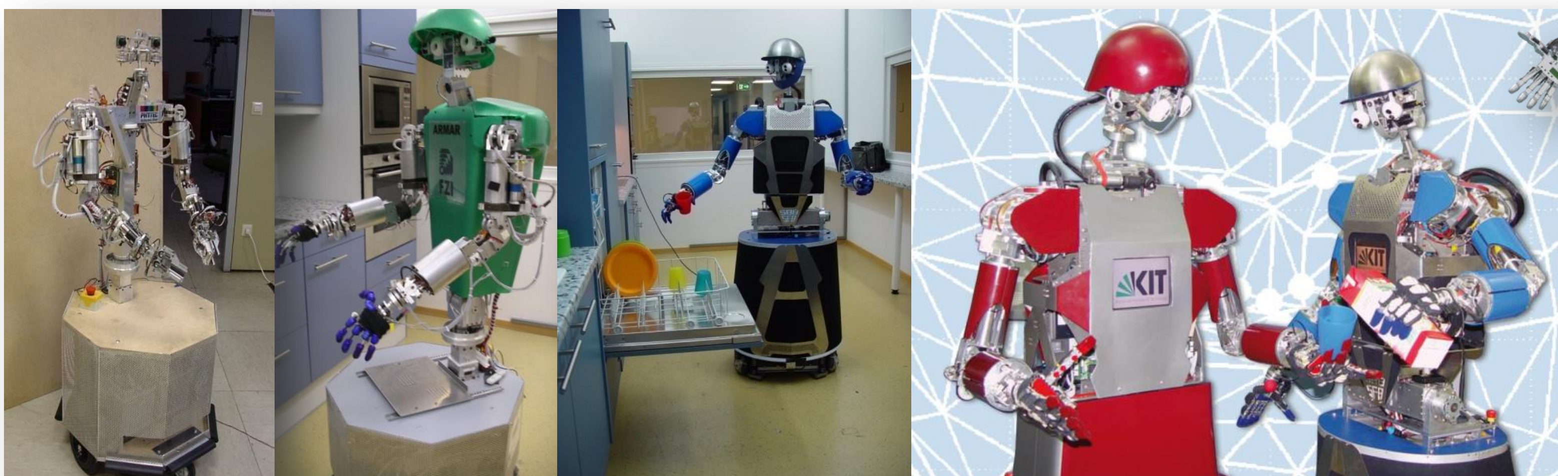
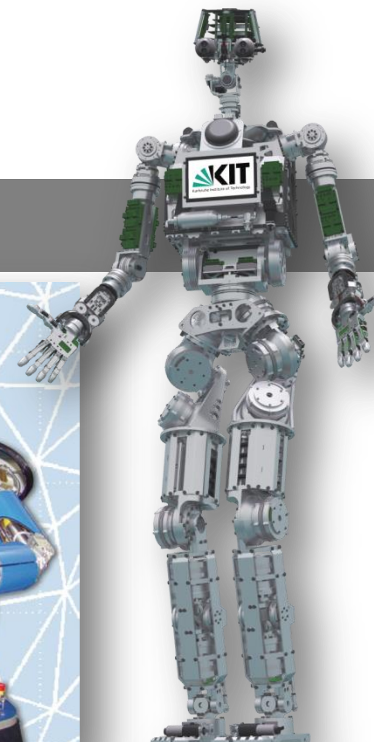
ICRA 2014 – Workshop on „Active Visual Learning and Hierarchical Visual Representations for General-Purpose Robot Vision”

Active Visual Perception for Humanoid Robots

Tamim Asfour
High Performance Humanoid Technologies (H^2T)

Institute for Anthropomatics and Robotics, High Performance Humanoid Technologies

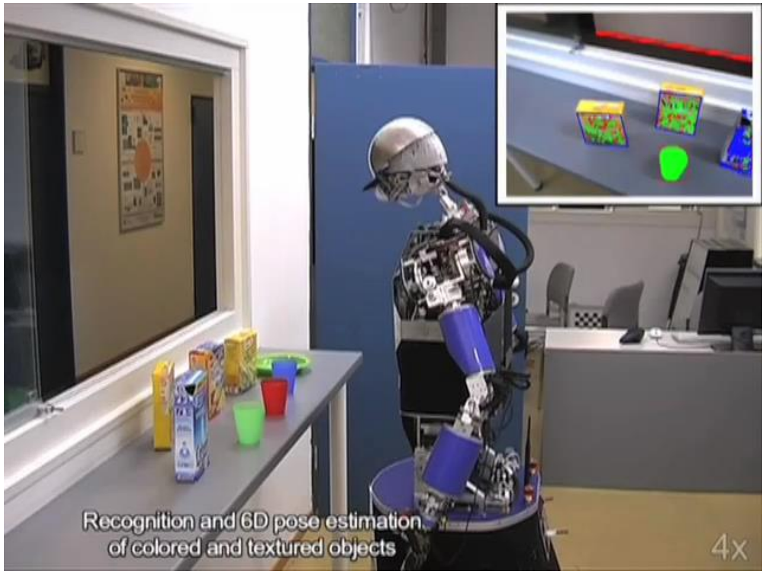




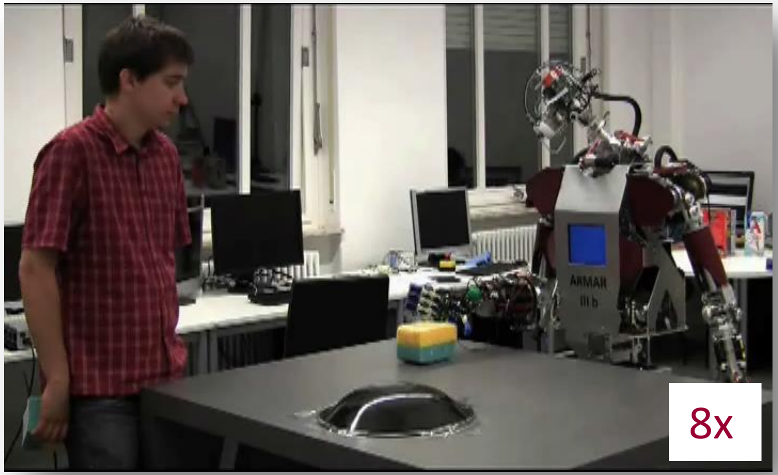
http://www.humanoid.kit.edu http://h2t.anthropomatik.kit.edu

Humanoids in the real world

- Grasping and manipulation



- Learning for human observation



Outline

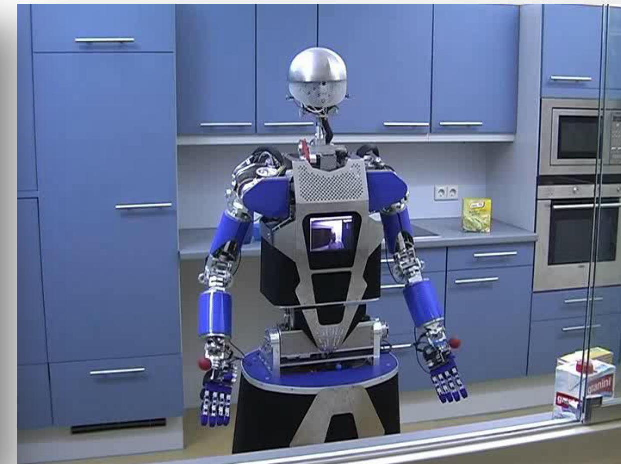
- Humanoid active head
- Visual perception for grasping and manipulation
- Active exploration for object learning and object search

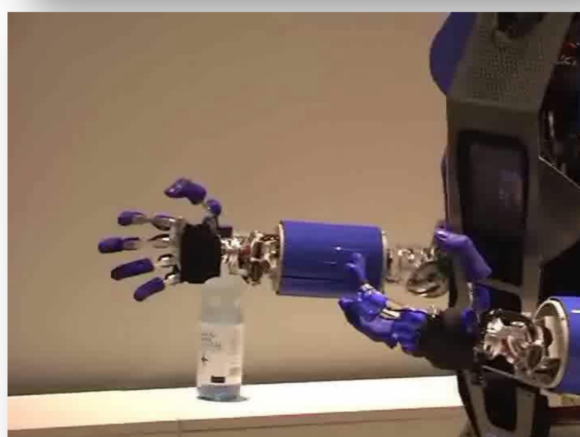
ARMAR-IIIa and ARMAR-IIIb

Karlsruhe Institute of Technology

- 7 DOF head with foveated vision
  - 2 cameras in each eye
  - 6 microphones
- 7-DOF arms
  - Position, velocity and torque sensors
  - 6D FT-Sensors
  - Sensitive Skin
- 8-DOF Hands
  - Pneumatic actuators
  - Weight 250g
  - Holding force 2,5 kg
- 3 DOF torso
  - 2 Embedded PCs
  - 10 DSP/FPGA Units
- Holonomic mobile platform
  - 3 laser scanner
  - 3 Embedded PCs
  - 2 Batteries
- Weight: 150 kg

Fully integrated humanoid system





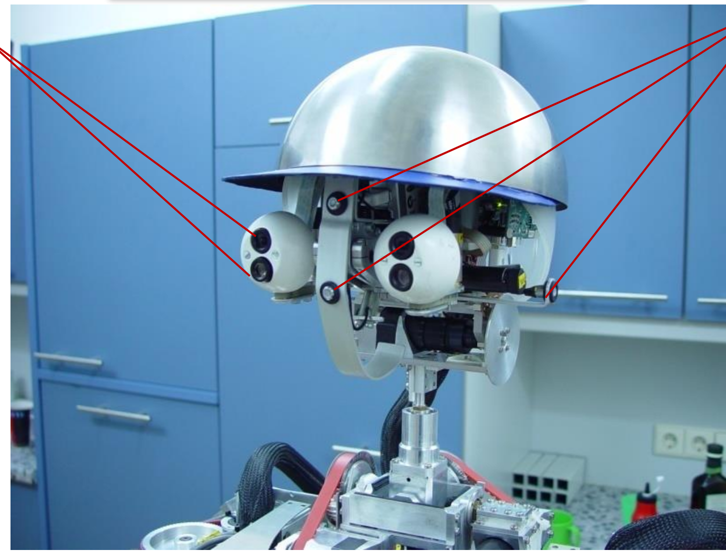
(Asfour et al. 2006, 2008)

ARMAR-III: Active Head

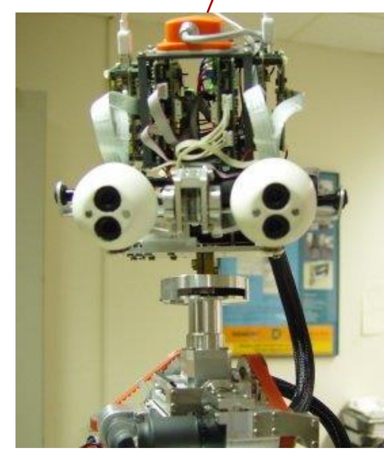
7 DOF
- 4 DOF neck
- 3 DOF eyes

Two cameras per eye
- wide-angle lens for peripheral vision
- narrow-angle lens for foveated vision

six microphones and six channel microphone pre-amplifier with integrated phantom power supply



6D inertial sensor



(Asfour et al. 2008)

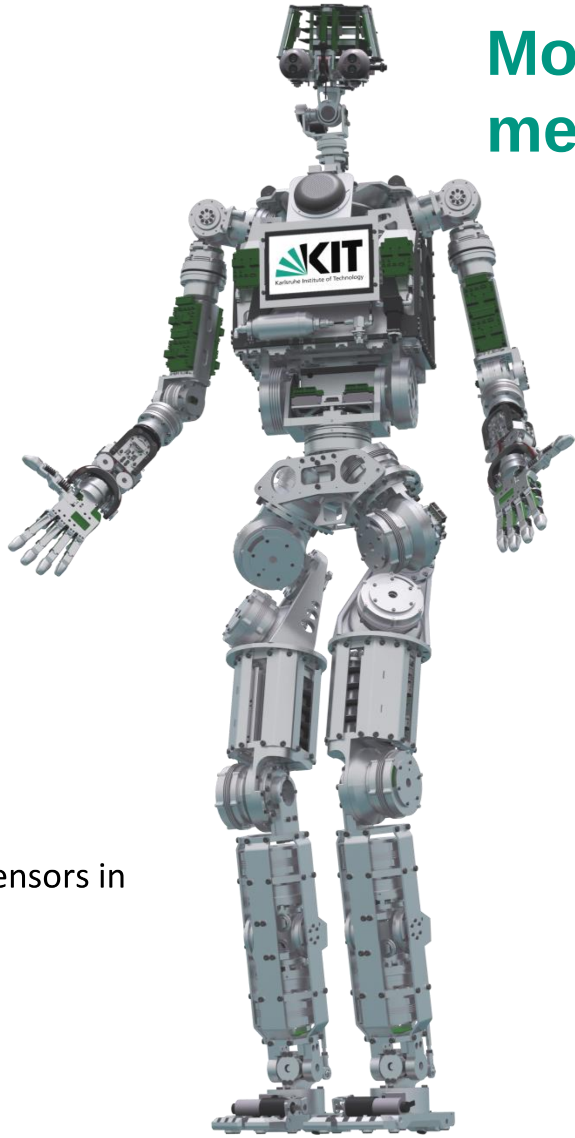
Copies of the head including the control software and basic vision processing library are used at Jozef Stefan Institute (Slovenia), KTH (Sweden), University of Bielefeld (Germany), University of Innsbruck (Austria), University of Pisa (Italy), University of Birmingham (UK), and at several labs at KIT

ARMAR-IV: Mechano-Informatics

- Torque controlled
- 3 on-board embedded PCs
- 76 Microcontroller
- 6 CAN Buses

- 63 DOF
  - 41 electrically-driven
  - 22 pneumatically-driven (Hand)

- 238 Sensors
  - 4 Cameras
  - 6 Microphones
  - 4 6D-force-torque sensors
  - 2 IMUs
  - 128 position (incremental and absolute), torque and temperature sensors in arm, leg and hip joints
  - 18 position (incremental and absolute) sensors in head joints
  - 14 load cells in the feet
  - 22 encoders in hand joints
  - 20 pressure sensors in hand actuators
  - …



More than mechatronics

ARMAR-IV

made@KIT

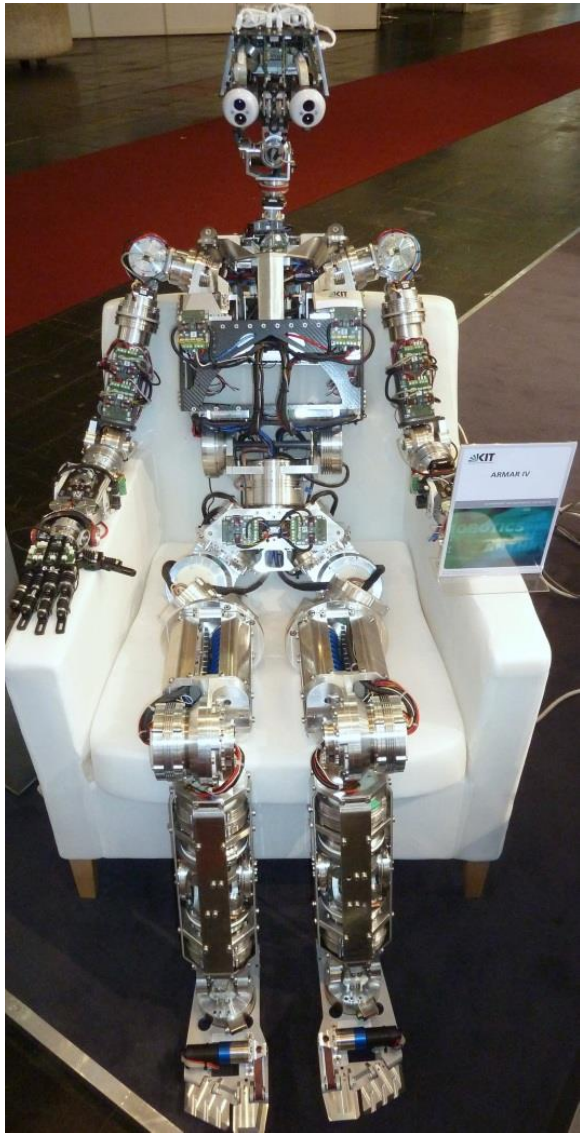
70 kg

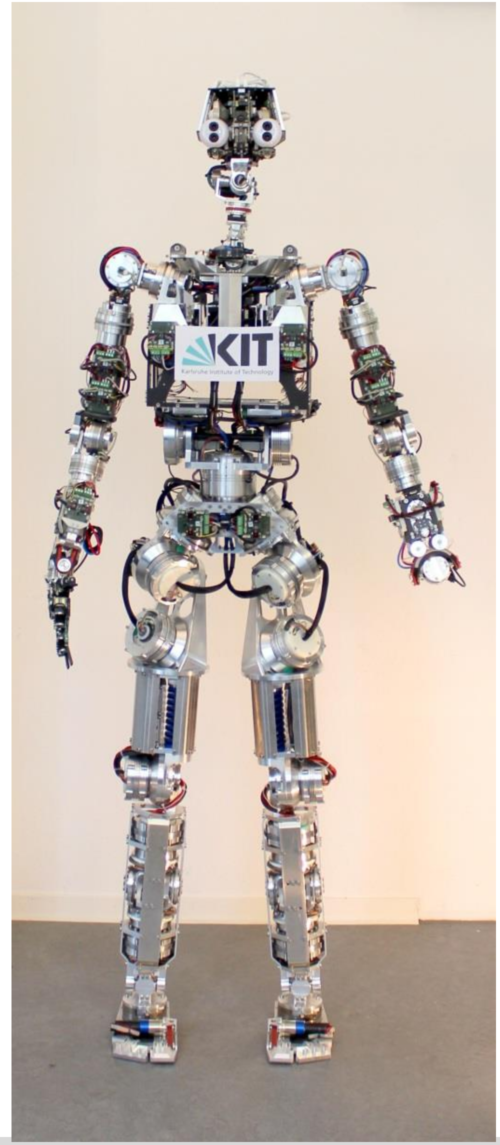
170 cm

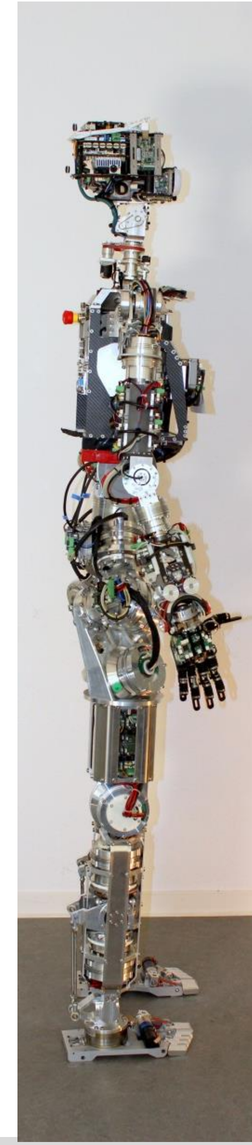
(Asfour et al. 2013)

ARMAR-IV

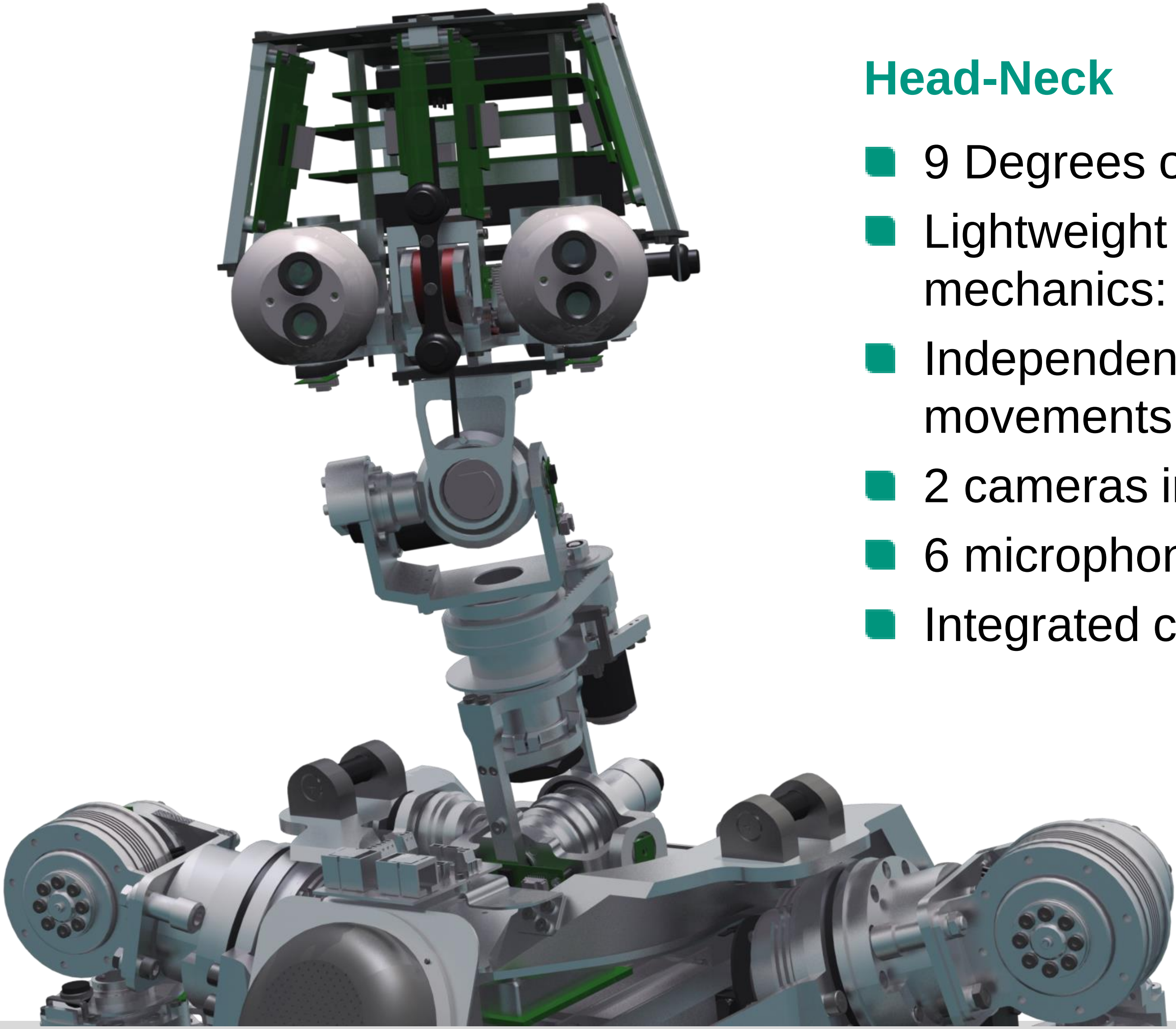
- 63 DOF
- 170 cm
- 70 kg
- Torque-controlled!







ARMAR IV - Head-Neck



Head-Neck

- 9 Degrees of freedom
- Lightweight design (weight of mechanics: 1412 g)
- Independent eye pan/tilt movements
- 2 cameras in each eyes
- 6 microphones
- Integrated computing power

ARMAR-III in the RoboKITchen

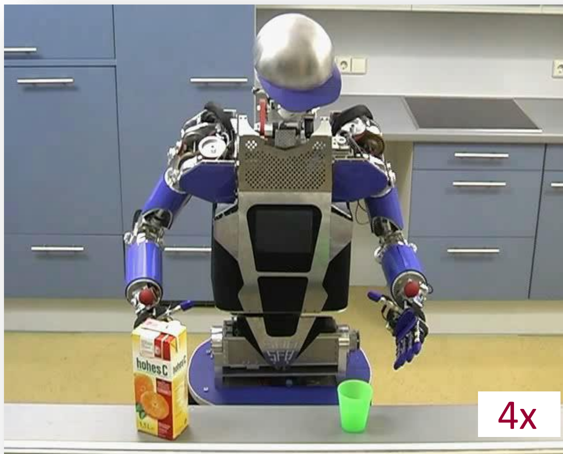
- Object recognition and localization
- Vision-based grasping
- Hybrid position/force control
- Combining force and vision for opening and closing door tasks
- Collision-free navigation
- Vision-based self-localisation
- Multimodal human-robot dialogs
- Continuous speech recognition
- Learning new objects, persons and words
- Audio-visual tracking and localization
- …

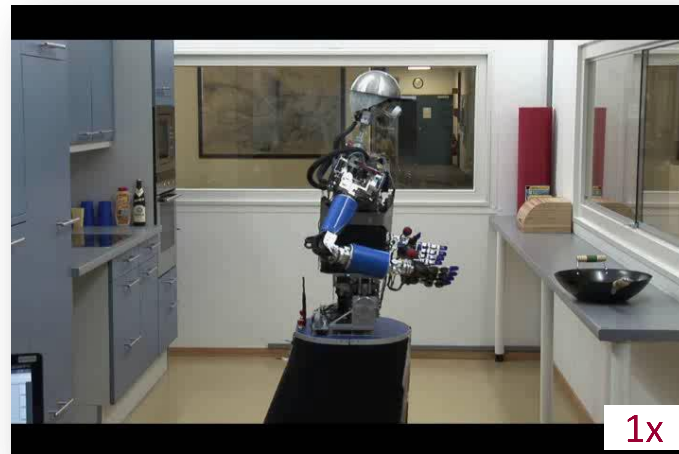
ARMAR-III in the RoboKITchen

- First step towards 24/7
  - 45 minutes demonstration
  - Shown more than 1000 times, since 03. February 2008, to experts and public
    - 75 times in 5 days for approx. 5000 visitors at CeBIT 2012
    - 50 times during the ICRA 2013 and EFFEKTE weekend, 2013 in Karlsruhe

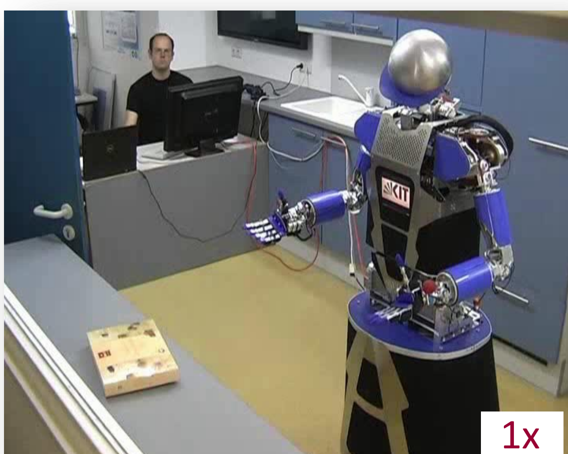
Advanced grasping capabilities

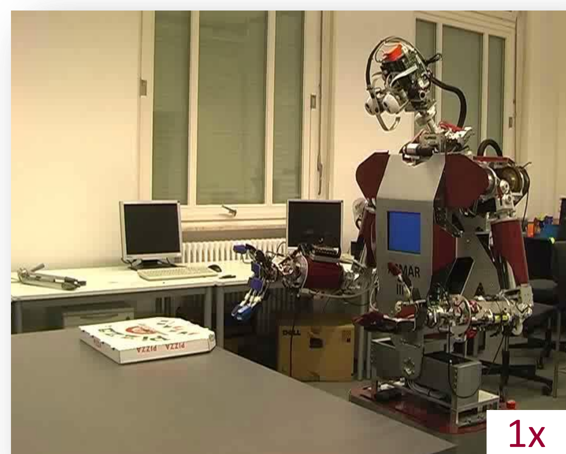
- Bimanual grasping and manipulation





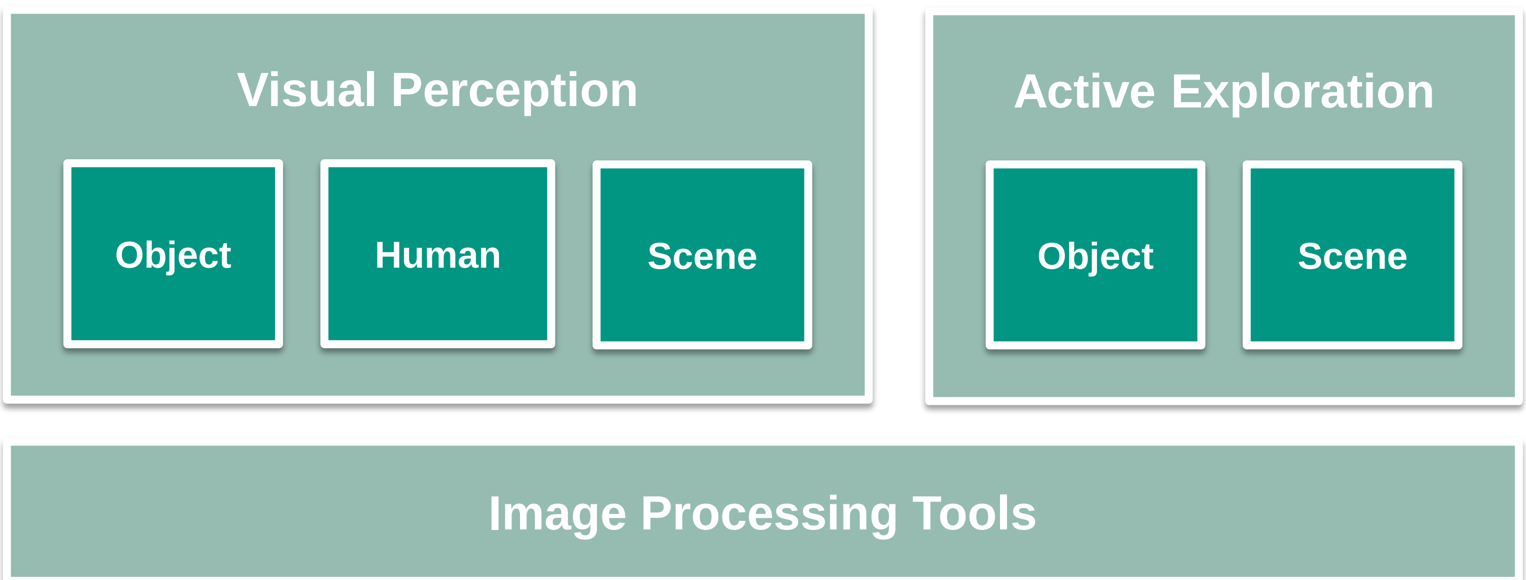
- Pre-grasp manipulation





RSJ 2013,
RAM 2012
IROS 2011
Humanoids 2010
Humanoids 2009
RAS 2008

Visual Perception and Active Exploration



Visual Perception and Active Exploration

Visual Perception

- Object
- Human
- Scene

Active Exploration

- Object
- Scene

Image Processing Tools

Object recognition and localization

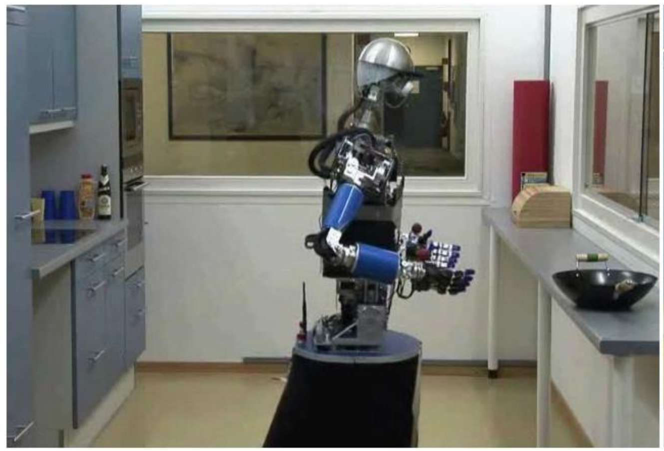
- Colored objects
  (Azad et al., 2008; 2009)
  - Segmentation by color
  - Appearance-based recognition using a global approach
  - Combination of stereo vision and stored orientation information for 6D pose estimation

- Textured objects
  (Azad et al., 2006; 2009)
  - Recognition using local features
  - 2D-localization using image point correspondences
  - 6D pose estimation using stereo vision

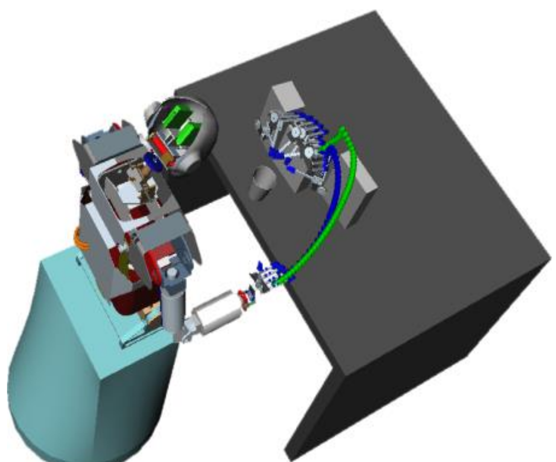
Correspondences between learned view and view in scene

Object grasping and manipulation (I)

- Visual Servoing
  (Vahrenkamp et al., 2008; 2009, Asfour et al. 2008, 2013)

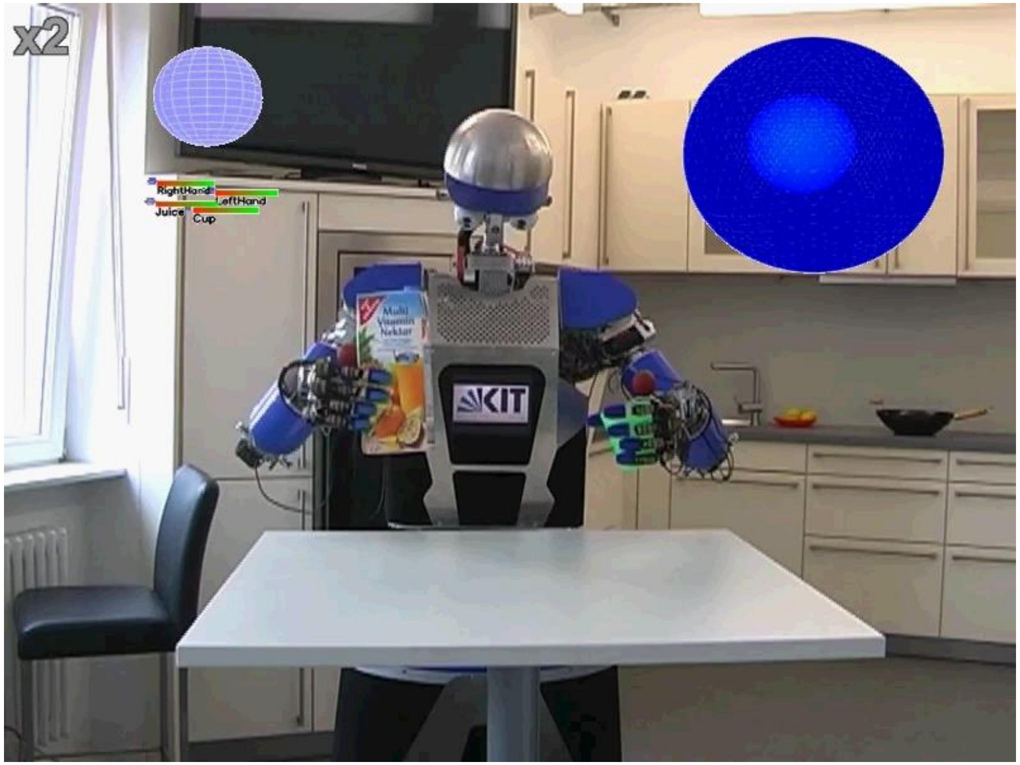


- Visually guided execution of planned tasks
  (Vahrenkamp et al., 2009)



Object grasping and manipulation (II)

- Gaze selection during manipulation
  (Welke et al., 2013)
  - Observe multiple objects during continuous tasks
  - Reduces localization uncertainty



Visual Perception and Active Exploration

Karlsruhe Institute of Technology

Visual Perception

- Object
- Human
- Scene

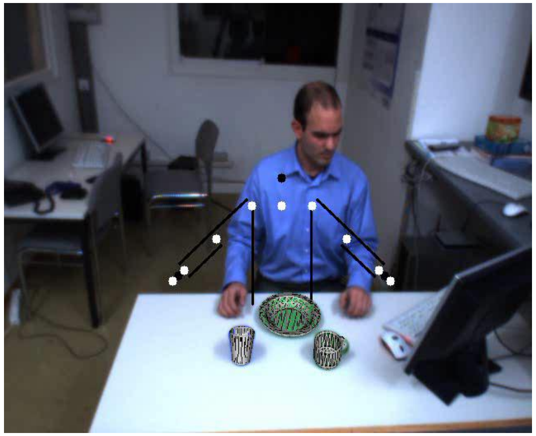
Active Exploration

- Object
- Scene

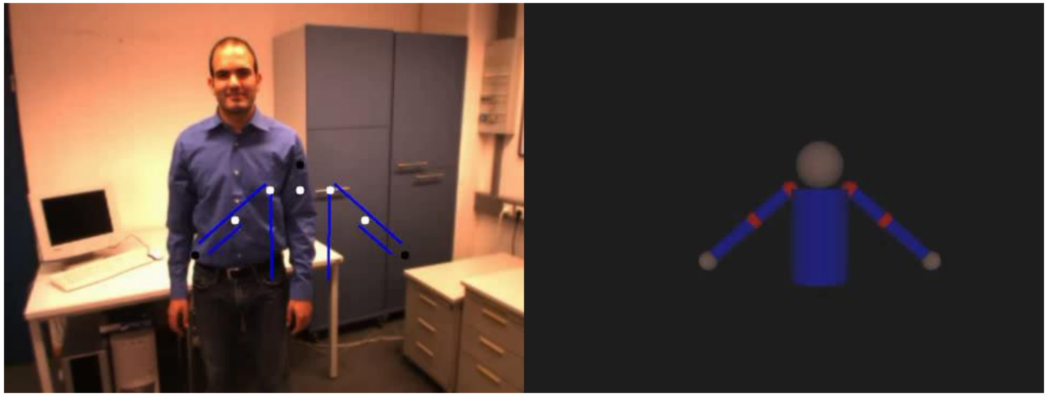
Image Processing Tools

Human Observation (I)

- Stereo-based 3D Human Motion Capture
  (Azad et al., 2008)
  - Hierarchical Particle Filter framework
  - Localization of hands and head using color segmentation and stereo triangulation
  - Fusion of 3d positions and edge information
  - Half of the particles are sampled using inverse kinematics

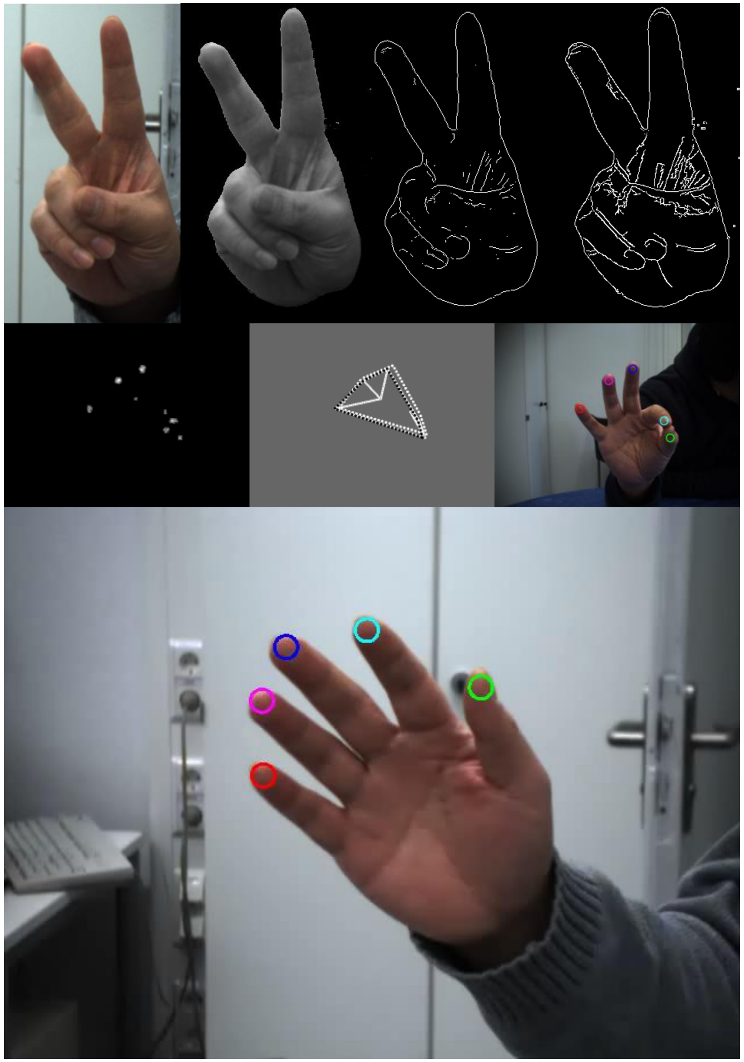


- Features
  - Automatic Initialization
  - 30 fps real-time tracking on a 3 GHz CPU, 640x480 images
  - Smooth tracking of real 3d motion



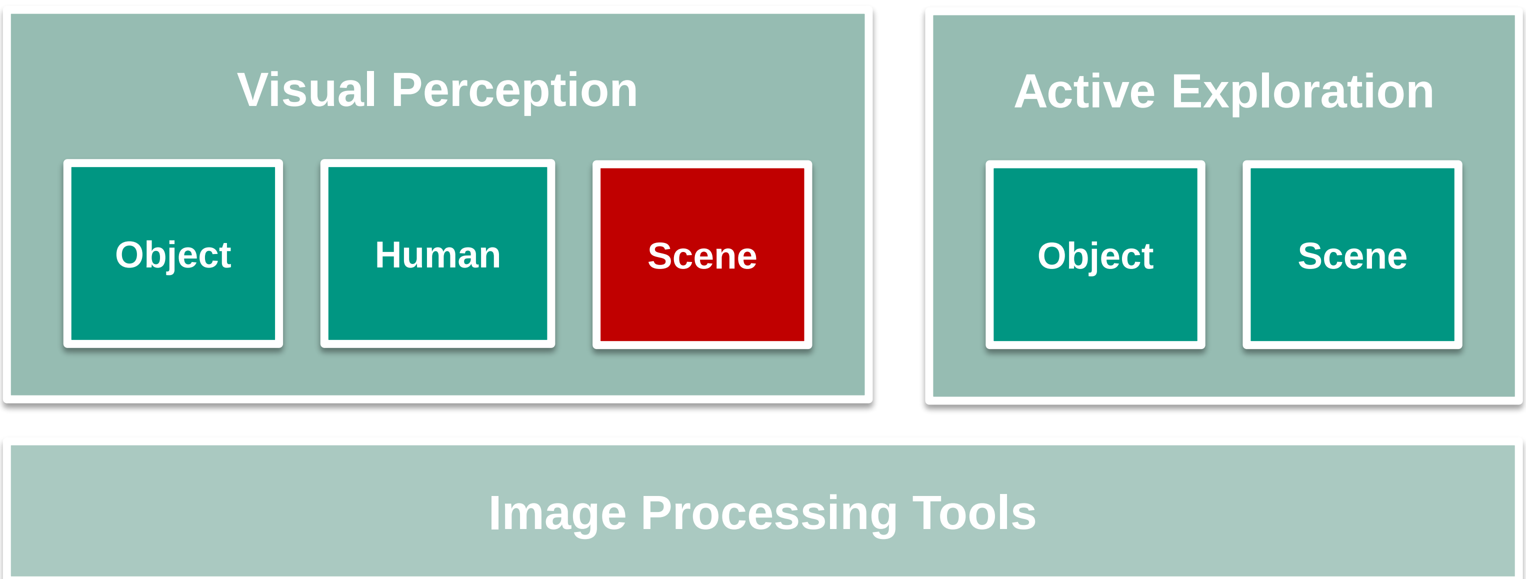
Human Observation (II)

- Markerless fingertip tracking
  (Do et al., 2011)
  - Edge map with multi-scale approach
  - Based on circular image features using Hough Transform
  - Tracking of a deformable contour using particle filter
  - Position correction with Mean Shift and radius adaption



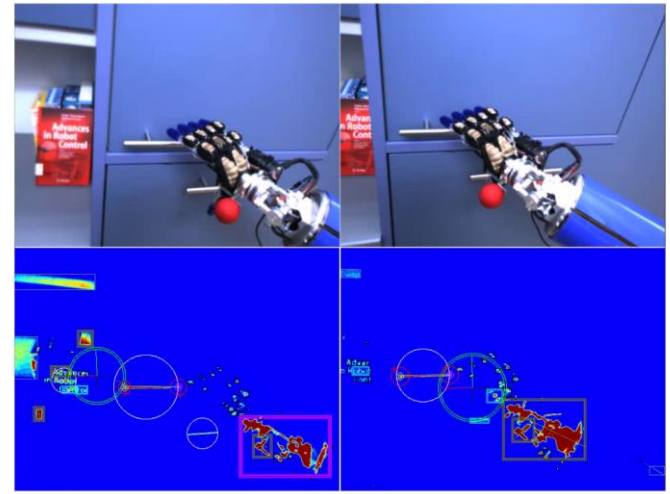
- Features
  - Automatic Initialization
  - 25 fps real-time tracking on a 3 GHz CPU, 640x480 images

Visual Perception and Active Exploration

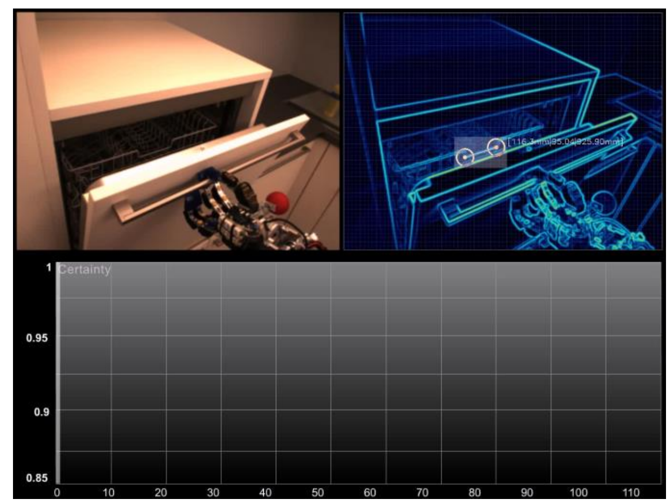


Environmental object grasping and manipulation

- Perception for environmental interaction
  (Wieland et al., 2009, Gonzalez et al., 2010; 2011)
  - Recognition of handles, doors, electric appliances, furniture
  - Based on computer aided geometric model
  - Combination of edge extraction and color segmentation



- Eccentricity Edge Graphs for Cluttered Object Recognition
  (Gonzalez et al., 2010)
  - CAD model-based approach
  - Extraction of geometric primitive

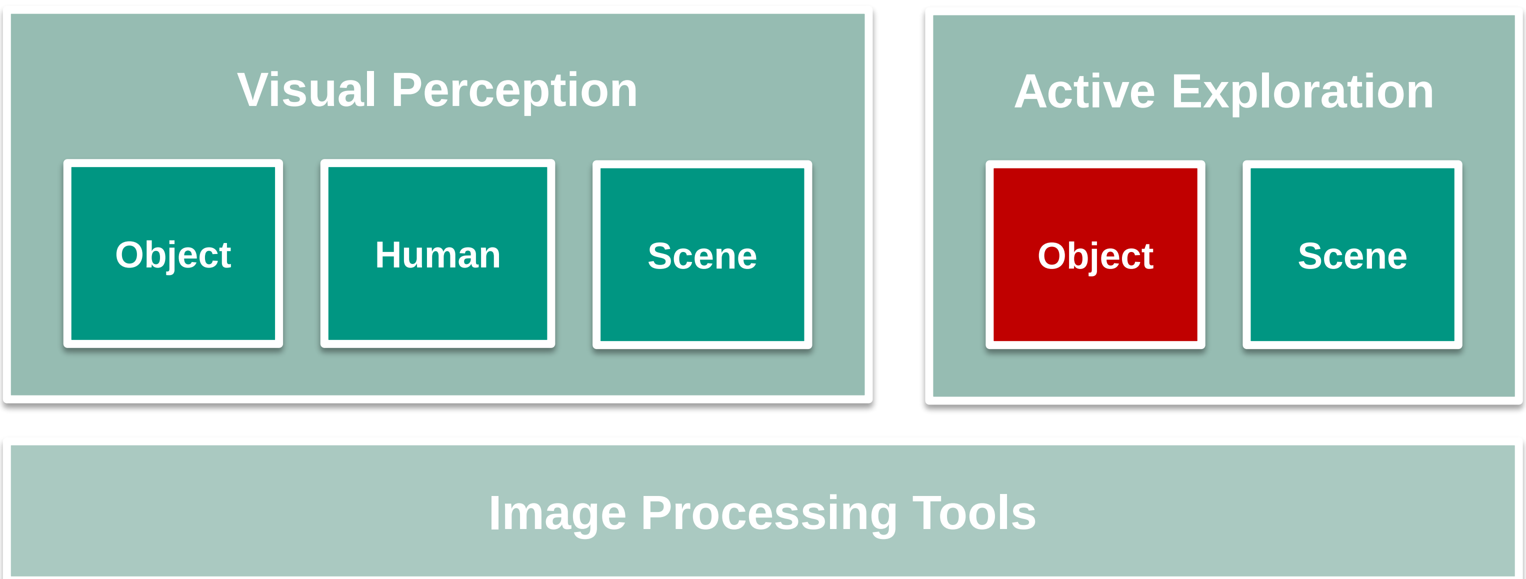


Self Localization

- Global and dynamic self localization
  (Gonzalez et al., 2008; 2009, 2010, 2012, 2014)
  - Complexity reduction based using visibility analysis
  - Pose estimation based on:
    - Sphere intersection
    - Particle-filter

Tuesday 11:10-11:30, (Paper TuB10.1, S228)
David Gonzalez, Michael Vollert, Tamim Asfour and Rüdiger Dillmann.
Robust Real-Time 6D Active Visual Localization for Humanoid Robots

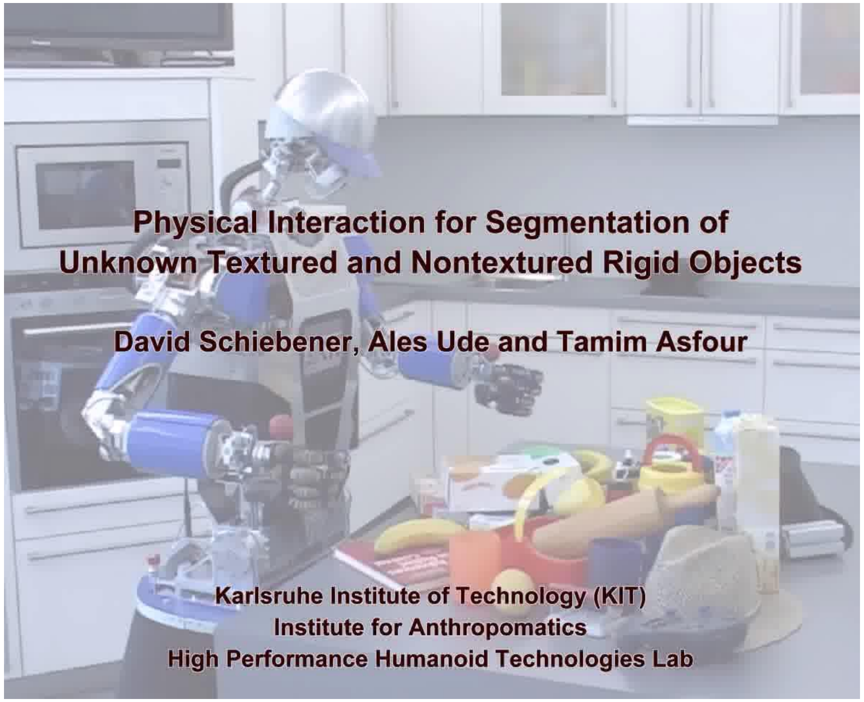
Visual Perception and Active Exploration



Active object exploration (I)

Karlsruhe Institute of Technology

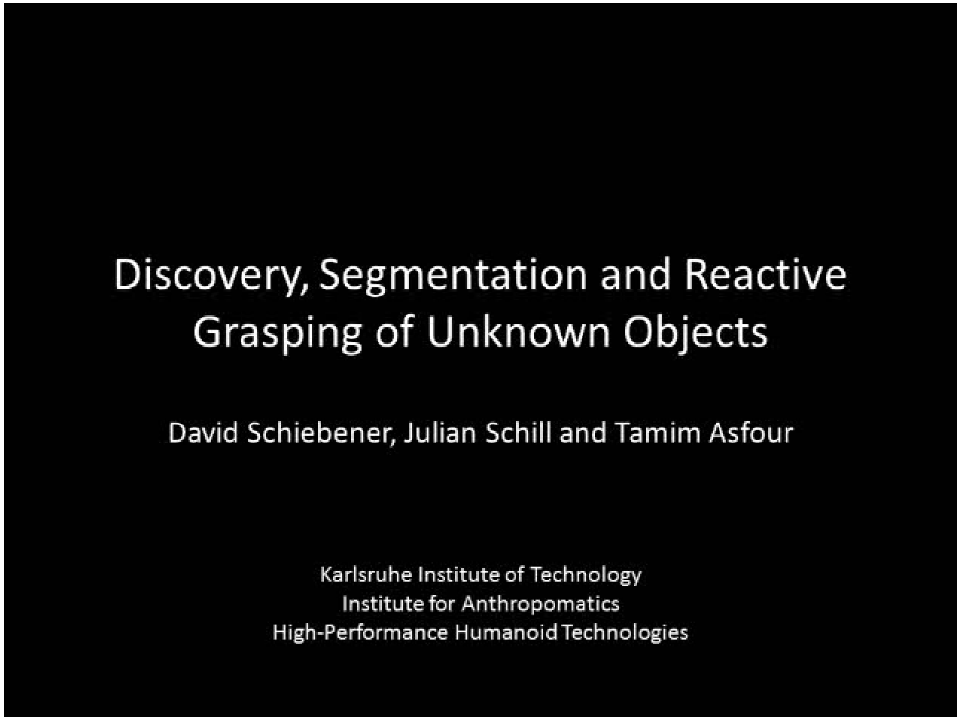
- Interactive object segmentation
  (Schiebener et al., 2011, 2012, 2013, 2014)
  - Interact with and learn about unknown objects
  - Physical interaction to support vision
  - Object segmentation using rigid body constraint



Wednesday 11:50-12:10 (Paper WeB02.3, Theatre 2)
David Schiebener, Ales Ude and Tamim Asfour
Physical Interaction for Segmentation of Unknown Textured and Non-Textured Rigid Objects

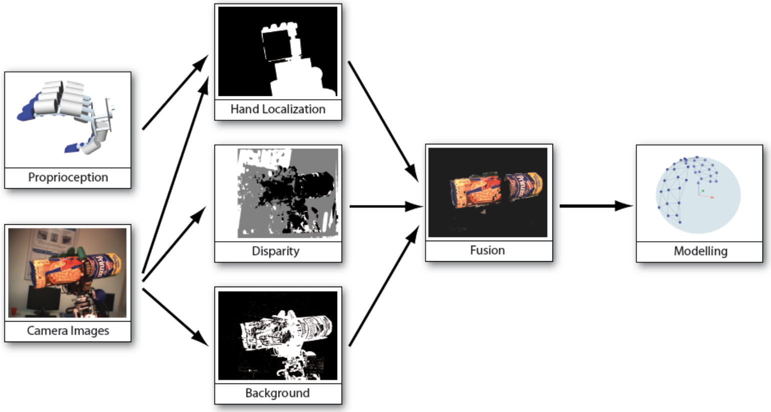
Active object exploration (II)

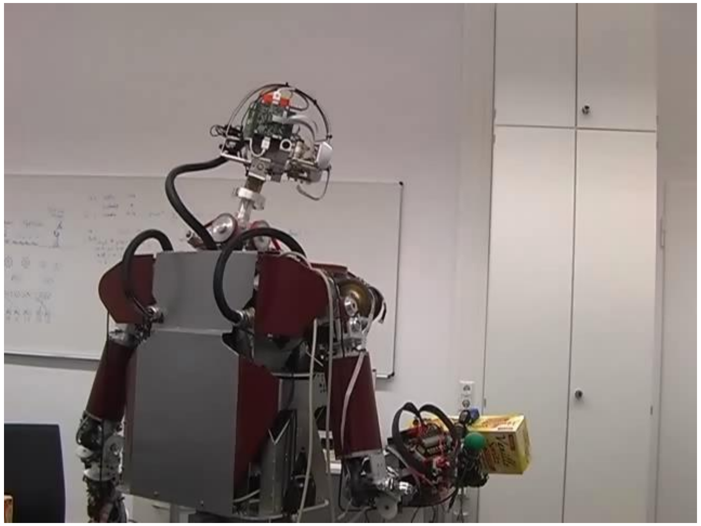
- Segmentation and reactive grasping
  (Schiebener et al., 2012)
  - Use hypotheses from interactive object segmentation
  - Reactive grasping based on tactile information
  - Corrective movements on failure

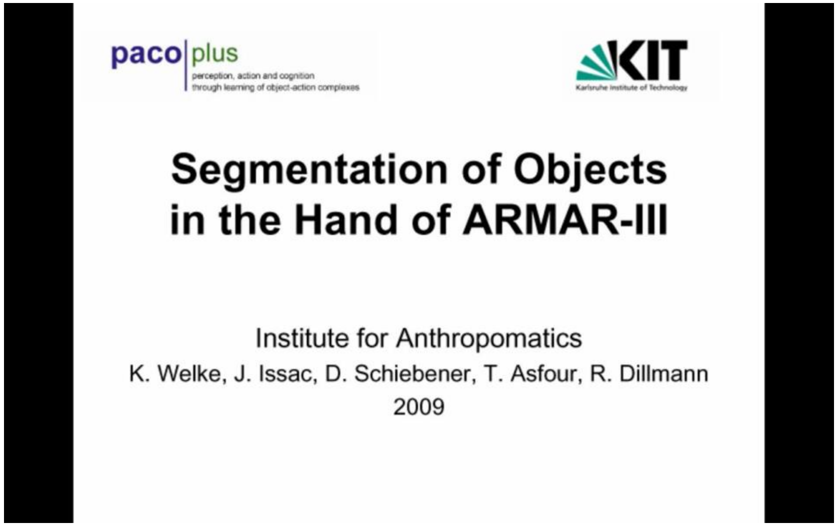


Active object exploration (III)

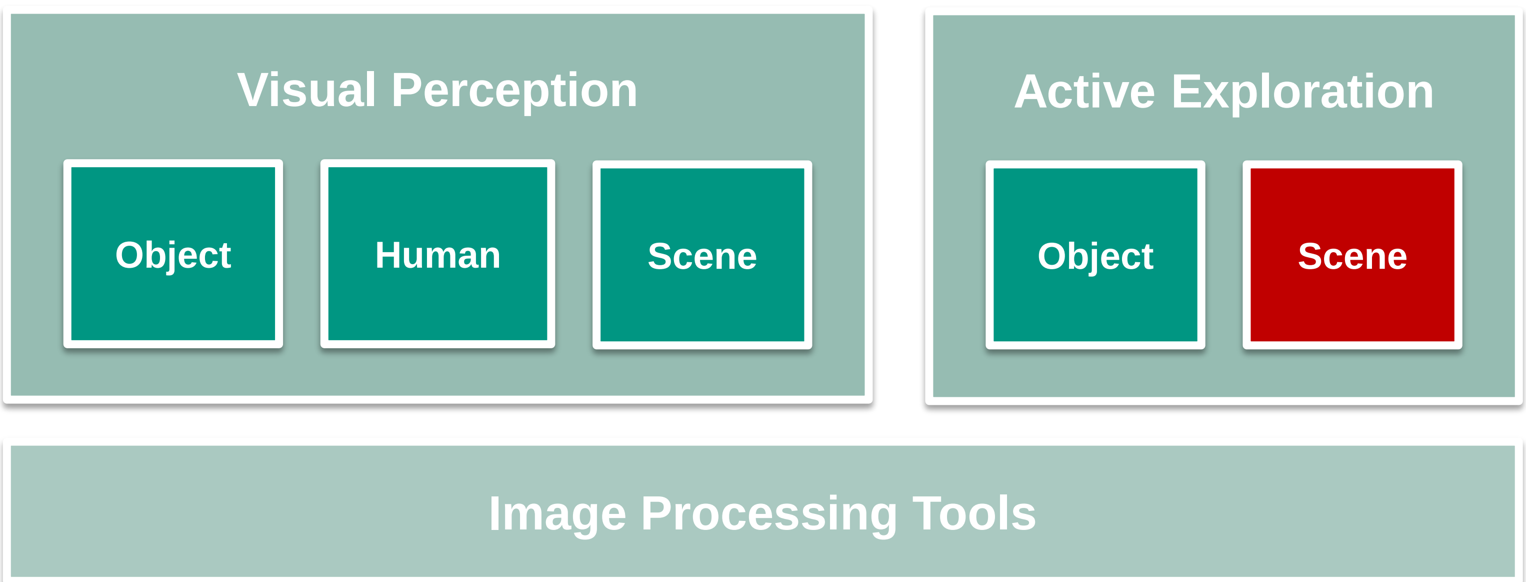
- Generation of multi-view representations
  (Welke et al., 2009, 2010)
  - Build model of objects in the hand using vision and prorioception
  - Segmentation of background, hand and object
  - Aspect graph as multi-view representation



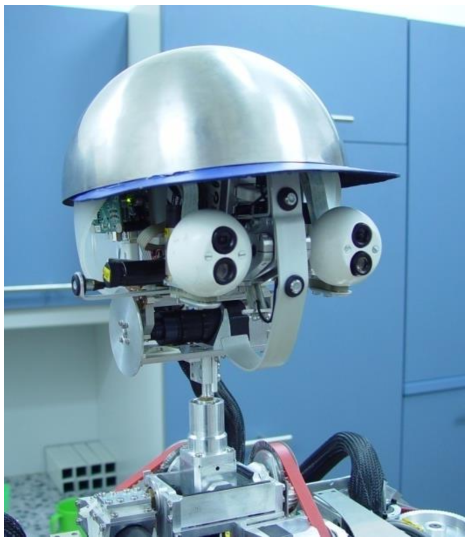




Visual Perception and Active Exploration



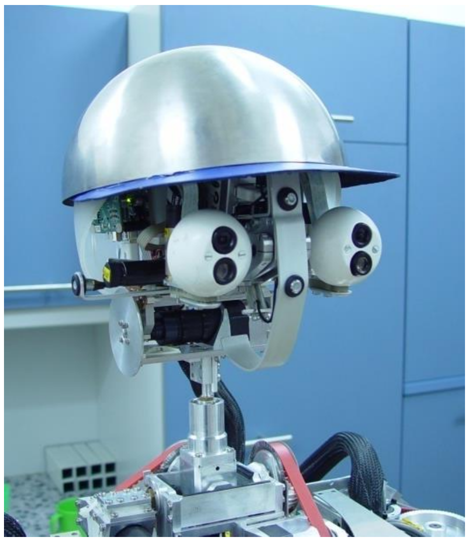
Active Visual Search



peripheral view

foveal view

Active Visual Search



peripheral view

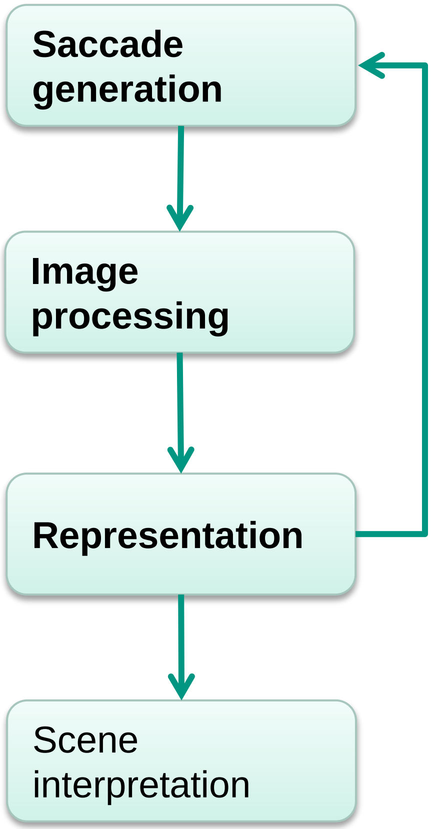
foveal view

- Tasks

Saccade generation → **Image processing** → **Representation** → Scene interpretation

Active Visual Search and Representation

- Active visual search
  - Search for known target object
  - Generation of saccadic eye movements
  - Object detection and recognition
- Representation
  - Transsaccadic memory
  - Perception as continuous process

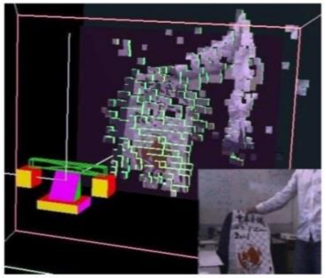


Kai Welke "Memory-Based Active Visual Search for Humanoid Robots", phd thesis, KIT, 2011

Related Work

- Foveal Vision
  - Search and pursuit using signatures *[Ude et al., 2003]*
  - Search based on depth information *[Bjorkman and Kragic, 2004]*
  - Bottom-up saliency and wieghts *[Rasolzadeh et al., 2010]*
  - Saliency based on color *[Orabona et al., 2005]*
- Representations
  - Occupancy Grid (3D) *[Dankers et al., 2009]*
  - Sensory Egosphere (2D) *[Figueira et al., 2009]*

[Ude et al., 2003]



[Dankers et al., 2009]

[Figueira et al., 2009]

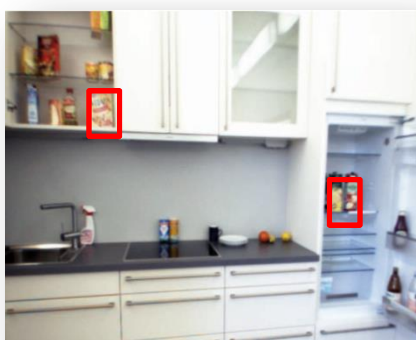
No integration of active visual search and representation.

Active Visual Search

- Complexity of visual search
  - General visual search problem: NP-complete

- Approach
  - Knowledge of the target object model: linearer complexity
  - Decomposition of the problem:

Coarse analyse → Validation



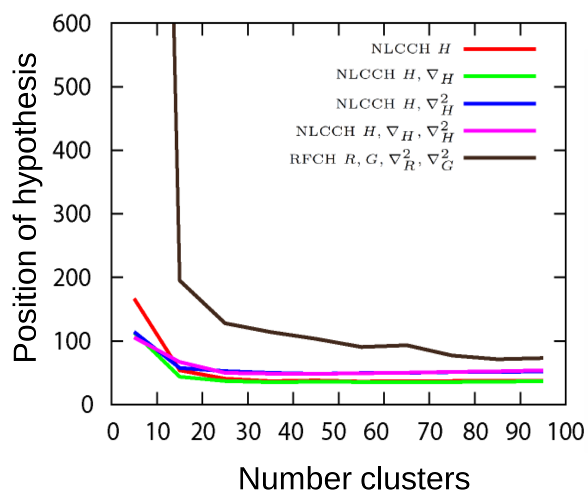
Peripheral view

Foveal view

Object search in the peripheral view

- **Goal:** Restriction of the search space
- **Approach**
  - Coarse analysis of the scene in peripheral view
  - Detection of object candidates

- **Methods**
  - Color Cooccurrence Histograms (CCH)
  - Search window for object candidate detection



Object recognition in the foveal view

- **Goal:** Validation of object candidates
  - Foveal view allows for detailed analysis
  - Elimination of false positive object candidates

➡ **Object recognition**
  - Texture-based recognition based on Harris-SIFT features *[Azad et al., 2008]*
  - Calculation of feature correspondences with object model
  - Classification of object candidates

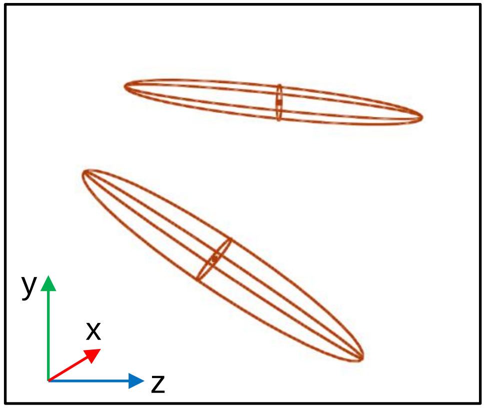
Saccade generation

- **Goal**
  - Minimal number of saccades until object recognition

  ➡ Gaze direction with maximum probability of recognition
- **Approach**
  - Saliency based on the Bayesian Strategy *[Torralba, 2003]*

$$p(O=1,X|F)=\frac{1}{p(F)}\cdot\underbrace{p(F|O=1,X)}_{\text{Object model}}\cdot p(X|O=1)\cdot p(O=1)$$

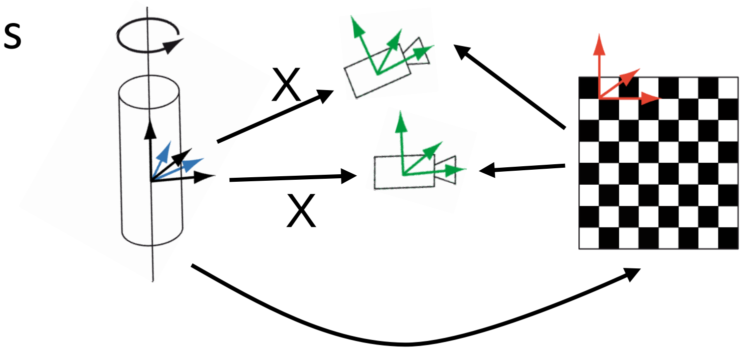
- **Representation of saliency**
  - Landmark-based map of candidates
    - Localization uncertainty
    - Probability of existence
  - Approximates $p(O=1,X|F)$



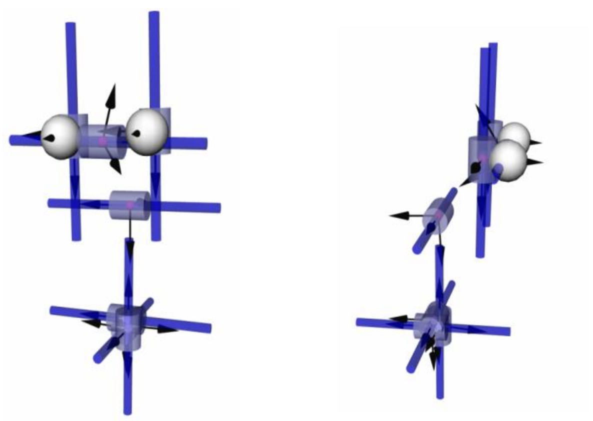
Localization uncerainty for 2 candidates

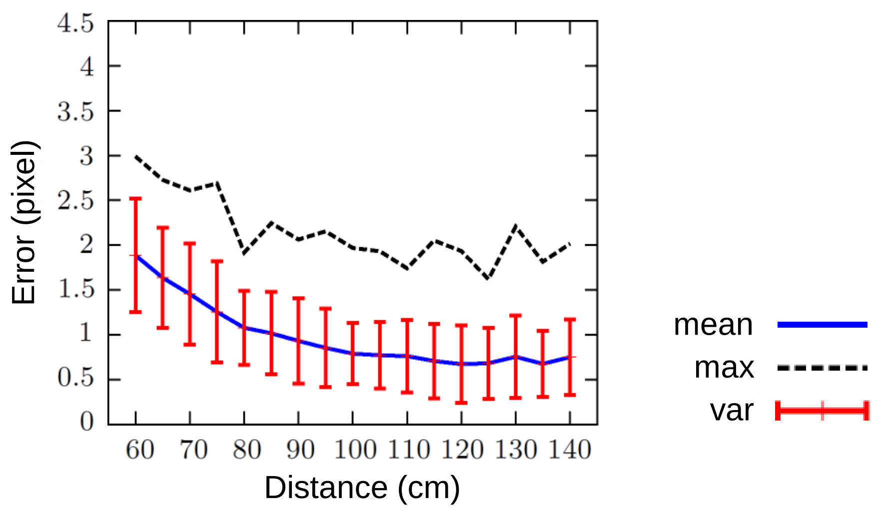
Execution of saccades

- **Kinematic model for saccade execution**
  - Pose of the camera coordinate systems unknown
  - Inaccuracies in CAD model



➡ Kinematic Calibration
  - Visual aided
  - Calibration of all joints



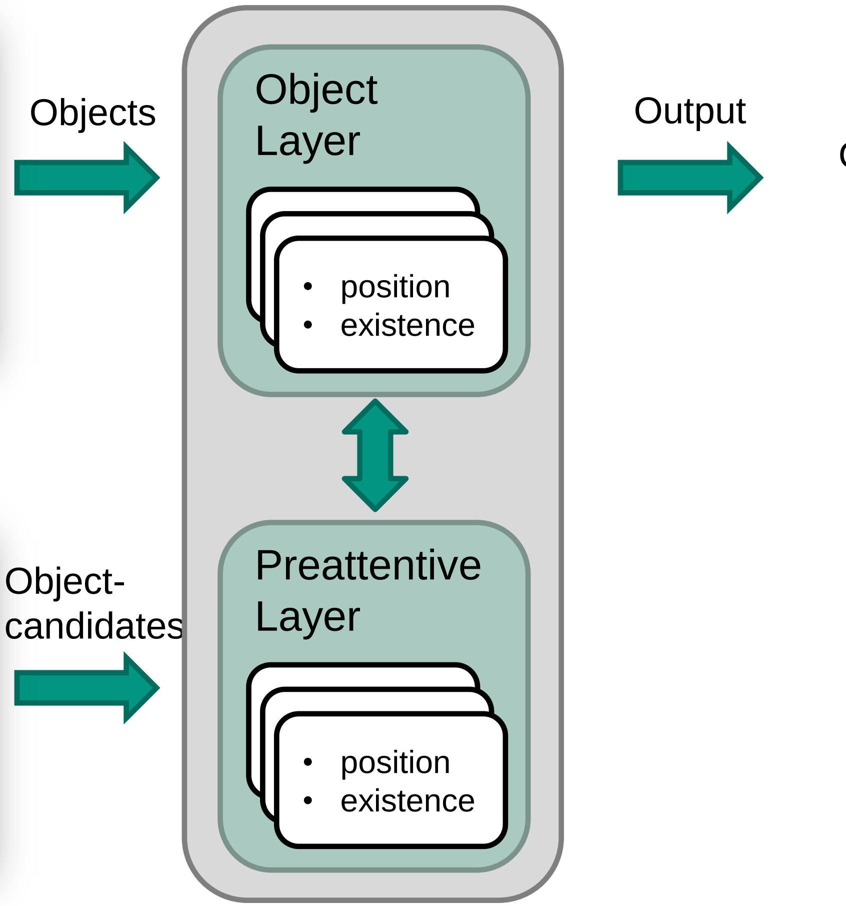


Transsaccadic Memory

Foveal camera image

Peripheral camera image

Transsaccadic Memory



Transsaccadic Memory – Update

- Update of the Preattentive Layer

- Update of the Object Layer

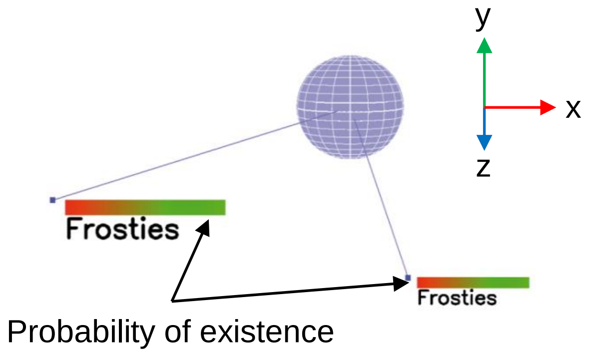
- Consistency of scene and memory

Update of the Object Layer

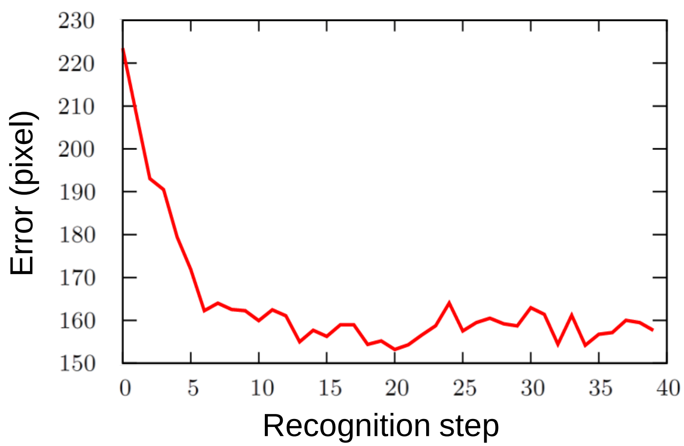
- **Prerequisite**
  - Object candidate fixated in foveal cameras

  ➡ Correspondence solved

- **Update of object existence**
  - Match probability
  - Update using Bayes Filter



- **Update of object position**
  - Closed loop
  - 2D position error in left and right camera



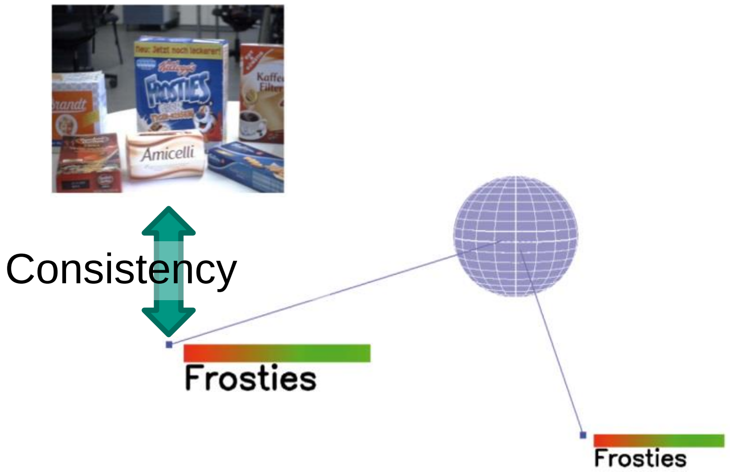
Memory and Saccade Generation (I)

- **Requirement**

  Consistency of scene and memory

  - For each object instance a corresponding representation exists in memory
  - For each representation in memory a corresponding object instance exists

- **Approach**
  - Consistency is assured using foveal validation



Memory and Saccade Generation (II)

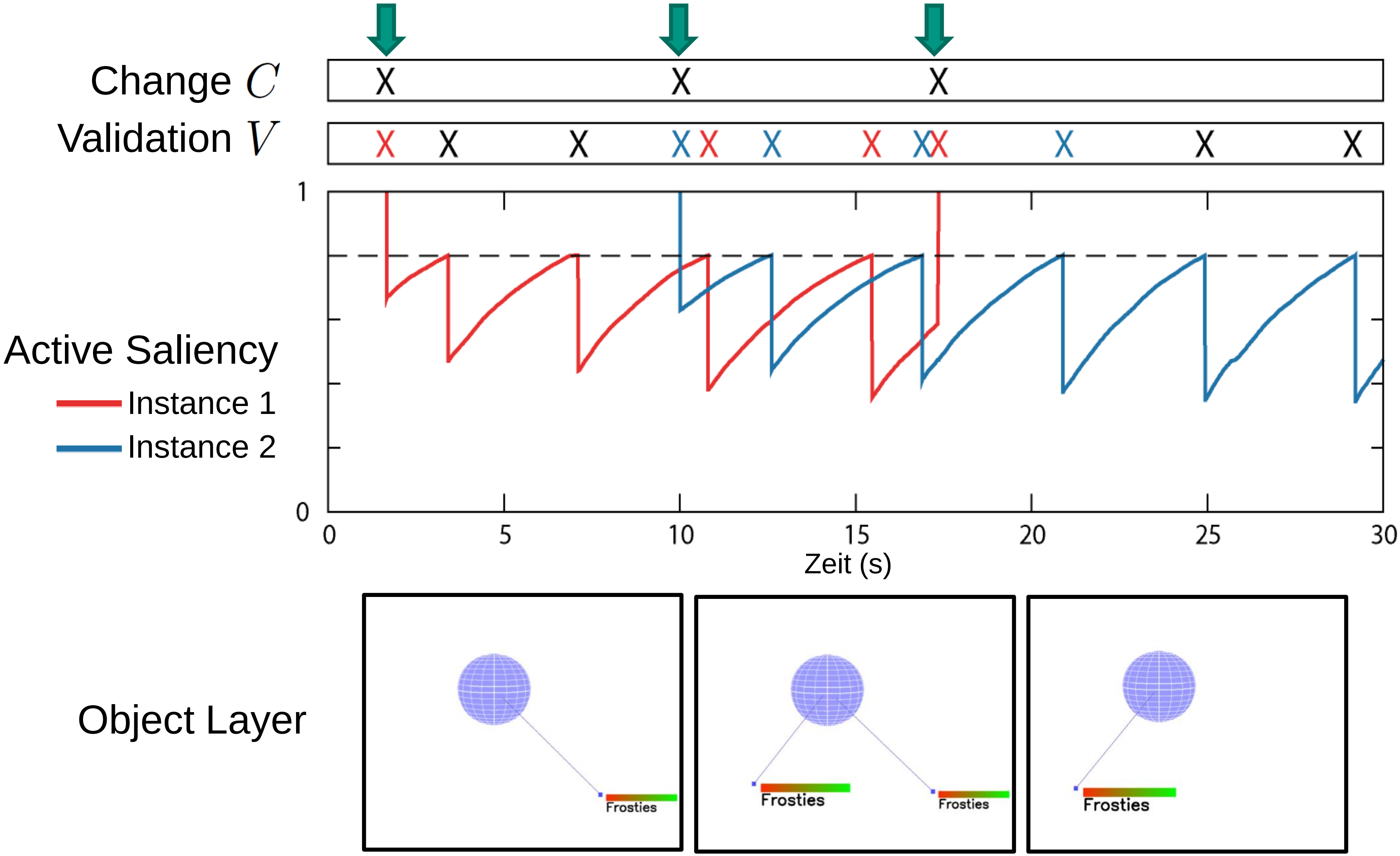
- **Consequences for Saccade Generation**
  - Account for consistency of Object Layer
  - Gaze directions towards inconsistent memory entities

- Inconsistency I depends on
  - Validation using foveal object recognition V
  - Change of the world C

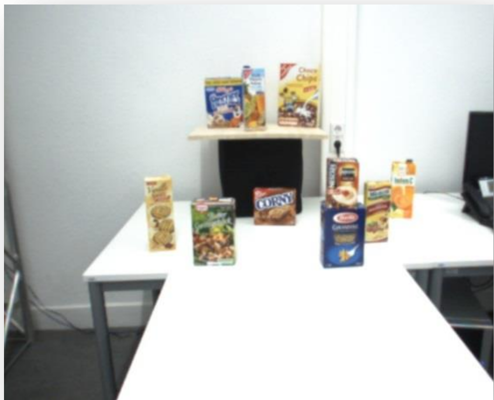
- **Active Saliency**

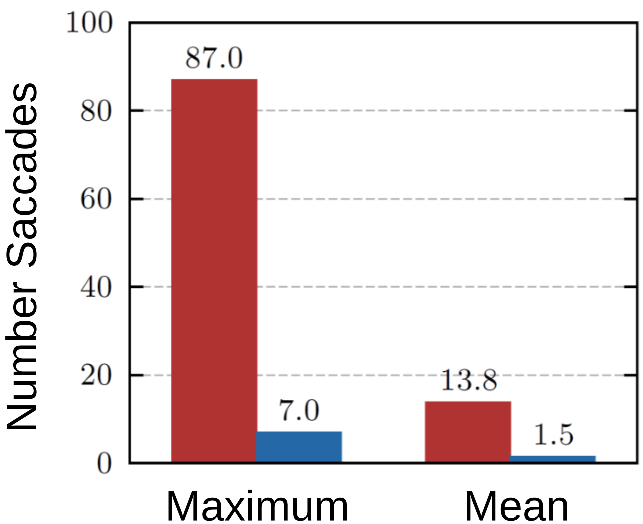
$$
\begin{aligned}
s_a &= p(O=1, X, I=1|Z) \\
&= \underbrace{p(O=1, X|F)}_{\text{Bayesian Strategy}} p(I=1|C,V)
\end{aligned}
$$

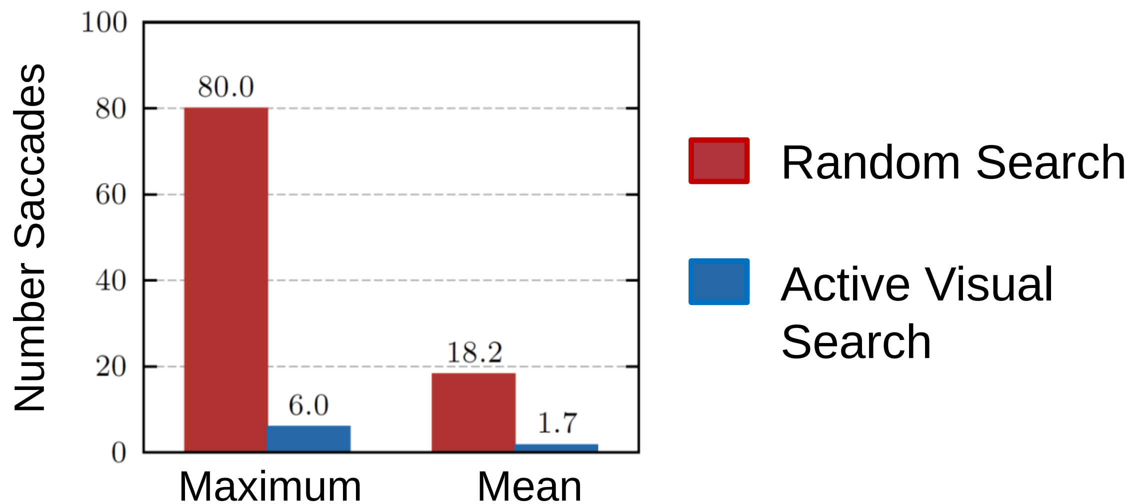
Active Saliency: Example



Active Visual Search: 10 objects in 20 scenes

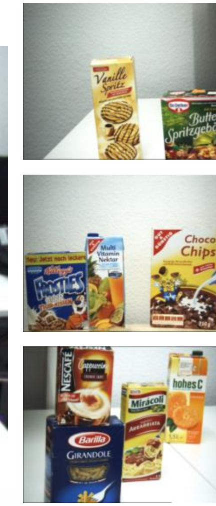






Active scene exploration

- **Active visual search**
  (Welke et al., 2009; 2011)
  - Analyze scene exploiting active foveal camera system
  - Build consistent scene representation
  - Continuous perception in changing environments



Conclusions

- Integrated results on visual perception for humanoids in real world scenarios

- Active vision difficult but promising

Thanks to ...

- German Research Foundation (DFG)
  - SPP 1527 autonomous-learning.org (2010 -)
  - SFB/TR 89 www.invasic.de (2009 -)
  - SFB 588 www.sfb588.uni-karlsruhe.de (2001 - 2012)

- European Commission
  - Xperience www.xperience.org (2012-2015)
  - Walk-Man www.walk-man.eu (2013-2017)
  - KoroiBot www.koroibot.eu (2013-2016)
  - GRASP www.grasp-project.eu (2008-2012)
  - PACO-PLUS www.paco-plus.org (2006-2011)

- Karlsruhe Institute of Technology (KIT)
  - Professorship “Humanoid Robotic Systems”
  - Heidelberg-Karlsruhe Research Partnership (HEiKA)

Thanks for your attention