Active Object Recognition under Gaze Control

Jan-Olof Eklundh, Mårten Björkman

CVAP, KTH, Stockholm

The KTH Head of 1991

- Independent eye and neck movements
- Eye rotations around optical center
- Eccentric neck
- Drive towards symmetry constrained redundancy
- Monocular stabilization and pursuit, binocular stereopsis and accommodation independent but integrated
- Binocular fixation at lateral speeds up to 115°/s, 5 m/s in depth. Saccades up to 360°/s

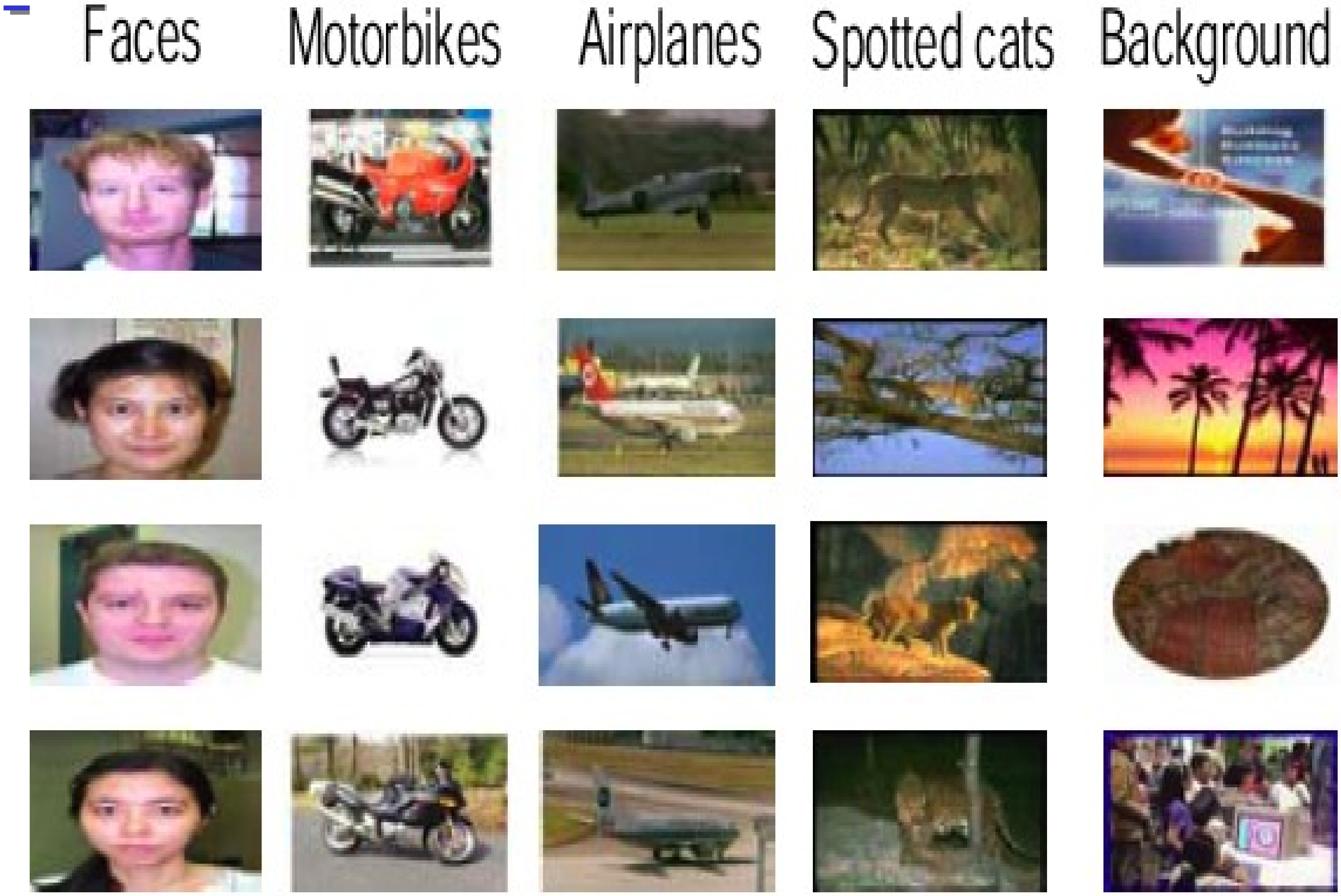
QuickTime™ and a
H.261 decompressor
are needed to see this picture.

QuickTime™ and a
H.261 decompressor
are needed to see this picture.

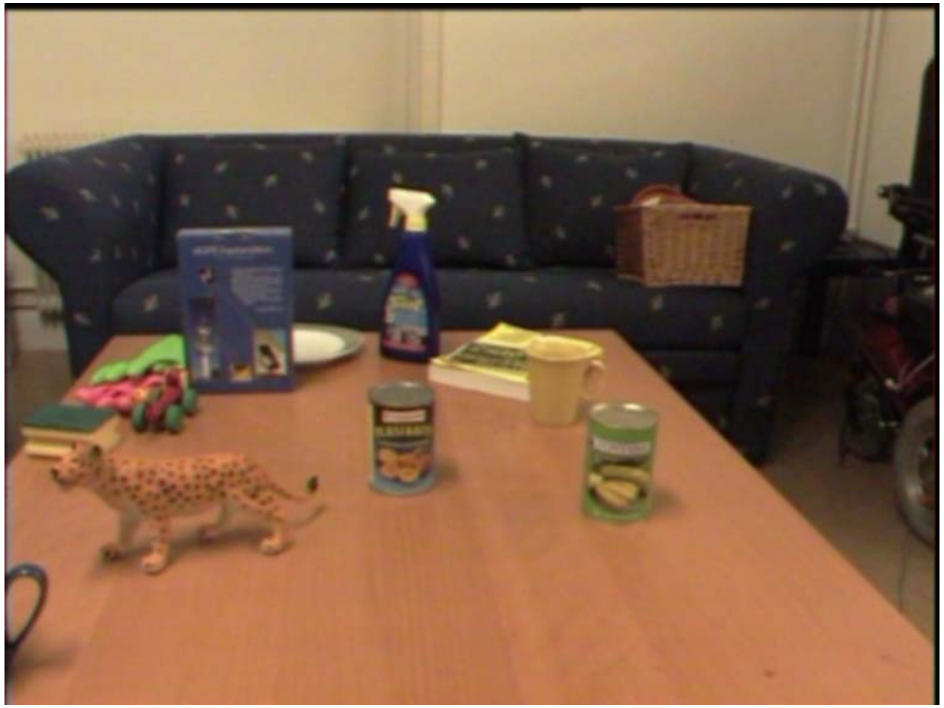
What did this system “see”?

- High performance through tightly integrated hardware. No resources left
- Information about ego-motion, independent object motion and depth available, but couldn’t be utilized
- Appearance of 3D objects too, as also to some extent pose
- Goal of current work to do that in visual search and hand-eye coordination tasks

S-o-t-a object classification



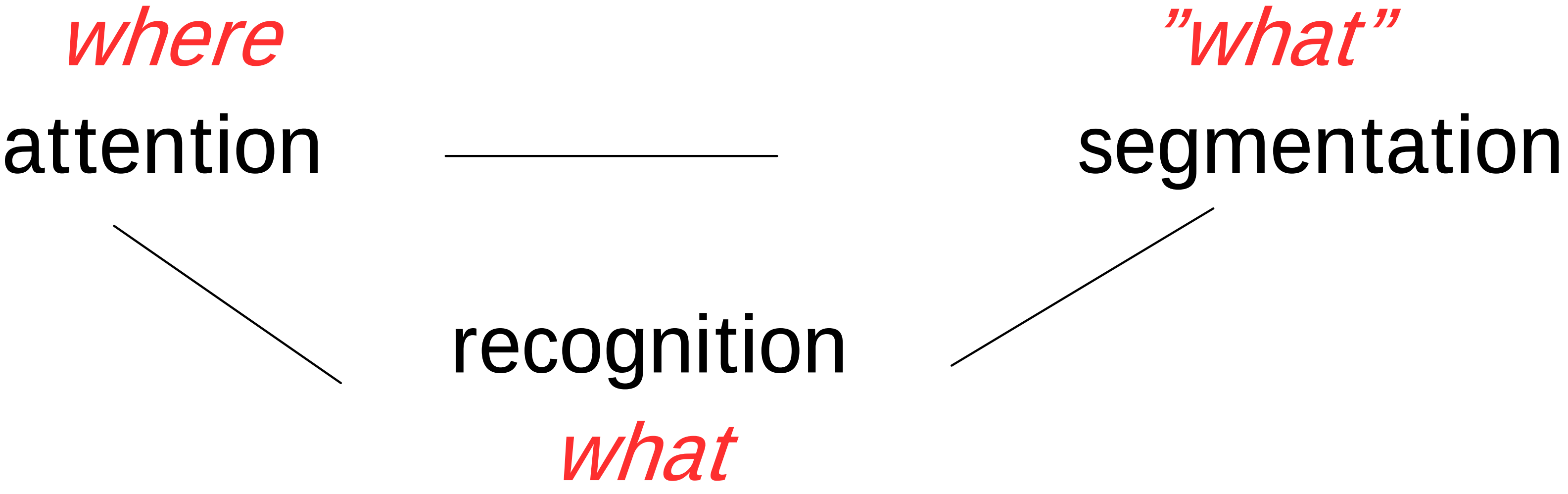
A robot looking at a table at 1.5 m



Objects subtend only a fraction of the scene and are not centered (unless attentional step)

Computational Vision and Active Perception

Desirable system structure



where

attention

"what"

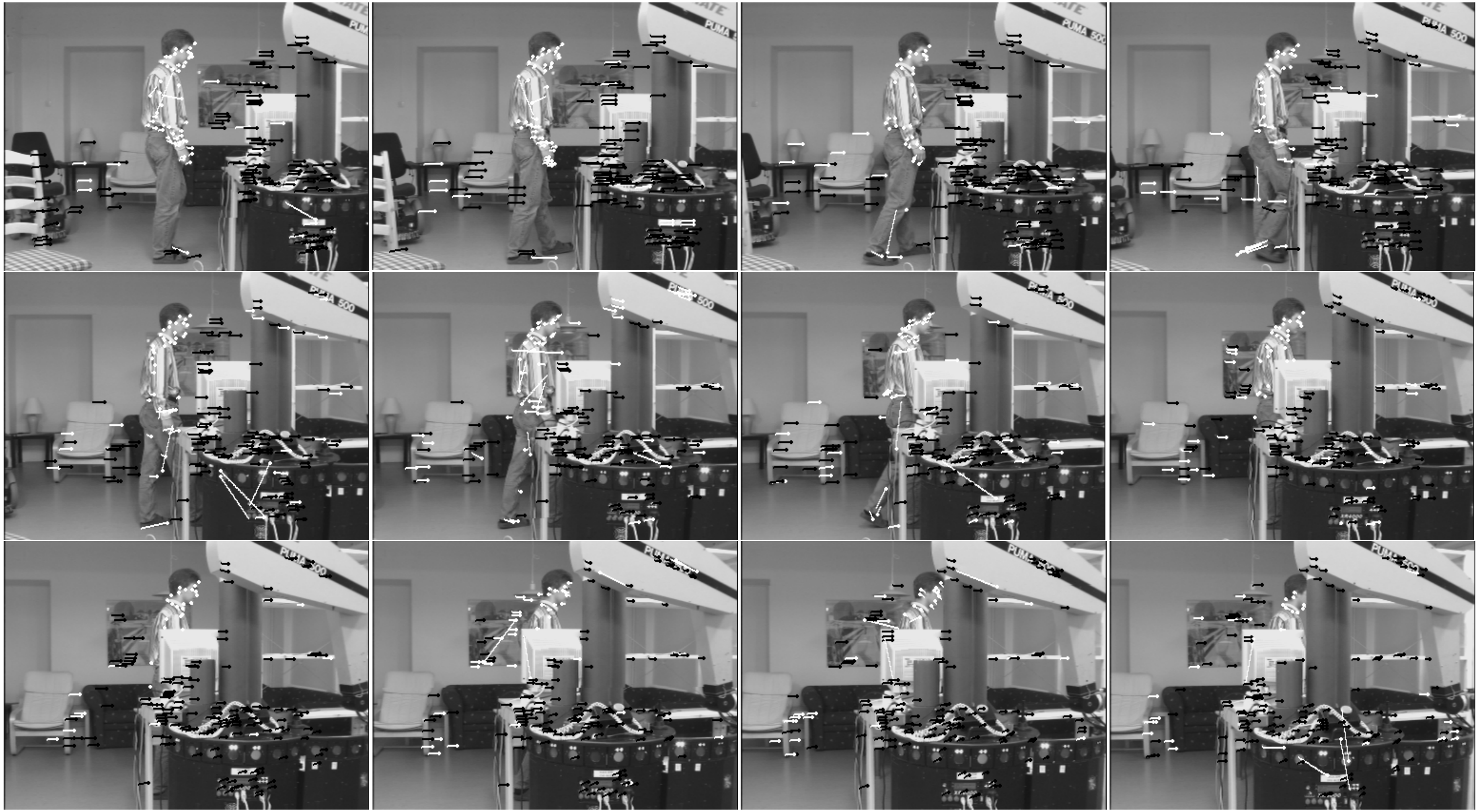
segmentation

recognition

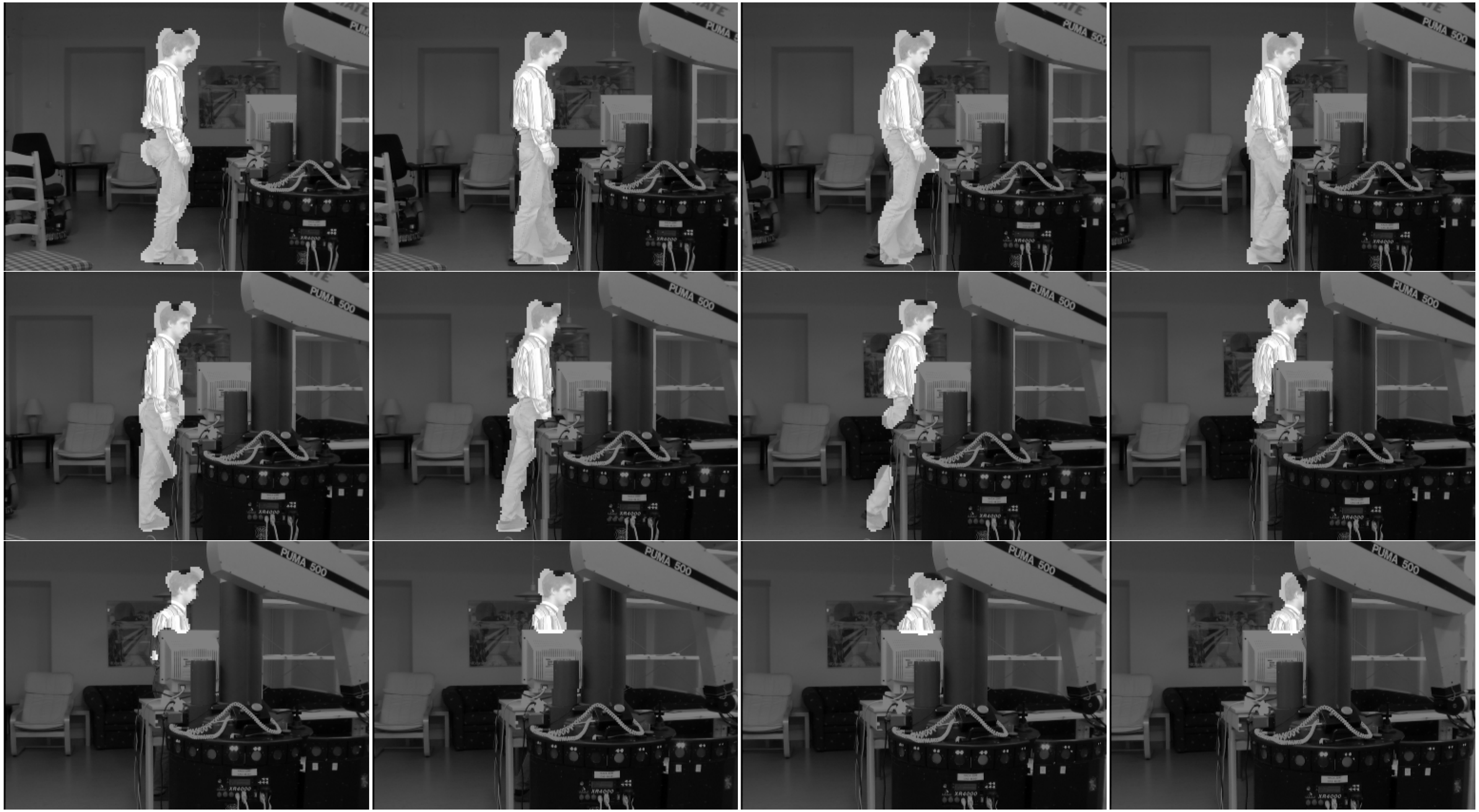
what

Run concurrently. Motion powerful in bootstrapping, but static objects often as important.

With stereo and motion



What the system “sees”



Computational Vision and Active Perception

F-g-s by integration of multiple cues from motion and appearance

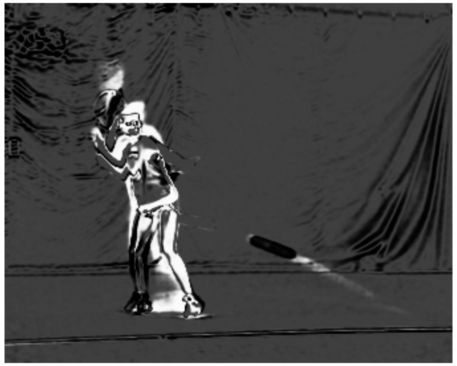
QuickTime™ and a
YUV420 codec decompressor
are needed to see this picture.

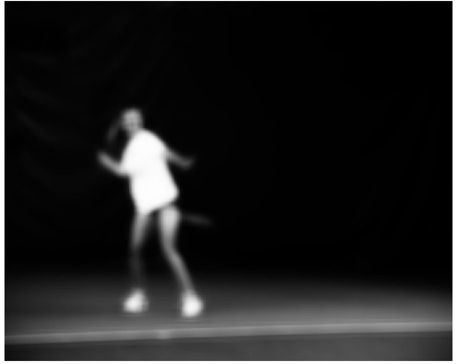
Original

Foreground mask

The cues

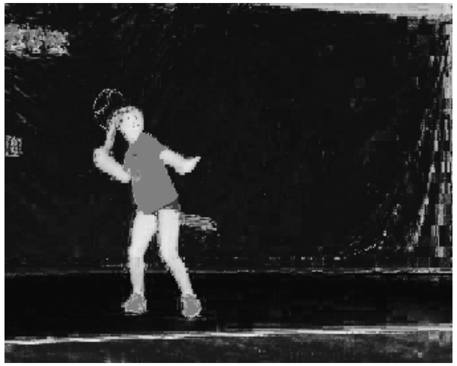
Motion

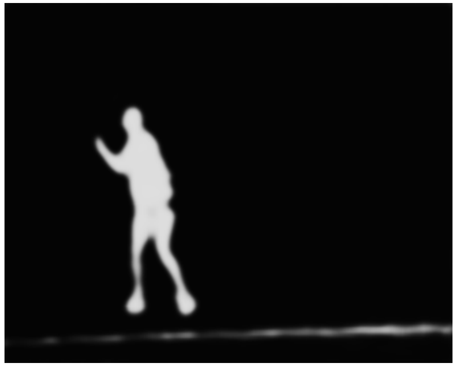




Texture (contrast)

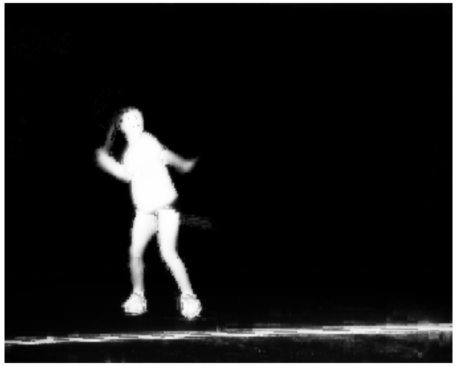
Colour





Prediction

Combined



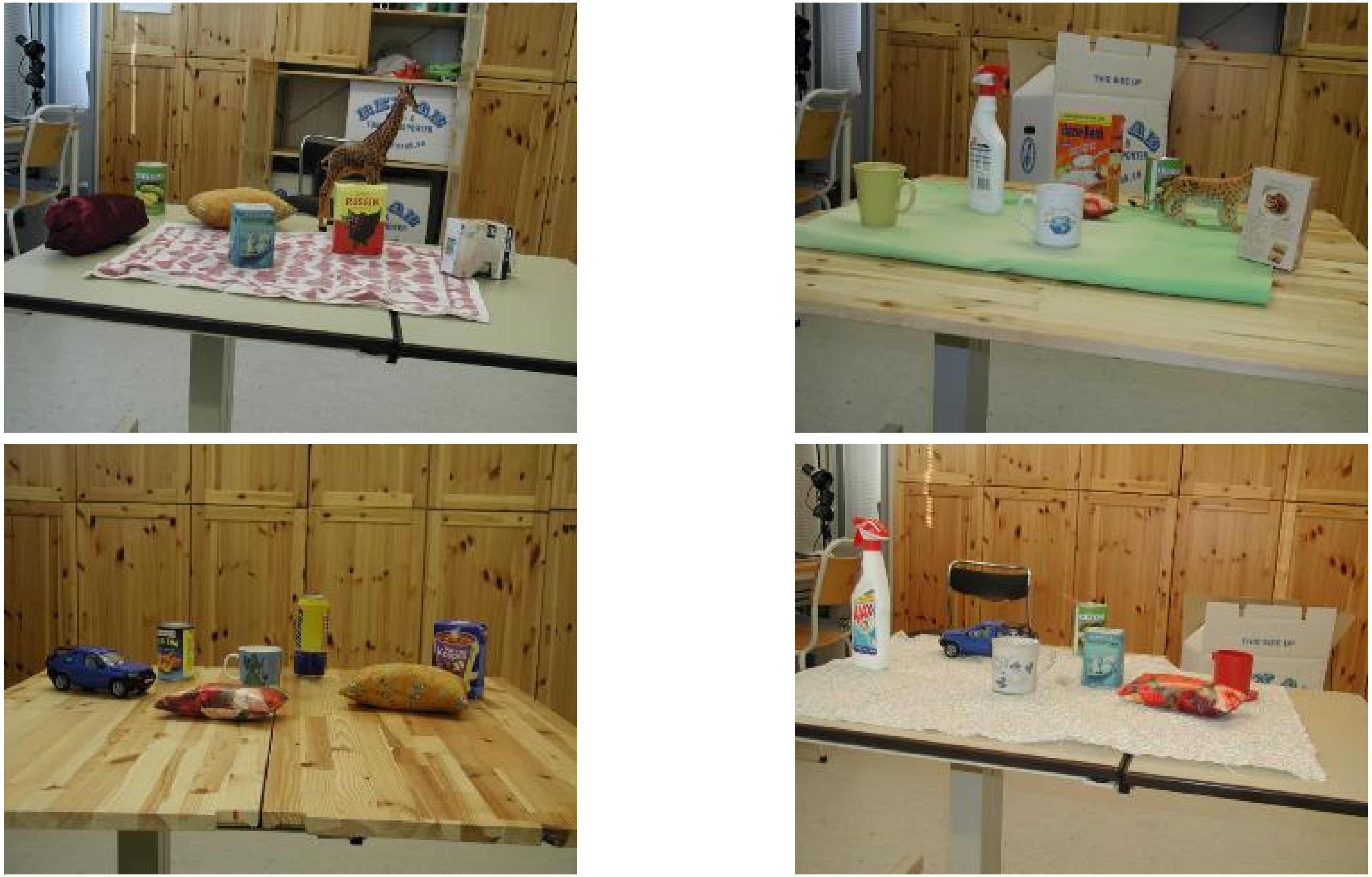
F-g-s by integration of multiple cues from motion and appearance

QuickTime™ and a
YUV420 codec decompressor
are needed to see this picture.

Original

Foreground mask

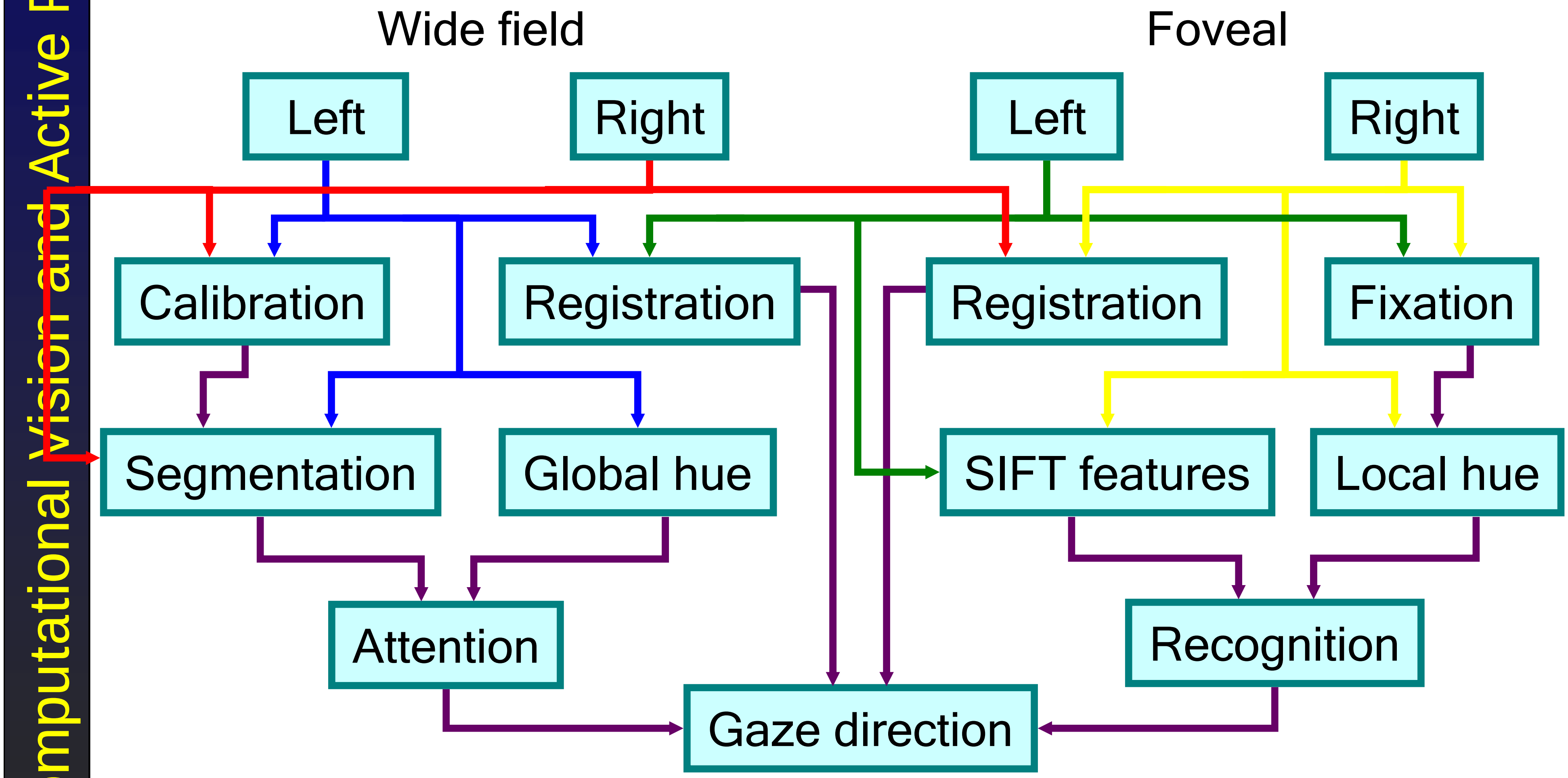
Typical static scenes



A system for searching for static objects

- A wide field of view for attention
- Recognition in foveated view
- Steps:
  - Divide scene into depth layers
  - Select candidate objects through attention
  - Fixate and track objects of potential interest
  - Recognize/classify objects in foveal view, possibly after a second binocularly based segmentation
- Technically: two pairs of stereo cameras
- Problem: transfer of views

Flow of information



Stereo computations

Relative orientations have to be known to

- relate disparities to depths
- simplify estimation of disparities

Using corner features and optical flow model

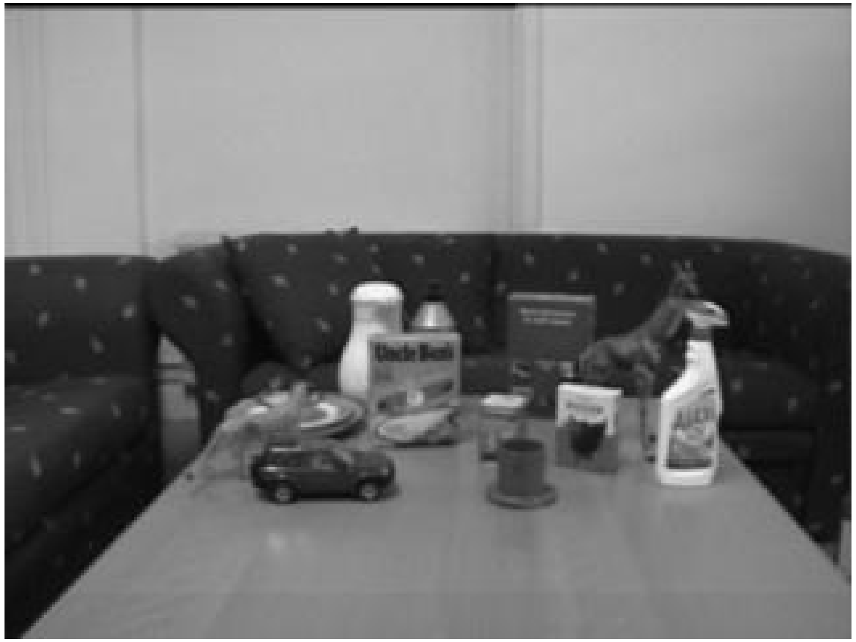
$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} (1+x^2)\alpha - y\, r_z \\ xy\, \alpha + r_y + x\, r_z \end{pmatrix} + \frac{1}{z} \begin{pmatrix} 1 - x\, t \\ - y\, t \end{pmatrix}$$

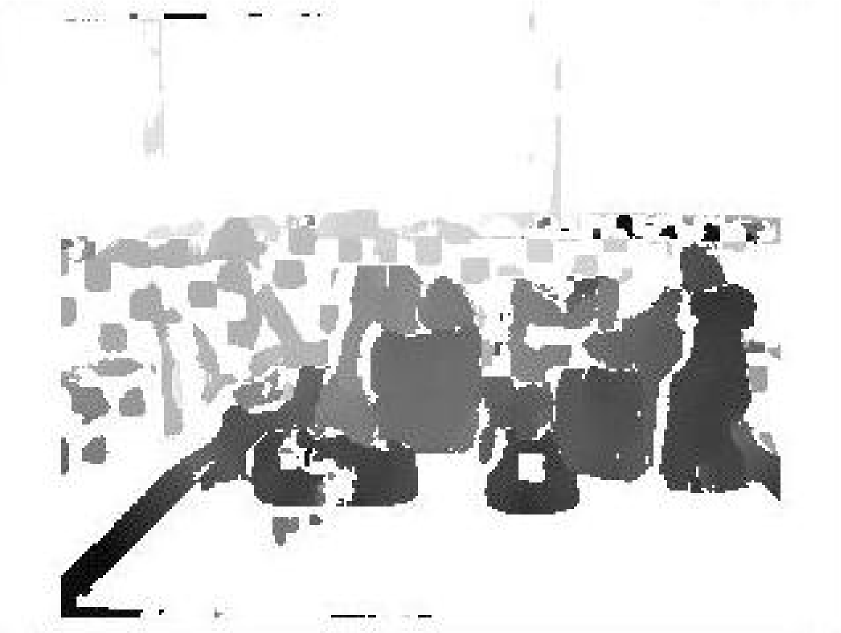
Unstable process => use robust methods

First assume r_z and r_y to be zero

On-line calibration allows the use of expected retinal size

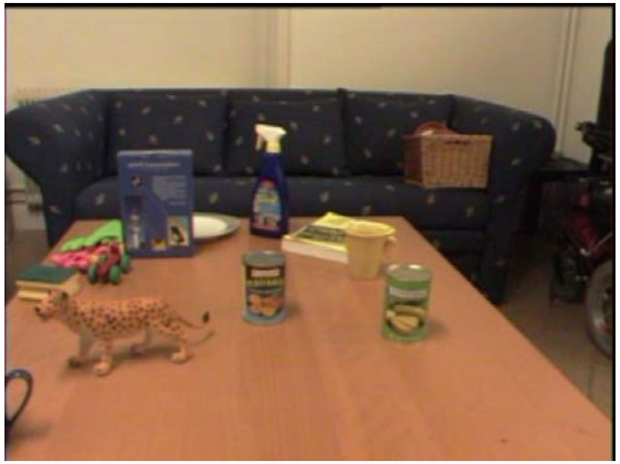
Figure-ground segmentation

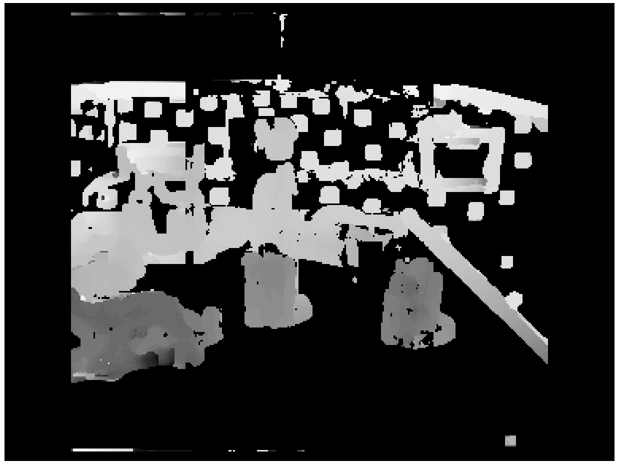




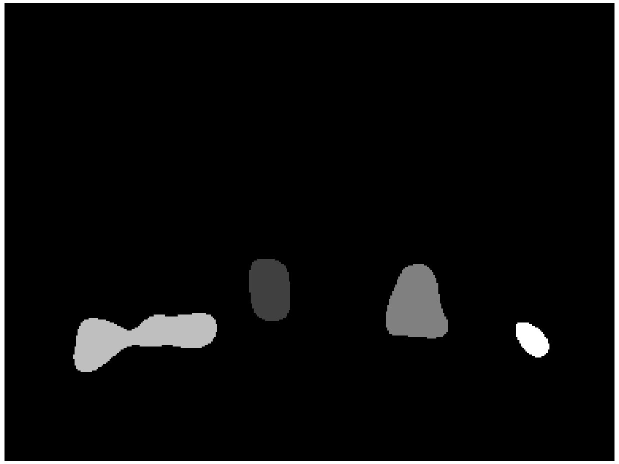
Disparity map is sliced into layers.
Widths are set after objects searched for

Figure-ground segmentation



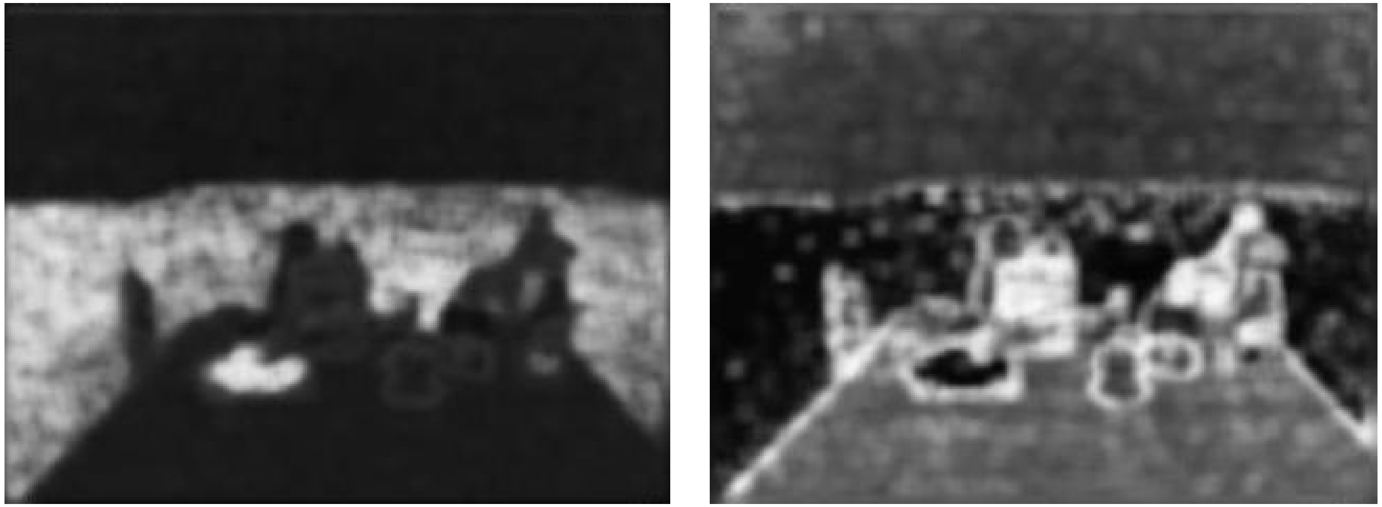


BinoCues



BinoAttn

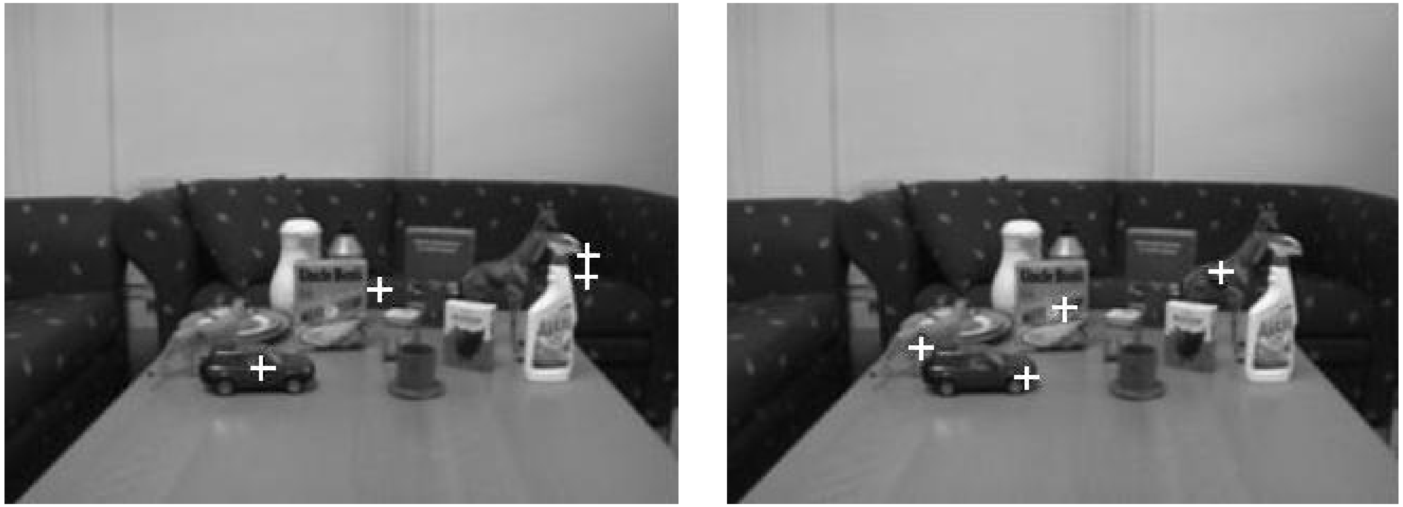
Appearance based attention



Local hue histograms correlated with that of requested object.

Fast implementation using rotating sums.

Saliency peaks



Peaks from blob detection of depth slices.
Based on Differences of Gaussians.
Hue saliency map used for weighting.
Random value added before selection.
Inhibition on return

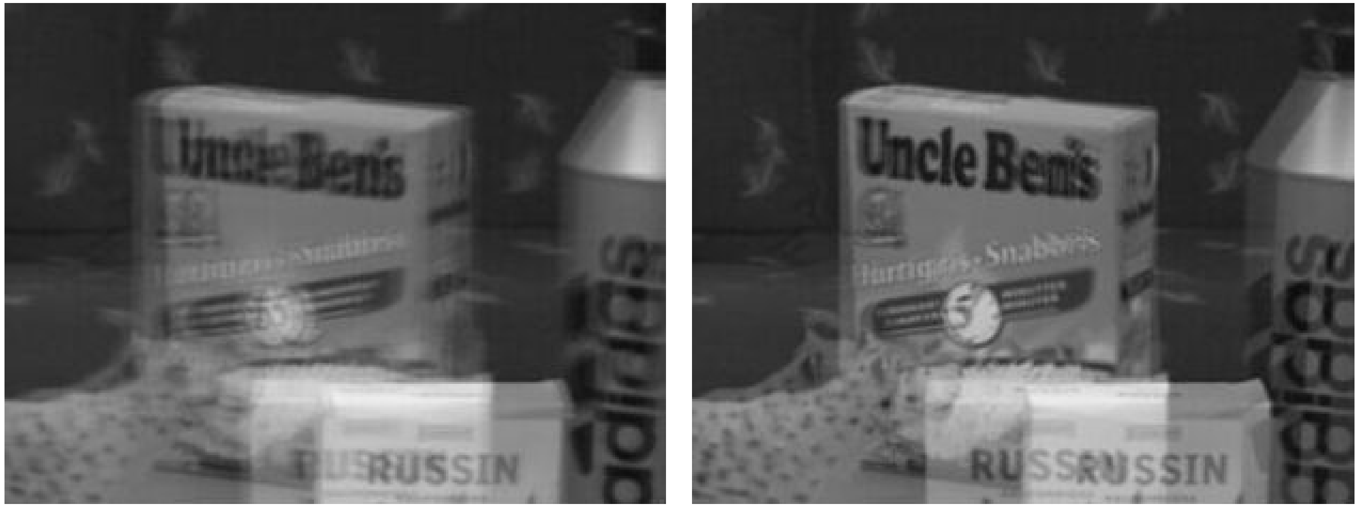
Fixation

The foveal system continuously tries to fixate

- done using corner features
- and affine essential matrix

$$F = \begin{pmatrix} 0 & 0 & c \\ 0 & 0 & d \\ a & b & e \end{pmatrix}$$

Zero disparity filters won’t work

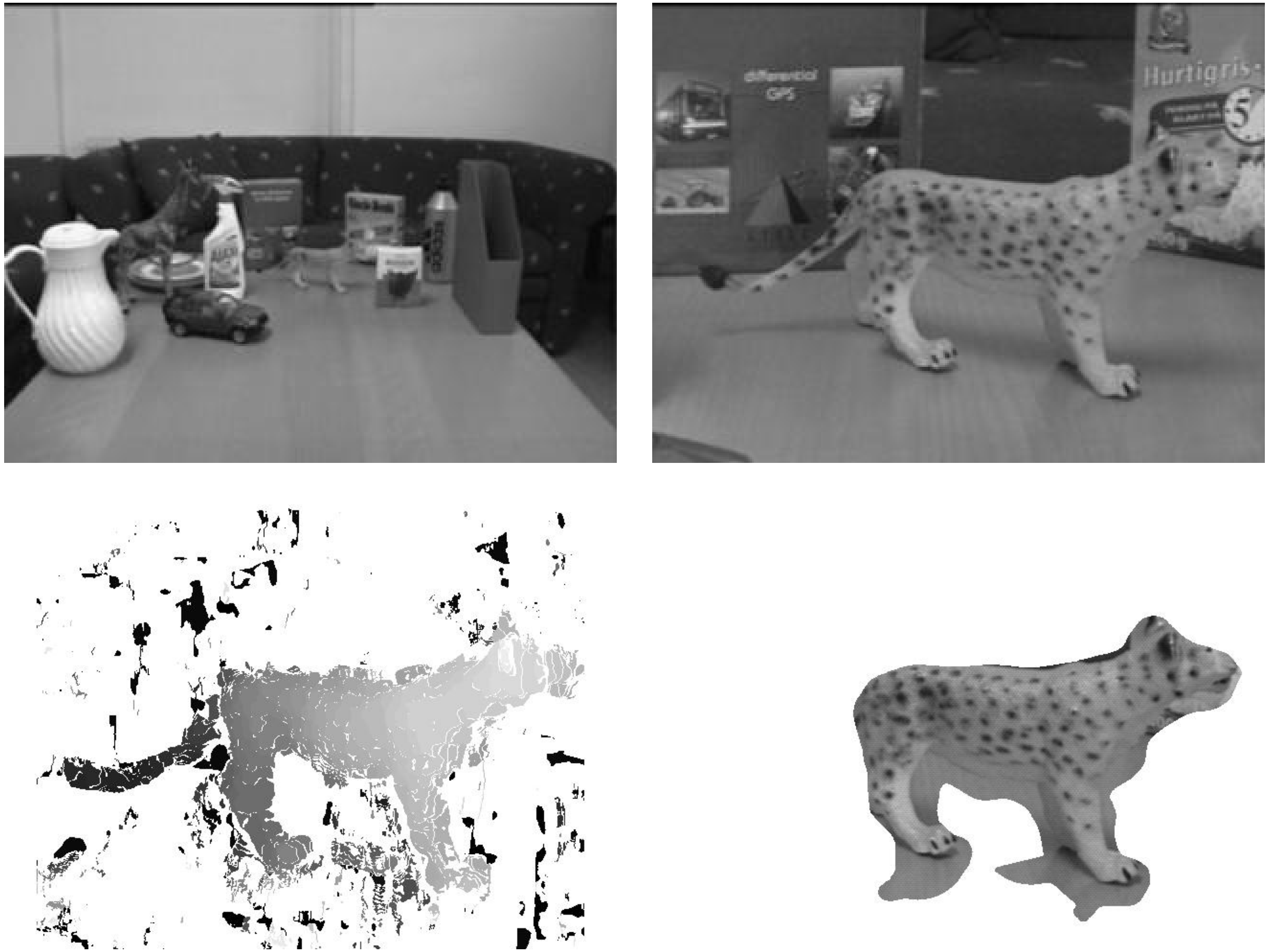


Foveated segmentation

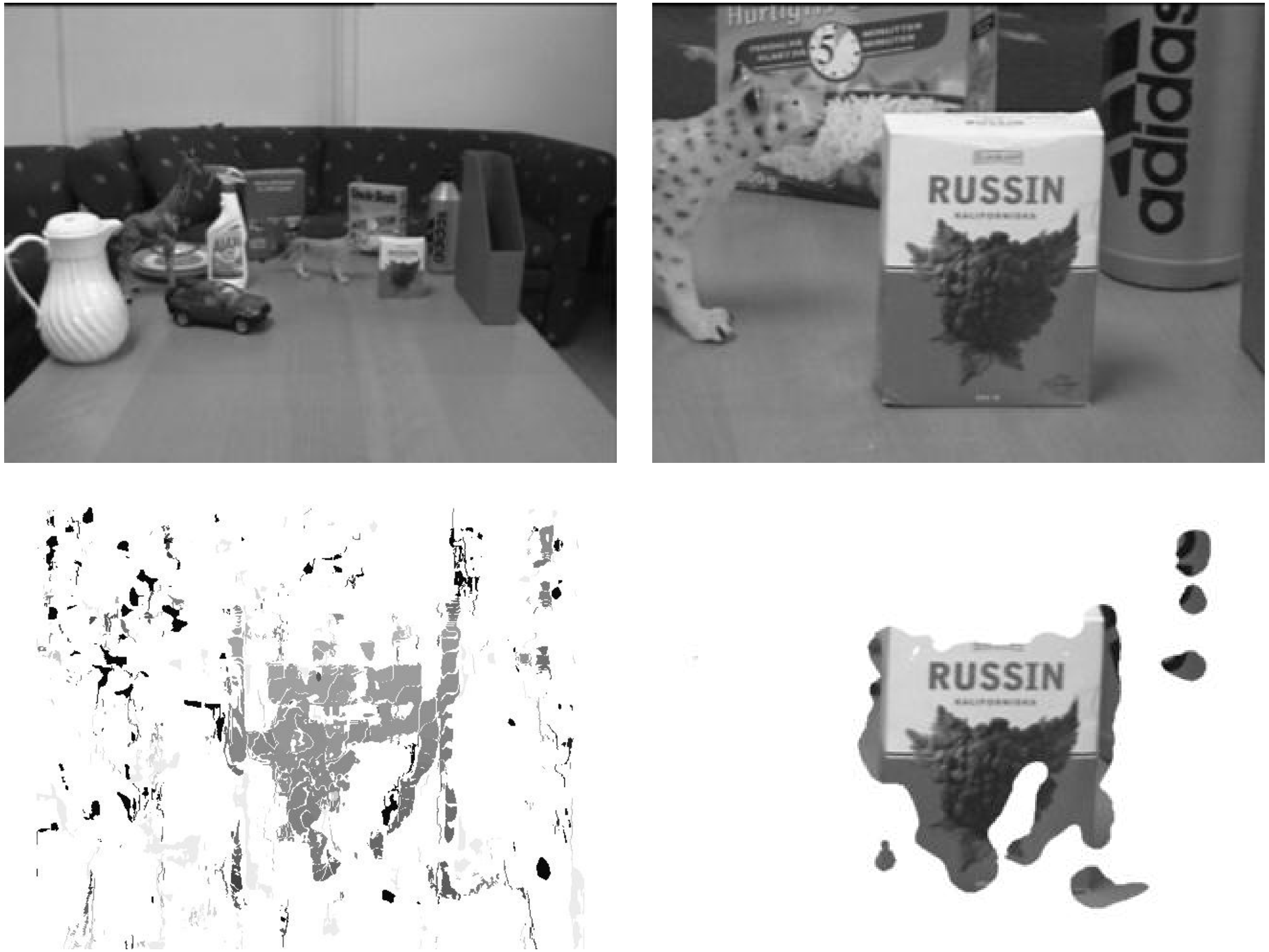
To boost the ensuing recognition/classification

- Foveal segmentation based on disparities
- Rectification using affine fundamental matrix
    - Only search for disparities around zero => Large number of false positives
    - Points clustered in 3D using mean shift

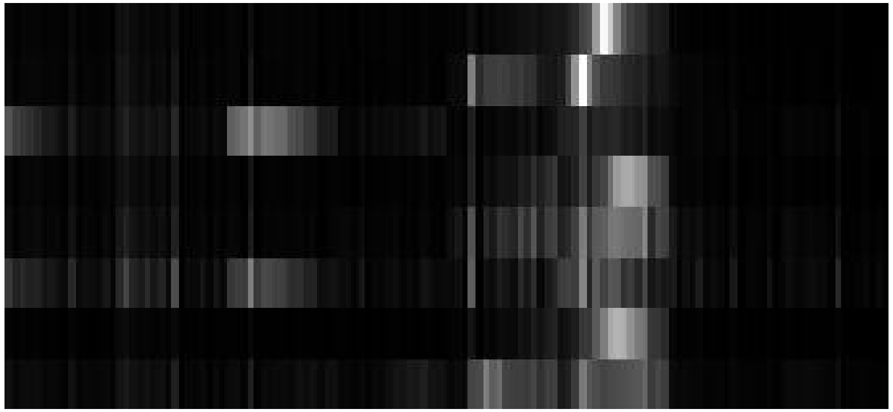
Foveated segmentation



Foveated segmentation



Small example object database in real-time experiments - in total 24

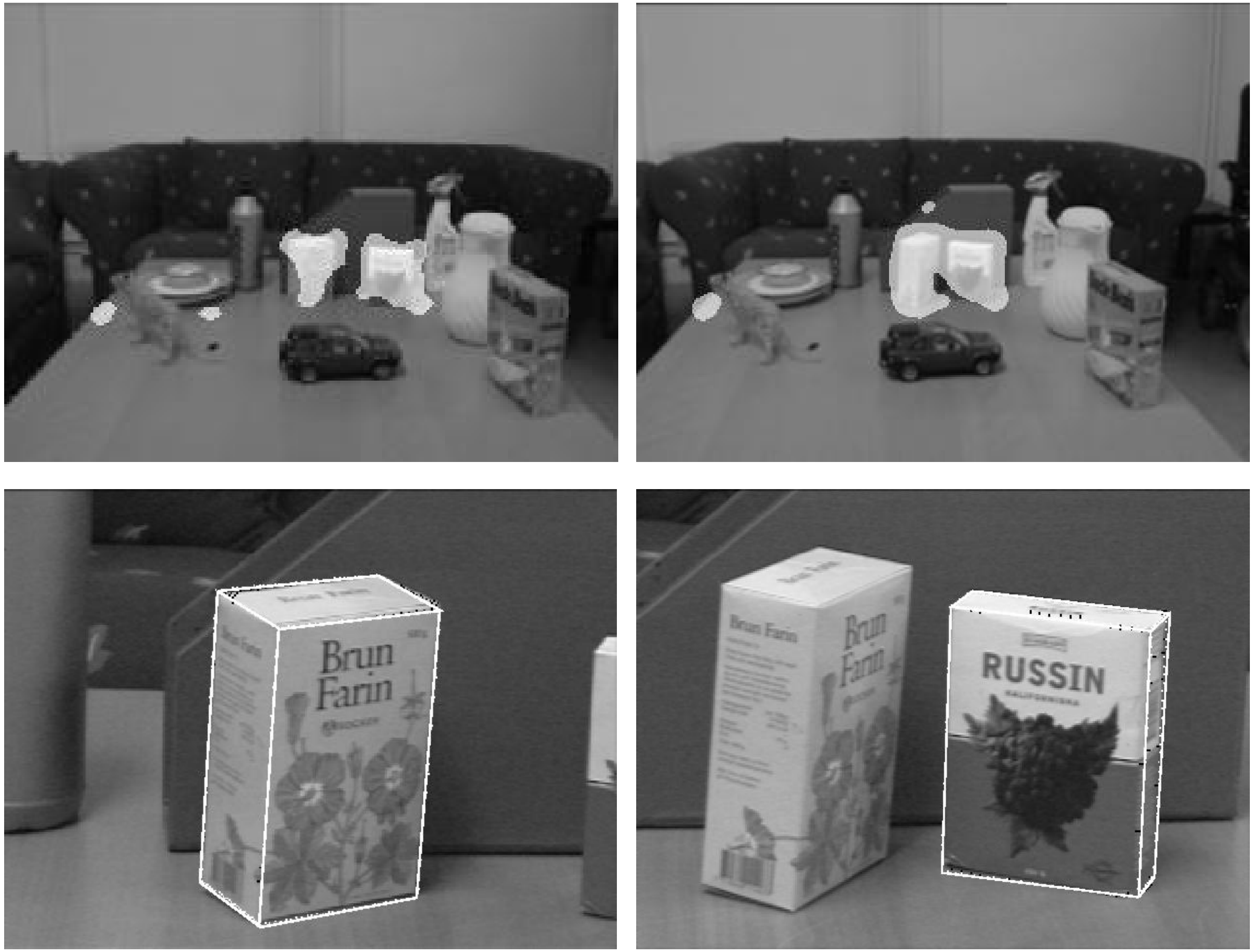


Here models of SIFT features and hue histograms. Texture descriptors also included now.

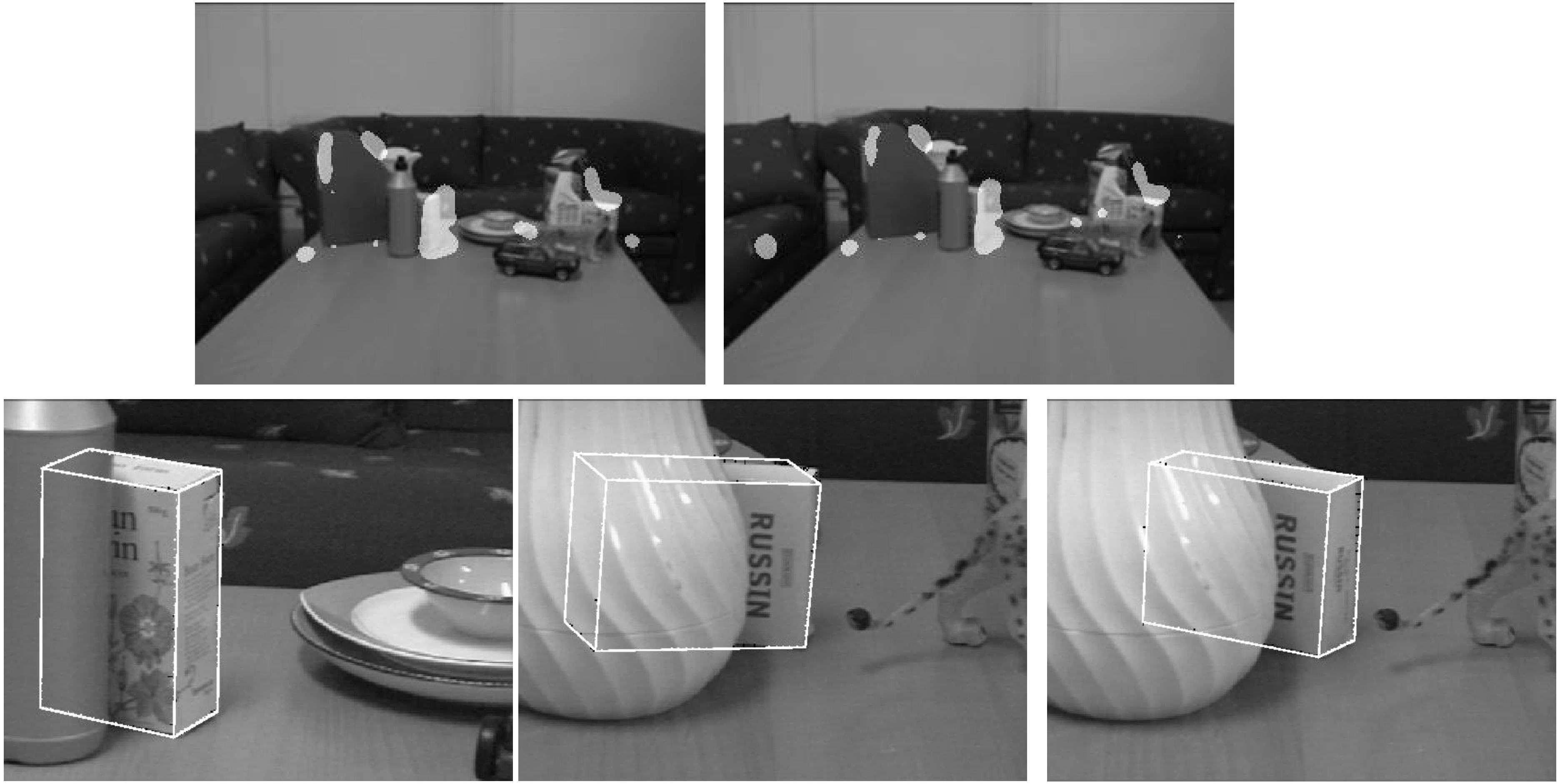
Visual scene search

QuickTime™ and a
YUV420 codec decompressor
are needed to see this picture.

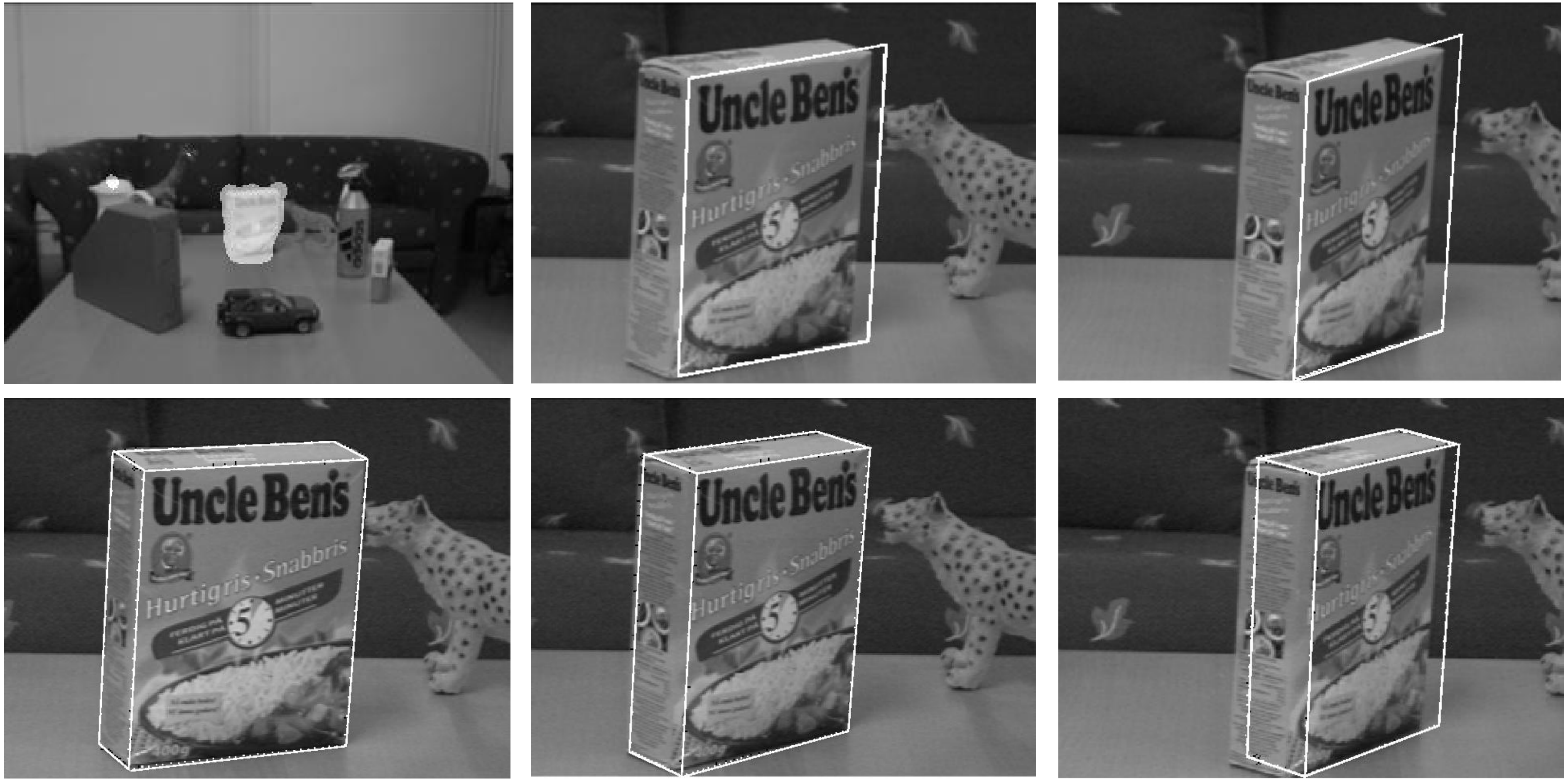
Segmentation robustness



Effect of occlusions



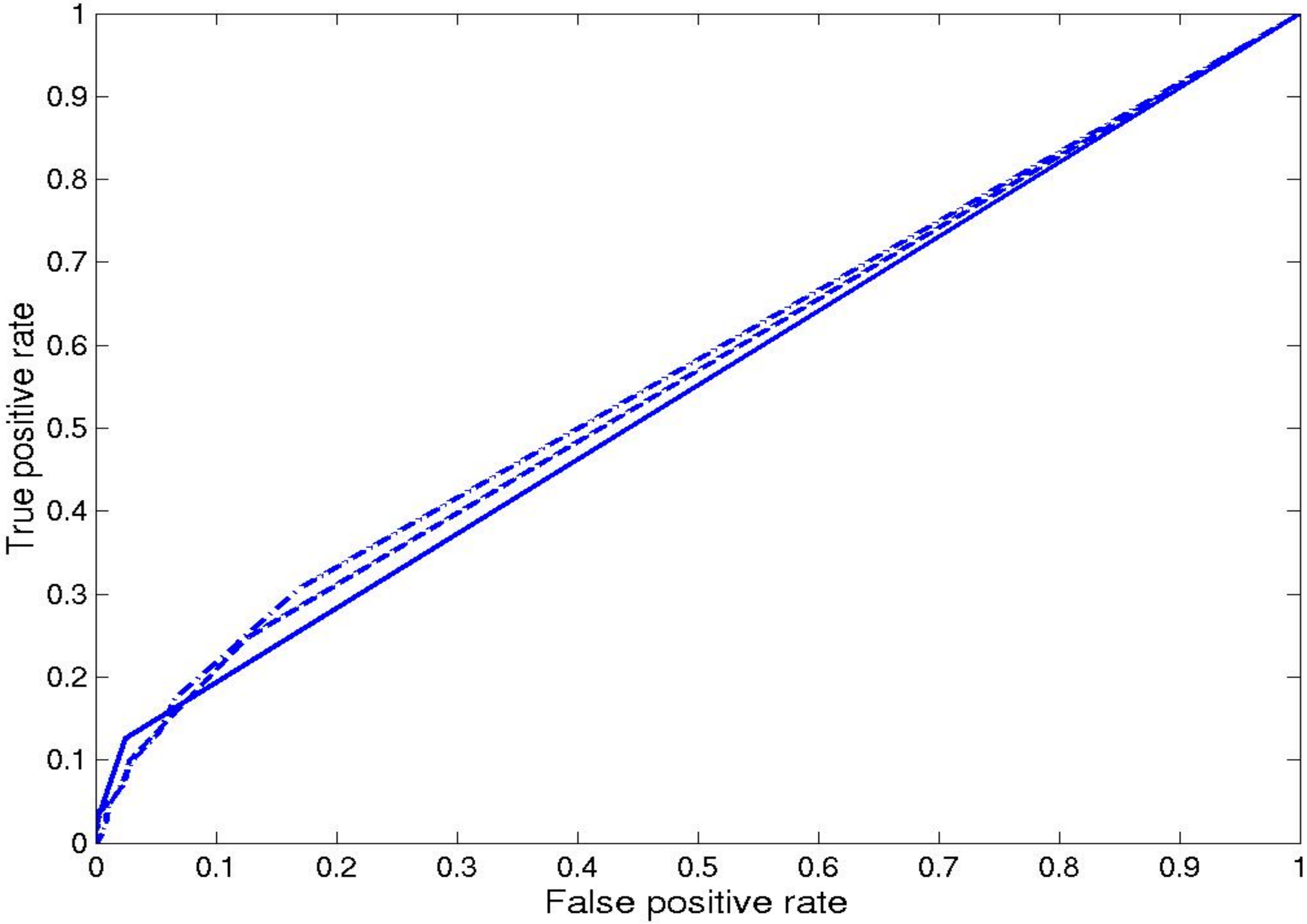
Effect of rotations



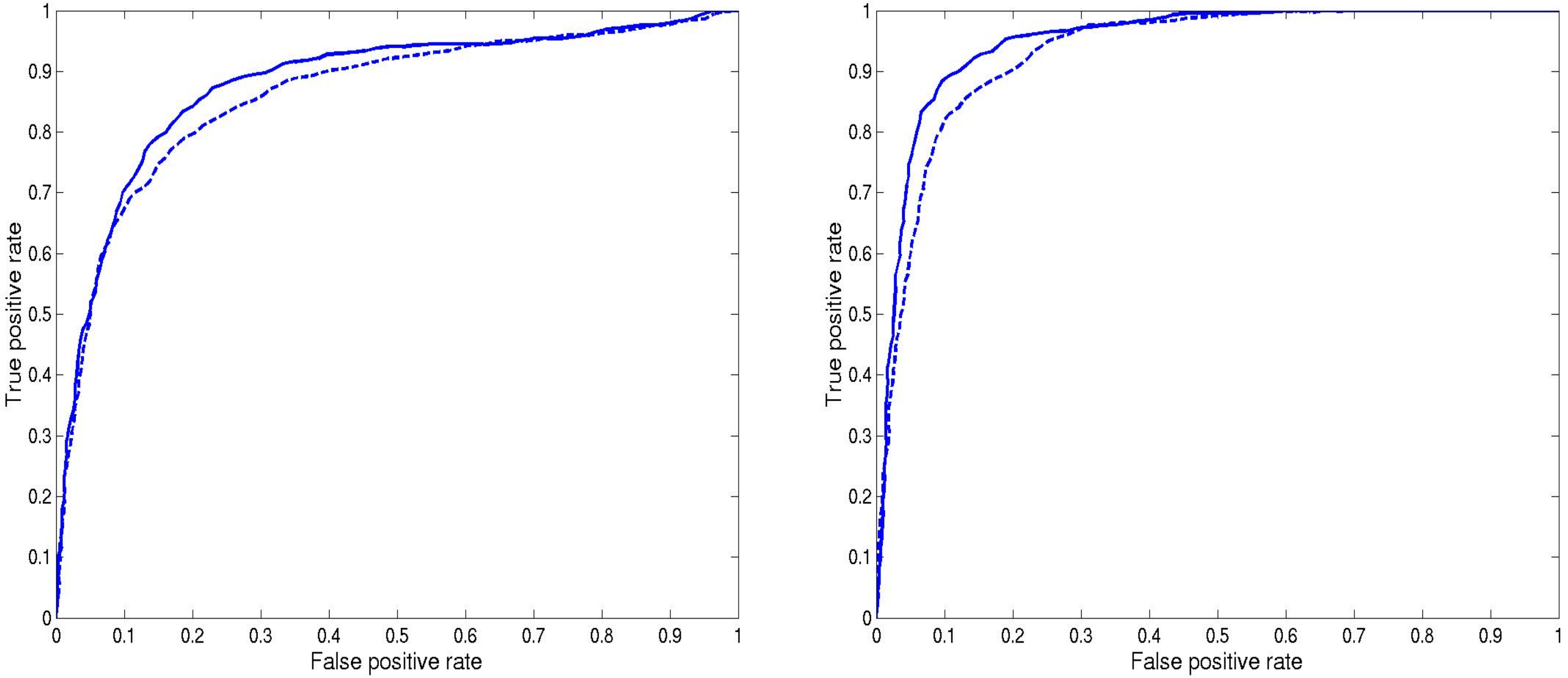
Recognition experiments

- 24 objects
- Learned over a range of views, represented by two features
- Arranged in 24 “scenes”
- “Is X in the scene?”
- 3 fixations allowed

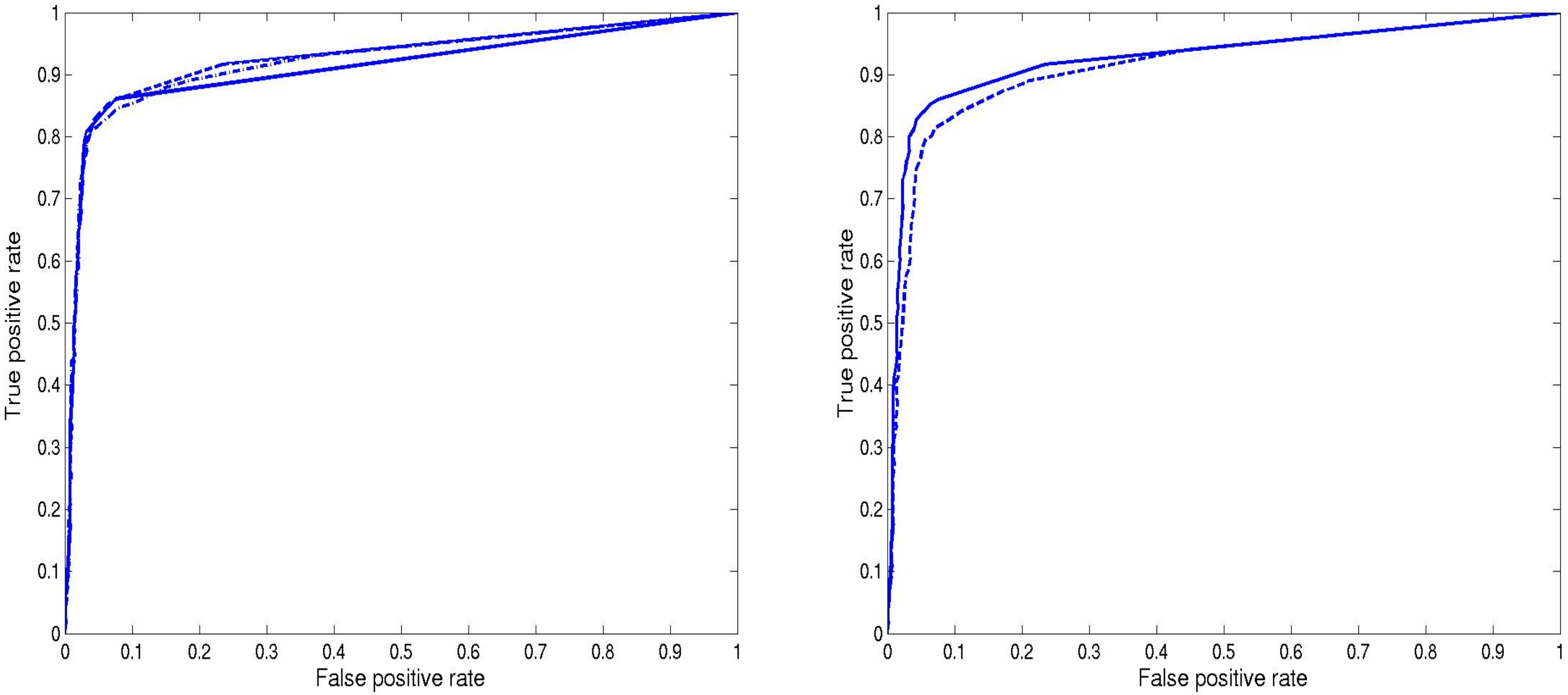
SIFT features in wide field, no disparity based segmentation



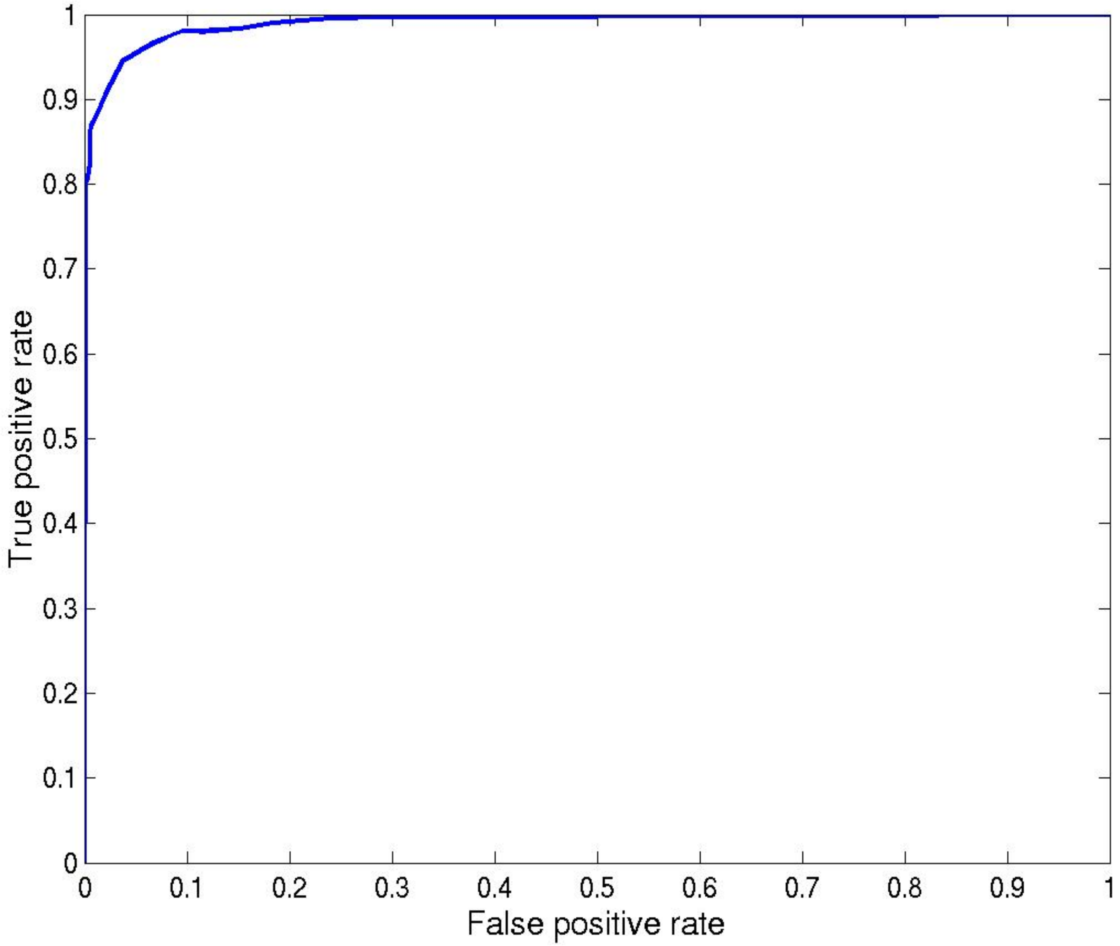
Colour cue, wide field vs wide + central field disparity segmentation



SIFT features, wide field vs wide + central field disparity segmentation



Wide field + central disparity based segmentation, combined features



Conclusions

- Gaze control essential. In fact, many current methods assume foveation or something with similar effects
- 3D cues powerful for figure-ground segmentation (informs about the scene)
- 3D cues thereby also support recognition and categorization
- Integration of multiple cues essential

Comments. Future work

- We have a running system, that normally finds objects within three saccades
- Experiments tedious (learning, scene setups)
- More cues being added, especially texture
- Focus on classification and eventually categorization
- Applications to hand-eye coordination and manipulation
- Potential for computing both local and global shape properties