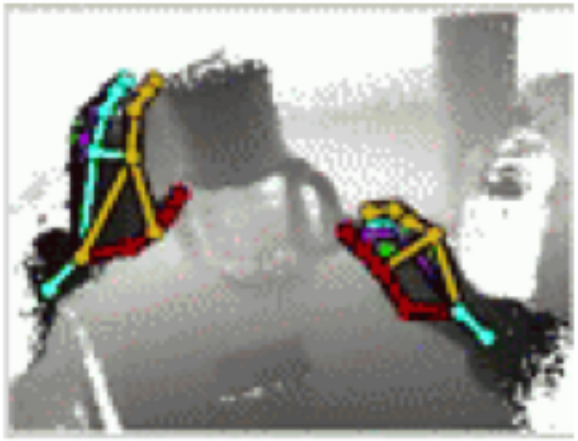
Understanding Everyday Hands in Action from a Wearable RGB-D Sensor

Grégory Rogez
(Inria Rhône-Alpes - Thoth team)



Collaborators

Deva Ramanan
CMU

J.M.M. Montiel
University of Zaragoza

James Supancic
UC Irvine

Maryam Khademi
UC Irvine

Funding

Work supported by the European Commission through Marie Curie grant PIOF-GA-2012-328288.

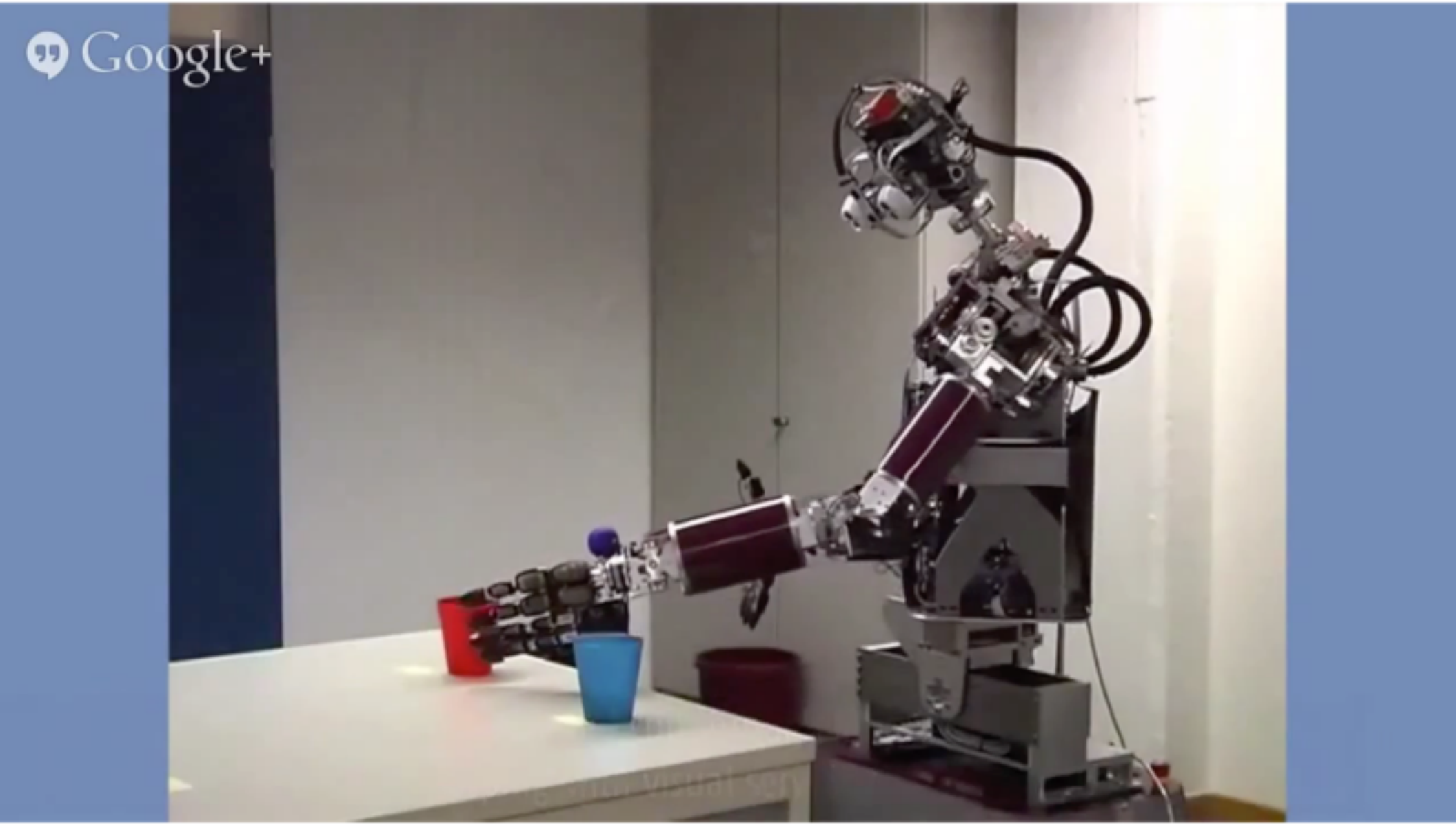
SEVENTH FRAMEWORK PROGRAMME

MARIE CURIE ACTIONS

Ego vision Health

Motivating scenarios and applications

Application in Robotics: Imitation Learning (or learning by demonstration)



courtesy of Prof. Dr. Tamim Asfour - MOOC on Humanoid Robotics Systems

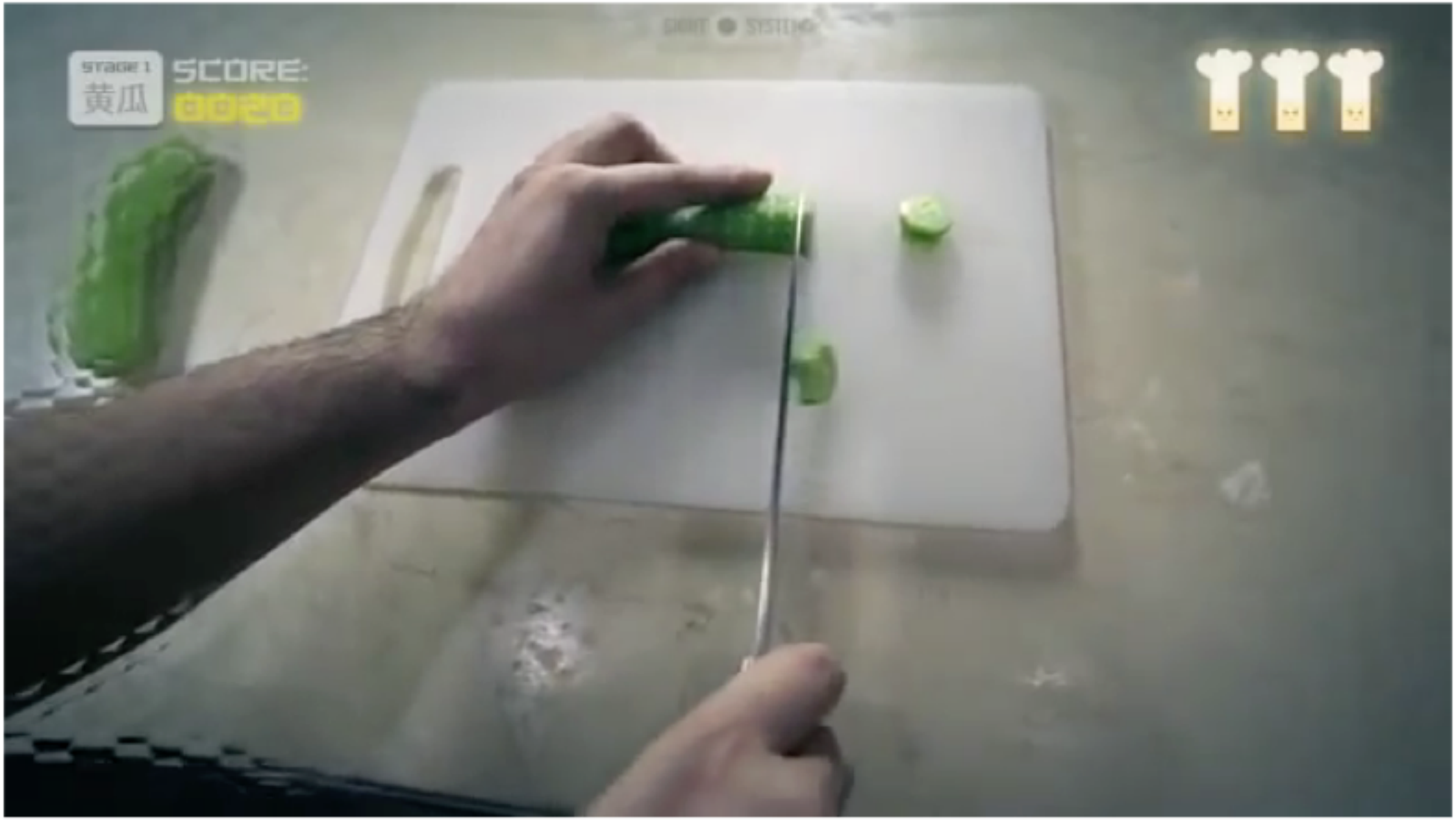
Motivating scenarios and applications

Application in Augmented Reality: Manipulation of virtual objects

Youtube video: "Meta Glasses are the future of computing"

Motivating scenarios and applications

Application in Augmented Reality: Manipulation of real objects



Youtube video: “Sight: Contact Lenses with Augmented Reality - Futuristic Video“

Motivating scenarios and applications

Application in Life-logging, Healthcare & Assistive technology

[Detecting Activities of Daily Living in First-person Camera Views, Pirsiavash & Ramanan, CVPR 2012]

Motivating scenarios and applications

Why a RGB-D sensor?

- Because it's possible..

Chest-mounted ToF camera (Intel Creative)

Motivating scenarios and applications

Why a RGB-D sensor?

- Because it's possible..

Chest-mounted ToF camera (Intel Creative)

[Rogez et al, ECCV Workshop 2014]

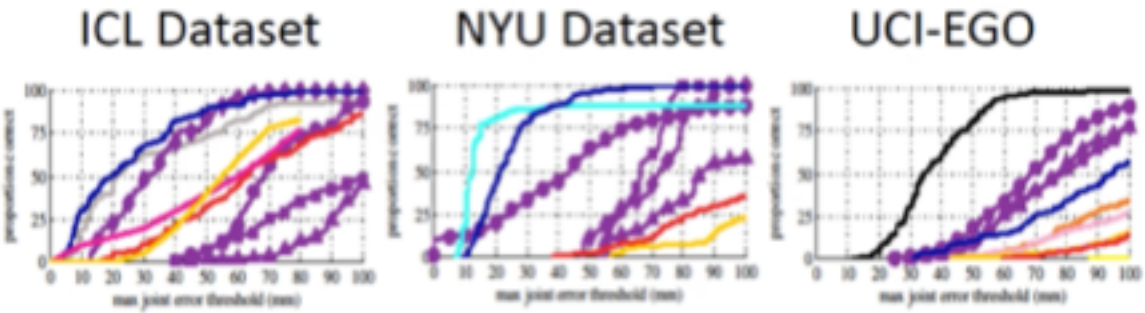
- **Accurate depth over "near-field"** reachable workspace

- Mimic near-field depth from human vision (stereopsis)

Motivating scenarios and applications

Why is it a challenging problem?

- Hands often **leave the field of view**, preventing current tracking algorithms to work.
- **Many occlusions** due to object manipulation and egocentric viewpoint (self-occlusions of the finger by the palm)
- **Cluttered** background.



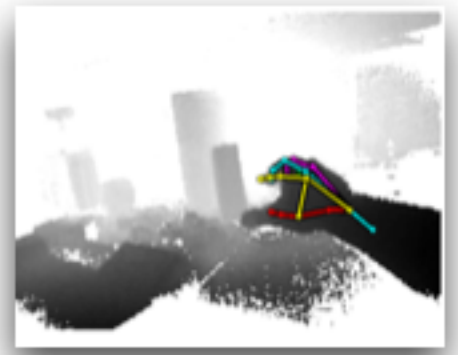
Egocentric everyday hands is the most challenging scenario

[Supancic et al, ICCV 2015]

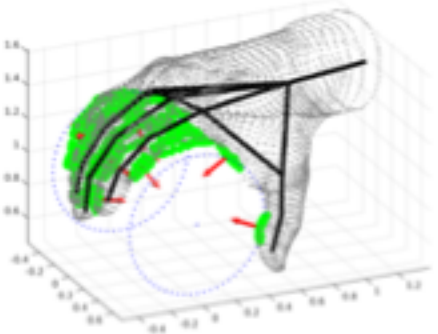
Outline

Part I: Data synthesis

Part II: Hand pose estimation



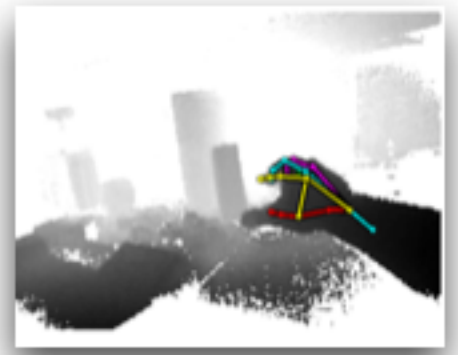
Part III: Functional understanding



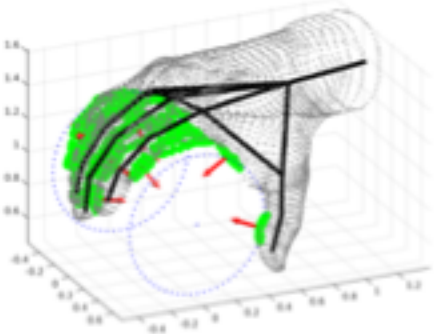
Outline

Part I: Data synthesis

Part II: Hand pose estimation



Part III: Functional understanding

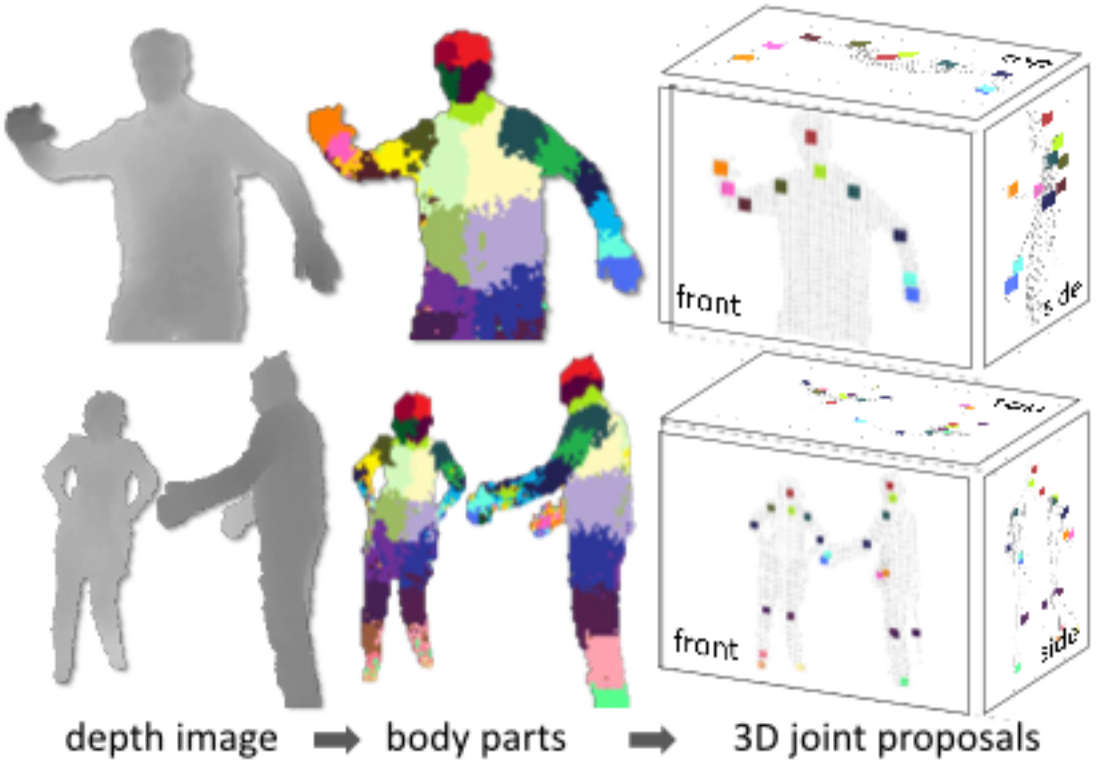


Data synthesis

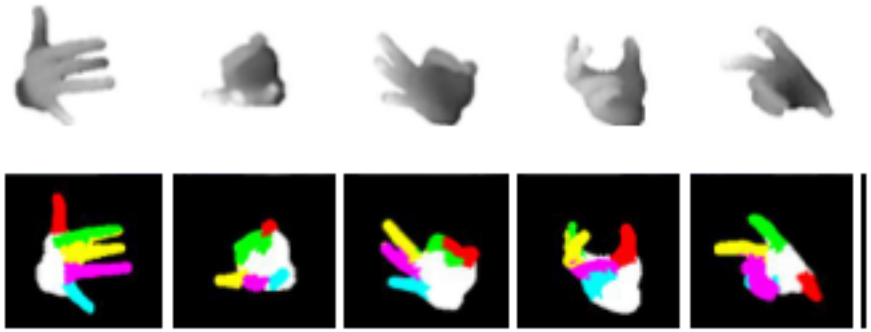
Significant advantage of Depth:

1) **Easier synthesis of training data (compared to RGB)**

Example: Microsoft Kinect Pose Estimation System: generate millions of synthetic images using 3D models of people



[Shotton et al, CVPR 2011]

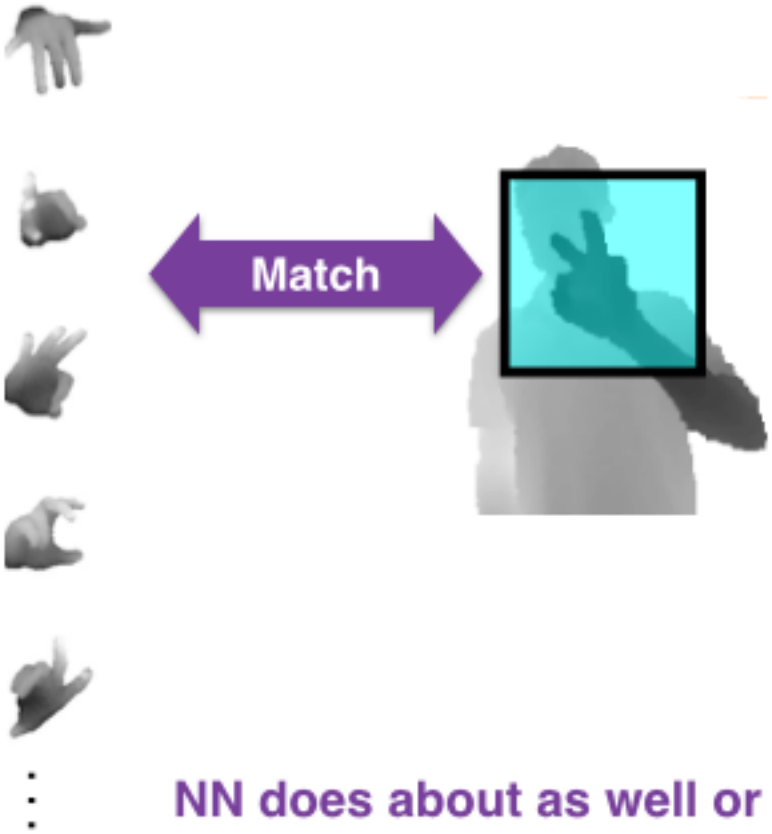


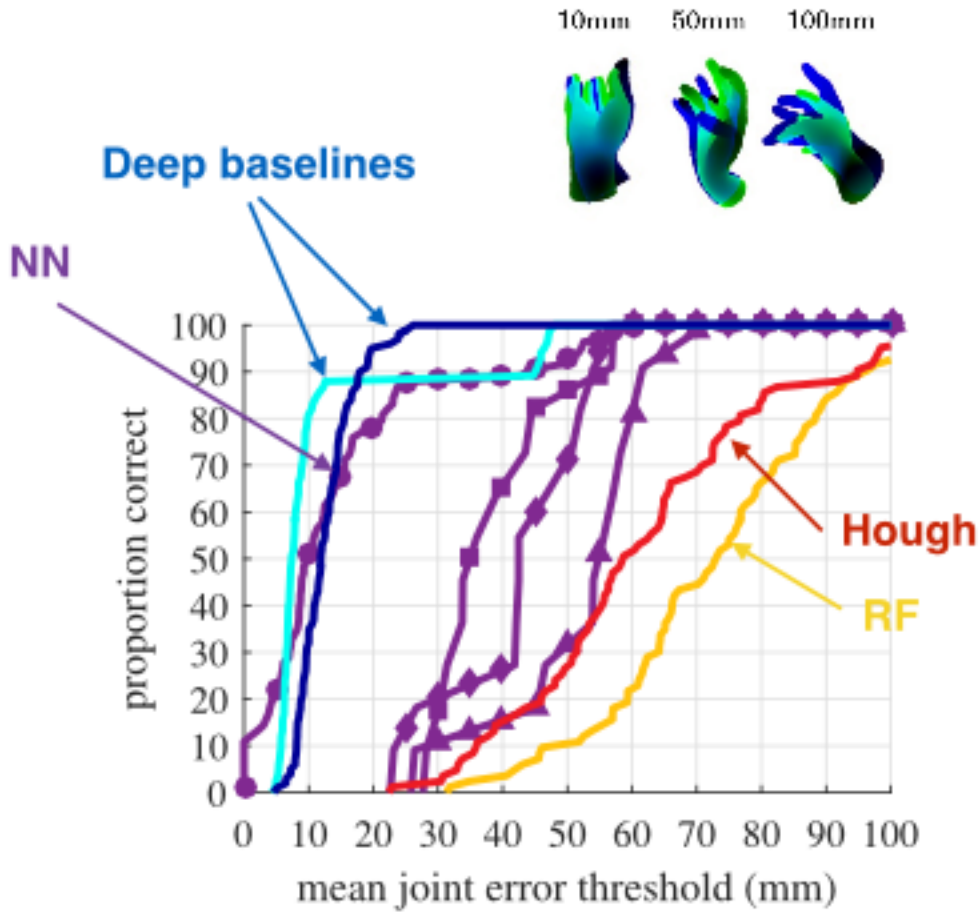
[Quian et al CVPR 2014]

Data synthesis

Significant advantage of Depth:

2) Recognition with simple matching





NN does about as well or better than almost all existing methods
It's All About the DATA !!

[Supancic et al, ICCV 2015]

Data synthesis

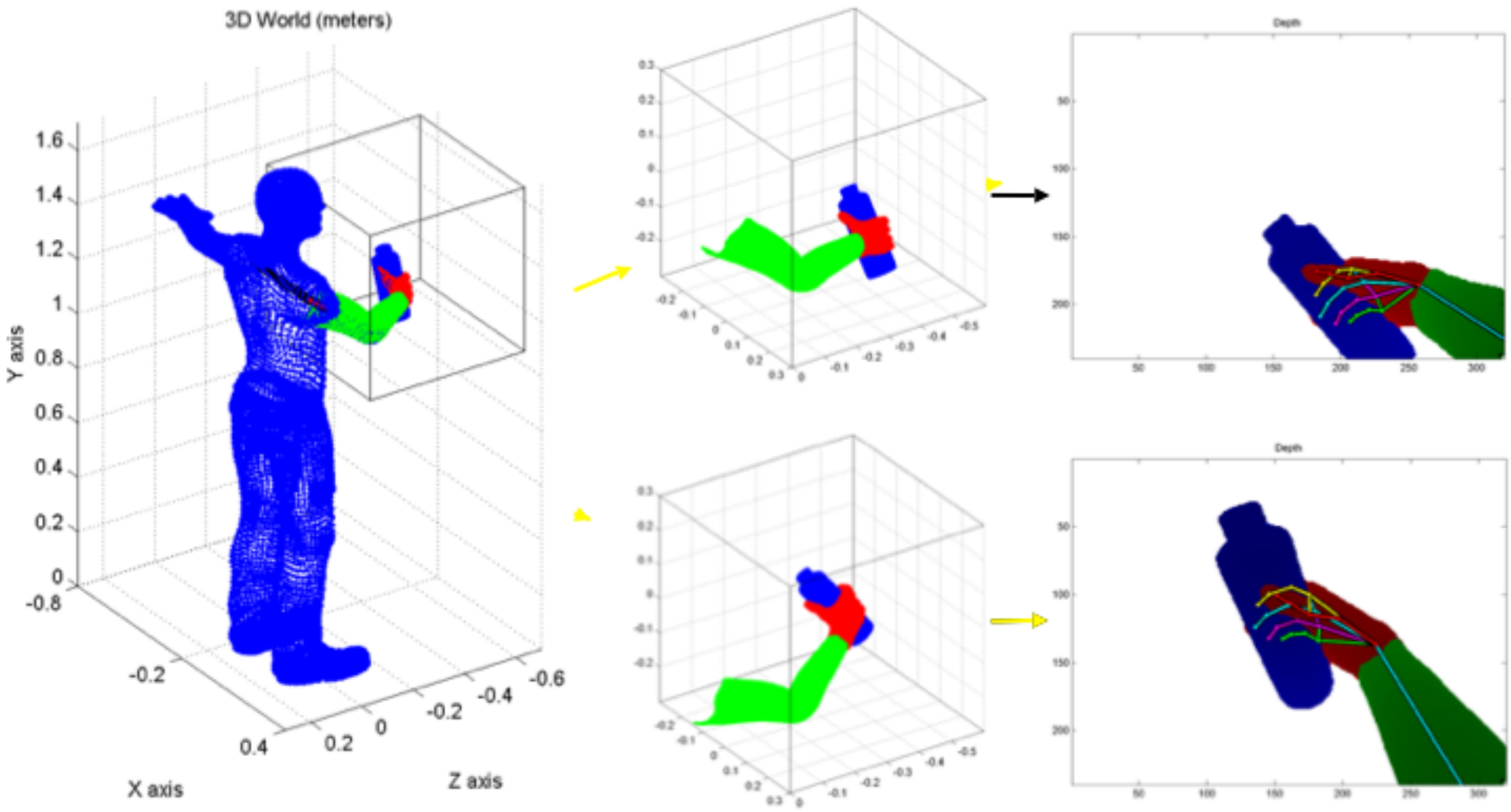
Idea: operationalize **viewpoint and pose priors** provided by egocentric settings to synthesize realistic exemplars/scenes of what a virtual character could do

- Mount avatar with **virtual egocentric camera**

[Rogez et al, ECCVW 2014]

- Use animation library of **household objects**
- Model **functional hand** movement (e.g. grasping pose)

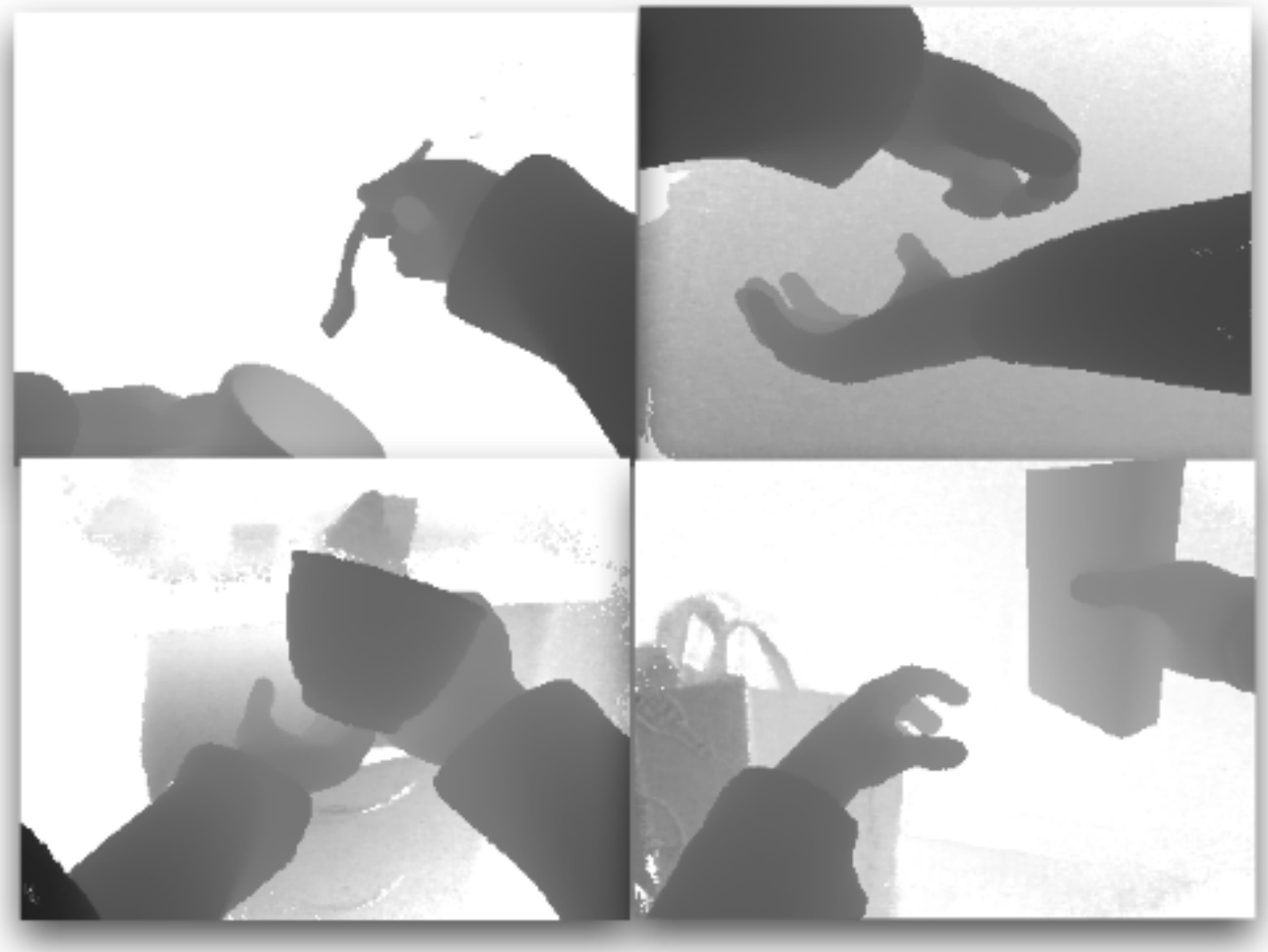
Data synthesis



For each object/grasp, we randomly perturb shoulder, arm and hand joint angles

Data synthesis

Add a **real background and a second hand**+arm
Depth map created by using a **real-world camera proj. matrix**.



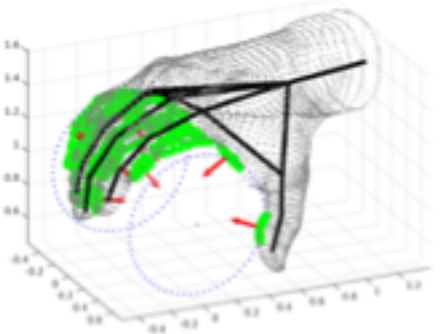
[Rogez, Supancic, Ramanan, CVPR 2015]

Outline

Part I: Data synthesis

Part II: Hand pose estimation

Part III: Functional understanding



Pose Classification

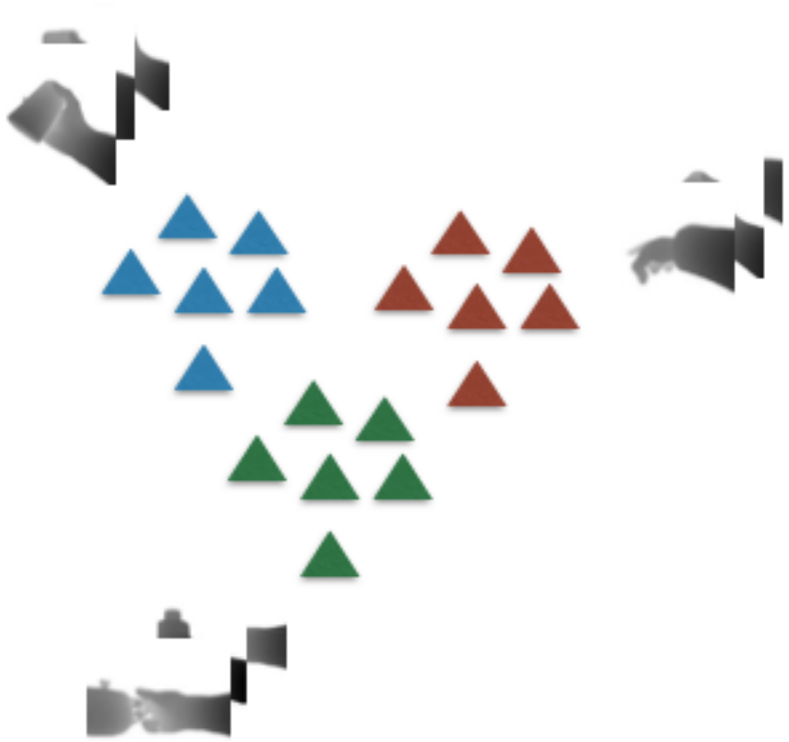
The advantage of data synthesis is that we can:

1) generate a large-scale training dataset

Pose Classification

The advantage of data synthesis is that we can:

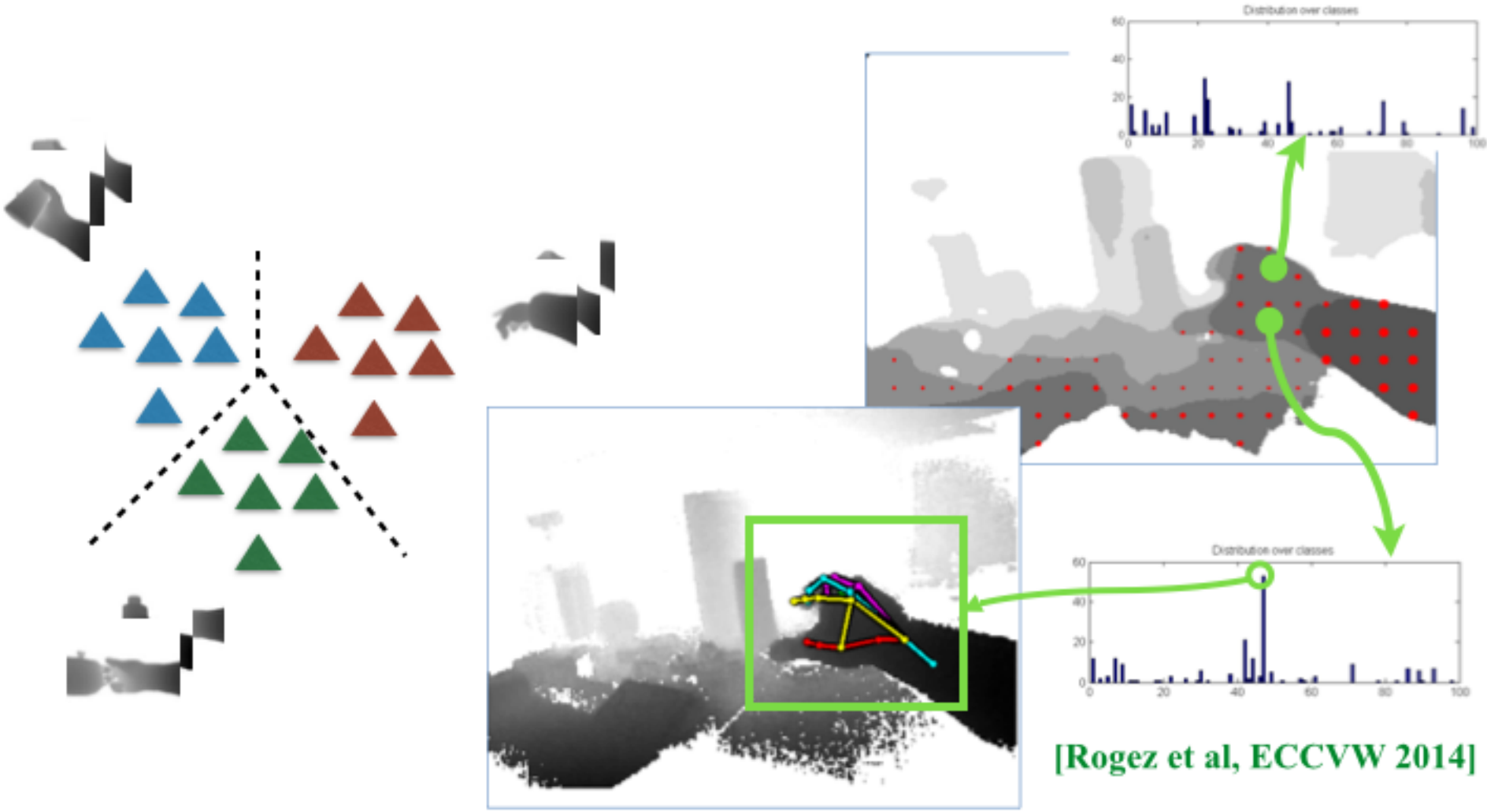
1) generate a large-scale training dataset
2) partition the entire pose space into k pose clusters



Pose Classification

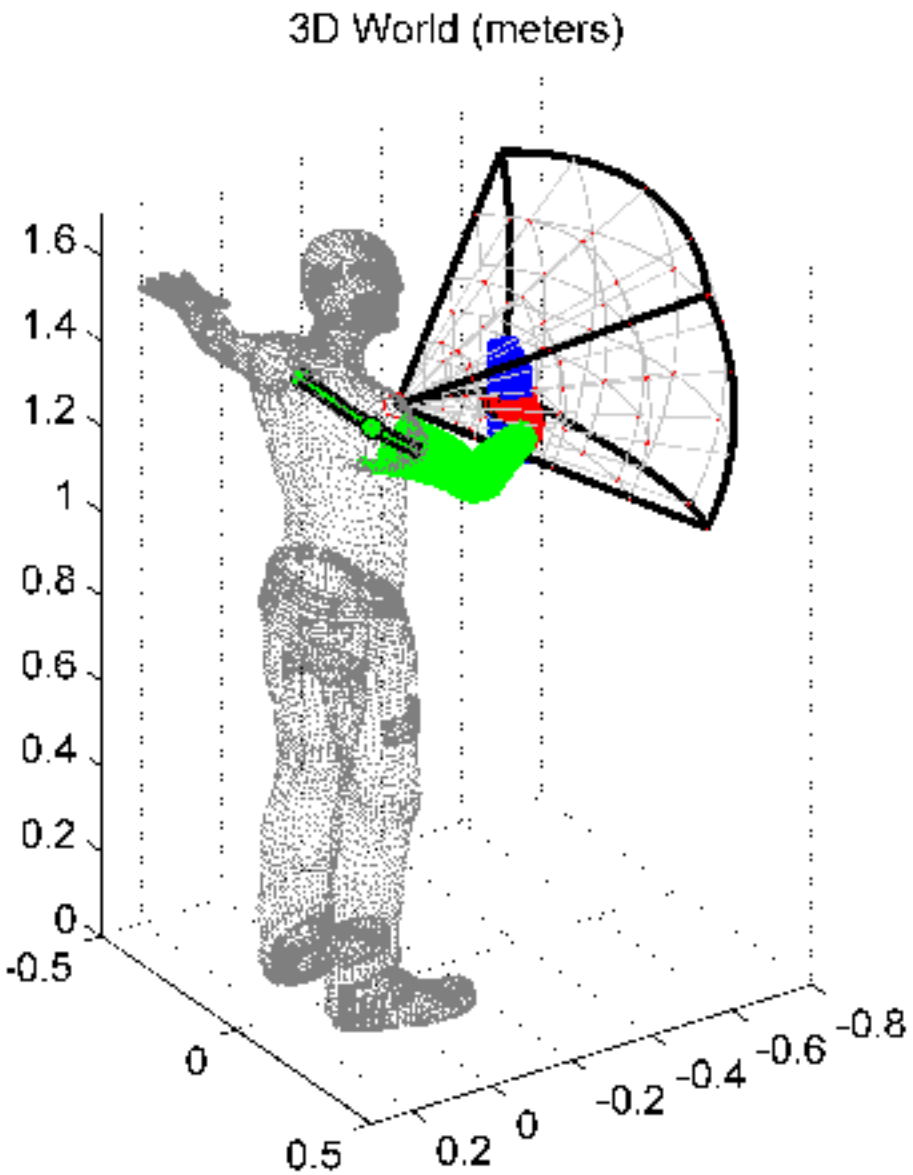
The advantage of data synthesis is that we can:

1) generate a large-scale training dataset
2) partition the entire pose space into k pose clusters
3) train a k-way classifier and tackle pose estimation/detection as a classification problem



[Rogez et al, ECCVW 2014]

Egocentric Workspace



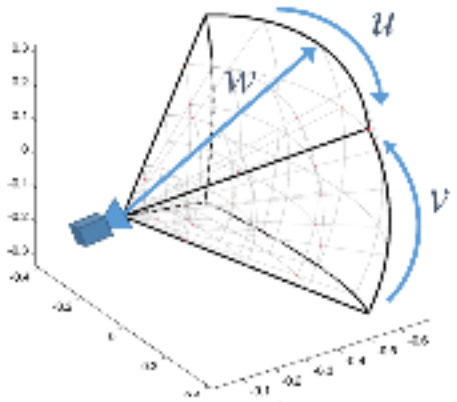
(1) Hands & Arms live in a well-defined observable volume in front of the camera: the "**Egocentric Workspace**"

(2) Hand & Arm appearance correlates with workspace location: **To exploit this, we classify arm+hand configurations in the global egocentric workspace (rather than a local scanning window).**

[Rogez, Supancic, Ramanan, CVPR 2015]

Perspective-aware binarized depth feature

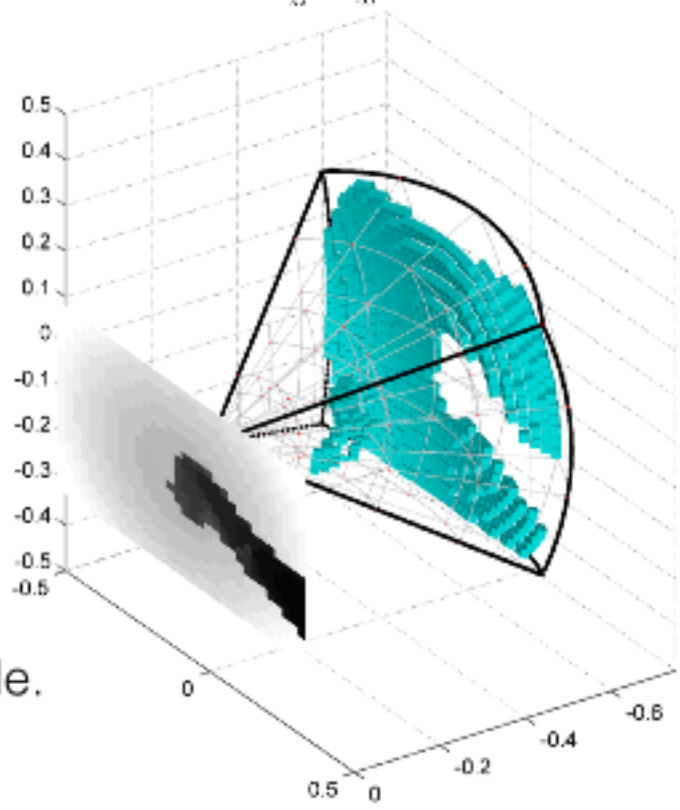
- We use a **spherical binning structure centered at the camera** where all voxels project to the same image area, increasing accuracy (better occlusions reasoning) and speed (sparse volumetric comp).



- Let's choose spherical bins F (u, v, w) that project to a single pixel (u, v) in the depth map.

- This allows one to compute the binary voxel grid b[u,v,w] by simply **"reading off" the depth value** for each z(u,v) coordinates, quantizing it to z':

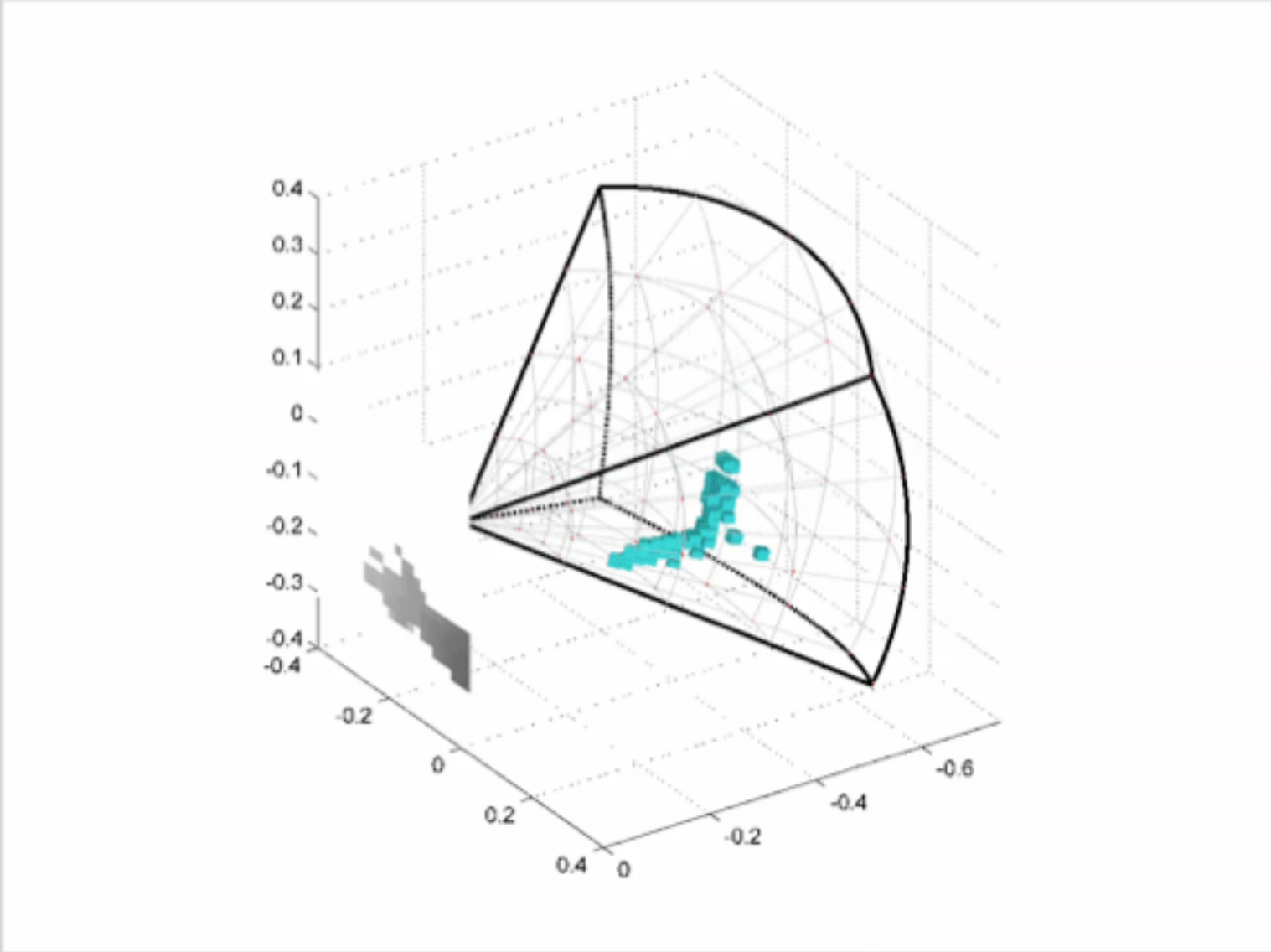
$$b[u, v, w] = \begin{cases} 1 & \text{if} \quad w = z'[u, v] \\ 0 & \text{otherwise} \end{cases}$$



- If depth observed at position b[u′, v′, w′] = 1, all voxels behind **(w ≥ w′) are occluded & defined to be "1"**.

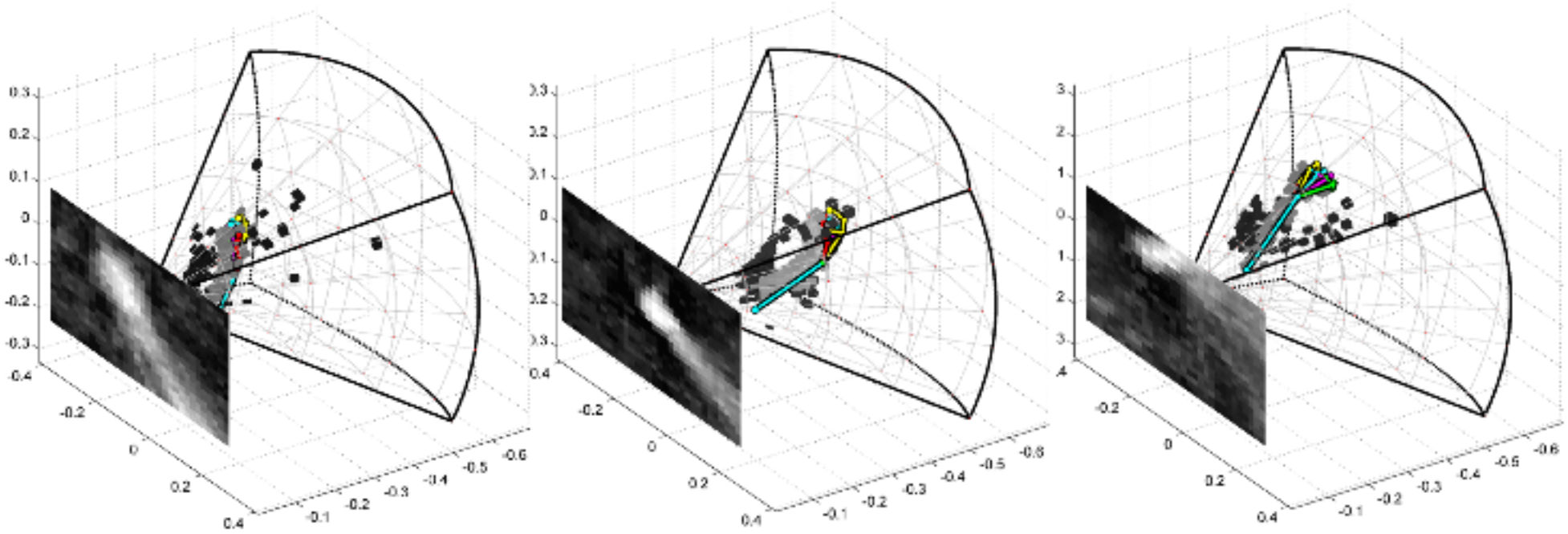
- A **coarse Nu × Nv discretization z'** (quantized in the z-direction) is considered to make the problem more tractable.

Perspective-aware binarized depth feature



Global pose classification

- We quantize the training poses into **K coarse classes** and train a K-way pose-classifier.
- We classify global depth maps quantized into our binarized depth feature b[u, v, w].
- For each class k ∈ {1, 2, ...K }, we train a **one-vs-all SVM classifier** obtaining weight vector which can be re-arranged into a $N_u \times N_v \times N_w$ tensor $\beta_k[u,v,w]$:

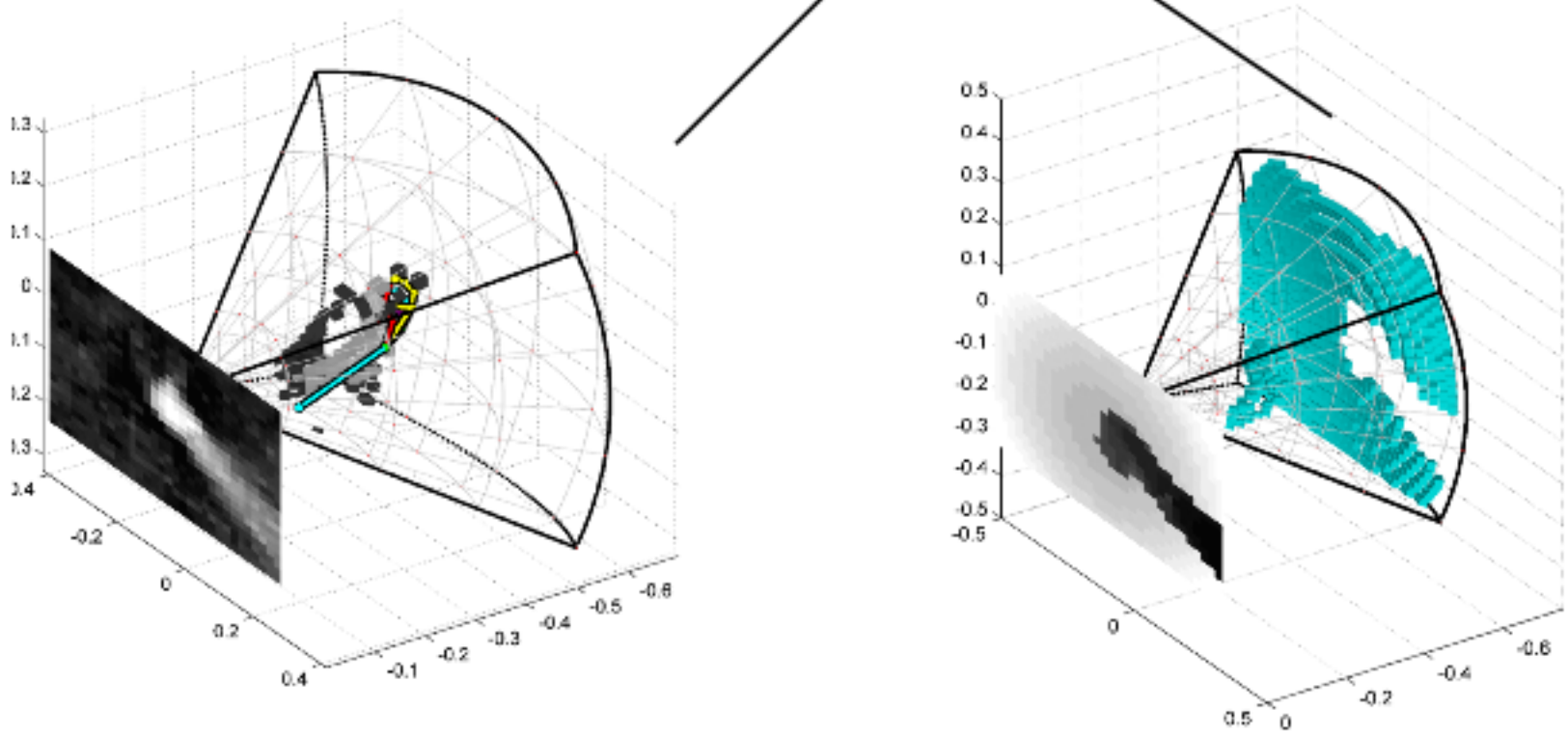


β_1 β_2 ..... β_K

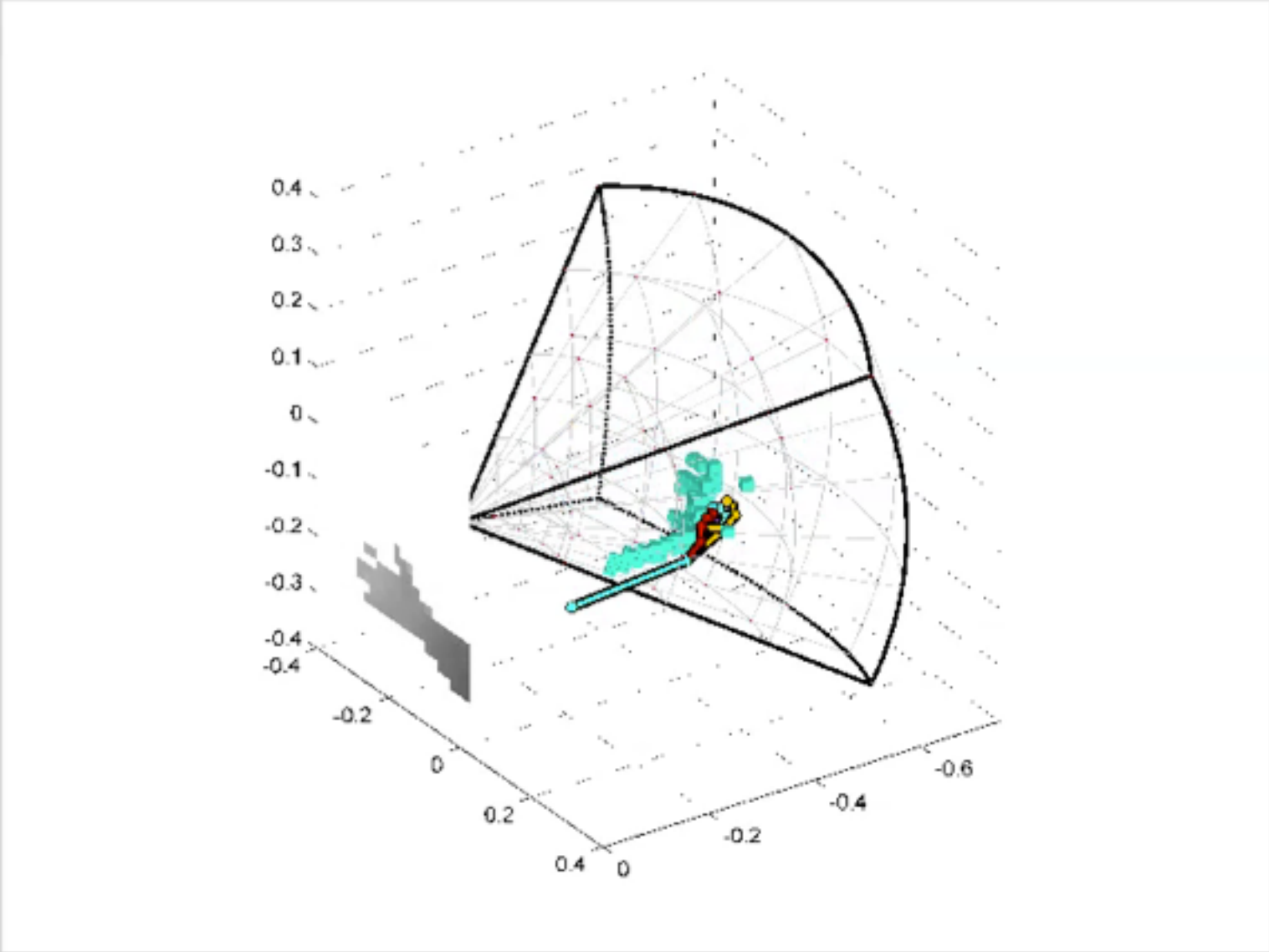
Global pose classification

- The score for class k is then obtained by a **simple dot product** of the weight tensor and b[u, v, w]:

$$\text{score}[k] = \sum_u \sum_v \sum_w \beta_k[u, v, w] \cdot b[u, v, w].$$

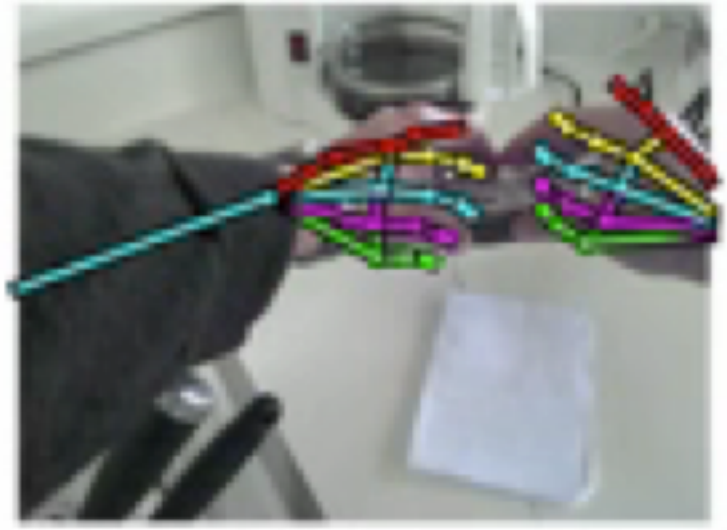


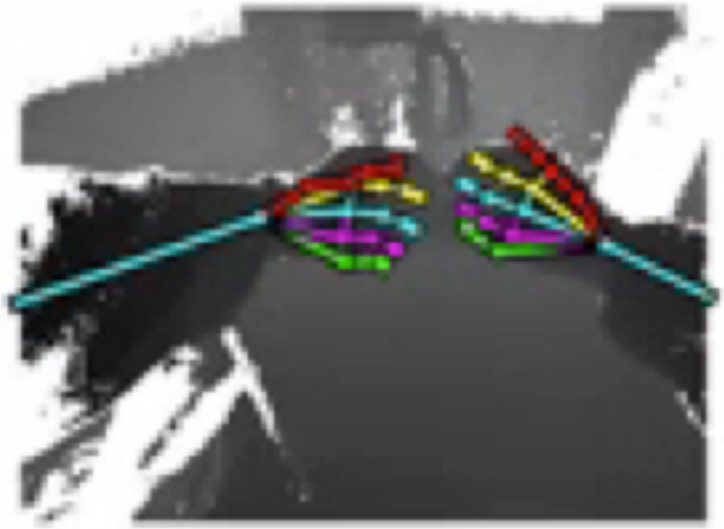
Global pose classification: qualitative results in “free space”

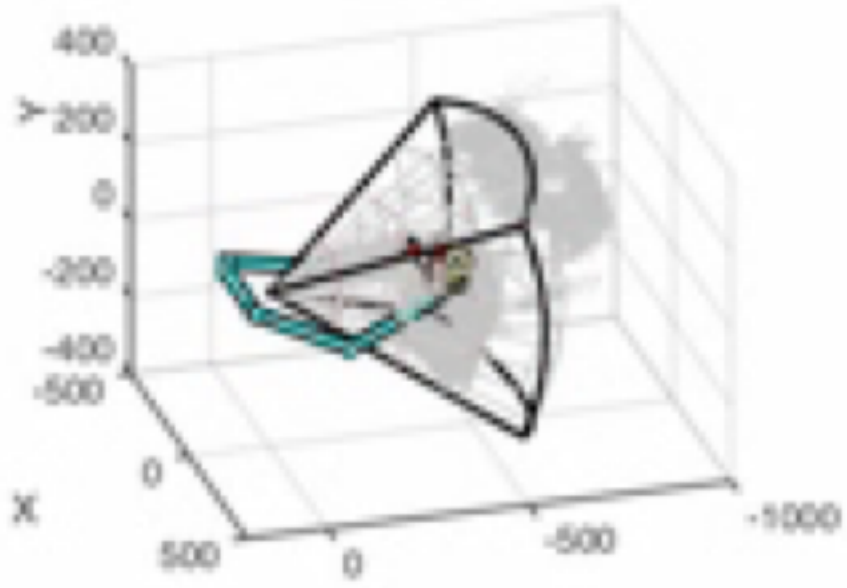


Our implementation runs at 300 fps

Global pose classification: qualitative results “in the wild”

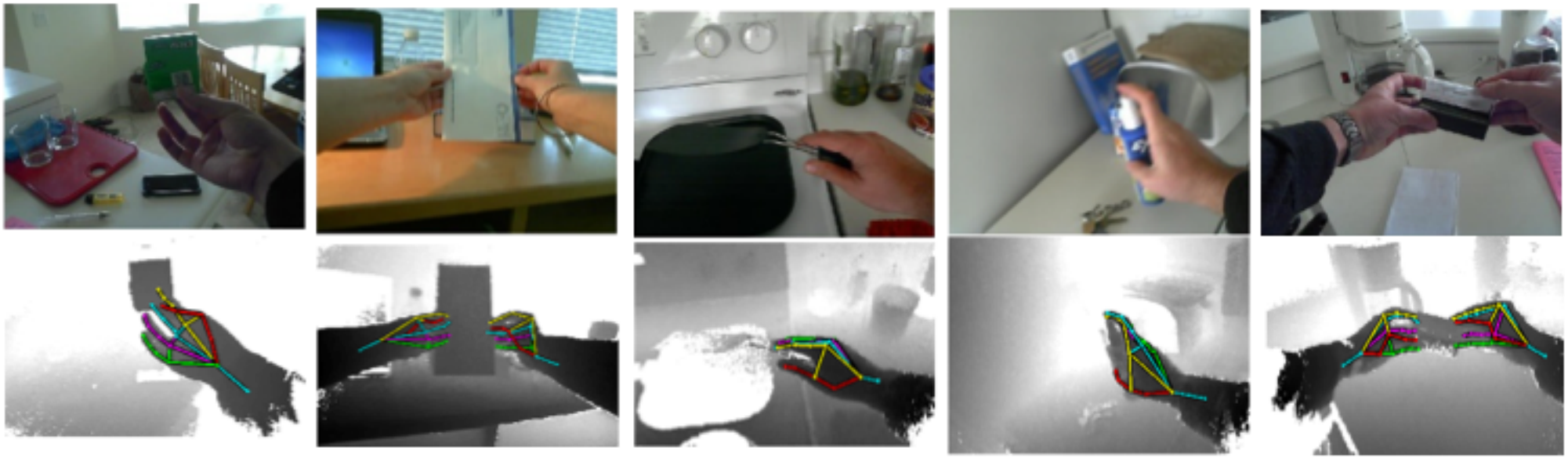






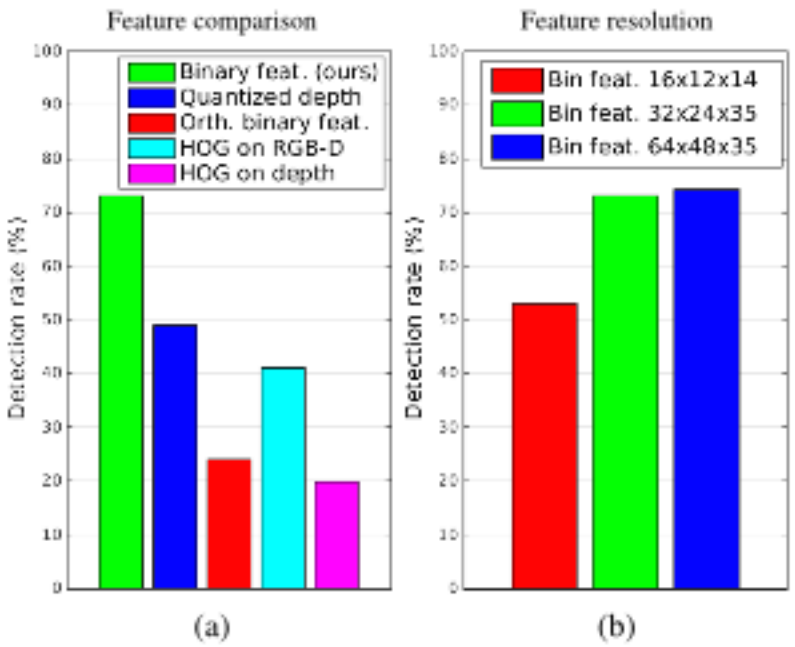
Our implementation runs at 300 fps

Numerical evaluation



- We captured & annotated videos of **real egocentric object manipulation scenes** (~4,000 frames)
- We developed a semi-automatic labelling tool that **accurately annotates partially-occluded hands** and fingers in 3D.

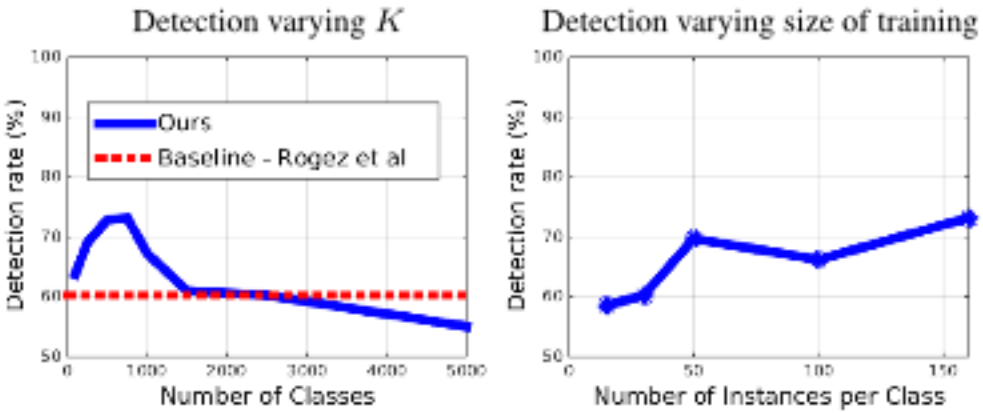
Numerical evaluation



Feature Evaluation: We compare our feature encoding to different variants (for K = 750 classes) in (a). **Our feature outperforms all other baselines**.

We also vary the resolution of our feature in (b), again for K = 750. A size of **32×24×35 is a good trade-off** between size and performance. Doubling the resolution in u, v marginally improves accuracy.

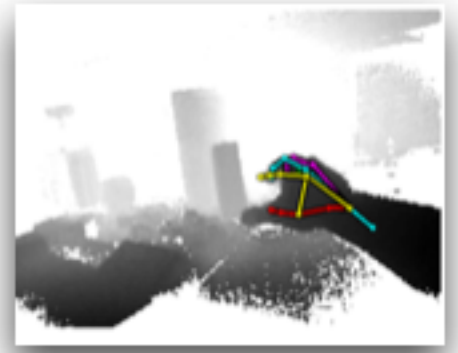
Clustering and size of training set:
Both results suggest that our system may perform **better with more training data and more quantized poses**.



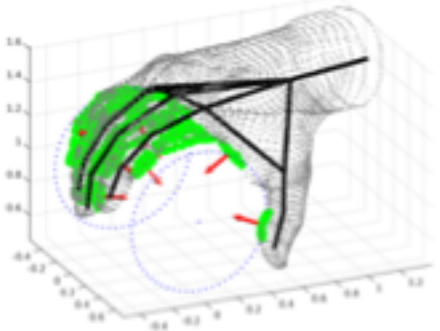
Outline

Part I: Data synthesis

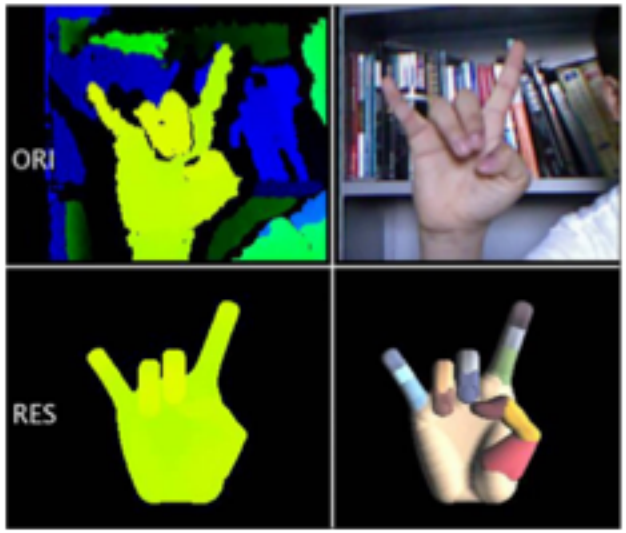
Part II: Hand pose estimation

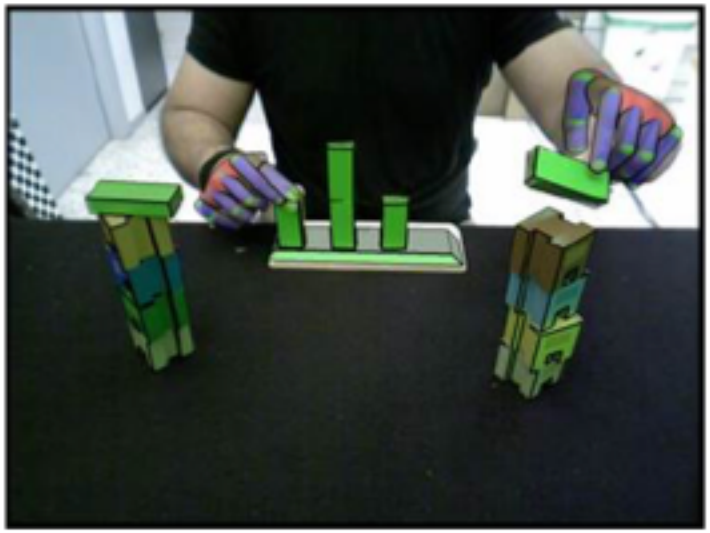


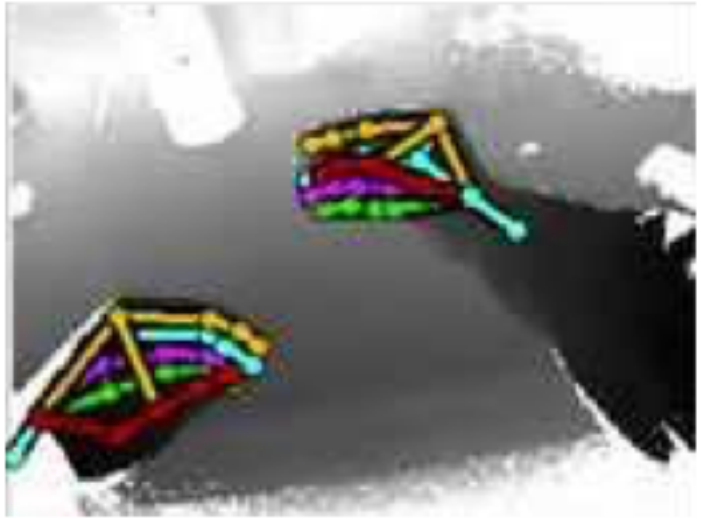
Part III: Functional understanding



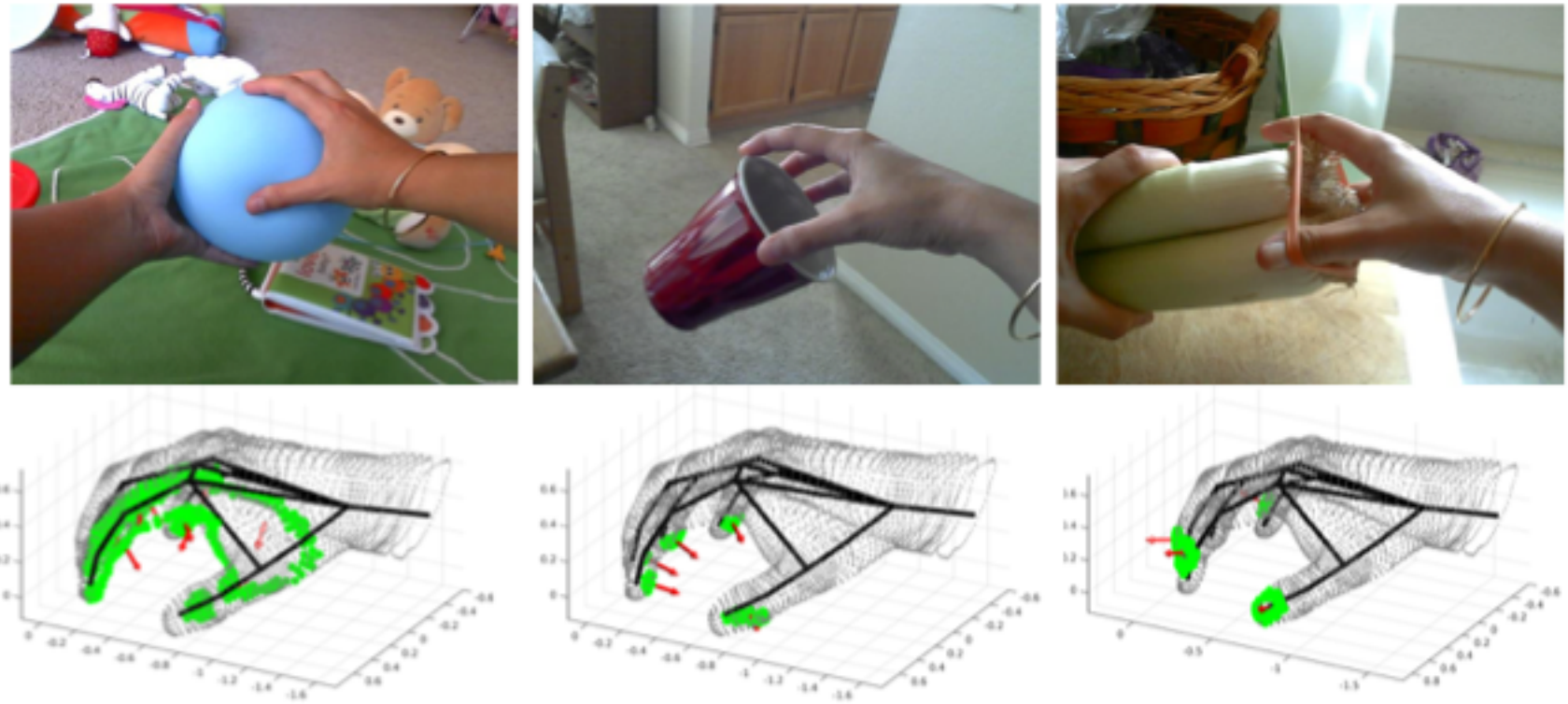
So far, we have focused (like past work) on **kinematic pose** estimation…







But the **same kinematic pose** can be used for dramatically **different functional manipulations**



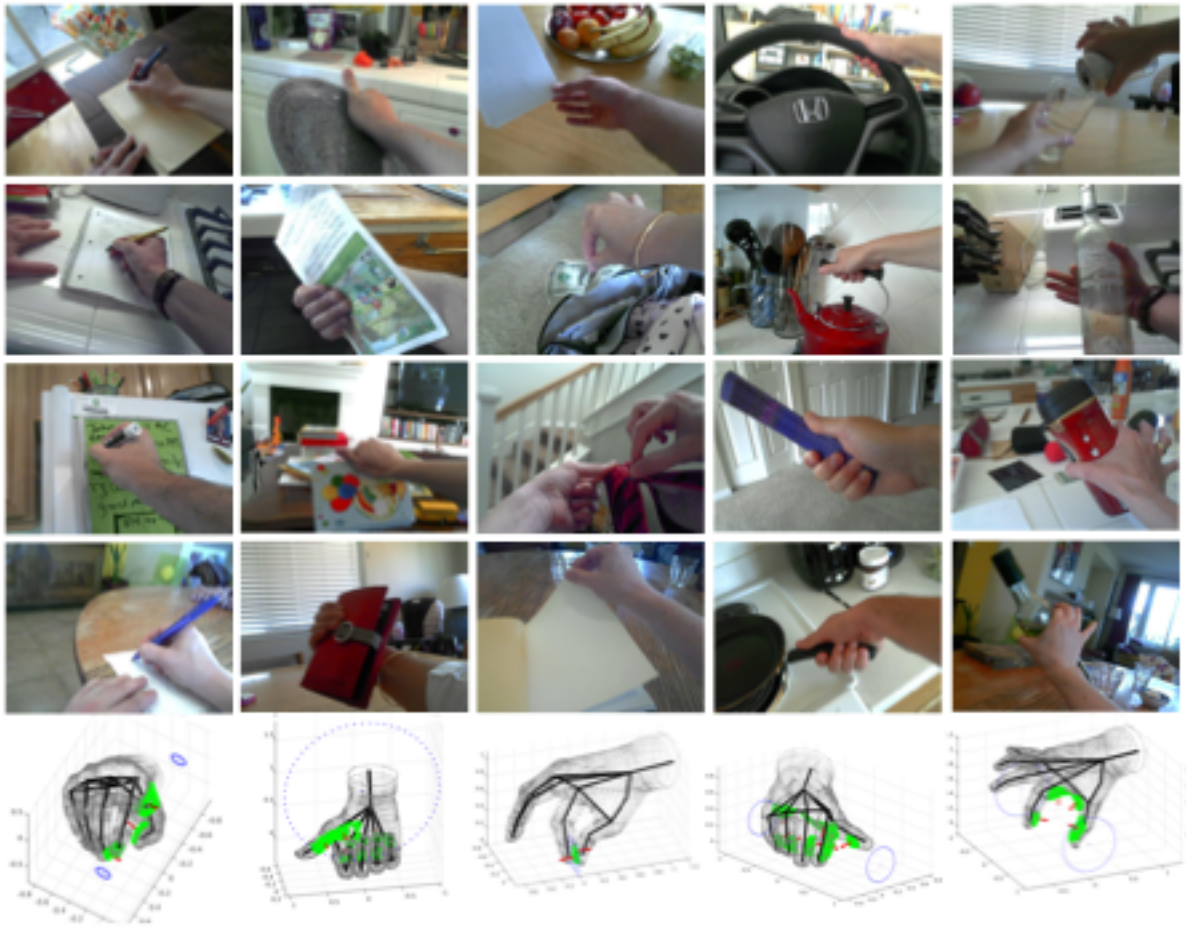
where differences are manifested in terms of distinct **contact points and force vectors!**

IDEA: We address **pose+contact+force** prediction as a **discrete fine-grained classification** task based on a **taxonomy of hand-object interactions** developed from the robotics community

We use the 71 grasps from the latest taxonomy **[Liu et al, HUMANOID 2014]**

Grasp UNderstanding (GUN-71) dataset

- Our classification engine is **data-driven**: we put forth considerable effort toward assembling a **large collection of diverse images** that span the taxonomy of classes.



We tailored our new dataset to "fill the gap" in terms of overall **scale, diversity, annotation detail**, and combination of **RGB and depth** data:

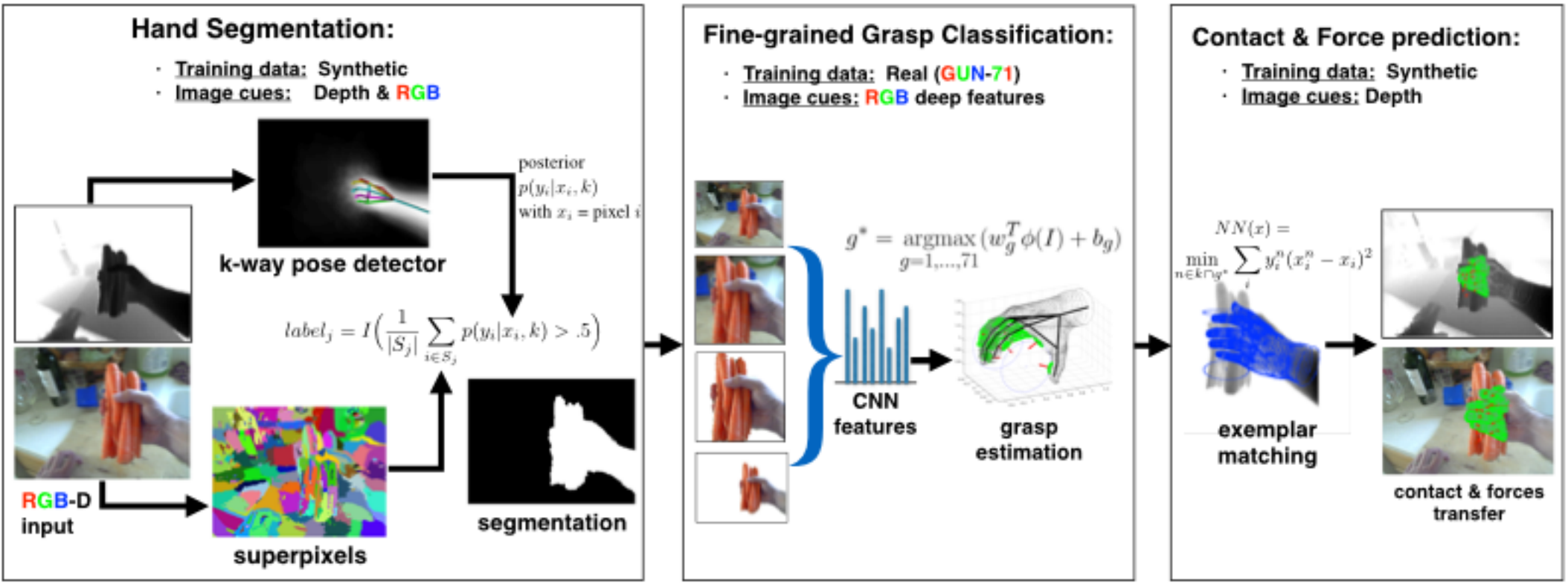
- 12,000 1st-person RGB-D images
- 71 grasps/ 28 objects per grasp
- 8 different subjects
- 5-6 views for each configuration
- Real object manipulation
- 5 different home environments

For each grasp, a **3D hand model** is used to generate synthetic data. We compute contact points and force directions by intersecting the triangulated hand and object 3D meshes.

[Rogez, Supancic, Ramanan, ICCV 2015]

Recognition pipeline

- We develop a pipeline for fine-grained grasp classification exploiting **depth and RGB data**, training on (deep) features extracted from both **real and synthetic training data.**



- Our simple **post-processing exemplar framework** predicts contacts and forces associated with hand manipulations by simply transferring them from the selected grasp model to the exemplar location in the 3D space.

[Rogez, Supancic, Ramanan, ICCV 2015]

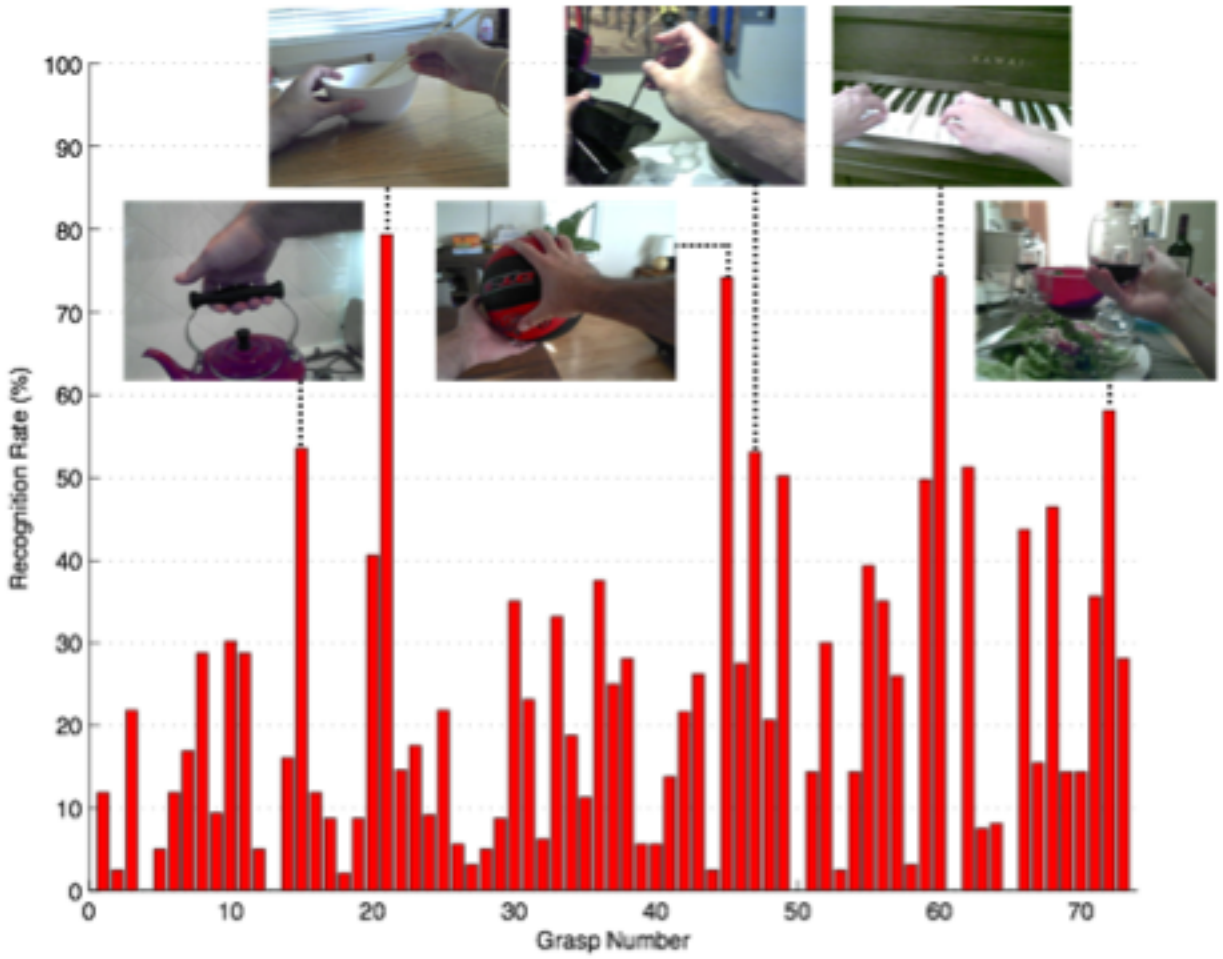
Numerical results

- We present extensive experimental results with state-of-the-art baselines.
- Our final system (making use of depth+rgb+real+synthetic data) produces a **near 2X improvement over prior work and a naive CNN baseline**.
- Overall, **fine-grained grasp classification accuracy approaches 20%** for 71 classes. This is a challenging and important fine-grained recognition problem!

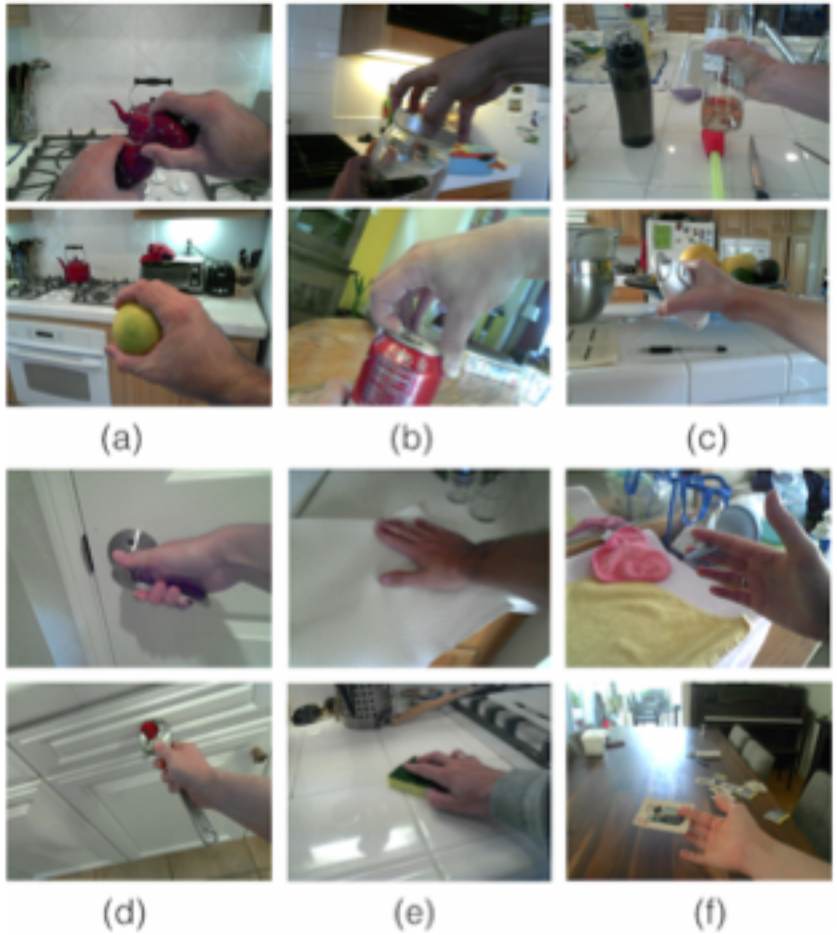
Features	Acc.	top 20	top 10	min	max
HOG-RGB	3.30	7.20	9.59	0.00	28.54
HOG-Depth	6.55	12.96	15.74	0.66	26.18
HOG-RGBD	6.54	13.76	19.24	0.00	45.62
Deep-RGB [3]	11.31	25.92	35.28	0.69	61.39
Deep-RGB(segm.)	11.10	21.56	26.51	0.69	29.46
HOG-RGB (cropped)	5.84	11.22	14.03	0.00	27.85
Deep-RGB (cropped)	13.67	27.32	36.95	1.22	55.35
HOG-RGB (crop.+segm.) [4]	7.69	15.23	18.65	0.69	30.77
HOG-Depth (crop.+segm.)	10.68	22.04	27.99	0.52	42.40
Deep-RGB (crop.+segm.)	12.55	22.89	27.85	0.69	37.49
Deep-RGB (All)	**17.97**	**36.20**	**44.97**	**2.71**	**68.48**

Qualitative analysis

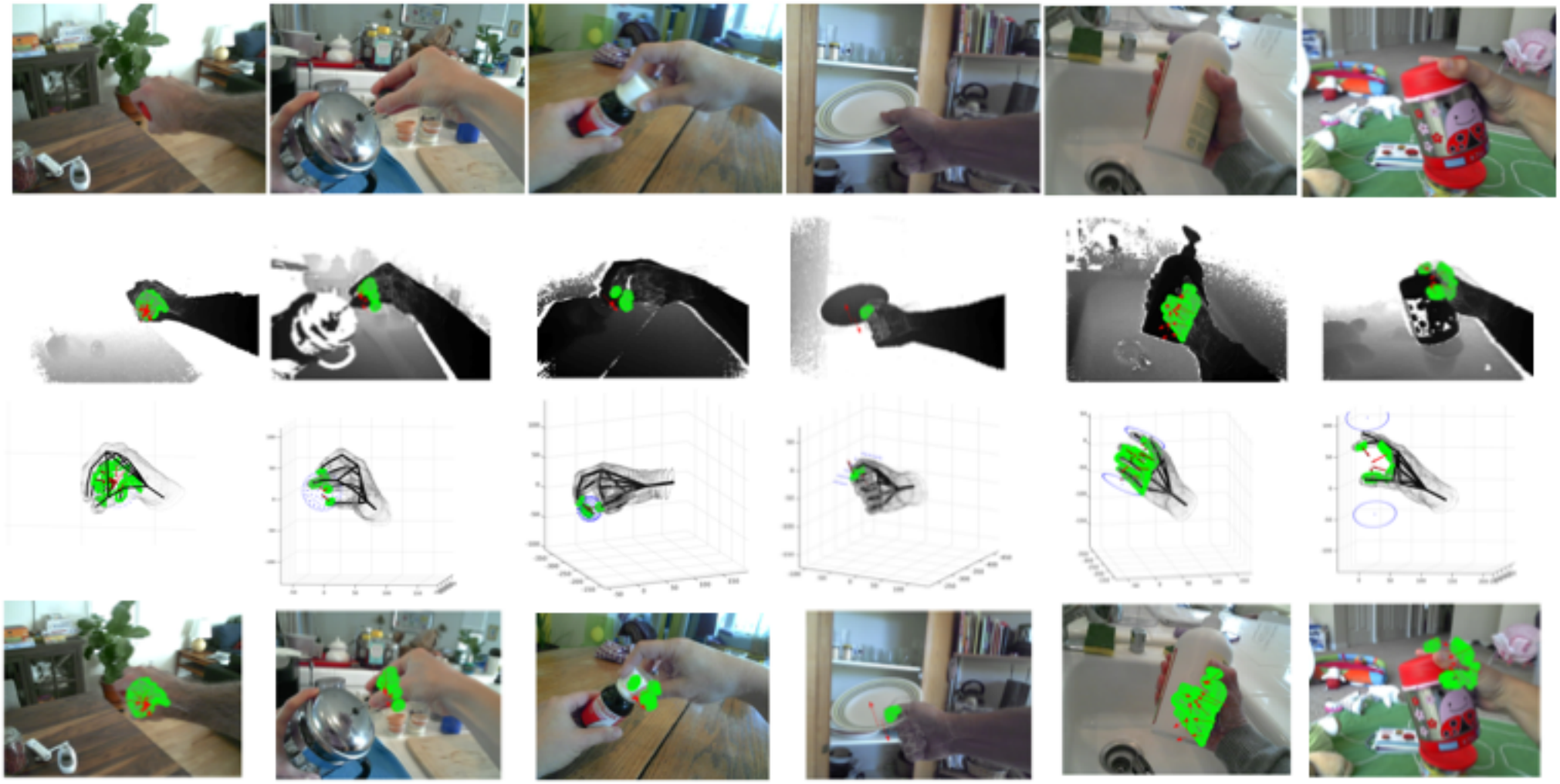
Easy cases



Common confusions



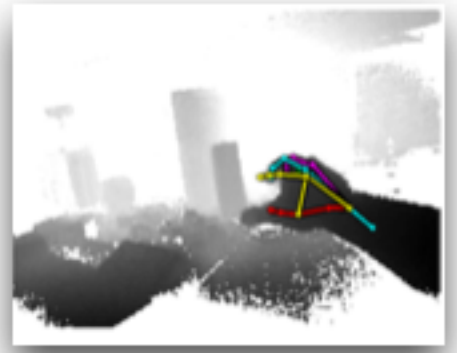
Qualitative results of contact and force prediction:



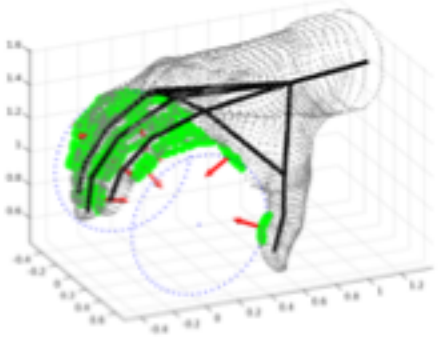
Take-Home message

- Data synthesis: we control what if being synthesized. **Make it as realistic as possible!**

- **Egocentric settings have specific properties**. We show how to take advantage of them for fast pose recognition.



- Fine-grained grasp recognition is still an open problem. A combination of **RGB + Depth data, real + synthetic data** reaches the best performance.



Thanks for your attention…

Questions?