From scene segmentation to task based grasping

Danica Kragic

Centre for Autonomous Systems
Computer Vision and Active Perception Lab
School of Computer Science and Communication
KTH – Royal Institute of Technology
Stockholm, Sweden

Scope

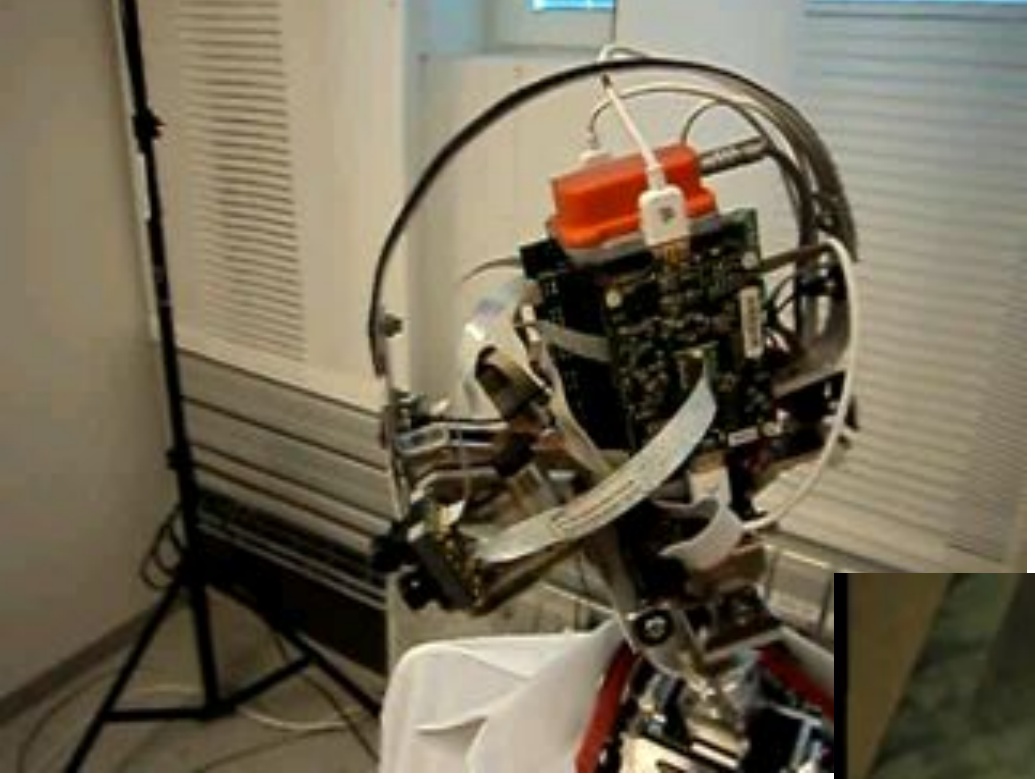
- ”Robot, pick up this”
- ”Robot, bring me a can of beans“
- “Robot, fetch me something to drink from”
- “Robot, fetch me something to drink“
- ”Robot, clear the table”

Problems we work on

- Scene segmentation
- Interactive perception
- Object categorization
- Grasp synthesis and grasp transfer
- Grasp stability assessment
- Object manipulation

Grasping in realistic scenes

- Noisy data: What is an object?
- Many grasping hypotheses: Which one to choose?
- Many parameters: friction, mass, scene, embodiment, task
- Learning from experience: What to represent?



An active vision system for detecting, fixating and manipulating objects in real world
B. Rasolzadeh, M. Björkman, K. Huebner, D. Kragic,
International Journal of Robotics Research, 2010

Active 3D Segmentation through Fixation of Previously Unseen Objects,
M. Björkman, D. Kragic
British Machine Vision Conference, 2010

Modeling

$\boldsymbol{m}$ Measurements (positions, disparities and colors)

$\boldsymbol{l}$ Labelling (foreground, background or flat surface)

$\theta = \theta_f \cup \theta_s \cup \theta_b$ Model parameters

Joint probability given by:

$$p(\mathbf{m}, \mathbf{l}|\theta) = p(\mathbf{m}|\mathbf{l}, \theta)p(\mathbf{l}|\theta)$$

Measurement distribution given by:

$$p(\mathbf{m}|\mathbf{l}, \theta) = \prod_i p(m_i|\theta_f)^{I_i^f} p(m_i|\theta_b)^{I_i^b} p(m_i|\theta_s)^{I_i^s}$$

Prior label probabilities:

$$p(\mathbf{l}|\theta) = \prod_k p(l_k) \prod_i \prod_{j \in N_i} p(l_i, l_j).$$

Modeling

"Active 3D scene segmentation and detection of unknown objects", (M. Bjorkman, D. Kragic), ICRA 2010

- For all scene parts we model the distributions of image point positions, disparities and colors.

- The set of the model parameters is:

$$\theta_f = \{p_f, \Delta_f, c_f\},$$
$$\theta_b = \{d_b, \Delta_b, c_b\},$$
$$\theta_s = \{d_s, \Delta_s, c_s\},$$

- The measurement conditionals are:

$$p(m_i|\theta_f) = n(p_i; p_f, \Delta_f)\, H_f(h_i, s_i),$$
$$p(m_i|\theta_b) = N^{-1} n(d_i; d_b, \Delta_b)\, H_b(h_i, s_i),$$
$$p(m_i|\theta_s) = N^{-1} n(d_i; \alpha_s x_i + \beta_s y_i + \delta_s, \Delta_s)\, H_s(h_i, s_i).$$

- Estimating model parameters:

$$p(\mathbf{m}|\theta) = \sum_{\mathbf{l}} p(\mathbf{m}, \mathbf{l}|\theta).$$

Approximate Expectation-Maximization

Instead of looping over all labellings, loop over all labels assuming they are independent:

$$Q_2(\theta|\theta') = \sum_i \sum_{l_i \in L} P(\boldsymbol{l}_i|\boldsymbol{m}_i, \theta') \log P(\boldsymbol{m}_i, \boldsymbol{l}_i|\theta)$$

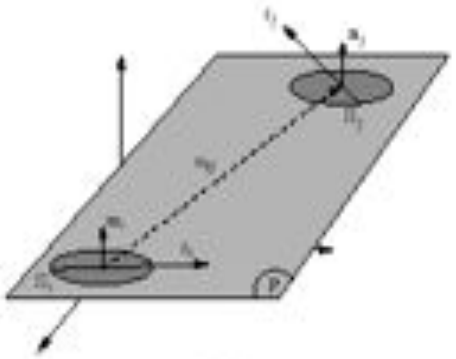
Two steps (similar to Expectation-Maximization):

1) Compute the marginals per pixel (segmentation) with Belief Propagation.

$$P(\boldsymbol{l}_i|\boldsymbol{m}_i, \theta')$$

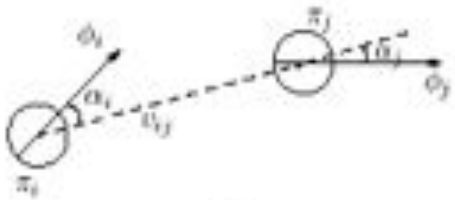
2) Update model parameters, maximizing $Q_2(\theta|\theta')$

Grasping Reflex and Second Order Relations



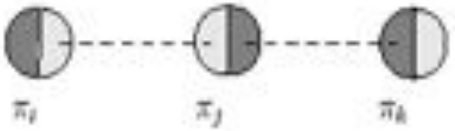
(a)

Co-planarity



(b)

Co-linearity



(c)

Co-colority

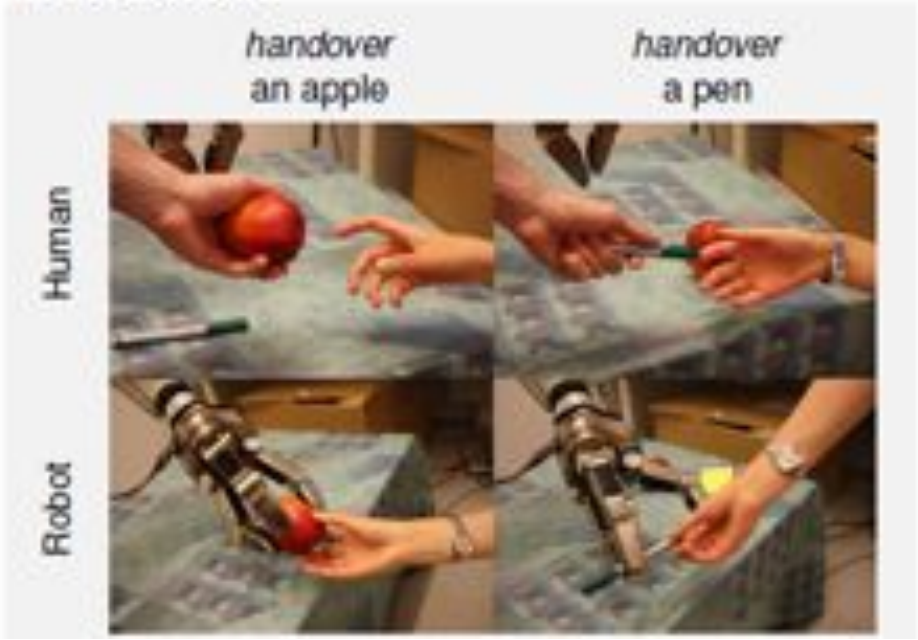
Enabling grasping of unknown objects through a synergistic use of edge and surface information , G. Kootstra, M. Popovic, J.A. Jorgensen, K. Kuklinski, K. Miatliuk, D. Kragic, N. Kruger, Int Journal of Robotics Research, 31(10), pp.1190-1213, 2012

Grasping Unknown Objects using an Early Cognitive Vision System for General Scene Understanding

Mila Popović, Gert Kootstra, Jimmy Alison Jørgensen, Danica Kragic, Norbert Krüger

What is a good grasp?

If the task is:



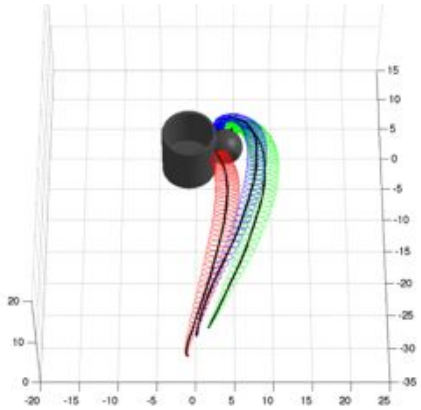
Problems:

1) How does a post-grasp action (or task) constrain which object to use, and how to grasp it? Greeno (1994)

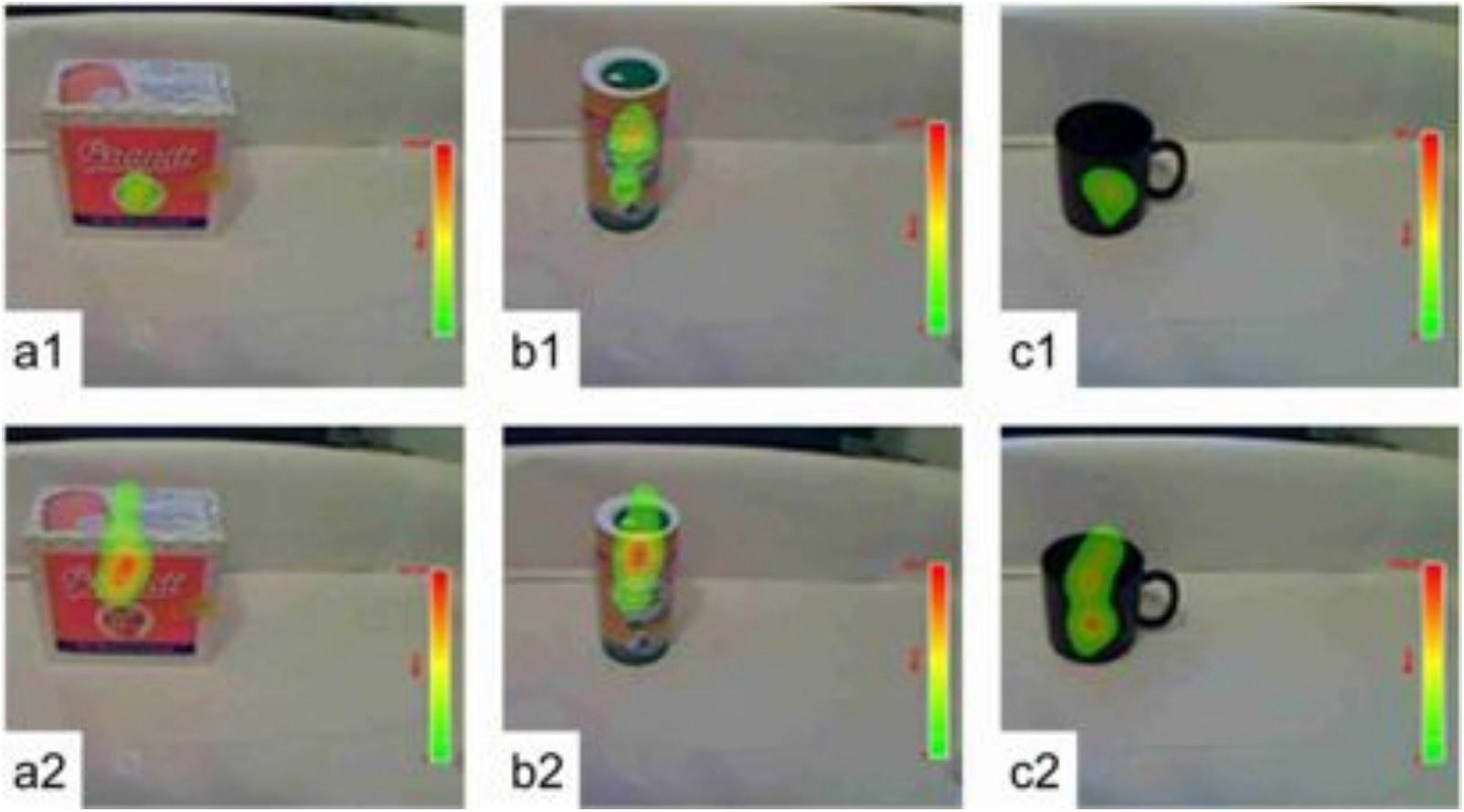
2) How to transfer this task knowledge across different embodiments?
Alissandrakis *et al.* (2007)

3) What can be the role of human teacher in task constraint learning?
Calinon and Billard (2007)

Eye fixations during grasping



H. Deubel et al: www.grasp-project.eu



Grasping point detection

- How to go from 2D to 3D?
- How to cope with uncertainty?
- Where is semantics?

Learning Grasping Points with Shape Context, J. Bohg, D. Kragic, Robotics and Autonomous Systems, 2009.

Task Based Grasping

- A semi-supervised method for encoding task-related grasps. Mixed BN encodes relation between tasks, objects and actions.
- The approach used to solve:
  - Action recognition
  - Task based planning
  - Grasp stability assessment

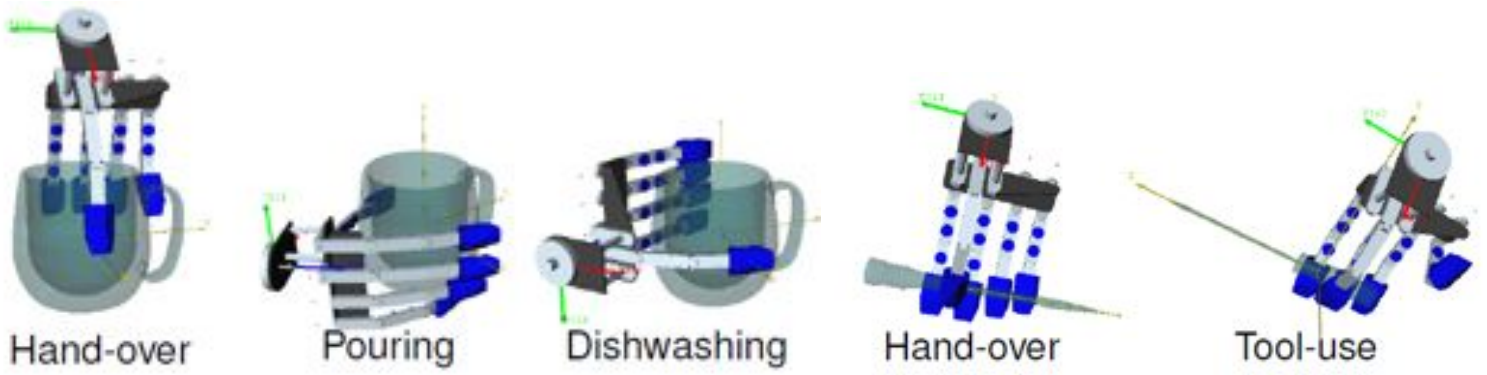
  in an integrated framework

Embodiment-Specific Representation of Robot Grasping using Graphical Models and Latent-Space Discretization, D. Song, CH Ek, K. Huebner, D. Kragic, *IEEE IROS* 2011

From Object Categories to Grasp Transfer Using Probabilistic Reasoning, M. Madry, D. Song, D. Kragic, *IEEE ICRA* 2012

Task Based Grasping

How to grasp object depends on goal/task to achieve.

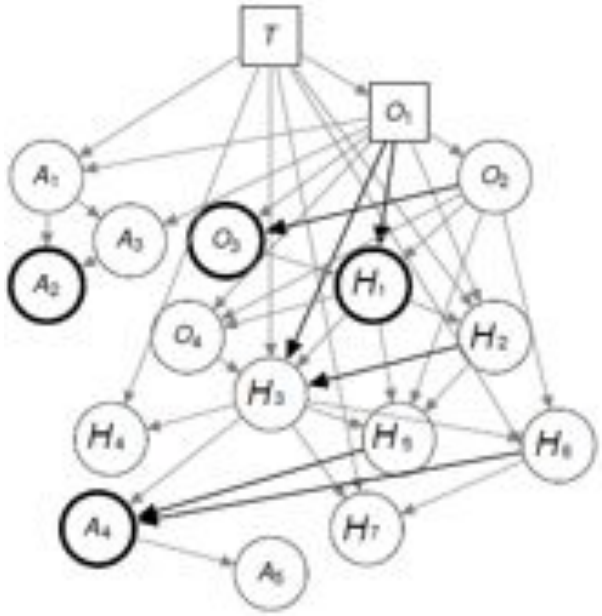


- Method combining exploration and supervision
  - Exploration enables the robot to learn about own SM abilities (how to grasp object to lift and manipulate),
  - Human tutoring helps the robot to associate its SM abilities to high-level goals.

Bayesian Networks

- Bayesian Networks (BNs) to integrate discrete semantic task information with the continuous representation of sensory data.
- Grasp space is composed of a set of sensory features *X* relevant for grasping tasks *T*.
- *X* originates from
  - *O - object feature set* from simulated visual sensing,
  - *A - action feature set* (gripper configurations) from proprioception and
  - *H - haptic feature set*.

Bayesian Networks



$$P(\mathbf{Y}) = P(\mathbf{Y}|\theta, S) = \prod_{i=1}^{N} P(Y_i|\mathbf{pa}_i, \theta_i, S)$$

$$\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_N\} \leftarrow \{O, A, H, T\}$$

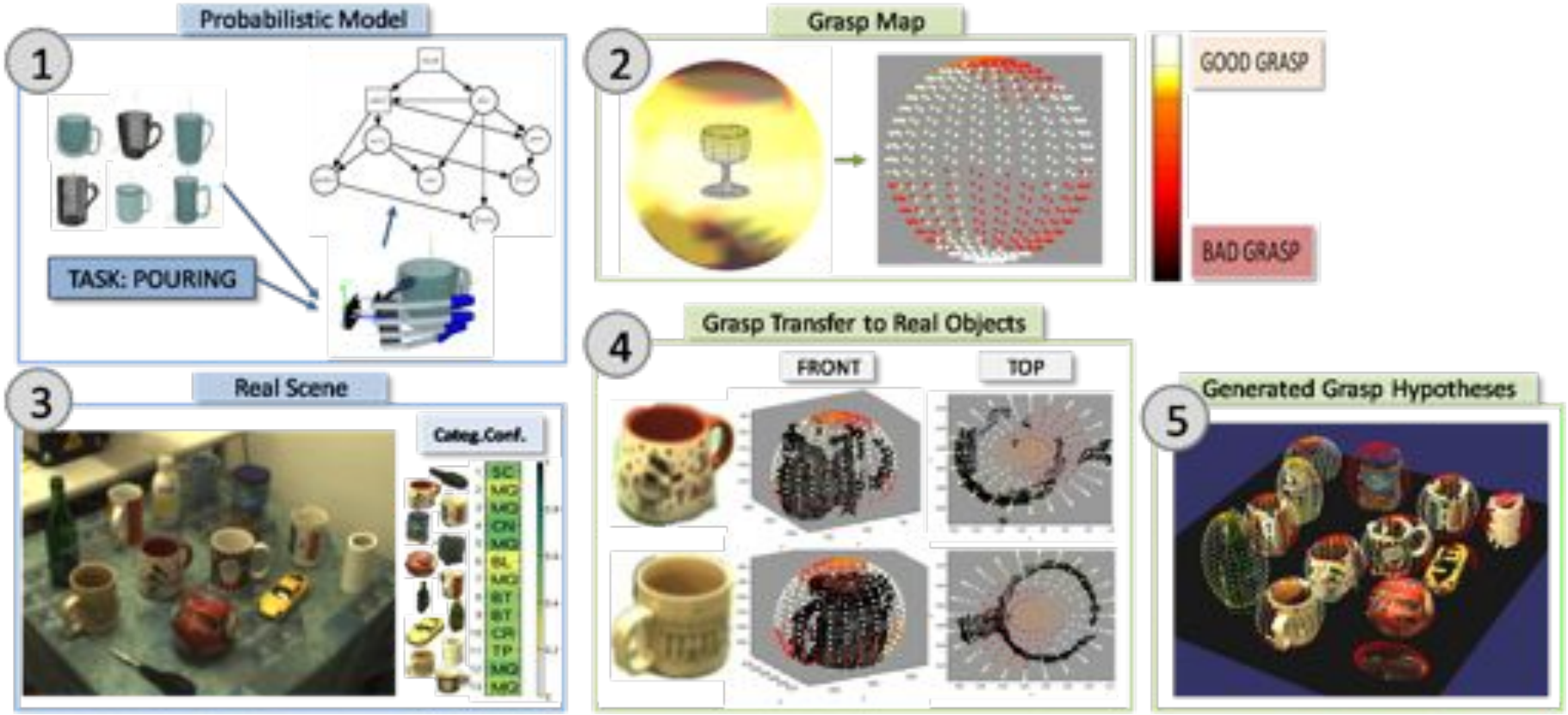
- Once trained, we can infer $P(Y_i|others)$.

- We can use a BN to
  - predict success of a grasp to achieve a task given observed object and action features by inferring ***P*(*T*|*O*,*A*)**.
  - given an assigned task find, the distribution of the object ***P*(*O*|*T*)** and/or grasp action features ***P*(*A*|*T*,*O*)**.
- Thus a robot can select objects that afford a given task and plan an optimal grasp strategy.

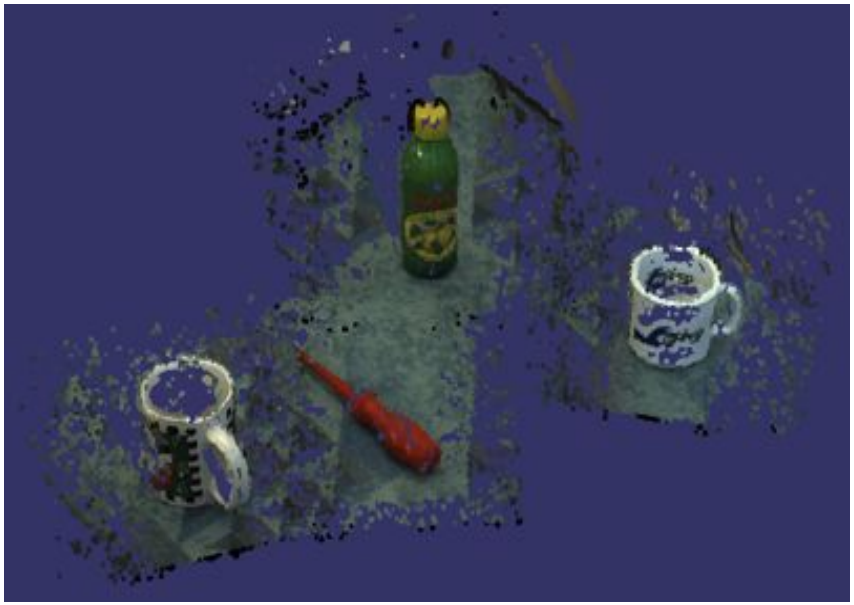
“Robot, bring me something to drink!”

ACTION OBJECT TASK

- Representation links information about: action + object + task
  to **transfer task-specific grasps from a known to a novel object**



Grasp hypotheses maps for scenes

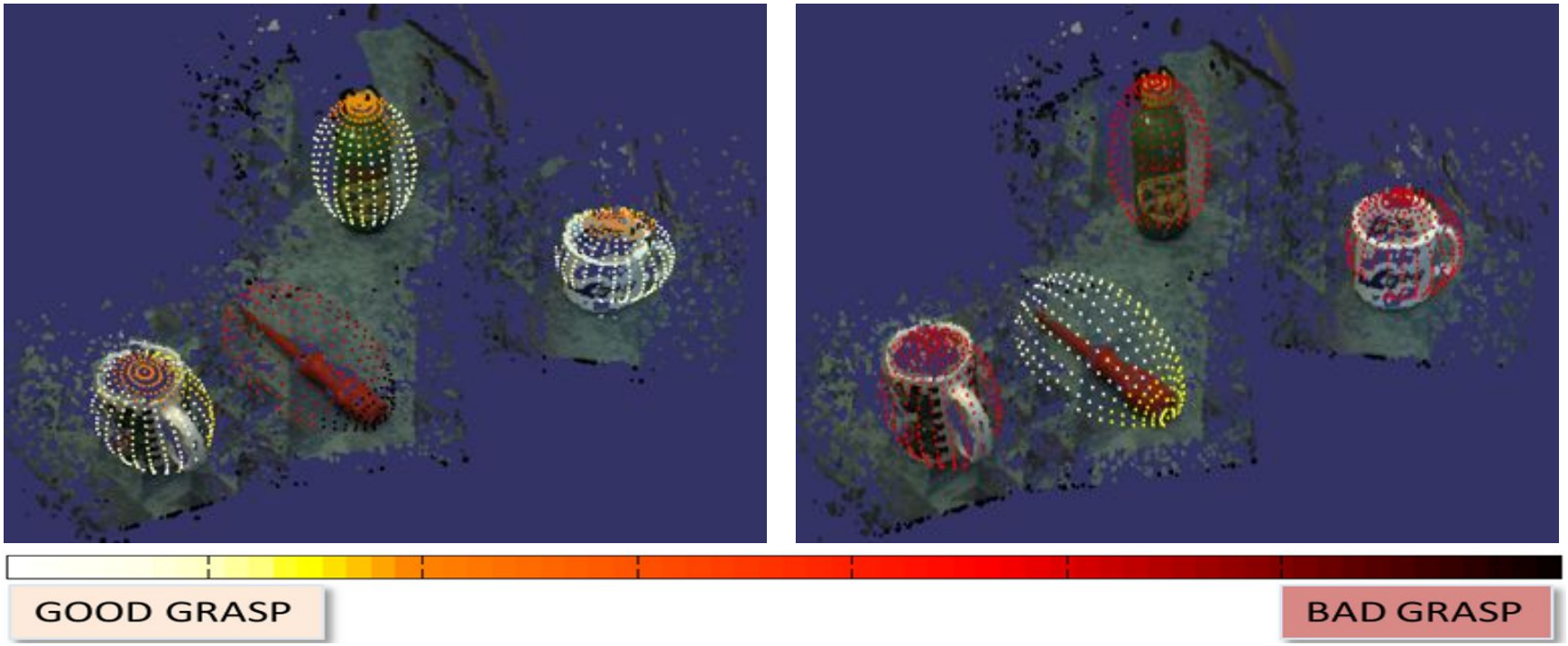


- 5 tasks:
  - hand-over, pouring, tool-use, playing, dishwashing
- 14 object categories

Grasp hypotheses maps for scenes

POURING

TOOL-USE

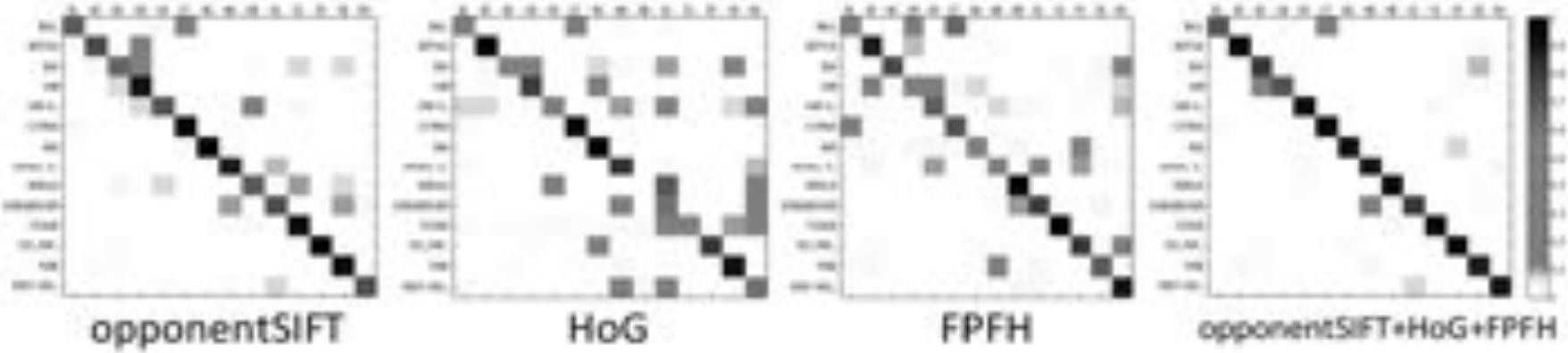


- Selecting objects that afford a TASK (categorization)
- Planning grasps that satisfy the constraints posed by the
- Grasp knowledge can be transferred from a known to a new object

- Challenges:

- Integration of different:
  - object properties (color, texture, contour, 3D shape)
  - modalities: 2D and 3D

  the best trade-off between discrimination and generalization



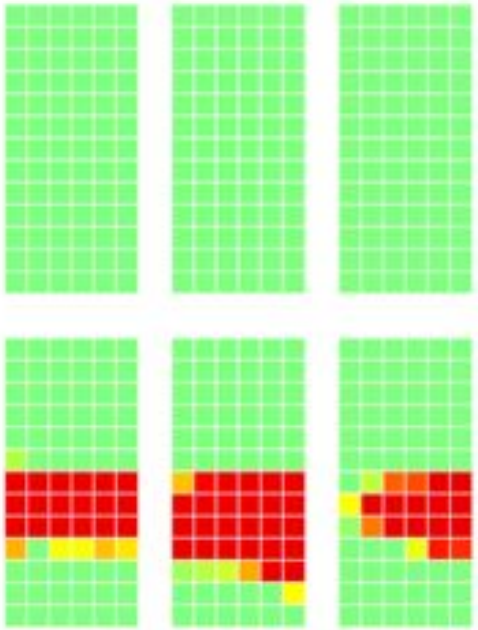
M. Madry, C.H. Ek, R. Detry, K. Hang, D. Kragic. Improving Generalization for 3D Object Categorization with Global Structure Histograms. In IEEE IROS 2012

A stable grasp

P(stable|vision) = 0.72

P(stable|touch) = 0.99

P(stable|vision, touch) = 0.97

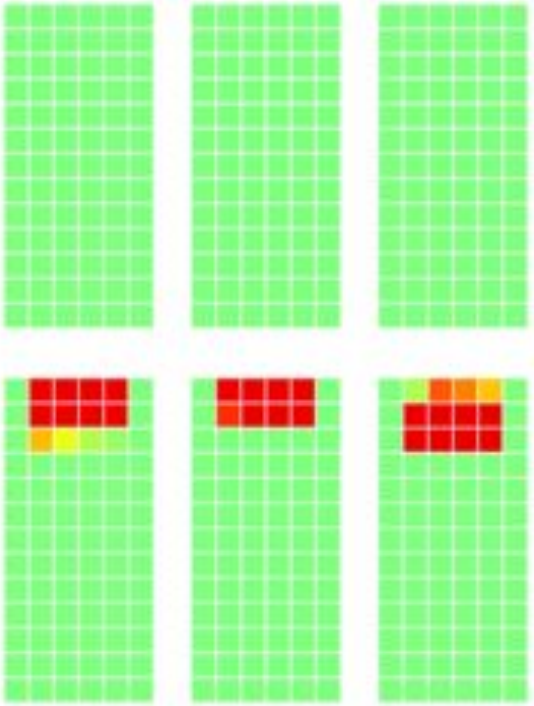


An unstable grasp

P(stable|vision) = 0.22

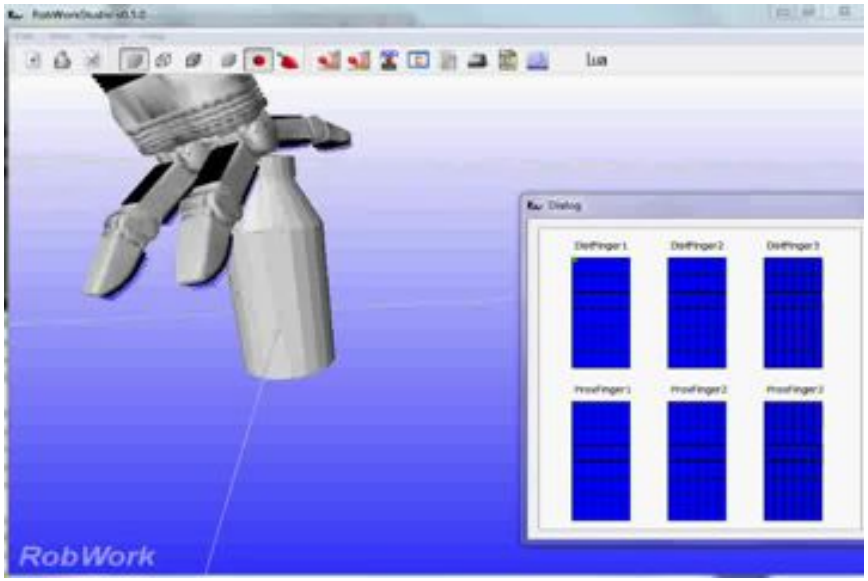
P(stable|touch) = 0

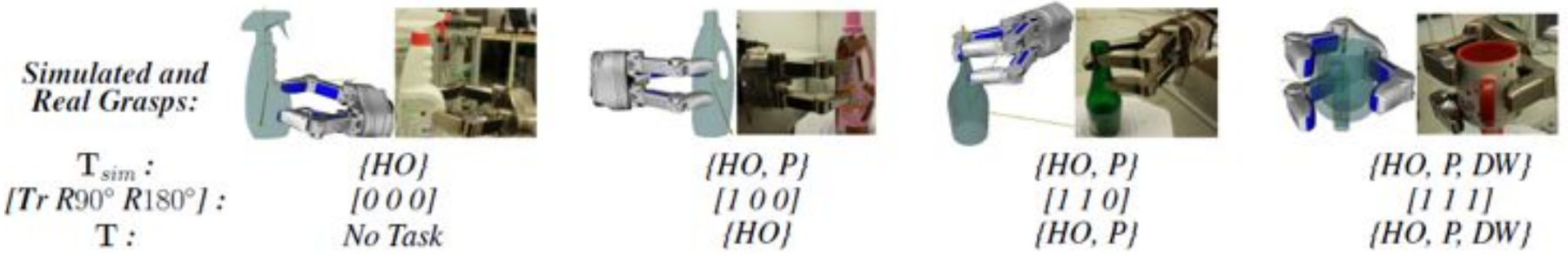
P(stable|vision, touch) = 0.02

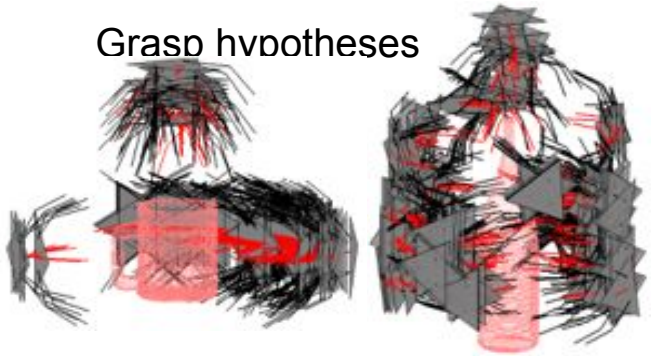


Data acquisition

- Grasps are planned in simulation and labeled manually for each task.
- Task labels (T) are refined by applying the required manipulations.
- Features (O,A,H) are extracted from simulation **and** real executions.



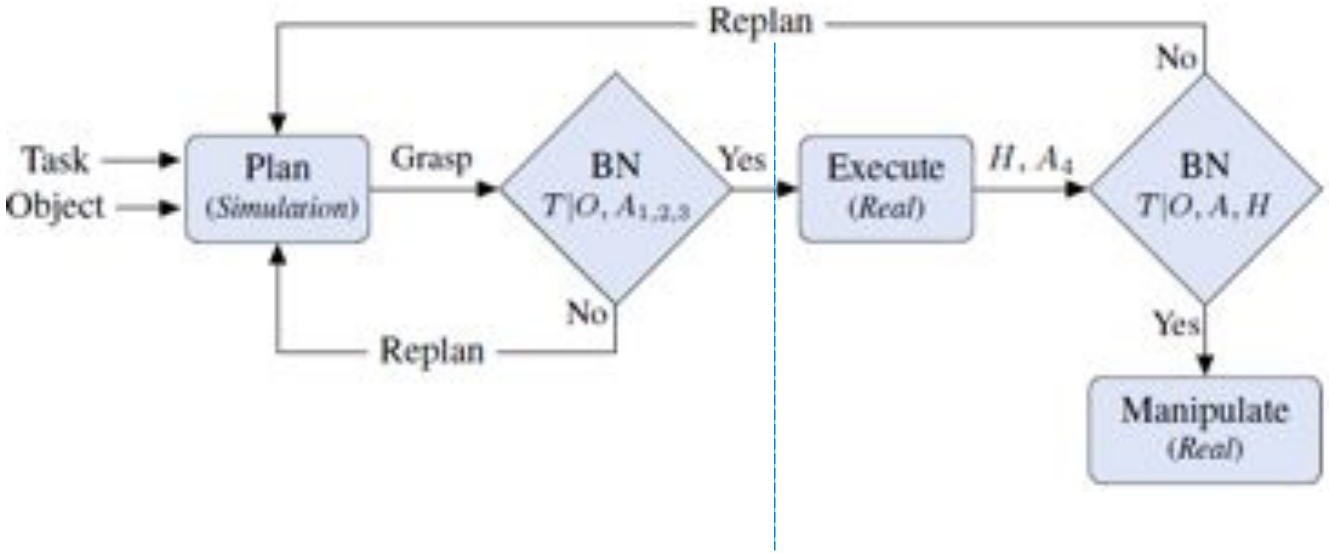




Objects used for training

Task based grasp adaptation

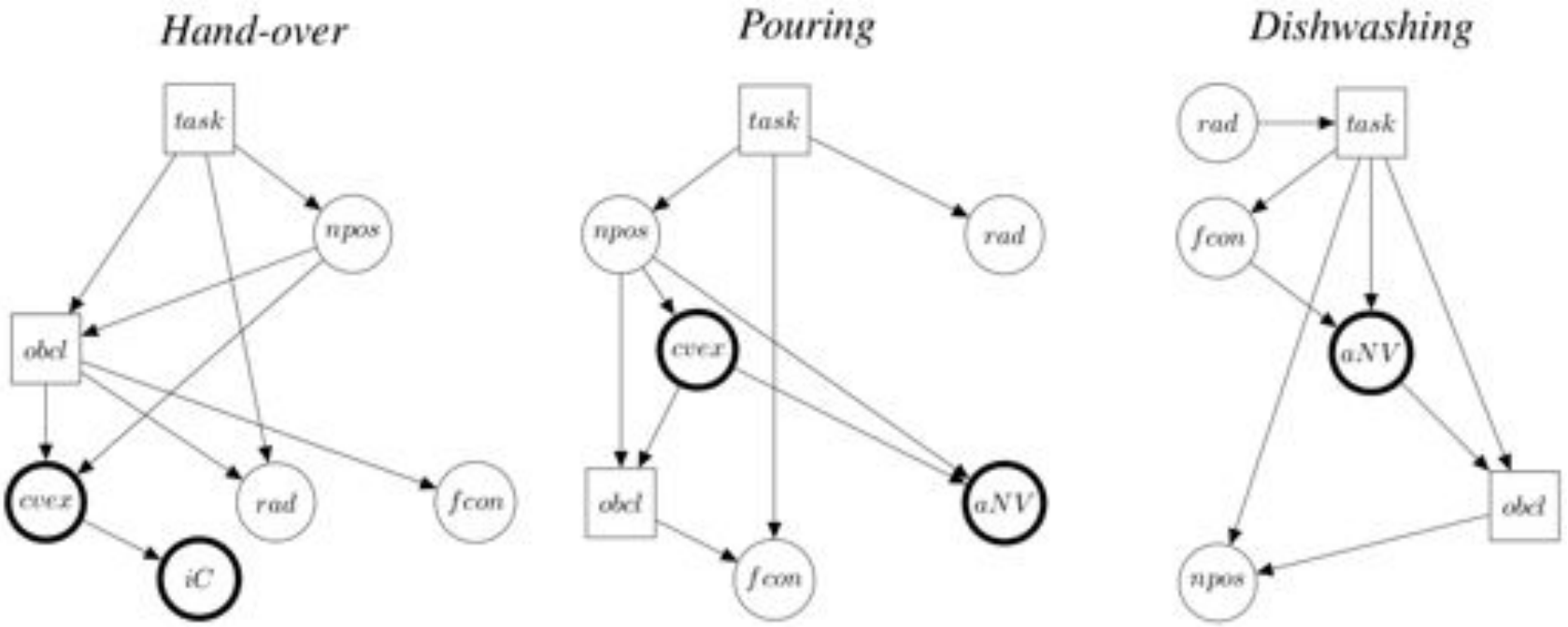
- Grasp success inferred from partial observations and training data
- We can initiate grasp replanning before a grasp is executed



- The first step predicts if a grasp hypothesis affords an assigned task
- The second step predicts if the grasp affords manipulation required for the task once the grasp has been executed on a robot

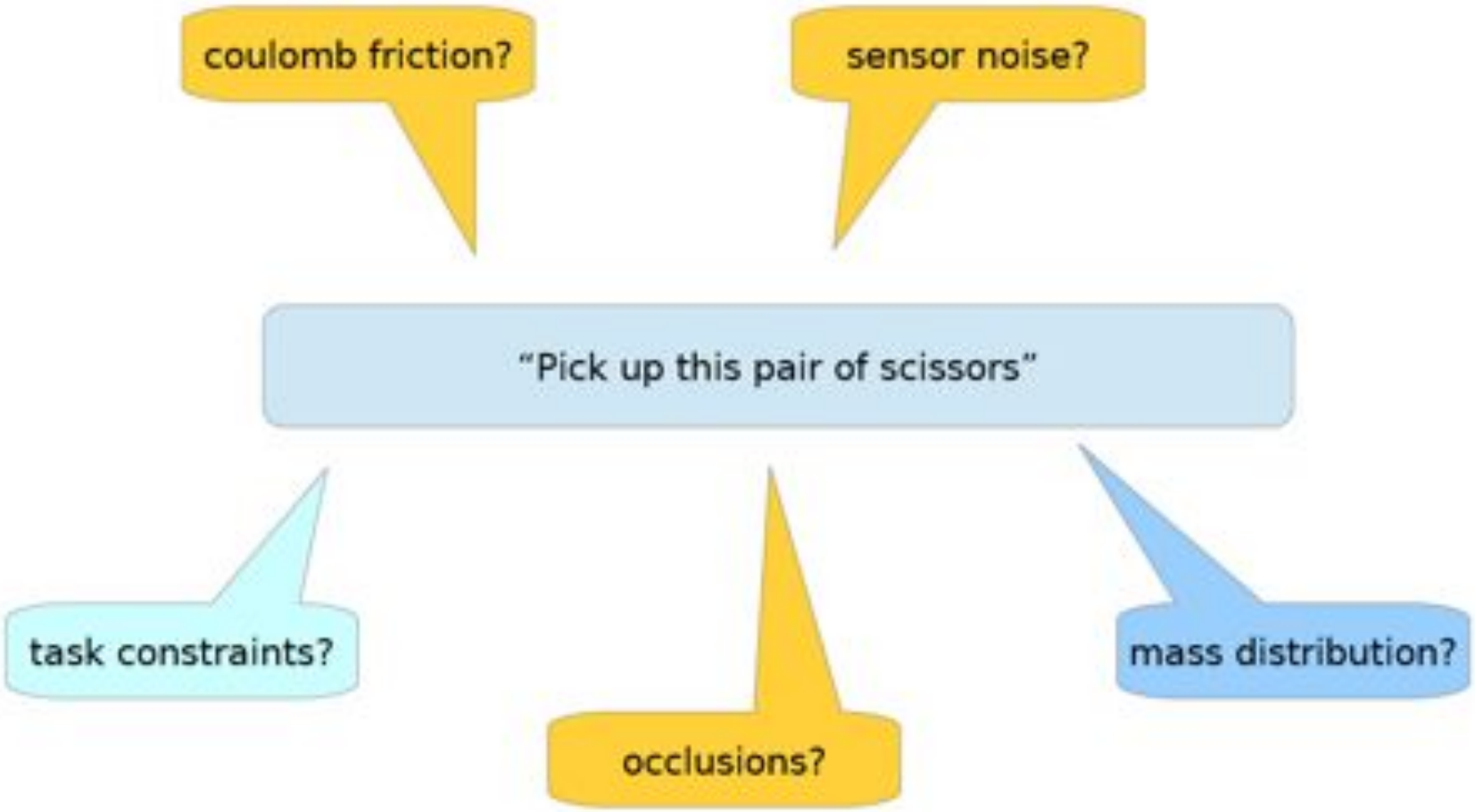
Learning network structure

- Dimensionality reduction, variable selection and discretization applied.
- Network structure learned from the data.



A probabilistic framework for task-oriented grasp stability assessment, Y. Bekiroglu, D. Song, L. Wang, D. Kragic, *IEEE ICRA,* 2013.

A classical problem:



“Local” methods

surface contacts?

noise?

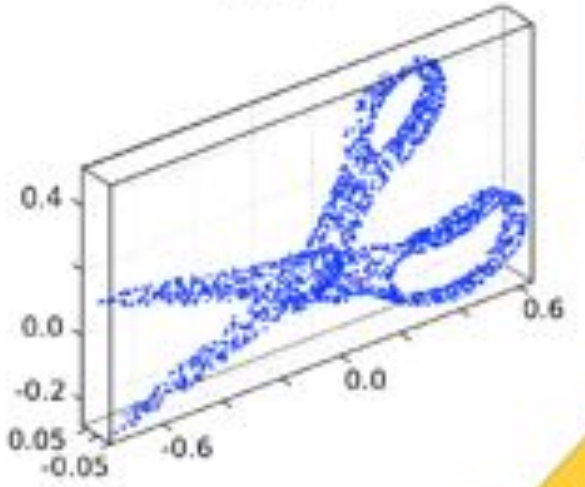
State of the art methods typically only synthesize grasps with point contacts

Beyond precision grasps?

deformable objects?

Images from openrave.org, cs.columbia.edu/~cmatei/graspit and simox.sourceforge.net respectively

A difficult challenge:



sensor noise?

Coulomb friction?

"Pick up **any** pair of scissors in *an arbitrary* configuration",
"Pick up a **soft deformable** bag by its handle"
"Carry an object by **wrapping** your arms around it"

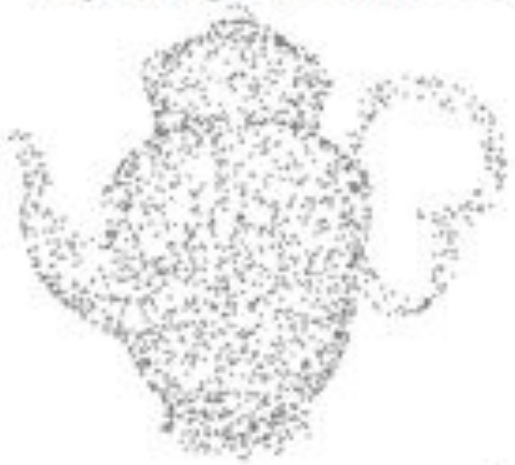
mass distribution?

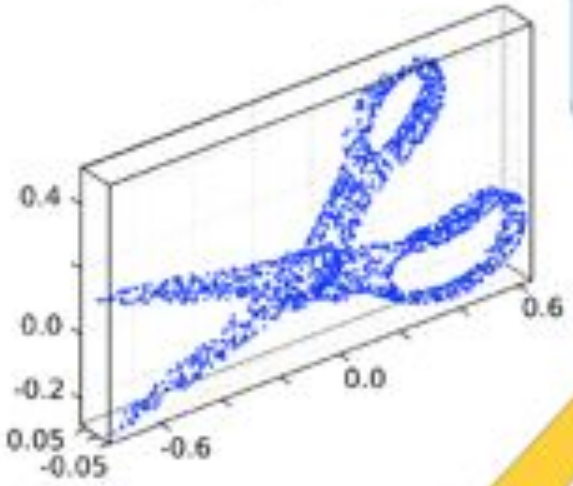
task constraints?

kinematic structure?

occlusions?

We propose to work with ***equivalence classes*** inspired by algebraic topology and based on global features:





sensor noise?

Coulomb friction?

"Pick up **any** pair of scissors in *an arbitrary* configuration",
"Pick up a **soft deformable** bag by its handle"
"Carry an object by **wrapping** your arms around it"

occlusions?

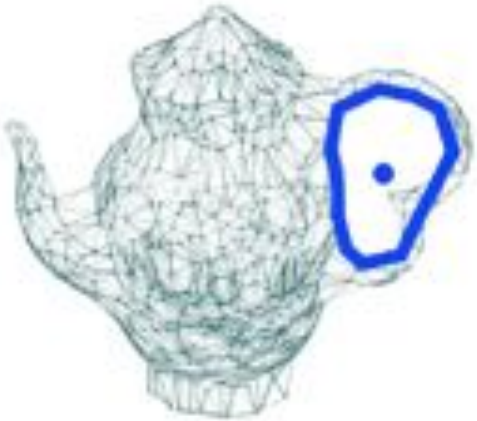
task constraints?

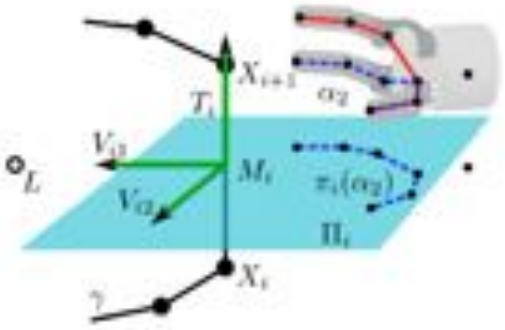
kinematic structure?

mass distribution?

Key ingredients

Topological object features	Topologically inspired robot hand representation	Topologically inspired control approach	Verification based on topological ideas

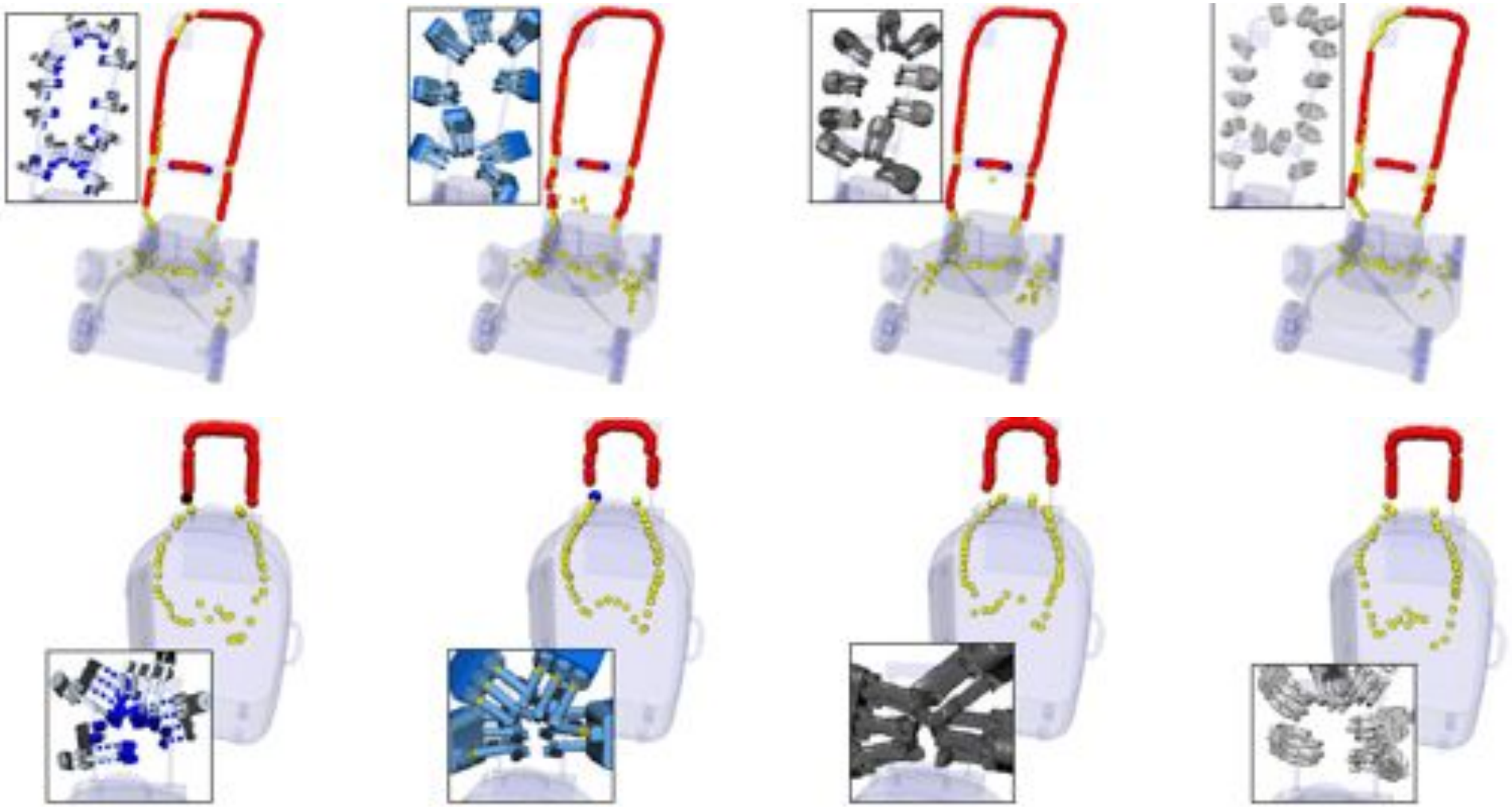




$$Lk(\gamma_1, \gamma_2)$$

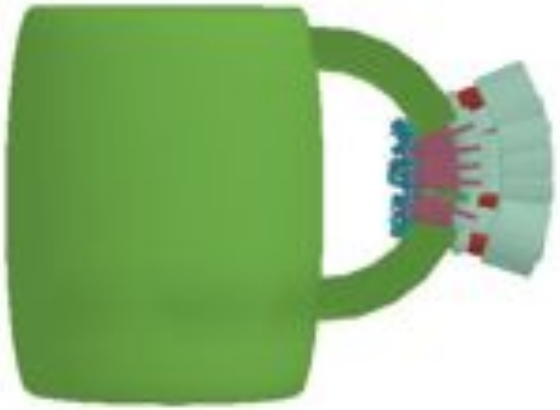
Holes and Loops are Everywhere

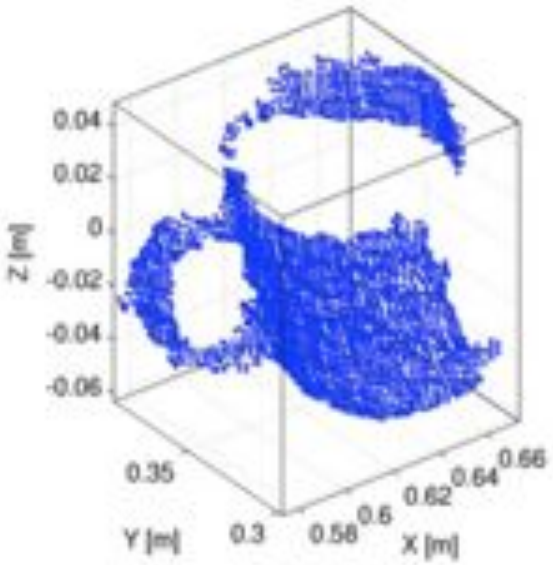
Clasping, latching, hooking

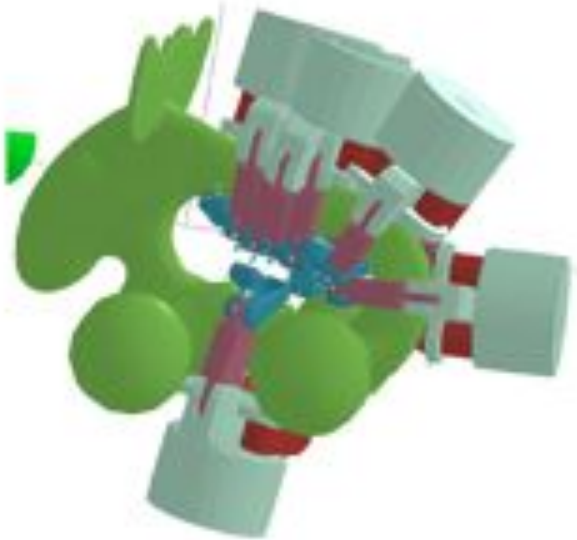


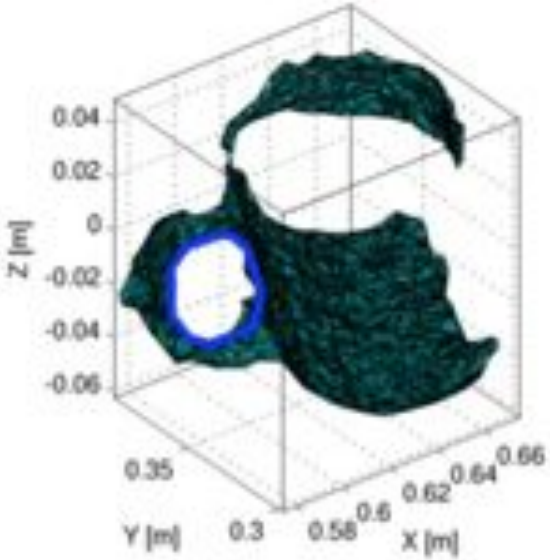
Grasping objects with holes: A topological approach, F. Pokorny, J. Stork, D. Kragic, *IEEE International Conference on Robotics and Automation,* 2013.

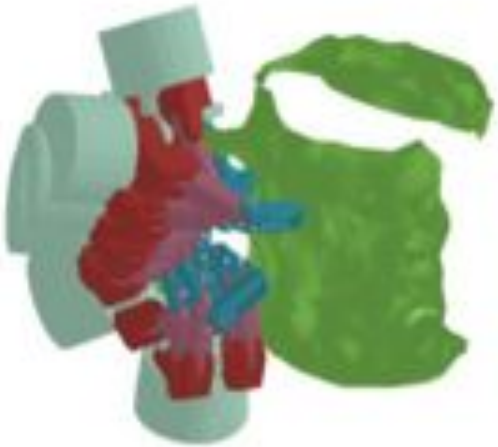
Examples




0.04
0.02
0
-0.02
-0.04
-0.06
Z [m]
0.35
0.3
Y [m]
0.58
0.6
0.62
0.64
0.66
X [m]




0.04
0.02
0
-0.02
-0.04
-0.06
Z [m]
0.35
0.3
Y [m]
0.58
0.6
0.62
0.64
0.66
X [m]



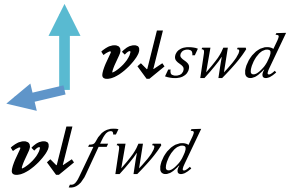
Object Representation: Point Cloud to Loops

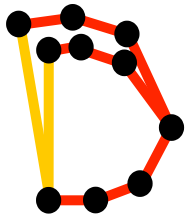
- Full view point cloud (PC)
- Refined Delaunay complex
- Shortest loops around holes [Busaryev *et al,* 2010]
- Topology-based global feature

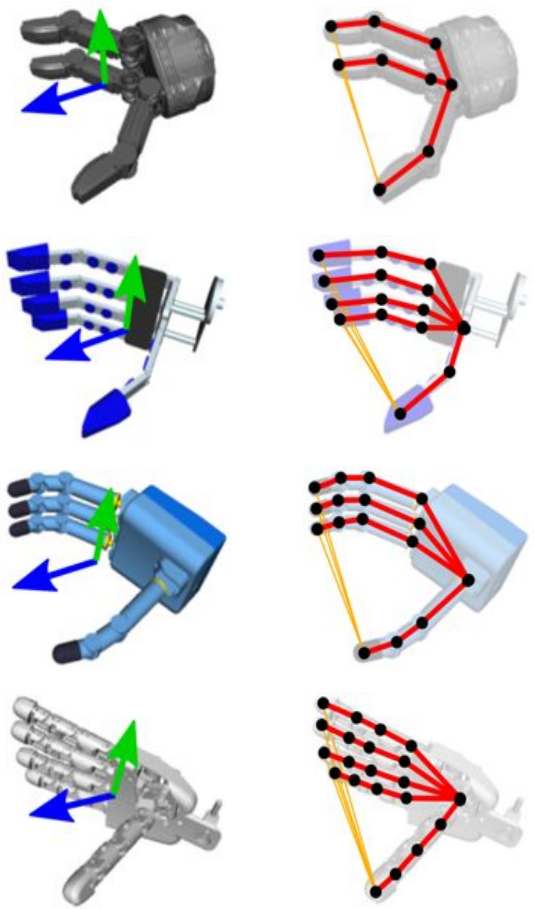
Integrated Motion and Clasp Planning with Virtual Linking, J. Stork, F.T. Pokorny, D. Kragic, *IEEE IROS* 2013.

Hand Representation

- Grasp center point frame
- Finger curves
- Virtual chords
- Finger loops
- Pre-shape and automatic closing







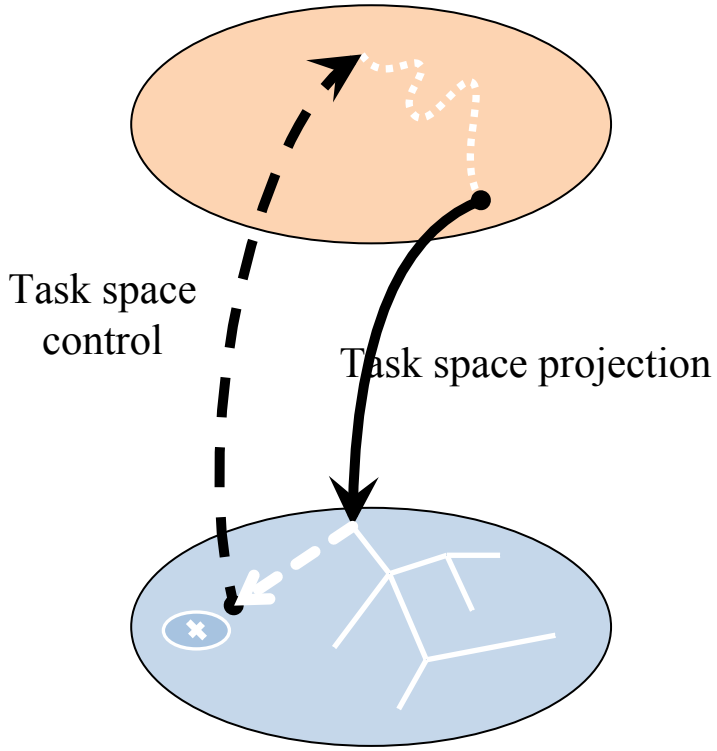
Task Space RRT: Sampling-based Planning and Control

12-21 dim **Joint space**

- Arm & Hand joints
- Collision checking
- Trajectory smoothing

4 dim **Task space**

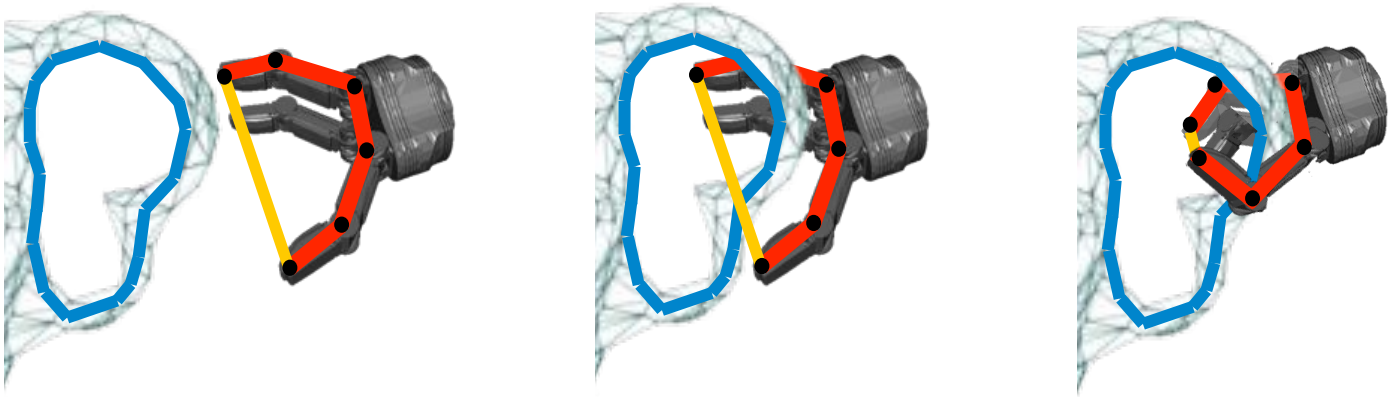
- Position, Virtual linking
- Alignment (implicit)
- RRT-Extend (random goal)
- RRT-Connect (closest loop position)



Virtual Linking

$$VLk\left(\gamma,\alpha\right)=\left(1-\frac{\delta_{\alpha}}{|\hat{\alpha}|}\right)Lk\left(\gamma,\hat{\alpha}\right)$$

Fingertip distance(■) x Gauss linking(■, ■■)



$|\hat{\alpha}|$ - length of a virtual finger loop $\hat{\alpha}$

δ_{α} - the Euclidian distance between the first and last node of the finger curve α

A Topology-based Object Representation for Clasping, Latching, and Hooking

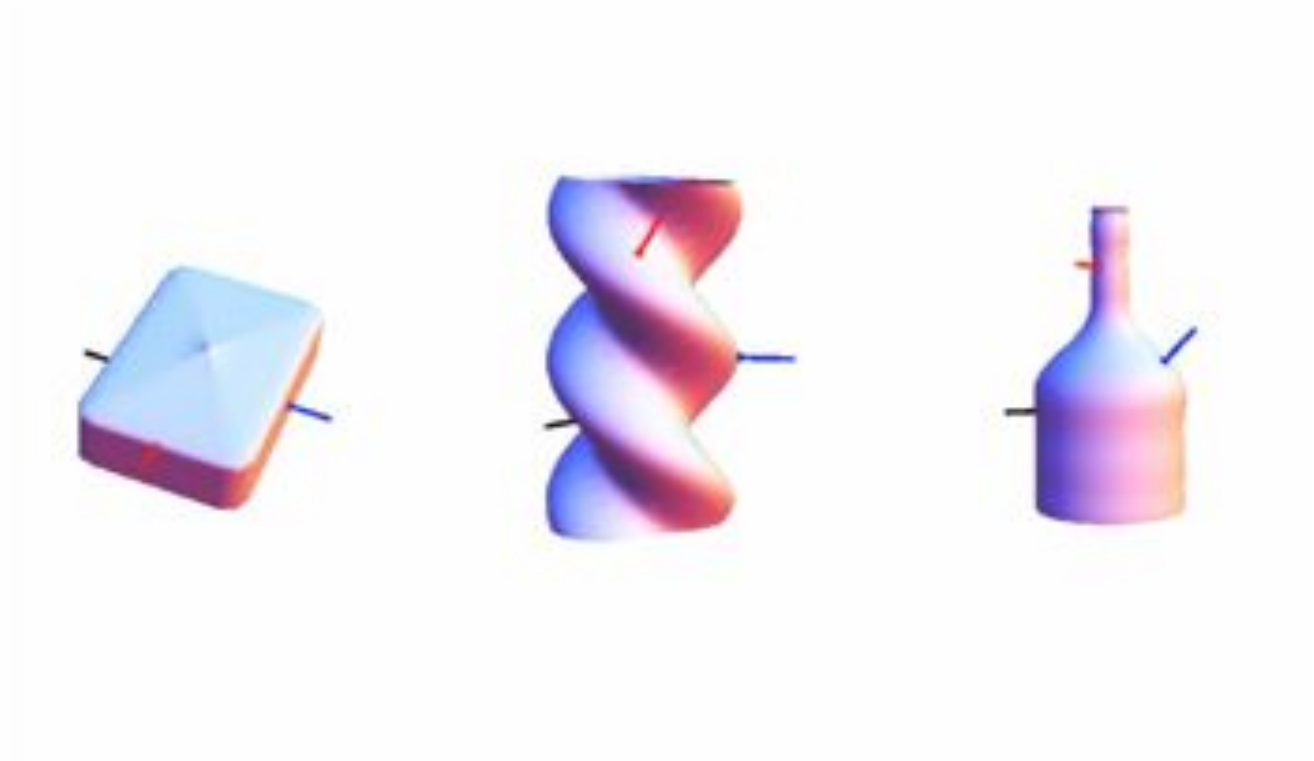
We have described an approach for grasping potentially deformable objects with holes

How can we represent a space of grasps and shapes:

- up to continuous deformations and
- based on observed grasp and shape data?

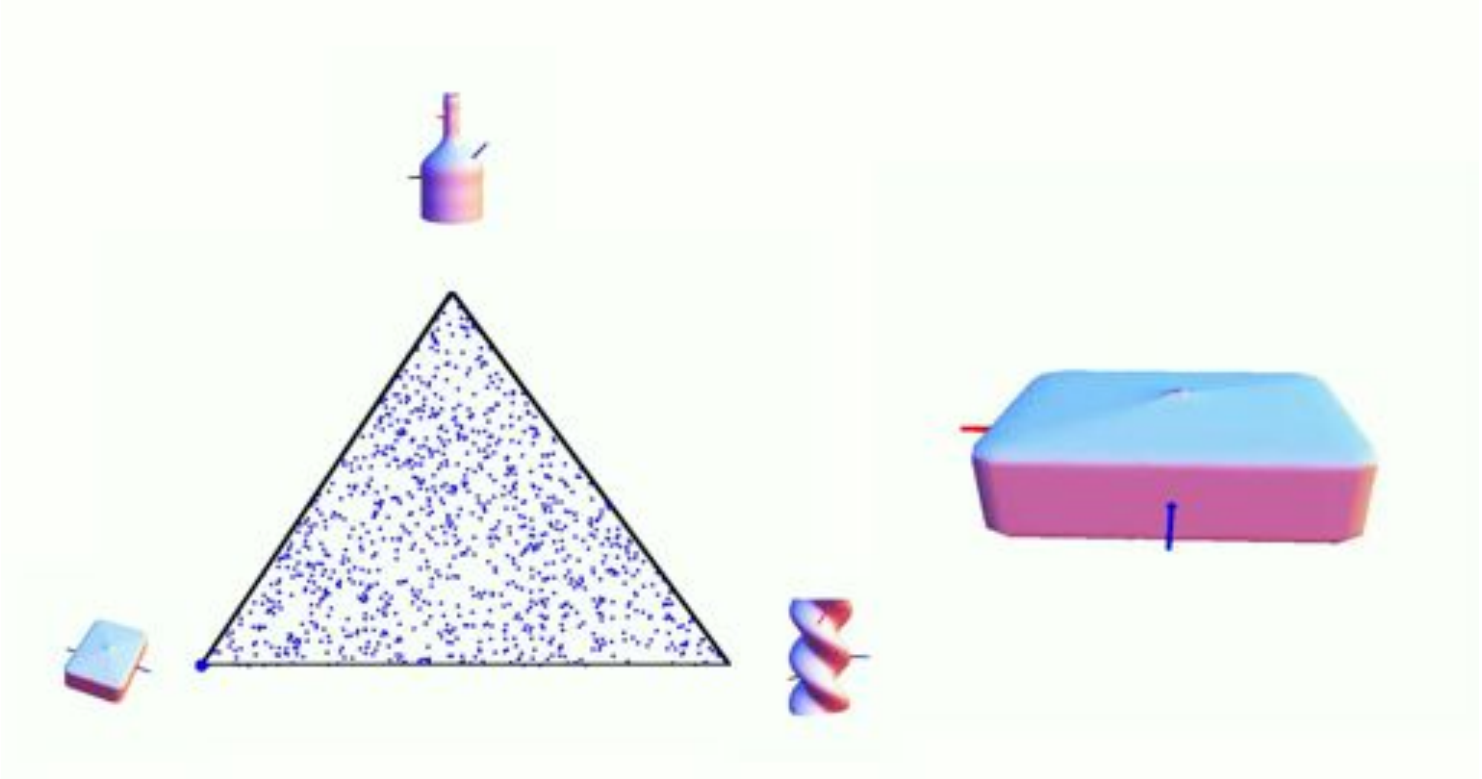
Grasp Moduli Spaces, F.T. Pokorny, K. Hang, D. Kragic, *Robotics: Science and Sytems,* 2013.

Grasp Moduli Spaces



Grasp Moduli Spaces

- Combined grasp/shape space
- Ability to deform continuously between configurations
- Transfer grasps to new shapes
- Interpolate between grasps/shapes

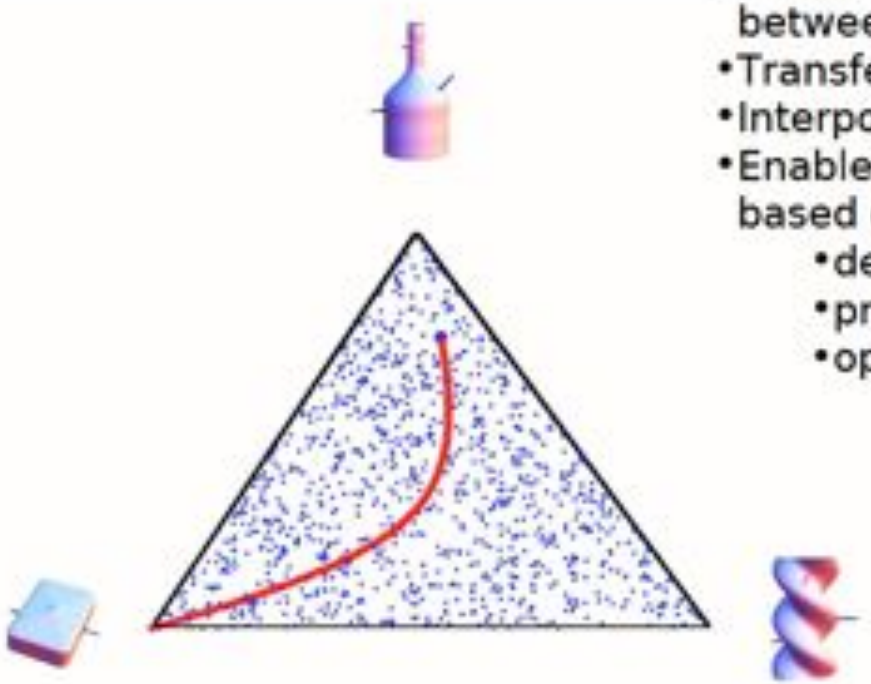


We can continuously move between configurations.

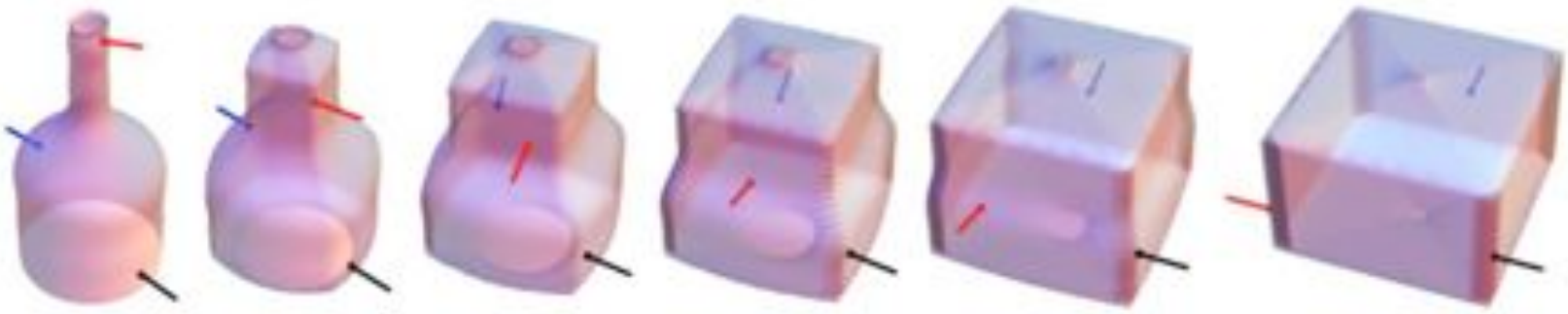
Grasp Moduli Spaces

We would like:

- Combined grasp/shape space
- Ability to deform continuously between configurations
- Transfer grasps to new shapes
- Interpolate between grasps/shapes
- Enable radically new approaches to grasp synthesis based on
  - deformation to "simple representatives"
  - probabilistic techniques
  - optimization



Grasp Moduli Spaces



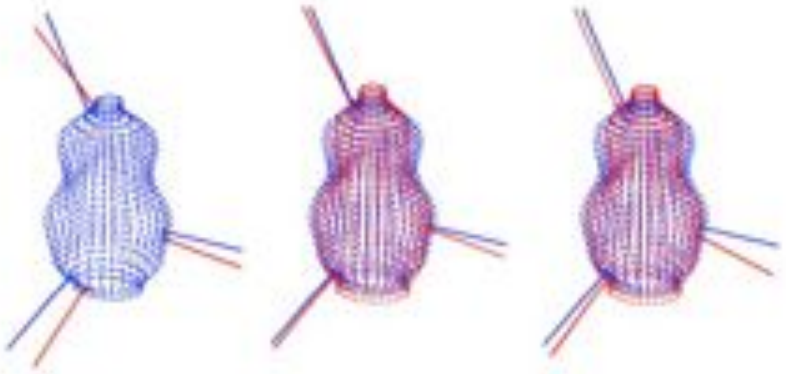
We define a **metric** on this space and show how to deform between arbitrary

- grasps on a single cylindrical surface
- grasps with identical coordinates but on different cylindrical surfaces
- different grasps on different cyl. surfaces

Grasp Moduli Spaces

We can ask precise questions about **local deformations**, e.g. if for a grasp quality measure Q,

- stable grasps remain stable under small perturbations in surface shape (object sensor noise)?
- stable grasps remain stable under small deformations in grasp position (control noise)?
- grasps remain stable under small deformations in both pose **and** shape in $\mathcal{G}^{cyl}(m)$



We quantify the above using a thorough statistical analysis of 100 million grasps and for an example set of shapes. Analytic proofs are currently in submission.

Conclusions

- Active vision for scene understanding: figure –ground segmentation remains an open problem
- Attention can help but we need to explore more cues both for detection and grouping
- Interaction with objects requires a representation that can be trasferred accross embodiements
- Task based grasping provides a basis for generating grasps relevant for the subsequent task
- Caging a viable way of interaction with deformable objects

Thanx to

- Mårten Björkman
- Javier Romero
- Jeannette Bohg
- Carl Henrik Ek
- Yasemin Bekiroglu
- Renaud Detry
- Christian Smith
- Petter Ögren
- Yiannis Karayiannidis
- Florian Pokorny
- Johannes Stork

- Swedish Foundation for Strategic Research
- Swedish Research Council
- European Commission

Journal papers

- Non-Parametric Hand Pose Estimation with Object Context, J. Romero, H Kjellstrom, CH Ek, D. Kragic, Image and Vision Computing, 2013
- Detecting, segmenting and tracking unknown objects using multi-label MRF inference, M. Bjorkman, N. Bergstrom, D. Kragic, Computer Vision and Image Understanding, 2013
- Enabling grasping of unknown objects through a synergistic use of edge and surface information, G. Kootstra, M. Popovic, J.A. Jorgensen, K. Kuklinski, K. Miatliuk, D. Kragic, N. Kruger, International Journal of Robotics Research, 31(10), pp.1190-1213, 2012
- Assessing grasp stability based on learning and haptic data, Y. Bekiroglu, J. Laaksonen, J.A. Jorgensen, V. Kyrki, D. Kragic, IEEE Transactions on Robotics, 27(3), pp.616-629, 2011
- Learning Grasping Points with Shape Context, J. Bohg, D. Kragic, Robotics and Autonomous Systems, Volume 58, Issue 4, pp. 362-377, 2010

Conference papers

- Enhancing Visual Perception of Shape through Tactile Glances, M. Bjorkman, Y. Bekiroglu, V. Hogman, D. Kragic, *IEEE IROS* 2013.
- Integrated Motion and Clasp Planning with Virtual Linking, J. Stork, F.T. Pokorny, D. Kragic, *IEEE IROS* 2013.
- Online Kinematics Estimation for Active Human-Robot Manipulation of Jointly Held Objects, Y. Karayiannidis, C. Smith, F. Vina, D. Kragic, *IEEE IROS* 2013.
- Grasp Moduli Spaces, F.T. Pokorny, K. Hang, D. Kragic, *Robotics: Science and Sytems,* 2013.
- Learning a dictionary of prototypical grasp-predicting parts from grasping experience, R. Detry, M. Madry, CH Ek, D. Kragic, *IEEE ICRA* 2013.
- A probabilistic framework for task-oriented grasp stability assessment, Y. Bekiroglu, D. Song, L. Wang, D. Kragic, *IEEE ICRA,* 2013.
- Generating Object Hypotheses in Natural Scenes through Human-Robot Interaction, N. Bergstrom, M. Bjorkman, D. Kragic, *IEEE IROS* 2011.
- Tracking people interacting with objects, H. Kjellstrom, D. Kragic, M. Black, *IEEE CVPR* 2010.