Perception

Introduction to HRI

Simmons & Nourbakhsh

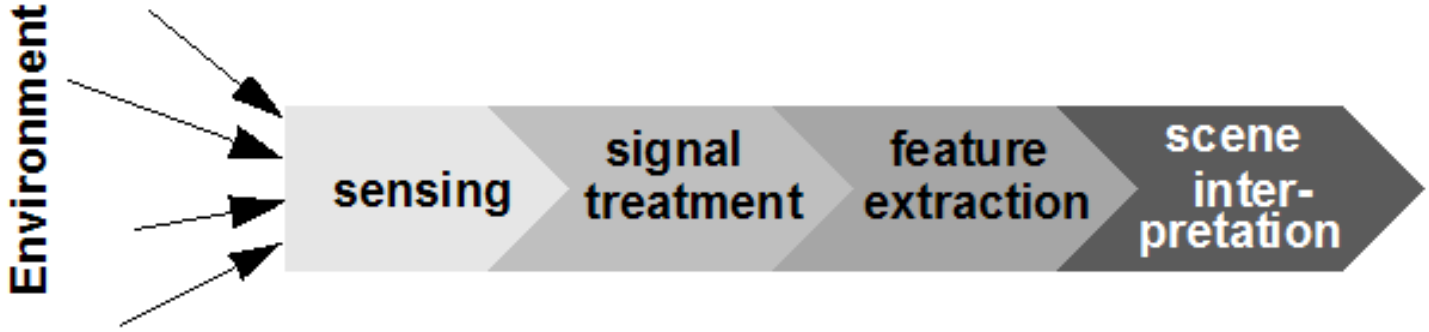
Spring 2015

Perception – my goals

- What is the state of the art boundary?
- Where might we be in 5-10 years?

The Perceptual Pipeline

- *The classical approach: a serial pipeline*
- *Weak link analysis: each step depends on predecessors*
-



Social Perception

- What features do we perceive for sociality?
- Is social perception a serial pipeline?

1. HRI for Human Perceptual Shifting

Insect Telepresence

Educational telepresence designed using formal HCI inquiry tools.

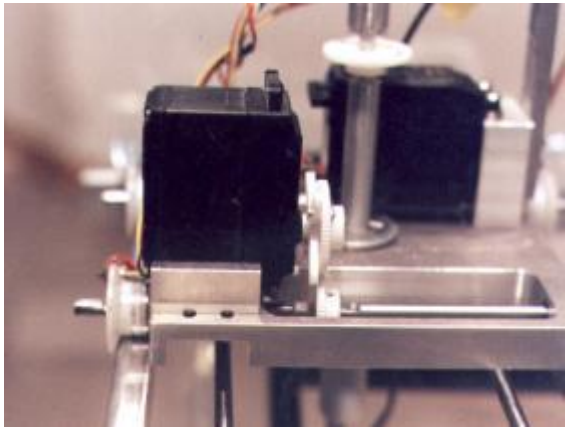
Insect Telepresence Robot

- **Problem**
- Increase visitors' engagement with and appreciation of insects in a museum terrarium at CMNH.

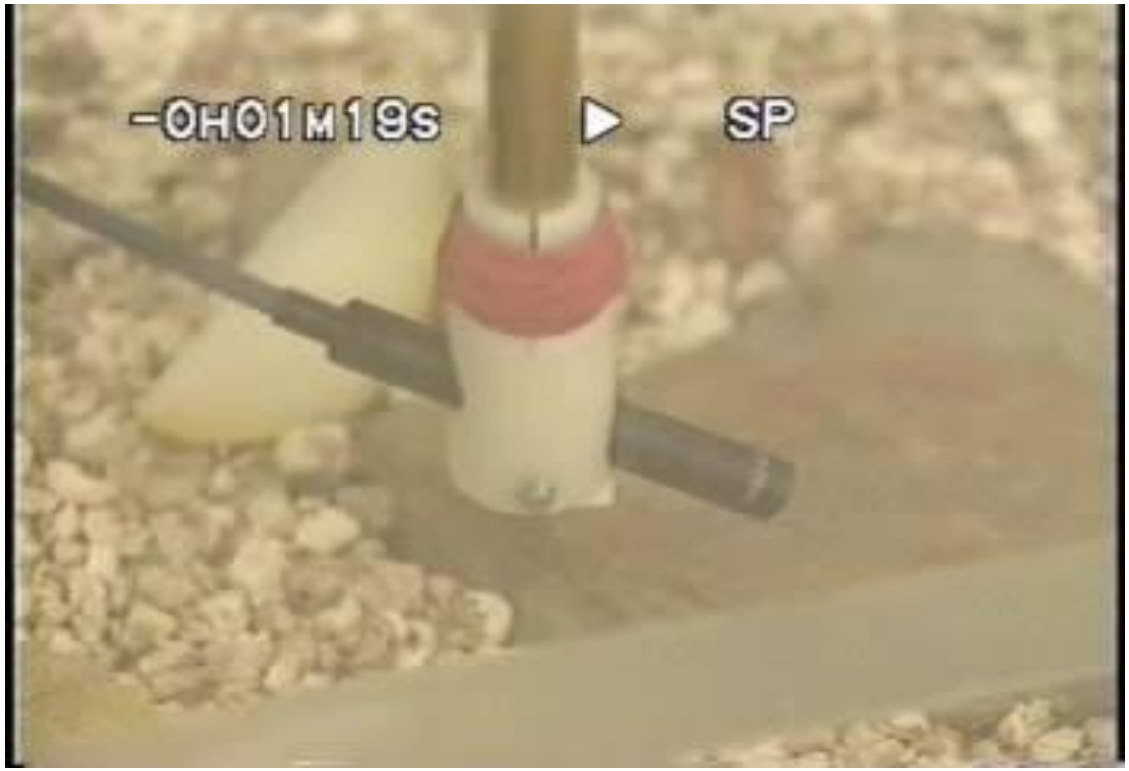
- **Approach**
- Provide a scalar telepresence experience with insect-safe visual browsing
- Apply HCI techniques to design and evaluate the input device and system
  - Cultural modeling, expert interview, baseline observation
- Measure engagement indirectly by 'time on task'
- Partner with HCII, CMNH

Insect Telepresence Robot

- **Innovations**
- Asymmetric exhibit layout
- Mechanical transparency
- Clutched gantry lever arm
- FOV-relative 3 DOF joystick



Insect Telepresence Robot



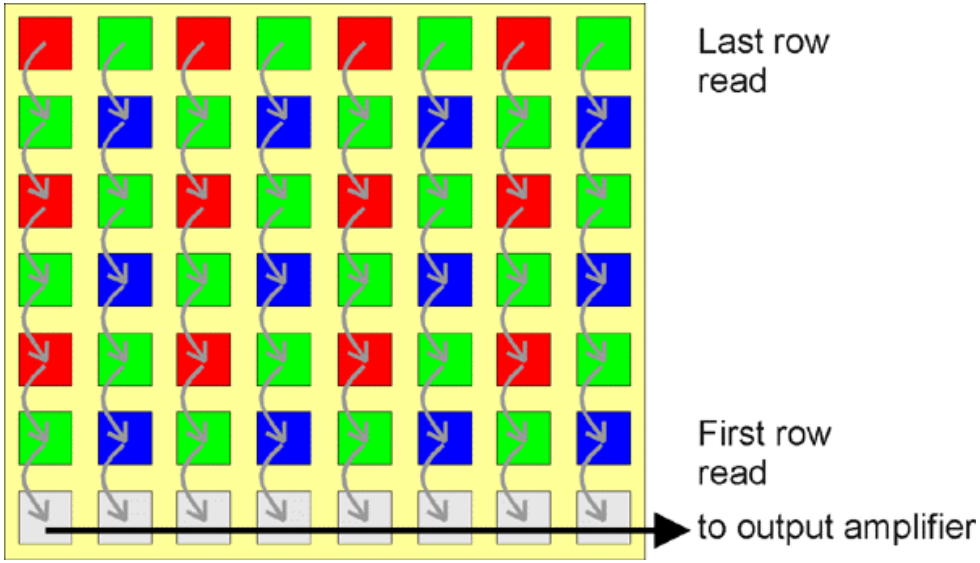
Insect Telepresence Robot

- **Evaluation Results:**
- Average group size: 3
- Average age of users: 19.5 years
- Three age modes: 8 years, 10 years, and 35 years
- Average time on task of all users: **60 seconds**
- Average time on task of a single user: 27 seconds
- Average time on task for user groups: 93 seconds

2. Vision Sensors

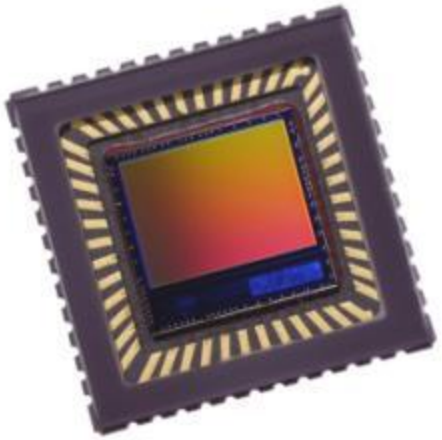
The CCD (Charged Couple Device)

- - Exotic timing circuitry required
- - Uneven frequency response in electron wells
- - Color separation: filters versus splitting
- - Lossy data formats: NTSC and “digital video”
- > Credit: http://www.shortcourses.com/how/sensors/ccd_readout.gif



The CMOS (Complementary Metal Oxide Semiconductor)

- - Standard chip fabrication techniques
- - Far lower power consumption overall (1:100)
- - Pixel/well measurement circuitry at along pixel
- - Real estate problems ; efficiency of photon usage



Human Vision

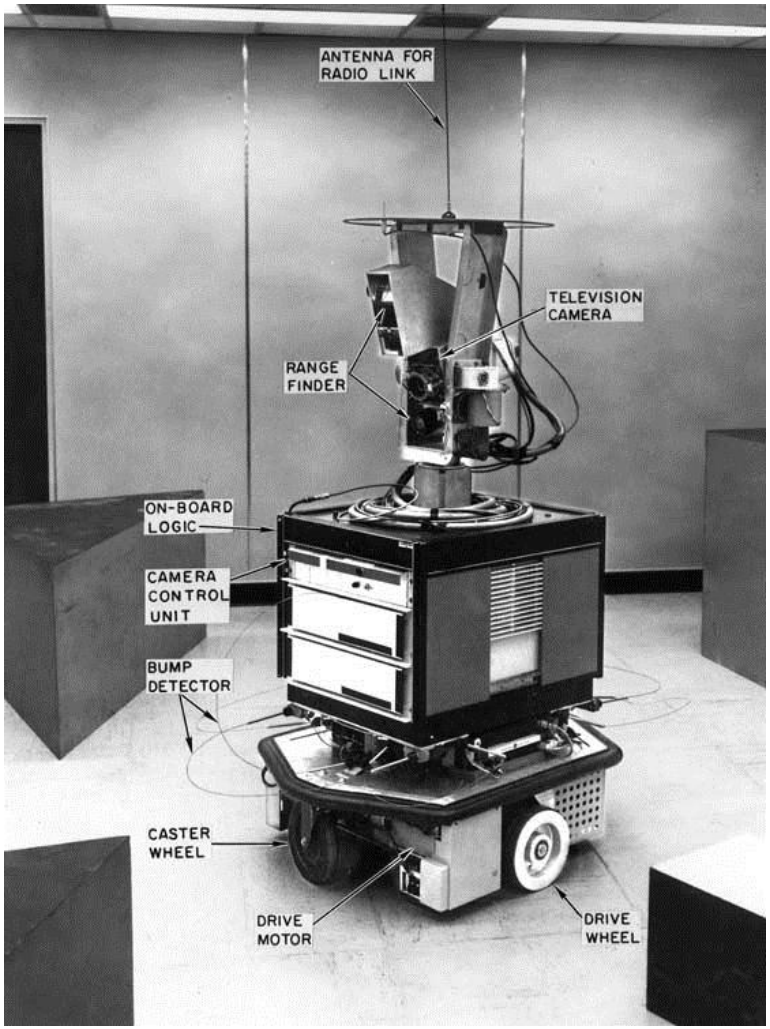
- High quality sensors
  - color depth, dynamic range, light sensitivity, etc.
- Massive information fusion
  - parallelism
  - context-based reasoning
  - active foveation and selective attention
  - selective sensor fusion over space, capability and time
  - tuned feedback from interpretation to first computation
  - elegant and gradual failure characteristics

3. Machine Vision

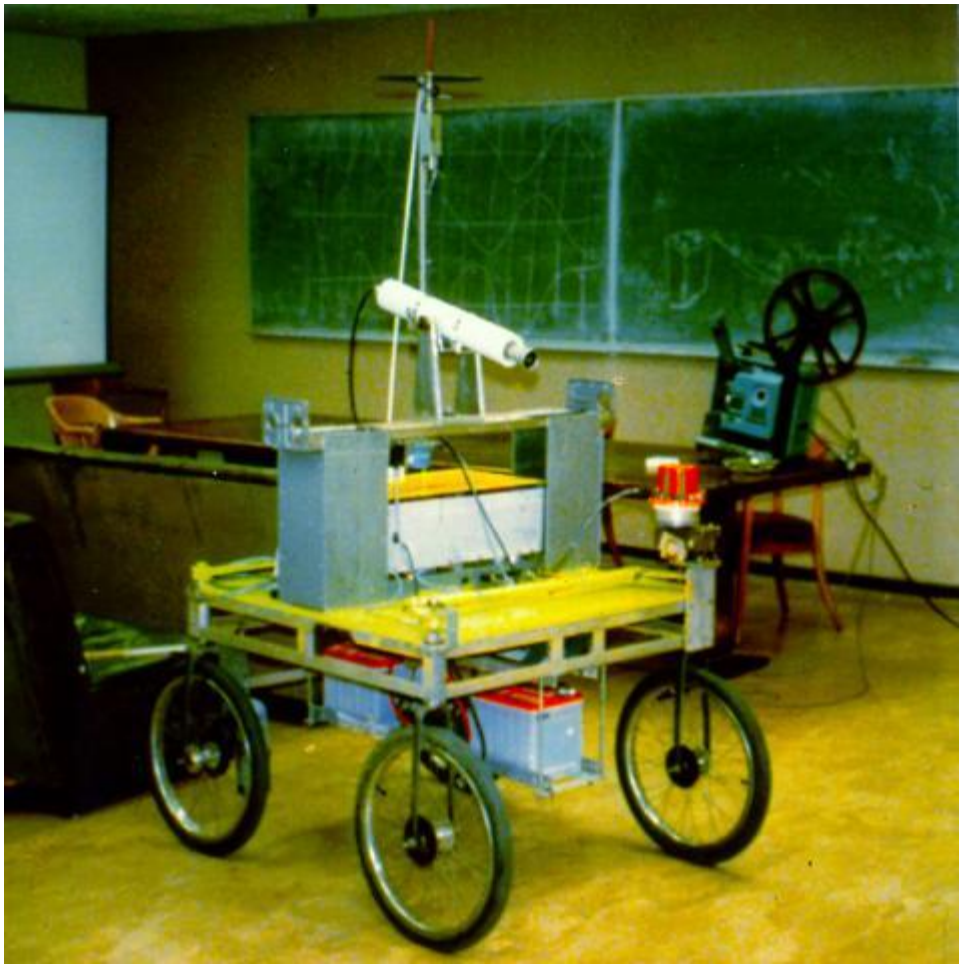
- Poor-performance sensors
- 8/24 bits of color, little dynamic range, inaccuracy and warp, inconstant properties

- Narrow, shallow, fragile
- serial information processing
- information context typically as assumptions that violate
- little sensor fusion across type
- little sensor feedback loops across levels of interpretation
- very little temporal filtering and interpretation

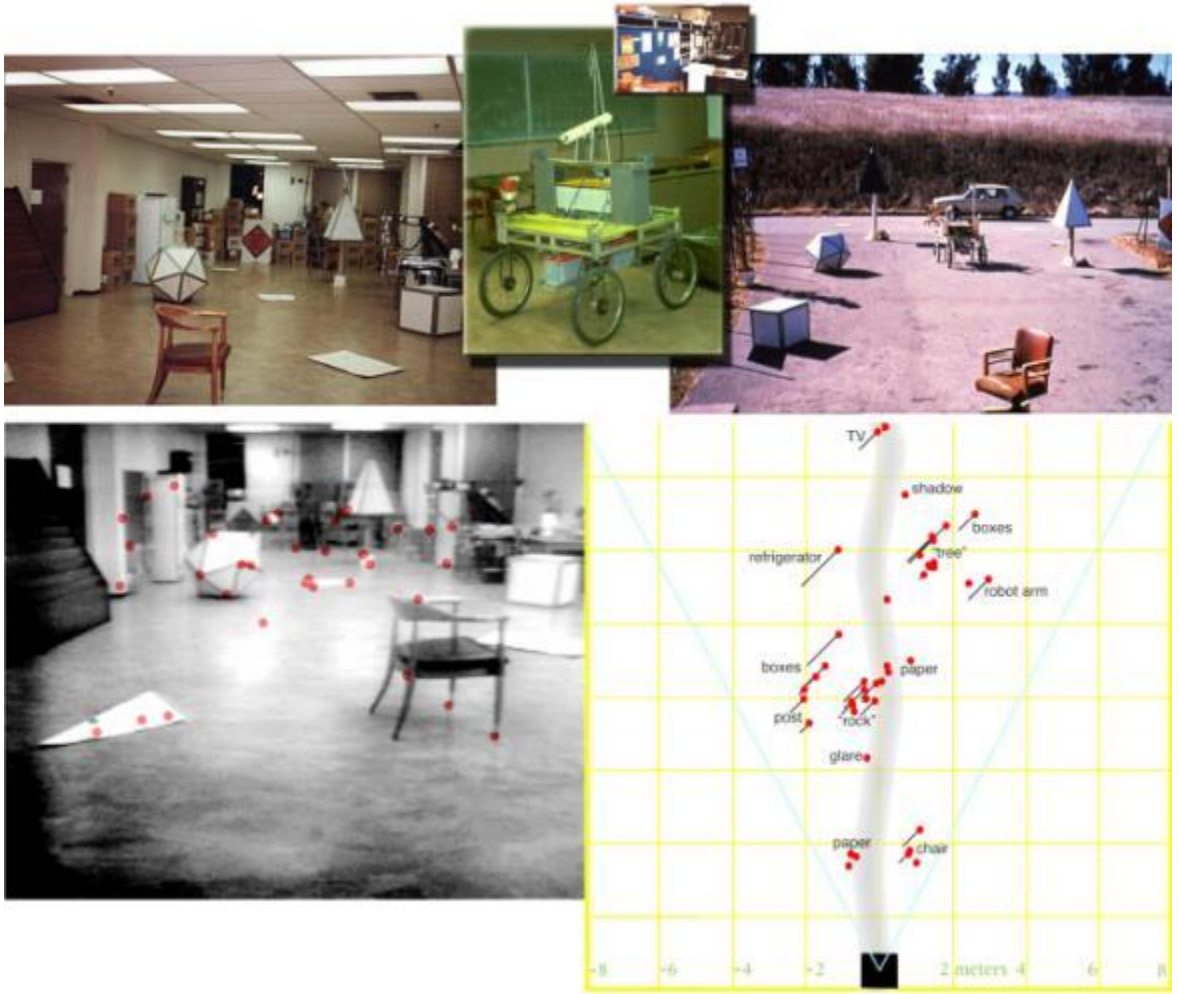
Origins: Shakey



Origins: The Stanford Cart



Origins: The Stanford Cart



Passive versus Active Tradeoff

The Passive/Active Design Question

- Sufficiency of natural contrast
- Interference between multiple robots
- System works in the dark
- System works in bright sunlight

Visual Ranging for Social Interaction

- Totally safe obstacle detection
- Human-body spatial interaction
- Arms and gesture recognition
- Human-designed environment engagement

Vision-based Rangefinding

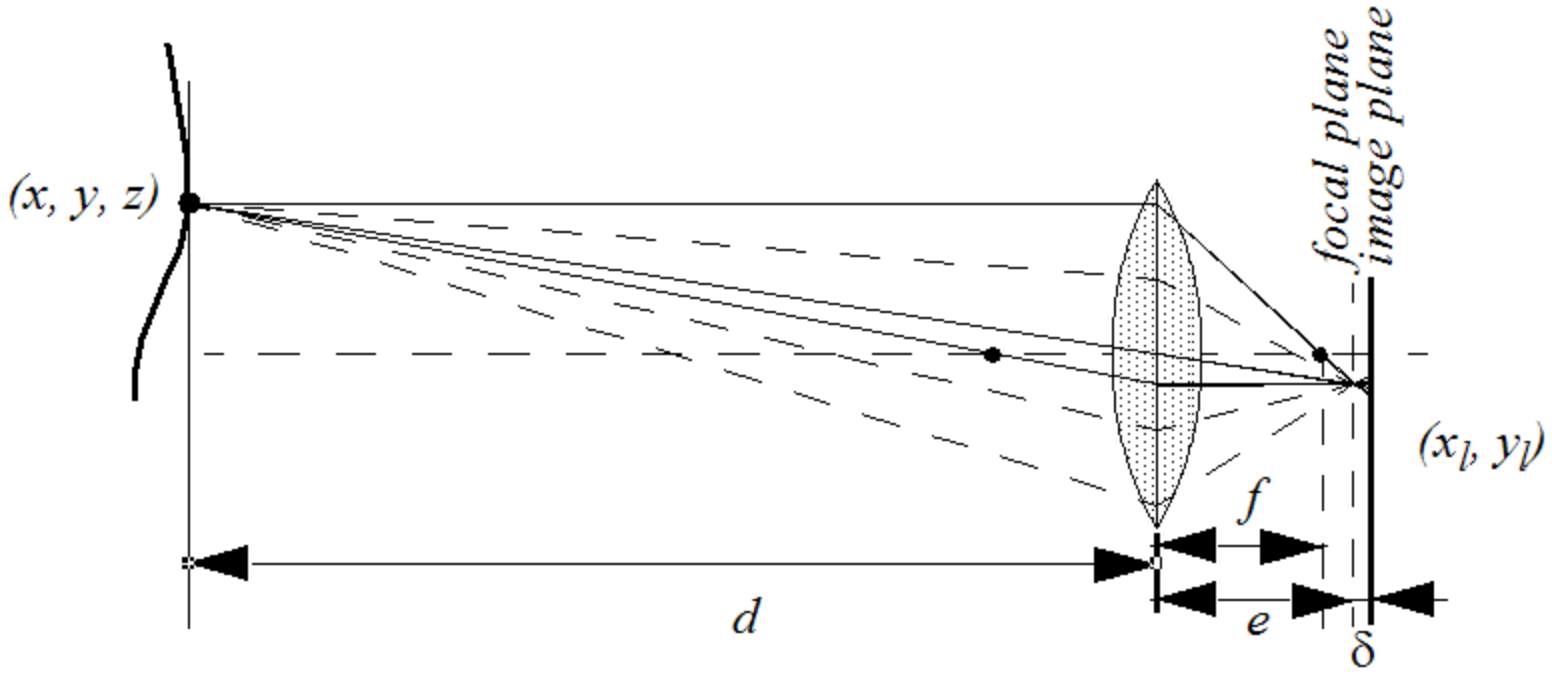
- ***Imaging chips collapse a 3D world onto a 2D plane***

- Range inference from world knowledge / logical reasoning
- Range inference from camera parameters
- Range inference from disparity / matching

Depth from Defocus

- 1/f = 1/d + 1/e

$$R = \frac{L\delta}{2e}$$

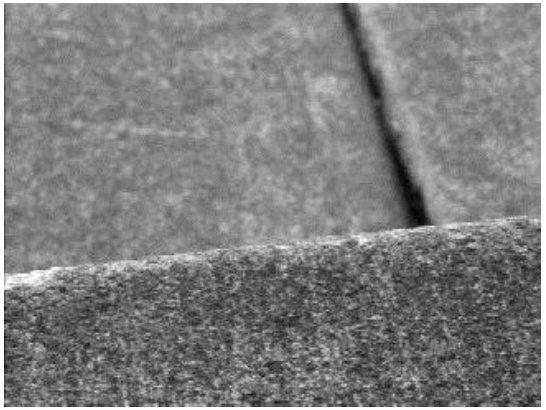


Depth from Defocus

Pinhole camera – no blurring

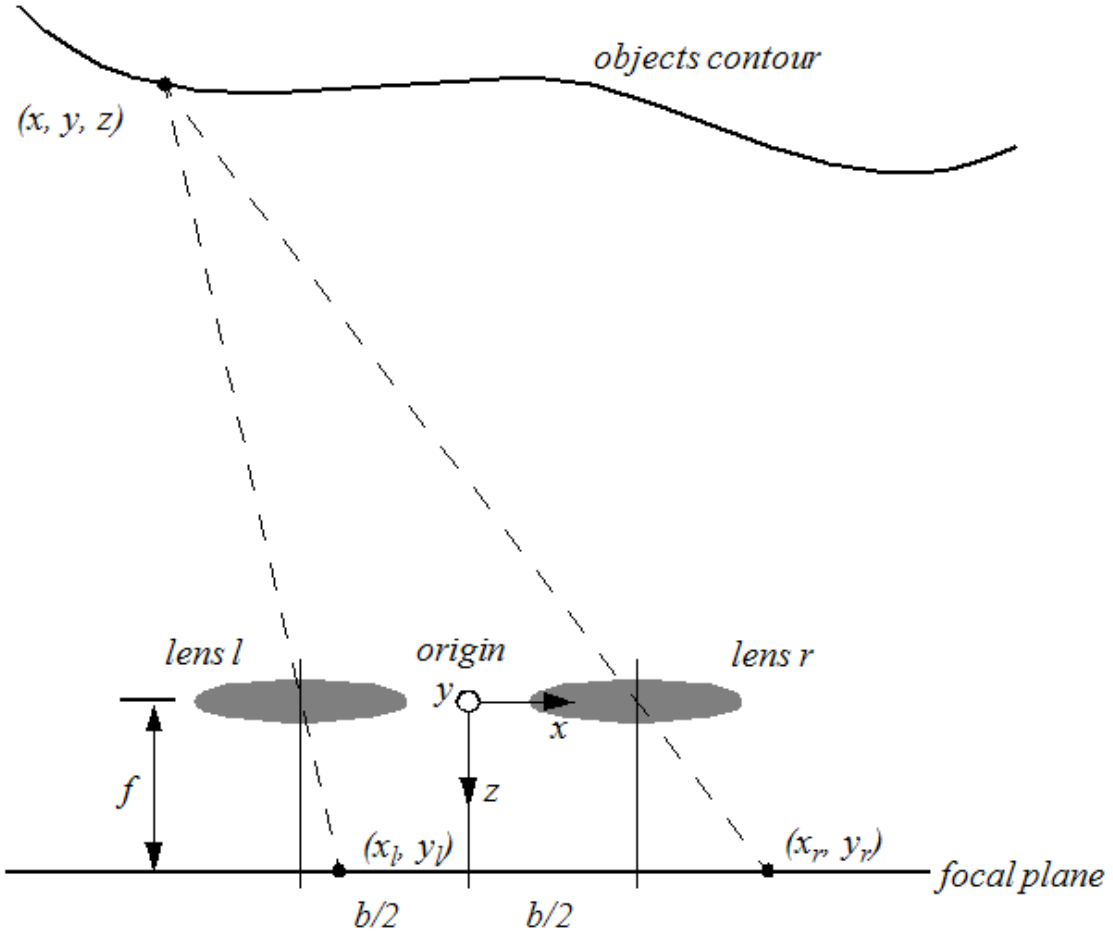
Blur circle sensitivity inversely proportional to distance

To calculate distance we must know focused image



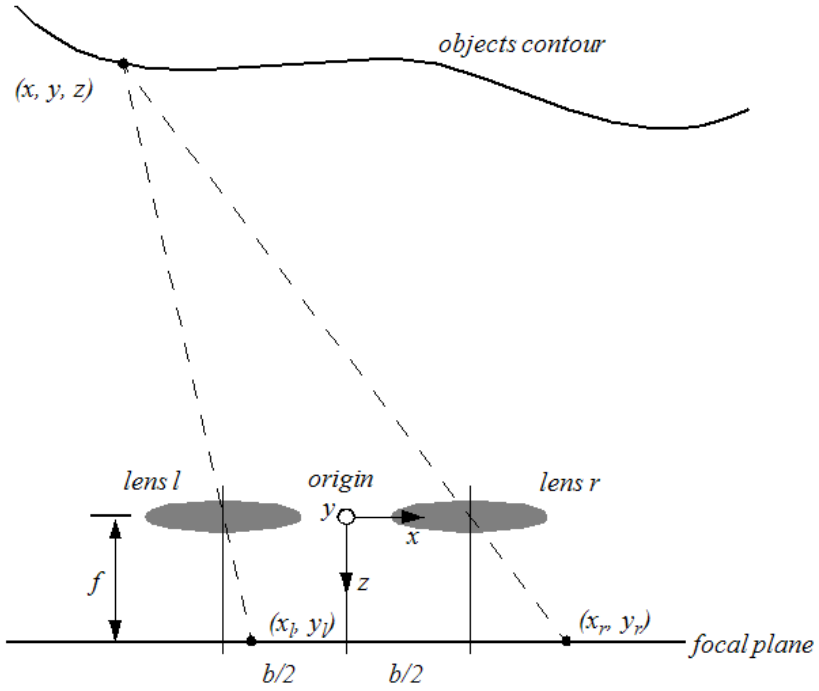
Depth from Defocus

Depth from Disparity



Depth from Disparity

$$x = b\frac{(x_l + x_r)/2}{x_l - x_r} \quad ; \quad y = b\frac{(y_l + y_r)/2}{x_l - x_r} \quad ; \quad z = b\frac{f}{x_l - x_r}$$



- Distance is inversely proportional to disparity
- Disparity is proportional to baseline
- Large baselines offer a tradeoff across range

The Feature Challenge

Features must:

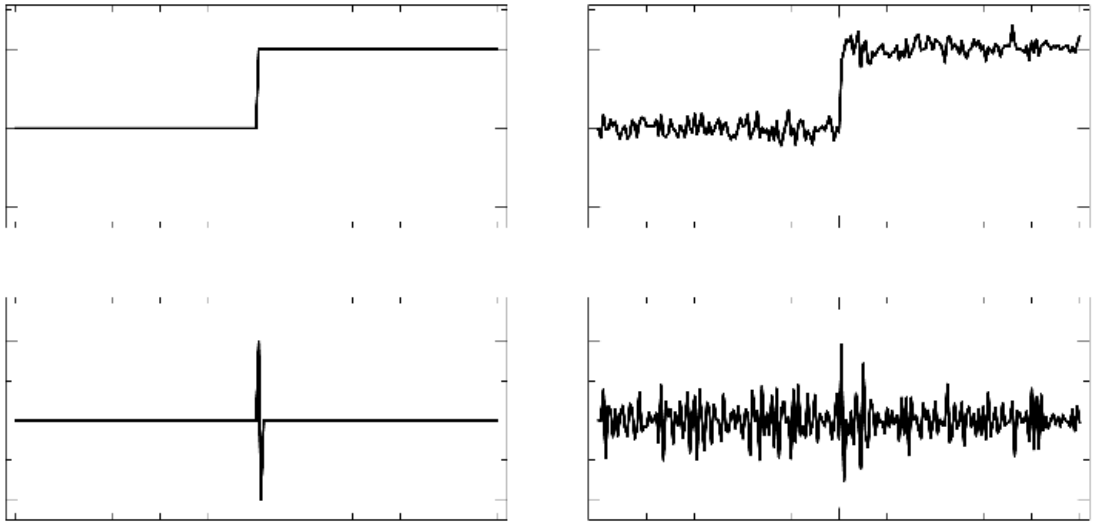
- provide sufficient density
- match across small viewpoint changes
- match across partial occlusions
- identify confidence

Features must not:

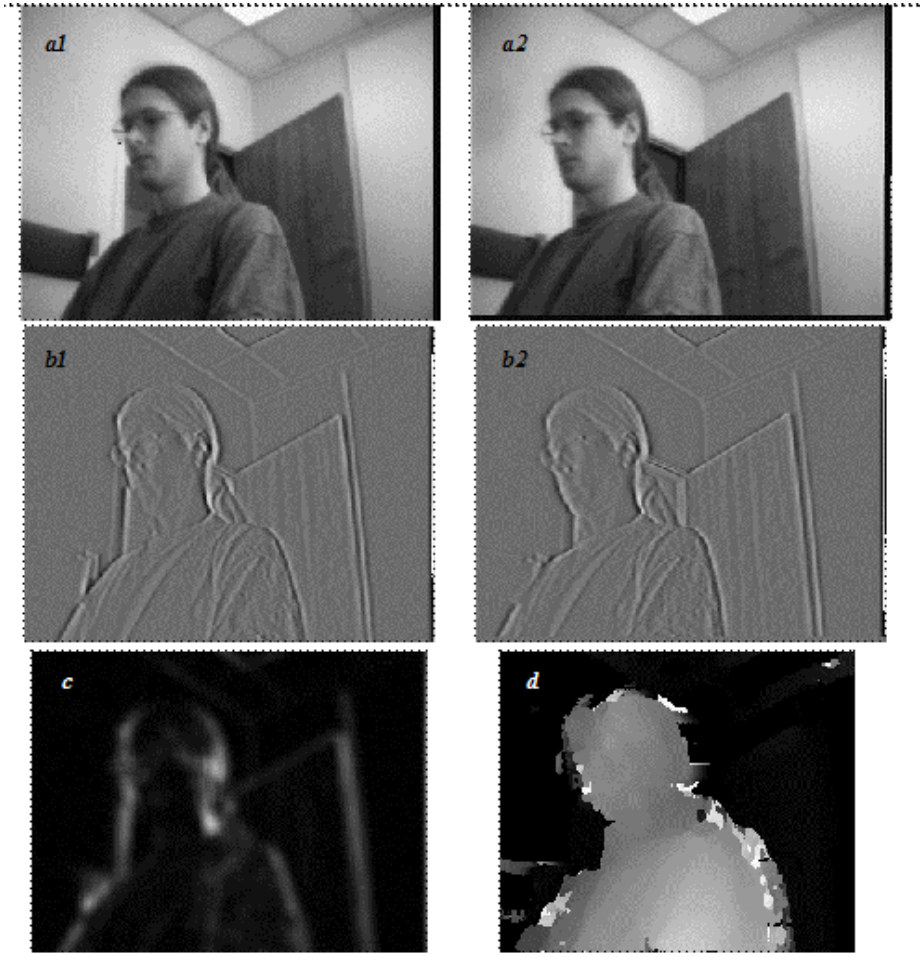
- trigger false positive matches
- prove too sparse for the robot's task
- require on-line human tuning

Example: ZLoG

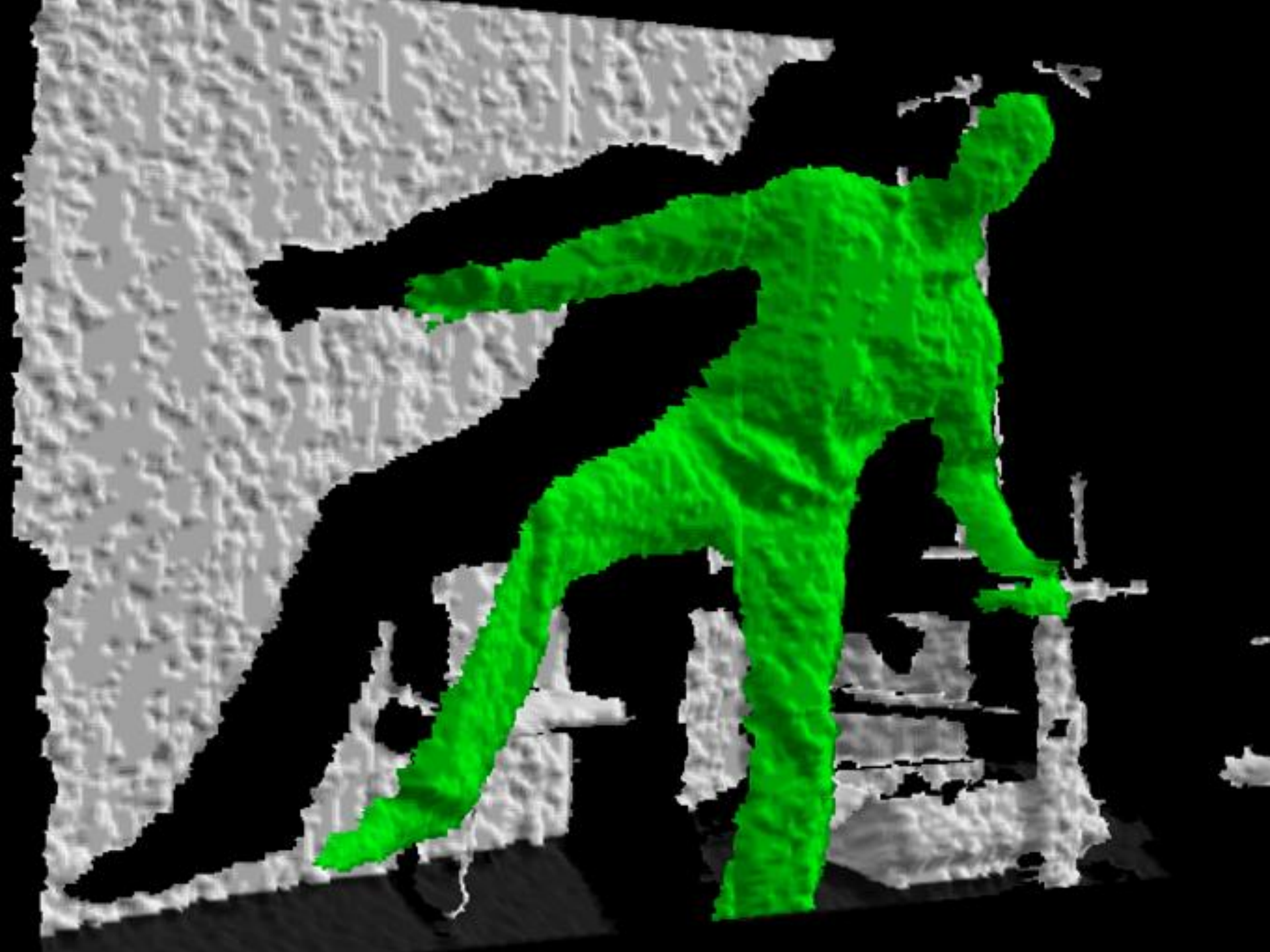
- Zero crossings of Laplacian of Gaussian
- Laplacian: “second derivative convolution”
- Gaussian: “smoothing convolution”
- Zero crossings: a *sharp* feature for interpolation



Stereo: Pictorial Example



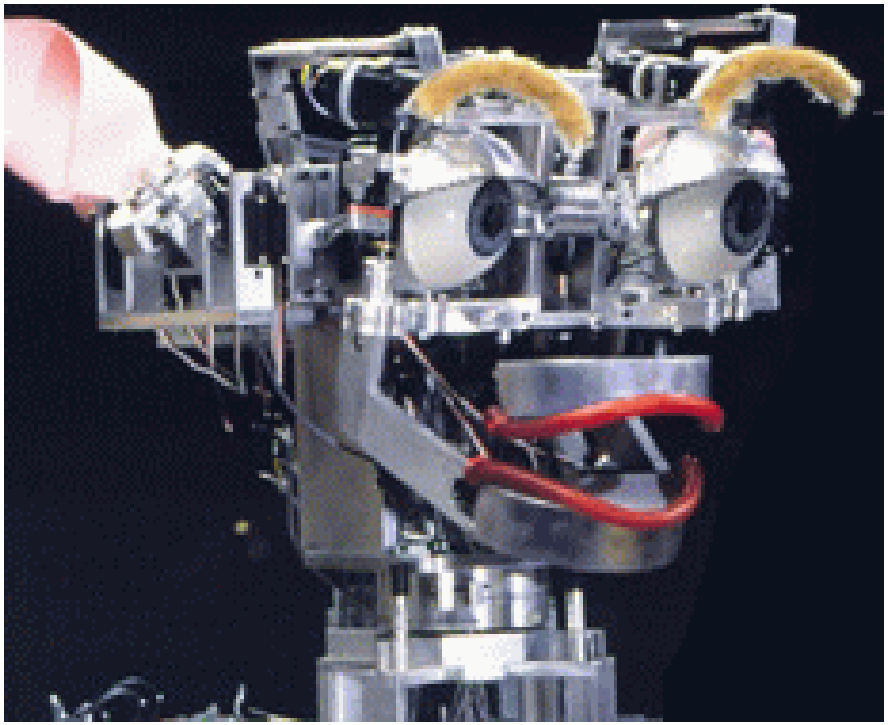
Active Rangefinding



HRI Vision: the special-case approach

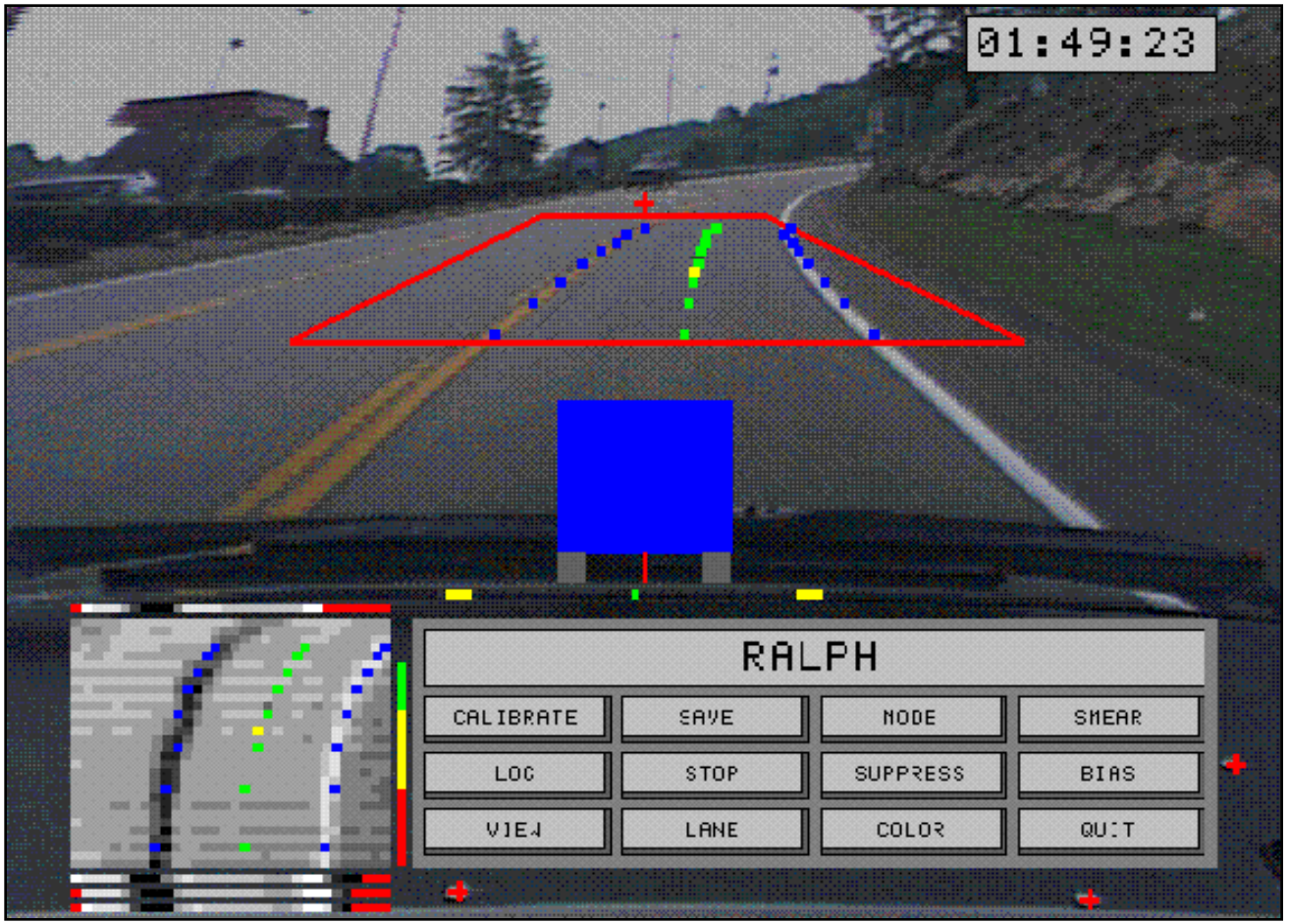
Example: Cueing in Kismet

- Color-based human-robot interaction
- Cueing, orthogonal events, child-based interaction
- Challenges: constancy, illumination, human expectation

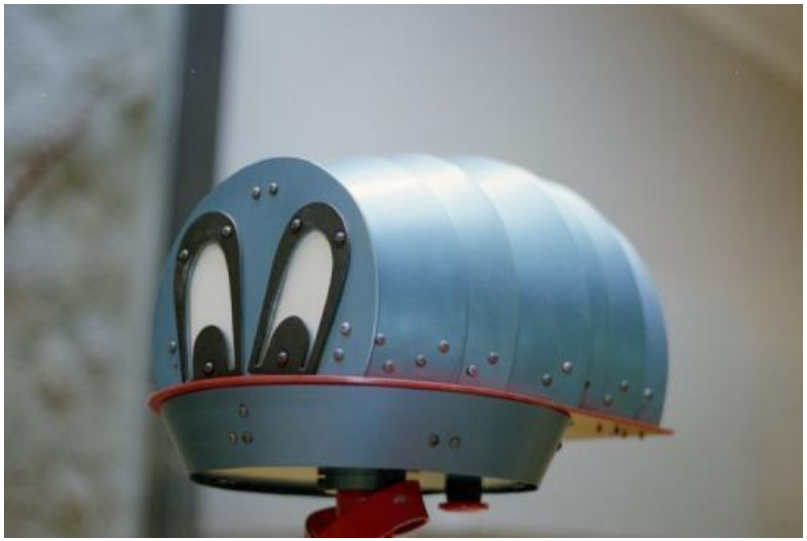


Motivational example: RALPH

Navlab on Streets



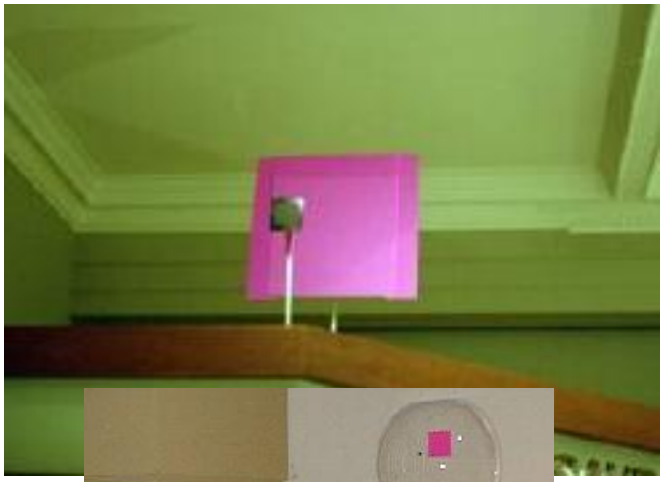
Museum Edubot - *Chips*



- Carnegie Museum of Natural History
- Autonomy
  - 5 years, > 500 km navigated, auto-docking
  - MTBF convergence at 1 week
  - Proactive health state identification

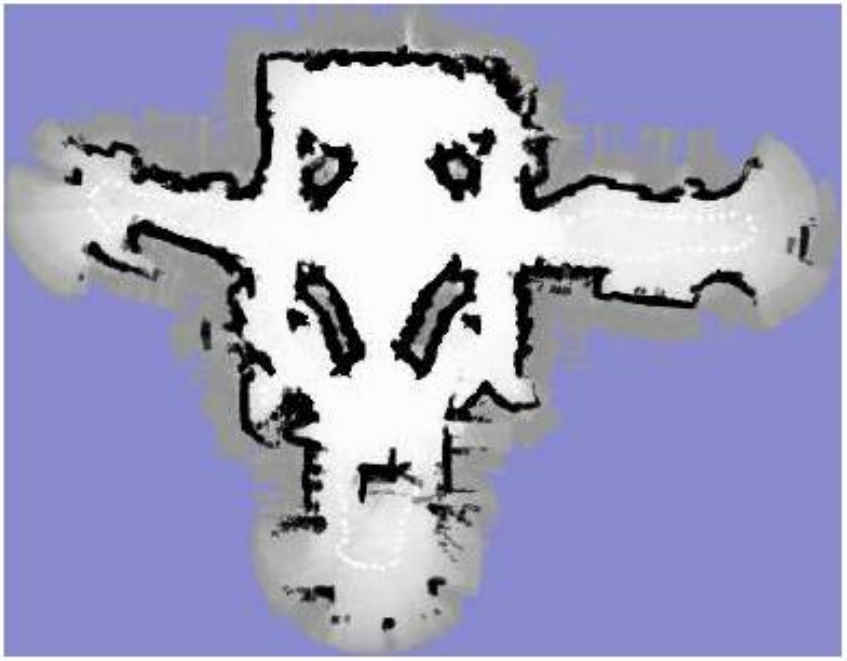
Museum Edubot - *Chips*

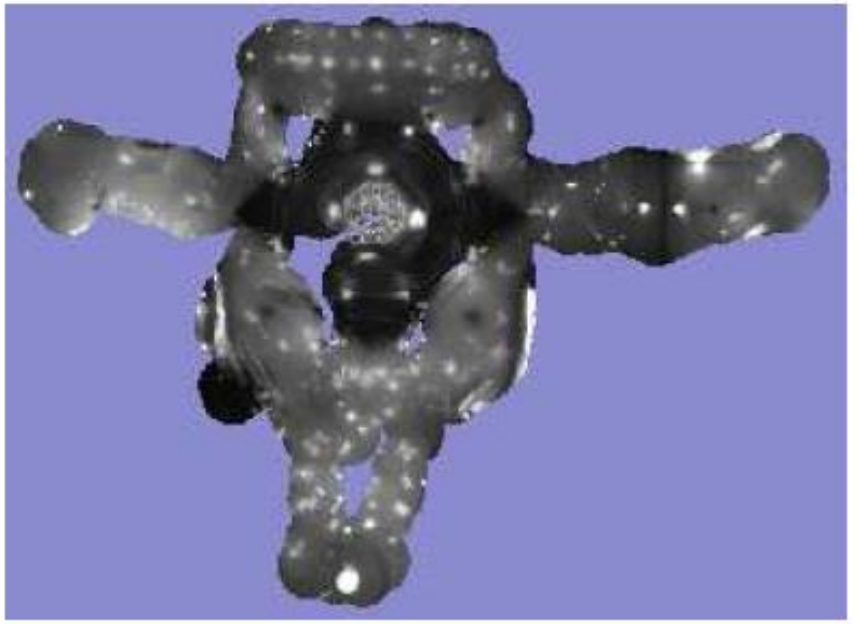
Landmarks: Visual Fiducials



Minerva: an example of “focused vision”

Minerva: an example of “focused vision”





When special-case fails…

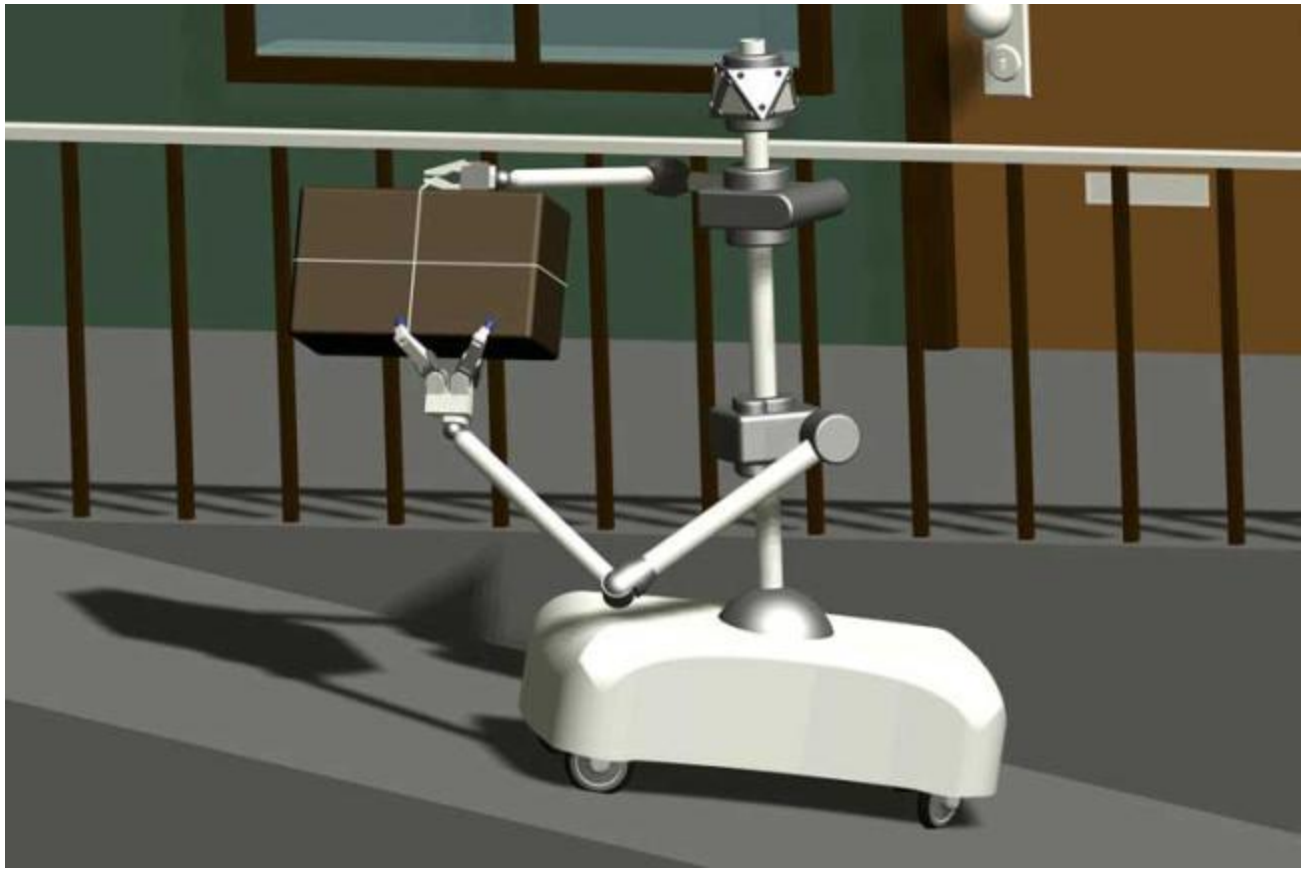
SLAM

Visual SLAM Considerations

- *Repeatable “landmark” recognition*
- *Feature locale*
- *Map-making*
- *Tracking robot position*

The Future of Visual Navigation

- *Hans Moravec's stereo-based voxel grid*



SIFT

- **Features**: image contents coded so they can be found again on other images of same scene, …

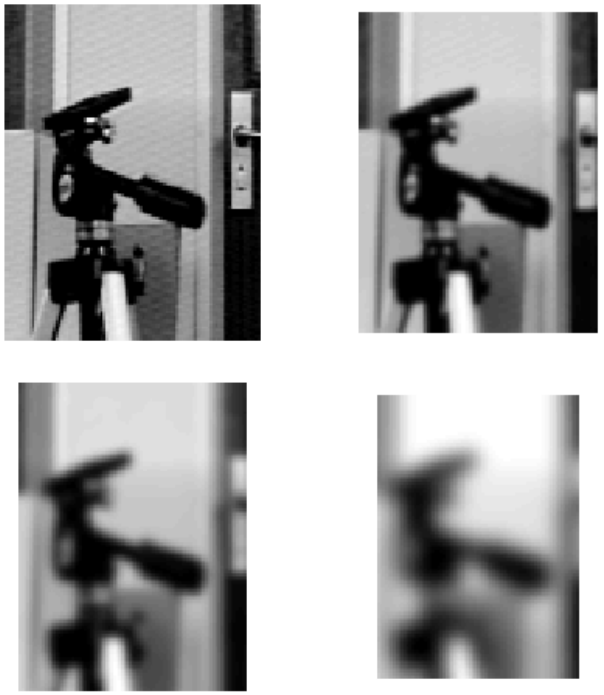
- **Invariant**: …despite many changes:
  - **rotation**, translation
  - camera viewpoint: **scale**, perspective
  - illumination
  - noise
  - occlusion

➡ Image matching by comparing invariant features

Notion of **Interesting points** and **Keypoints**

Gaussian pyramid

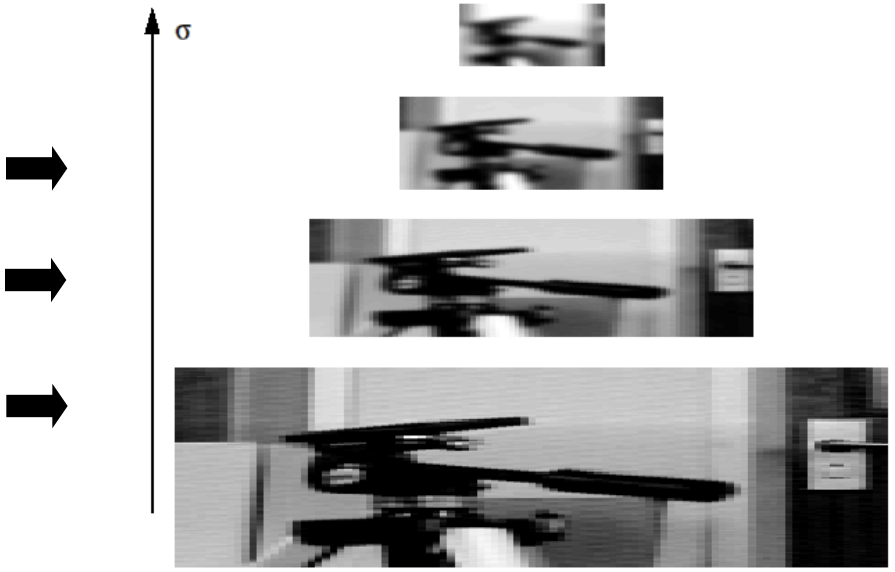
Scale ≡ smoothing parameter σ



Increasing sigma

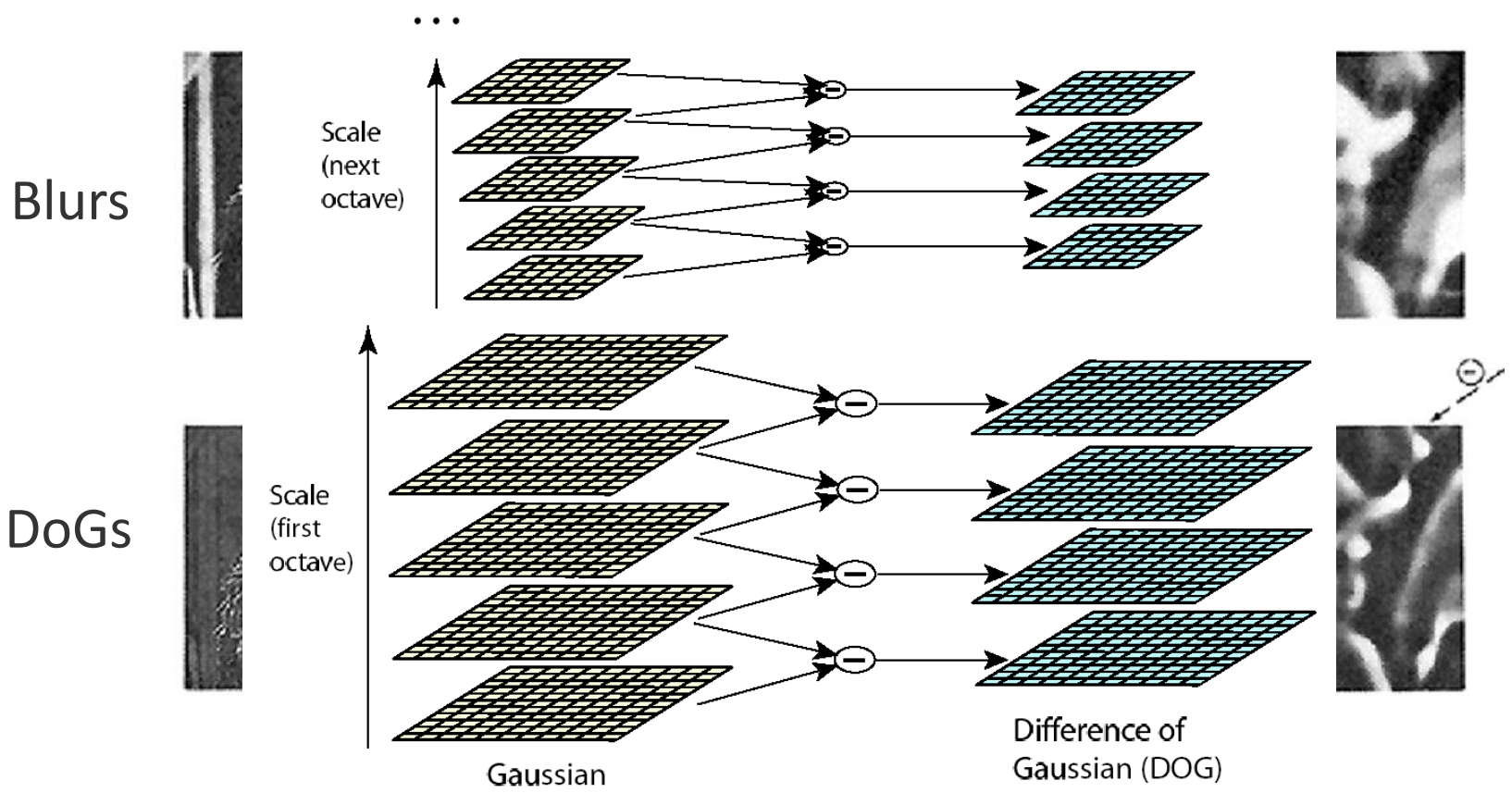
Increase σ -> no need to retain all pixels

Stored image can be reduced in size

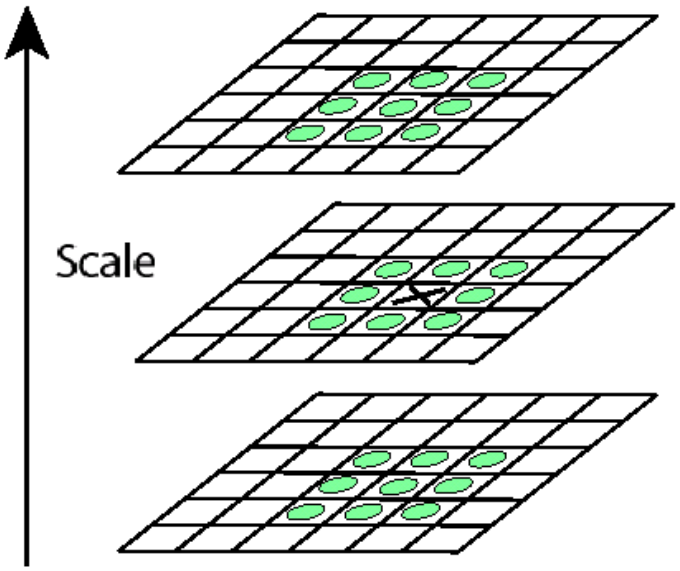


Gaussian pyramid

Gaussian Pyramid processed one octave at a time



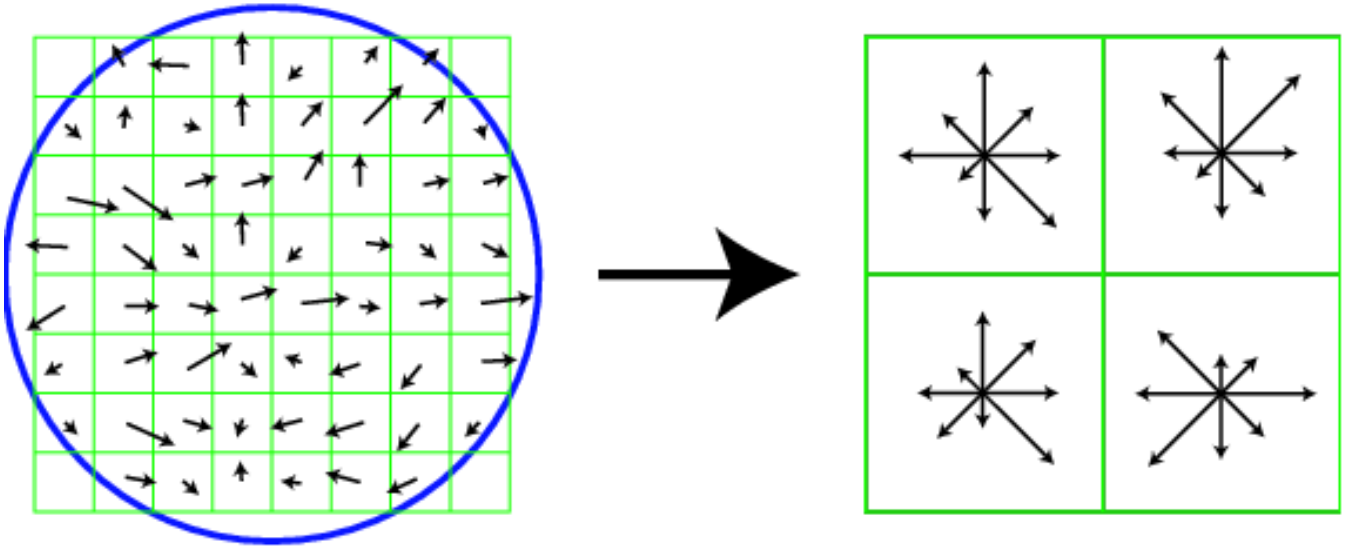
- Detect maxima and minima of difference-of-Gaussian in scale space
- Reject points lying on edges
- Fit a quadratic to surrounding values for sub-pixel and sub-scale interpolation



Thresholded image gradients are sampled over 16x16 array of locations in scale space

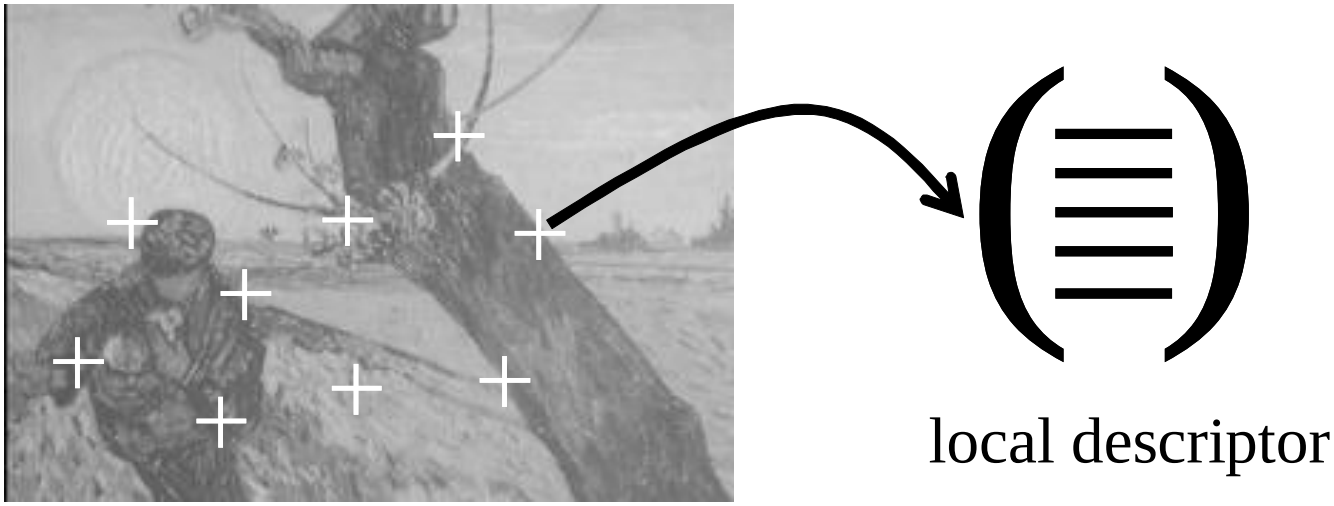
Create array of orientation histograms

8 orientations x 4x4 histogram array = 128 dimensions



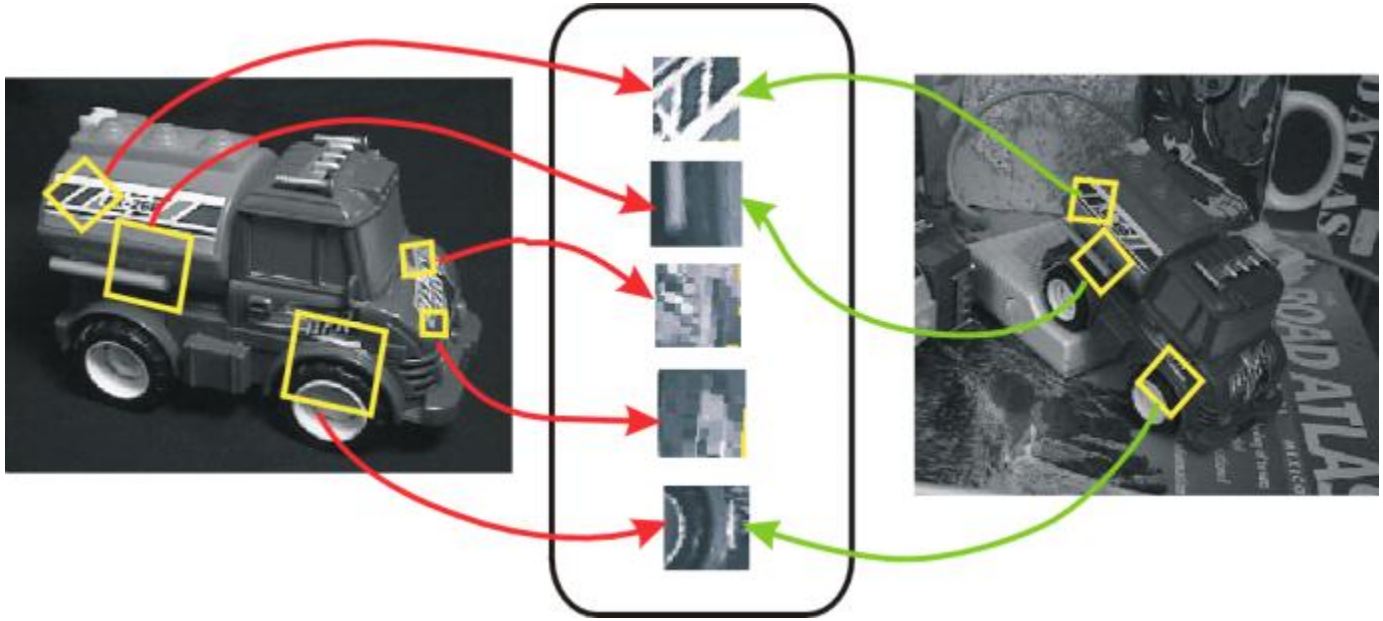
Sampled regions located at interest points

Local invariant descriptors to scale and rotation



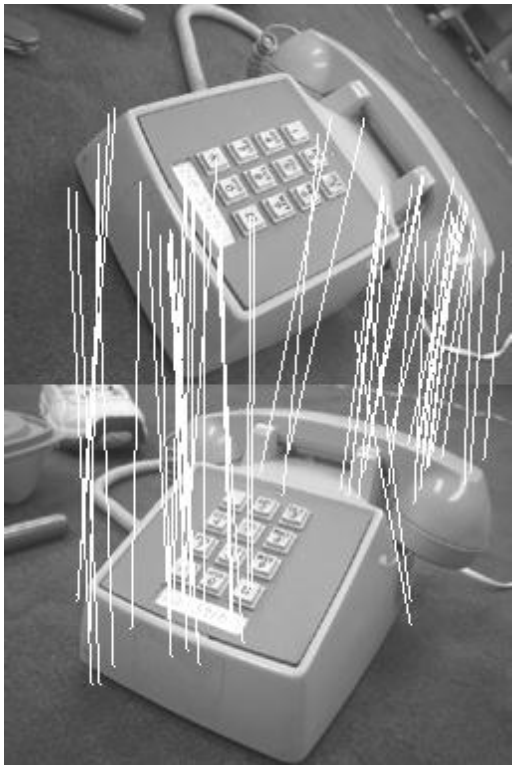
Local:	robust to occlusion/clutter	+	no segmentation
Invariant:	to image transformations	+	illumination changes

- Very powerful method developed by David Lowe, Vancouver
- Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters



SIFT Features

SIFT



Example: K9 Science Rover

Example: K9 Science Rover's SIFT

4. Social Vision State of Art

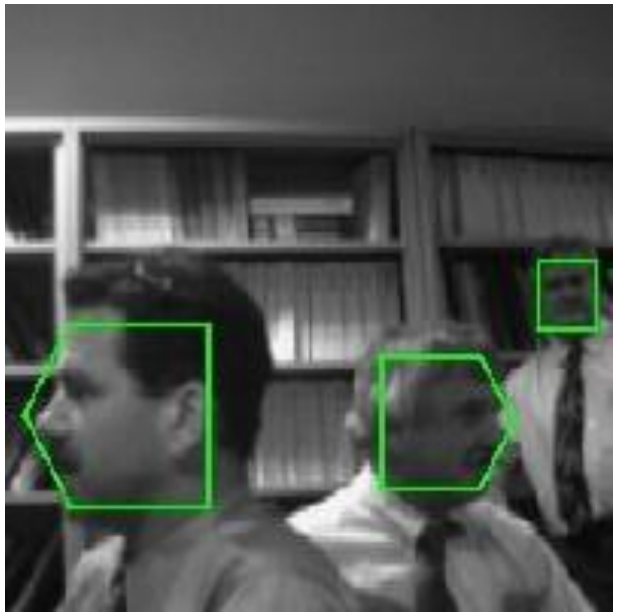
- Face detection, recognition
- Speech understanding
- Gesture understanding

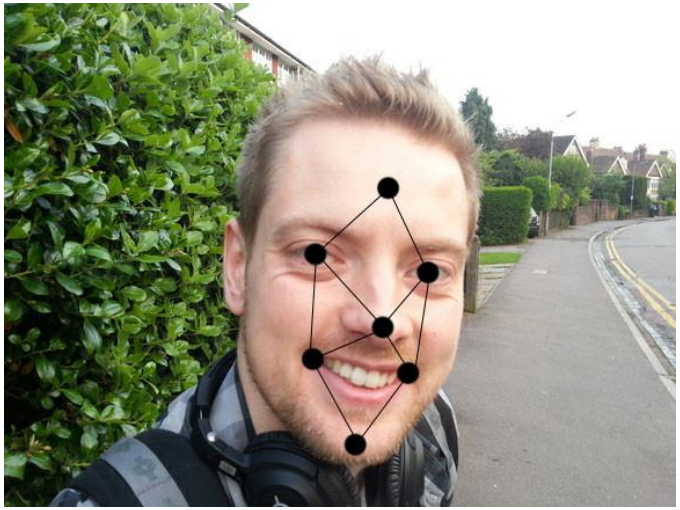
Face Detection

- How would you detect faces in images?

Face Detection

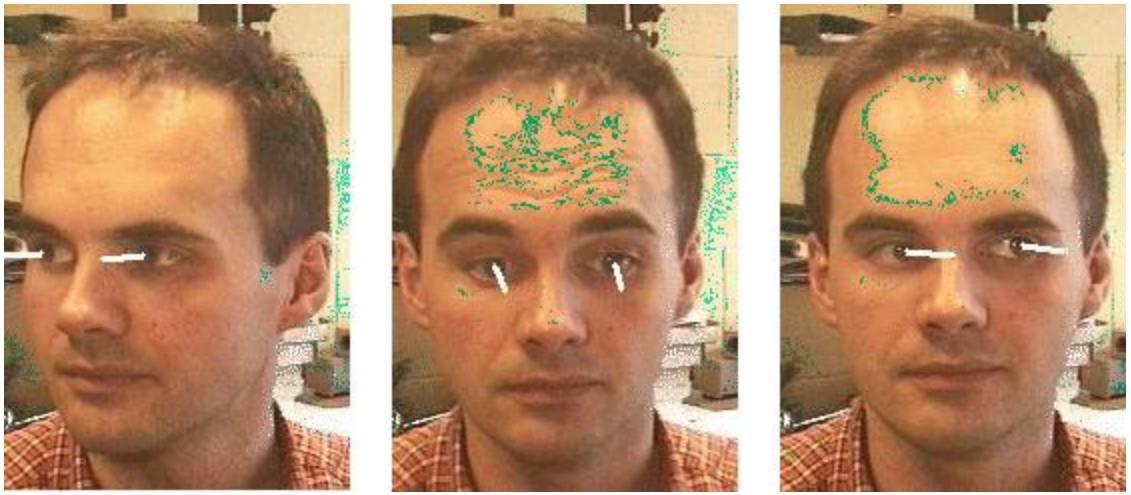
- How would you detect faces in images?





Face Detection

- How would you detect faces in images?



Expression Detection

First Person Vision

Speech and Gesture Understanding

- Time for some fun:
  - http://www.youtube.com/watch?v=1s-PiIbzbhw