Scene Representation and Object Grasping Using Active Vision

Xavi Gratal Jeannette Bohg Mårten Björkman Danica Kragic

Centre for Autonomous Systems
Computer Vision and Active Perception Laboratory
KTH, Stockholm, Sweden

Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics, International Conference on Intelligent Robots and Systems 2010, Taipeh, Taiwan

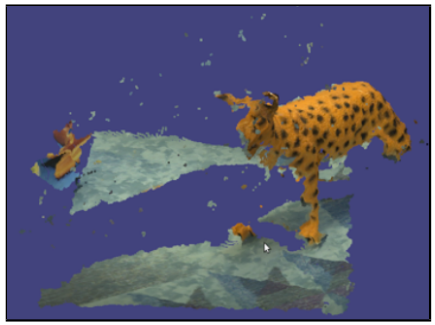
Introduction

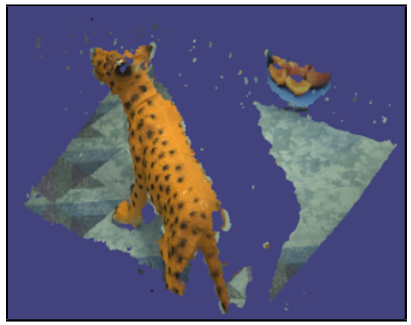
- Large research field:
  Object-Centered Grasp Inference
  $\longrightarrow$ affordances
- Given an Object o, which grasp g can be applied to it?

$$g = f(o) \qquad (1)$$

- o: known, unknown or familiar
- function f: stable grasp, imitation, task specific, heuristics, etc.

?





Questions in Grasp Inference and Executation

- How do we detect object hypotheses o ?
  - foreground/background segmentation $\rightarrow$ occlusions, viewing constraints
  - recognition & pose estimation
  - categorisation
- How to constrain f to obtain a grasp g feasible in a whole scene? $\rightarrow$ collision avoidance
- Given a task, how to plan a whole sequence of manipulation tasks?
  - Prepare the dinner table!
  - Pour me a cup of coffee!
  - Clean the table!
  - Unload the dishwasher!
- **Robot needs to understand the scene it is facing!**

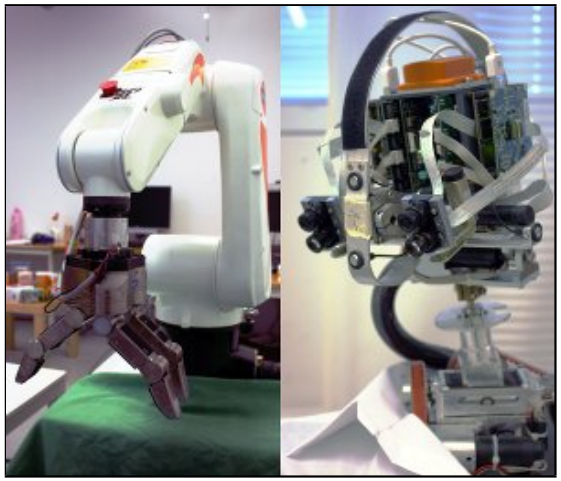
What do we mean by scene?

- Embodiment
- Obstacles (furnitures, supporting planes, ...)
- Object Hypotheses

Our Approach Towards Scene Understanding

Active Vision System integrating different **computational processes** for **incremental** scene understanding

- Hardware:
  Armar III Robotic Head (7 DoF), 4 Cameras
  Kuka Arm (6 DoF)
  Schunk 3-fingered Dexterous Hand (7 DoF)
- Scene:
  one supporting plane
  several object hypotheses



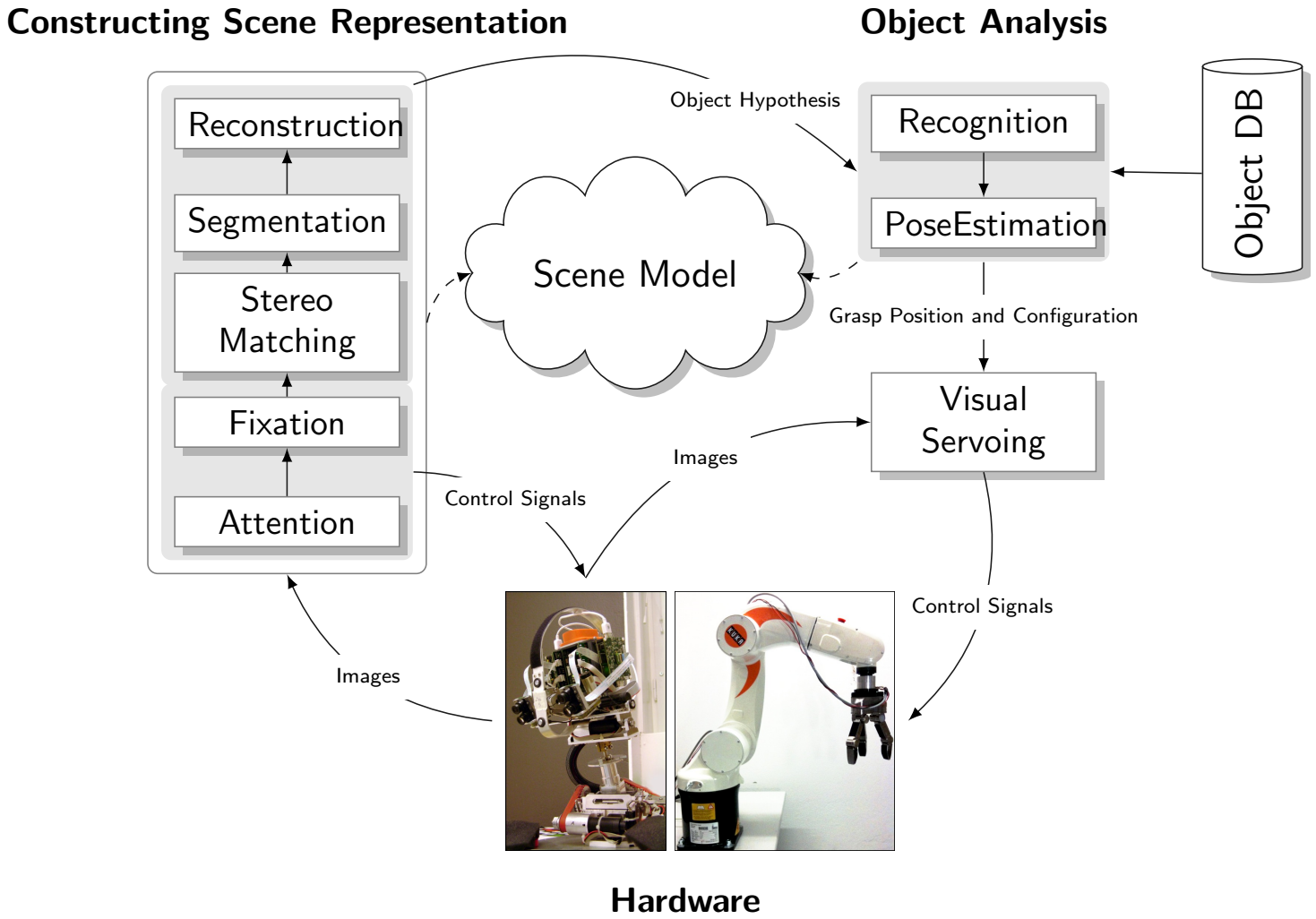
Why do we need four cameras?

- Wide field of view - scene oriented
  - Easier to find objects and their relations
  - Suitable for attention
- Narrow field of view - object oriented
  - Easier to analyze objects and perform learning
  - Suitable for recognition/categorization

Computational processes involved

Task: Grasping Known Objects

Building up scene model independent of object knowledge



Offline Calibration

- Neccessary for:
  - Stereo Calibration for image rectification prior to stereo matching
  - Head-Eye Calibration to bring an attention point in wide field cameras to center of foveal cameras
  - Hand-eye calibration for visual servoing
- Transformations to determine
  1. Between left & right camera for fixed joint configuration
  2. Between one camera system in two different joint configurations
  3. Between camera and arm

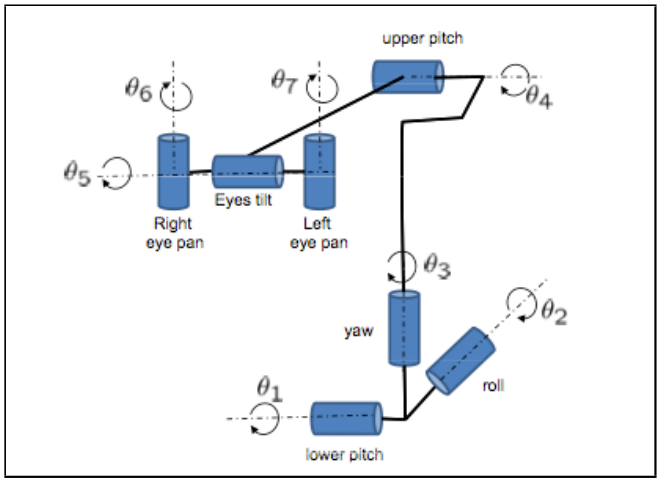
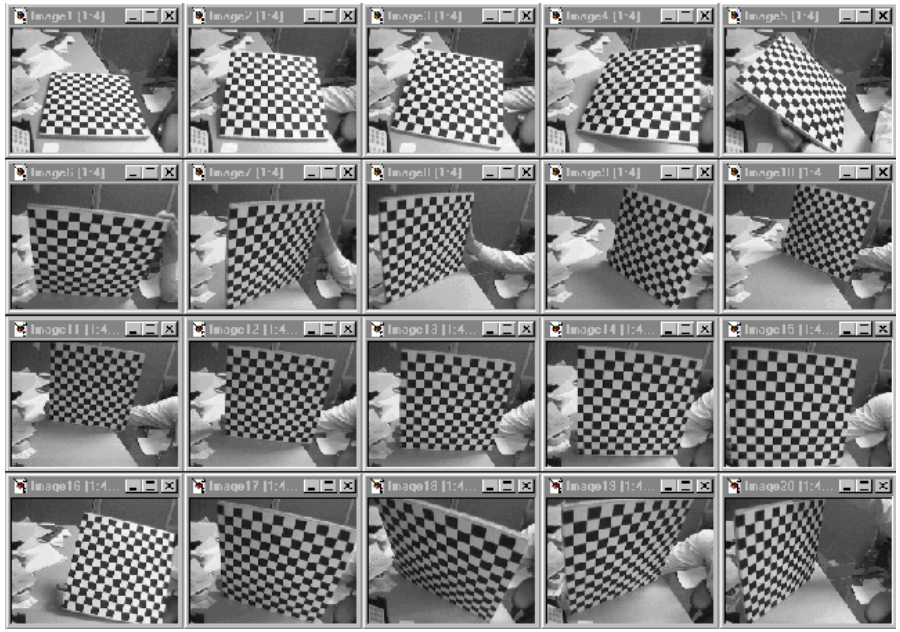


Figure: Armar III Head Kinematic Chain

Stereo Calibration - Classic

- checkboard pattern $\longrightarrow$ own coordinate system
- Each transformation between checkboard and left or right camera system determinable $\longrightarrow$ transformation between left and right camera



Stereo Calibration - Our Approach

- Kuka arm with high precision
- Tracking of LED attached to end effector
- Advantages:
  1. Camera-to-Arm transformation for free
  2. No restriction of point positioning $\longrightarrow$ Checkerboard has to be visible for both cameras
  3. Pattern uniform in image space

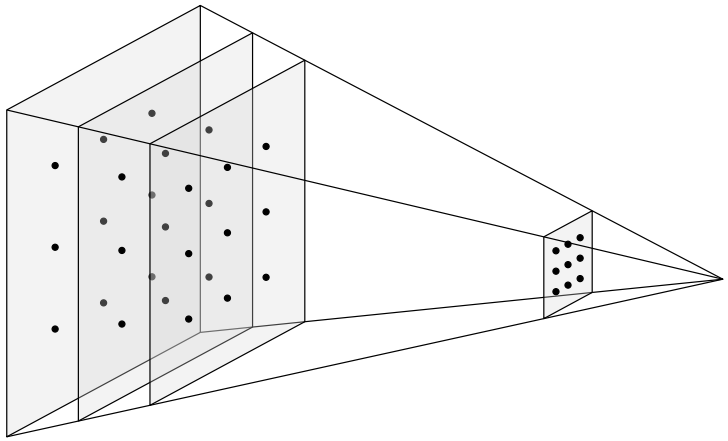


Figure: Movement pattern of end effector.

Stereo Calibration - Example Video

See `www.csc.kth.se/~bohg/calibVideo.mp4`

Head-Eye Calibration

- Head with 7 DoF $\longrightarrow$ 3 change epipolar geometry, 4 change pose of cameras relative to neck
- Mechanical inaccuracies affect center and axis of rotation
- repeatability issues remain $\longrightarrow$ online calibration & visual servoing

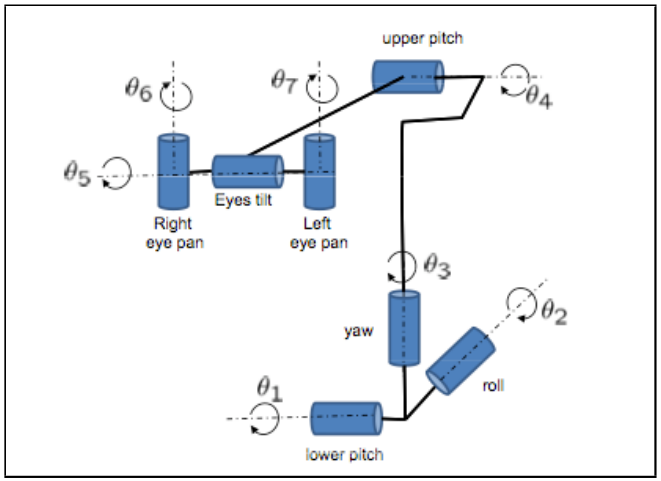


Figure: Armar III Head Kinematic Chain

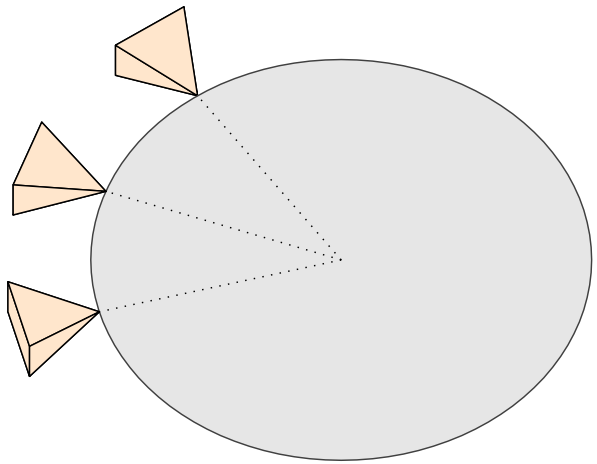
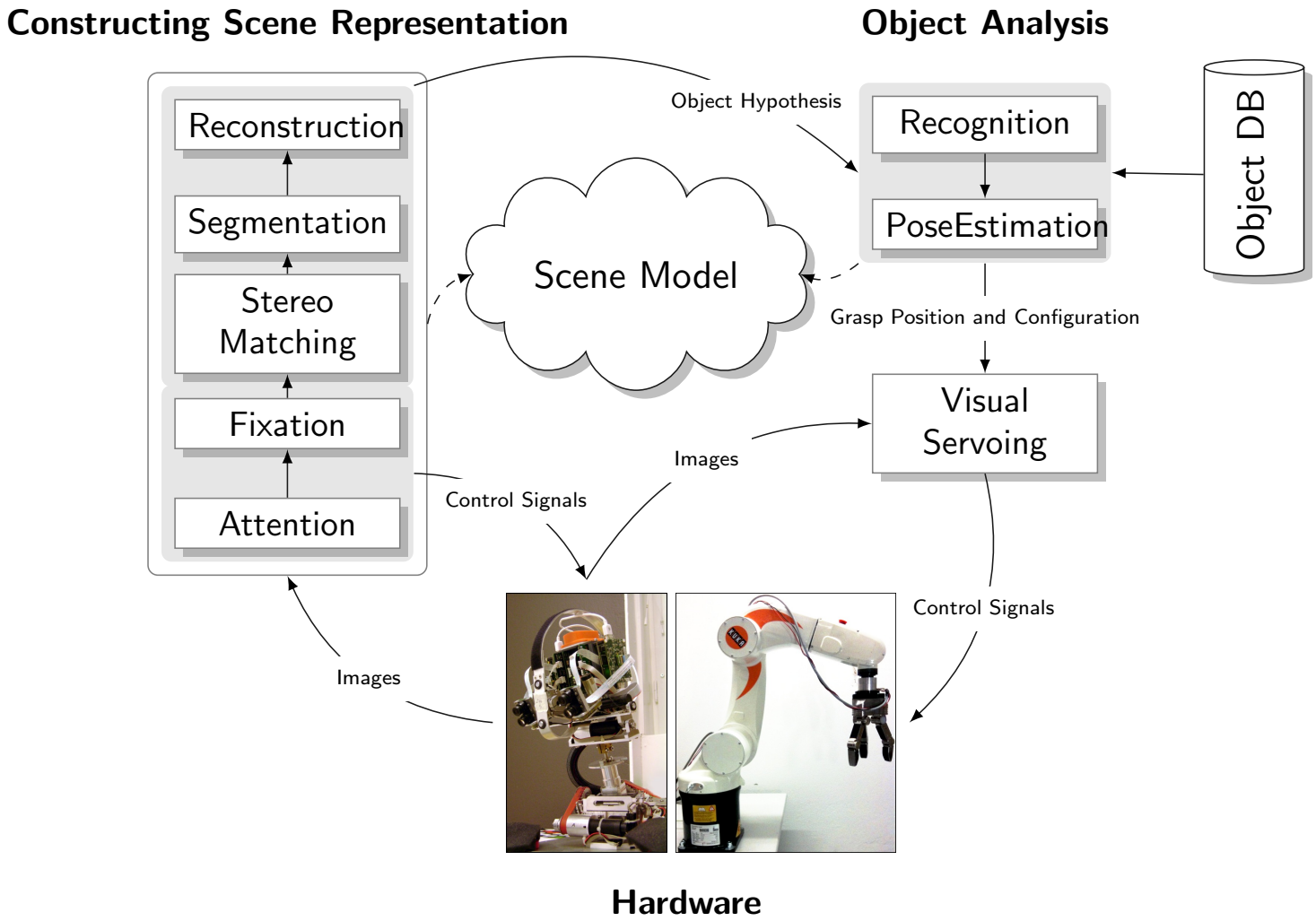


Figure: Three viewing directions of a camera rotated around one axis.

Computational processes involved

Task: Grasping Known Objects

Building up scene model independent of object knowledge



Attention for Scene Search

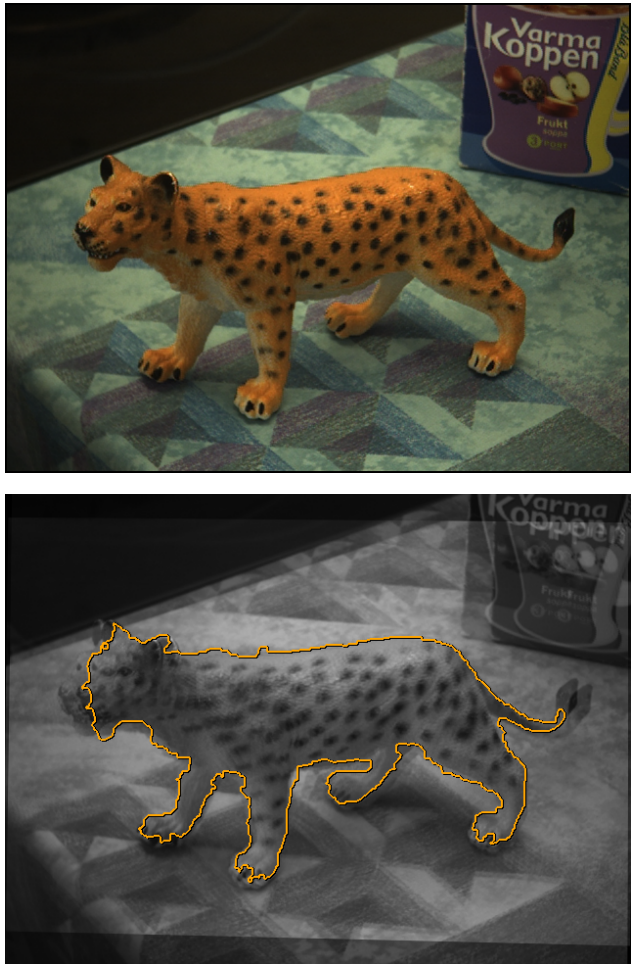
- Saliency Map on Wide-Field View (Itti & Koch Model)
- Peaks in the map = object hypotheses
- Gaze shift and Fixation triggered
- B. Rasolzadeh et al., IJRR 2009

Fixation, Online Calibration and Stereo Matching

- After gaze shift fixation to improve reconstruction
- Adjustment of vergence angle
- Goal: Highest density of points close to the center of images is at zero disparity
- Initial rectification from offline calibration
- Parallel to Fixation: Refinement through online calibration
  - Matching of Harris Corner features in left and right images
  - Affine essential matrix
- Stereo Matching with OpenCV
- B. Rasolzadeh et al., IJRR 2009, `www.csc.kth.se/~celle`

Foveated Segmentation of unknown objects

- Hard to segregate unknown object from supporting surface
  - In household environment, objects commonly placed on flat surfaces
  - Three hypotheses for each pixel: Foreground, background or surface
- Object Model: 3D shape and colour distribution
- Iterative process to strengthen hypotheses over time
- Similar to Expectation-Maximisation; approximate technique for real-time capability
- Initialisation through fixation point
- Björkman and Kragic, ICRA 2010, BMVC 2010

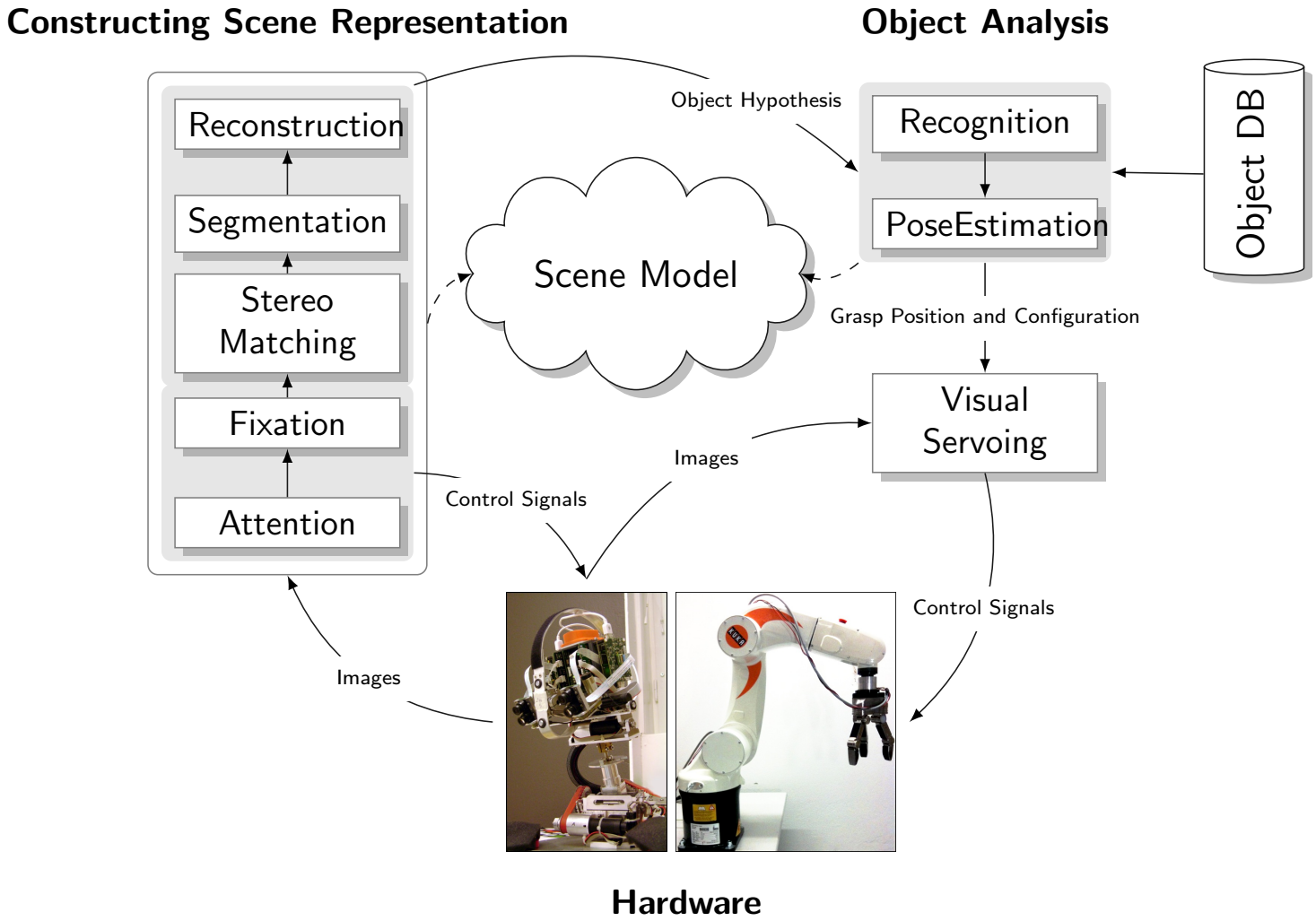


Resulting Scene Model

Five views merged together

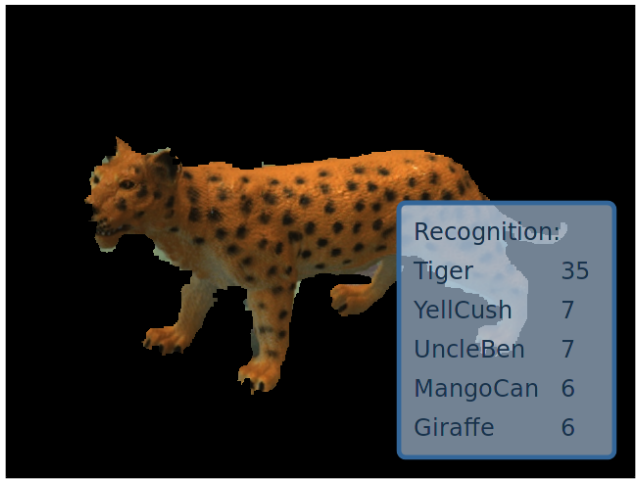
Initial Scene Model

Task: Grasping Known Objects



Recognition of Objects

- Two complementory cues: Color Co-occurence Histogram (CCH) and SIFT features (Bag-of-Words model)
- Object model learned offline from several viewpoints
- Online: Model matched against database with 25 objects
- M. Björkman and J-O. Eklundh, International Journal of Imaging Systems and Technology, 2006 and BMVC 2005



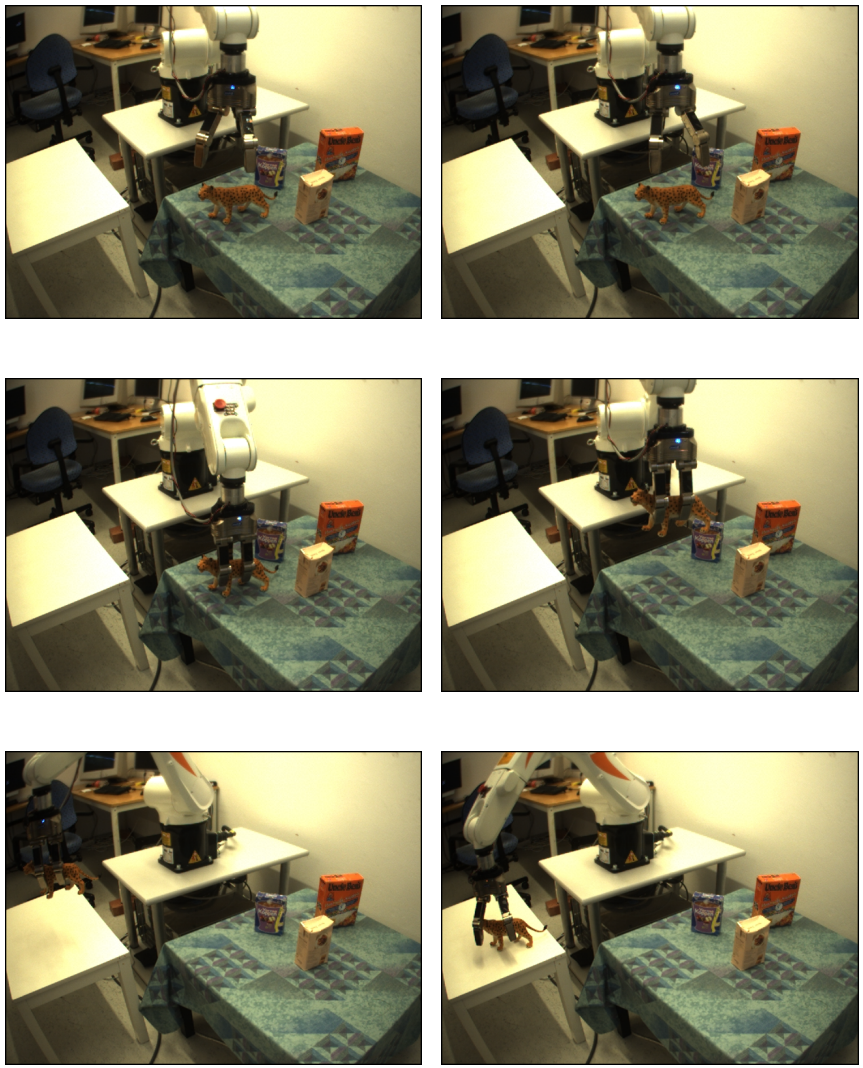
Pose Estimation

- Assumptions:
  - Objects can be approximated as either cylinders or boxes of known dimensions
  - Standing upright (cylinder) or on one of its surfaces (box)
- Given point cloud of a recognised object projected on the table $\rightarrow$ either rectangle or circle can be fitted to in 2D
- Initial guess with RANSAC
- refinement via minimisation of sum of absolute errors of each point to model (efficiently through a 2D distance map)
- another solution C. Papazov and D. Burschka, ISVC 2009, ACCV 2010

Figure: 3D Point Cloud (from two Viewpoints) and Estimated Object Pose.

Grasping the Objects

- Main focus of this talk: How to get to a scene representation suitable for grasping and manipulation?
- Grasping simplified to top grasps
- Pre-Shaping based on overall object shape
- position based visual servoing
  - LED on Hand aligned with desired position above object in the image
  - Hand lowered based on known object height
  - object is picked up and moved to side table



And now all together .....

See `www.csc.kth.se/~bohg/2010_IROS_WS.mpg`

Conclusions and Future Work

- We proposed an **Active Vision System** that **incrementally** builds up a **scene model** suitable for manipulation and grasping
- achieved through integrating different computational processes like
  - Attention
  - Fixation
  - Stereo Reconstruction
  - Segmentation
  - Recognition
  - Pose Estimation
  - Visual Servoing
- Initial scene model independent of object knowledge

Future Work - Grasping Unknown objects

- Really fresh work in collaboration with UJI Castellon, Spain
- Exploring the scene with the same processes
- Filling in the holes
- Manipulation planning in whole scene

Future Work - Augmenting Initial Scene Model with Haptic Information

- Talk on Thursday here at IROS 2010
- No Grasping, but haptic exploration to fill in the holes in the scene model
- See `www.csc.kth.se/~bohg/IROS2010Grasp.mp4`

Thank you for your attention!

This works has been supported by the EU project GRASP.