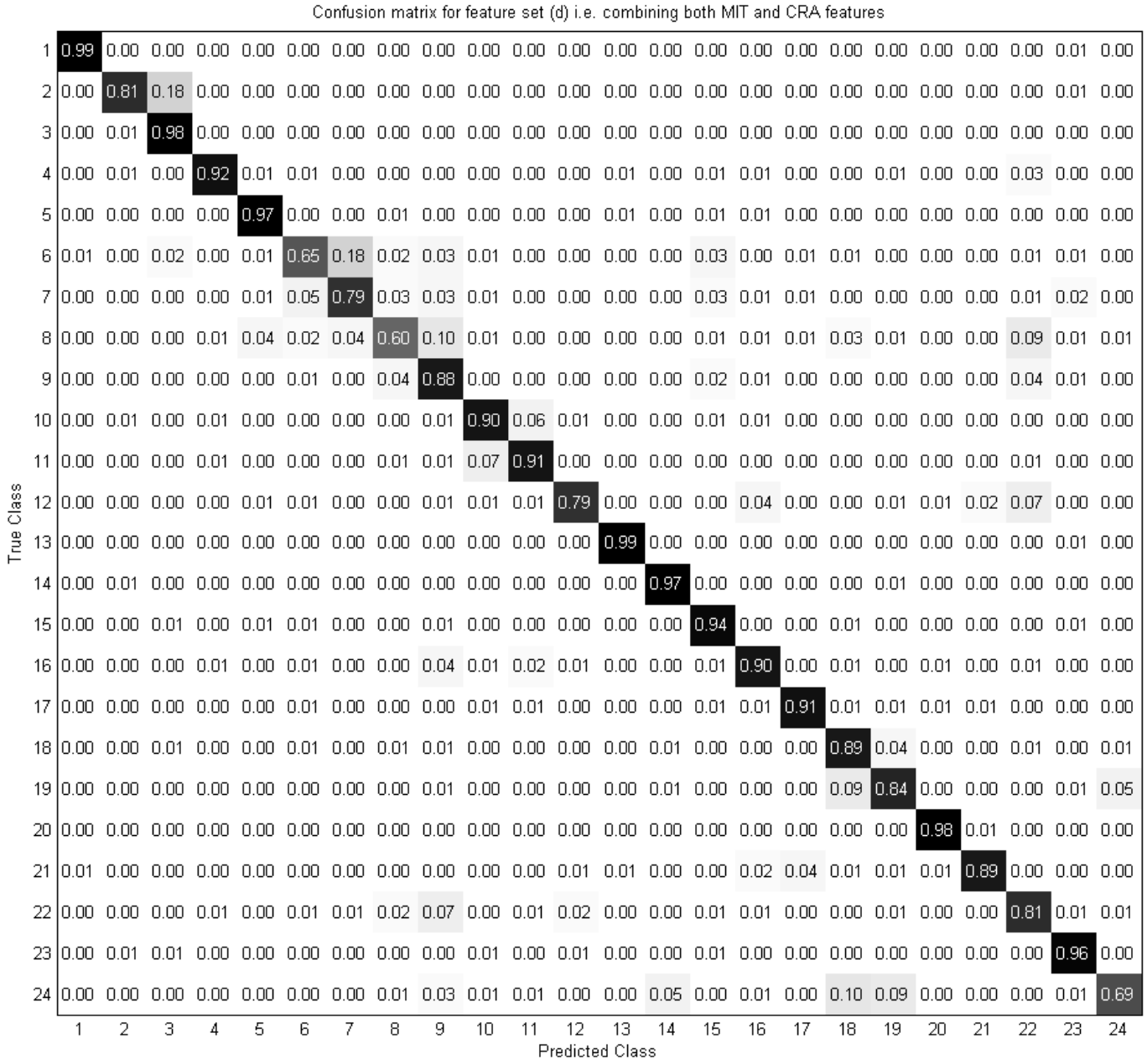
CHALEARN



- 20 unique Italian cultural and anthropological signs
- Development data: 7,754 labelled gestures
- Validation data: 3,363 gestures
- Test data: 2,742 gestures
- Dataset includes RGB color images, depth maps, mask images and skeletal information

5. EXPERIMENTAL RESULTS

- NATOPS
  - Classification only
- CHALEARN
  - Classification and Segmentation



NATOPS classification accuracy averaged over all subjects and gestures: **87.35%**

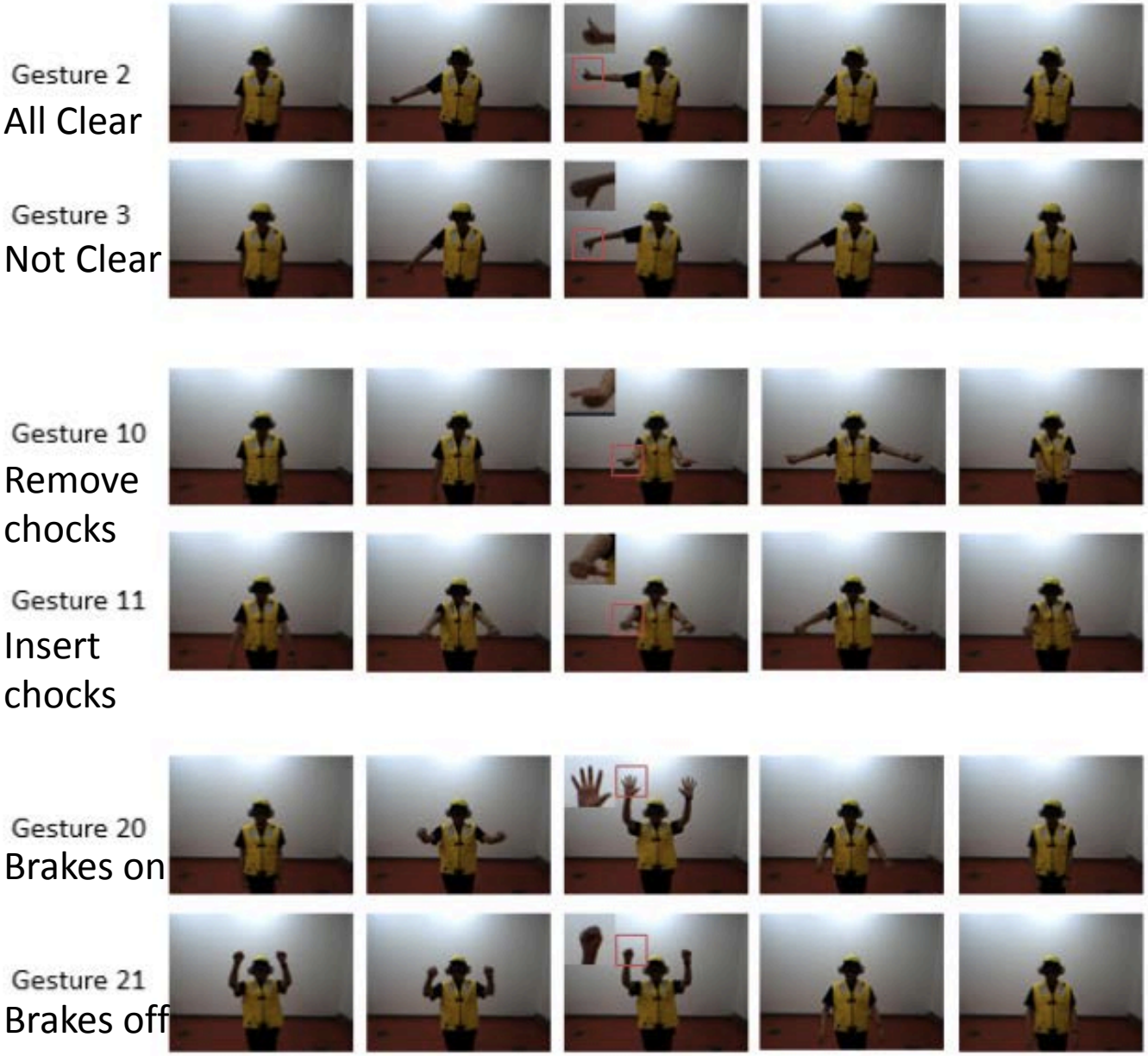
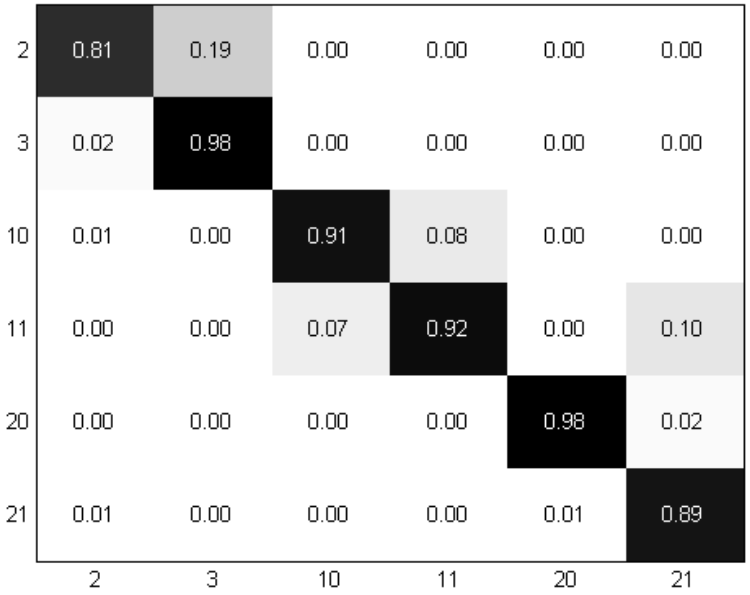


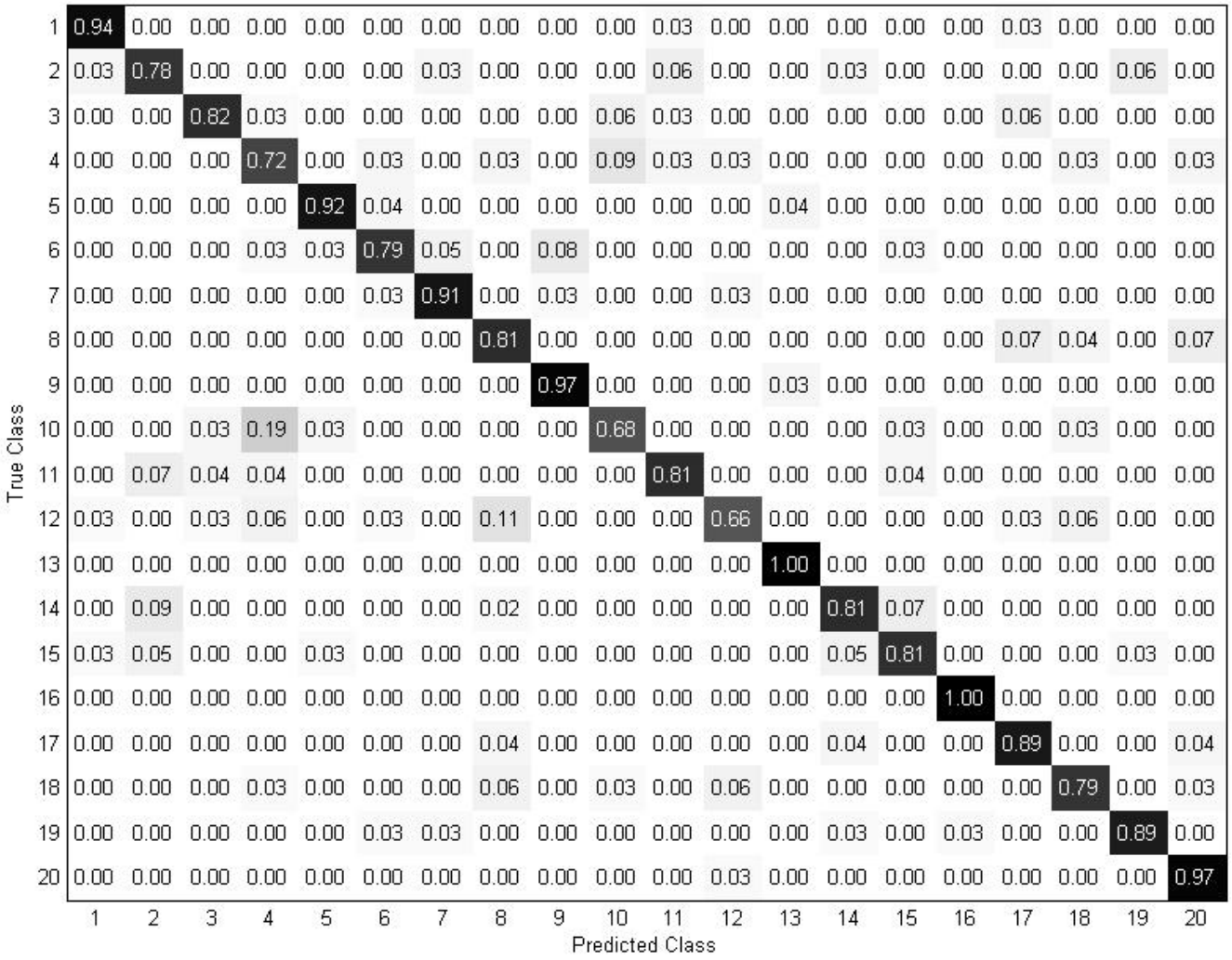
Figure: Pairs of similar gestures



Confusion Matrix for pairs of similar gestures

Classifier	Average Classification Accuracy
*HMM	77.67%
*HCRF	87.0%
Random Forest (ours)	**88.1%**

* As shown in:
Song, Yale, L. Morency, and Randall Davis. "Multi-view latent variable discriminative models for action recognition." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.



CHALEARN classification accuracy averaged over all subjects and gestures: **88.91%**

$$J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}}$$

Here, $A_{s,n}$ is the vector describing the ground truth indices of gesture n at sequence s, whereas $B_{s,n}$ is the vector describing the predicted indices of gesture n at sequence s.

Classifier	Jaccard score
Deep Neural Network	0.84
Random Forest (our)	0.68
Competition baseline	0.37

Table: CHALEARN 2014 Competition results

6. DISCUSSION

- We have presented a simple and efficient random forest framework.
- Reliable classifier that generalizes well
- The task of simultaneously detecting and classifying gestures is a more difficult challenge than solely classifying correctly segmented gestures.

THANK YOU!