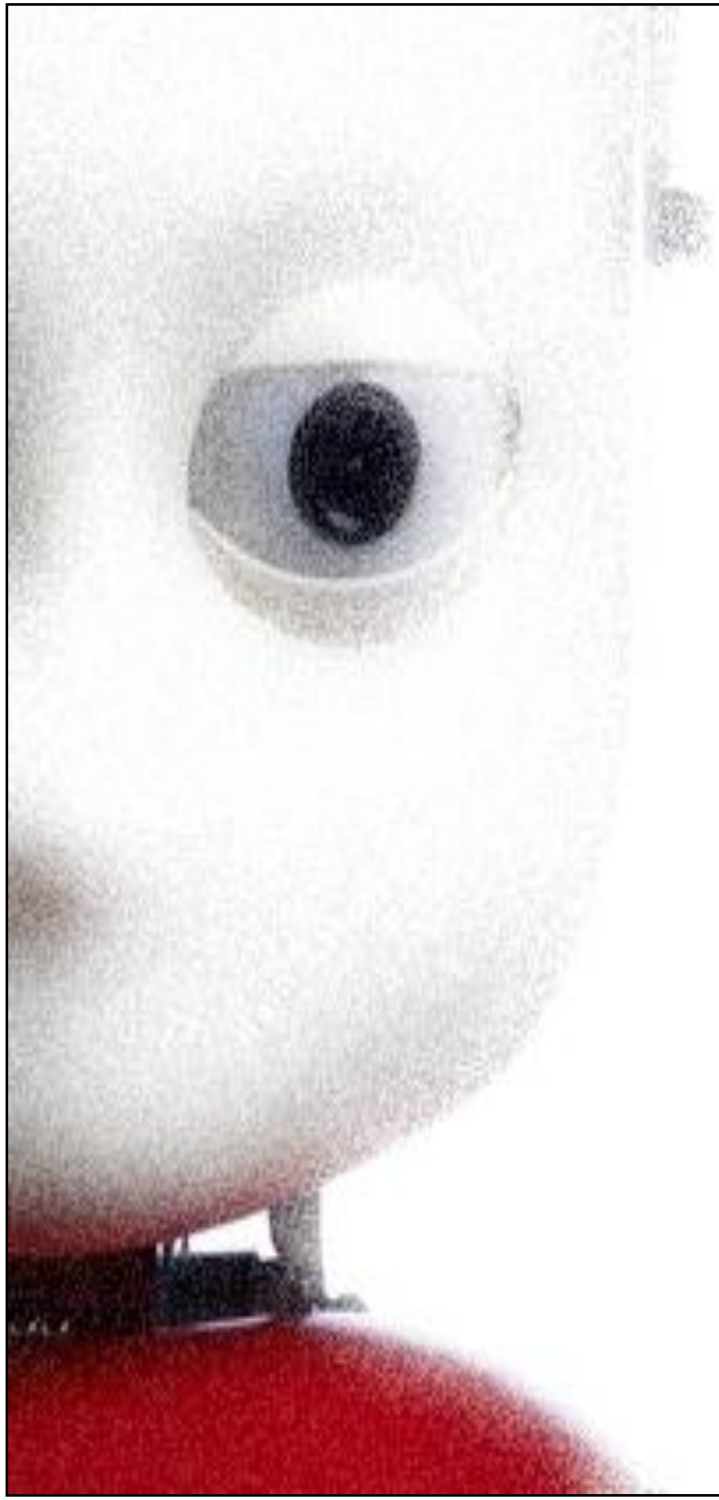


Teaching iCub to see (its world)

Francesca Odone
Università degli Studi di Genova
francesca.odone@unige.it

Joint work with

Carlo Ciliberto

Sean Ryan Fanello (now at MS)

Ilaria Gori

Giorgio Metta

Nicoletta Noceti

Francesca Odone

Lorenzo Rosasco

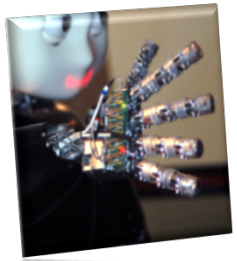
…

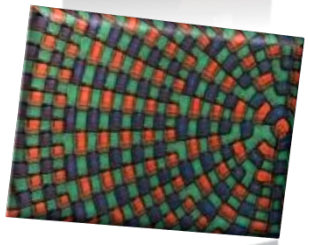
iCub is an **open source** international endeavour initially funded by the EU project **RobotCub**

- a full humanoid robot
- is 104cm, weighs 25 kg
- has 53 degrees of freedom
- can crawl, sit and manipulate
- open design under GPL/LGPL

open hardware

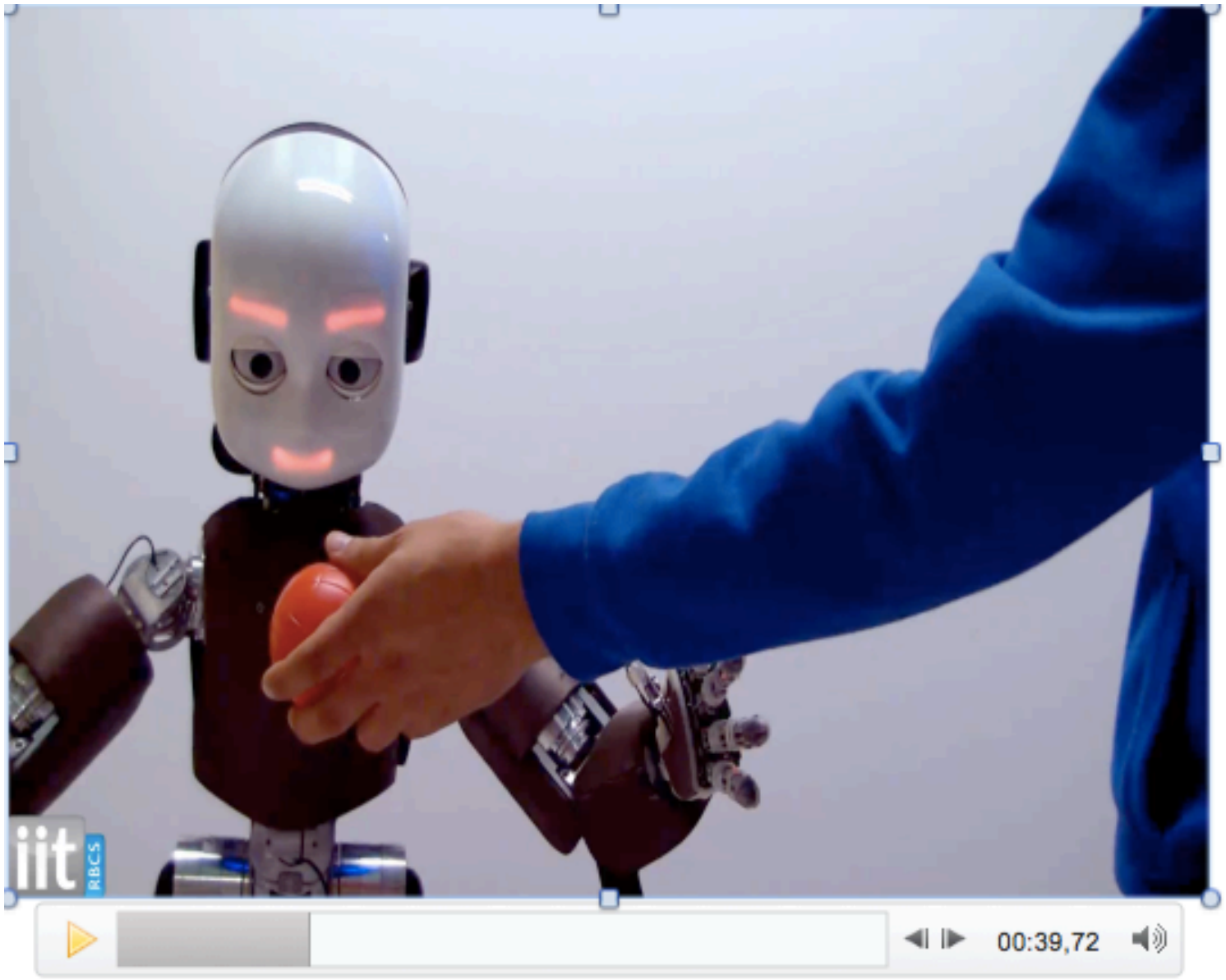
why is the iCub so special?



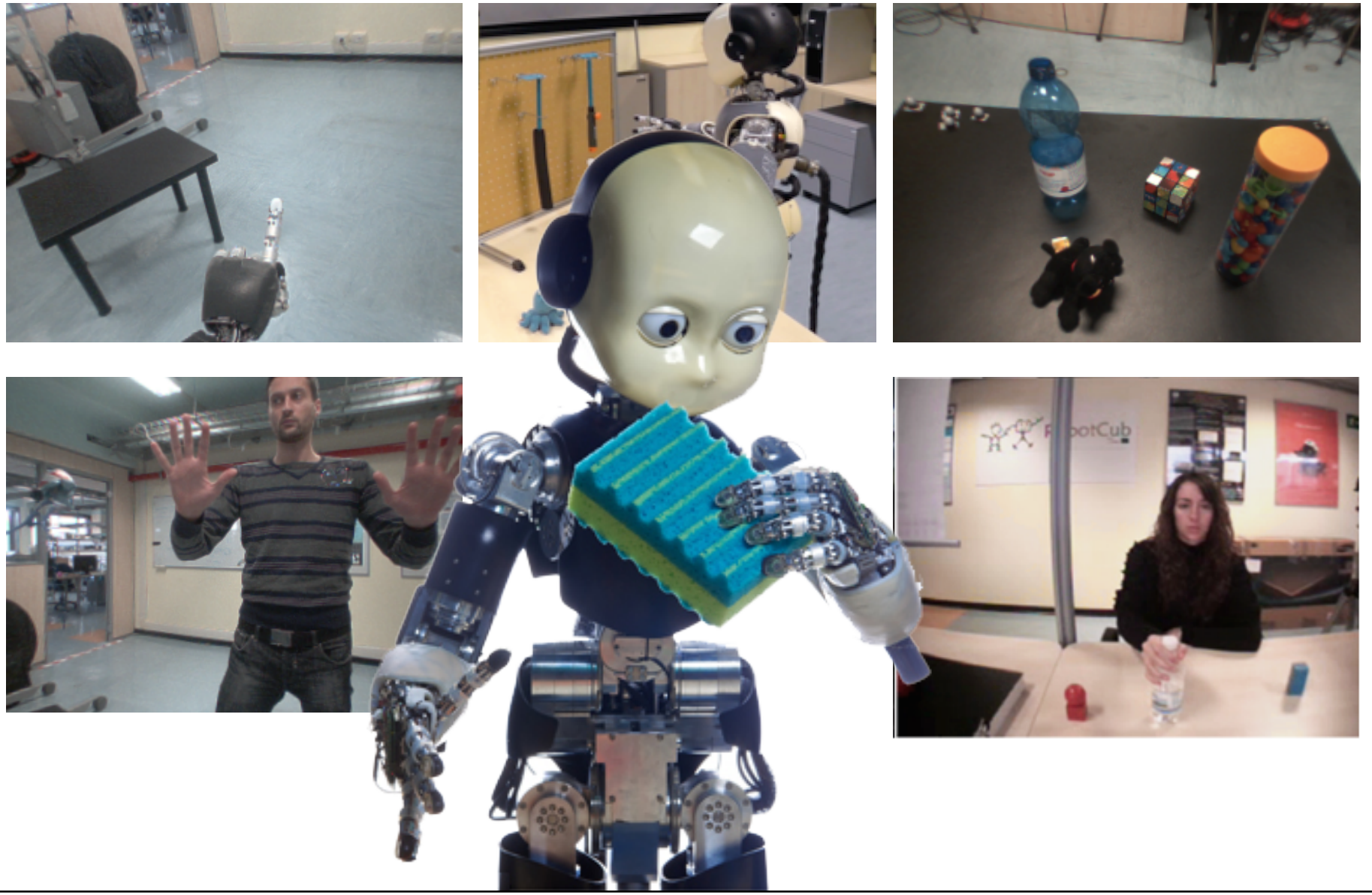


- hands: the design started from the hands
  - 5 fingers, 9 degrees of freedom, 19 joints
- sensors: human-like, e.g. no lasers
  - cameras, microphones, gyros, encoders, force, tactile…
- electronics: flexibility for research
  - custom electronics, small, programmable (DSPs)
- reproducible platform: community designed
  - reproducible & maintainable yet evolvable platform
  - large software repository (>1M lines of code)

Seeing is important



iCub world is still quite repetitive



HRI can help

- The interaction between human and robot allows us to gather significant amount of data quite easily
- Also we may label (most of) them in a straightforward way

The challenges of a real setting

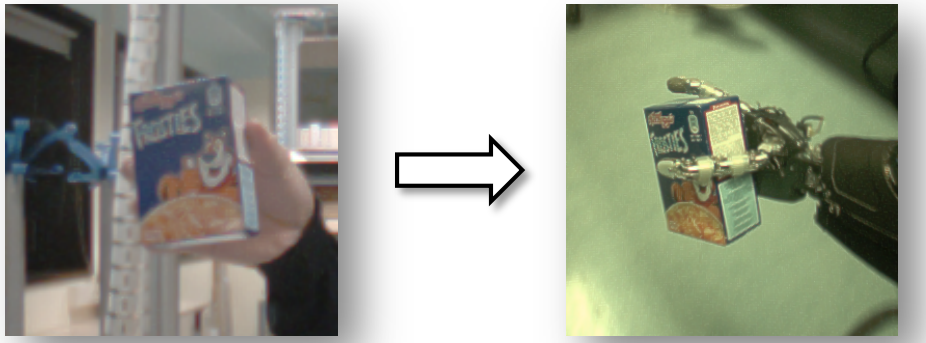
- Real time requirements
- Unbalanced Data
- Structured Clutter
- Incremental Learning
- Integration among different modules

So far…

we are addressing two main higher level vision problems

- Object recognition / categorization
- Action recognition

Object recognition

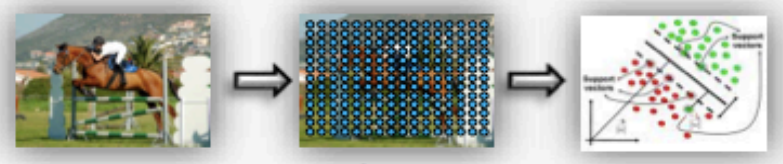


We mainly focus on representation

- ✓ *No context (so far)*
- ✓ *Weak supervision*
- ✓ Invariance

Learning data representations

2-Layer Systems



G. Csurka et al. – **Visual categorization with bags of keypoints** ECCVW 2004

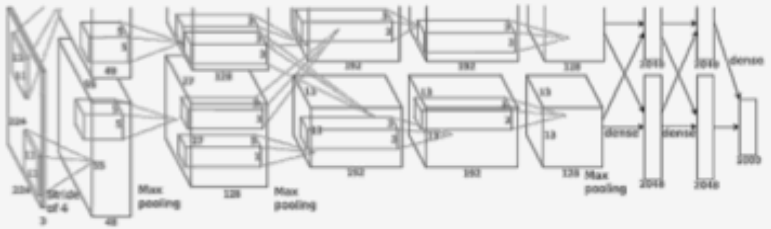
PRO:

Simple
Fast
Online
Easy to Generalize

CONS:

No invariance

Deep Architectures



A. Krizhevsky et al. – **ImageNet Classification with Deep Convolutional Neural Networks** NIPS 2012
T. Serre et al. – **Robust object recognition with cortex-like mechanisms** TPAMI 2007

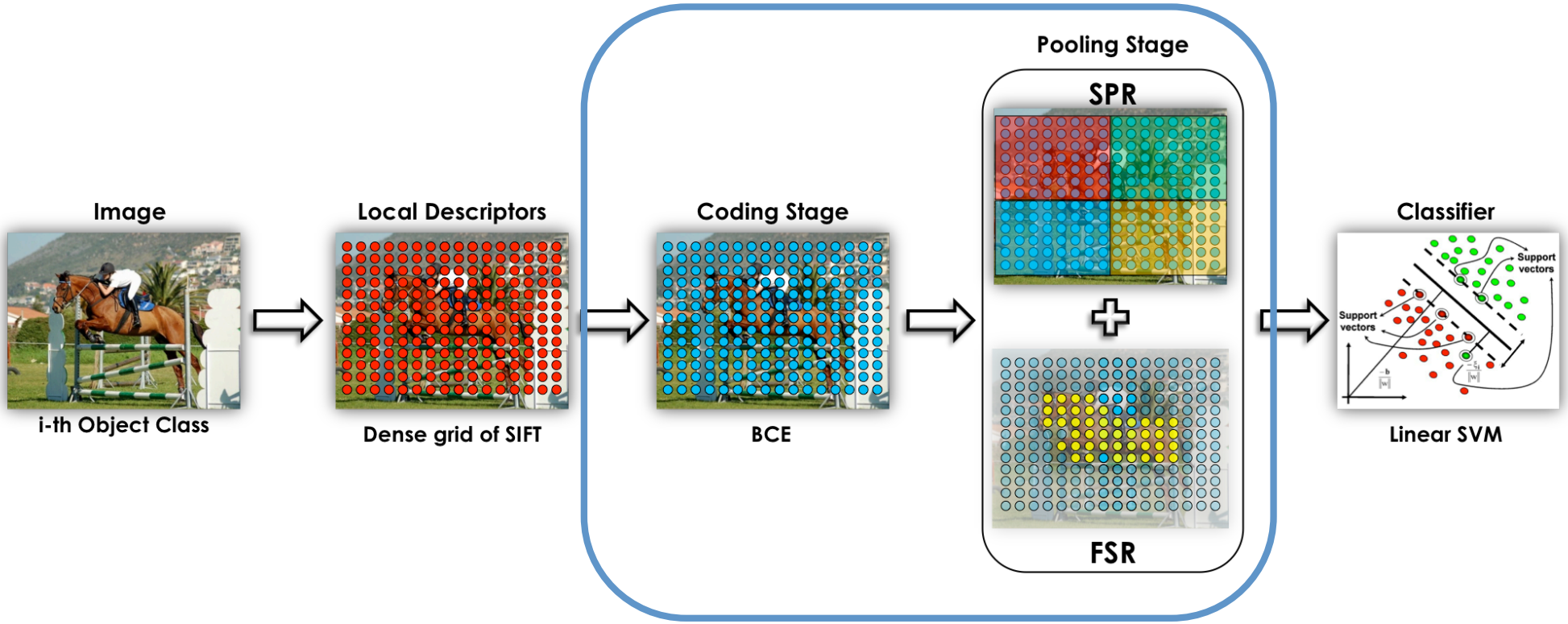
PRO:

Very Accurate

CONS:

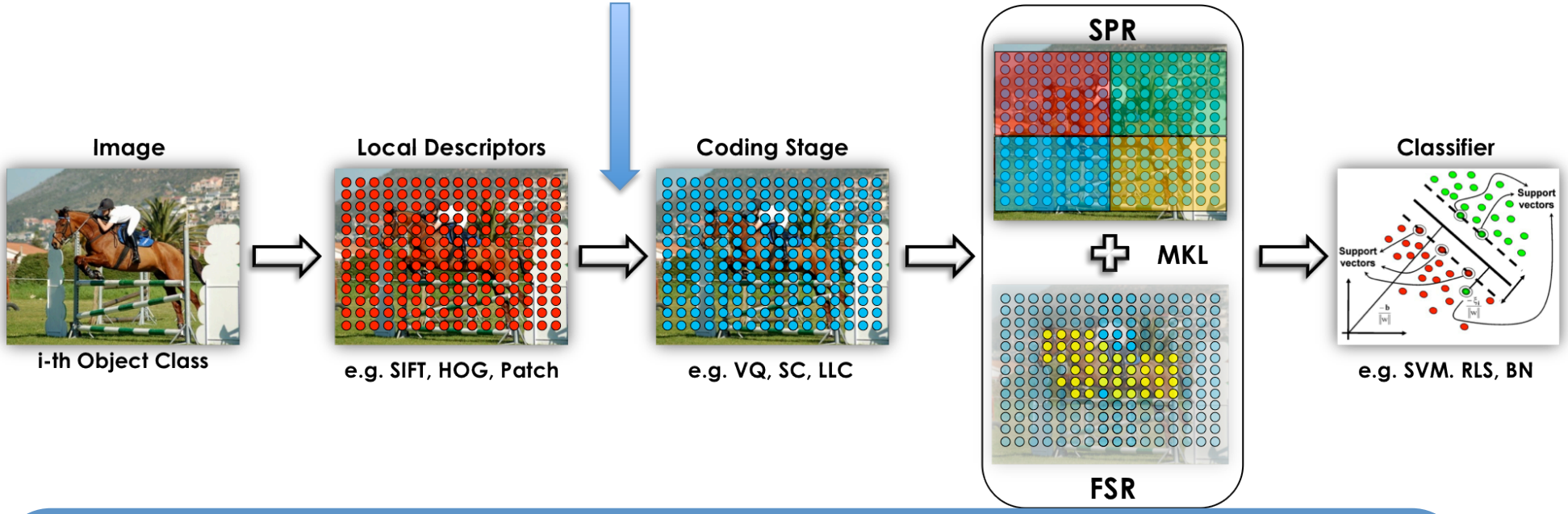
Big Data needed
Training Time
Parameter Tuning

iCub object recognition - today



- ✓ It is real-time
- ✓ It exploits supervision whenever it is possible

Unsupervised dictionary learning



K-means

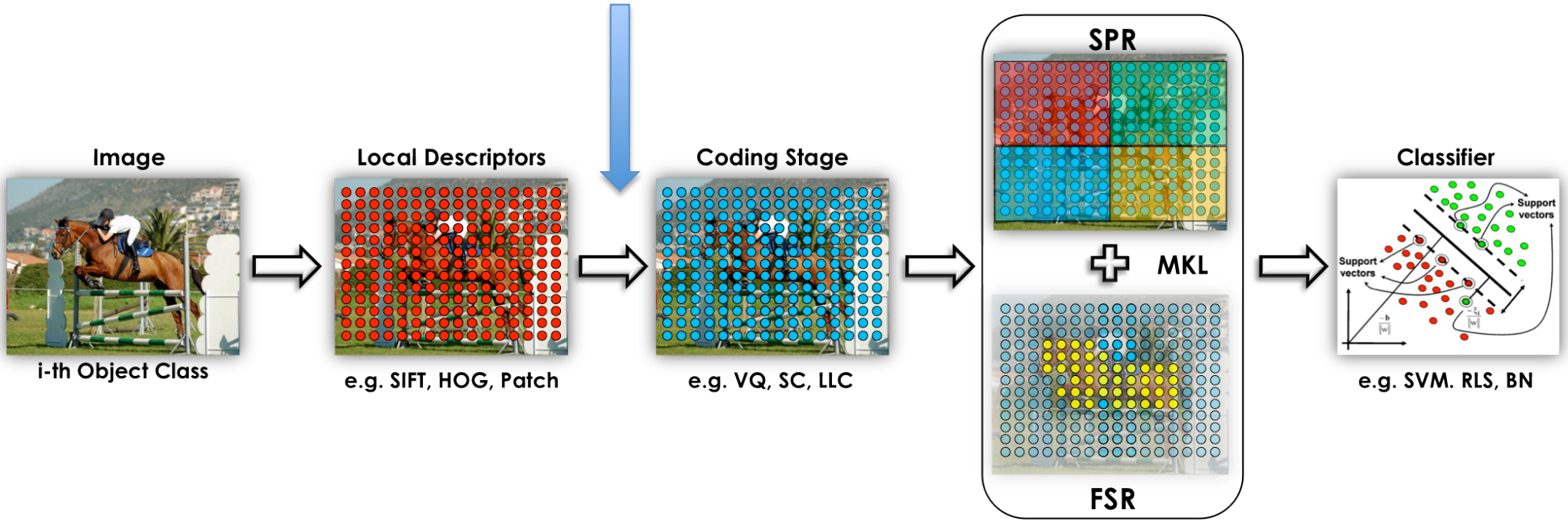
$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{X} - \mathbf{D}\mathbf{U}\|_F^2$$

$$\text{s.t.} \quad \text{Card}(\mathbf{u}_i) = 1, |\mathbf{u}_i| = 1, \mathbf{u}_i \succeq 0, \forall i = 1, \ldots, T$$

Sparse coding

$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{X} - \mathbf{D}\mathbf{U}\|_F^2 + \lambda \|\mathbf{U}\|_1$$

Discriminative dictionary learning



If supervision is available we may learn a Dictionary $\mathbf{D}^i$ for each class i.

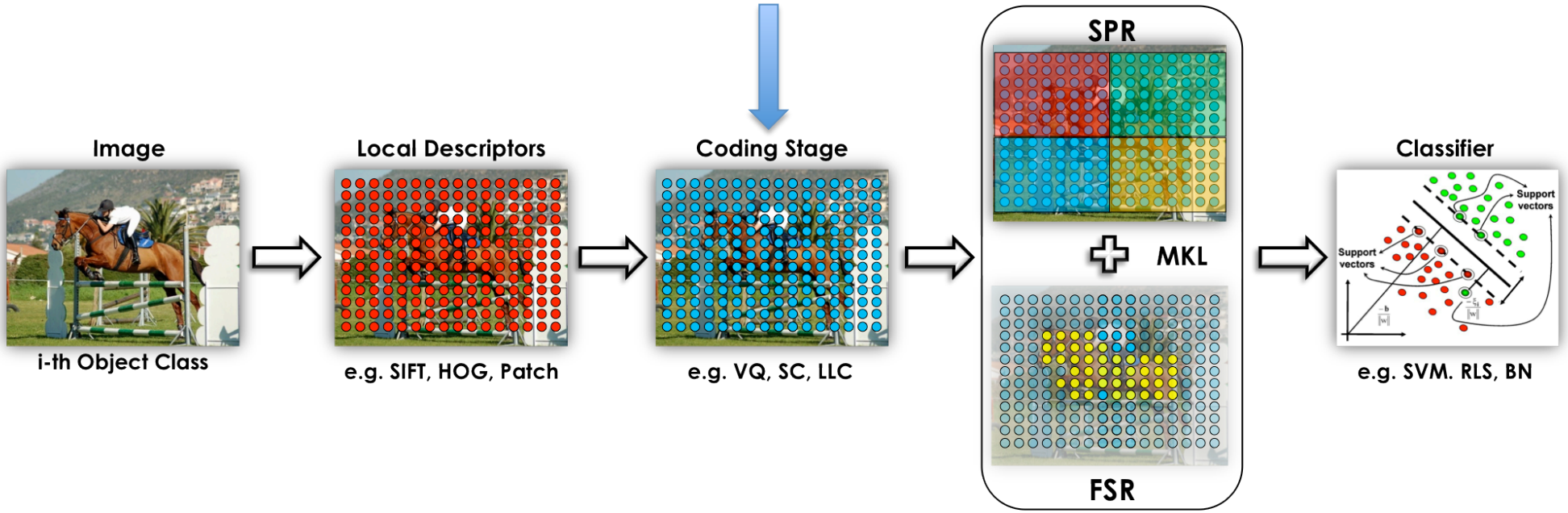
$\mathbf{X}^i$ are the examples of the positive class

$\overline{\mathbf{X}}^i$ are the examples of the negative class

$$E = ||\mathbf{X}^i - \mathbf{D}^i\mathbf{U}^i||_F^2 + ||\overline{\mathbf{X}}^i - \mathbf{D}^i\overline{\mathbf{U}}^i||_F^2 + \lambda||\mathbf{U}^i||_1 + \mu||\overline{\mathbf{U}}^i||_2$$

S.R. Fanello N. Noceti, G. Metta, F. Odone – Multi-Class Image Classification: Sparsity Does It Better, VISAPP 2013

Coding – Reconstruction error min.



Map input features $\mathbf{x}_1, \ldots, \mathbf{x}_M \in R^d$ into a new, possibly overcomplete, space of codes $\mathbf{u}_1, \ldots, \mathbf{u}_M \in R^K$ by minimizing the reconstruction error

$$\mathbf{u}_i = \arg\min_{\mathbf{u}} \|\mathbf{x} - \mathbf{D}\mathbf{u}\|^2 + \lambda \mathrm{R}(\mathbf{u})$$
$$\text{s.t.} \quad \mathrm{C}(\mathbf{u})$$

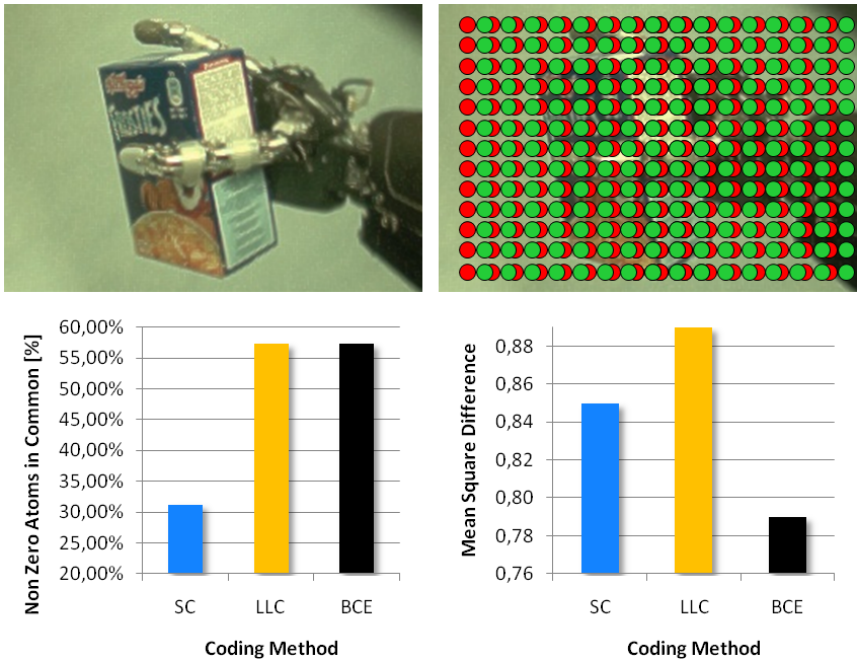
Examples: Vector Quantization (VQ), Sparse Coding (SC, Yang 2009), Locality-constrained Linear Coding (LLC, Want et al 2010)

Coding - Best Code Entries (BCE)

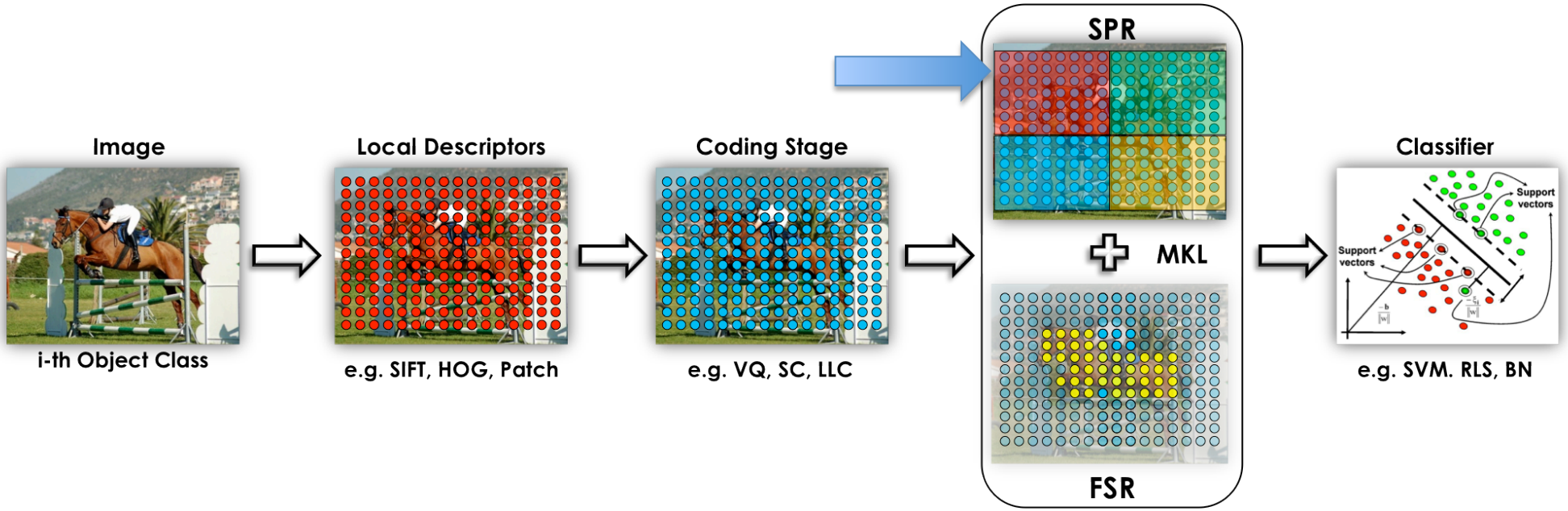
- BCE finds the atoms which are more similar to the input descriptor according to $\mathbf{u}_i = \mathbf{D}^\top \mathbf{x}_i$
- To ensure sparsity we consider only the *k* most similar atoms $\overline{\mathbf{d}}_j$

$$\overline{u}_{ij} = K(\mathbf{x}_i, \overline{\mathbf{d}}_j) \quad \forall j = 1, \ldots, k$$

Our coding ensures **stable codes** in terms of dictionary activations and responses

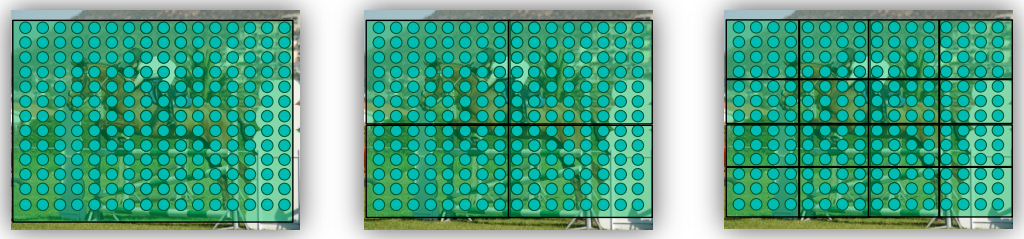


Pooling on the spatial pyramid



To capture statistical properties at higher scales, while encoding spatial relationships among elements of the scene (Lazepbik, 2006)

Image partioned in $2^l \times 2^l$ *segments*



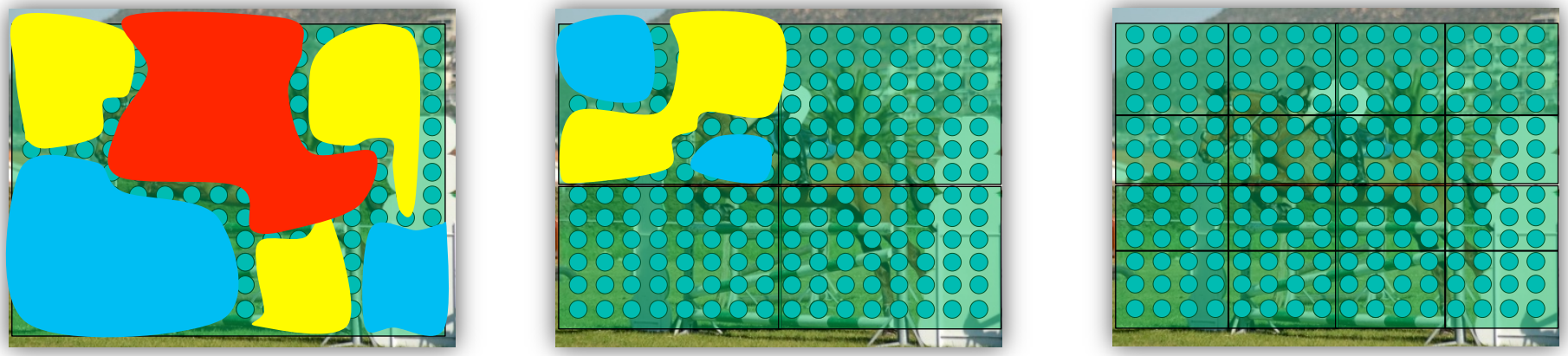
l=0 l=1 l=2

21 Spatial Bins

Pooling in the feature space

Main Idea:
Searching for the configuration of the features within a cell of the spatial pyramid.

S Spatial Pyramid Bins (S=21)
P Feature Bins (P=64)



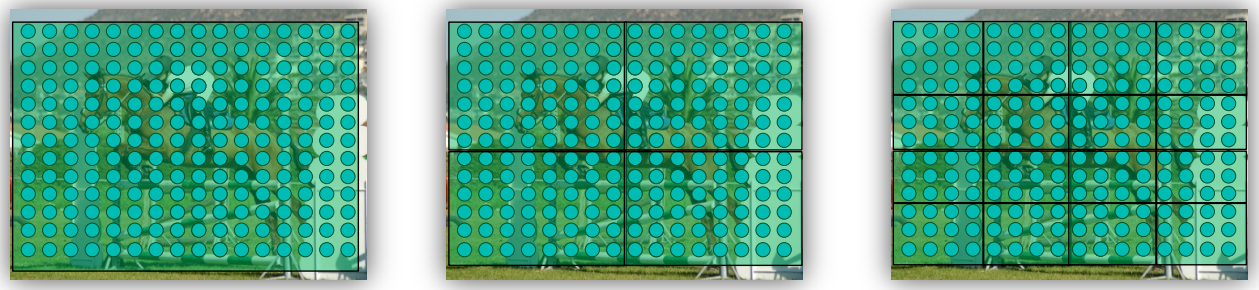
With dictionary size K=1024 the final descriptor is $\mathbf{z} \in \mathbb{R}^{K \times S \times P}$

The descriptor size is 1.376.256 per image!

Boureau et al. Ask the locals: multi-way local pool- ing for image recognition. In ICCV 2011

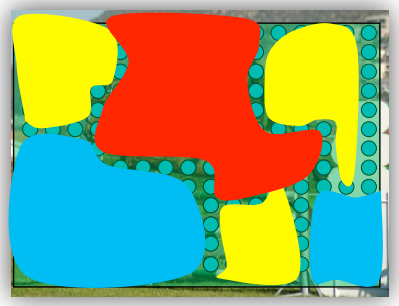
Pooling in the feature space at a lower dimension

Spatial Pyramid Representation



With dictionary size K=1024 the final descriptor is $\mathbf{z} \in \mathbb{R}^{K \times (S+P)}$

The descriptor size is 87.040 per image!

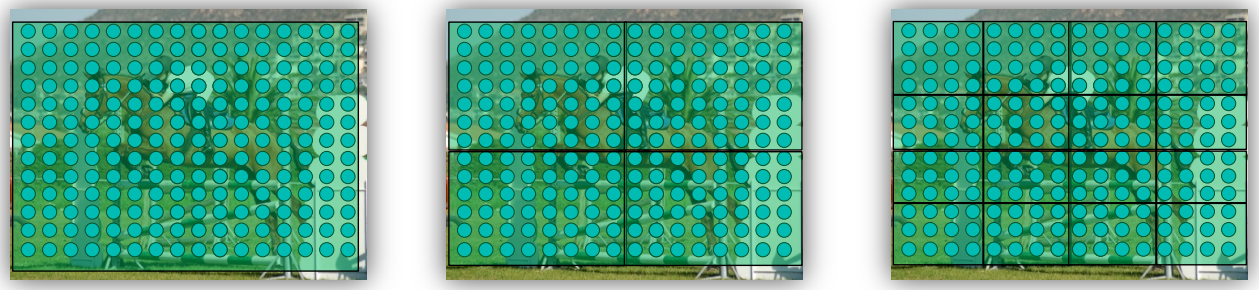


Feature Space Representation

Still Unsupervised..

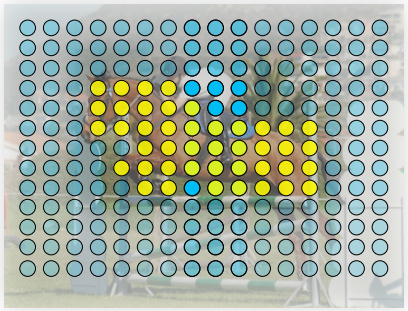
Dictionary-based pooling

Spatial Pyramid Representation



- *Each class p assigns a weight w for each code u*
- *The weight w is a confidence measure for the class p*

$$z_{p,j} = \max_i w_i^p u_{i,j} \quad \forall j = 1, \ldots, K$$



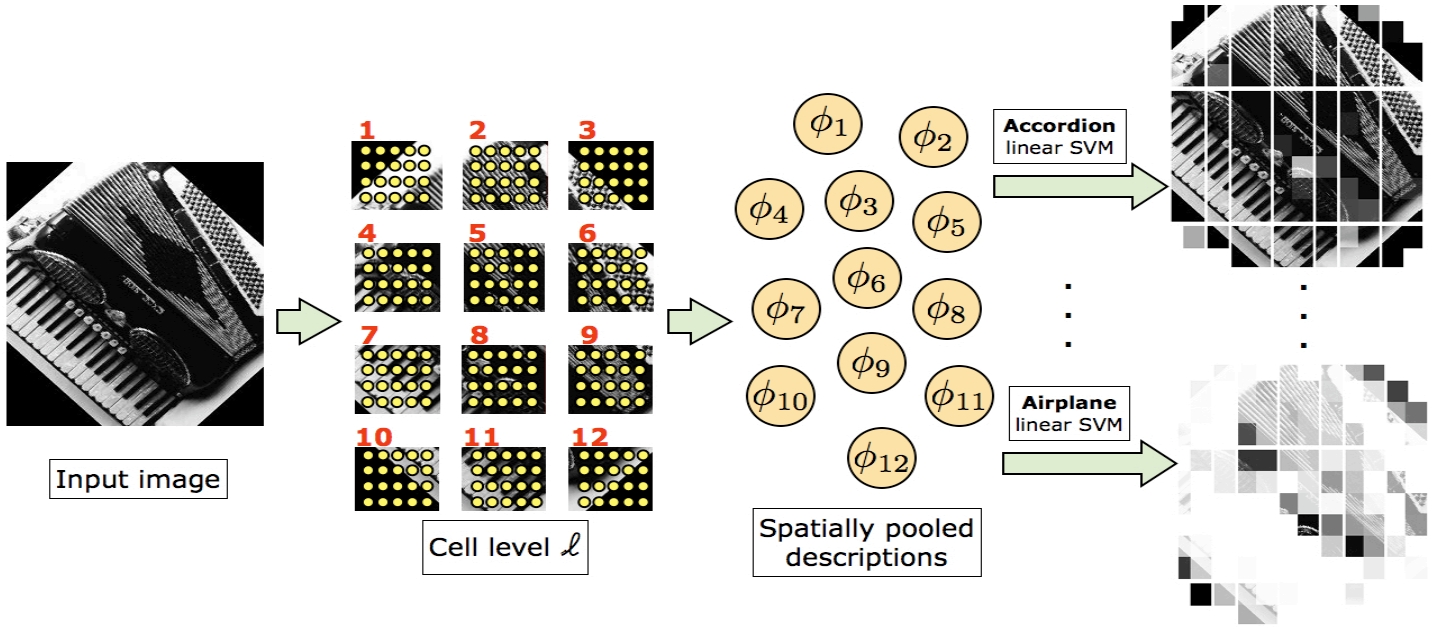
Feature Space Representation

We exploit the Supervision!

S. Fanello, N. Noceti, C. Ciliberto, G. Metta, F. Odone - **Ask the image: supervised pooling to preserve feature locality**. CVPR 2014

How to choose the weights

- How likely it is to observe a code u_i in an image of class *p*?



S. Fanello, N. Noceti, C. Ciliberto, G. Metta, F. Odone - **Ask the image: supervised pooling to preserve feature locality**. CVPR 2014

Quantitative results

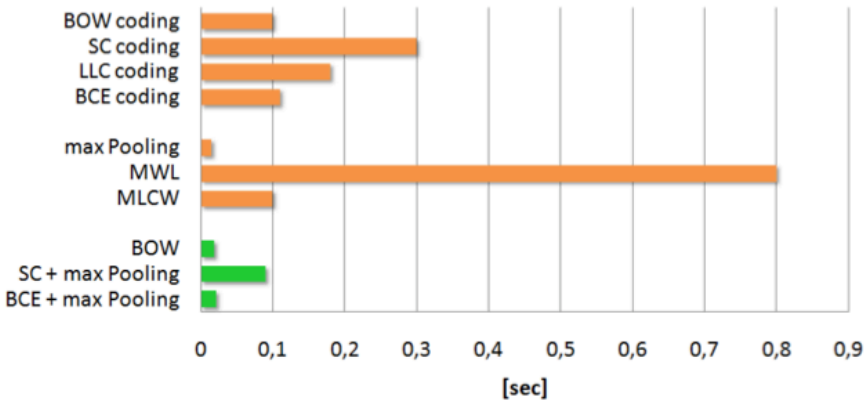
Caltech-256

	Method	Training Class Size 15	30	45	60
CODING	BOW	30.7	36.4	39.6	41.2
	SC	27.7	34.0	37.4	40.1
	LLC	32.0	38.4	42.2	44.3
	BCE	**32.4**	**39.0**	**43.1**	**44.9**
POOL.	max	32.0	38.4	42.2	44.3
	MWL	32.8	39.0	42.8	45.4
	MLCW	**35.2**	**40.1**	**44.9**	**47.9**

Pascal VOC 2007

CODING	BOW	51.5%
	SC	**54.0%**
	LLC	52.5%
	BCE	**53.1%**
POOLING	max	52.5%
	MWL	52.8%
	MLCW	**57.5%**

Processing Time



Method	#Train.	Feature	Pooling	K	Descr. length	Avg. acc. (%)
Wang et al.[31]	60	HOG	max pooling	4096	SK	47.7
Yang et al.[32]	60	SIFT	max pooling	1024	SK	40.1
Gemert et al. (from [31])	30(*)	SIFT	bag-of-words	128	K	27.2
Boureau et al.[3]	30(*)	SIFT	max pooling	1024	KPS	41.7
Gao et al. [14]	60	SIFT	max pooling	1024	SK	40.4
Harata et al. [16]	15(*)	SIFT	max pooling	1024	SK	30.2
Feng et al. [13]	45(**)	SIFT	geom. pooling	4096	SK	47.3
proposed	60	SIFT	superv. weight. max pool.	4096	K(N+S)	**47.9**

Comparison with State of the art (Caltech-256)

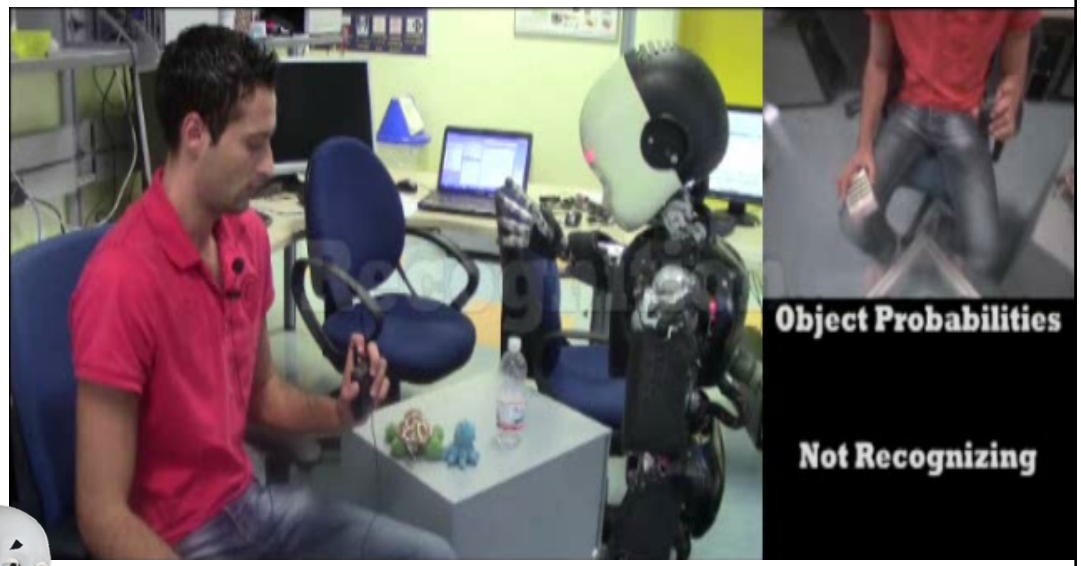
[BOW] S. Lazebnik et al.– **Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories**. CVPR 2006
[SC] J. Yang - **Linear spatial pyramid matching using sparse coding for image classification** . CVPR 2009
[LLC] J. Want et al. - **Locality-constrained Linear Coding for Image Classification.** CVPR 2010
[MWL] Y.L. Boureau et al. - **Ask the locals: multi-way local pooling for image recognition**. ICCV 2011

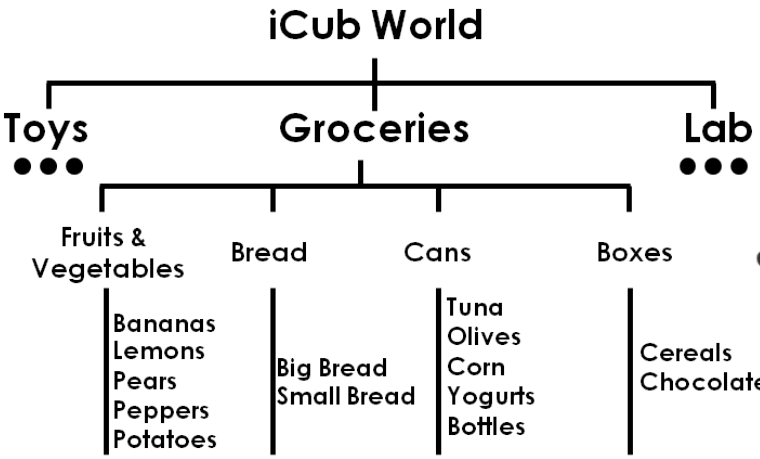
iCubWorld dataset

Self-Supervised Strategies

Kinematics

Motion





iCubWorld Dataset:
10 Categories
Bananas, Bottles, Boxes, Bread, Cans
Lemons, Pears, Peppers, Potatoes, Yogurt

HRI simplifies the **image labeling problem**: manual labeling is replaced with the use of gestures and speech

structured clutter
context cannot be exploited

iCubWorld available at: http://www.iit.it/it/projects/data-sets.html

Fanello et al. Icub world: friendly robot help building good vision datasets CVPR wshops 2013

Challenges of the iCubWorld dataset

iCubWorld Categorization Data-Set
10 Categories, 40 Objects, 2 humans, 2 domains

	T1(%)	T2(%)	T3(%)	T4(%)
BOW	78.6	27.8	29.8	14.4
SC	89.8	**38.2**	**44.0**	**19.2**
LLC	87.8	35.7	38.4	13.5
HMAX	**91.6**	36.0	41.9	17.2

T5	Whole(%)	Obj(%)	Bkg(%)
SC	99.0	80.0	68.0

Proof of concept

Issues in Real Applications:

Unbalanced Data

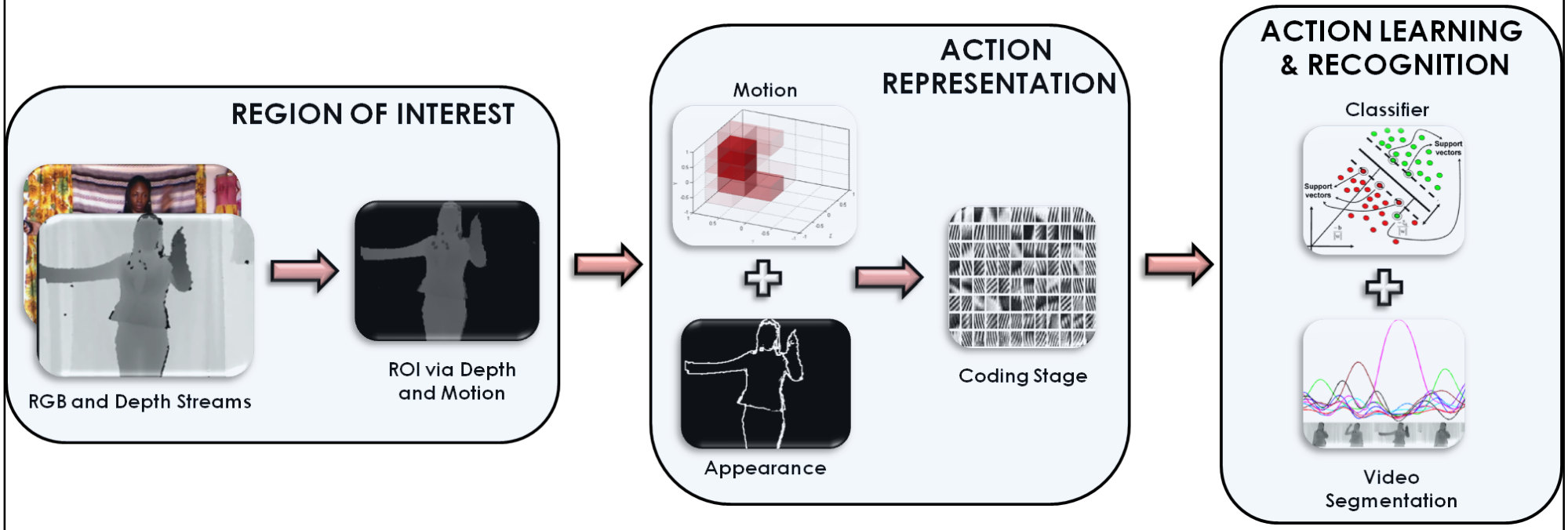
Incremental Learning

Structured Clutter

Real-Time Requirements

Integration within Different Modules

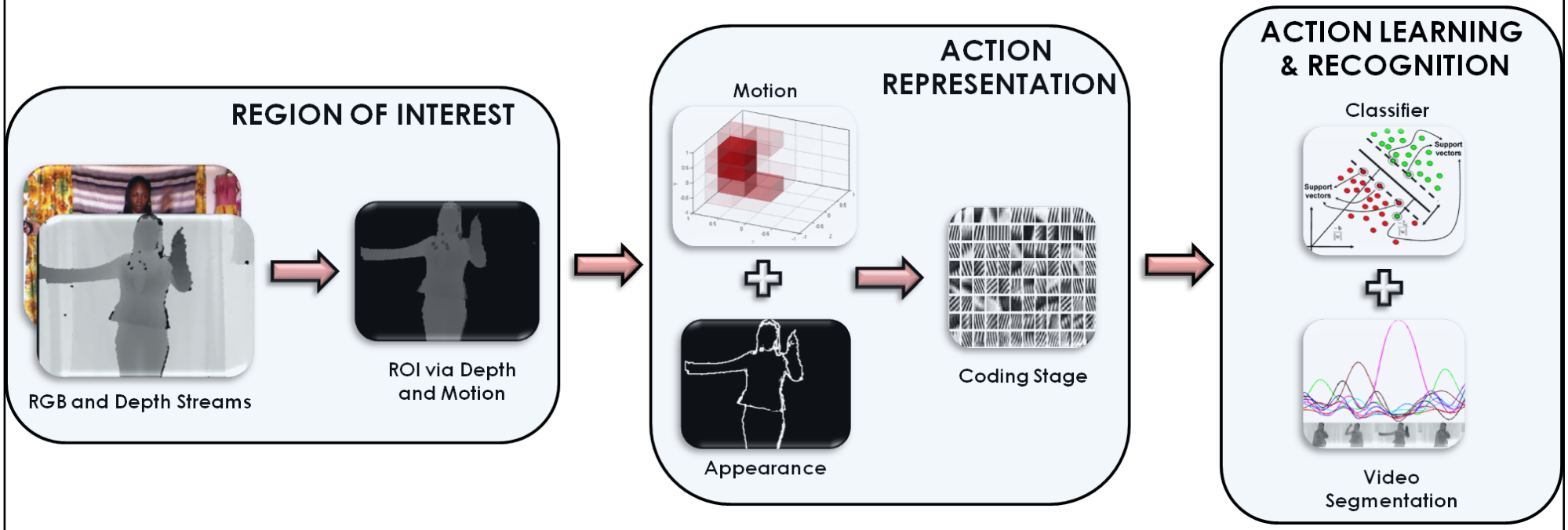
iCub action recognition - today

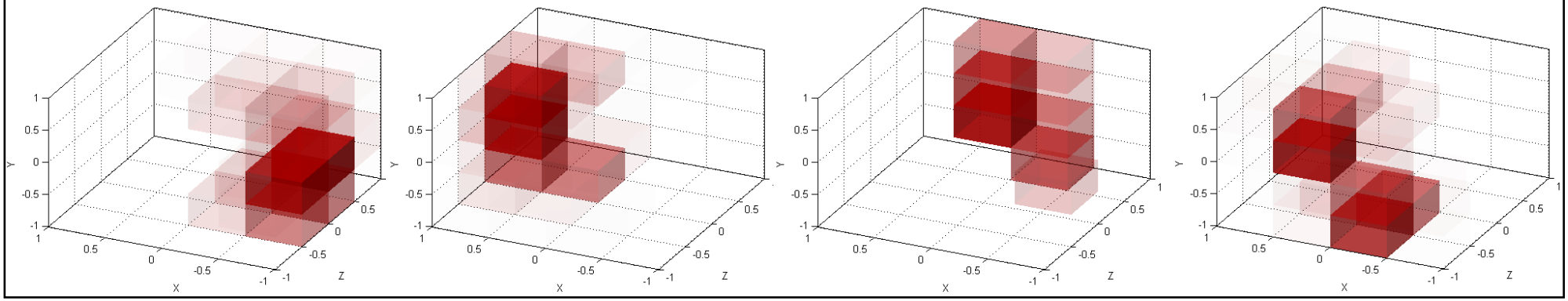


- One-shot (one video) learning
- Online video segmentation
- Real time performances

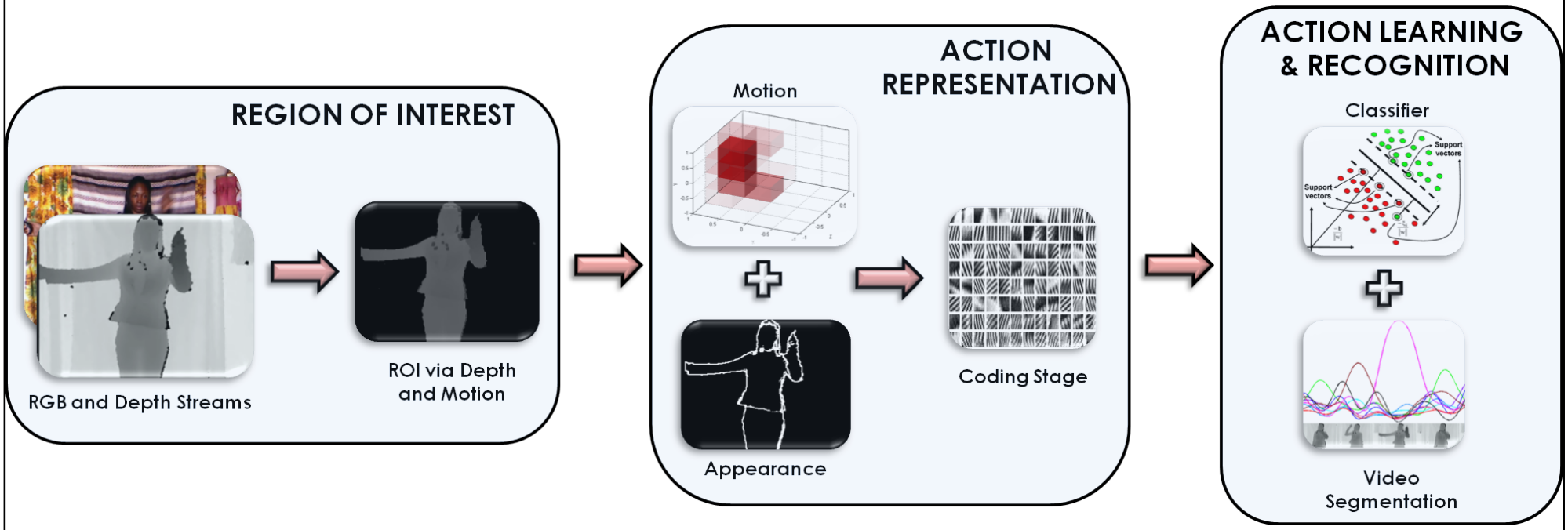
Fanello et al. Keep it simple and sparse: real-time action recognition, JMLR 2013

iCub action recognition





Motion and appearance description



$$\mathbf{m}(t) \quad \mathbf{m}(t) \sim \mathbf{D}_M \mathbf{u}_M(t)$$

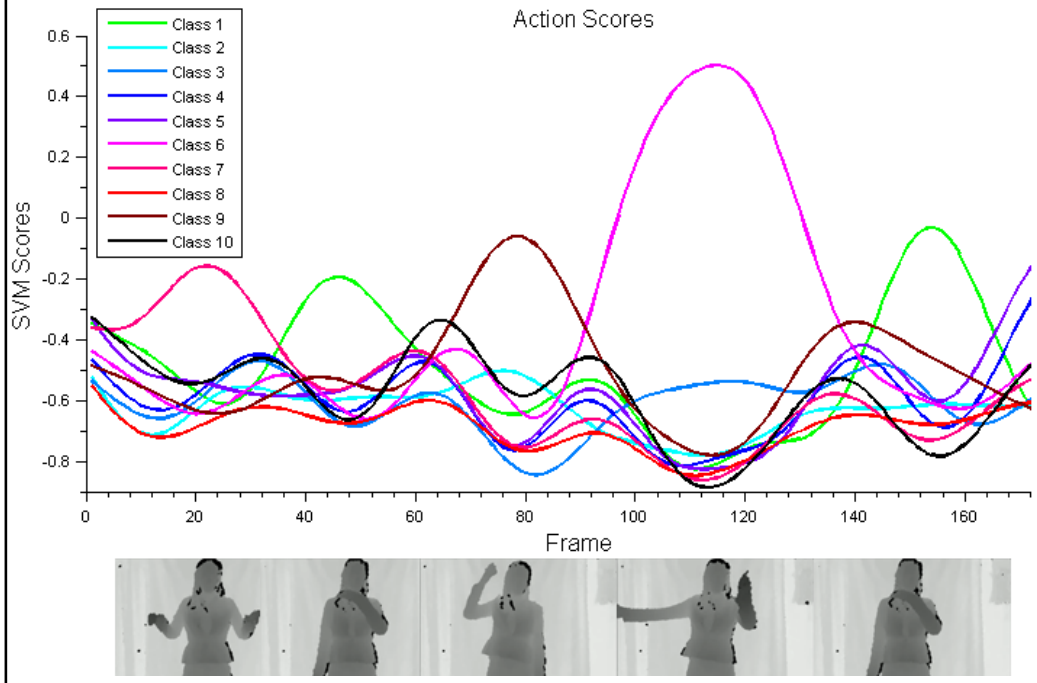
$$\mathbf{a}(t) \quad \mathbf{a}(t) \sim \mathbf{D}_A \mathbf{u}_A(t)$$

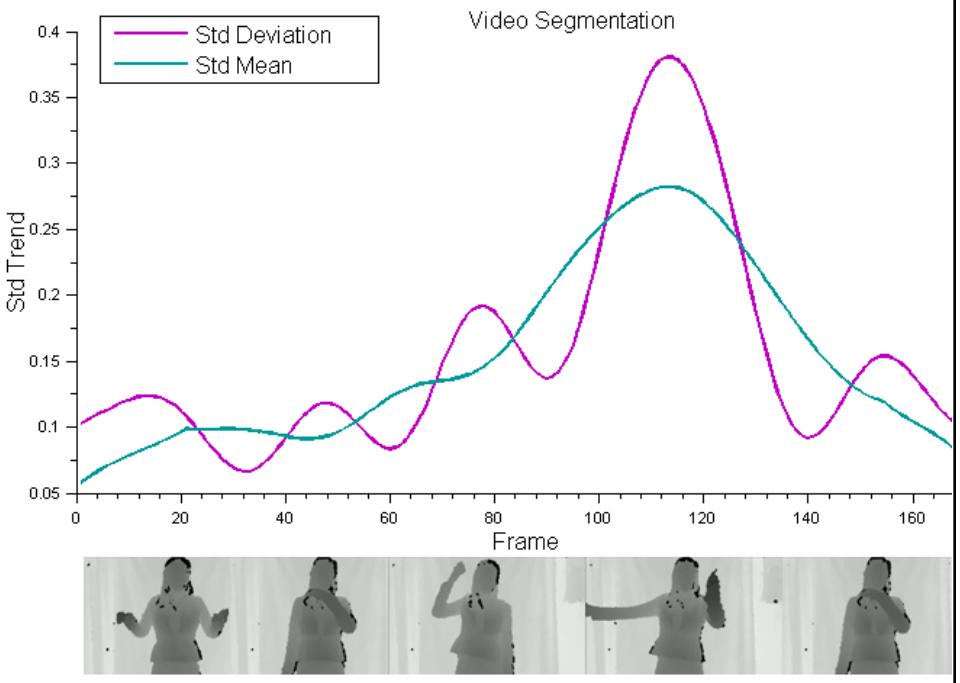
$$\mathbf{u}(t) = [\mathbf{u}_M(t), \mathbf{u}_A(t)]$$

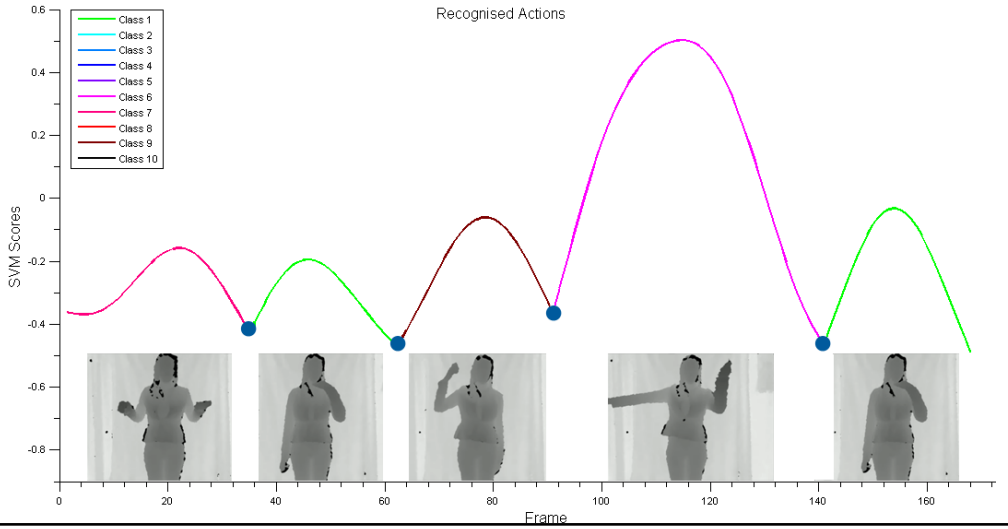
$$\mathbf{u}_T = [\mathbf{u}(T-t), \ldots, \mathbf{u}(T)]$$

Classification over a temporal window

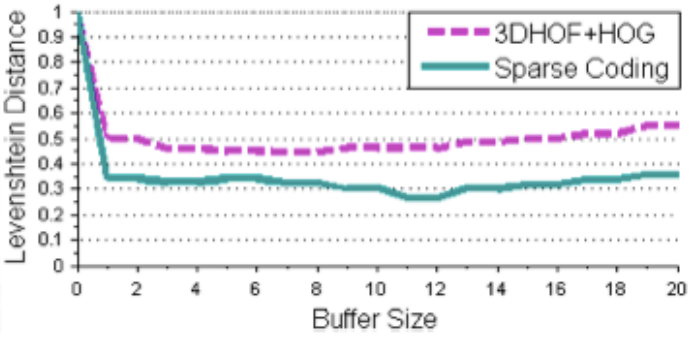
Online segmentation







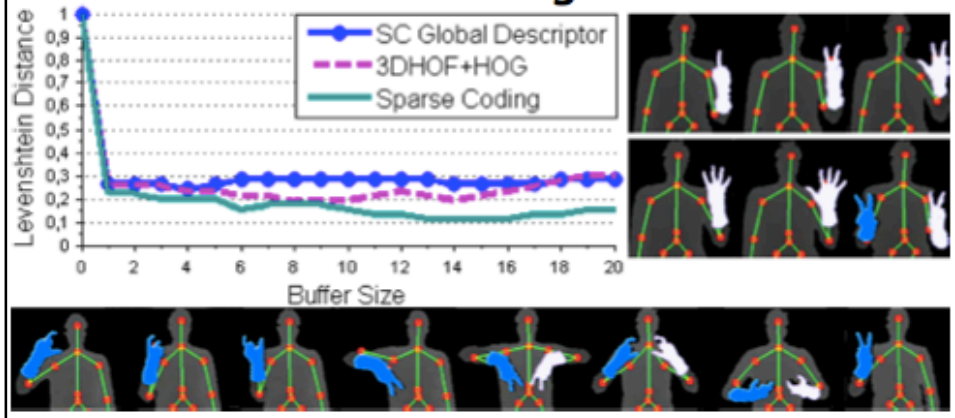
Quantitative results



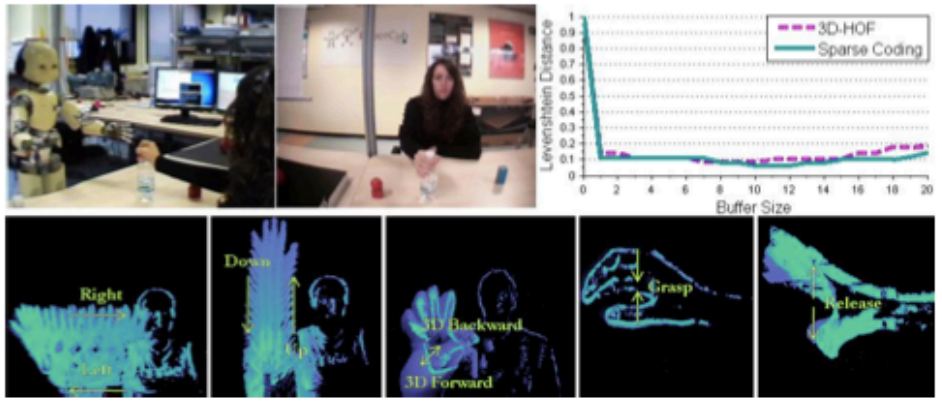
Chalearn Gesture Dataset (CGD)
Avg Err 25.11%

9th place (54 participants) at the ChaLearn Gesture Evaluation Challenge
http://gesture.chalearn.org/

Kinect Dataset Avg Err 10.11%



iCub Gesture Recognition Avg Err 9.81%



All gestures you can: a memory game

https://www.youtube.com/watch?v=U_JLoe_fT3I

Ranked II to the ChaLearn Gesture Challenge 2012 (CVPR 2012)

Wrap up

- We analysed the effectiveness of state of the art 2-layer representations for iCub

- We proposed variants which allowed us
  - To perform real-time
  - To exploit data supervision

(some) open questions

- Invariance
- What else can we learn from icub (other sensors, other senses, kinematics, …)
- Space-time representations
- ….

Thank you for your attention!