

# 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

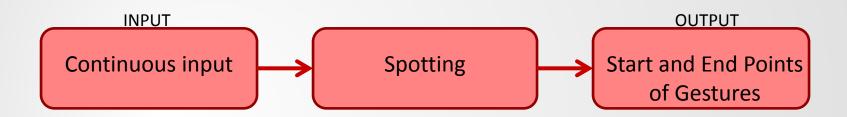




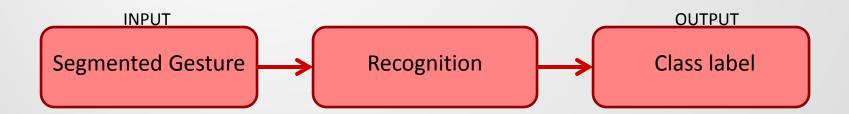
INTRODUCTION

# Gesture spotting and recognition

Gesture spotting



• Gesture recognition



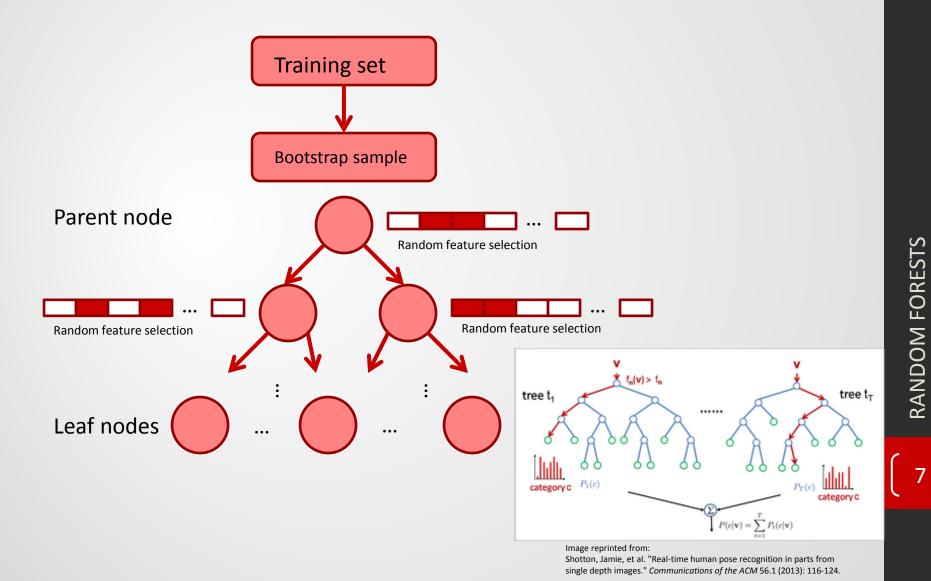
### **Problem Definition**

- 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**



### **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*)

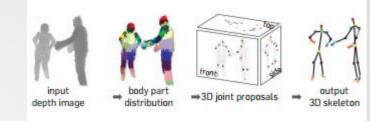
$$I_j = H(S_j) - \sum_{k \in (L,R)} \frac{|S_j^k|}{|S|} H(S_j^k)$$

 $S_j$ : 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

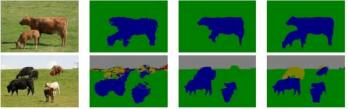
$$H(S) = -\sum_{c \in C} p_c \log(p_c)$$

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



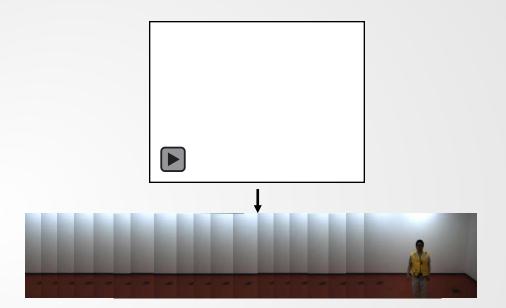
- object segmentation
  - Schroff et al., 2008



- image classification
  - Bosch et al., 2007

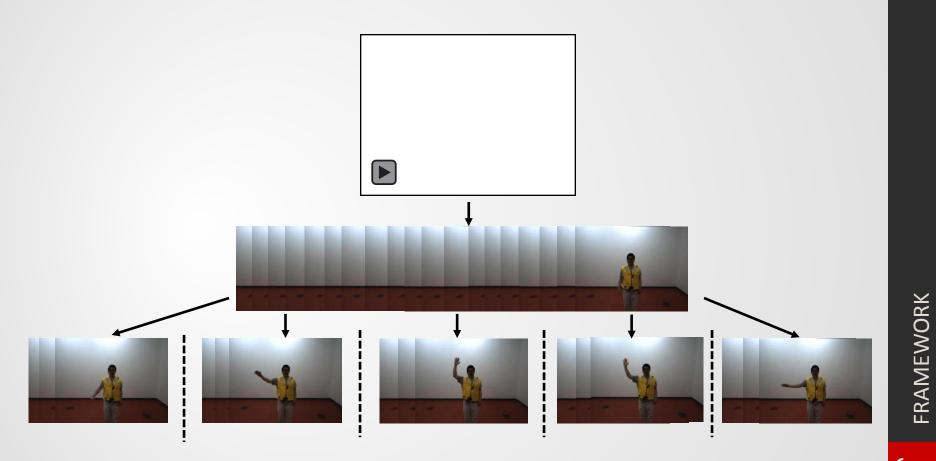


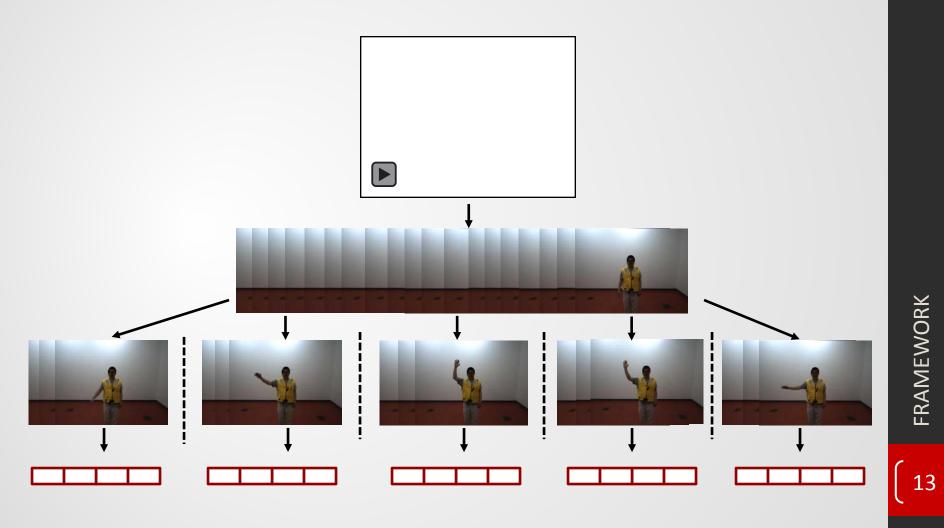


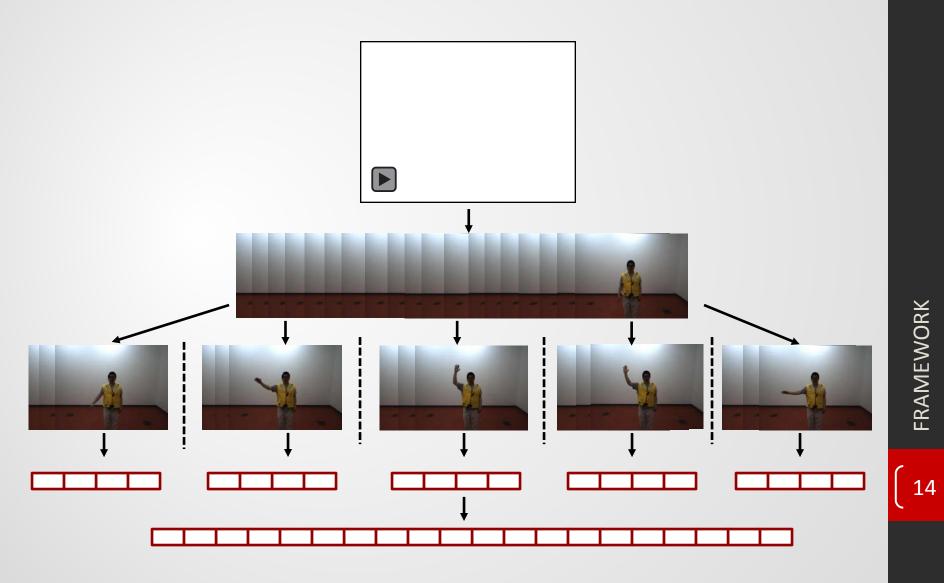












### Training: Input













FRAMEWORK

### **Training: Feature Extraction**

#### Feature Extraction

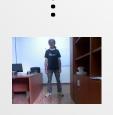


Input











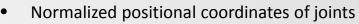
#### Joint-based features



#### Image-based features

#### Feature Fusion

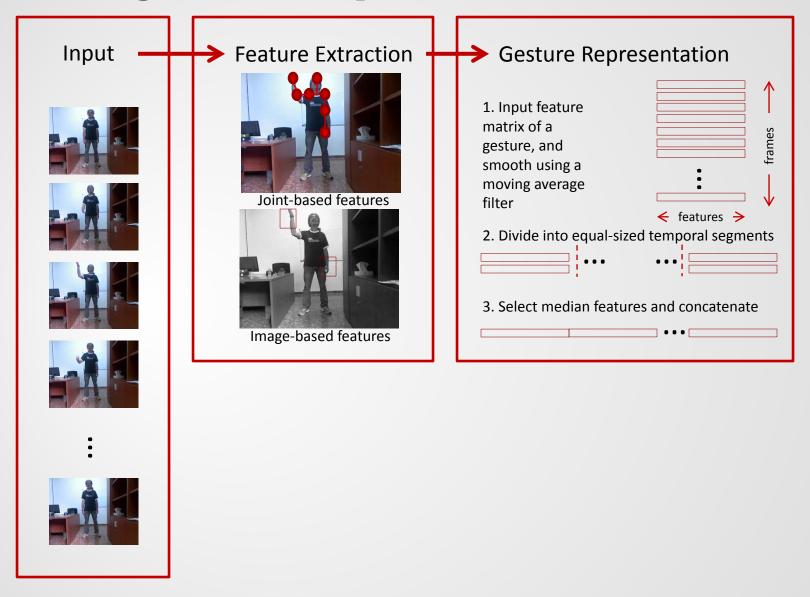
• Feature are fused to create a combined feature descriptor



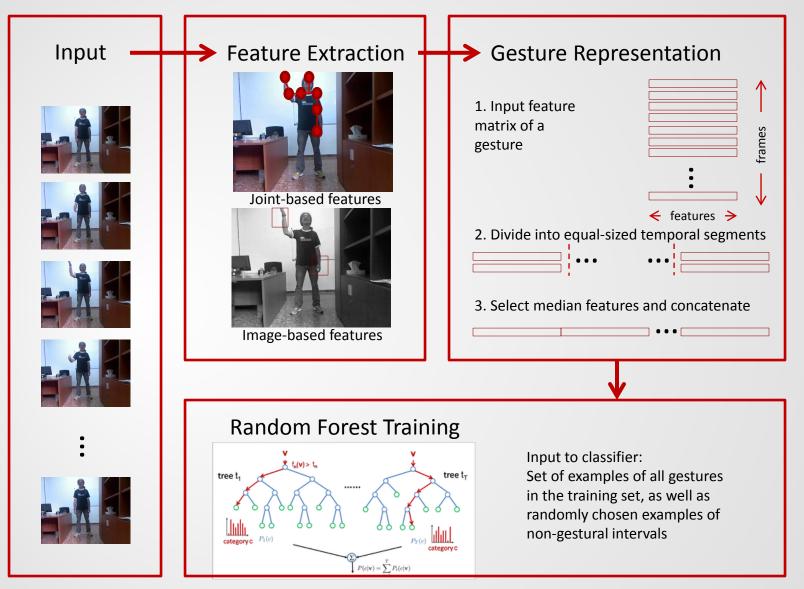
- Rotation angles of joints
- Positional and angular velocity of joints

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

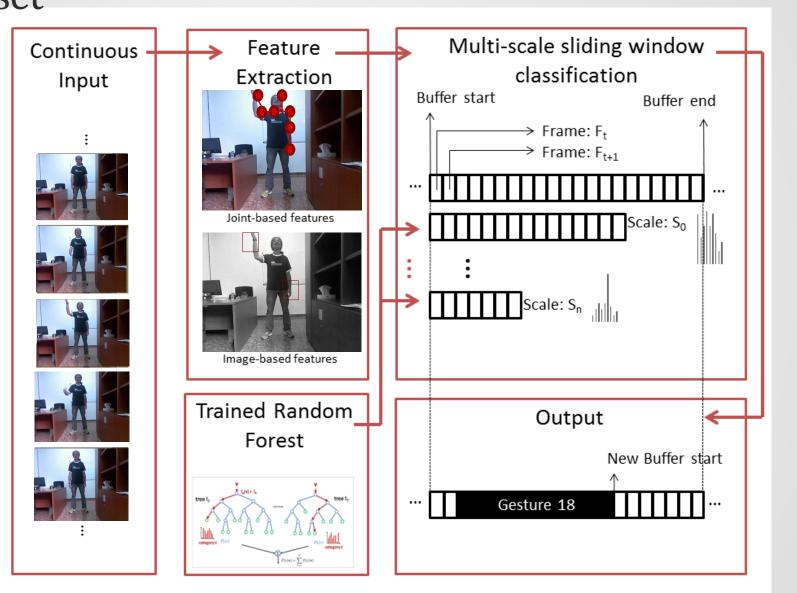
### **Training: Gesture Representation**



### **Training: Initial Random Forest Training**

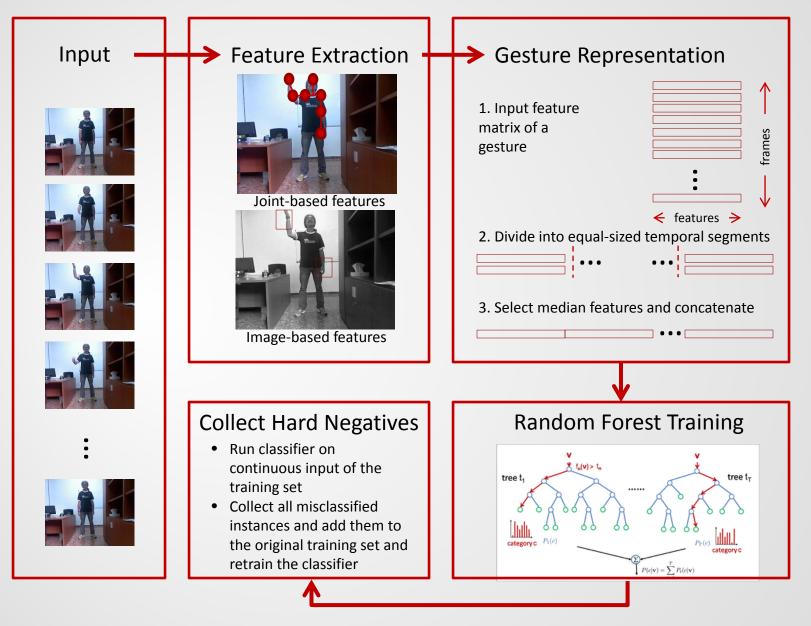


# Training: Test on continuous input of training set

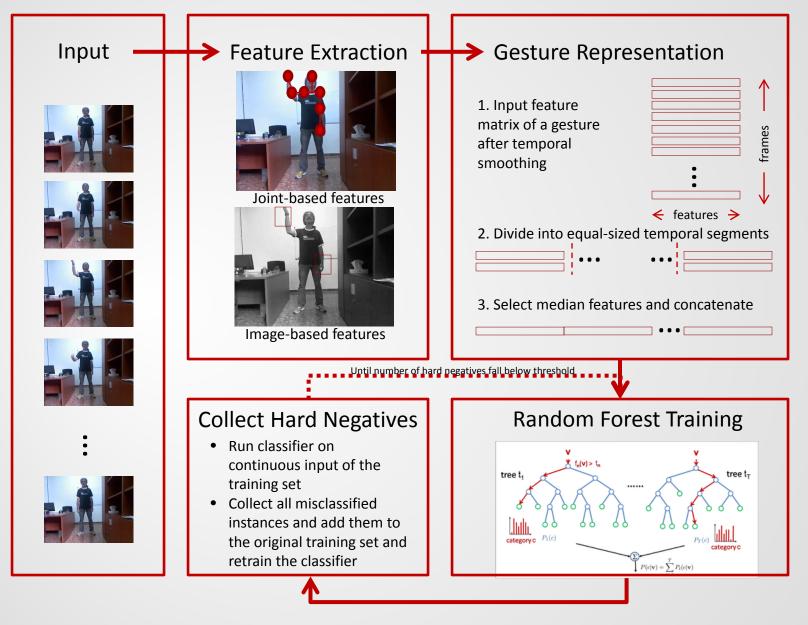


FRAMEWORK

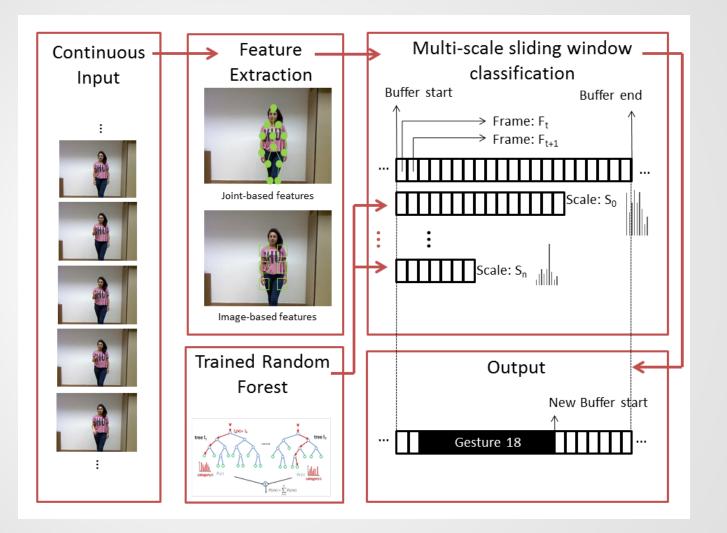
### Training: Collection of hard negatives



### **Training: Iterative Random Forest Training**

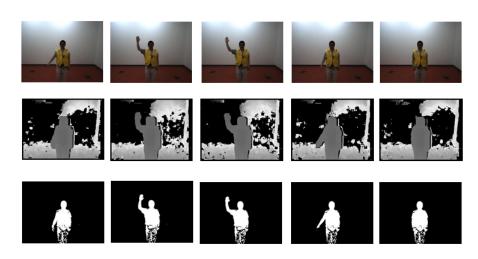


### Testing



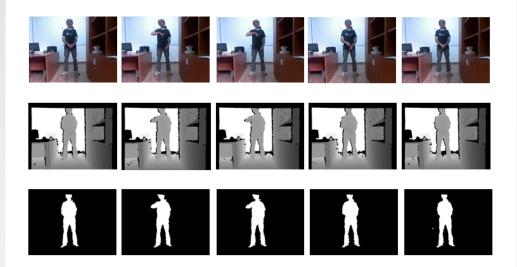
# 4. DATASETS

### 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



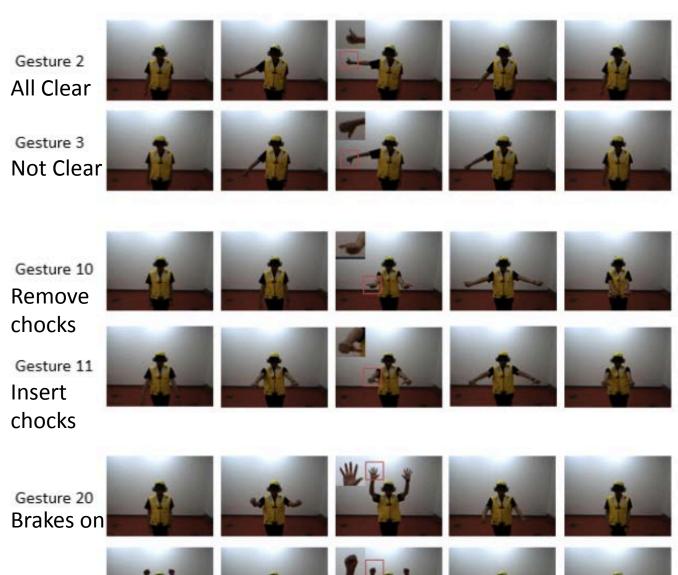
- 20 unique Italian cultural and anthropological signs
- Development data: 7,754 labelled gestures
- Validation data: 3,363 gestures
- Test data: 2,742 gestures
- Dataset includes RGB color images, depth maps, mask images and skeletal information

# **5. EXPERIMENTAL RESULTS**

- NATOPS
  - Classification only
- CHALEARN
  - Classification and Segmentation

	1	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	2	0.00	0.81	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	3	0.00	0.01	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.01	0.00	0.92	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.03	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6	0.01	0.00	0.02	0.00	0.01	0.65	0.18	0.02	0.03	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.00
	7	0.00	0.00	0.00	0.00	0.01	0.05	0.79	0.03	0.03	0.01	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.00
	8	0.00	0.00	0.00	0.01	0.04	0.02	0.04	0.60	0.10	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.03	0.01	0.00	0.00	0.09	0.01	0.01
	9	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.88	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.01	0.00
	10	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.90	0.06	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	11	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.07	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Class	12	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.79	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.01	0.02	0.07	0.00	0.00
True	13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	14	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	15	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
	16	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.04	0.01	0.02	0.01	0.00	0.00	0.01	0.90	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00
	17	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.00	0.00	0.00
	18	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.89	0.04	0.00	0.00	0.01	0.00	0.01
	19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.09	0.84	0.00	0.00	0.00	0.01	0.05
	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.01	0.00	0.00	0.00
	21	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.02	0.04	0.01	0.01	0.01	0.89	0.00	0.00	0.00
	22	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.02	0.07	0.00	0.01	0.02	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.81	0.01	0.01
	23	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00
	24							0.00					0.00						0.10			0.00	0.00	0.01	0.69
		1	2	3	4	5	6	7	8	9	10	11 Pi	12 redicte	13 d Clas	14 ss	15	16	17	18	19	20	21	22	23	24

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

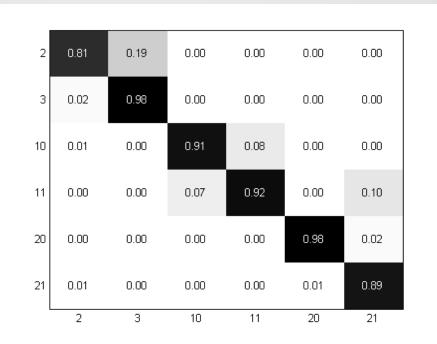


Gesture 21 Brakes off

27

Figure: Pairs of similar gestures

**EXPERIMENTAL RESULTS** 



#### Confusion Matrix for pairs of similar gestures

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

28

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.

	1	0.94	0.00	0.00			0.00										0.00	0.03		0.00	0.00
	2	0.03		0.00		0.00	0.0000000	0.03	0.000000	0.00 0.00		0.06		0.00	0.03	20000000 20000000	0.00	0.00	0.00	0.06	0.00
	4	0.00	0.00	0.00	0.72	0.00	0.03	0.00	0.03	0.00	0.09	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03
	5	0.00	0.00	0.00	0.00	0.92	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.03	0.03	0.79	0.05	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
	7	0.00	0.00	0.00		1000000	0.03		10000000	- 000 R	0.00	20220		0.00	0.00	0.00	0.00		0.00	0.00	0.00
		0.00	0.00		0.00	0.00		0.00						0.00		0.00		0.07		0.00	0.07
SS	क 	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	and a second	0.68	la construction de la constructi	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rue Class			0.00	0.03		0.00	0.00	0.00	0.00		0.00		6	0.00		0.03	0.00	0.00	0.00	0.00	0.00
Ц	12	0.03	0.00	0.03			0.03						(Annotation)	lan an a		0.00	0.00	0.03		0.00	0.00
	13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	14	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.81	0.07	0.00	0.00	0.00	0.00	0.00
	15	0.03	0.05	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.81	0.00	0.00	0.00	0.03	0.00
	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
	17	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00		0.00		0.04		0.00			0.00	0.04
	18	0.00	0.00	0.00		0.00			0.06	22552		00000		100000000	0.00	ononini Magazini		0.00		0.00	0.03
		0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00 0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.00		0.00 0.97
	20	1	2	3	4	5	0.00	7	0.00	9	10	11	12	13	14	15	16	17	18	19	0.97 20
		202	_			50 <del>1</del> 0		10		P	redicte	d Clas			5059		20. <del>T</del> .C			1000	

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

**EXPERIMENTAL RESULTS** 

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

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

# DISUCSSION



- 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!**

