3D object modelling, recognition and tracking with an application to action learning VVV18

Michael Zillich

Vision for Robotics Group
Institute of Automation and Control
TU Wien

Is this Vision?

"Now! *That* should clear up a few things around here!"

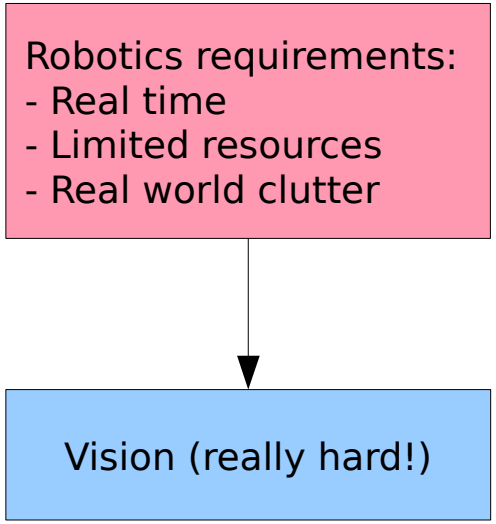
Is this Vision?

"Now! *That* should clear up a few things around here!"

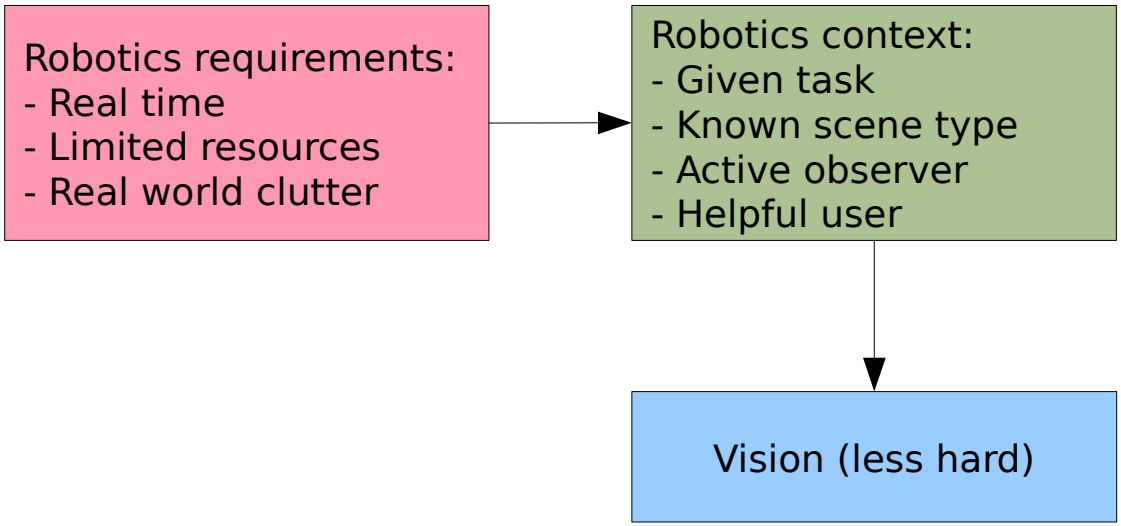
What is vision for *robotics?*

Vision (hard!)

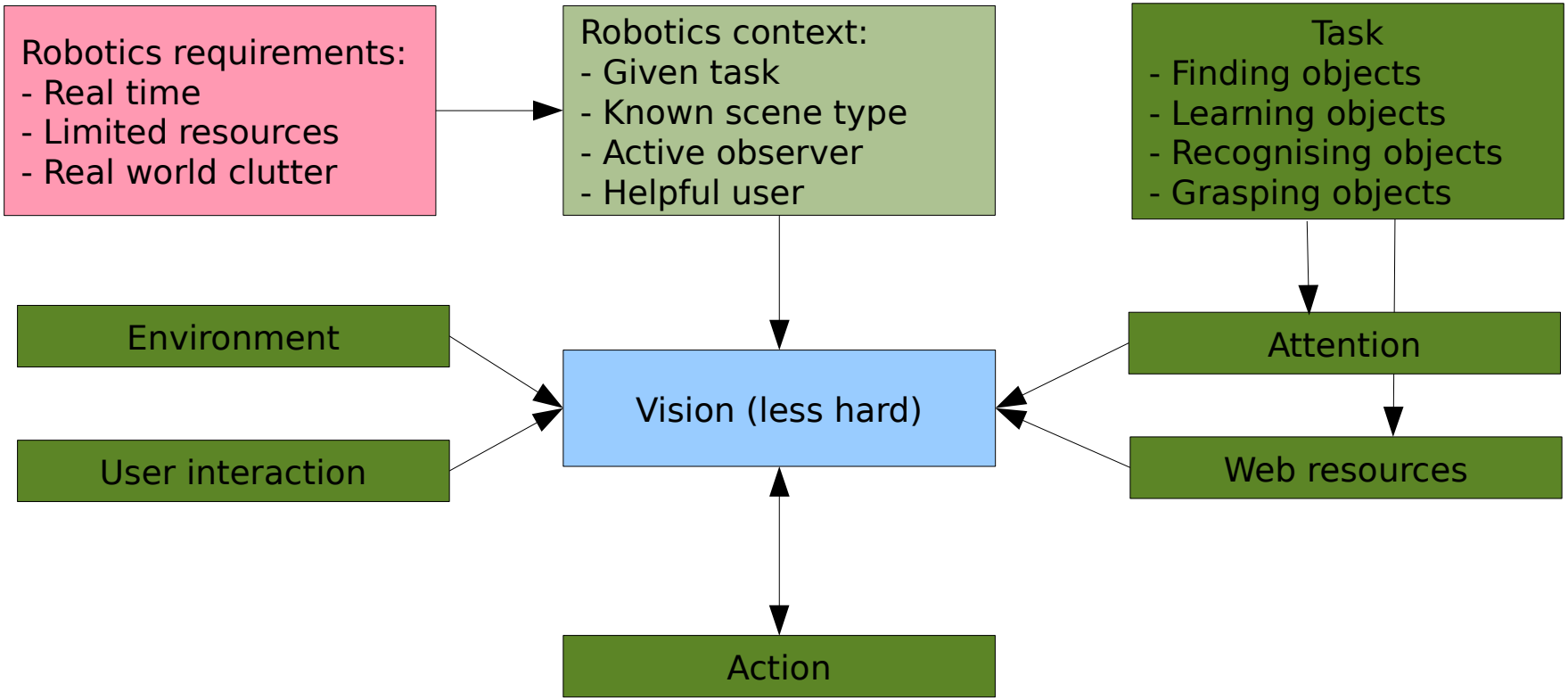
What is vision for *robotics?*



What is vision for *robotics?*



What is vision for *robotics?*



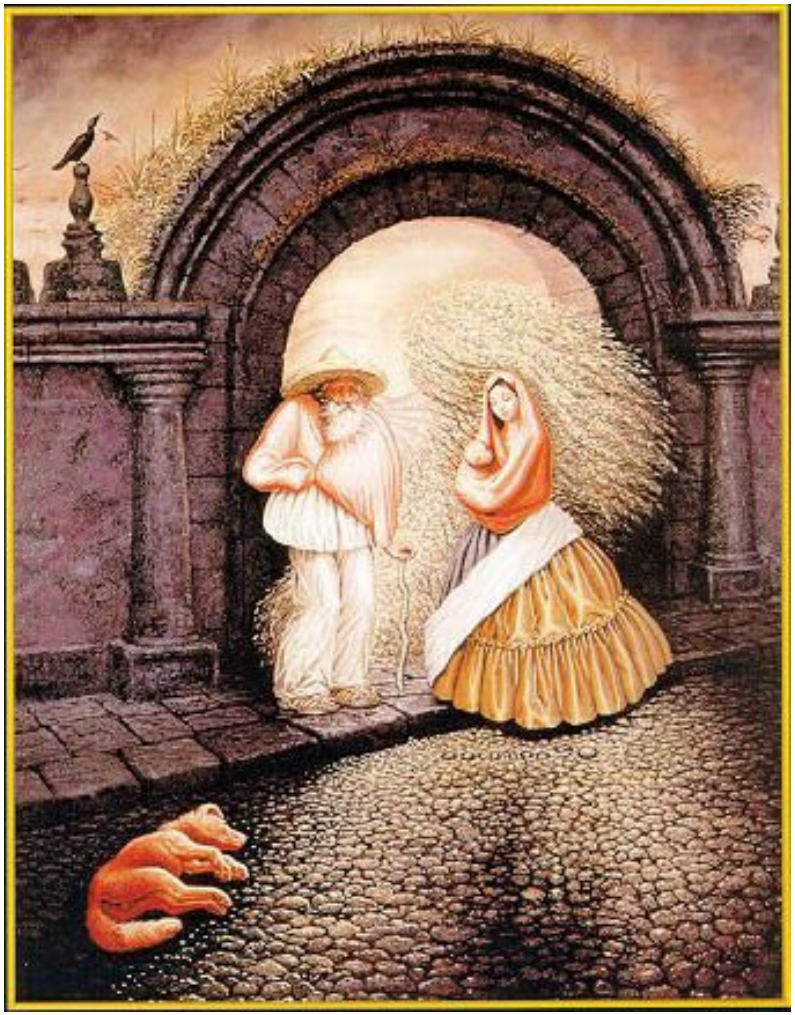
Object X

- Object **detection**, figure-ground **segmentation**, perceptual grouping = find relevant entities (to task)
- Object instance **recognition** = recognising one known object
- Object **categorisation/classification** = recognising objects belonging to a category (bottle, animal)
- Object **tracking** = recognise in image sequence while propagating state

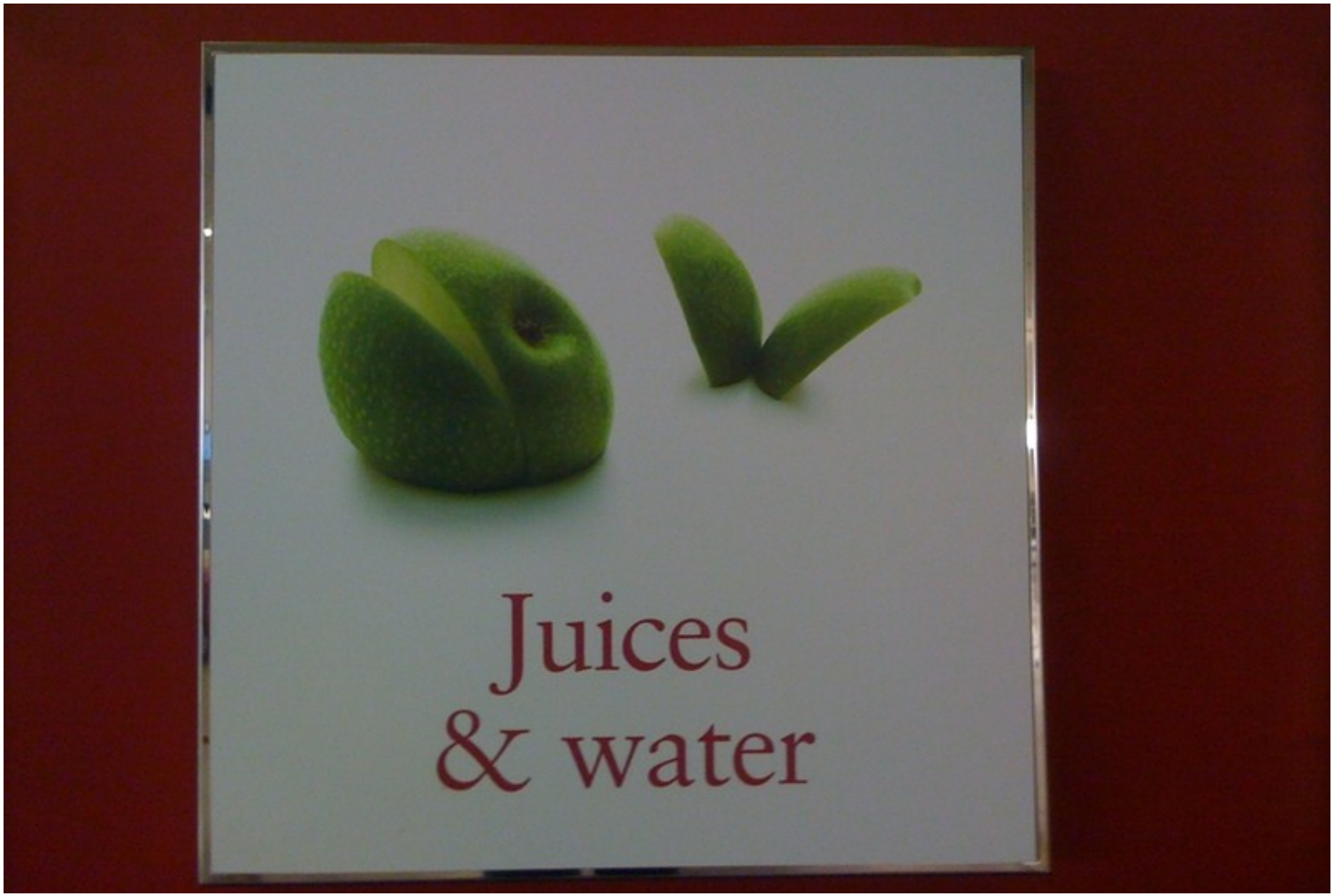
Overview

- **Detection / segmentation**
- Modelling
- Recognition
- Classification
- Tracking
- Attention
- Application to action learning

What is the object?

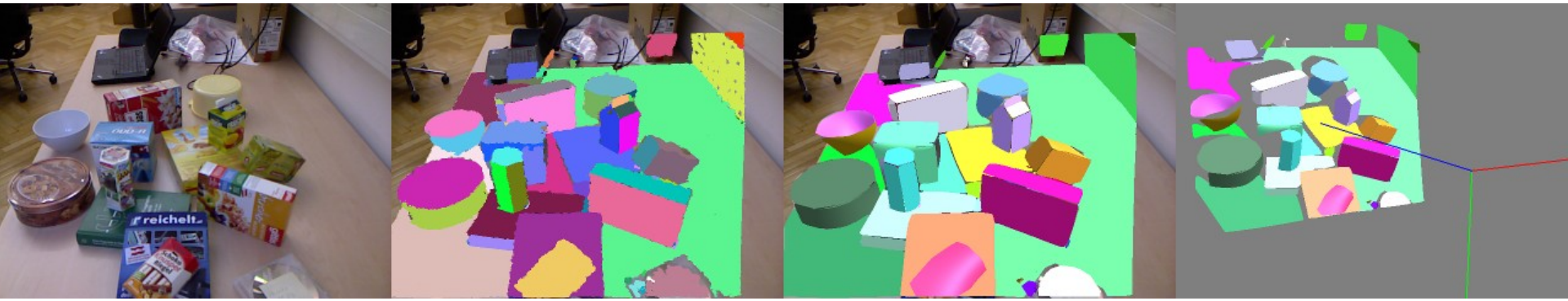


What is the object?



Object Segmentation

- Identify, in a **general** way, which bits of the scene could be **task relevant** objects
- Amidst **distractors, occlusions**
- [Ückermann ea IROS 2012]
- [Mishra ea ICRA 2012]
- [Katz ea RSS 2013]
- [Hager ea IJRR 2011]

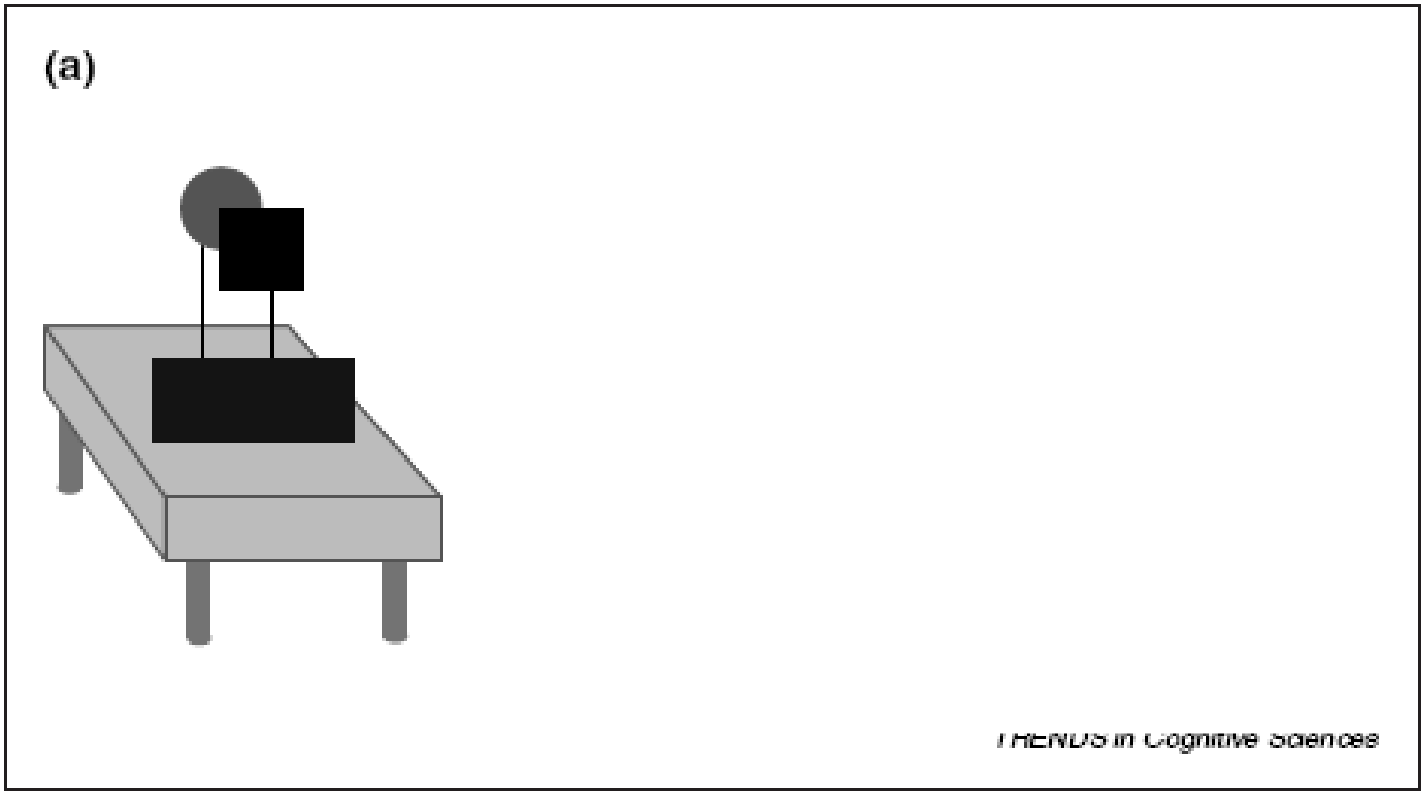


From coloured point clouds ... ... to separated object hypotheses

[Richtsfeld ea JVCI'14]

Generic view principle

"Qualitative (e.g. topological) image structure is stable with respect to small changes of viewpoint."



[M. K. Albert: Surface perception and the Generic View Principle, 2001.]

Object Segmentation

Gestalt principles

- Proximity
- Similarity
- Continuity
- Closure
- Symmetry
- Common region
- Element connectedness
- Common fate
- Good Gestalt

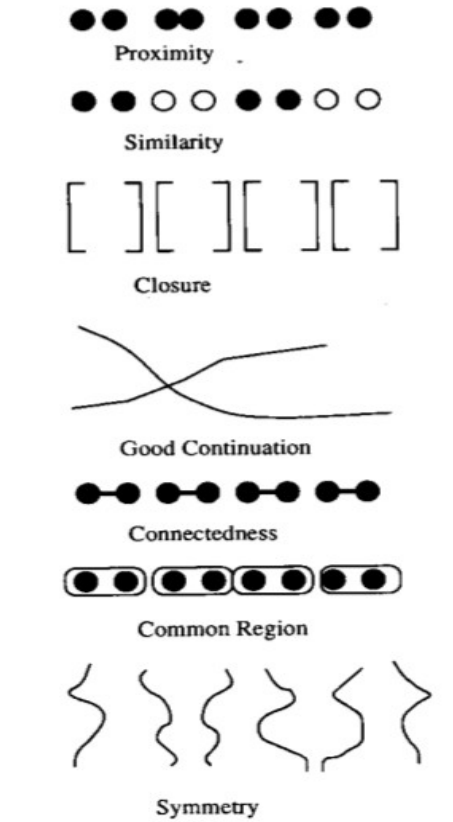
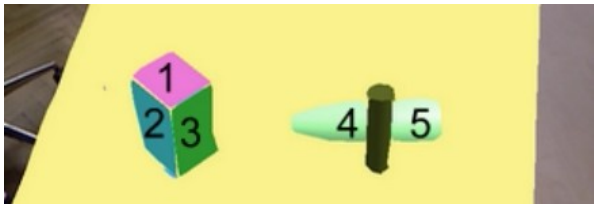


Fig. 3. Gestalt laws of grouping.

Object Segmentation: Grouping



Relations btw. neighboring surfaces

- r_co ... similarity of patch colour
- r_rs ... relative patch size similarity
- r_tr ... similarity of patch texture quantity
- r_ga ... gabor filter match
- r_fo ... fourier filter match
- r_co3 ... color similarity on 3D patch borders
- r_cu3 ... mean curvature on 3D patch borders
- r_cv3 ... curvature variance on 3D patch borders
- r_di2 ... mean depth on 2D patch borders
- r_vd2 ... depth variance on 2D patch borders

Relations btw. non-neighboring surfaces

- r_co ... similarity of patch colour
- r_rs ... relative patch size similarity
- r_tr ... similarity of patch texture quantity
- r_ga ... gabor filter match
- r_fo ... fourier filter match
- r_md ... minimum distance between patches
- r_nm ... angle between mean surface normals
- r_nv ... difference of variance of surface normals
- r_ac ... mean angle of normals of nearest contour p.
- r_dn ... mean distance in normal direction of nearest contour p.

Object Segmentation: Grouping

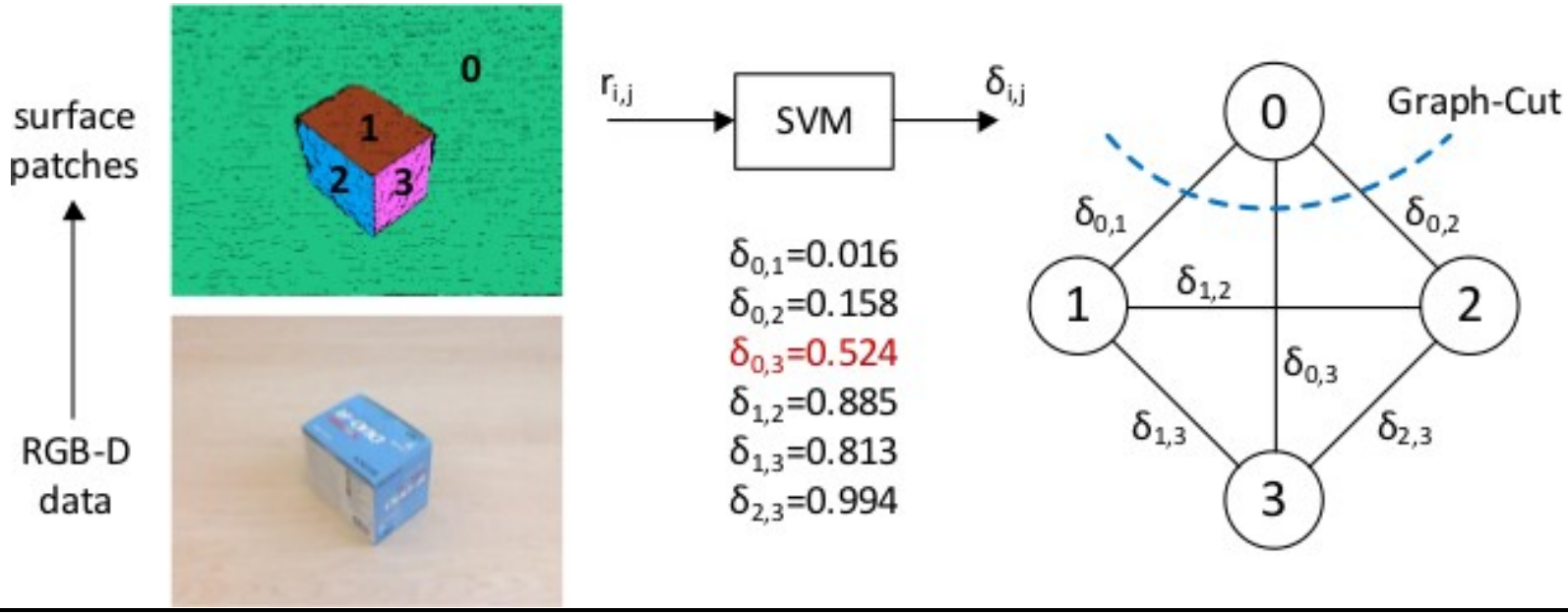
Global decision using graph cut

- Train Support Vector Machines (SVMs) on feature vectors, using annotated training data

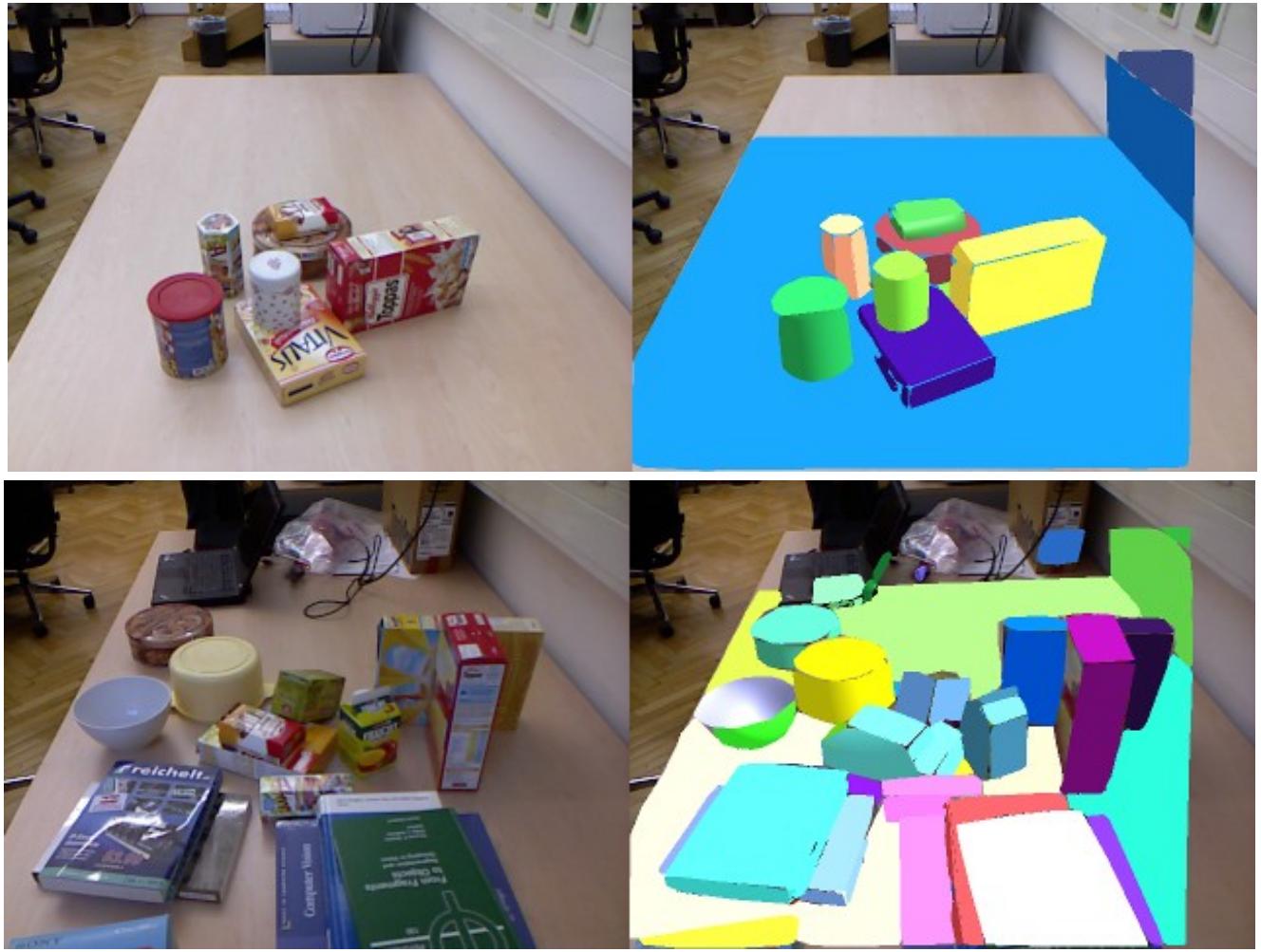
  r_st = (r_co, r_rs, r_tr, r_ga, r_fo, r_co3, r_cu3, r_cv3, r_di2, r_vd2)

  r_as = (r_co, r_rs, r_tr, r_ga, r_fo, r_md, r_nm, r_nv, r_ac, r_dn)

- Use predicted probability of “same object” as pairwise terms for graph cut



Object Segmentation



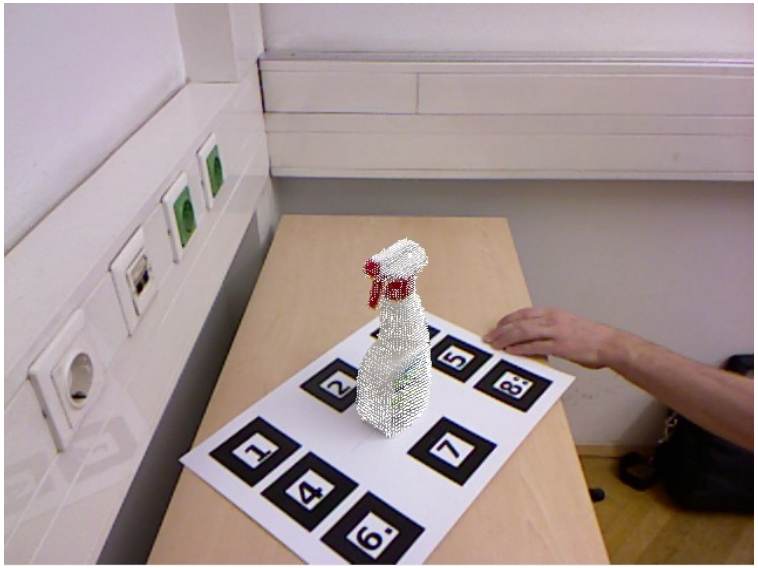
Object Segmentation Database (OSD) [Richtsfeld ea IROS'12]

Overview

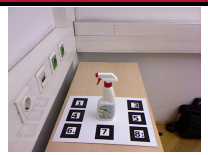
- Detection / segmentation
- **Modelling**
- Recognition
- Classification
- Tracking
- Attention
- Application to action learning

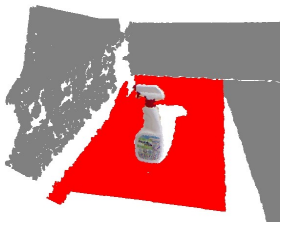
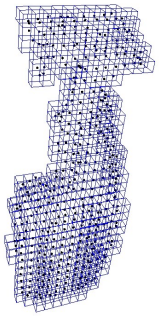
Object modelling

- Learn **individual** object models
- One shot to a **few views**
- Build database of known objects



Object modelling



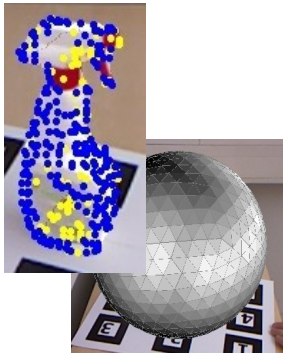
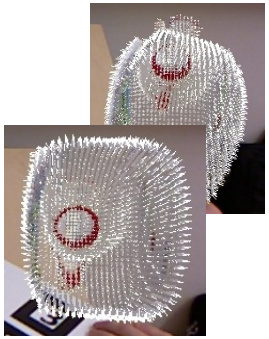



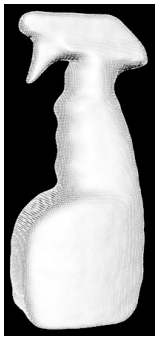

Input
- image
- point cloud

Segmentation
- Ground plane detection
- Euclidean clustering

Pose estimation
- Guess (SIFT)
- Scan alignment (ICP)

Voxel grid update
- Point weights
- Surface normals



Create recognition model
- Key-frame selection
- SIFT (yellow) [Lowe 2004]
- SHOT (blue) [Tombari 2010]

Loop closing
- Document indexing [Sivic 2003]
- Error distribution [Sprickerhof 2009]

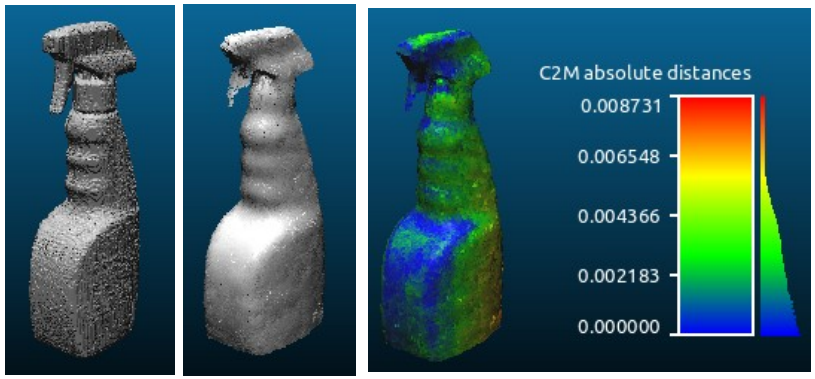
Point cloud
- Adaptive threshold

Surface modelling
- Poisson triangulation

Object Modelling

- RTM Toolbox (Recognition, Tracking, Modelling)

Object modelling

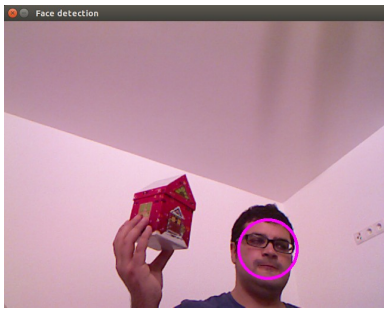


Reconstructed (left) vs. laser scanned (right): 2.16 ± 1.53mm [KIT object database]

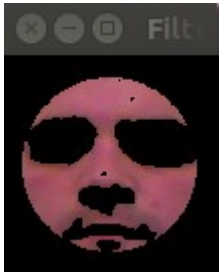
[Prankl ea IROS 2015]

In-hand Modelling

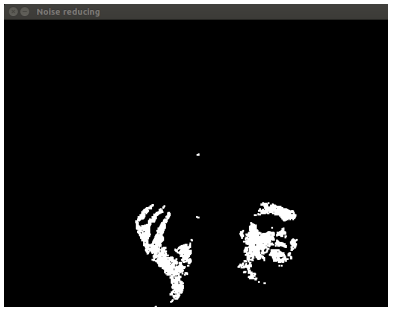
- User-assisted modelling while holding in hand [Streicher 2018]



Detect face



Get skin color



Mask skin regions

Subtract and find nearest cluster

Grab Cut for precise contours

RTM pipeline: 3D object model

Overview

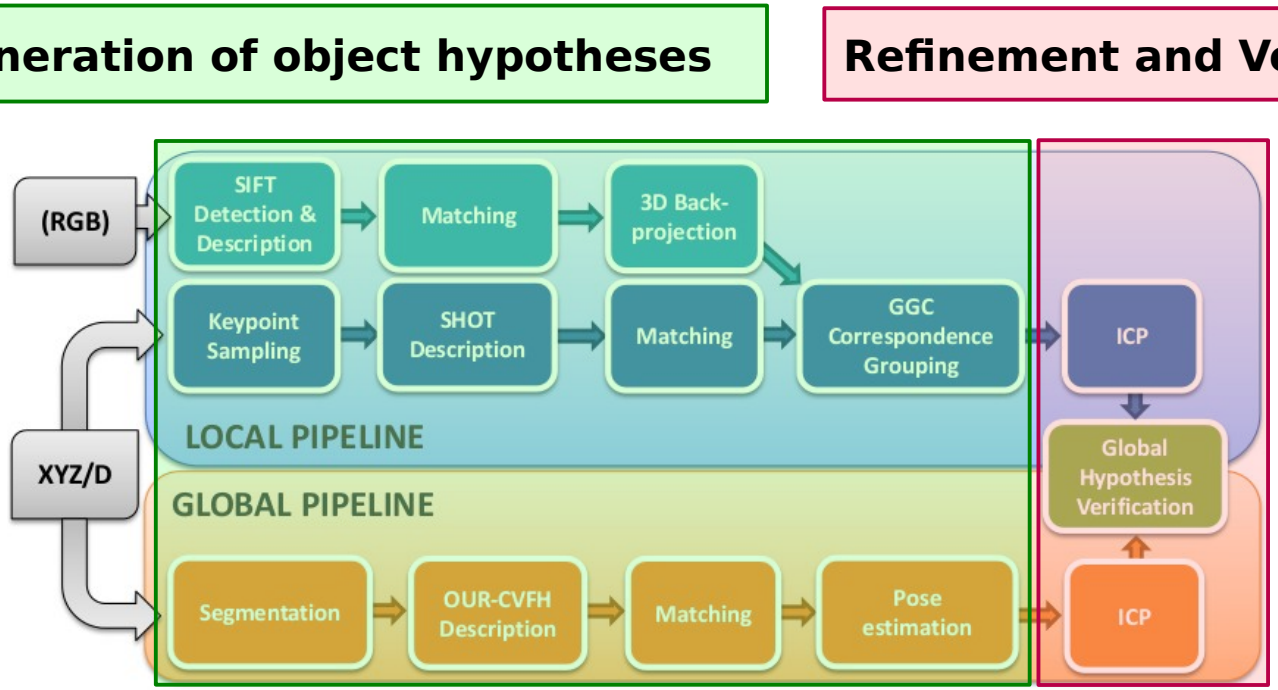
- Detection / segmentation
- Modelling
- **Recognition**
- Classification
- Tracking
- Attention
- Application to action learning

Object Recognition

- Robust recognition of **object instances** in uncontrolled environments: Partial occlusions, clutter, degenerate views, illumination conditions
- **Diverse object properties**: Textured or texture-less, distinctive or uniform shape
- => object **ID** and **6D pose**

Object Recognition

- Framework for multi-cue recognition



[Aldoma ea PAMI'16]

Features - 2D

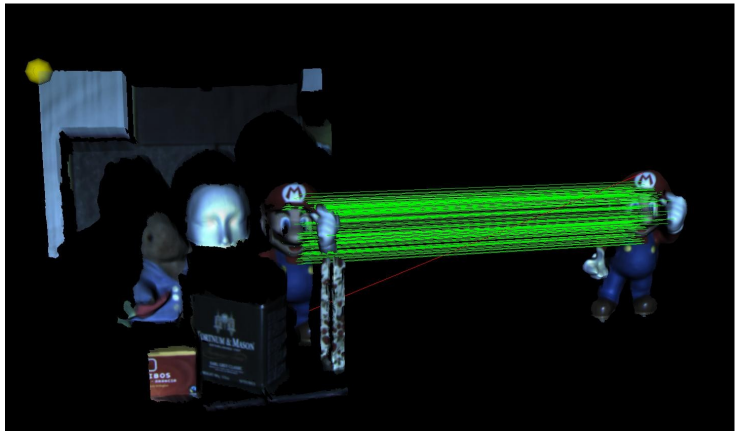
Classic feature based 2D recognition

- Find interest points in both images (e.g. DoG)
- Match corresponding point pairs using descriptors (e.g. SIFT)
- Align, RANSAC for obtaining 6D pose

Features - 3D

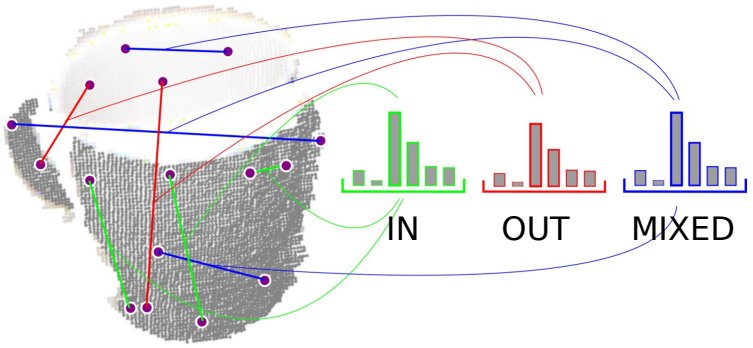
Local descriptors

- SHOT: Unique Signatures of Histograms for Local Surface Description (= 3D SIFT)



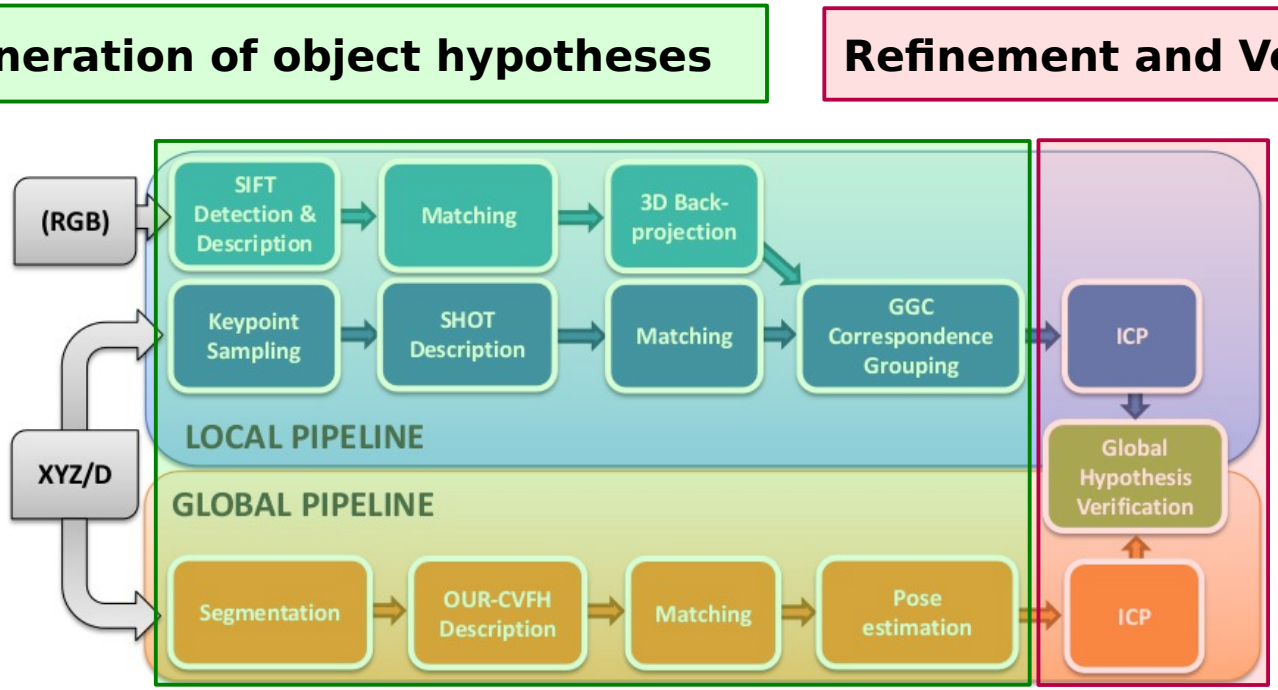
Global descriptors

- Ensemble of Shape Functions (ESF) [Wohlkinger 2011]
  Based on shape distributions [Osada ea 2001], inside/outside/mixed
  Additional histograms for ratio, area and angle



Object Recognition

- Framework for multi-cue recognition



[Aldoma ea PAMI'16]

Multi-hypothesis Recognition

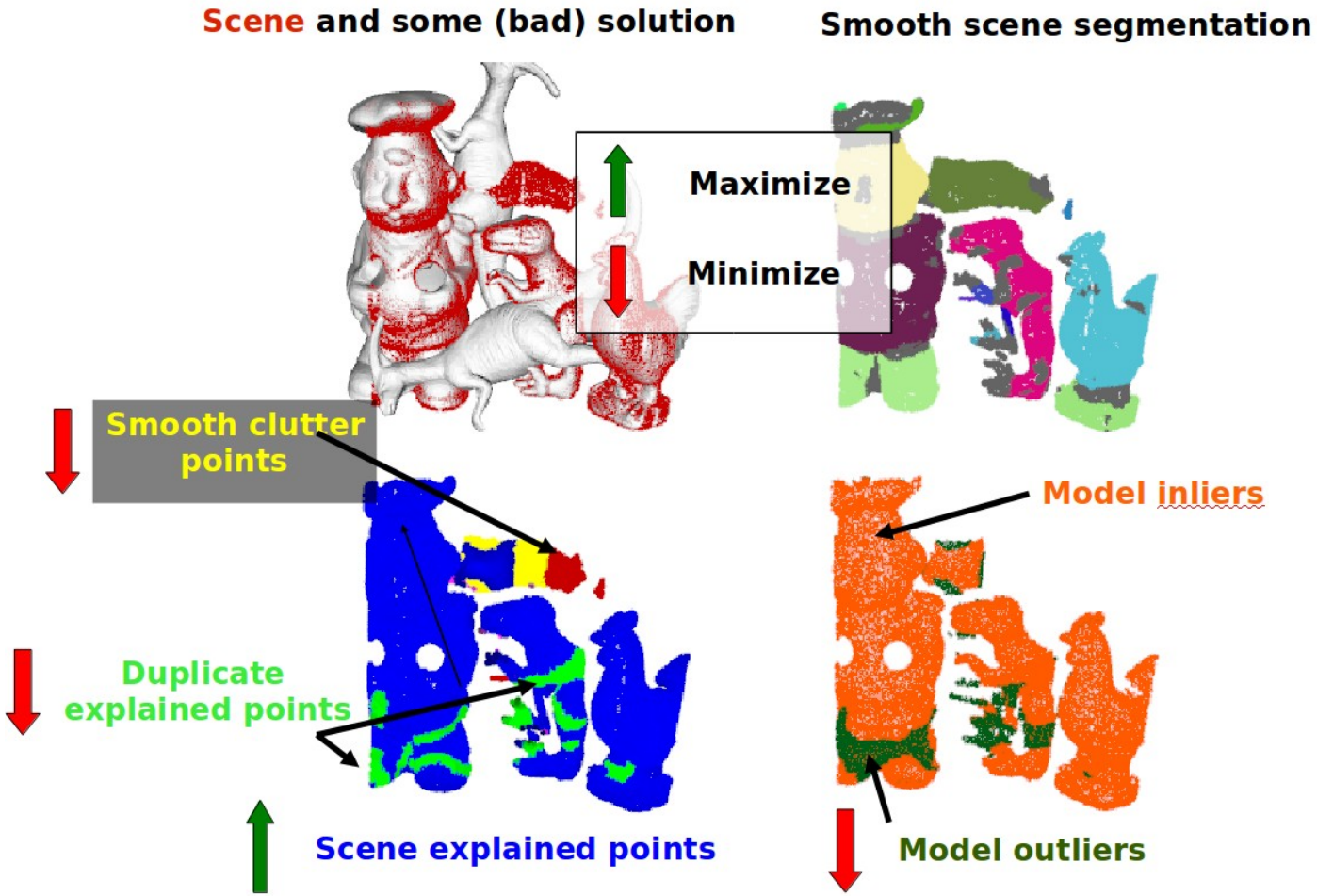
Common hypothesis verification paradigms

- Non-maxima supression to handle duplicate hypotheses, but what if two instance of same object in scene?
- Acceptance thresholds tricky, highly occluded objects vs. false positives
- Disregards interaction among hypotheses

=> Consider **all hypotheses simultaneously in a global model** of the scene in terms of objects

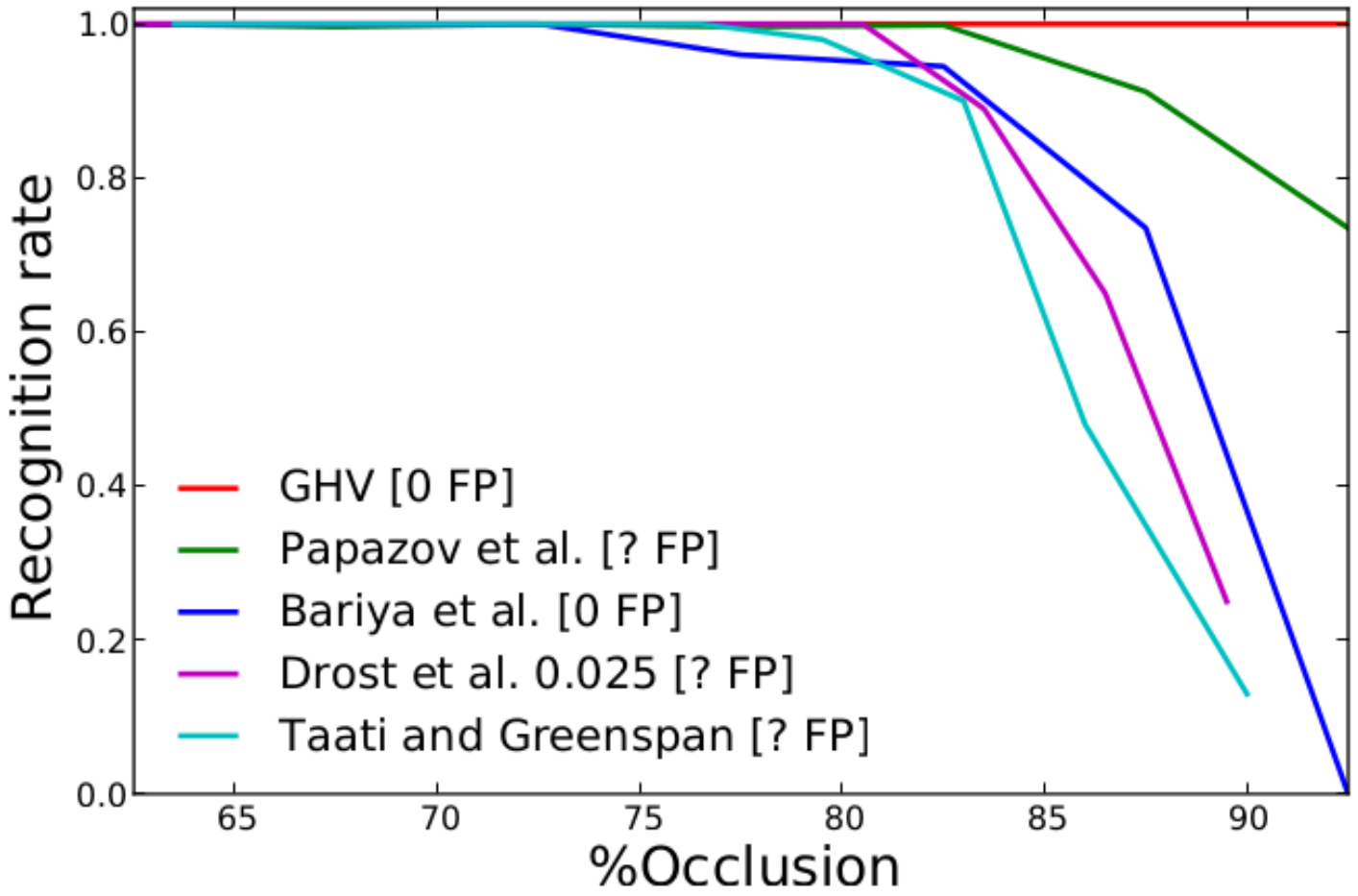
Global Hypothesis Verification

- Binary optimisation with tabu search



Result: high robustness

Laser Scanner dataset



Overview

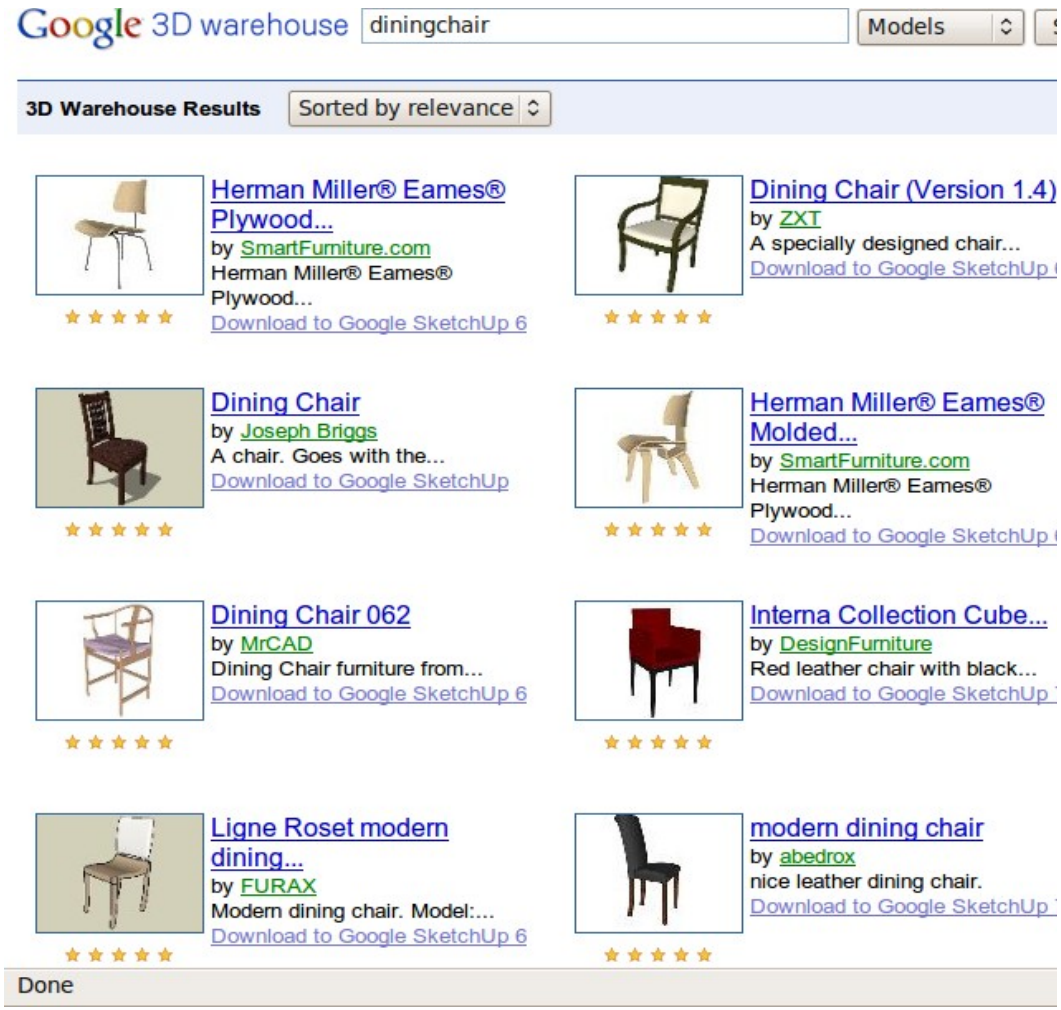
- Detection / segmentation
- Modelling
- Recognition
- **Classification**
- Tracking
- Attention
- Application to action learning

Object Categorisation

- Many objects sharing common characteristics
- Large amounts of **training data, training time**
- **Scalability** with number of classes

Offline Training from Web Models

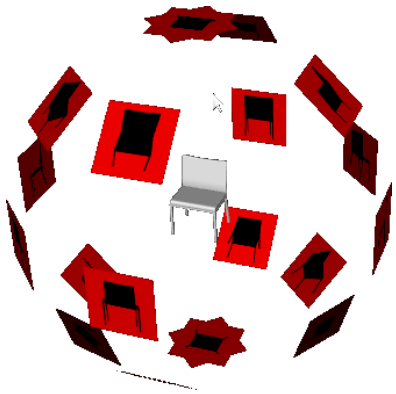
- E.g. “dining chair”
- Get many 3D CAD models, e.g. google 3D warehouse
- Find similar models from synonyms, e.g. Wordnet (mug, cup; chair, stool; etc.)

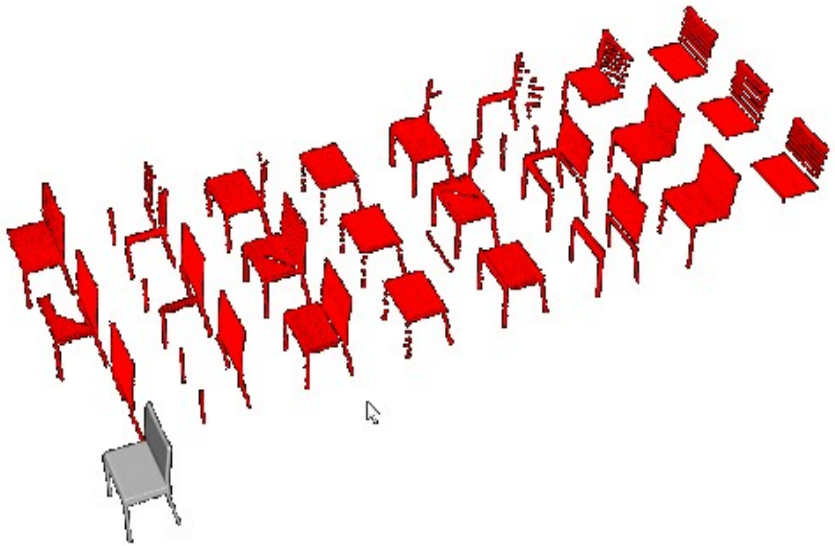


Offline Training from Web Models

Generate training views

- Objects are “perfect” 3D CAD data
- Actual data is 2.5D (RGB-D)
- Create views on object to simulate sensor view, incl. noise
- Dozens of views, for 100s of models

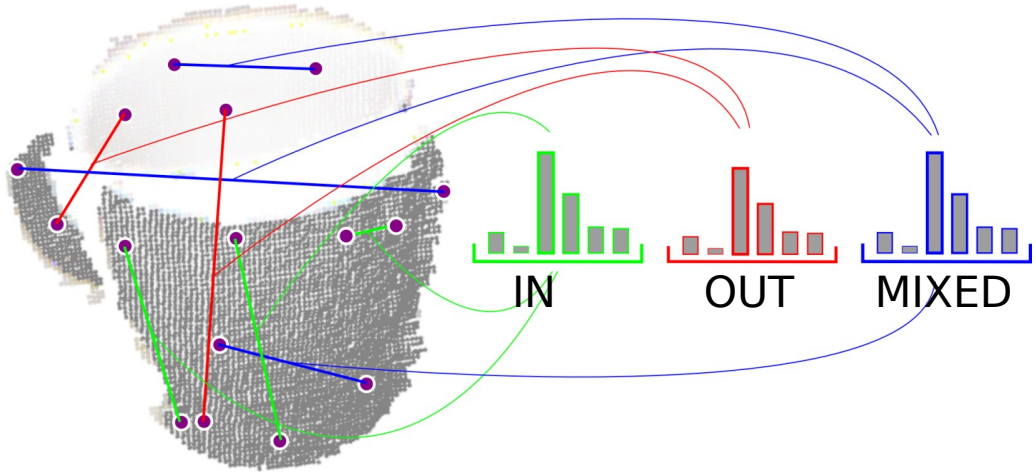


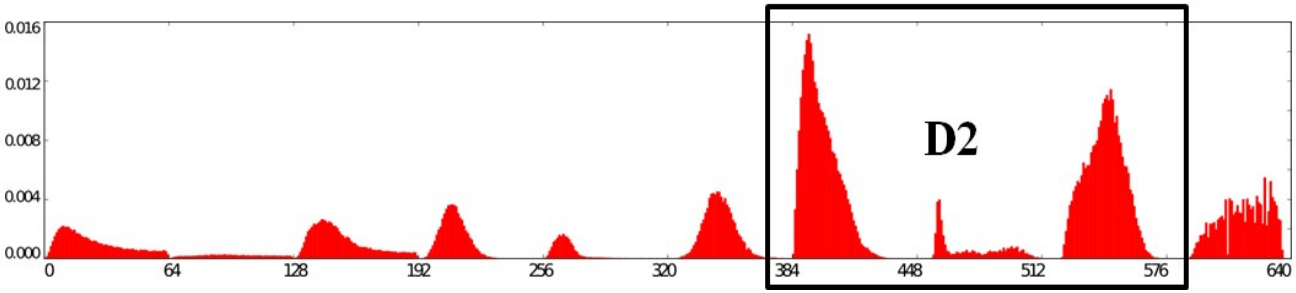


Offline Training

Robust Feature vector

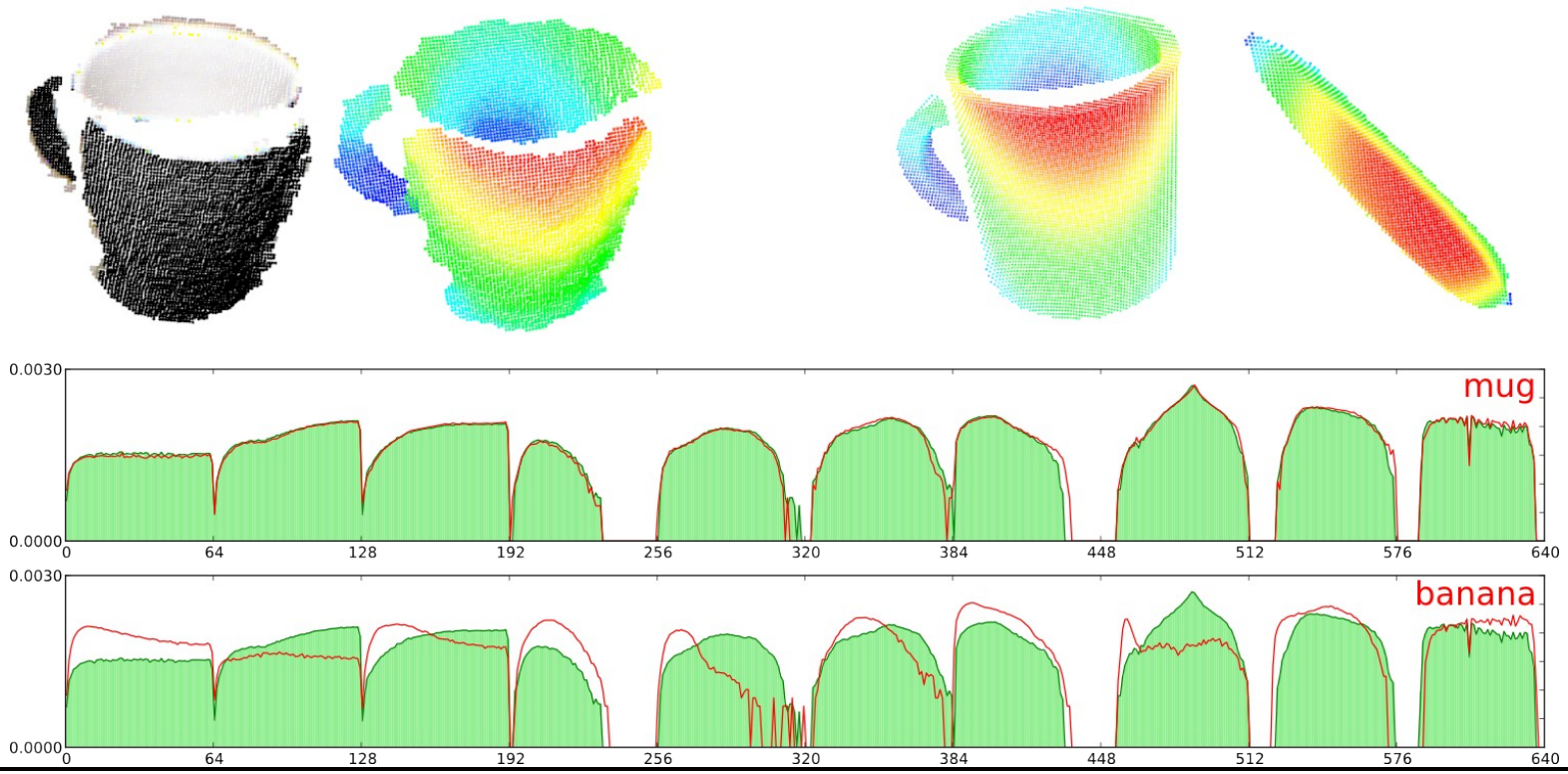
- Ensemble of shape functions (ESF)
- Based on shape distributions [Osada ea 2001]
  inside, outside, mixed
- Additional histograms for pint triplets: ratio, area, angle
- 640 dim. vector





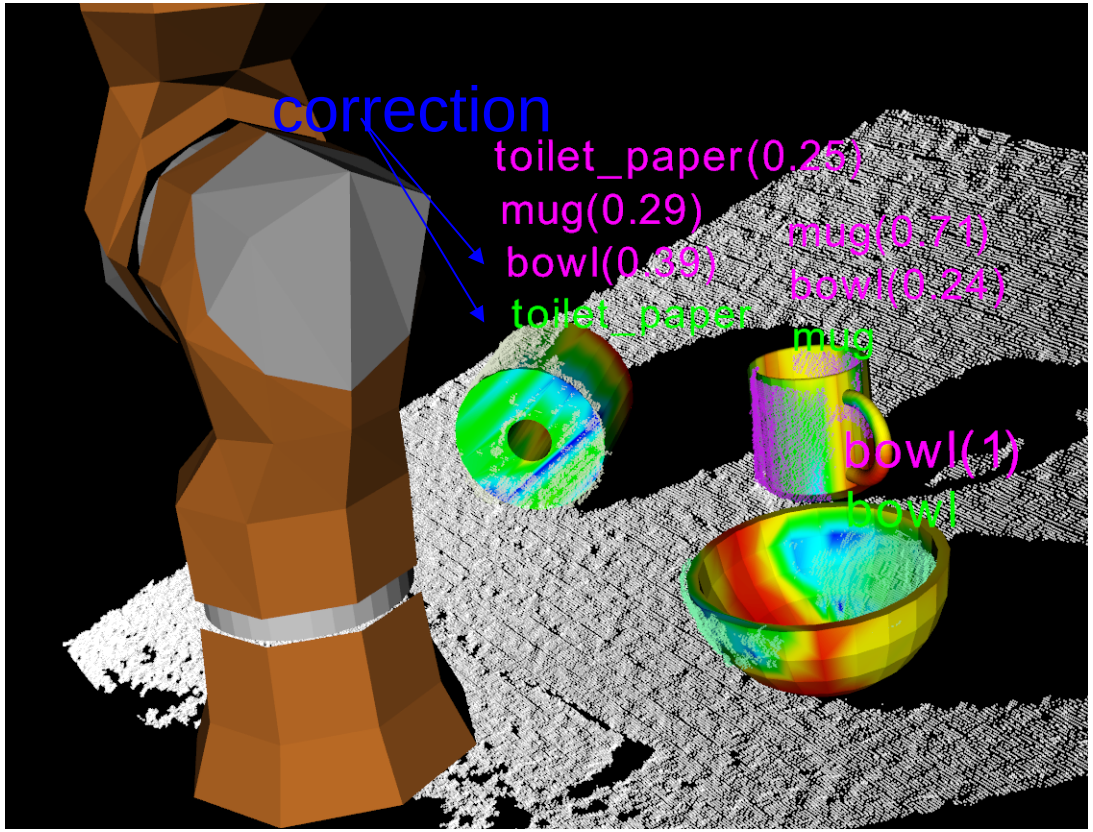
Online Matching: kNN classifier

- Find nearest neighbour in feature space
- Efficient indexing techniques to cope with large database (100,000s views)
- Majority vote from k nearest neighbours



Verification with Pose Fit

- Best view i of model j
- Fit 3D model j to point cloud
- Verify classification, precise pose



Initial classification hypotheses and verified after pose fit

Results: 200 classes

[Wohlkinger ea IROS'11]

Results: 200 classes

NEAREST NEIGHBOR CLASSIFICATION AND MOST CONFUSING CLASS

class name	1-NN	10-NN	confusing class
per scenes OVERALL	58.22 %	78.23 %	
per class OVERALL	49.10 %	71.39 %	
apple	81.40 %	98.45 %	pumpkin
banana	54.79 %	69.86 %	pistol
bottle	48.77 %	79.01 %	suv
bowl	50.00 %	76.47 %	hat
car	11.52 %	43.64 %	suv
donut	20.00 %	62.00 %	cap
hammer	83.41 %	96.10 %	axe
mug	91.96 %	99.46 %	watch
tetra pak	47.09 %	72.09 %	mug
toilet paper	2.11 %	16.84 %	armchair

Results on 3d-net Cat200 database using ESF

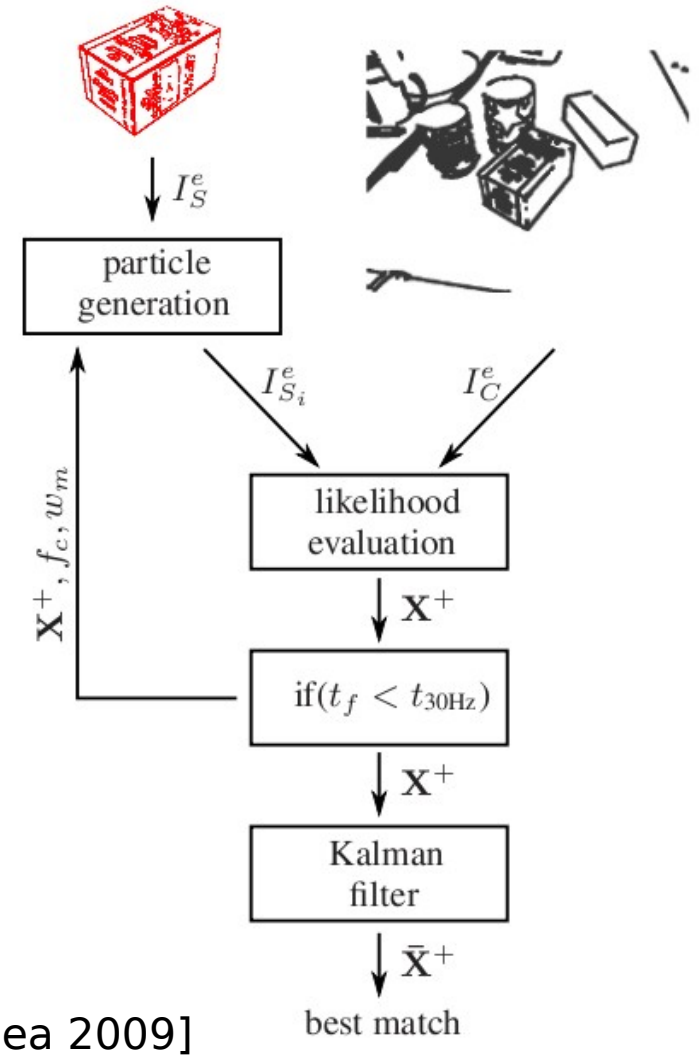
[Wohlkinger ea ICRA'12]

Overview

- Detection / segmentation
- Modelling
- Recognition
- Classification
- **Tracking**
- Attention
- Application to action learning

Model Based Object Tracking

- Given: 3D model, estimated 6D pose
- Represent pose estimate (PDF) with a number of hypotheses (particles)
- Propagate pose into next image
- Verify each particle (e.g. matching projected object edges to image edges), GPU implementation
- Weak particles are discarded, good ones are cloned (plus noise)
- Repeat ..



[Mörwald ea 2009]

Object Tracking and Modelling

[Mörwald ea 2011]

Overview

- Detection / segmentation
- Modelling
- Recognition
- Classification
- Tracking
- **Attention**
- Application to action learning

Human attention

- Test showing the necessity and effectiveness of attention for the human visual system
- In the following video, count how many times the players wearing white pass the basketball
- Just observe and count silently, don't distract the other participants
- Ready ...?

Human attention

Play video ..

Human attention

How many passes?

One more test ...

Look at the scene ...


Freiheit
Computer Vision

- How many boxes?
- How many objects had red in them?
- Was the laptop turned on?
- How many books? Gorillas ..?
- Speed of processing in the human visual system [Thorpe ea 1996]: ca. 150 ms to get scene gist

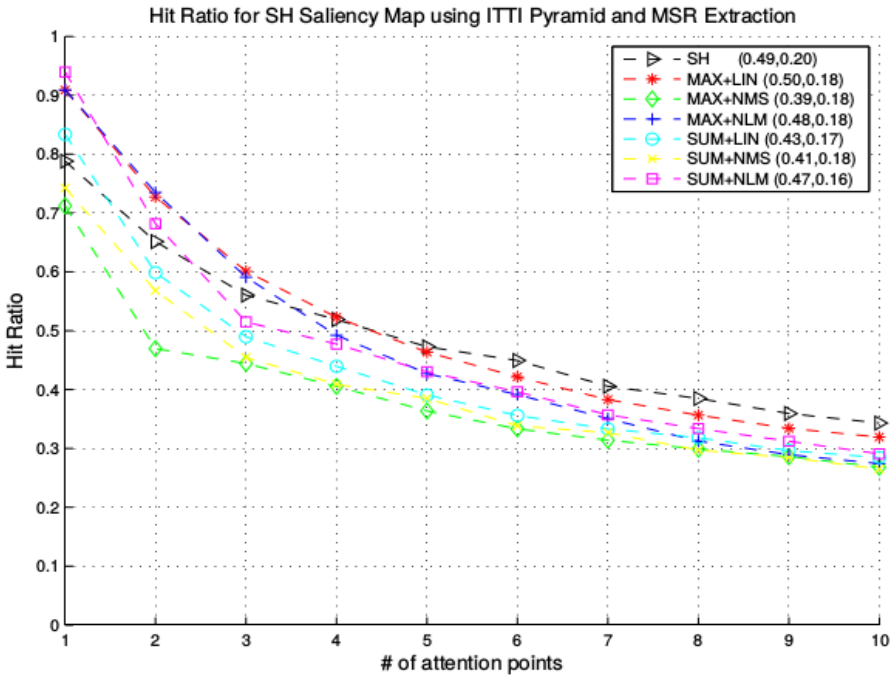
Visual attention

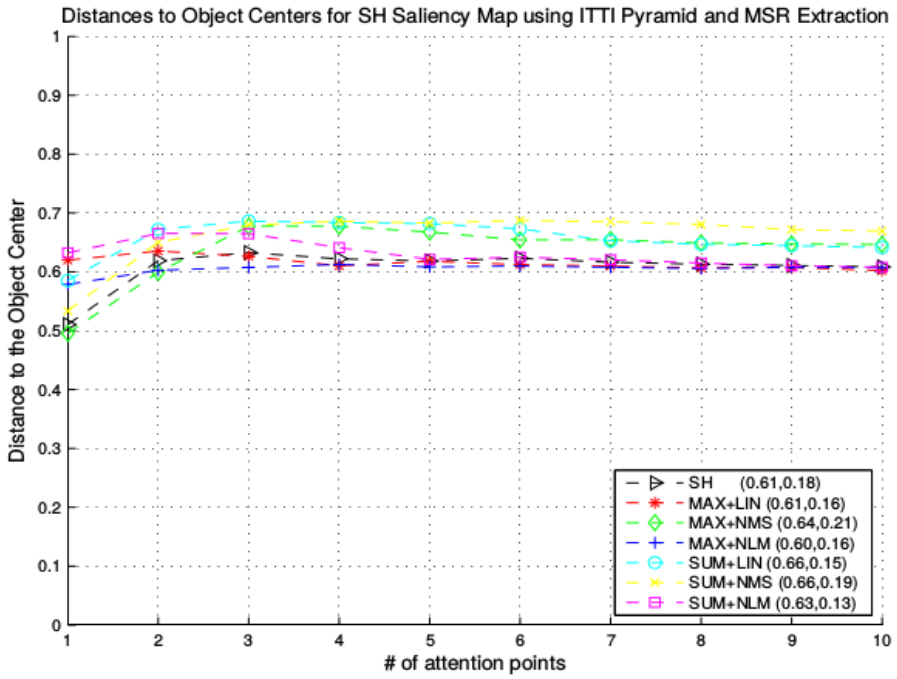
- Humans don't take in the whole scene at once in all details
- So why should we expect robots to?

- Many vision problems become a lot easier (or feasible at all) once the object is large in the image center
- Bottom up saliency (e.g. colour contrast)
- Top down, task-driven attention
- Many saliency computation methods, e.g. surveys [Borji & Itti PAMI'13, Potapova ea IJRR 2017]

Visual attention

Evaluating saliency methods (typical example)

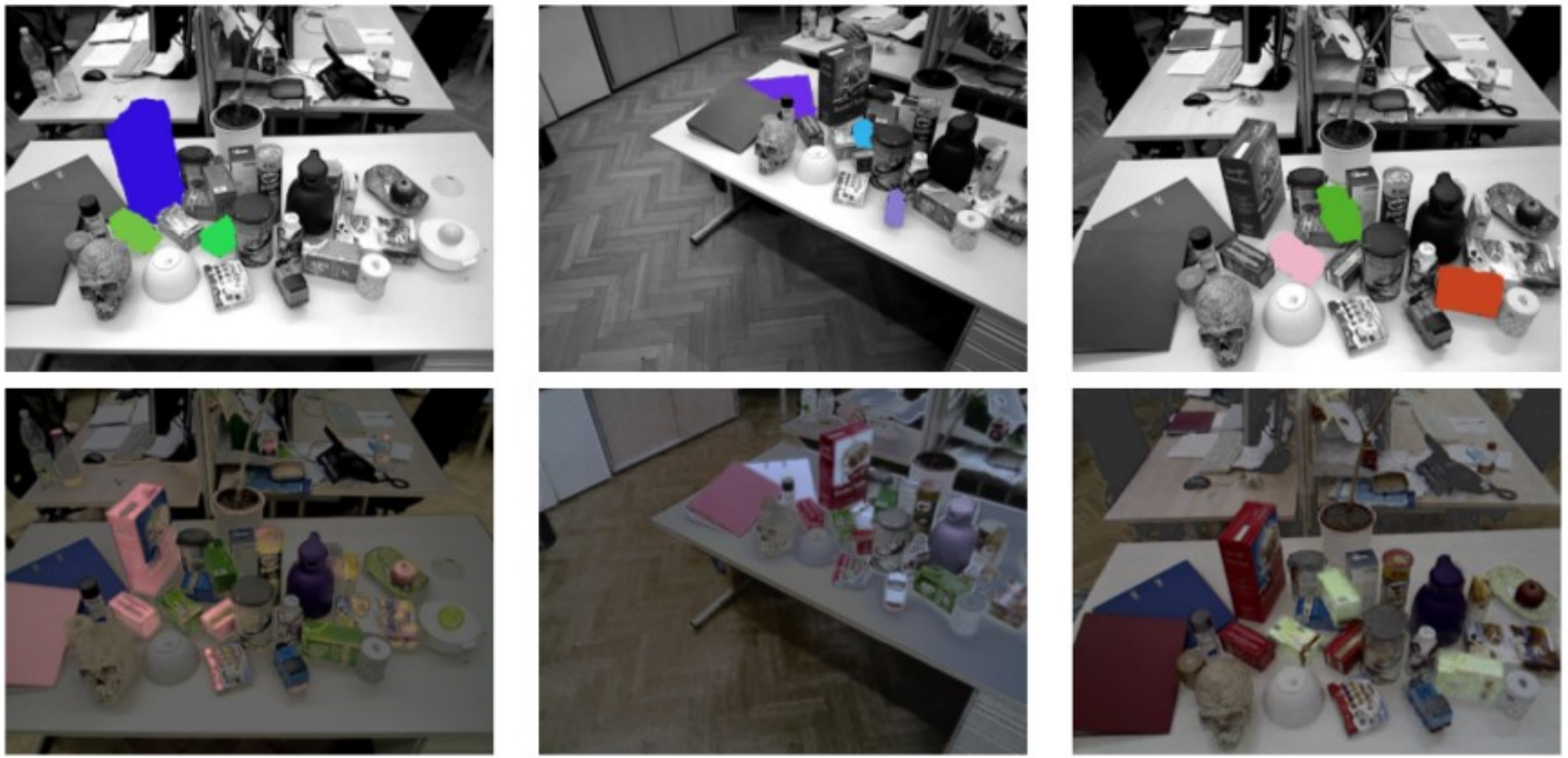




=> Attention per se is meaningless, needs to be understood in task context

=> How to *use* attention

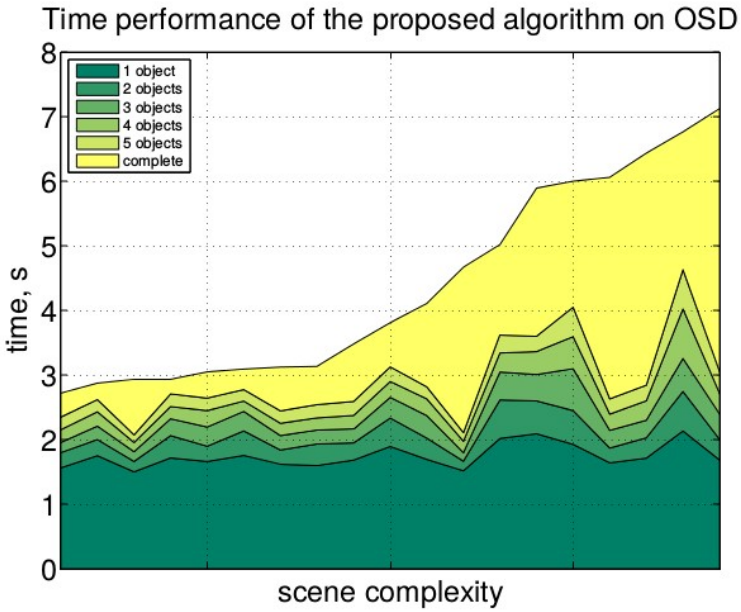
Attention-guided Segmentation

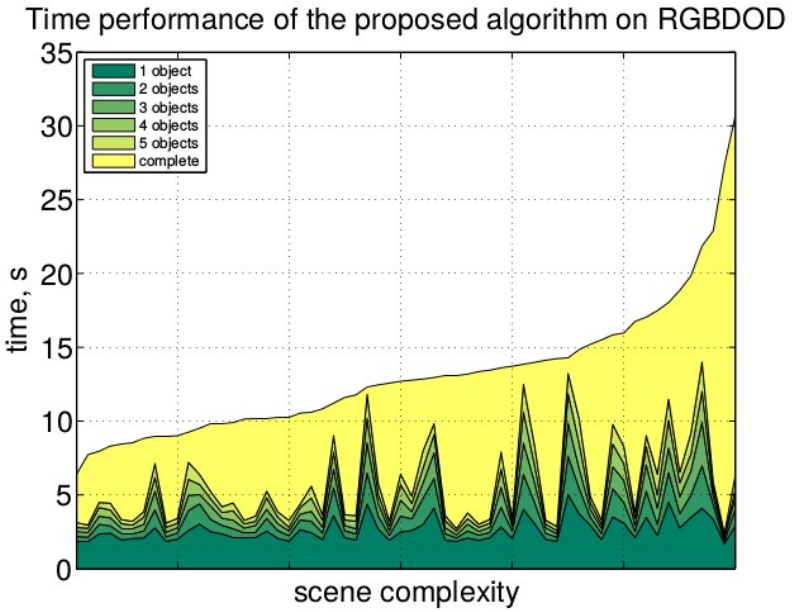


Red Objects *Blue Objects* *Green Objects*

Potapova et al.: Incremental Attention-driven Object Segmentation, Humanoids'14

Attention-guided Segmentation





low *middle* *high*

Object Segmentation Database [19]

low *middle* *high*

RGB-D Object Dataset [13]

Overview

- Detection / segmentation
- Modelling
- Recognition
- Classification
- Tracking
- Attention
- **Application to action learning**

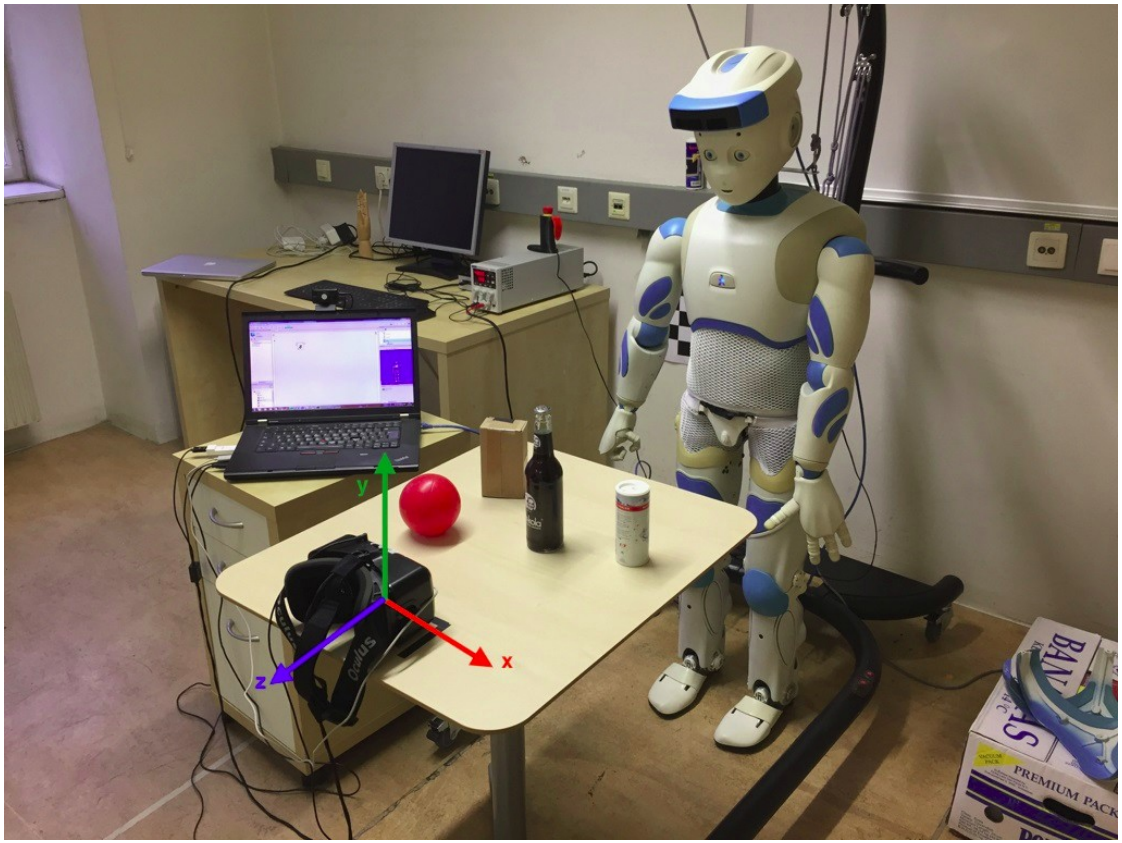
Project RALLI

Robotic Action-Language Learning through Interaction (OFAI, TU Wien, Tufts University)

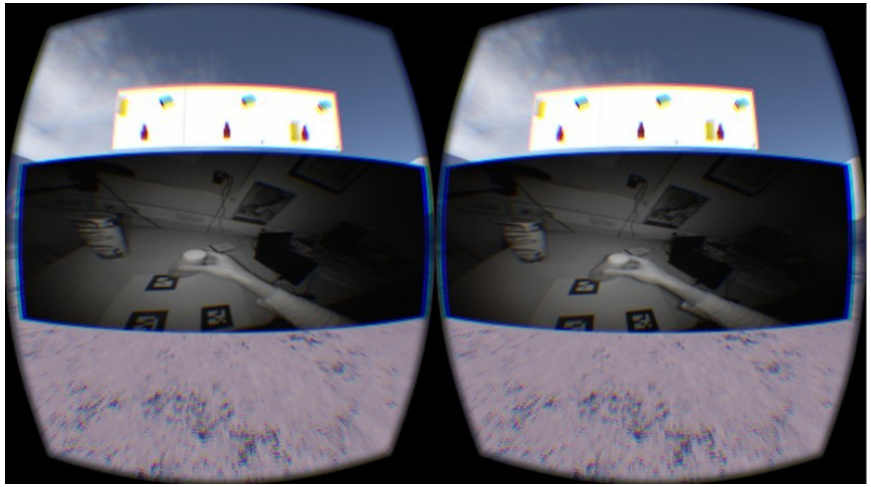
- Motivated by early child development
- Human tutor shows and explains action (“I am pushing the ball”, “I take the bottle” etc.
- Robot learns action verbs and their meaning in terms of object and hand trajectories
- Robot shows that it learned by repeating actions

Project RALLI

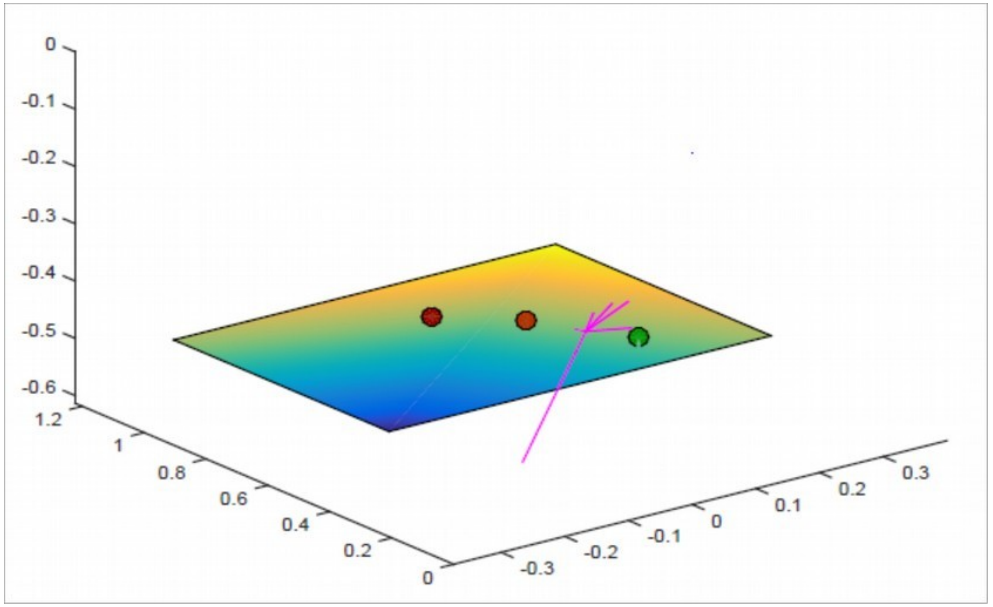
Experimental setup with Romeo (Asus in head), Oculus Rift headset (head tracking) and Leap motion (hand/finger tracking)



Project RALLI



View though the Oculus including the instructions



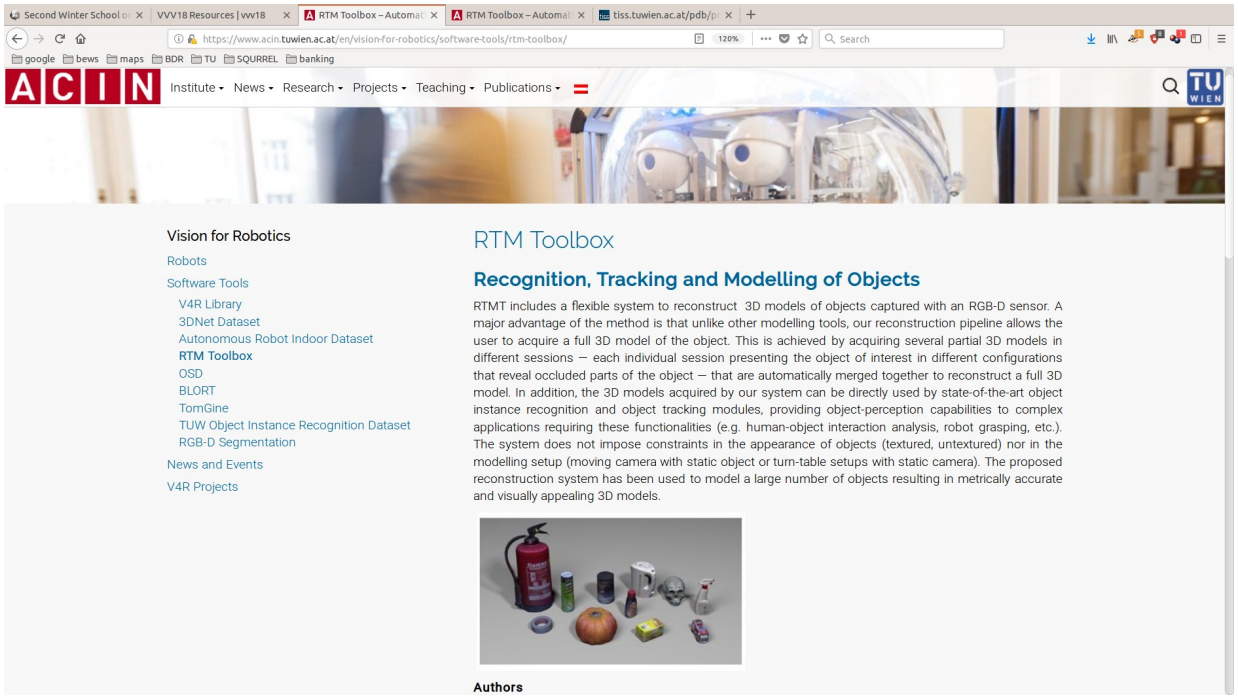
Visualisation of hands and tracked objects

Project RALLI

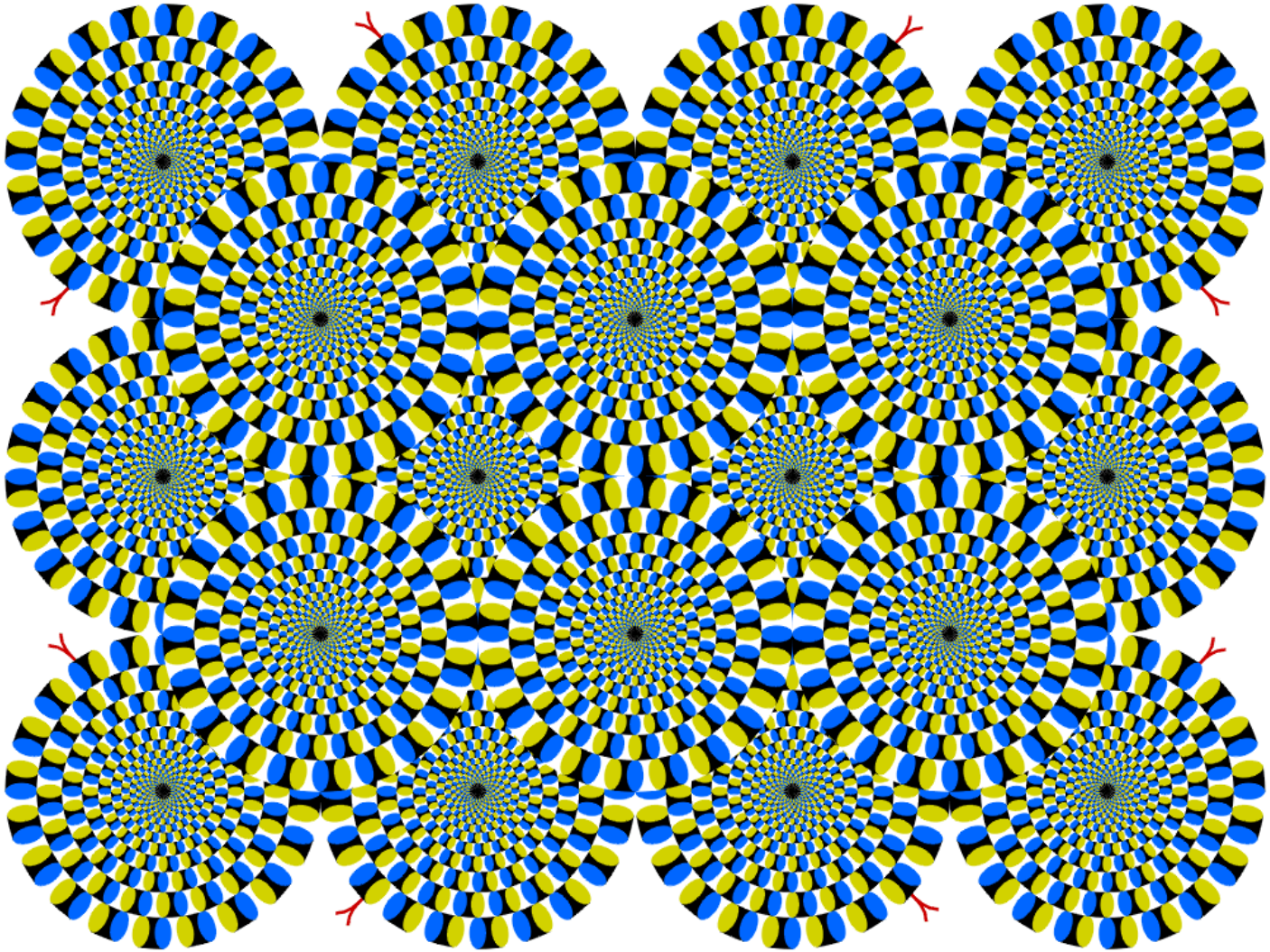
- Example of annotated speech, synchronised with object/hand tracks

Summary

- Present tool box for various robot vision tasks
- https://www.acin.tuwien.ac.at/en/vision-for-robotics/software-tools/v4r-library/
- https://www.acin.tuwien.ac.at/en/vision-for-robotics/software-tools/rtm-toolbox/



Questions?



Many thanks to my colleagues who did all the actual work (in no particular order)

Johann Prankl
Thomas Mörwald
Paloma de la Puente
Thomas Fäulhammer
Aitor Aldoma Buchaca
Ekaterina Potapova
David Fischinger
Karthik Mahesh Varadarajan
Peter Einrahmhof
Walter Wohlkinger
Andreas Richtsfeld

- Mörwald, T., Zillich, M., & Vincze, M. Edge Tracking of Textured Objects with a Recursive Particle Filter. 19th International Conference on Computer Graphics and Vision (Graphicon) 2009.
- Wohlkinger, W., & Vincze, M. Shape-Based Depth Image to 3D Model Matching and Classification with Inter-View Similarity. IROS 2011.
- Zillich, M., Prankl, J., Mörwald, T., & Vincze, M. Knowing Your Limits - Self-evaluation and Prediction in Object Recognition. IROS 2011.
- Mörwald, T., Zillich, M., Prankl, J., & Vincze, M. Self-Monitoring to Improve Robustness of 3D Object Tracking for Robotics. In IEEE International Conference on Robotics and Biomimetics (ROBIO) 2011.
- Mörwald, T., Kopicki, M., Stolkin, R., Wyatt, J., Zurek, S., Zillich, M., & Vincze, M. Predicting the Unobservable: Visual 3D Tracking with a Probabilistic Motion Model. ICRA 2011.
- Wohlkinger, W., Buchaca, A. A., Rusu, R., & Vincze, M. 3DNet: Large-Scale Object Class Recognition from CAD Models. ICRA 2012.
- Richtsfeld, A., Mörwald, T., Prankl, J., Zillich, M., & Vincze, M. Segmentation of Unknown Objects in Indoor Environments. IROS 2012.
- Mörwald, T., Richtsfeld, A., Prankl, J., Zillich, M., & Vincze, M. Geometric data abstraction using B-splines for range image segmentation. ICRA 2013.
- Prankl, J., Mörwald, T., Zillich, M., & Vincze, M. Probabilistic Cue Integration for Real-time Object Pose Tracking. In Proceedings of the 9th International Conference on Computer Vision Systems (ICVS) 2013.
- Aldoma, A., Tombari, F., Prankl, J., Richtsfeld, A., Di Stefano, L., & Vincze, M. Multimodal Cue Integration through Hypotheses Verification for RGB-D Object Recognition and 6DOF Pose Estimation. ICRA 2013.
- Buchaca, A. A., Tombari, F., Stefano, L. di, & Vincze, M. A Global Hypotheses Verification Method for 3D Object Recognition. ECCV 2013.
- Fischinger, D., Jiang, Y., & Vincze, M. Learning Grasps for Unknown Objects in Cluttered Scenes. ICRA 2013.