LAMoR 2015
Computer Vision

Michael Zillich

Automation and Control Institute
Vienna University of Technology

What is vision for *robotics?*

Vision (hard!)

What is vision for *robotics?*

Robotics requirements:
- Real time
- Limited resources
- Real world clutter

↓

Vision (really hard!)

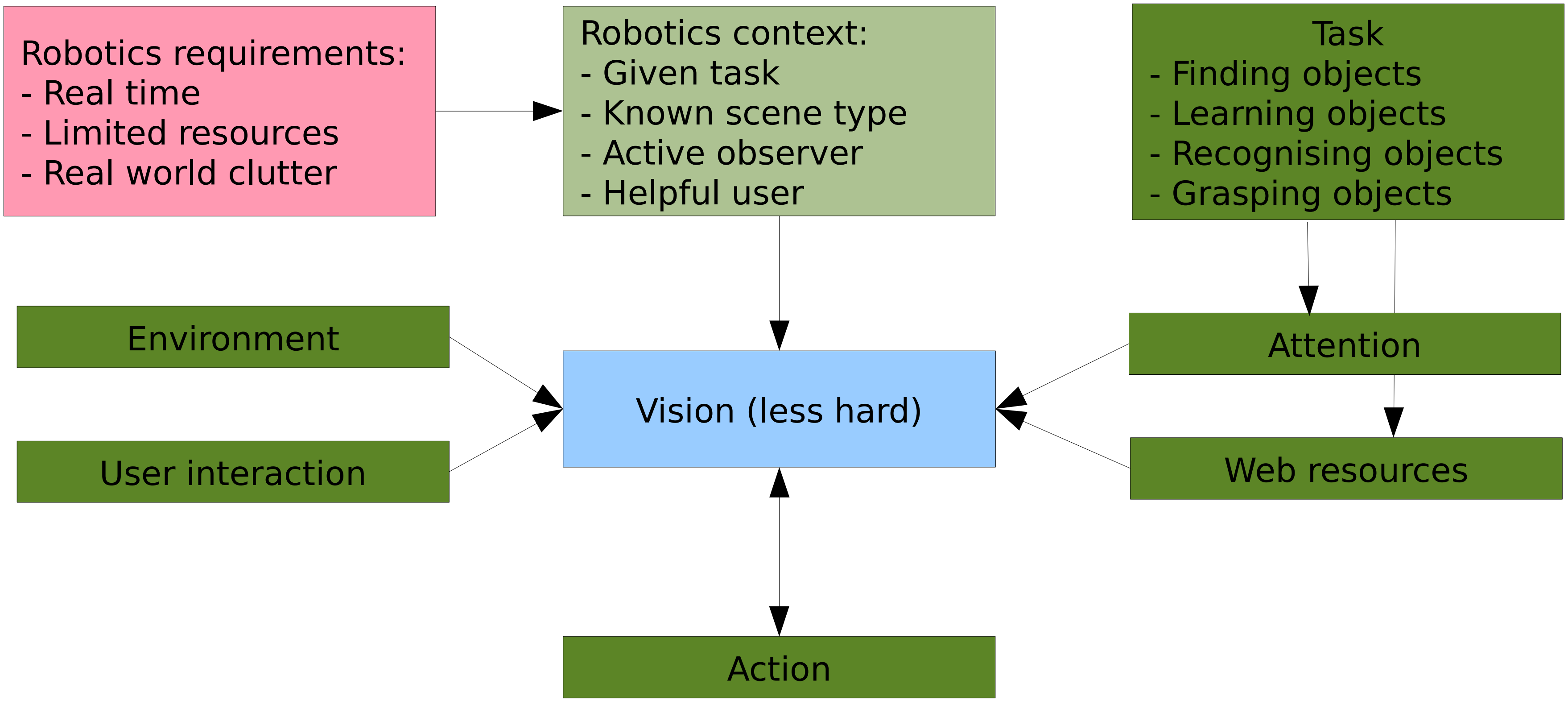
What is vision for *robotics?*

Robotics requirements:
- Real time
- Limited resources
- Real world clutter

Robotics context:
- Given task
- Known scene type
- Active observer
- Helpful user

Vision (less hard)

What is vision for *robotics?*



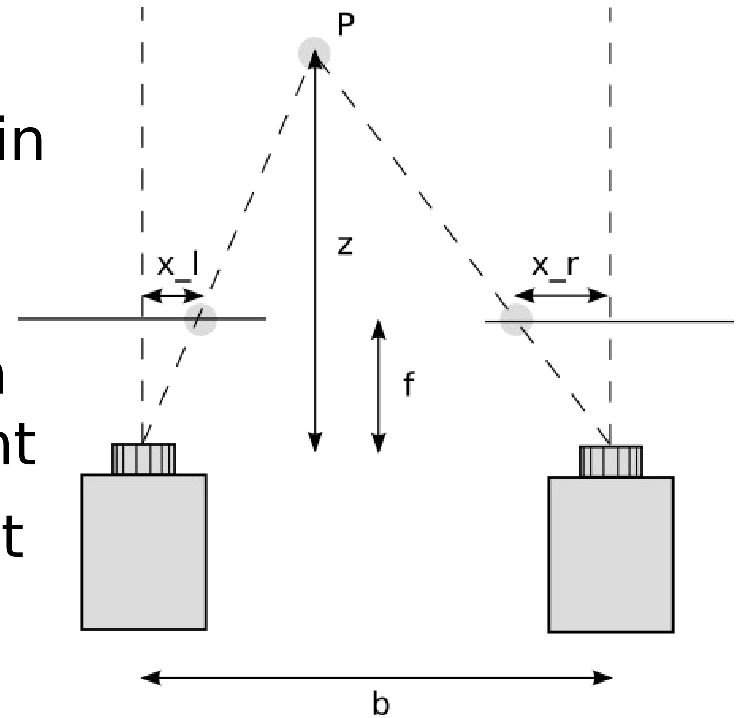
Overview

- **Sensors**
- Detection / segmentation
- Recognition
- Classification
- Tracking
- Attention

Sensors

Binocular stereo

- Find corresponding image features in left and right image
- With known camera intrinsic and extrinsic calibration calculate depth from disparity between left and right
- Possibly vergence, but often difficult to calibrate precisely



Disparity: $d = x_l - x_r$

Depth: $z = \frac{fb}{d}$

Sensors

Binocular stereo

- Any camera pair
- Point Grey Bumblebee

+ dirt cheap

+ works in any light

– requires texture

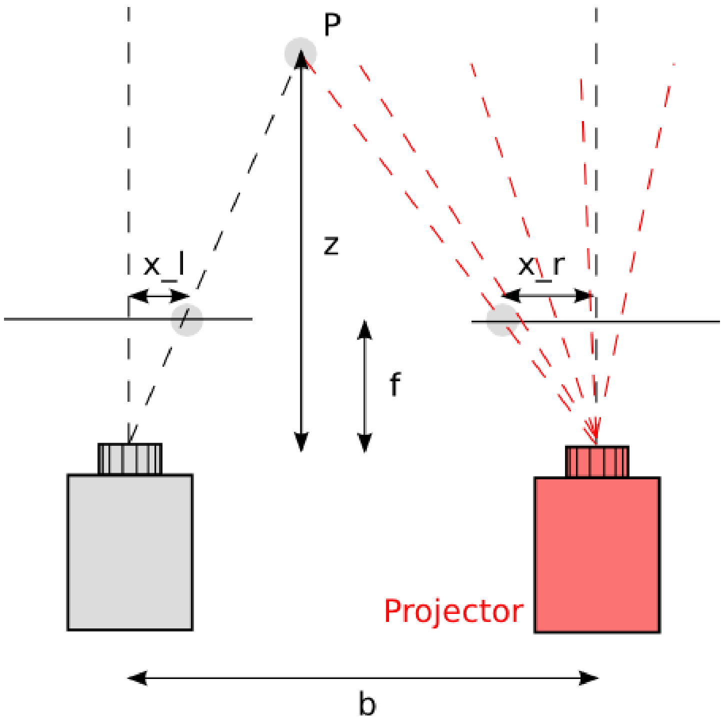
– selected disparity range limits range

Point Grey

Sensors

Projected light stereo

- Same principle
- Replace second camera with a pattern projector
- IR light with band-pass filter
- Combine with 3rd camera for RGBD



Sensors

Projected light stereo

- Microsoft Kinect
- Asus Xtion Pro Live
- Primesense Carmine (discontinued)

+ dirt cheap

+ fairly accurate

+ OK resolution (320x240)

– sensitive to external lighting

– minimum distance of e.g. 0.5 m (stereo disparity range)

– problems with reflective, dark, translucent surfaces

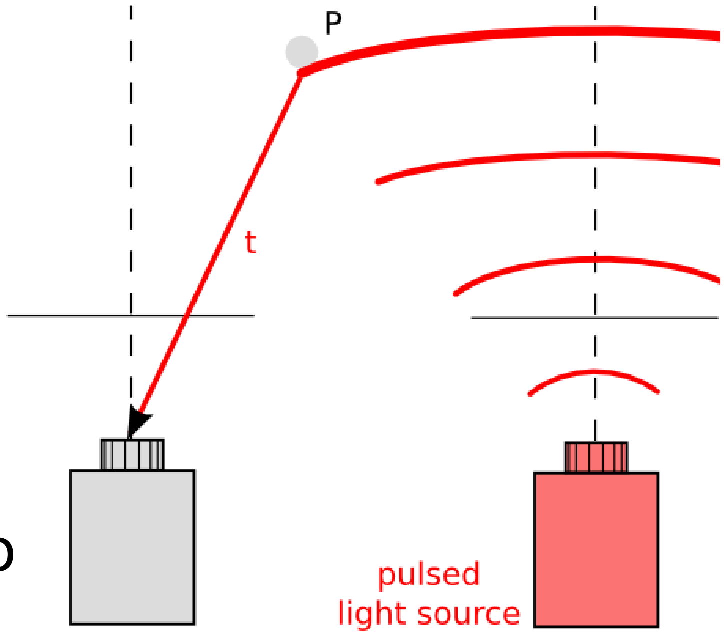
Microsoft

Asus/Primesense

Sensors

IR time of flight (TOF)

- Pulsed IR light source synchronised with camera
- Measure time of flight of light pulse per pixel
- With light speed calculate distance
- Varying source intensity to adjust to lighting conditions (avoid saturation)



Sensors

IR time of flight (TOF)

- MESA Imaging SwissRanger
- SoftKinetic DepthSense
- Fotonic
- BlueTechnix Argos
- Microsoft Kinect 2

\+ better robustness to external light

\+ from 1 cm to several m

\+ frame rate up to 160 Hz

– more noise

– slightly more expensive

– high energy consumption (IR LED lighting)

\- problematic artefacts

MESA

SoftKinectic

BlueTechnix

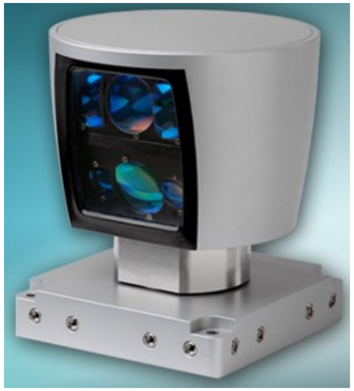
Fotonic

Microsoft

Sensors

Laser time of flight

- TOF principle again, but with array of sweeping laser beams
- Very precise measurement



- Velodyne HDL-64E

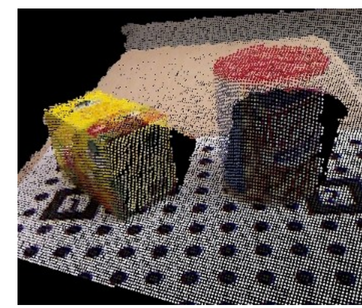
+ very robust to external lighting

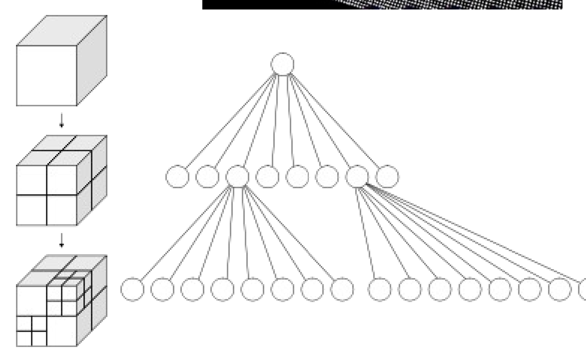
+ very accurate (up to 2 mm at 20 m)

– very expensive (> tens of thousands €)

Data Types

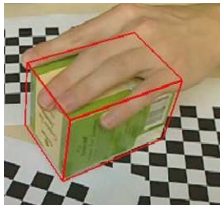
- Depth image + RGB image, plus known calibration
- Organised point cloud: image of XYZRGB data, efficient access to neighbours
- Unorganised point cloud
- 3D Voxel grid, possibly with varying resolution to save space (octree)

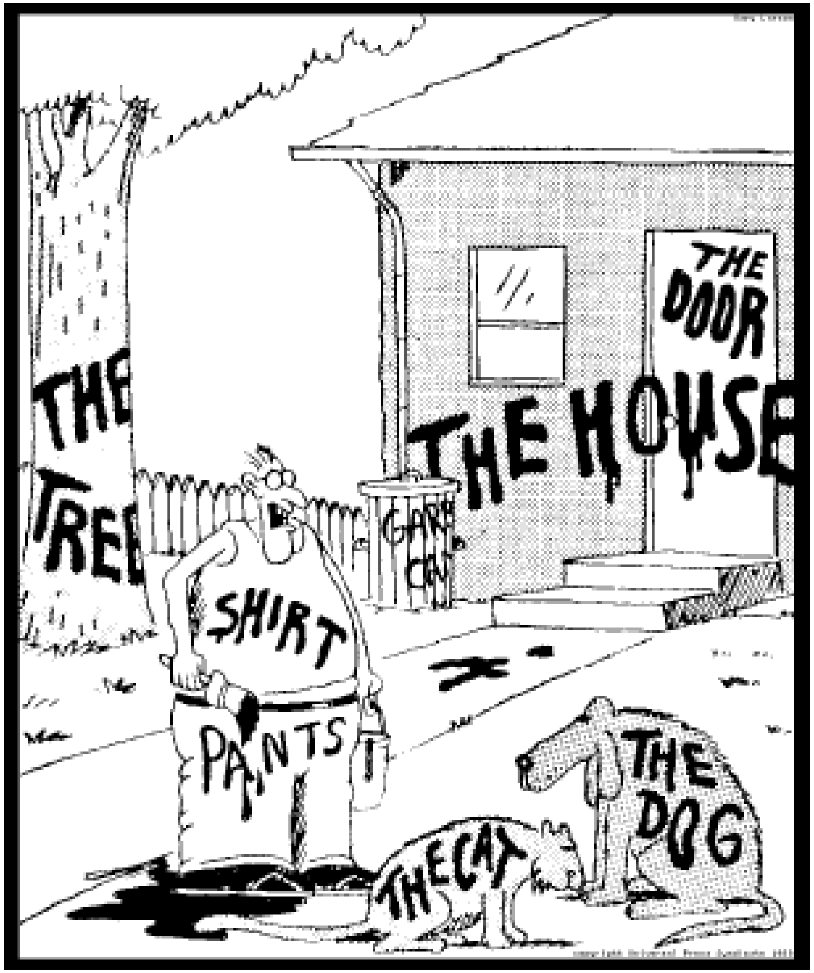




Object X

- 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



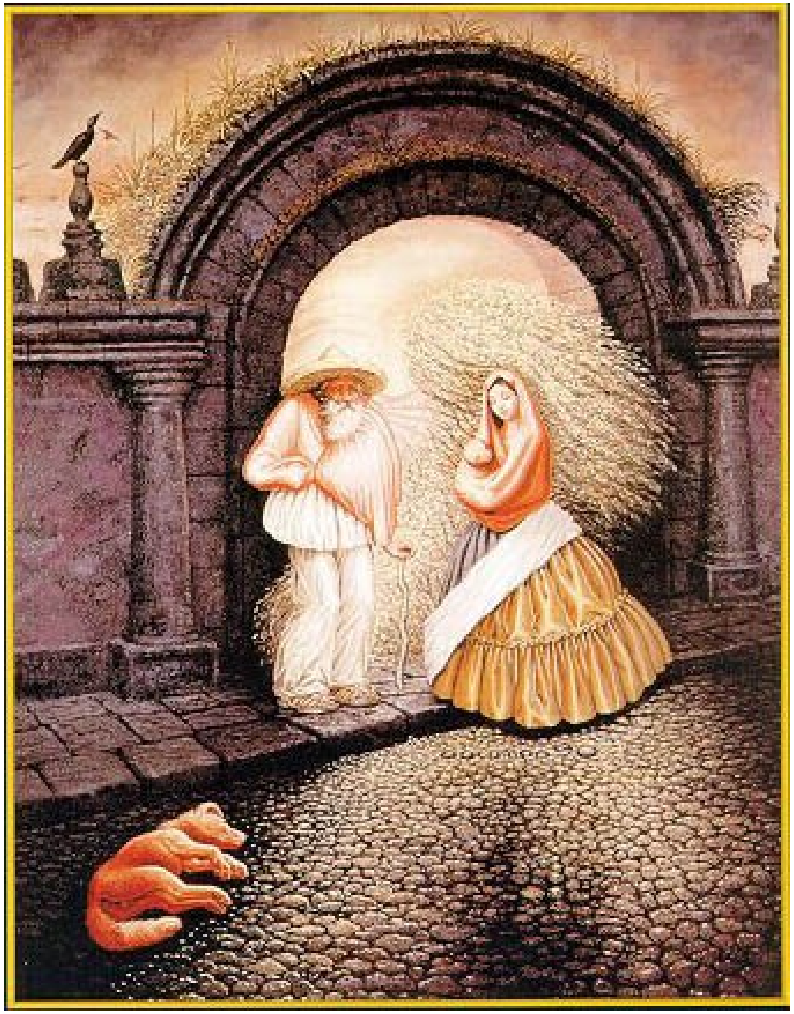


“Now! *That* should clear up a few things around here!”

Overview

- Sensors
- **Detection / segmentation**
- Recognition
- Classification
- Tracking
- Attention

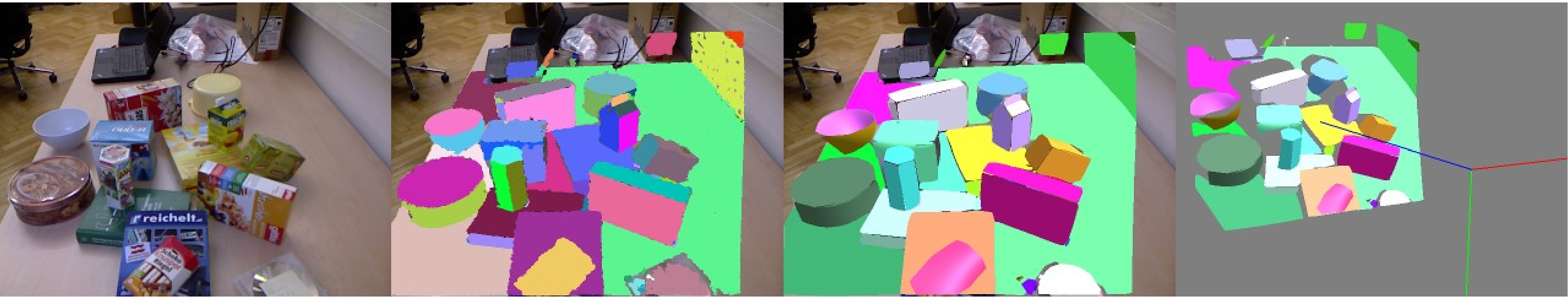
What is the object?



What is the object?

Object Segmentation

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

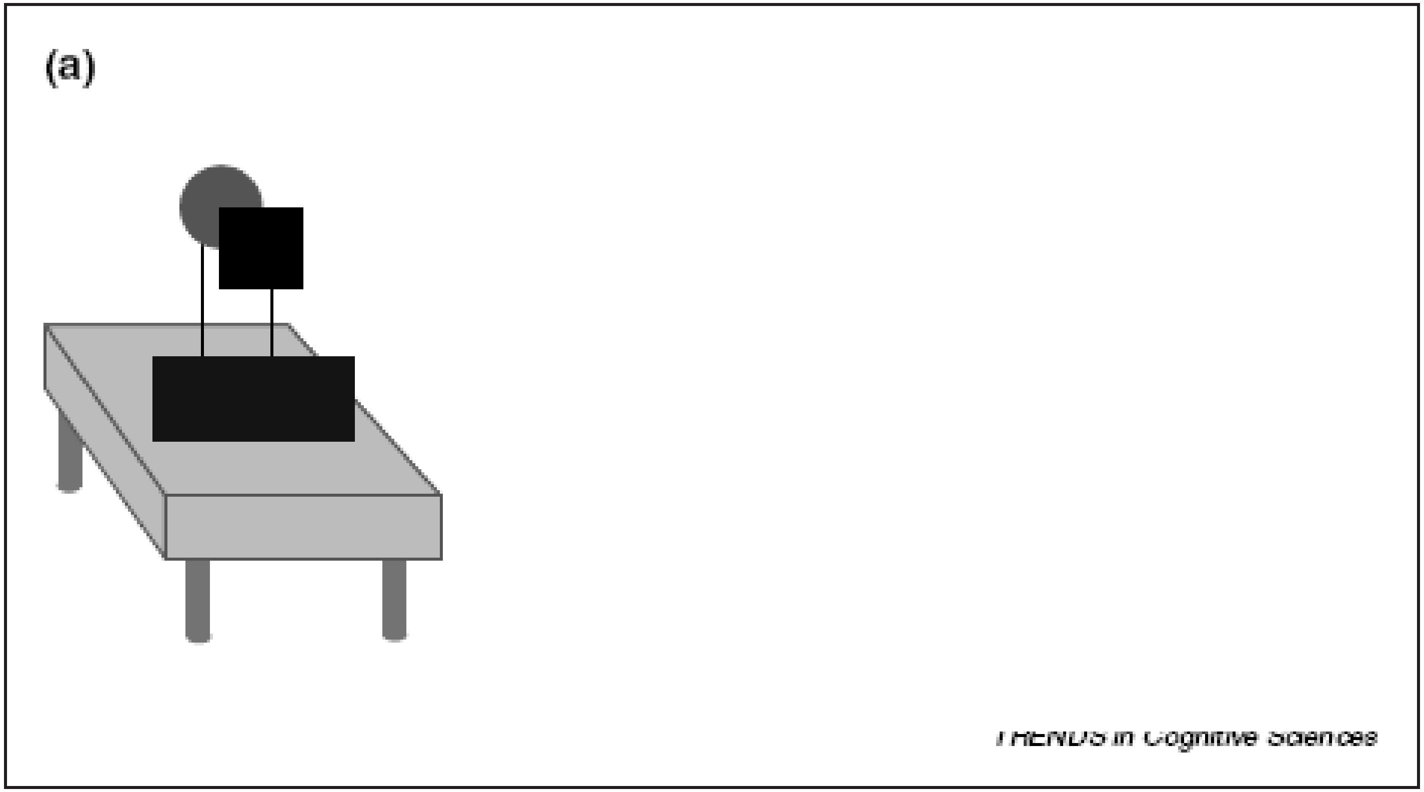


From coloured point clouds ... ... to separated object hypotheses

[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

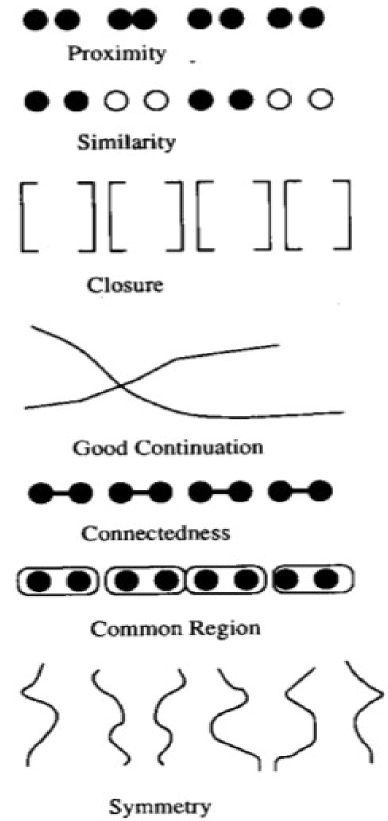
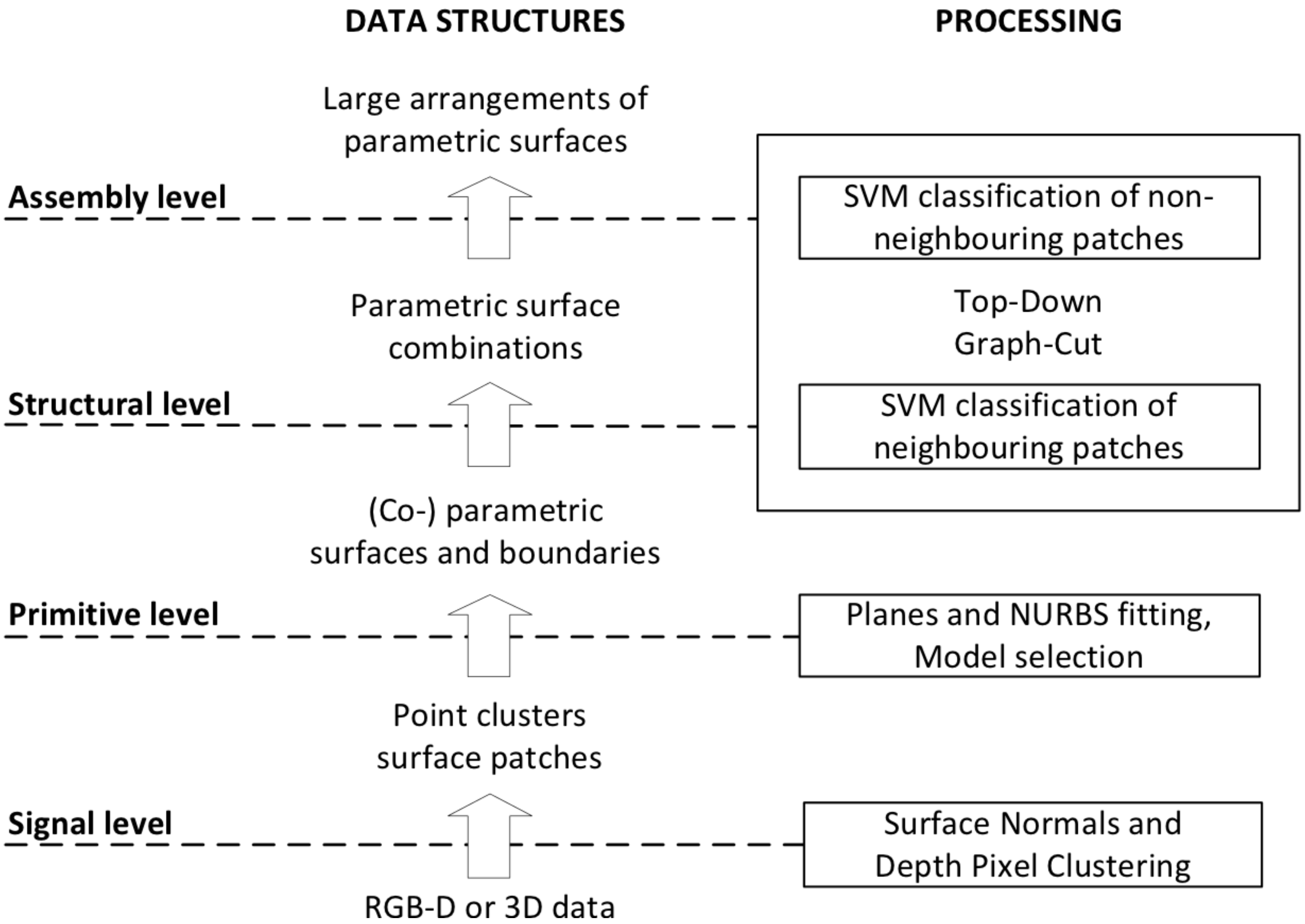


Fig. 3. Gestalt laws of grouping.

Object Segmentation

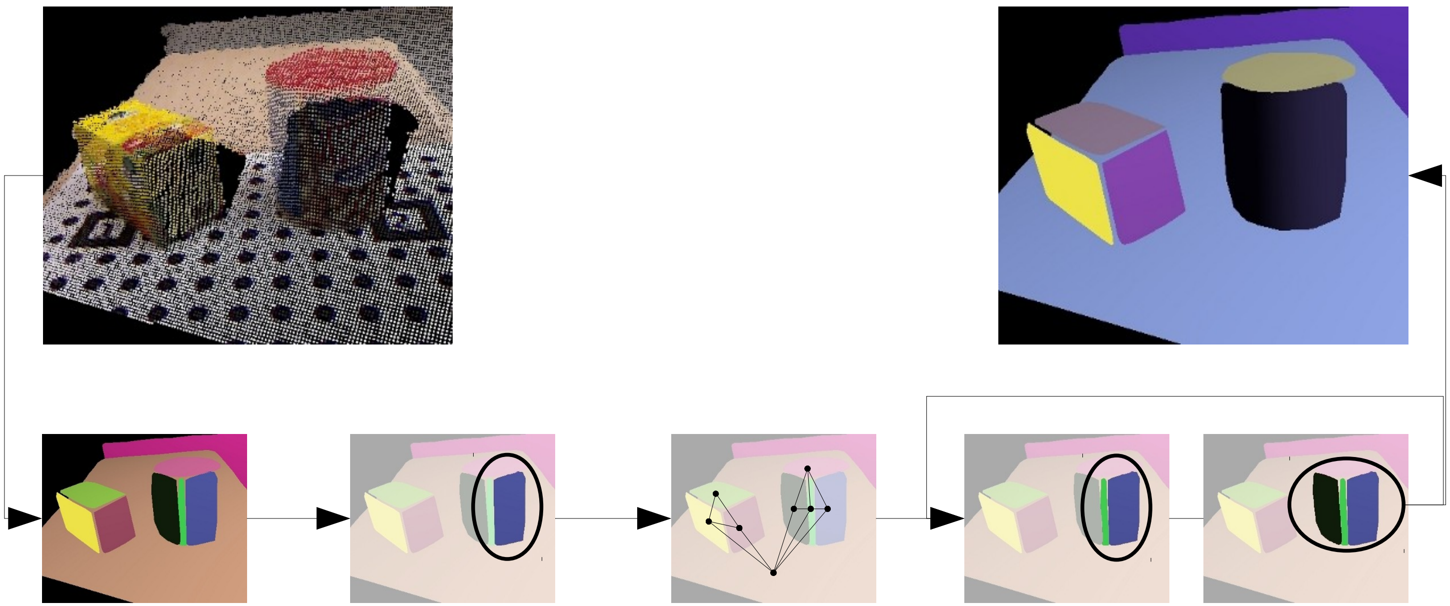


Object Segmentation: Surface

- Fitting surface patches
- Minimum Description Length (MDL) model selection [Leonardis ea 1995] to find optimal description

Input: point cloud

Output: Planes / NURBS



segment patches

select patch P_i compare models plane ↔ NURBS

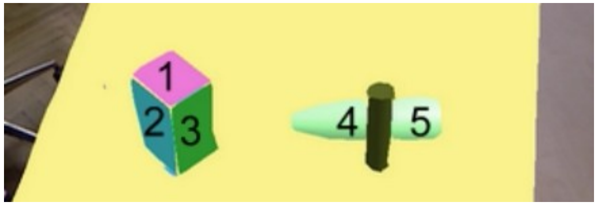
$S_N > S_P$

compute neighbours

Greedily fit NURBS to neighbours P_{ij} and compare models

$S_{ij} > S_i + S_j$

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

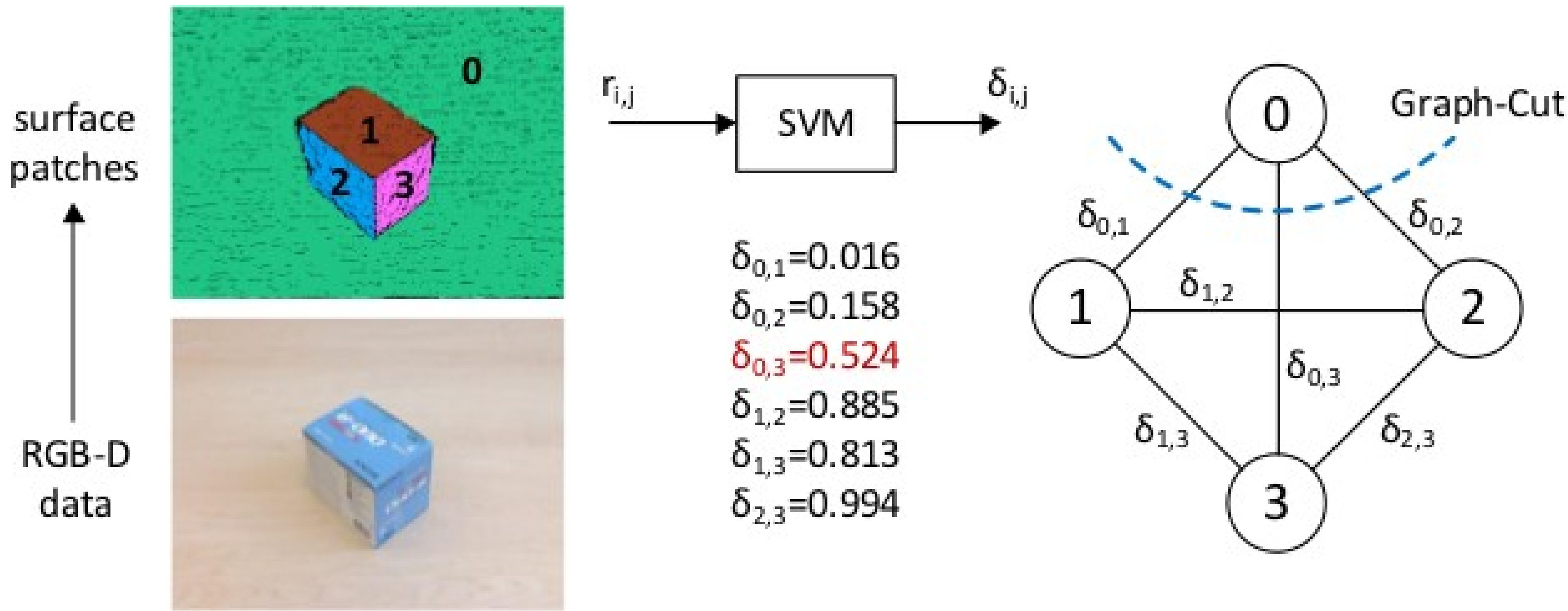
Global decision using graph cut

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

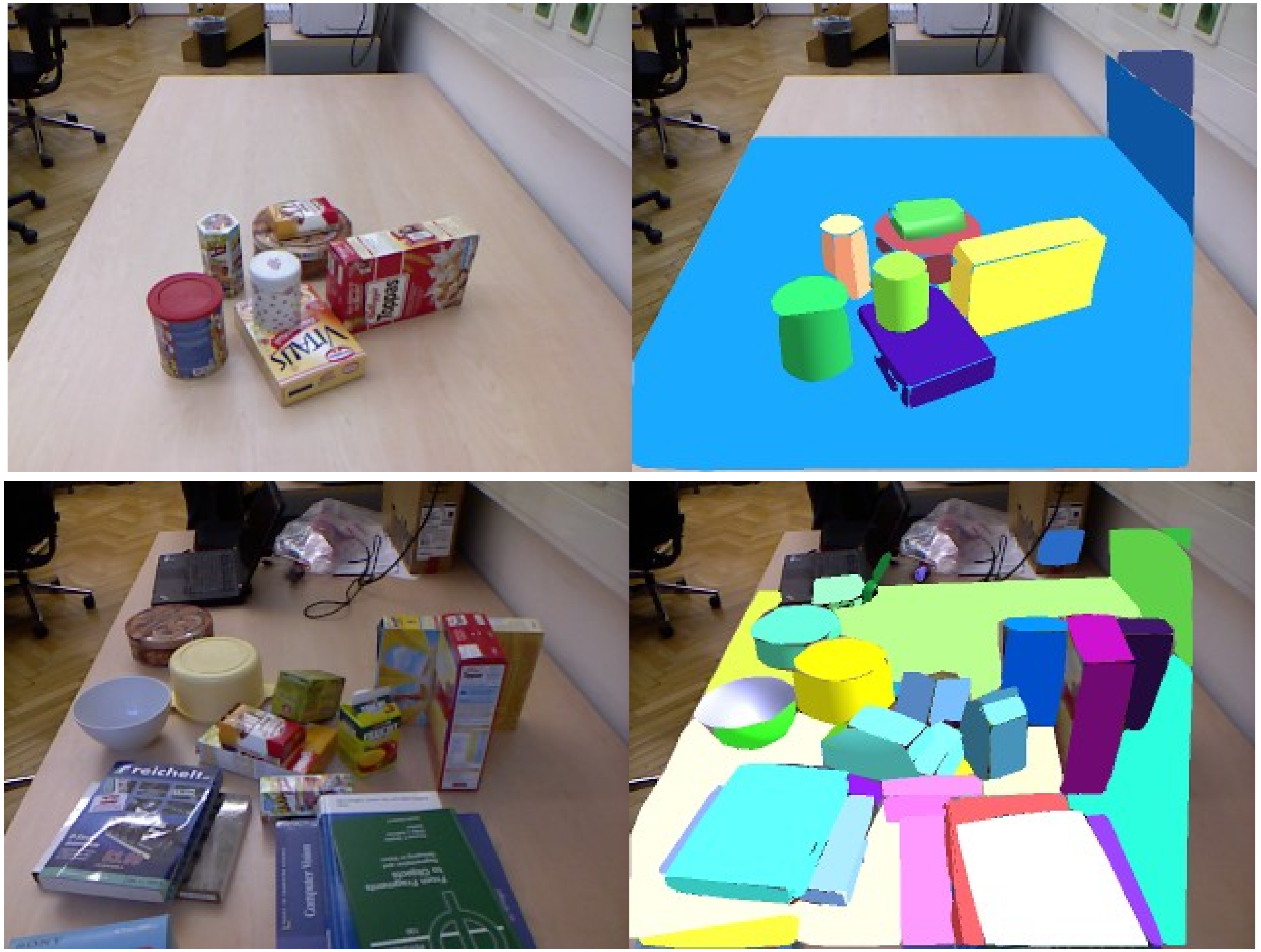
  r_st = (r_co, r_rs, r_tr, r_ga, r_fo, r_co3, r_cu3, r_cv3, r_di2, r_vd2)

  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]

Object Segmentation

Segmentation of Unknown Objects in Indoor Environments

A. Richtsfeld, J. Prankl, T. Mörwald, M.Zillich, M. Vincze

[Richtsfeld ea IROS'12]

Look at the scene ...


reichelt

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

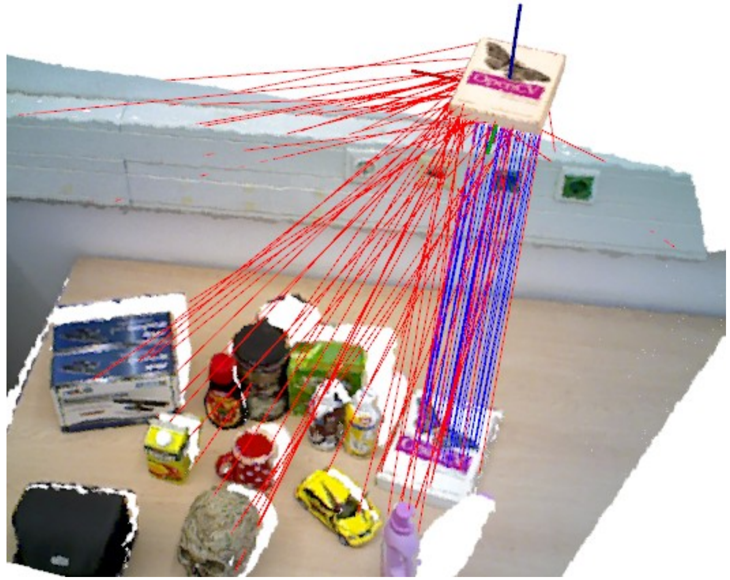
Overview

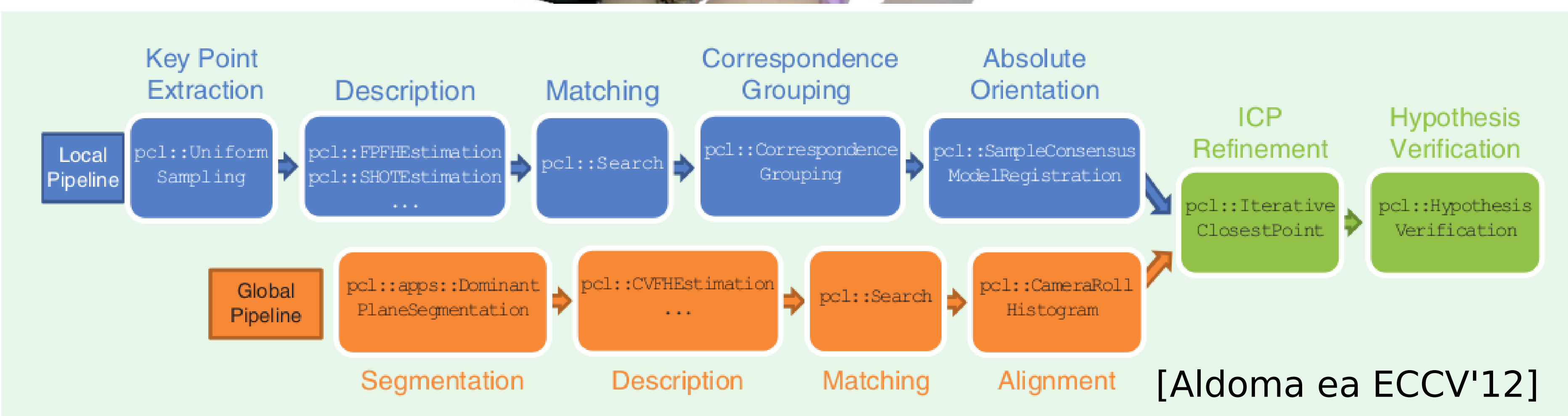
- Sensors
- Detection / segmentation
- **Recognition**
- Classification
- Tracking
- Attention

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**

Typical pipeline



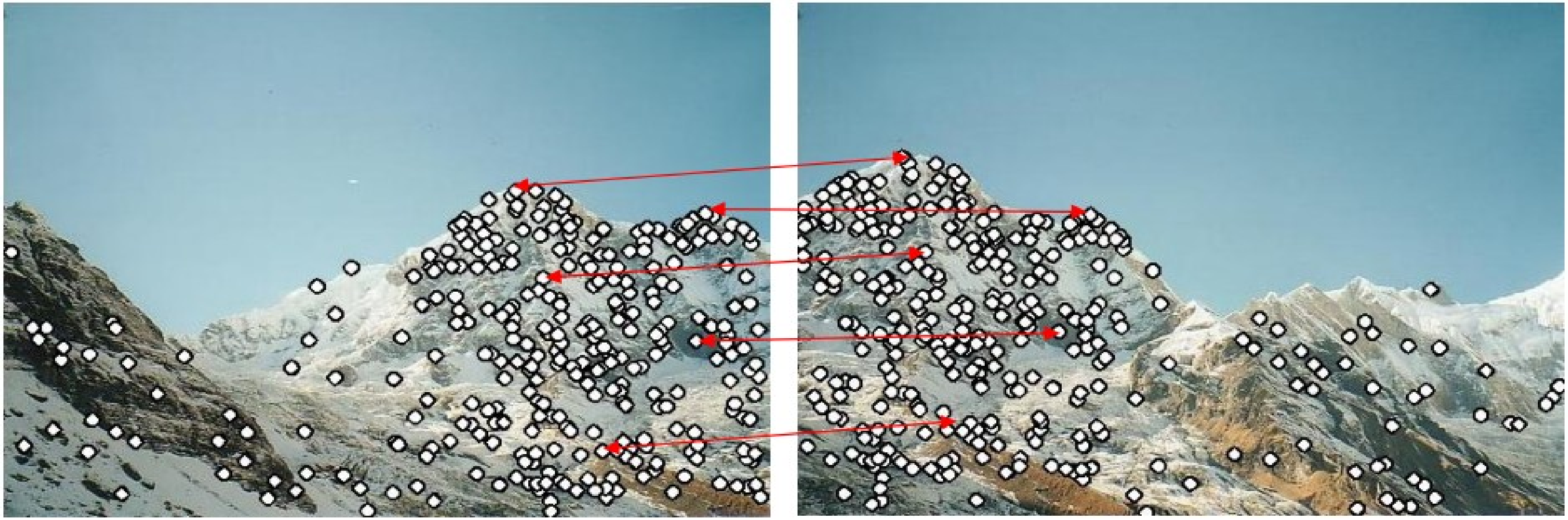


[Aldoma ea ECCV'12]

Features - 2D

Classic feature based 2D recognition

- Find interest points in both images
- Find corresponding point pairs
- Align



Features - 2D

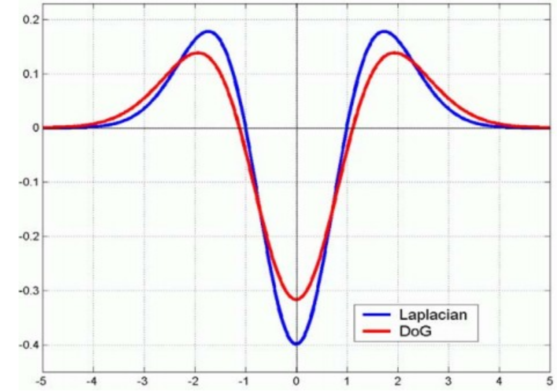
Classic feature based 2D recognition

- Find interest points in both images
- Find corresponding point pairs
- Align

Features 2D – Interest points

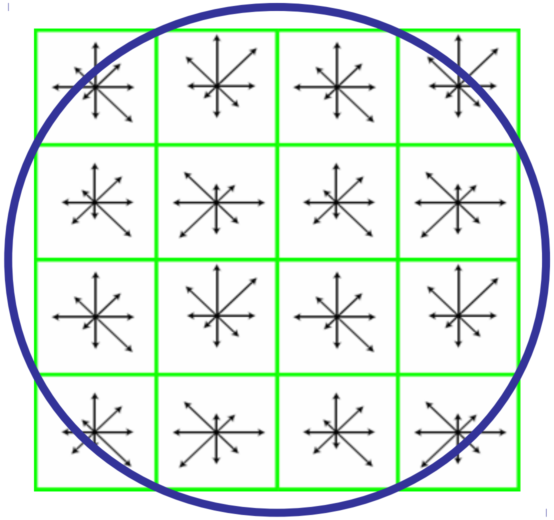
- Harris corners
  Autocorrelation in neighbourhood
  of points

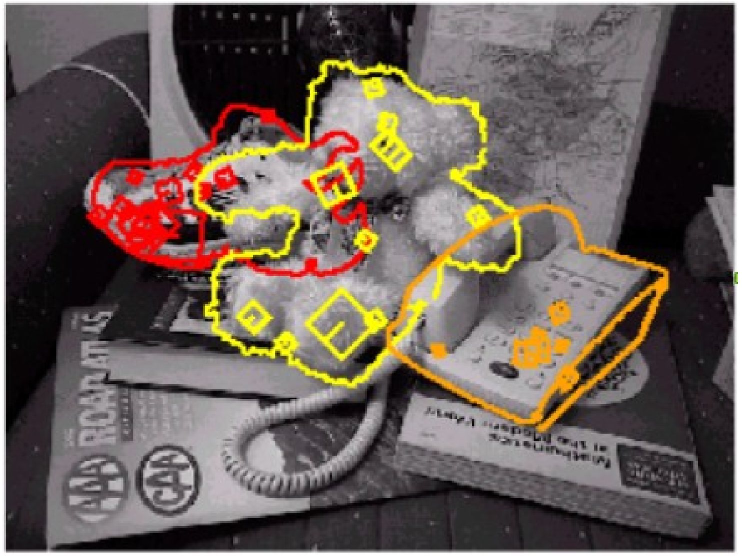
- Difference of Gaussians (DoG)
  Filter with “Mexican Hat” kernel



Features 2D - Descriptors

- Local description around interest point
- Classic: SIFT [Lowe 2004]
  Histograms of gradient orienations
  4 x 4 histograms, 8 orientations
  => 128 dim. vector

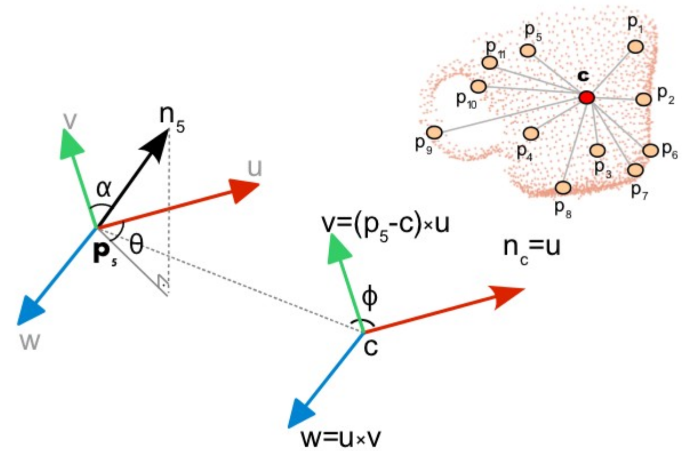




Features - 3D

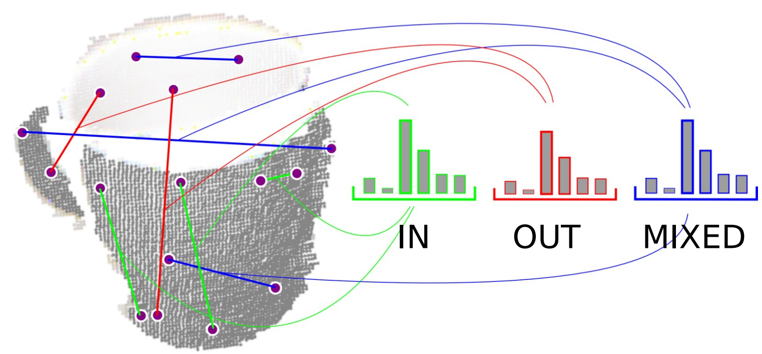
Local descriptors

- (FAST) Point Feature Histogram (PFH / FPFH) [Rusu ea 2008, Rusu ea 2009]
  3D Histogram of angles of key point and points in neighbourhood (angles between normals and distances)
  33 dim. Vector



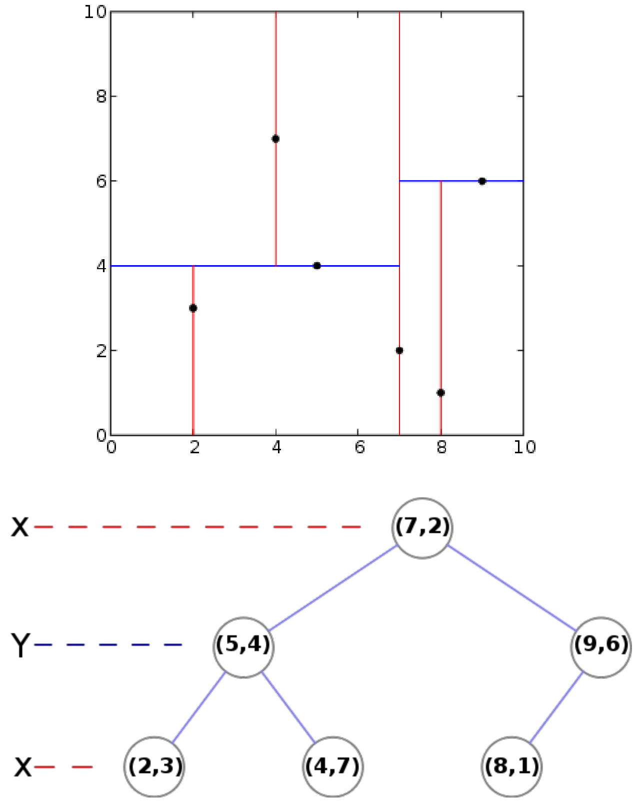
Global descriptors

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



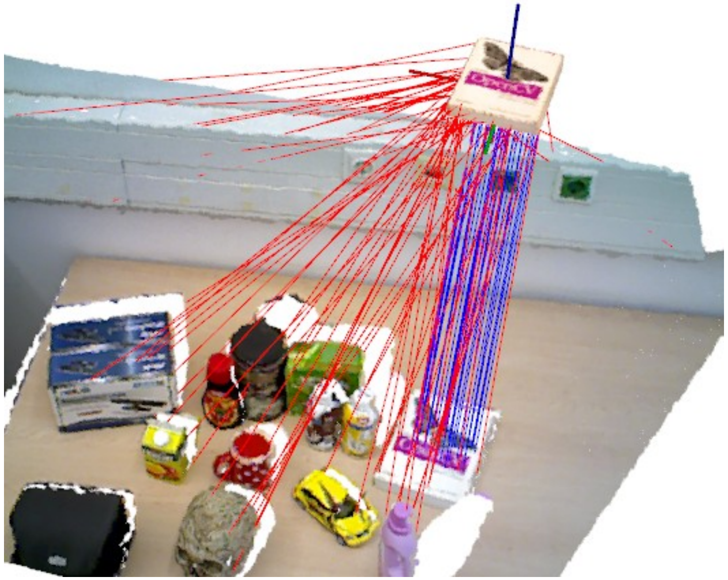
Matching

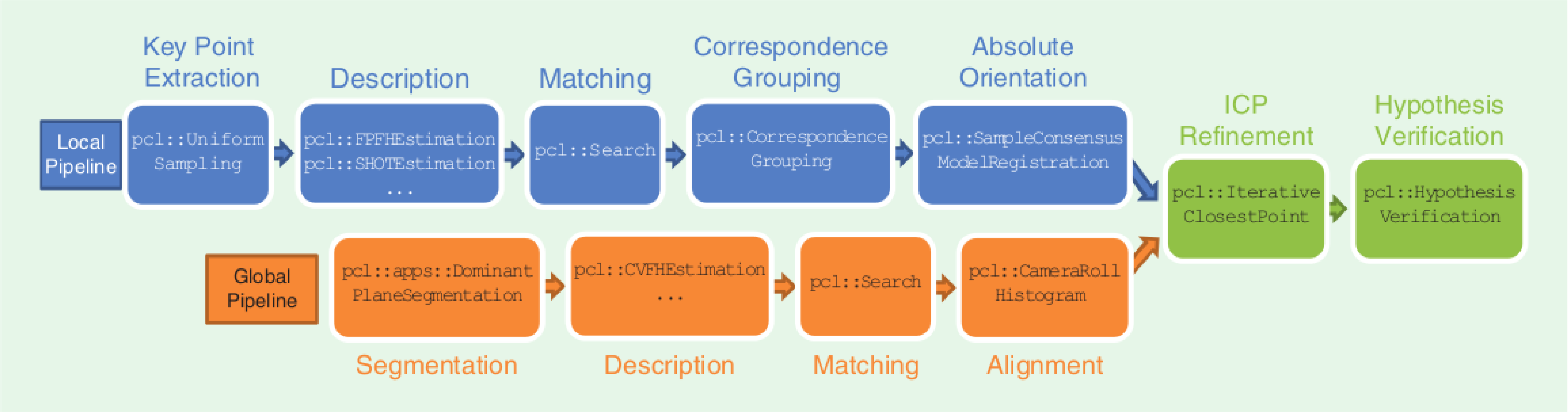
- Find point-to-point correspondences between query feature and feature in data base
- Nearest neighbour (NN) search in high-dimensional feature space, e.g. k-d tree, FLANN [Muja ea 2009] different distance norms (L1, L2, ..)
- Discard weak correspondences
  - Threshold (dangerous)
  - Ratio of distances closest / second nearest neighbour (should be small)
  - Just leave to later processing stage



k-d tree

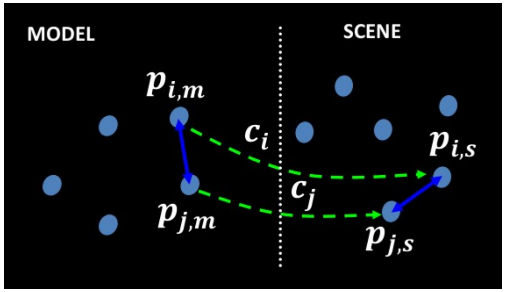
Typical pipeline



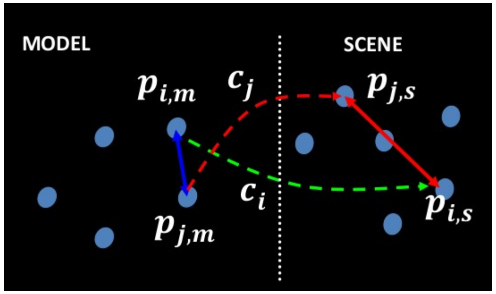


Pose estimation

- Correspondence grouping:
  Create groups of geometrically consistent point pairs
  Same distance between points in model and query data



consistent



inconsistent

- 6D pose fitting with RANSAC
  select minimum sample of point pairs to uniquely calculate
  6D pose [e.g. Horn 1987]
  gather consensus from other pairs
  best hypothesis wins

Refinement, verification

Iterative closest point (ICP) to align two point clouds

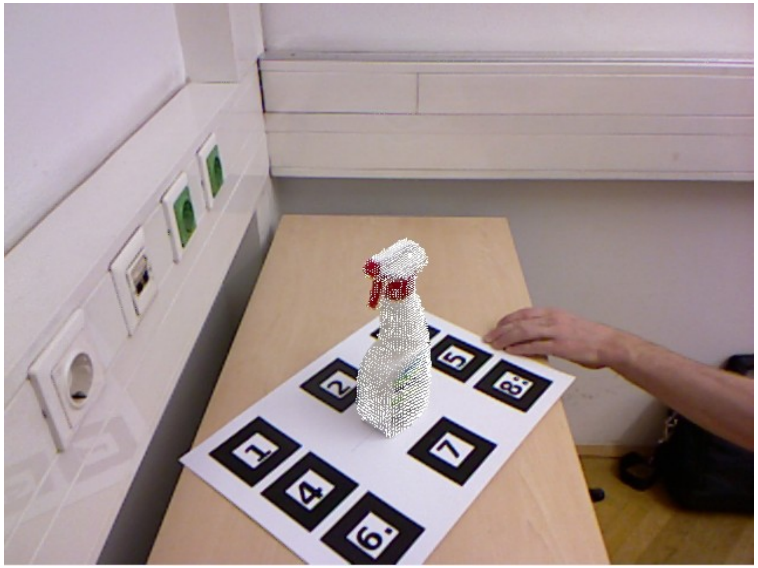
- For each point in the source point cloud, find the closest point in the reference point cloud
- Estimate the transformation that will best align each source point to its match found in the previous step
- Transform the source points using the obtained transformation
- Iterate (re-associate the points, and so on)
- Good initialisation is critical

Global hypothesis verification

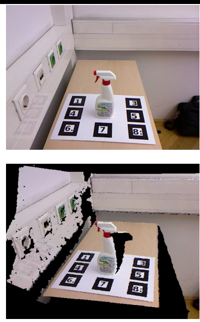
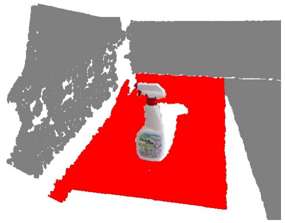
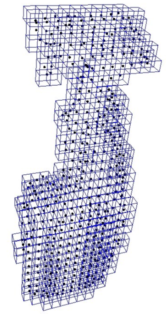
- Remove false positives, keep weak hypotheses if they make sense, decide between overlapping pose hypotheses using number of explained scene points, number of supporting points

Object modelling

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



Object modelling

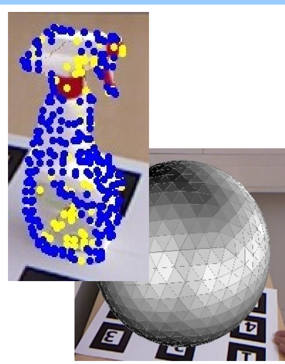
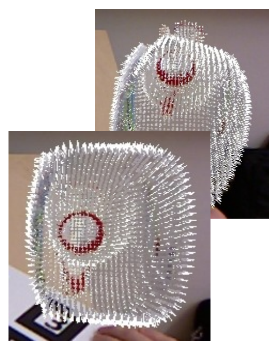




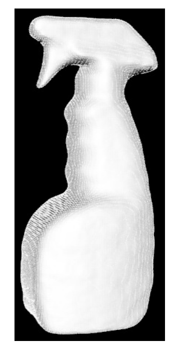

Input
- image
- point cloud

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

Object modelling

Object recognition: example scene

[Prankl 2010]

Overview

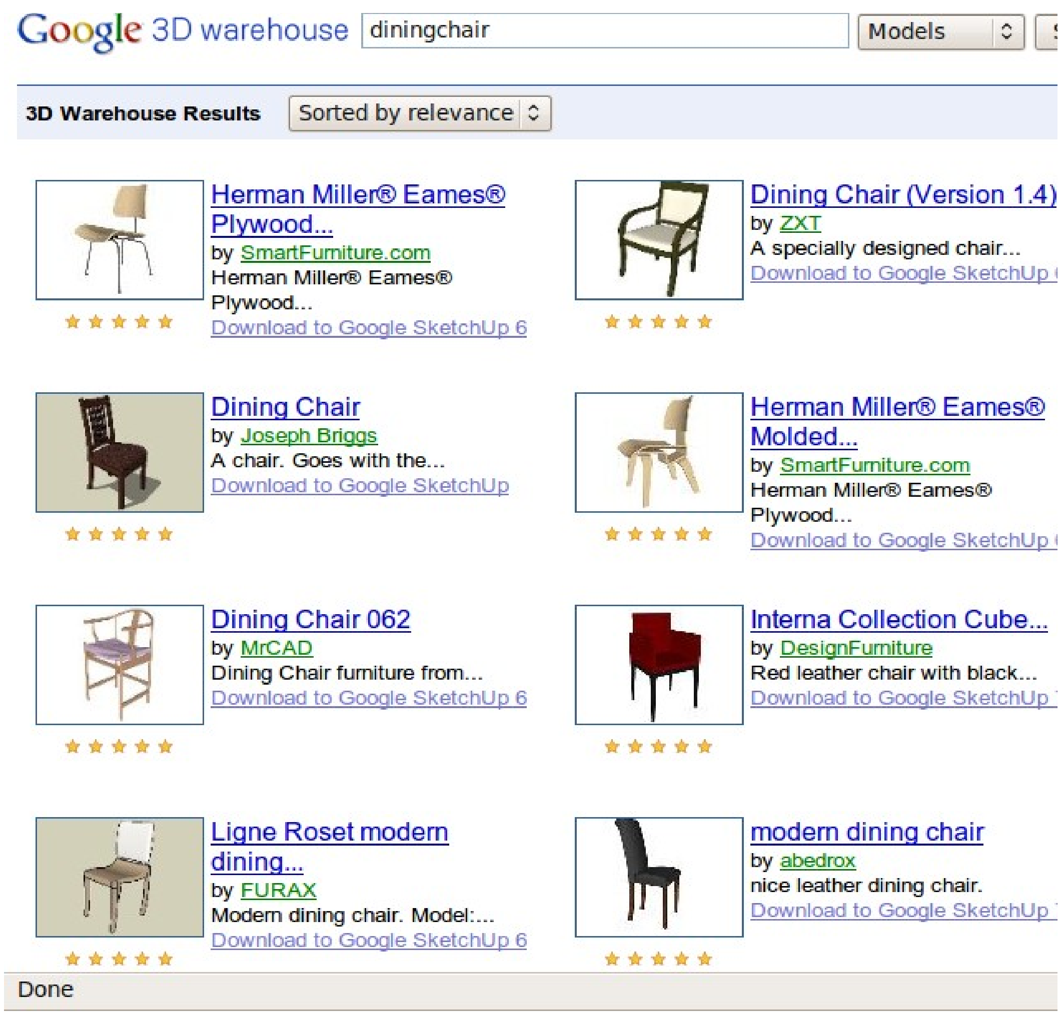
- Sensors
- Detection / segmentation
- Recognition
- **Classification**
- Tracking
- Attention

Object Categorisation

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

Offline training

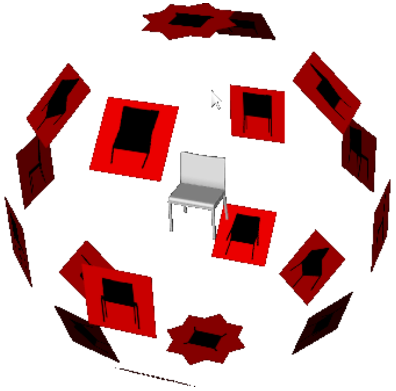
- E.g. “dining chair”
- Get many 3D CAD models, e.g. google 3D warehouse
- Find similar models from synonyms, e.g. Wordnet (mug, cup; chair, stool; etc.)

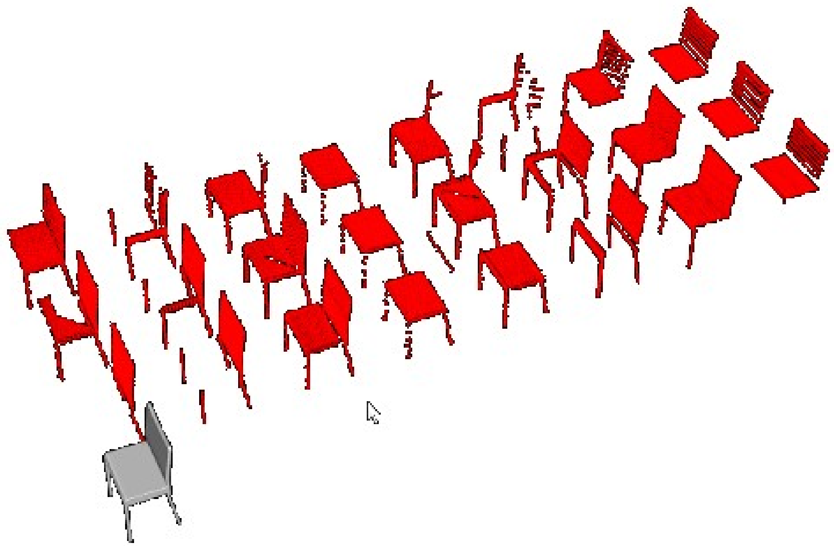


Offline training

Generate training views

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

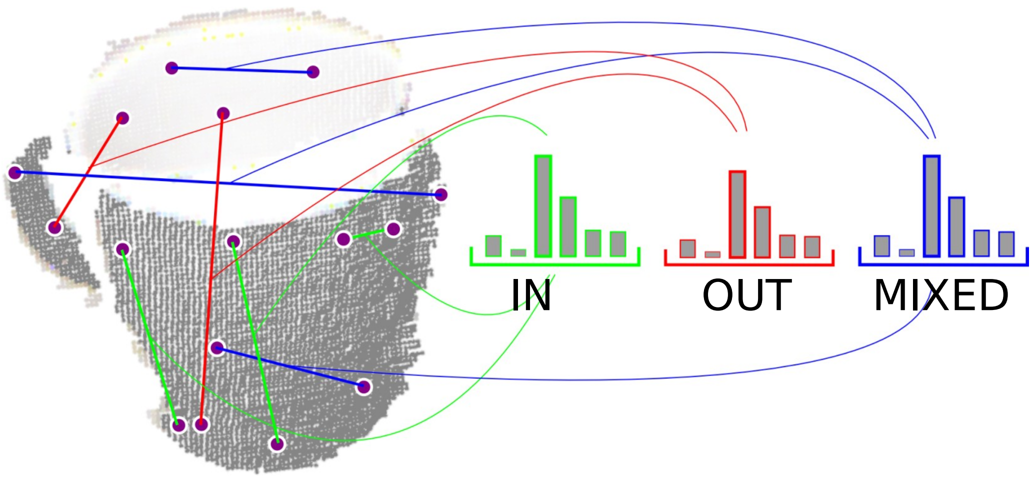


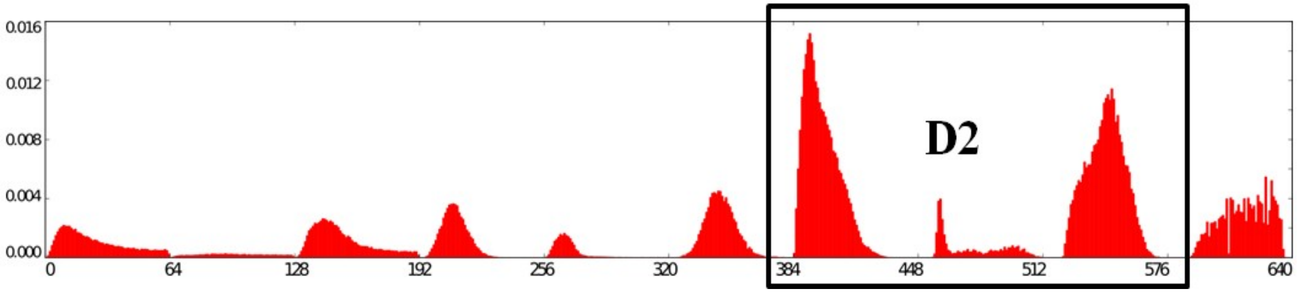


Offline training

Feature vector

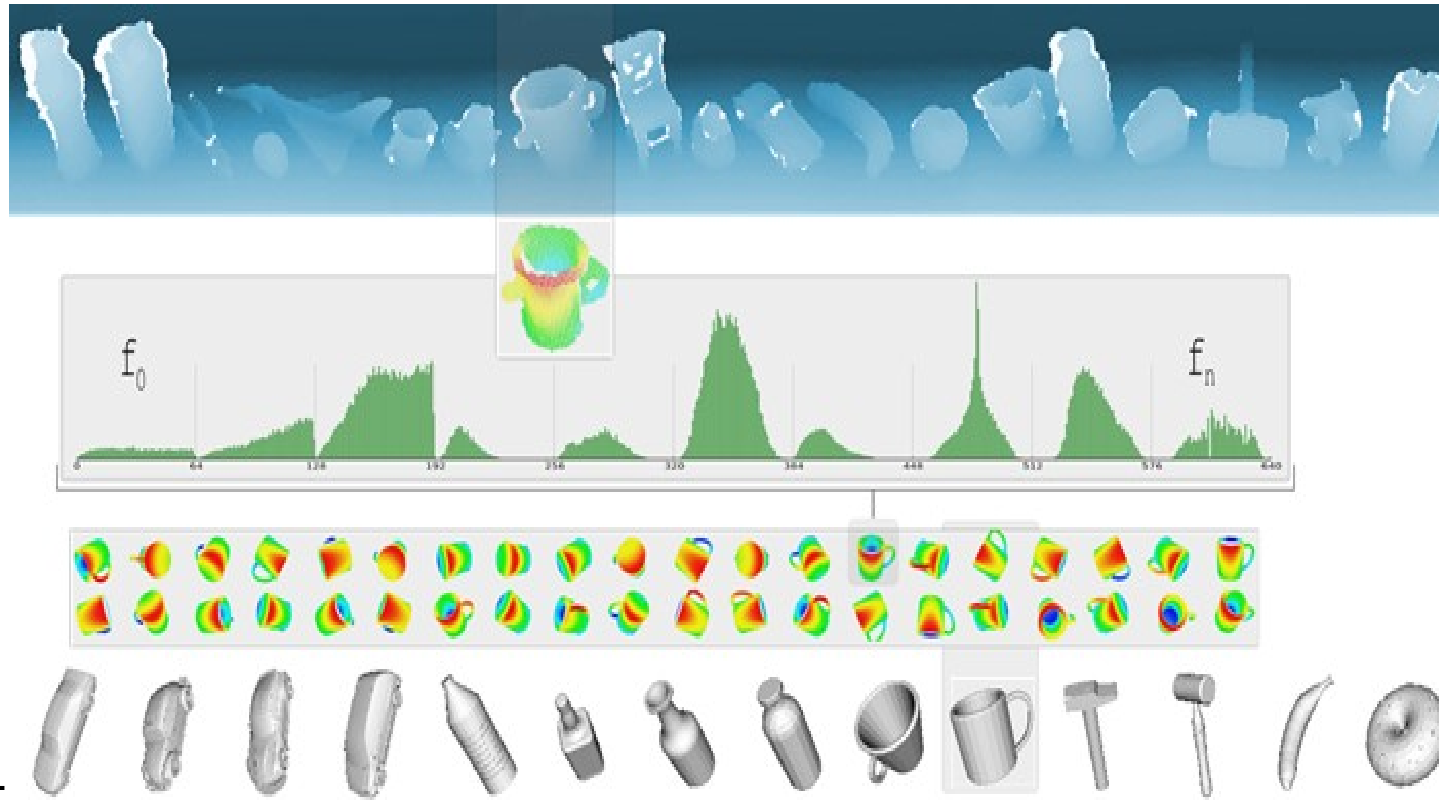
- Ensemble of shape functions (ESF)
- Based on shape distributions [Osada ea 2001]
  inside, outside, mixed
- Additional histograms for ratio, area, angle





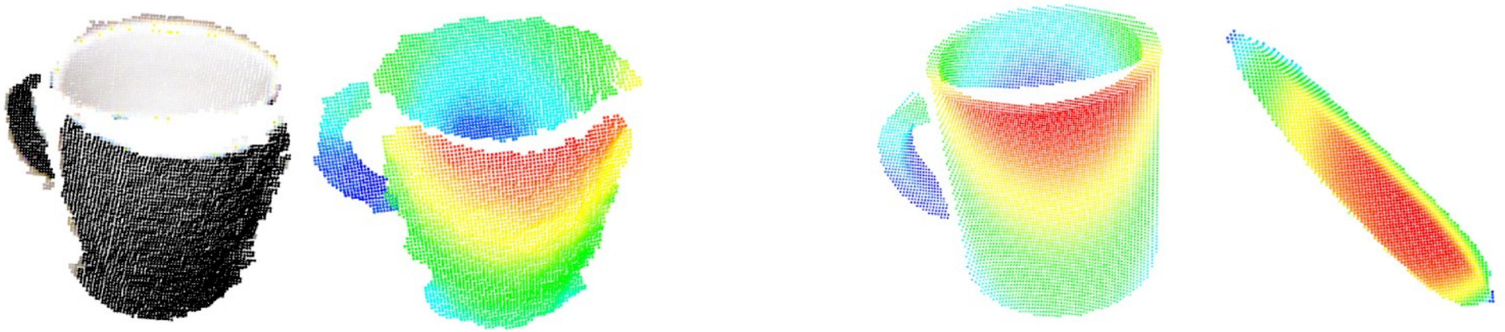
Online Object Categorisation

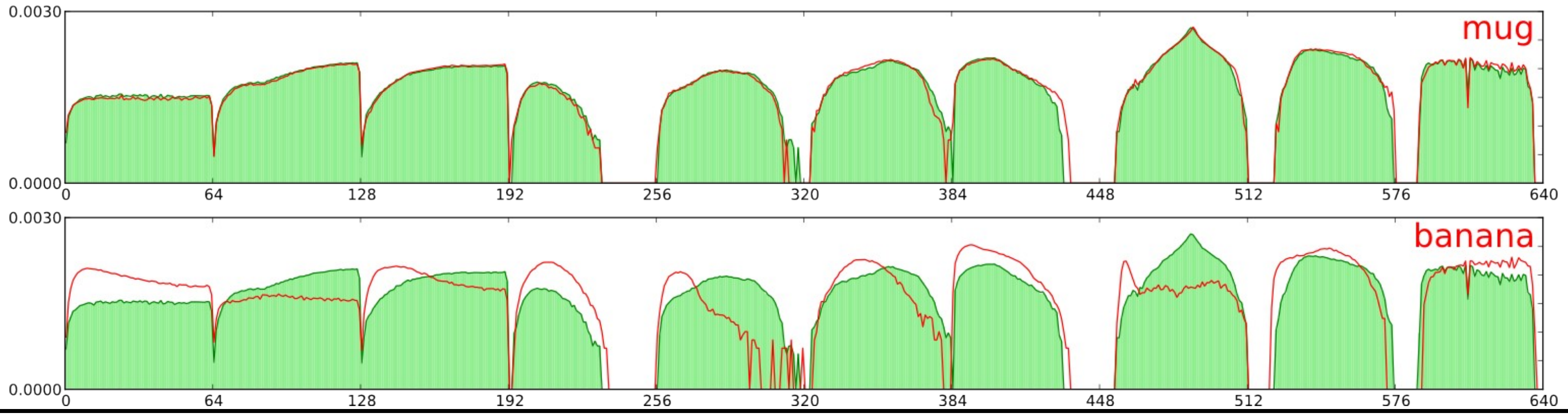
1) Point cloud
2) Segment objects
3) Feature vector
4) Find matching view
5) Verify with 3D model fit, pose estimation



Matching: kNN classifier

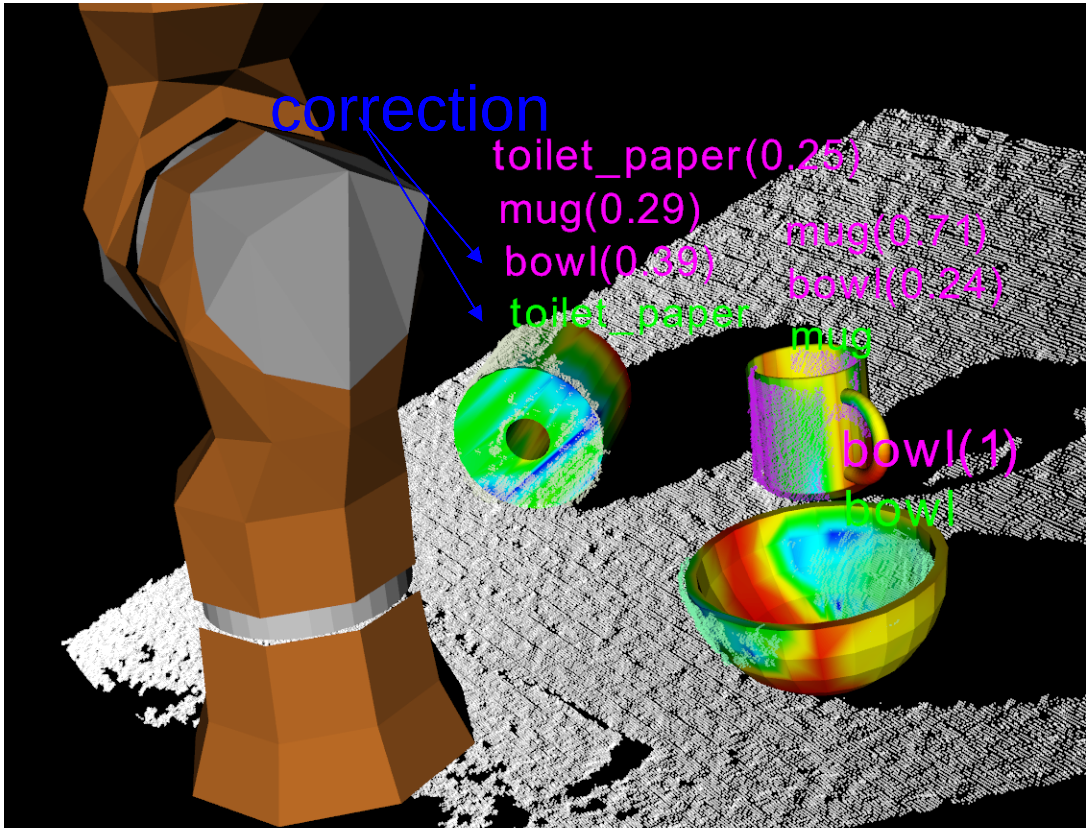
- Find nearest neighbour in feature space
- Efficient indexing techniques to cope with large database (100,000s views)
- Majority vote from k nearest neighbours





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

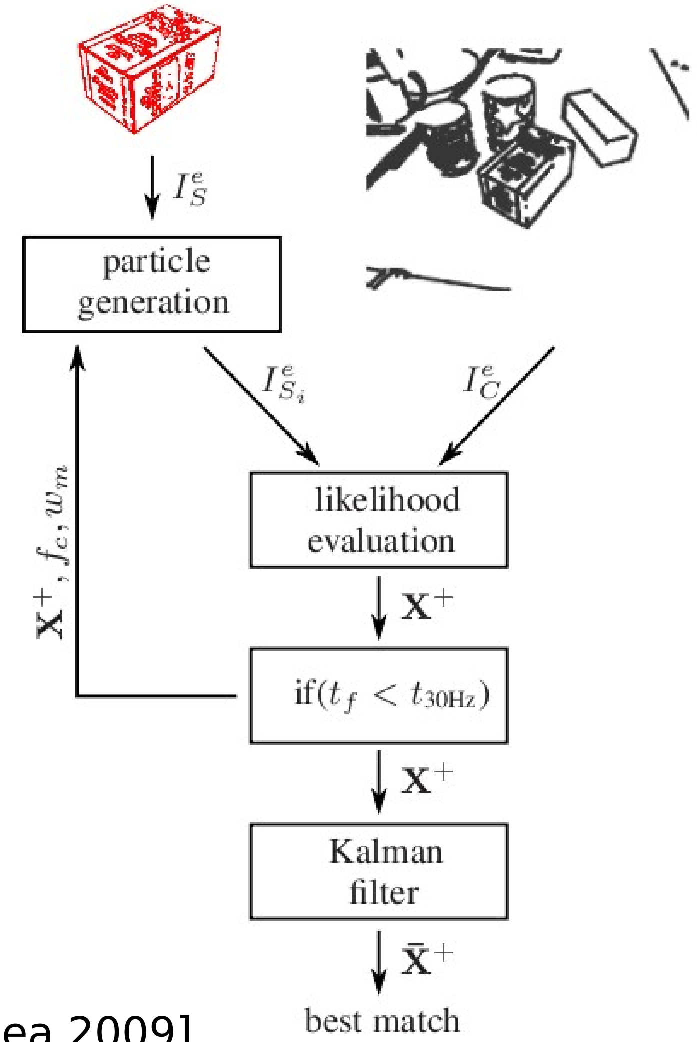
[Wohlkinger ea ICRA'12]

Overview

- Sensors
- Detection / segmentation
- Recognition
- Classification
- **Tracking**
- Attention

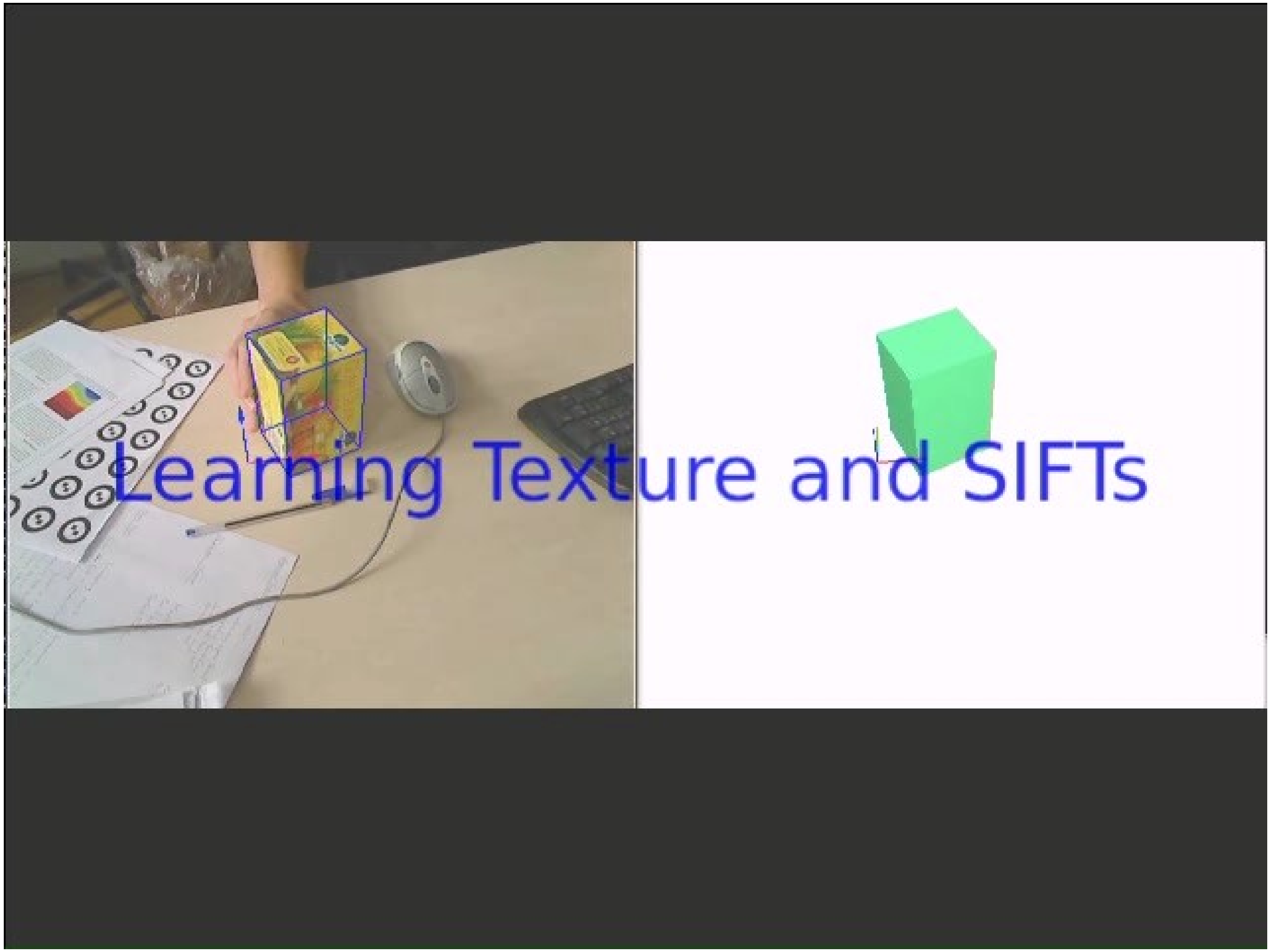
Object tracking: particle filter

- 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)
- Weak particles are discarded, good ones are cloned (plus noise)
- Repeat ..



[Mörwald ea 2009]

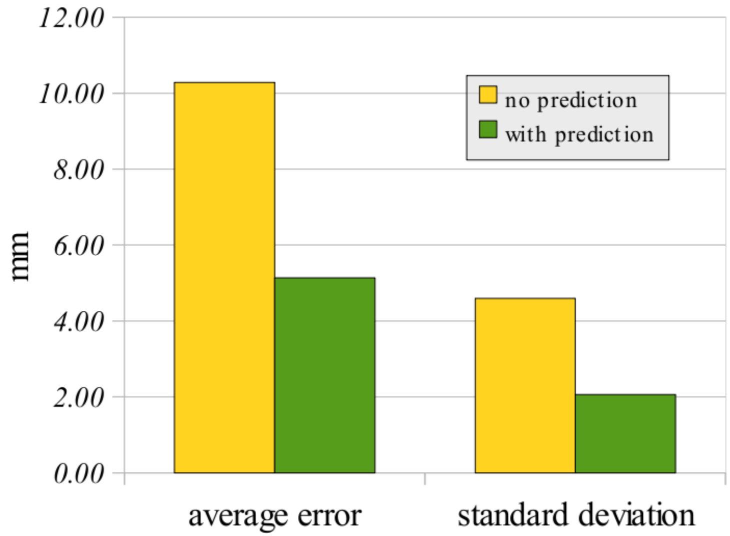
Object tracking



[Mörwald ea 2011]

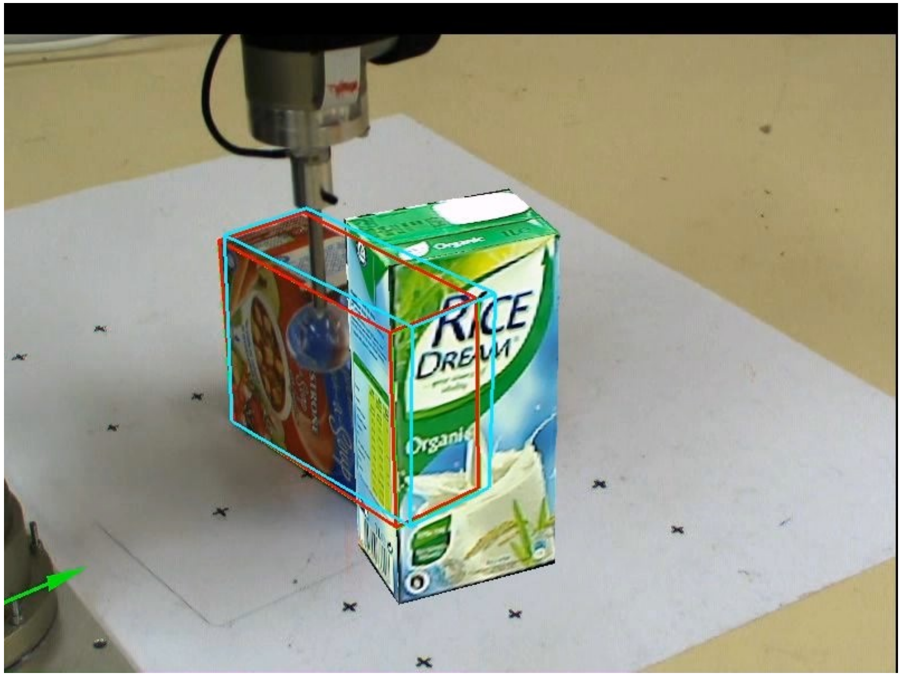
Object tracking with a physics model

- Replace simplistic motion model in particle filter with actual physics model
- Physics engines are difficult to parameterise => learn physics model
- KDE to learn predictive model of motion given a particular interaction [Kopicki ea ICAR'09] (Birmingham Univ.)



Improved accuracy ...

[Mörwald ea ICRA'11]



... and robustness

prediction, tracking, tracking + prediction

Overview

- Sensors
- Detection / segmentation
- Recognition
- Classification
- Tracking
- **Attention**

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?

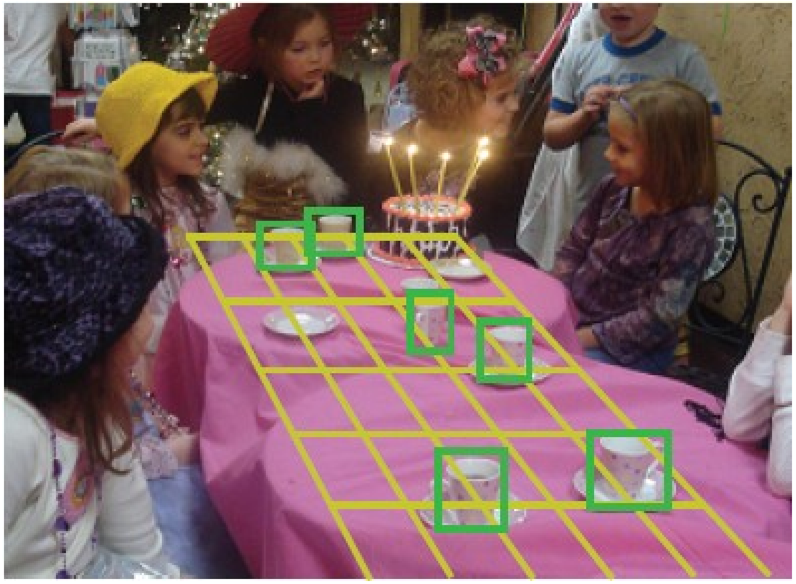
Visual attention

- 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

Scene context

- Detectors can produce many false positives
- But semantic/geometric information rejects false hypotheses

Mugs everywhere?



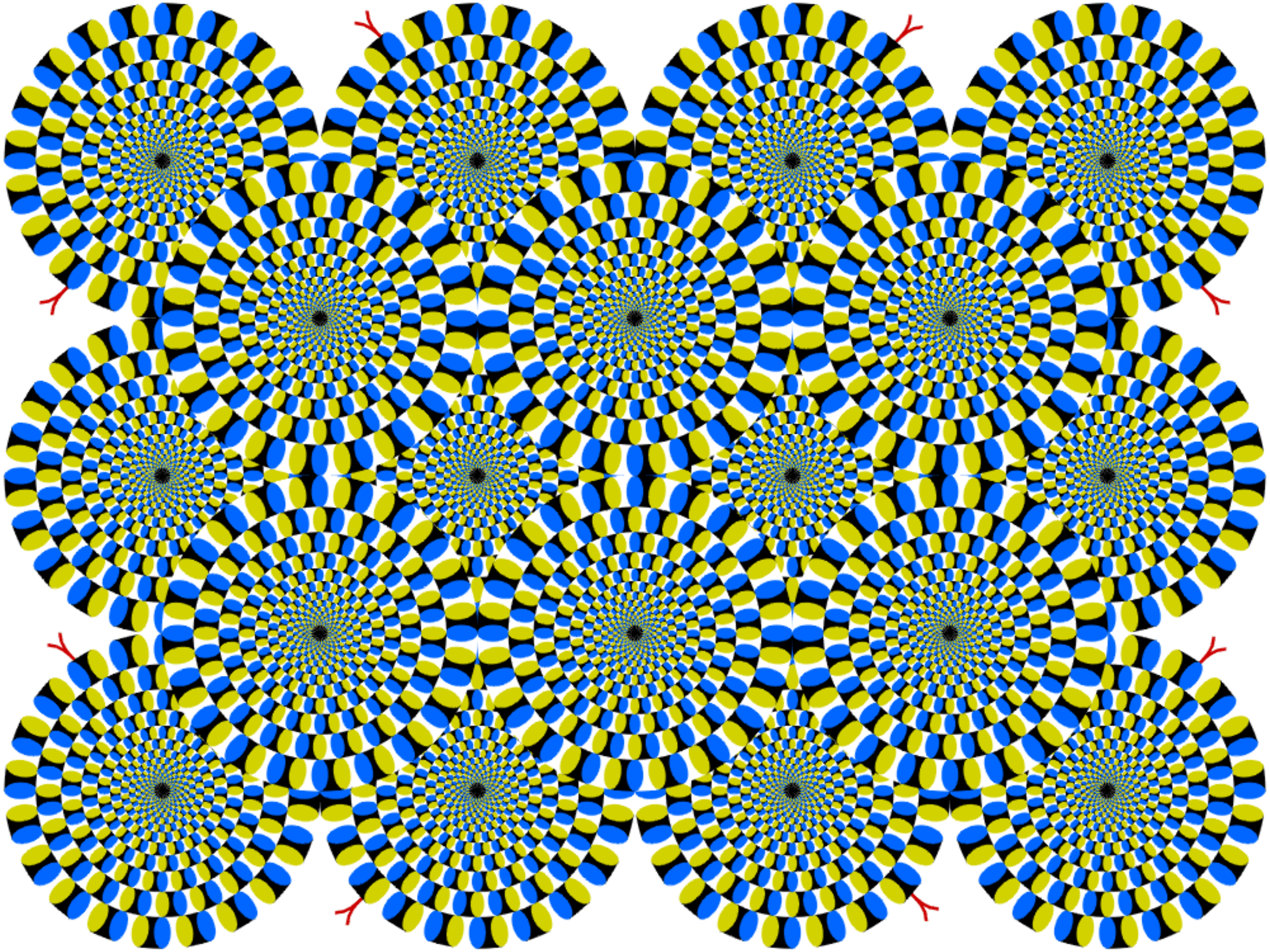
Mugs are on tables!

[Y.Z. Bao et al. 2010]

Recommended reading

- Parts of the lecture are based on:
  Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu Bogdan Rusu, Suat Gedikli, and Markus Vincze: *Point Cloud Library - Three-Dimensional Object Recognition and 6 DoF Pose Estimation*, Robotics and Automation Magazine, Sept. 2012
- PCL Tutorial, ICRA 2013:
  http://pointclouds.org/media/icra2013.html

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.