Perception, Planning, and Learning for Cognitive Robots

Sven Behnke

University of Bonn Computer Science Institute VI Autonomous Intelligent Systems



Many New Application Areas for Robots

- Self-driving cars
- Logistics
- Agriculture, mining
- Collaborative production
- Personal assistance
- Space, search & rescue
- Healthcare
- Toys

Need more cognitive abilities!











Sensory-motor Loop

- Environment perception
- Behavior generation
- Affecting the environment





Some of our Cognitive Robots

- Equipped with numerous sensors and actuators
- Complex demonstration scenarios



Soccer

Domestic service

ce Mobile manipulation

Bin picking

Aerial inspection



Some more of our Cognitive Robots

- Equipped with numerous sensors and actuators
- Complex demonstration scenarios



Rescue



Phenotyping



Human-robot collaboration



Telepresence



Soccer Robot NimbRo-OP2X

- 135 cm, 19 kg
- 18 DoF
 - 5 per leg (parallel kinematics)
 - 3 per arm
 - 2 in the neck
- 34 Dynamixel XH540 actuators
- Mini-ITX PC
- Nvidia SFF GPU
- Fisheye camera
- LiPo battery (14.8 V, 8 Ah)





6



RoboCup 2023 Humanoid AdultSize Final





Visual Perception for Soccer

Object detection, semantic segmentation, robot pose estimation



RoboCup 2023: Technical Challenges





Winning Team NimbRo @ RoboCup 2023



[Pavlichenko et al. RoboCup 2023]



Deep Learning

 Learning layered
 representations

Compositionality



UNIVERSITÄT BO

[Schulz; Behnke, KI 2012]

Neural Abstraction Pyramid



[Behnke, Rojas, IJCNN 1998] [Behnke, LNCS 2766, 2003]

UNIVERSITÄ

Iterative Image Interpretation

- Interpret most obvious parts first
- Use partial interpretation as context to iteratively resolve local ambiguities





[Behnke, Rojas, IJCNN 1998] [Behnke, LNCS 2766, 2003]

Neural Abstraction Pyramid for Semantic Segmentation of RGB-D Video

Recursive computation is efficient for temporal integration



[Pavel, Schulz, Behnke, Neural Networks 2017]

The Data Problem

- Deep Learning in robotics (still) suffers from shortage of available examples
- We address this problem in three ways:

1. Transfer learning:

Pre-training on large related data, self-supervised learning

Generating data: Online mesh databases, scene synthesis

3. Inductive biases:

3D projective geometry, camera motion, canonical frames, object relations, compositionality, ...







RGB-D Object Recognition and Pose Estimation

Transfer learning from large-scale data sets





[Schwarz, Schulz, Behnke, ICRA2015]

Canonical View, Colorization

- Objects viewed from different elevation
- Render canonical view





Colorization based on distance from center vertical







[Schwarz, Schulz, Behnke, ICRA2015]

Pretrained Features Disentangle Data

 t-SNE embedding



[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

Improved both category and instance recognition

	Category A	ccuracy (%)	Instance Accuracy (%)		
Method	RGB	RGB-D	RGB	RGB-D	
Lai <i>et al.</i> [1]	74.3 ± 3.3	81.9 ± 2.8	59.3	73.9	
Bo <i>et al.</i> [2]	82.4 ± 3.1	87.5 ± 2.9	92.1	92.8	
PHOW[3]	80.2 ± 1.8		62.8		
Ours	83.1 ± 2.0	88.3 ± 1.5	92.0	94.1	
Ours	83.1 ± 2.0	89.4 ± 1.3	92.0	94.1	

Confusion:



1: pitcher / coffe mug



2: peach / sp







[Schwarz, Schulz, Behnke, ICRA2015]

Amazon Robotics Challenge

- Storing and picking of items
- Dual-arm robotic system





[Amazon]



Object Capture and Scene Rendering

Turntable + DLSR camera



Insertion in complex annotated scenes





Semantic Segmentation and Grasp Pose Estimation

- Semantic segmentation using RefineNet [Lin et al. CVPR 2017]
- Grasp positions in segment centers









mouse_traps conf: 0.921731 windex 'conf: 0.861246 q-tips_500 conf: 0.475015 fiskars_scissors 'conf: 0.831069 ice_cube_tray conf: 0.976856

[Schwarz et al. ICRA 2018]



Amazon Robotics Challenge Finals





[Schwarz et al. ICRA 2018]

Dense Convolutional 6D Object Pose Estimation

- Extension of PoseCNN [Xiang et al. RSS 2018]
- Dense prediction of object center and orientation





Stillleben: Learning from Synthetic Scenes

- Cluttered arrangements from 3D meshes
- Photorealistic scenes with randomized material and lighting including ground truth
- For online learning & render-and-compare
- Semantic segmentation on YCB Video Dataset
 - Close to real-data accuracy
 - Improves segmentation of real data







[Schwarz and Behnke, ICRA 2020]



Learning from SAM: Sim2Real Domain Adaptation through Segment VI-Regularization

- Learns from synthetic scenes and unannotated real images
- Supervised training of semantic segmentation for synthetic scenes
- Segment Anything Model (SAM) used to generate many overlapping segments for real images
- Dense features from shared backbone
- Contrastive loss for segments
 - Features within a segment are trained to have low variance
 - Features for different segments trained to have high variance



[Bonani et al. arXiv:2309.15562]



Learning from SAM: Sim2Real Domain Adaptation through Segment VI-Regularization

 Learned dense features correspond well to objects, are stable under camera motion, and label sub-parts





[Bonani et al. arXiv:2309.15562]

Learning from SAM: Sim2Real Domain Adaptation through Segment VI-Regularization

- Good results on real images without need for real labels
- Better than training with real labels on VCB-Video



YCB-Video

[Bonani et al. arXiv:2309.15562]







Mean IoU

Method	YCB-Video [19]	HomebrewedDB [20]		
Imbusch <i>et al.</i> [4] - real labels - synthetic only - full	0.770 0.701 0.763	$0.737 \\ 0.481^1 \\ 0.558^1$		
Ours - real labels - synthetic only - CUT [4] only ² - full	0.839 0.807 0.814 0.853	0.883 0.748 0.787 ³		

Note: "real labels" is a baseline which has access to real supervision.

¹ Using Stillleben [2] synthetic data, where we use Blender-Proc4BOP.

² Training our backbone on CUT-refined synthetic data.

³ Model was trained for only 200k epochs.



MSPred: Video Prediction at Multiple Spatio-Temporal Scales

Coarser, more abstract predictions for longer time horizons in higher layers

Predict image itself, human pose joint keypoints, and human body position





MSPred: Video Prediction at Multiple Spatio-Temporal Scales

- Coarser, more abstract predictions for longer time horizons in higher layers
- Predict image itself, human pose joint keypoints, and human body position





[Villar-Corrales et al., BMVC 2022]

Hierarchical Planning in the Now

- Use predicted state on different layers of abstraction for planning
- Coarse-to-fine planning makes actions more concrete as they come closer to execution
- Plan consists of few steps on each layer



Centauro Robot





- Serial elastic actuators
- 42 main DoFs
- Schunk hand
- 3D laser
- RGB-D camera
- Color cameras
- Two GPU PCs

[Tsagarakis et al., IIT 2017]



Hybrid Driving-Stepping Locomotion Planning: Abstraction

- Planning in the here and now
- Far-away details are abstracted away





Hybrid Driving-Stepping Locomotion Planning: Abstraction

Level	M	ap Resolution		Map Features	Robot Representation		Action Semantics	
1		• 2.5 cm • 64 orient.	\land	• Height			\bigwedge	 Individual Foot Actions
2		• 5.0 cm • 32 orient.		HeightHeight Difference				• Foot Pair Actions
3	\bigvee	10 cm16 orient.		HeightHeight DifferenceTerrain Class	\bigvee			• Whole Robot Actions







[Klamt and Behnke, IROS 2017, ICRA 2018]

CENTAURO Evaluation @ KHG: Locomotion Tasks





[Klamt et al. RAM 2019]

Transfer of Manipulation Skills





[Rodriguez and Behnke ICRA 2018]
Learning a Latent Shape Space

- Non-rigid registration of instances and canonical model
- Principal component analysis of deformations





[Rodriguez and Behnke ICRA 2018]

Interpolation in Shape Space





[Rodriguez and Behnke ICRA 2018]

Shape-aware Non-rigid Registration





[Rodriguez and Behnke ICRA 2018]

Shape-aware Registration for Grasp Transfer





Collision-aware Motion Generation

Constrained Trajectory Optimization:

- Collision avoidance
- Joint limits
- Time minimization
- Torque optimization





[Pavlichenko et al., IROS 2017]

Grasping an Unknown Power Drill and Fastening Screws





CENTAURO: Complex Manipulation Tasks





Regrasping for Functional Grasp

- Direct functional grasps not always feasible
- Pick up object with support hand, such that it can be grasped in a functional way





[Pavlichenko et al. Humanoids 2019]

Regrasping Experiments







ANA Avatar XPRIZE Competition

- Organized by XPRIZE Foundation
- Sponsored by All Nippon Airways (ANA)
- Objective: Create a robotic avatar system that can transport human senses, actions, and presence to a remote location in real time
 - Expanding human connection
 - Transferring skills
 - Exploring dangerous or inaccessible places
- Panel of 22 expert judges
- Launched 03/2018
- Prize purse of \$10M
- 99 teams registered by 09/2019











Finals 11/2022 in Long Beach, CA, USA



- 17 teams from 10 countries: Leading research groups and companies
- Mission on a distant planet, 10 tasks in a given order
- Qualification day, two test days with down-selection of teams
 - ➔ System reliability required



Finals Tasks

- Three domains:
 - Connectivity
 - Exploration
 - Skill transfer
- Incl. judging object weight and remote feeling of texture
- One point per task
- Operator experience:3 points
- Recipient exp.: 2 pts
- Trial time to break ties



Start



1997t

2. Ir







4. Activate switch 5: Travel planet 6. Identify full canister 7: Place it





9: Use drill 10







[XPRIZE]

NimbRo Avatar System





[Lenz et al. International Journal of Social Robotics 2023]

ANA Avatar XPRIZE Finals Test Run Day 1





Operator Face Animation

- Recording of the operator before putting on the head-mounted display (HMD)
- Live animation using video of eyes in HMD and mouth region



Face Animation @ ANA Avatar XPRIZE Finals

Team UNIST





Northeastern [12]

Ours (NimbRo)

Team AVATRINA [13]





i-BOTICS





Pollen Robotics



Source: Official XPRIZE Avatar live stream

ANA Avatar XPRIZE Finals Day 2 Testing



UNIVERSITÄT BO

ANA Avatar XPRIZE Finals Results



Rank	Team name	Time	Task score	Judged score	Total
1	NimbRo (DE)	5:50	10	5	15
2	Pollen Robotics (FR)	10:50	10	5	15
3	Team Northeastern (US)	21:09	10	4.5	14.5
4	AVATRINA (US)	24:47	10	4.5	14.5
5	i-Botics (NL)	25:00	9	5	14
6	Team UNIST (KR)	25:00	9	4.5	13.5
7	Inbiodroid (MX)	25:00	8	5	13
8	Team SNU (KR)	25:00	8	4.5	12.5
9	AlterEgo (IT)	25:00	8	4.5	12.5
10	Dragon Tree Labs (SG)	25:00	7	4	11
11	Avatar Hubo (US)	25:00	6	3.5	9.5
12	Last Mile (JP)	25:00	5	4	9
			[XPRIZI	E]	





Motivation for Autonomy

- Longer latencies require less direct control
 - Use autonomous skills, such as grasping an object or navigating to a waypoint
 - Shared autonomy where the operator controls highlevel behavior and autonomy fills-in the low-level details (horse metaphor, Flemisch 2003)
- Operator might not always be available
 - 1:1 control often too costly
 => one operator must supervise many robots
 - Issues of privacy and of being in operator's dept
- AI: Understanding intelligence by creating intelligent artefacts



[Photographer: Volker Lannert]



Unmatched Human Operators





Humans can solve many tasks by teleoperation

- Can cope with novel situations, quickly learn new tasks
- Recognize and mitigate errors
- Far beyond the capabilities of autonomous robots







Human Cognitive System

Cognitive architecture of the human mind has evolved to continuously interact with changing environments and self-monitor



Cognitive functions according to Kahneman (2011) and Dehaene (2017)

UNIVERSITÄ

My Objective

Develop methods for learning perception and planning for service robots, which go beyond unconscious routine tasks by incorporating conscious processing to cope with novel situations and self-monitor





Overall Approach

Equip service robots with key elements of human cognitive architecture
 Bottom-up approach ensures grounding of conscious processing



Cognitive functions according to Kahneman (2011) and Dehaene (2017)



Unconscious Perception & Tracking

- 1. Learning hierarchical representations
- 2. Learning 3D multimodal scene models
- 3. Learning object models & relations
- 4. Learning prediction and tracking

Scene compositionality

- Objects and scenes described by their constituent parts and their relations
- Infinite variants from a finite set of building blocks
- Exploit inductive biases like canonical frames, 3D projective geometry, camera motion, object relations, compositional structure, hierarchical categorization, ...





Object-centric Video Prediction Decoupling Dynamics and Interaction



[Villar-Corrales et al. ICIP 2023]

- Scene parsing into object slots
- Video synthesis from objects and masks
- Predictor decouples temporal and relational attention



UNIVERSITÄT

Object-centric Video Prediction: Scene Segmentation

- Objects are represented in different slots
 - → Unsupervised scene segmentation





[Villar-Corrales et al. ICIP 2023]

YOLOPose: Multi-Object 6D Pose Estimation using Keypoint Regression





[Amini et al. IAS 2022, Best Paper Award]



YOLOPose: Multi-Object 6D Pose Estimation using Keypoint Regression

Encoder self-attention



Object detections and decoder cross-attention



[Amini et al. IAS 2022, Best Paper Award]



64

MOTPose: Attention-based Temporal Fusion for Multi-object 6D Pose Estimation

Propagating object embeddings, object descriptors, and poses



[Periyasamyn and Behnke, ICRA 2024]



Unconscious Prediction and Control

- 1. Learning action-conditioned prediction
- 2. Learning to control in the now
- 3. Learning reusable skills
- 4. Learning from imitation and real-robot experience

Action compositionality

- Activities consists of sequence of actions, which can be decomposed into movement primitives
- Exploiting inductive biases like hierarchical structure, object binding, planning in the now, ...





Learning Interactive Functional Grasping



[Mosbach and Behnke CASE 2023, Best Paper Award]

Learning Pre-grasp Manipulation for Human-like Functional Grasping



 Dense multi-component reward function encodes desired functional grasp



Learns to reposition and reorient objects to achieve functional grasps







Grasp Anything: Augmenting Reinforcement Learning with Instance Segmentation to Grasp Arbitrary Objects

Teacher training



Teacher-guided sensorimotor learning



 Real-world deployment of promptable grasping policy



UNIVERSIT

Grasp Anything: Augmenting Reinforcement Learning with Instance Segmentation to Grasp Arbitrary Objects

Learned policy with improved object visibility is real-world deployable





Conscious Prediction and Planning

- 1. Learning a working memory
- 2. Learning working memory predictions
- 3. Learning conscious planning
- 4. Learning new conscious concepts

Systematic generalization



- Reuse task knowledge in infinitely many novel situations in which irrelevant items change
- Working memory as communication bottleneck
 - Focus on few items, ignore all others which are irrelevant for the task
 - Must combine multiple lower-level items to larger, composite items



Conscious Self-monitoring

- 1. Representing uncertainty
- 2. Predicting multiple plausible futures
- 3. Error detection and mitigation
- 4. Interactive learning

Self-aware

72

- Being aware of own capabilities and limitations, dangers, etc.
- Systematically model and use uncertainty
 - Collect more information when needed
 - Avoid dangerous situations
 - Detect System 1 errors and mitigate them




Potential Impact

Consciousness is not a bug, but a feature!

- Will bring service robots to the next level
 - **Systematically generalize** skills and cope with novel situations
 - Self-monitor: obtain more information when needed, avoid risks, detect errors, and mitigate them
- Applicable to a large variety of openended domains
- Will enable personal service robots



[XPRIZE]



NimbRo@Home RoboCup 2023 in Bordeaux, France

LLM Integration

- Utilization of Function Calling Models (GPT / Mistral)
- Map available robot functionalities to user commands
- Given Information:
 - Environment info
 - Callable robot functionalities
 - Can request and query additional image data (counting, descriptions, ...)

User: Please bring me a sponge from the kitchen. System: Briefly summarize the next steps to accomplish the task in text form. Assistant: 1. Drive to the kitchen. 2. Search for a sponge in the kitchen. 3. Grasp the sponge. 4. Drive back to the parlor. 5. Hand over the sponge to the user. 6. Call the exit function to complete the interaction. System: From now on only use function calls to execute the task. **Assistant:** drive to location: {"location": "kitchen"} Tool: You successfully arrived in the new location kitchen. **Assistant:** find object: {"object name list": ["sponge"]} Tool: The following items were found in the kitchen: 3 Sponges **Assistant:** grasp object: {"object name": "sponge"} Tool: You successfully grasped the object sponge. **Assistant:** drive to location: {"location": "parlor"} Tool: You successfully arrived in the new location parlor. **Assistant:** place object: {"object name": "sponge"} Tool: You successfully placed the object sponge.

Assistant: exit: {}



Cooking Advice

 Q: Please describe the content of this image. Focus your description on food items, drinks, and ingredients that could be used for cooking.



A: The image shows a variety of food items and drinks arranged on a white surface. There is a can of SPAM, which is a type of canned cooked pork product.
Next to it is an open can of Campbell's Tomato Soup...

Cooking Advice

- Q: How can I cook dinner from these ingredients? <PROMPT>
- A: You can create a simple and creative meal using these ingredients. Here's a suggestion:
 - <u>SPAM and Orange Glaze</u>: Slice the SPAM into thin pieces and fry them in a pan until they are crispy on both sides. Create a glaze by squeezing the juice of the orange into a small saucepan, adding a bit of the Coca-Cola Zero Sugar for sweetness, and reducing it over medium heat until it thickens. Pour the glaze over the fried SPAM slices.
 - <u>Tomato Soup Enhancement</u>: Warm the Campbell's Tomato Soup on the stove. To add more depth to the flavor, you could mix in a small amount of the orange juice and a pinch of the JELL-O powder for a hint of sweetness and tanginess.



Conclusions

- Developed capable robotic systems for challenging scenarios
 - Soccer
 - Bin picking
 - Disaster response, telepresence
 - Domestic service
- Challenges include
 - 4D semantic perception
 - High-dimensional motion planning
 - Human-robot interaction
- Promising approaches
 - Prior knowledge (pretrained models, inductive bias, LLMs)
 - Shared experience (fleet learning)
 - Shared autonomy (human-robot)
- 78 Instrumented environments



