2D-3D Object Categorization for Task-based Grasping

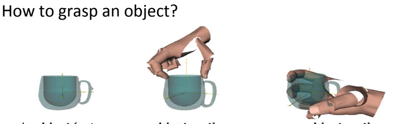
Marianna Madry Dan Song Danica Kragic

Computer Vision and Active Perception Lab, KTH, Stockholm, Sweden

{madry,dsong,danik}@csc.kth.se

Motivation

- How to grasp an object?



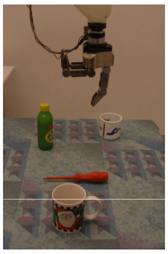
only **object** features: not sure | **object+action**: moving or hand-over | **object+action**: pouring or drinking

- Humans classify objects according to their functionality, i.e. depending on tasks they afford [Greeno'94]
- Grasp knowledge can be transferred between objects that belong to the same category
- Fusion of 2D and 3D may provide more robust system

Goal

- Where to grasp to *pour* a liquid or *hand over* an object?
- Goal: Finding grasping points for a desired task in a natural scene
- Task-based grasp depends on:
  I. Embodiment
  II. Scene content

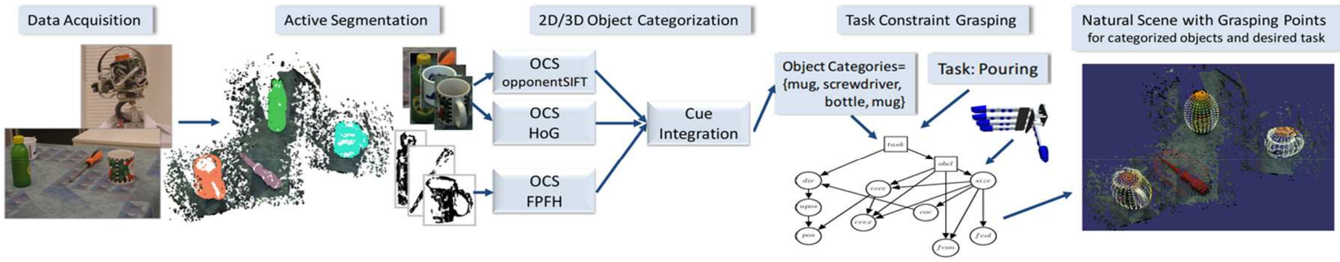
Mid-goal: Finding object categories



Contributions

- Object Categorization:
  - Evaluation of several 2D and 3D appearance, color and shape descriptors on real stereo data
  - Fusion of 2D and 3D object categorization with high categorization rate (up to 92% for 11 object categories)
- Task-based Grasping :
  - Robot can choose objects that afford a desired task
  - Robot can plan the grasp that satisfies the constraints posed by the task
- Integration of the 2D-3D Object Categorization System with the active segmentation module [*Björkman '10*] and the probabilistic grasp reasoning system [*Song '10*]

System Overview



Active Segmentation

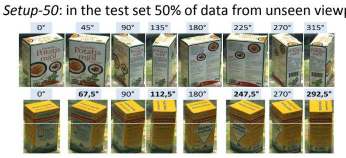
- Attention mechanisms in the peripheral view direct the foveal cameras towards region of interest

- Imperfect segmentation in ~10% of cases

110 Object Stereo Database

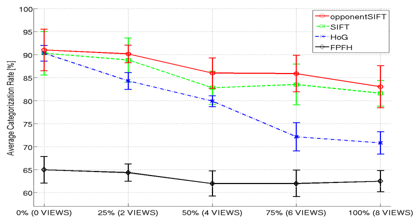
- 11 object categories x 10 object intances per category
- Data: 2D (RGB image) and 3D (point cloud)

- Data collected in 16 views around an object (every 22,5°)
- Experimental Setup: 8 views per object for traning and testing
  - *Setup-50*: in the test set 50% of data from unseen viewpoint



Object Representation

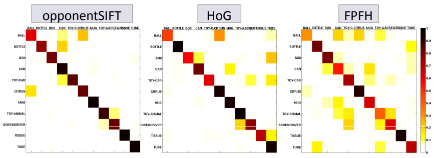
- Evaluation of several 2D and 3D descriptors encoding different object properties: appearance (SIFT), color (opponentSIFT), 2D shape (HoG), 3D shape (FPFH)
- We build a single-cue OCS for each descriptor:
  - Spatial pyramid for 2D; Bag-of-words for 3D
  - Classification: SVMs with a χ^2 kernel
- Results:
  - Performance under varying viewpoint condition



- Setup-50:

Descriptor	Av. Categ.Rate	σ
SIFT	82.8%	3.6%
opponentSIFT	**86.0%**	3.3%
HoG	79.9%	1.2%
FPFH	62.0%	2.8%

- Confusion matrices – complementarity for cue integration:



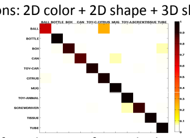
2D-3D Object Categorization

- Confidance Measure:
  - normalized distance of a sample to the hyperplane for OaA
- Cue Integration:
  - Fusion of evidences from the single-cue OCSs at the high level
  - Evaluation of the linear and nonlinear integration methods:

Descriptors D1+D2(+D3)	Max Rule Average Categ.Rate	Max Rule σ	Max Rule Av. Gain D1,D2,(D3)	Product Rule Average Categ.Rate	Product Rule σ	Product Rule Av. Gain D1,D2,(D3)	Sum Rule Average Categ.Rate	Sum Rule σ	Sum Rule Av. Gain D1,D2,(D3)
opponentSIFT+HoG+FPFH	84.5%	5.8%	-1.5,4.6,22.4%	89.9%	4.9%	3.9,10.0,27.9%	**92.0%**	2.8%	**6.0**,12.0,29.9%
opponentSIFT+FPFH	86.6%	3.7%	0.6%, 24.5%	87.8%	4.3%	1.9%, 25.9%	90.9%	3.2%	4.9%, 28.9%
opponentSIFT+HoG	81.9%	3.4%	-4.1%, 2.0%	86.0%	0.8%	0%, 6.1%	87.4%	0.6%	1.4%, 7.5%
HoG+FPFH	79.9%	1.5%	0.0%, 17.9%	83.1%	6.3%	3.1%, 21.0%	83.4%	4.6%	3.5%, 21.4%

Descriptors D1+D2(+D3)	RBF Kernel Average Categ.Rate	RBF Kernel σ	RBF Kernel Av. Gain D1,D2,(D3)	χ^2 Kernel Average Categ.Rate	χ^2 Kernel σ	χ^2 Kernel Av. Gain D1,D2,(D3)	Hist. Intersection Kernel Average Categ.Rate	Hist. Intersection Kernel σ	Hist. Intersection Kernel Av. Gain D1,D2,(D3)
opponentSIFT+HoG+FPFH	87.6%	6.5%	1.6,7.7,25.6%	90.4%	0.3%	4.4,10.5,28.4%	89.8%	1.8%	3.8,9.9,27.8%
opponentSIFT+FPFH	87.2%	2.0%	1.2%, 25.2%	87.6%	1.6%	1.6%, 25.6%	87.2%	2.8%	1.2%, 25.2%
opponentSIFT+HoG	84.3%	5.4%	-1.7%, 4.3%	84.8%	5.3%	-1.1%, 4.9%	84.7%	3.1%	-1.3%, 4.7%
HoG+FPFH	79.1%	1.8%	-0.8%, 17.0%	79.5%	3.0%	-0.4%, 17.5%	80.4%	1.2%	0.5%, 18.4%

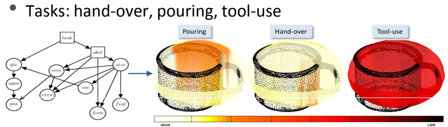
- Results:
  - Best combinations: 2D color + 2D shape + 3D shape descriptors



  - 2D-3D OCS significantly outperforms the best single-cue OCSs
  - Linear weighted summation is better than the complex methods

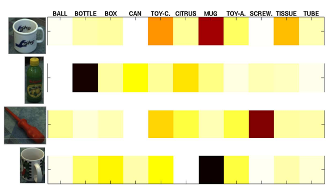
Task-constrained Bayesian Network

- Training on the synthetic 3D object models from the Princton Shape Benchmark
- Tasks: hand-over, pouring, tool-use



Where to grasp to perform the desired task in the real scene?

Categorization Confidence



Hand-over | Pouring | Tool-use