Vision and Robotics in LCI

Jim Little
for
David Lowe
Alan Mackworth
Bob Woodham
(and Nando, Kevin, …)

Background

When we build systems that interact with the world, with humans, and with other agents, we rely upon all of the aspects of cognitive vision:

- knowledge representation
- descriptions of the scene and its constituent objects
- models of agents and their intentions
- learning
- adaptation to the world and other agents
- reasoning about events and about structures
- interpretation of other agents' and users' interactions
- recognition and categorization

Robot Partners

The Robot Partners project focuses on the design and implementation of visually guided collaborative agents, specifically interacting autonomous mobile robots.

- localization and mapping
- user modeling
- interpretation of gestures and actions
- interaction with human agents

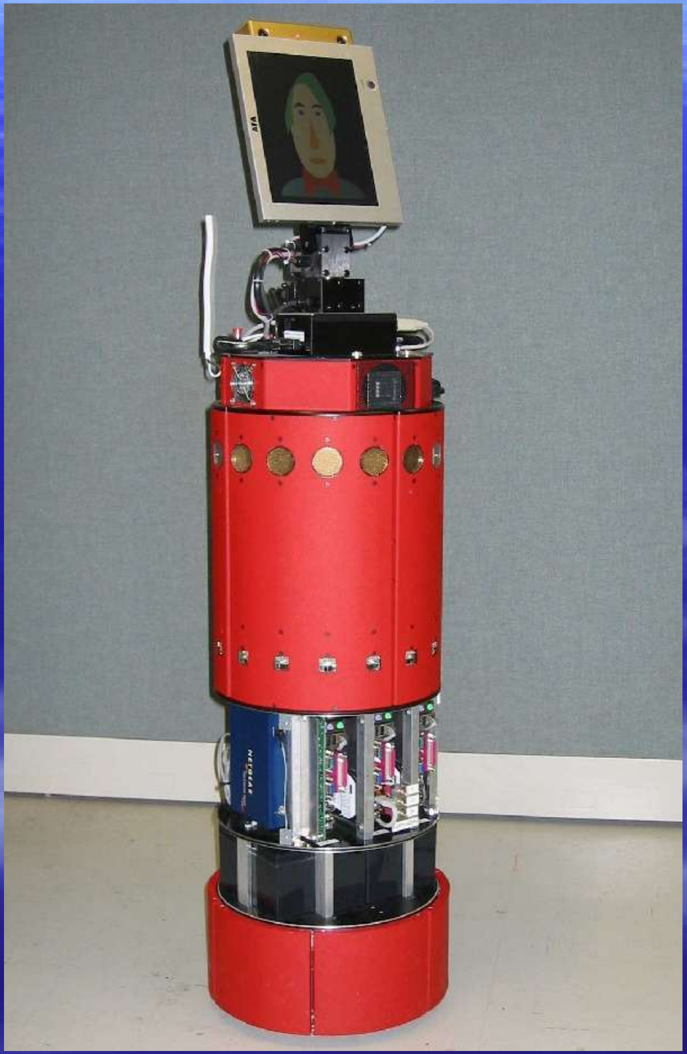
- Jim Little
- Nando de Freitas
- David Lowe
- Alan Mackworth
- Rob Sim

A robot waiter at work

Plan

- I will tell you about what sort of visual capabilities our robots have
- Then I’ll describe several roles the robots have played – jobs they have done.
- Finally I’ll describe our recent work where we learn what visual inputs matter to a robot when interacting with people.

Robot physical structure

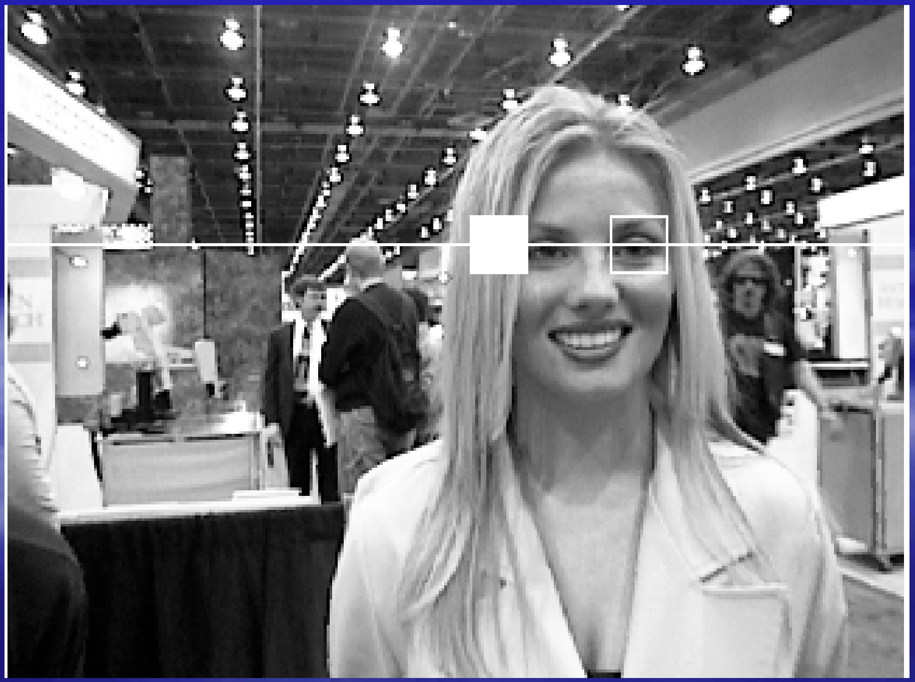


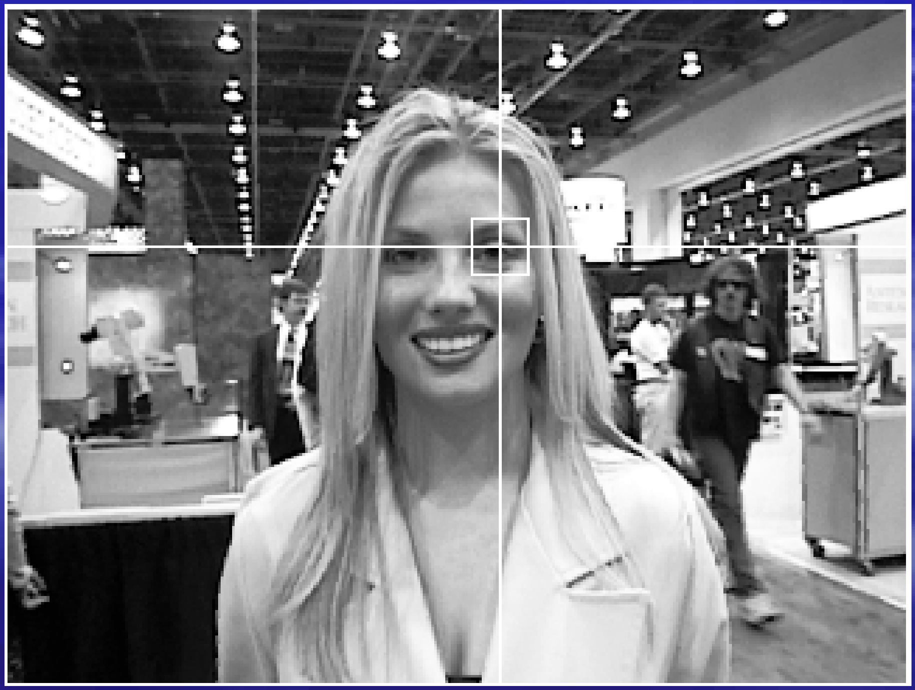
Erik: face display on PTU, Bumblebee stereo (gold), brain belt

Capabilities for visually-guided robots

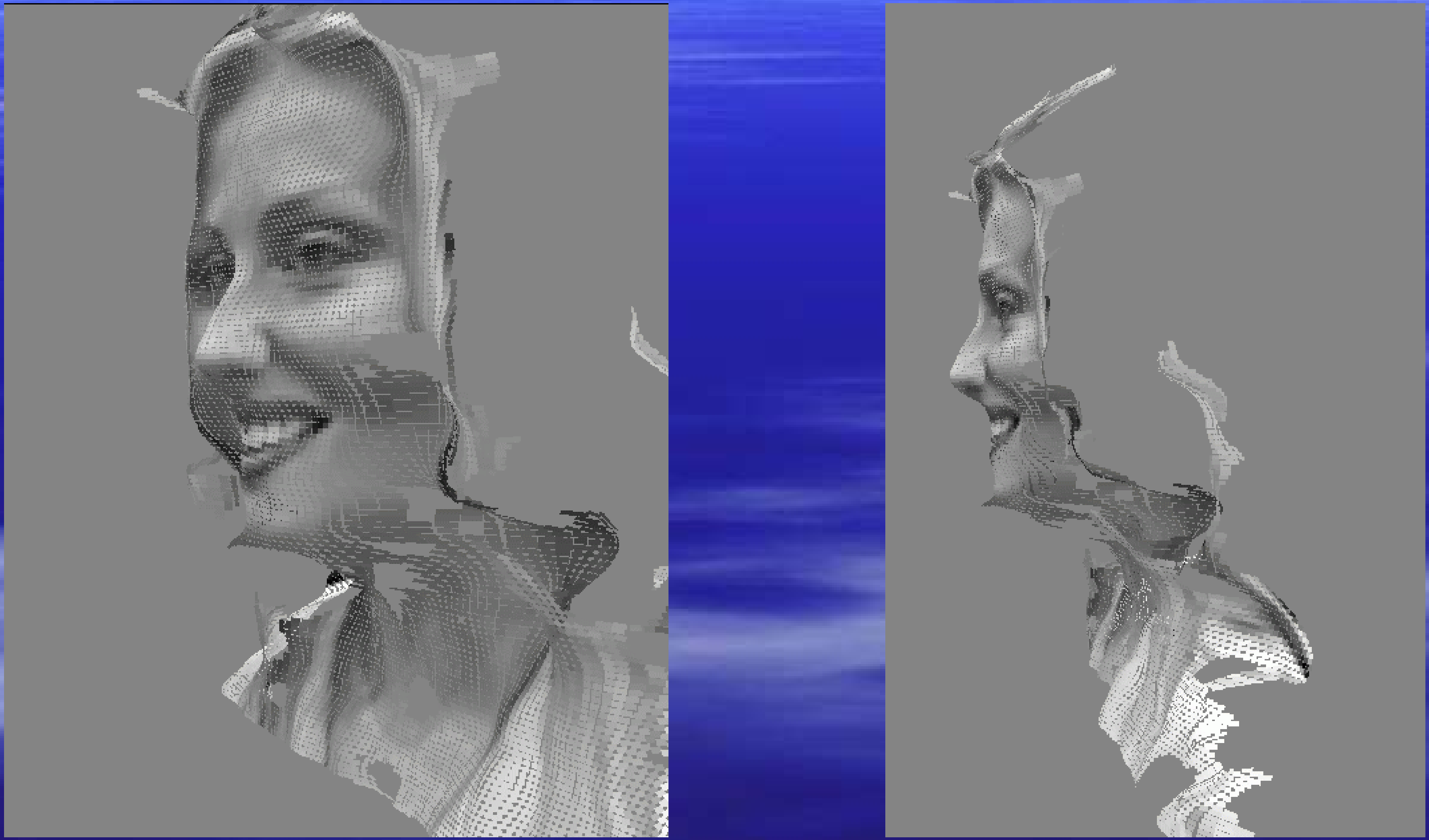
- Seeing in depth
- Navigation - Avoiding dynamic obstacles
- Local Spatial Context
- Localization
- Face Recognition
- Gestures and Expressions
- Speech Recognition and Synthesis
- Interaction
- Associating Names with Visual Structures

Stereo: seeing in depth

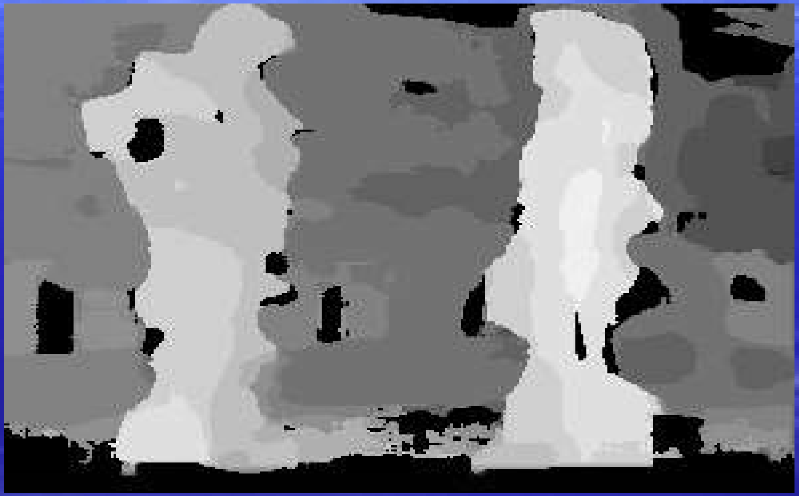


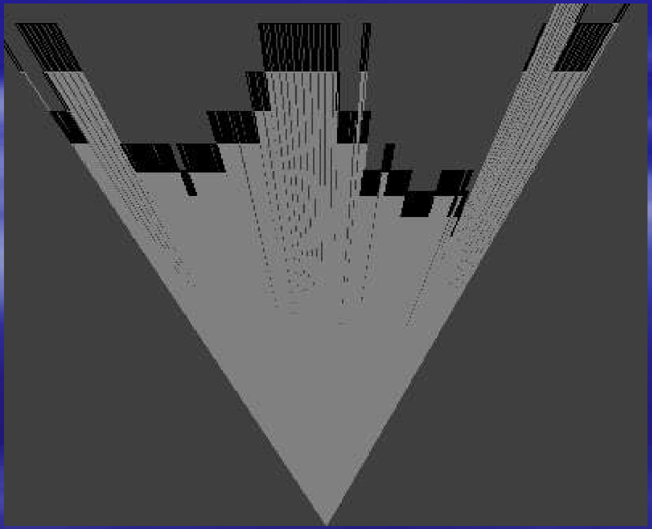


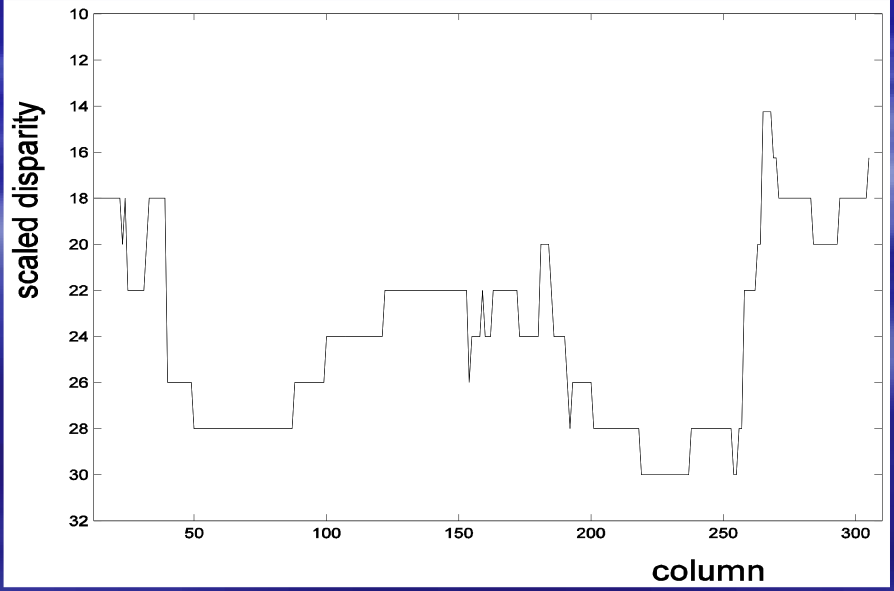
Stereo results (with patchlets)



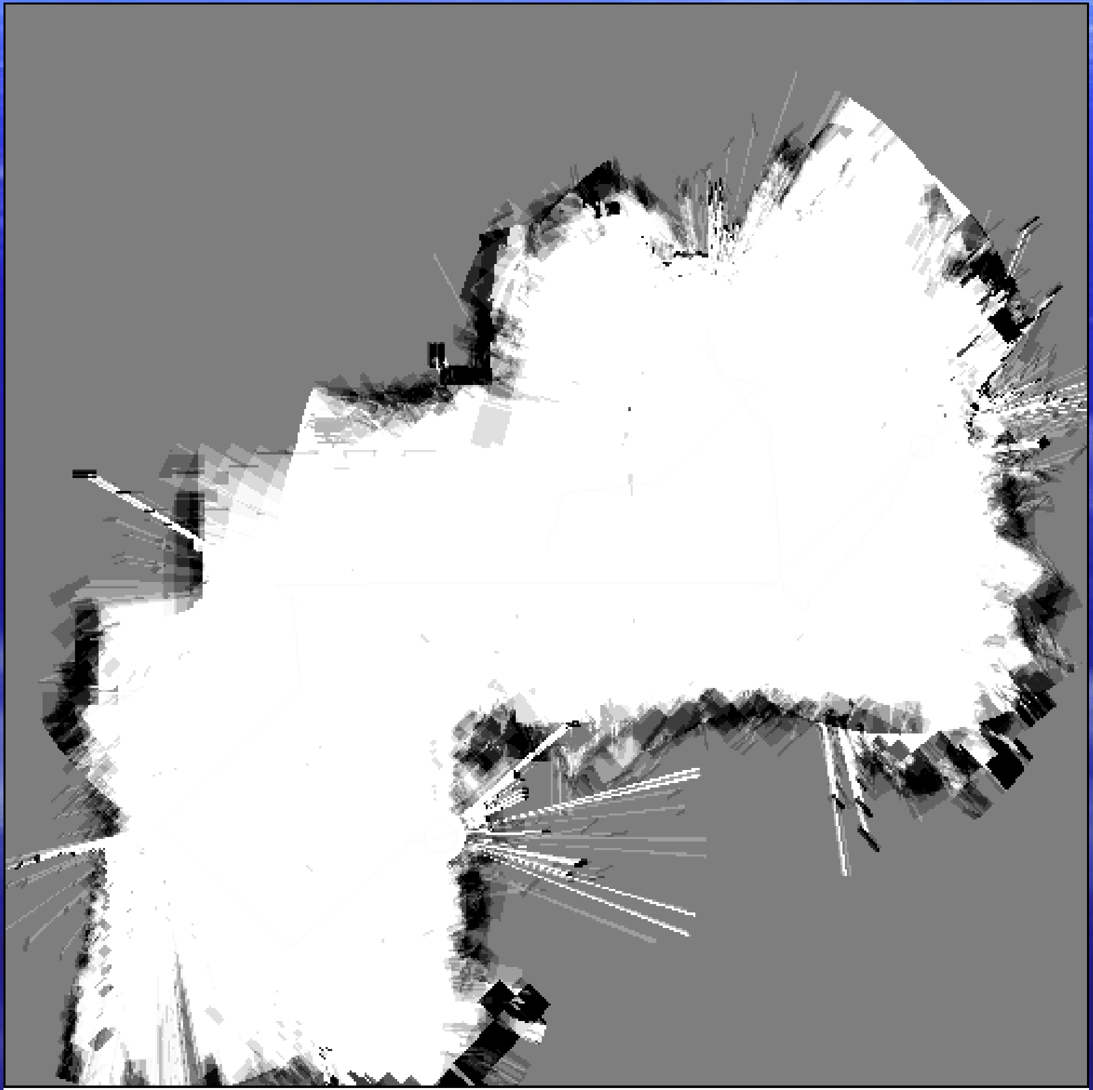
From stereo to maps



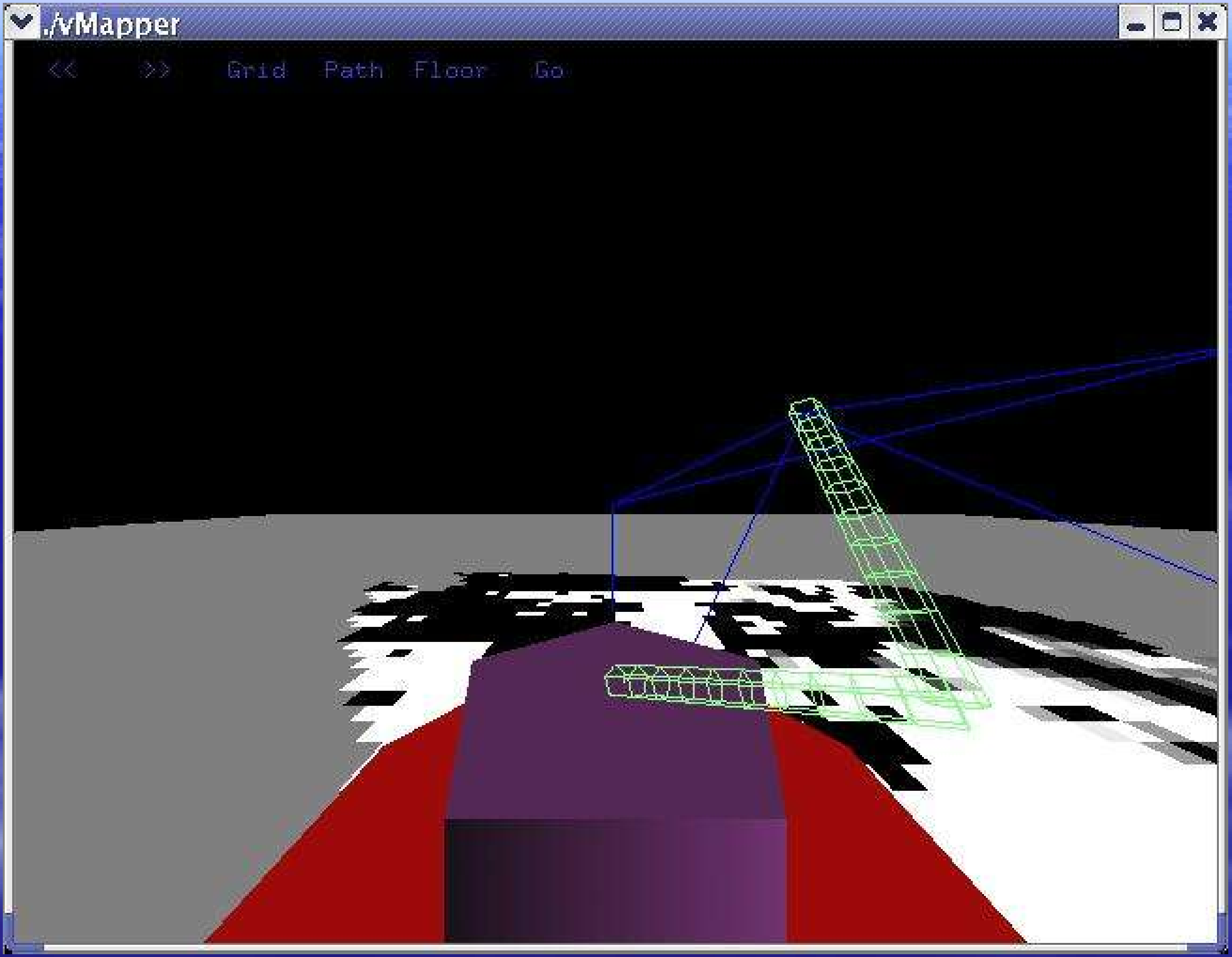




Local map generated by autonomous exploration

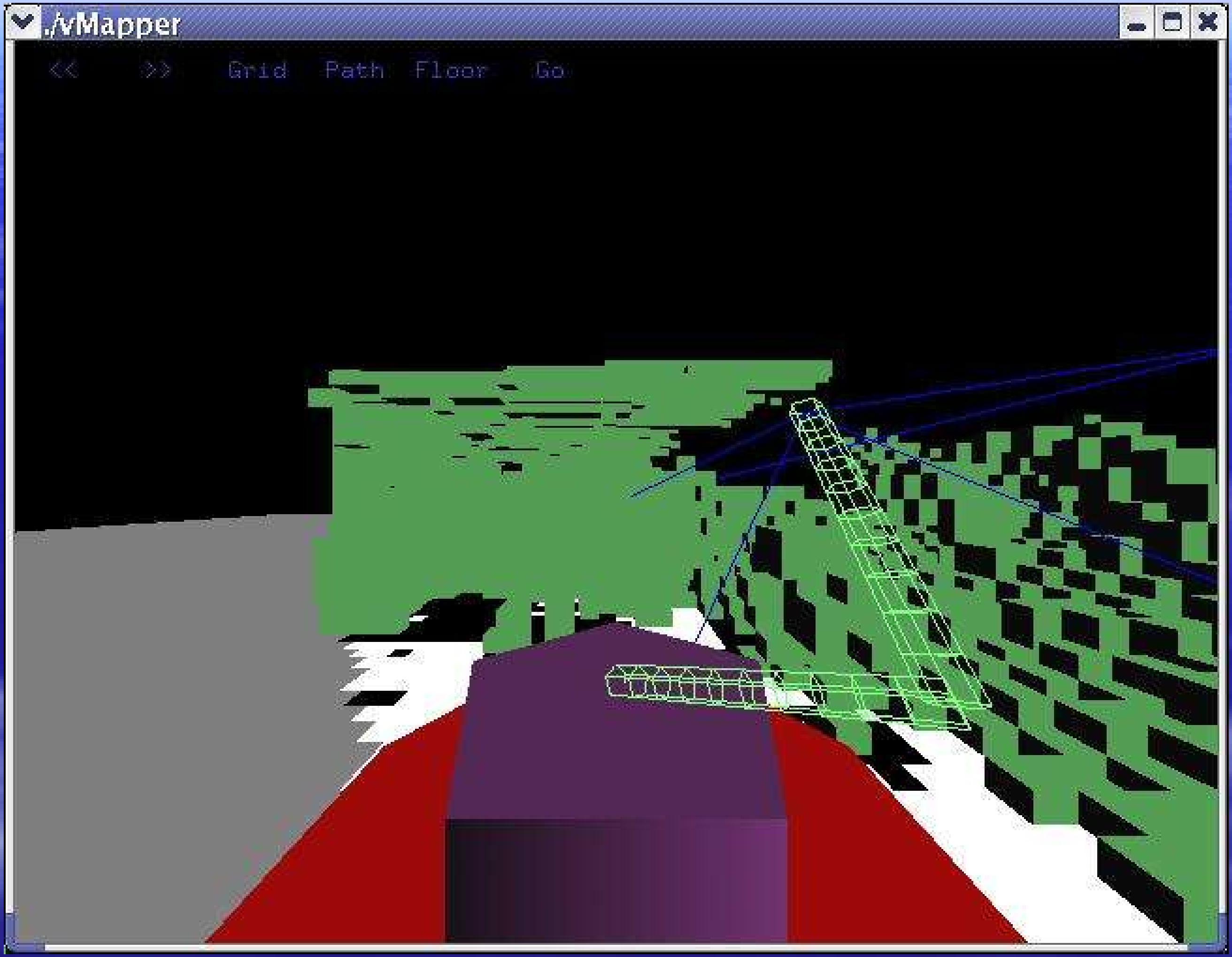


Robot viewing wall and desk with computers to its right.

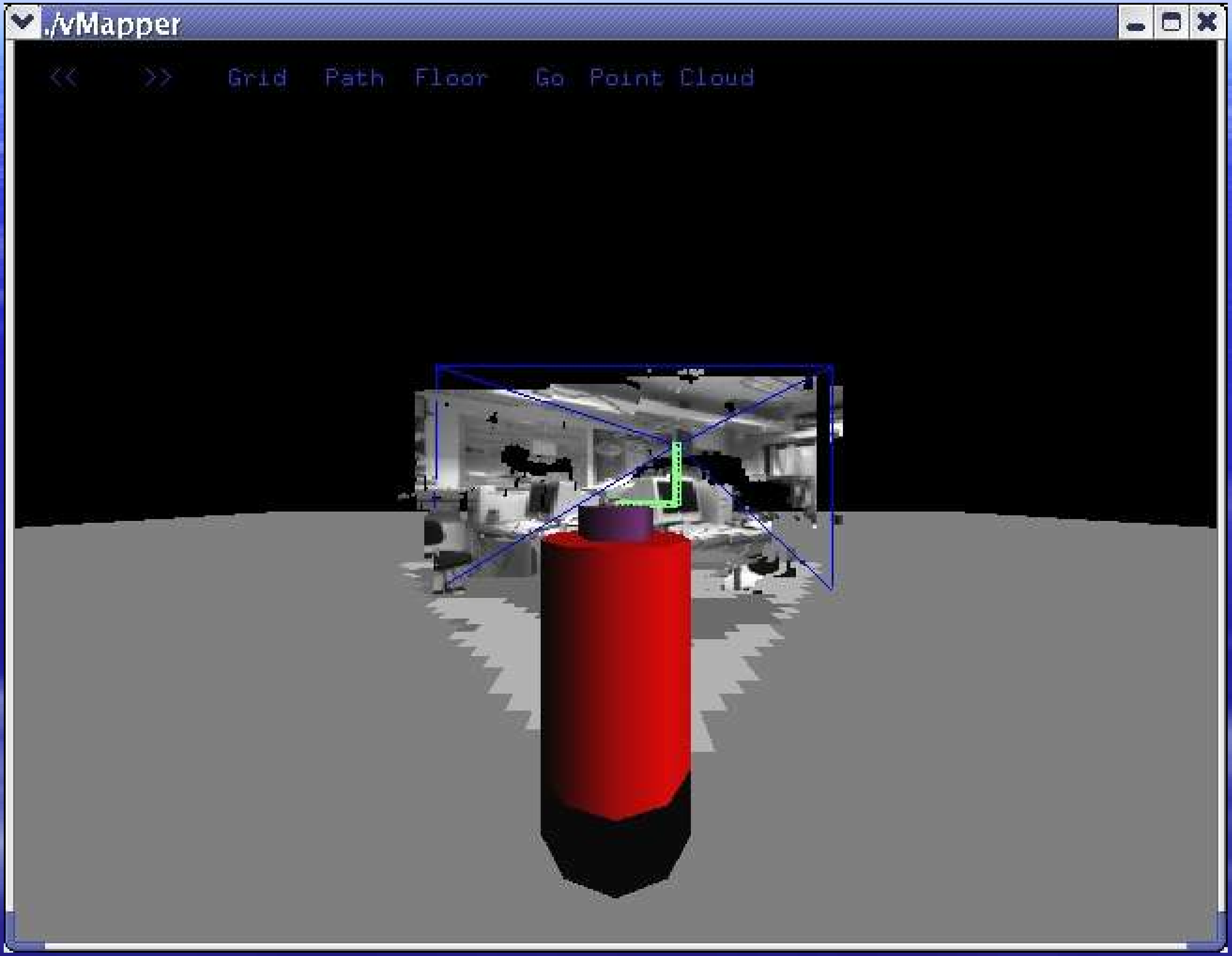


Floor: occupancy grid map ; Blue: field of view

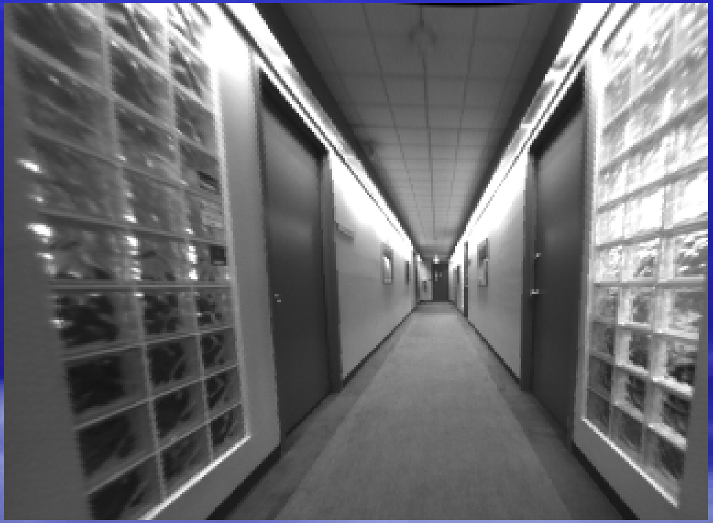
Black: at least one cell above is occupied; White: empty space

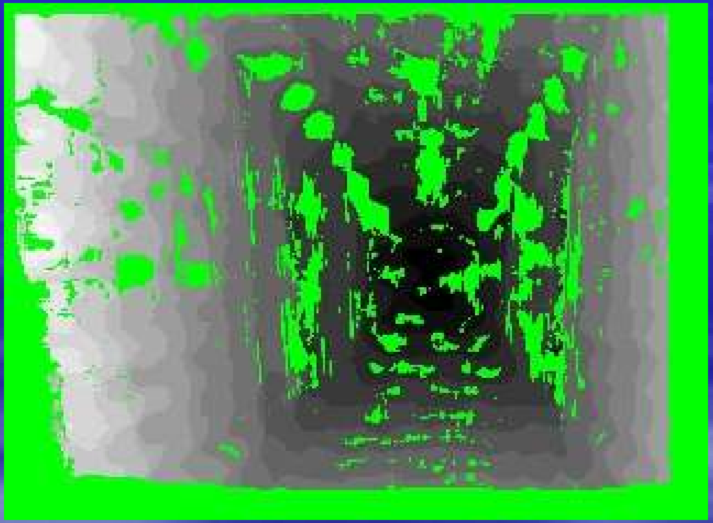


Green: cells 60% occupied

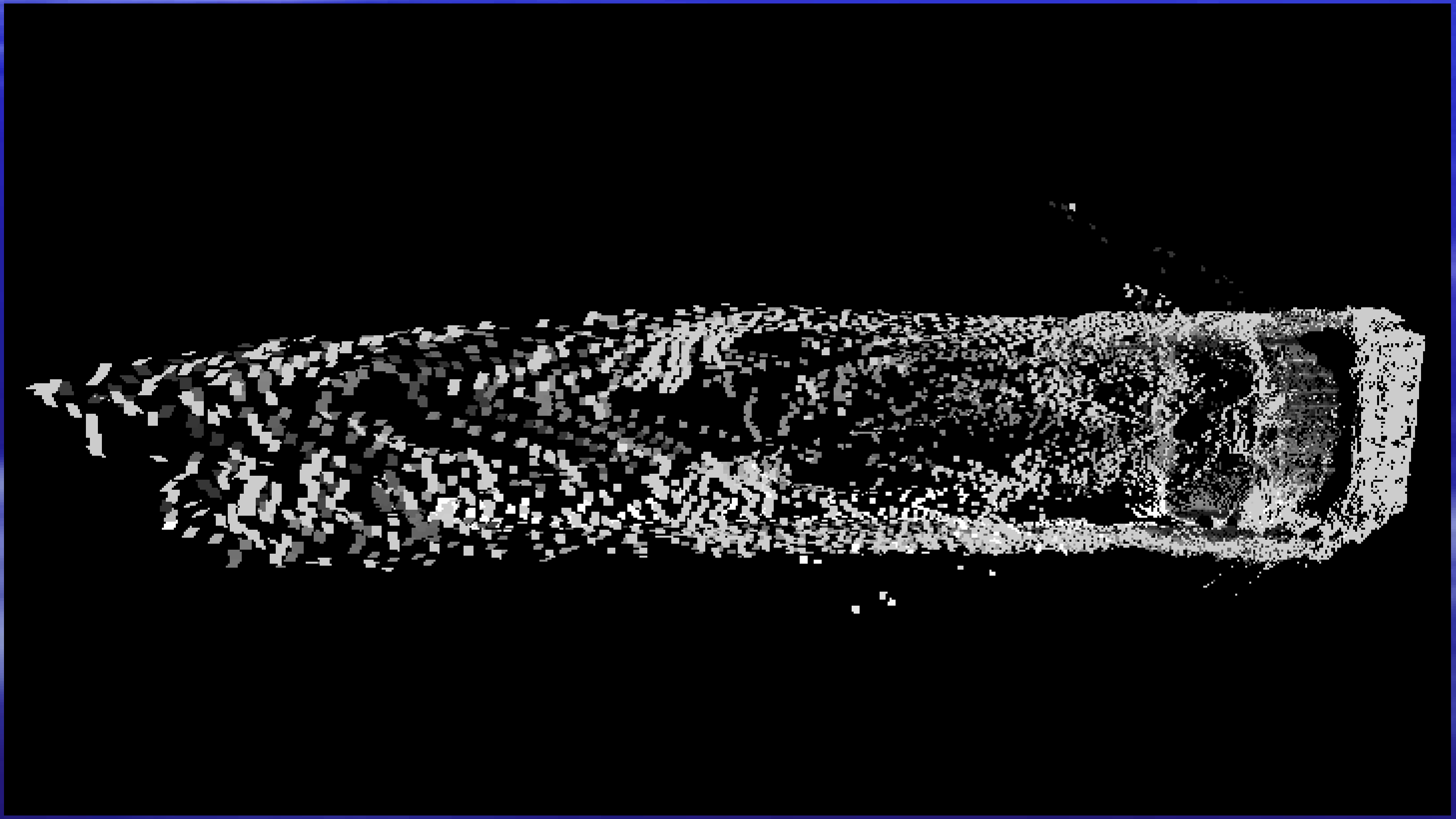

./vMapper
<<
>>
Grid
Path
Floor
Go
Point Cloud

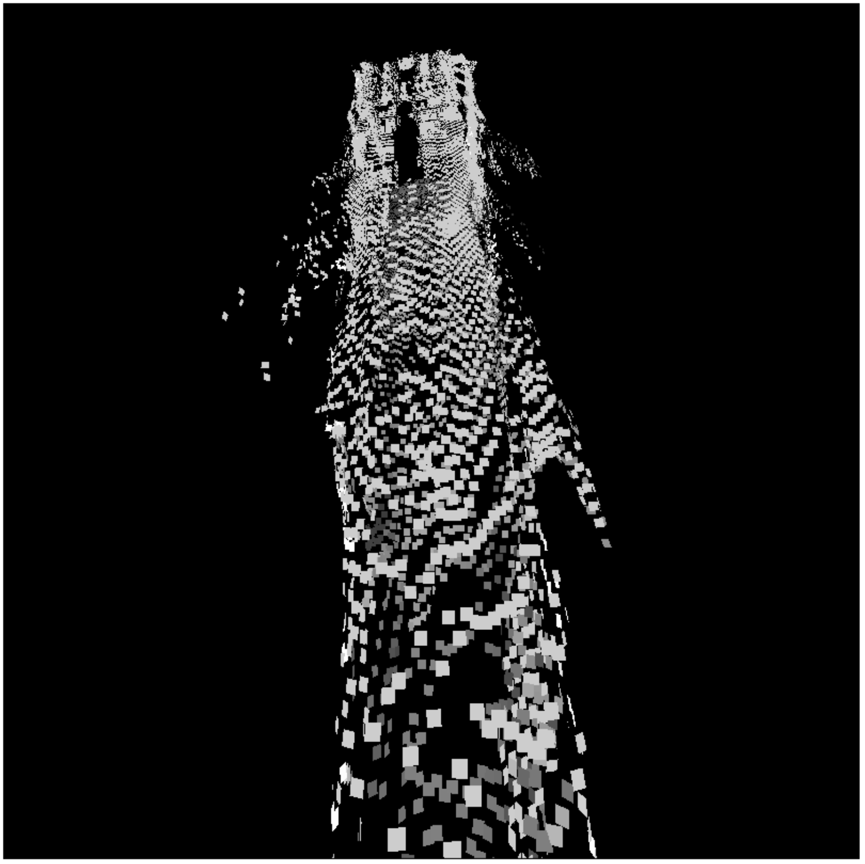
Brightness and Depth





Depth Pixels and Scale



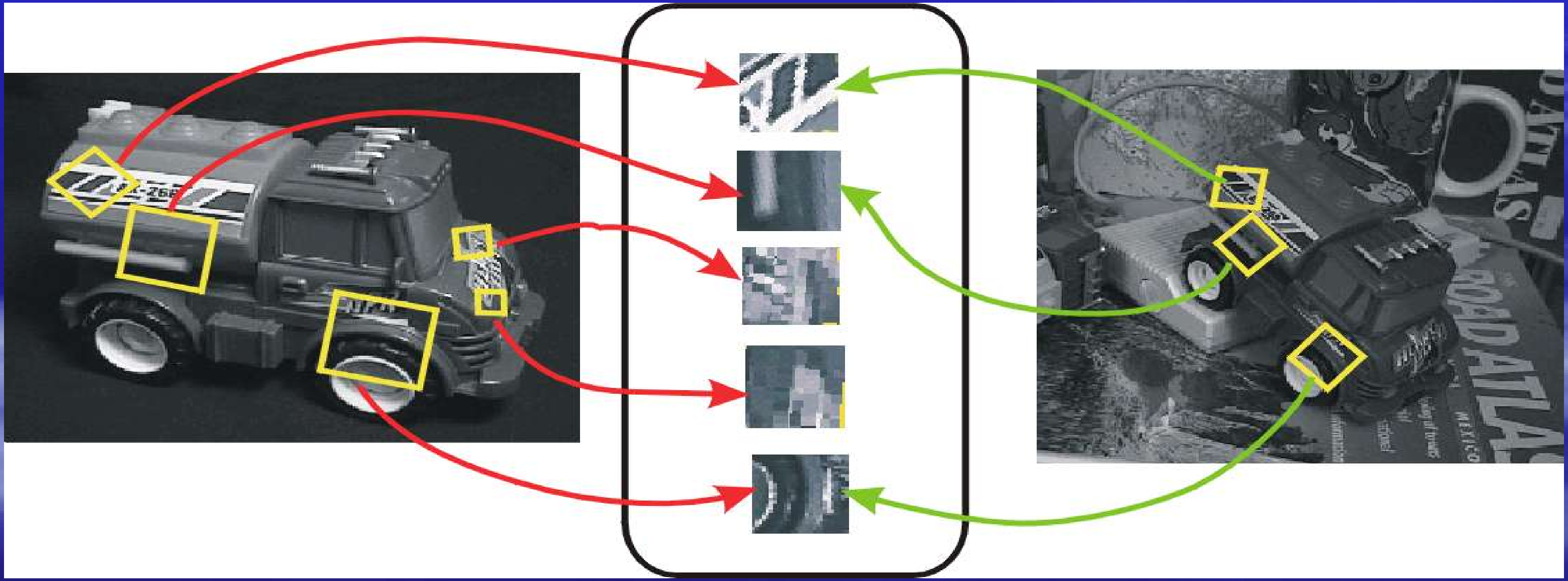


Where am I?
Simultaneous Localization and Mapping

- In order to collaborate with other robots and humans – its partners – a robot needs to determine its location.
- SLAM provides a map built from observed, distinguishable landmarks.
- We use visual features of the environment to build re-usable maps.
- The features are David Lowe’s Scale-Invariant Feature Transform image descriptors which are use to recognize objects and determine where they are.
- Where am I? = where is the world?

Scale Invariant Feature Transform (SIFT)

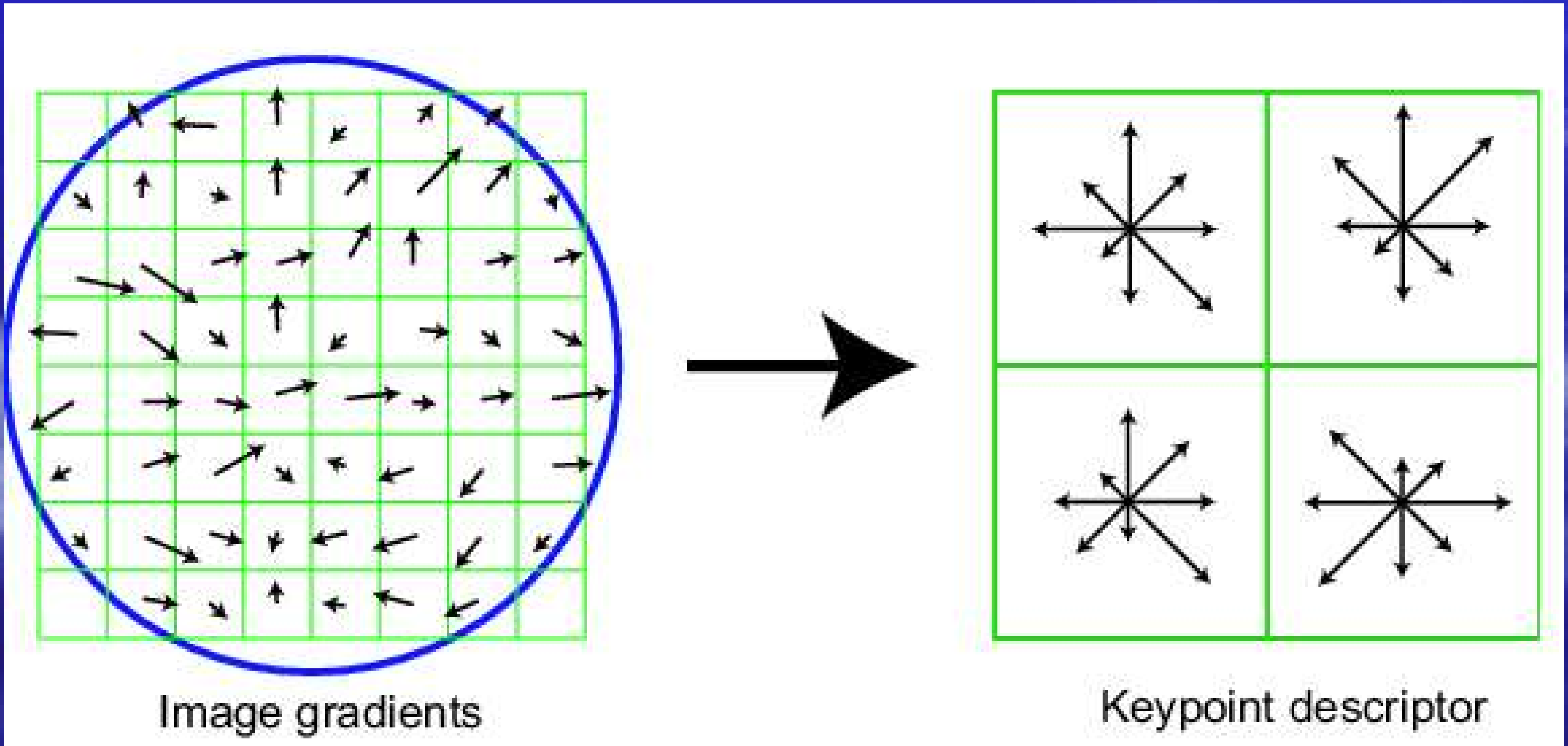
- Image content is transformed into local feature descriptions that are invariant to translation, rotation, scale, and other imaging parameters



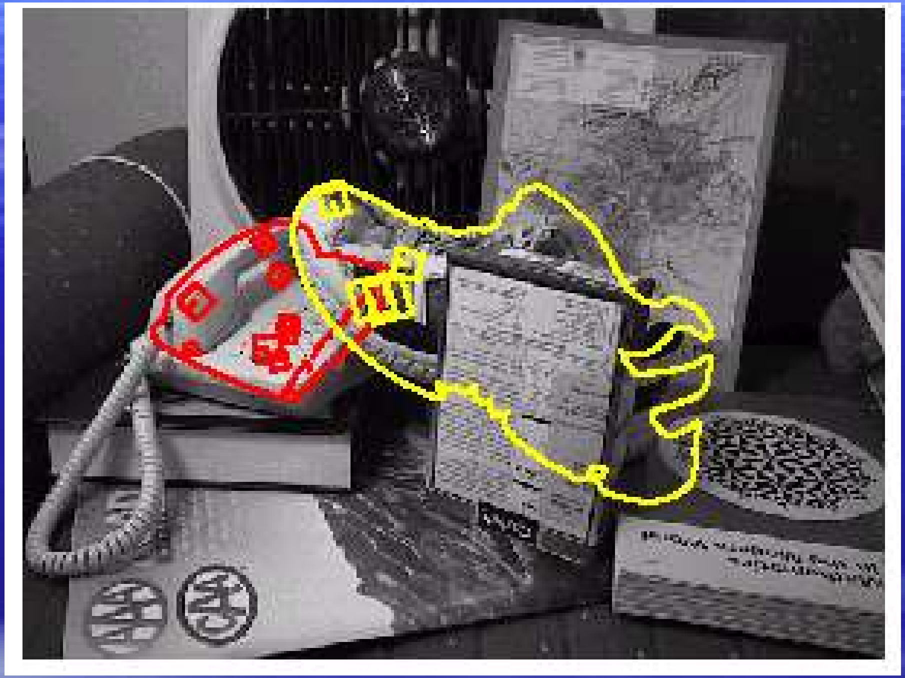
SIFT Features

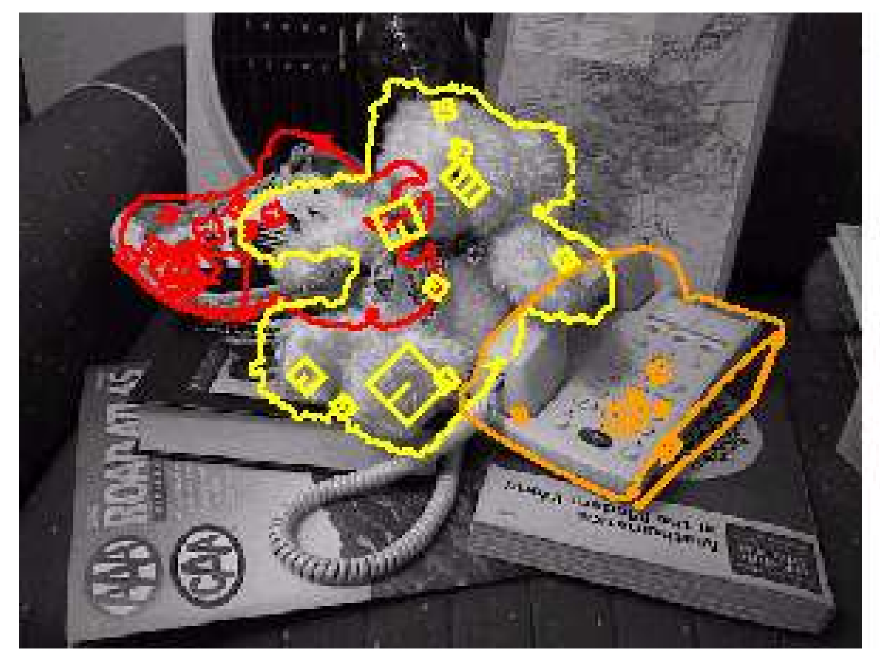
SIFT vector formation

- 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



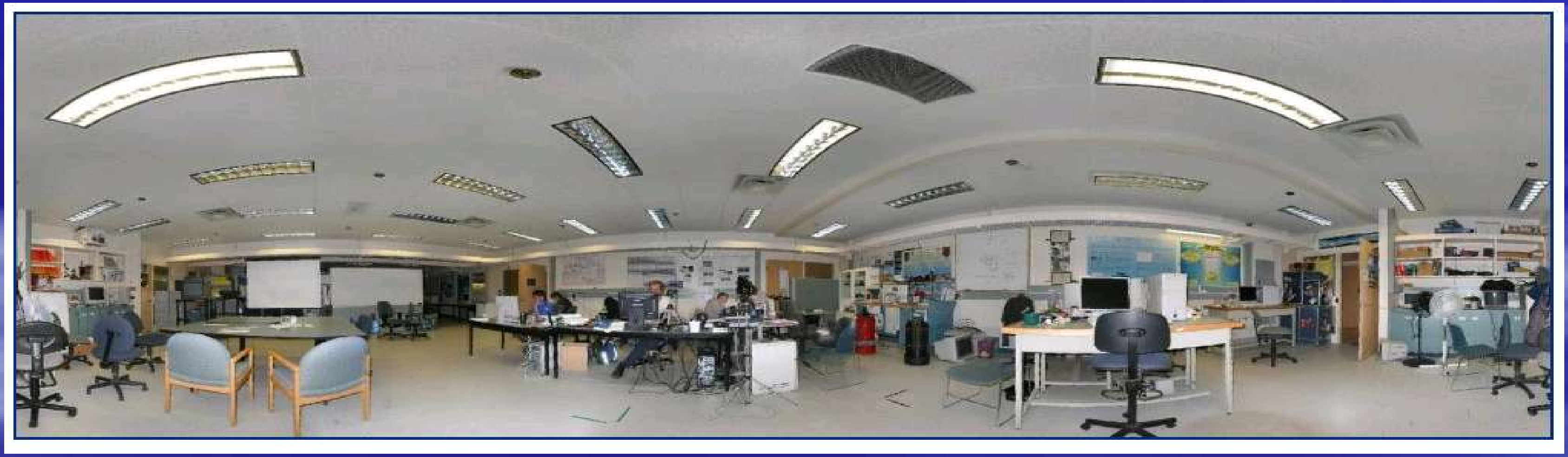
Recognition under occlusion





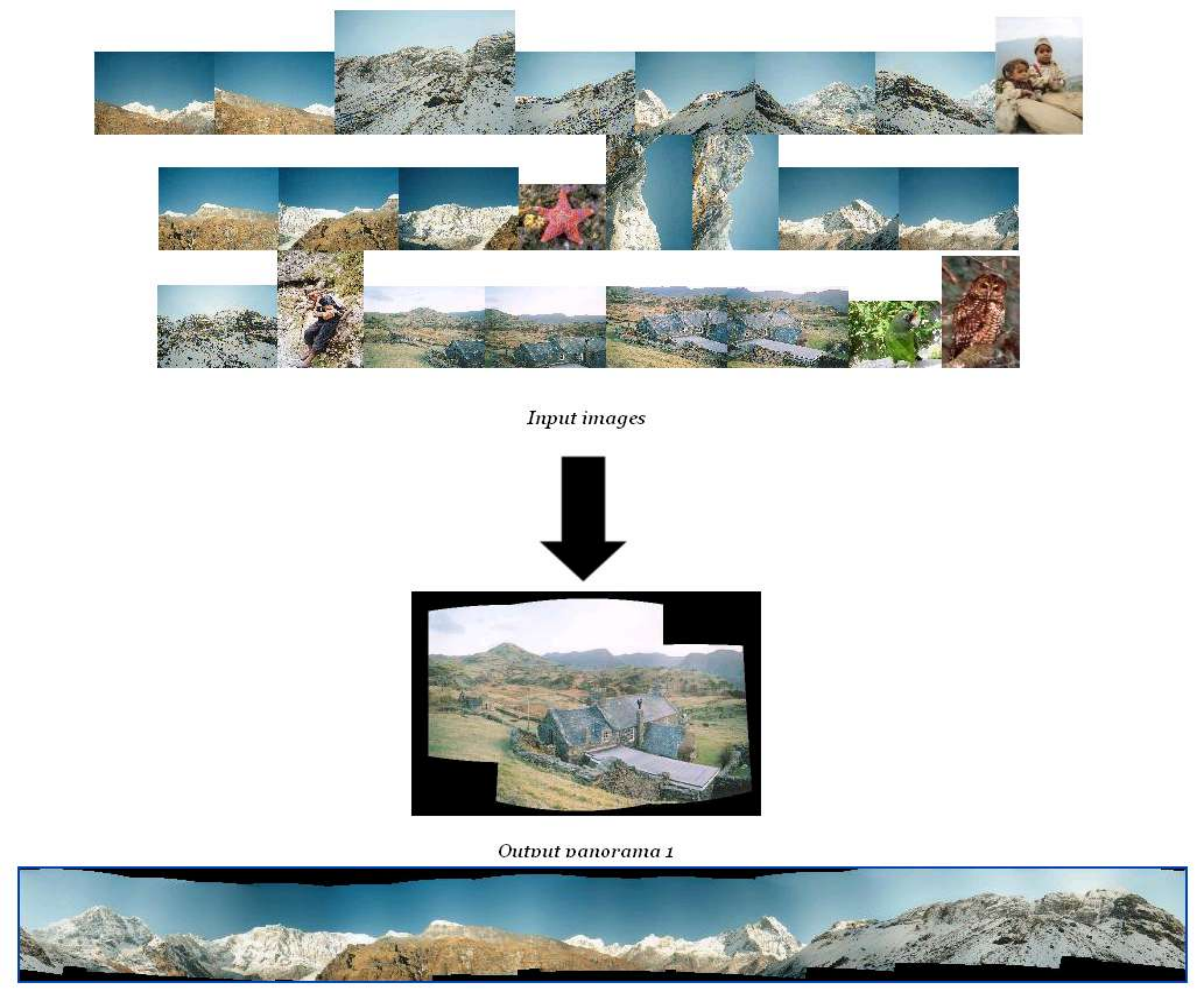
Recognizing Panoramas

- Matthew Brown and David Lowe
- Recognize overlap from an unordered set of images and automatically stitch together
- SIFT features provide initial feature matching
- Image blending at multiple scales hides the seams



Panorama of our lab automatically assembled from 143 images

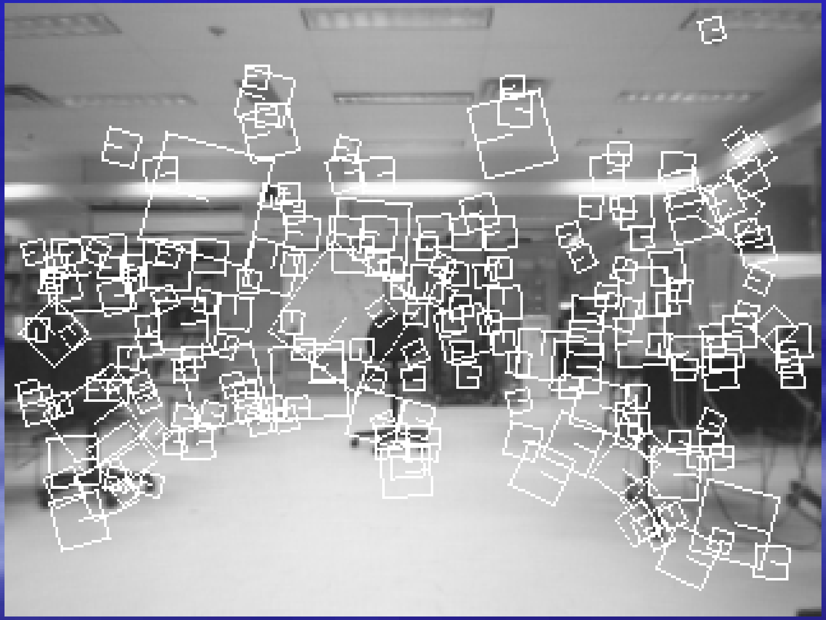
Multiple panoramas from an unordered image set



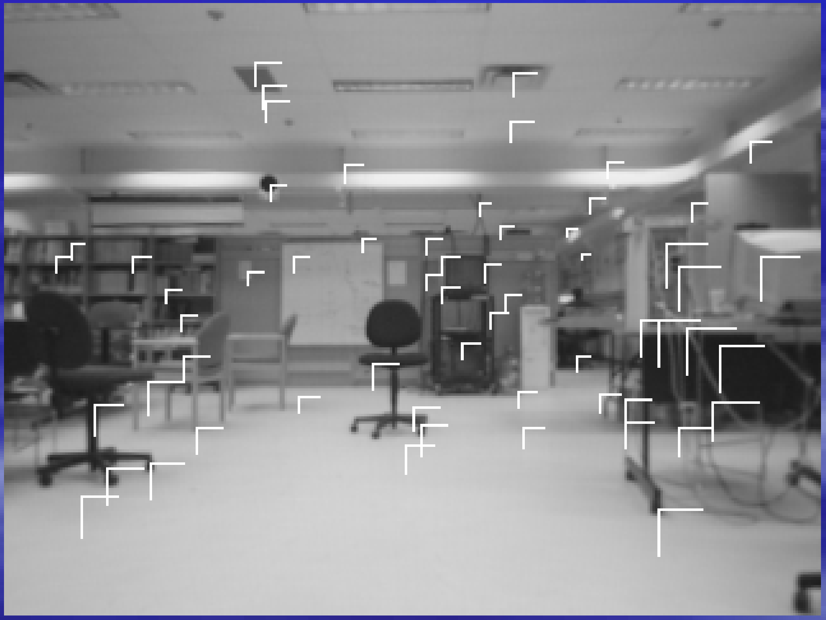
Input images

Output panorama 1

SIFT-base localization

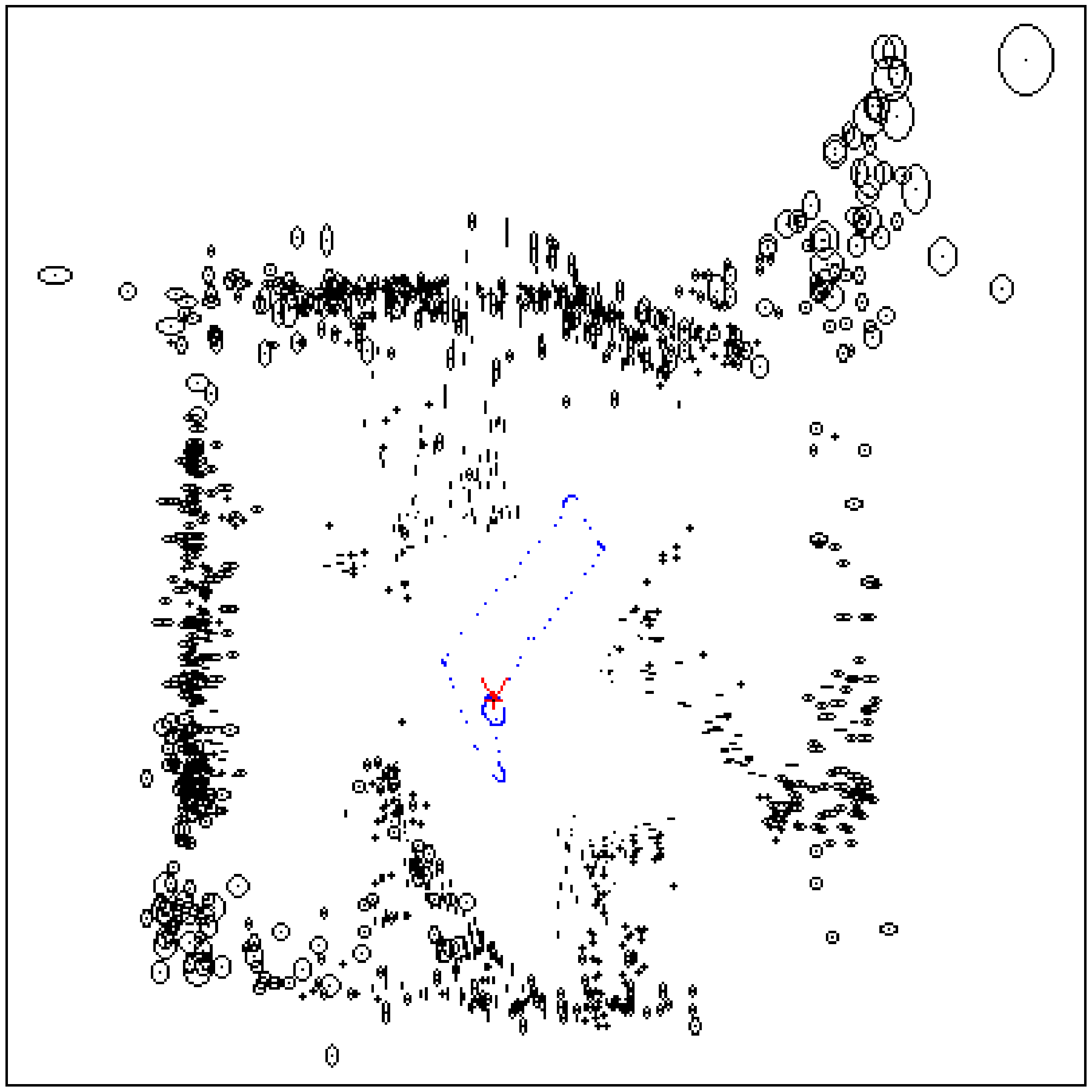


SIFT features: scale, orientation



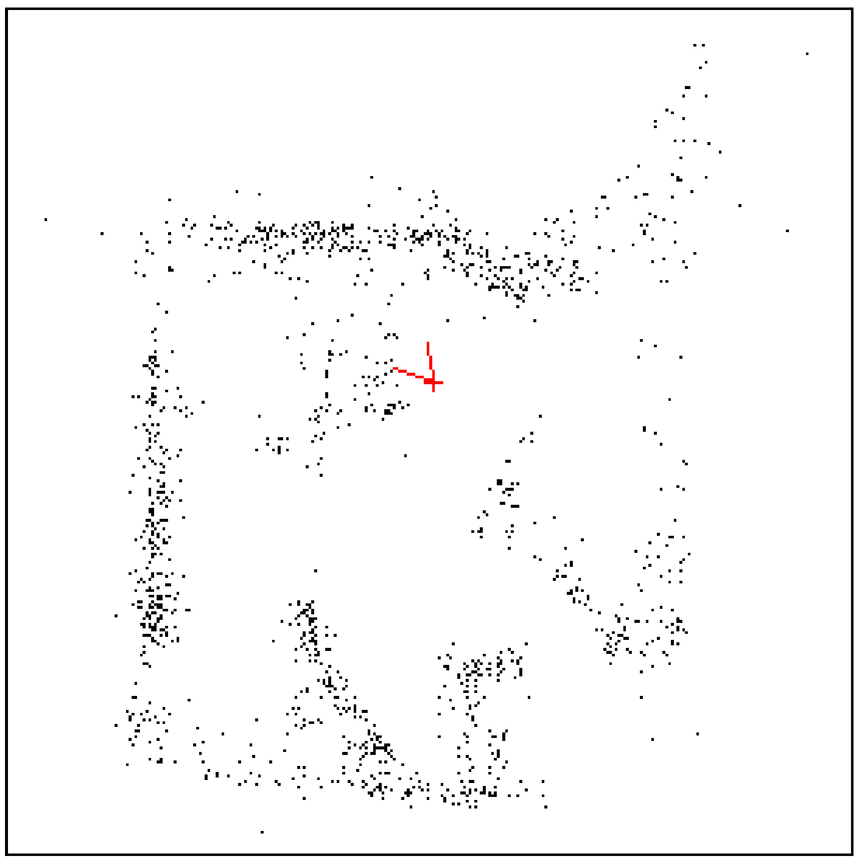
SIFT stereo: distance indicated by size

Map continuously built over time



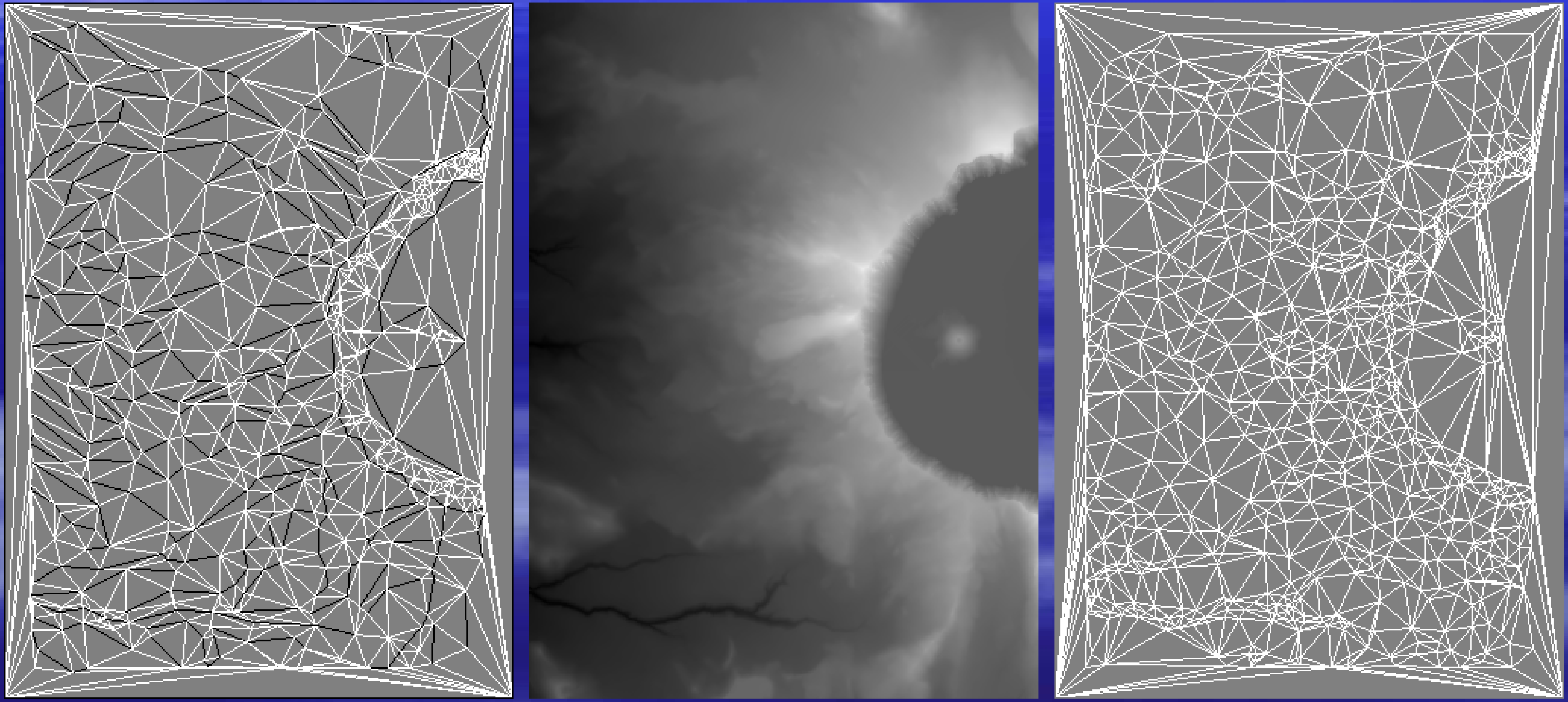
Global Localization

- Kidnapped robot problem
- Recognize robot pose relative to a pre-built map

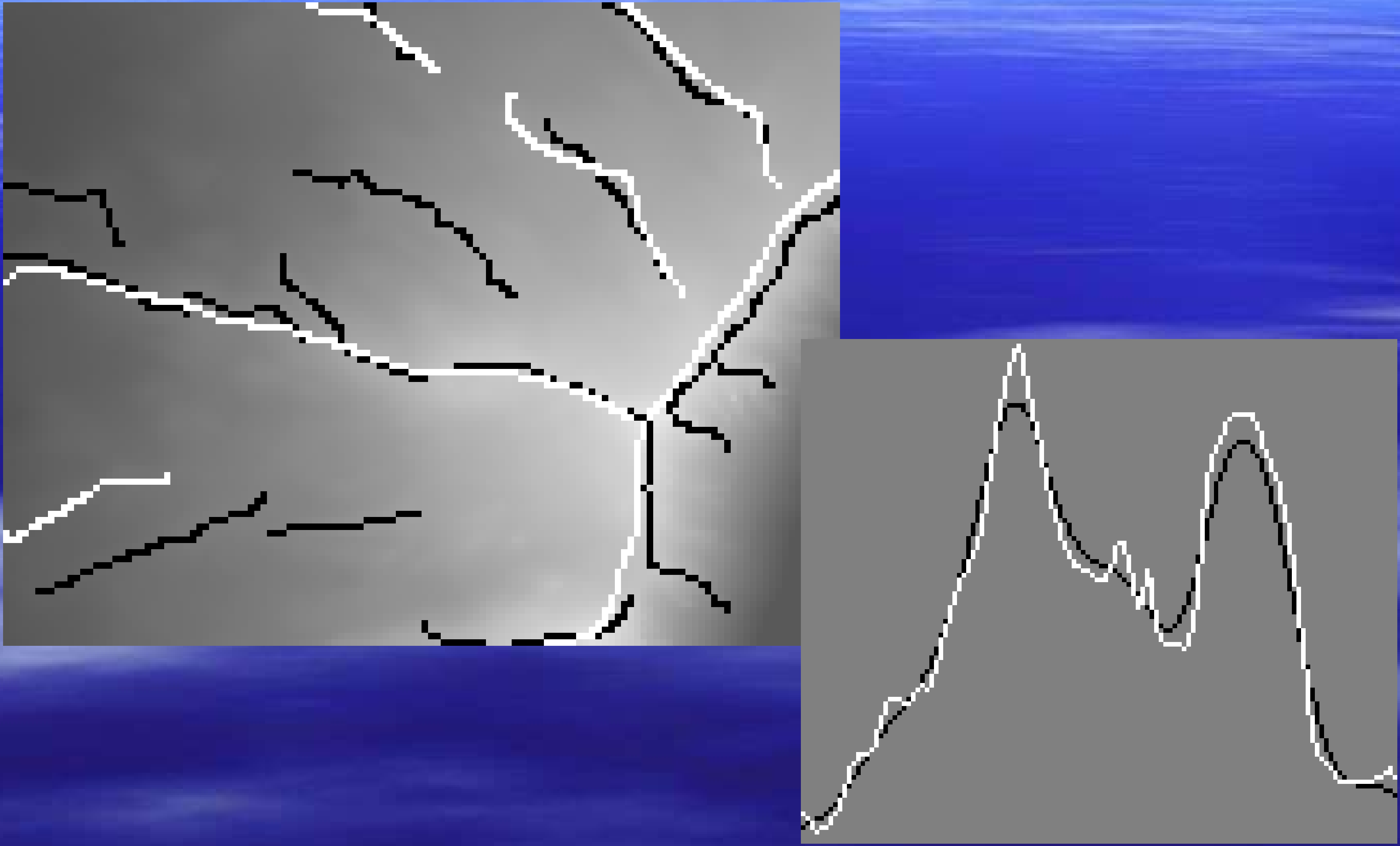


Measured Pose :
(70, 300, -40°)
Estimated Pose :
(75.8, 295.9, -41.1°)

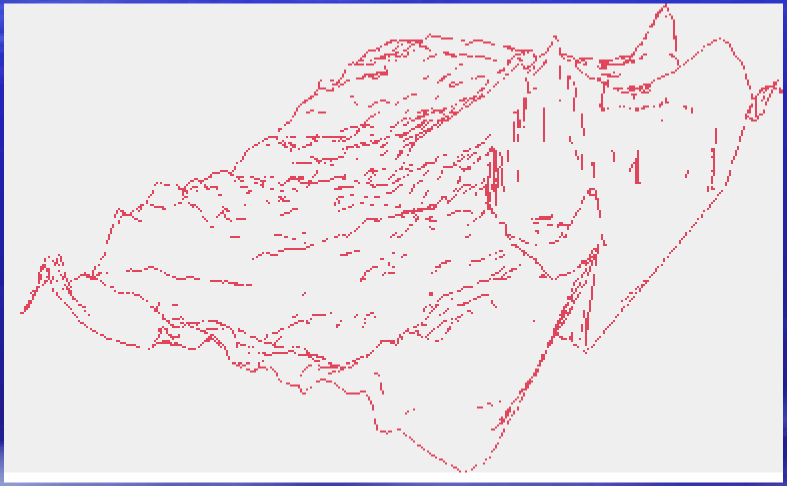
Maps of terrain

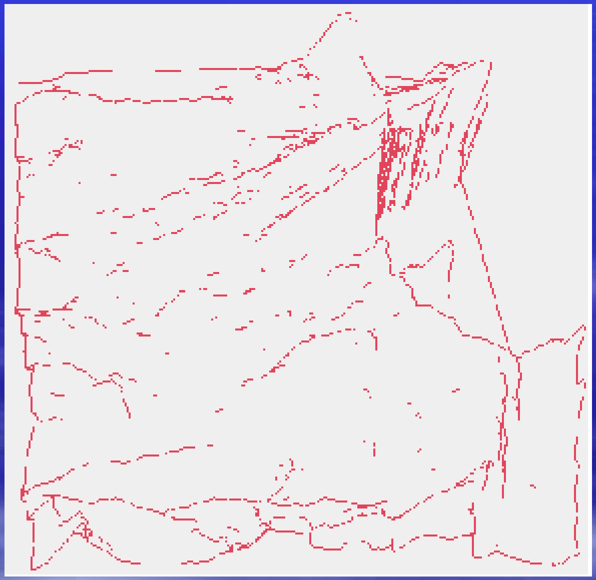


Simplification by anisotropic smoothing



Silhouettes for viewing





Depending on task requirements, a detailed silhouette in a coarsely modeled surface may be sufficient for recognition and navigation purposes

Seeing people and their actions

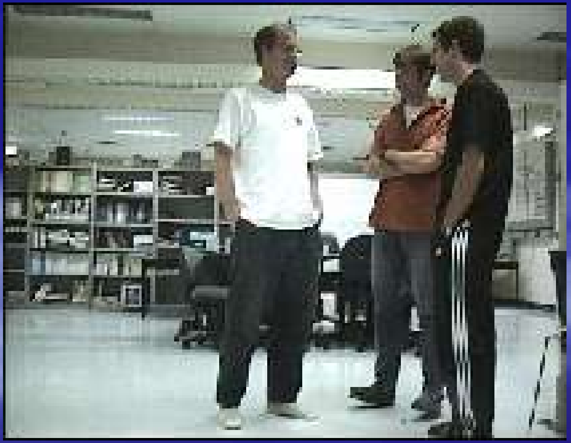
- The robot needs to locate people in order to assist them.
- We find faces in images and find how they move so that we can judge their expressions.
- A robot needs to know what people are doing, and what they may be telling the robot, either through gestures or motions that can be interpreted to know their attitude or feelings, or what they intend to do.

Finding People

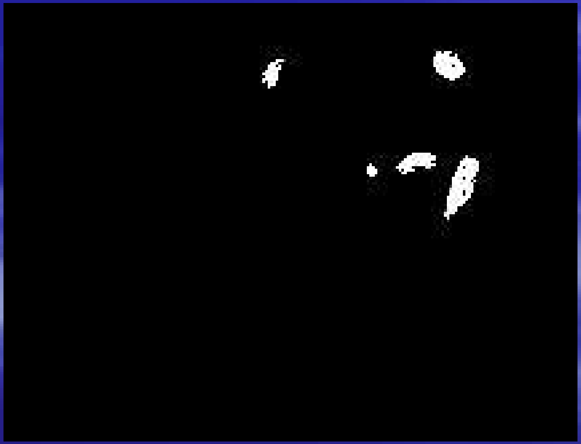
Use occupancy grid probability map to decide where to serve

Detect people using skin color

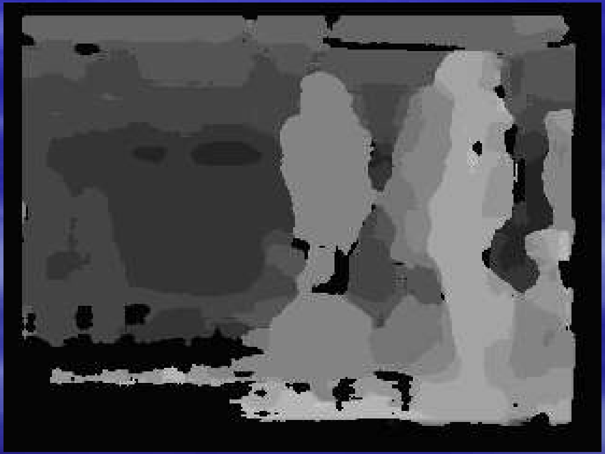
Use stereo data to compute 3D position of people



Color Image



Skin Regions

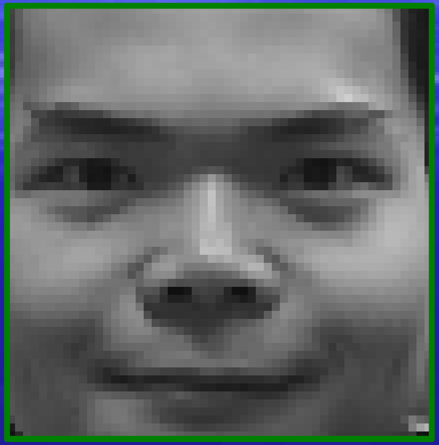


Stereo Data

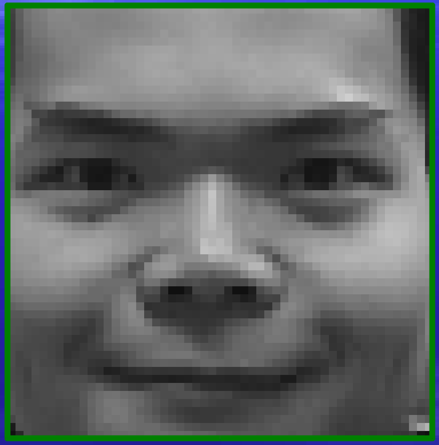
What the robot sees

Optical Flow

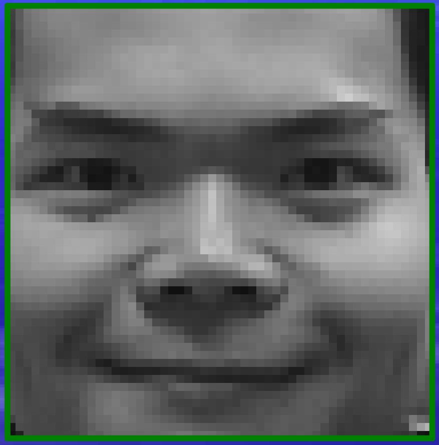
t=0

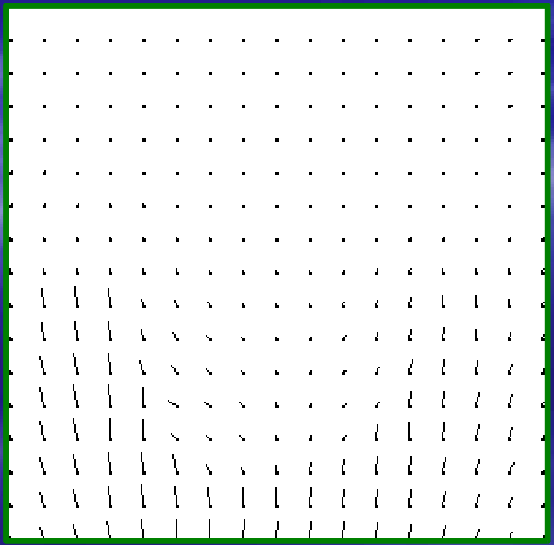


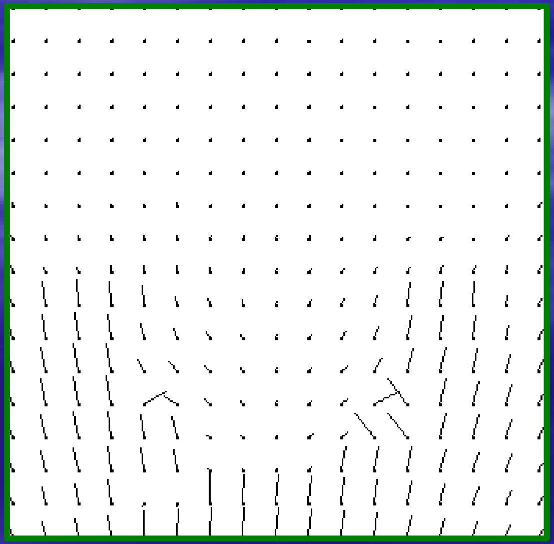
t=1



t=2







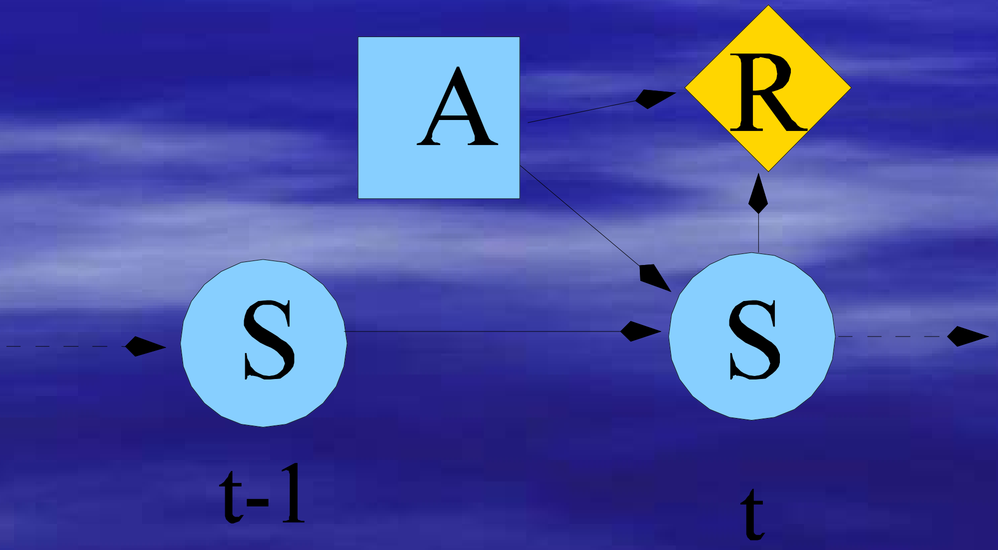
Human facial displays in interactions : Partially Observable Markov Decision Processes

The model is a partially observable Markov decision process, or POMDP.

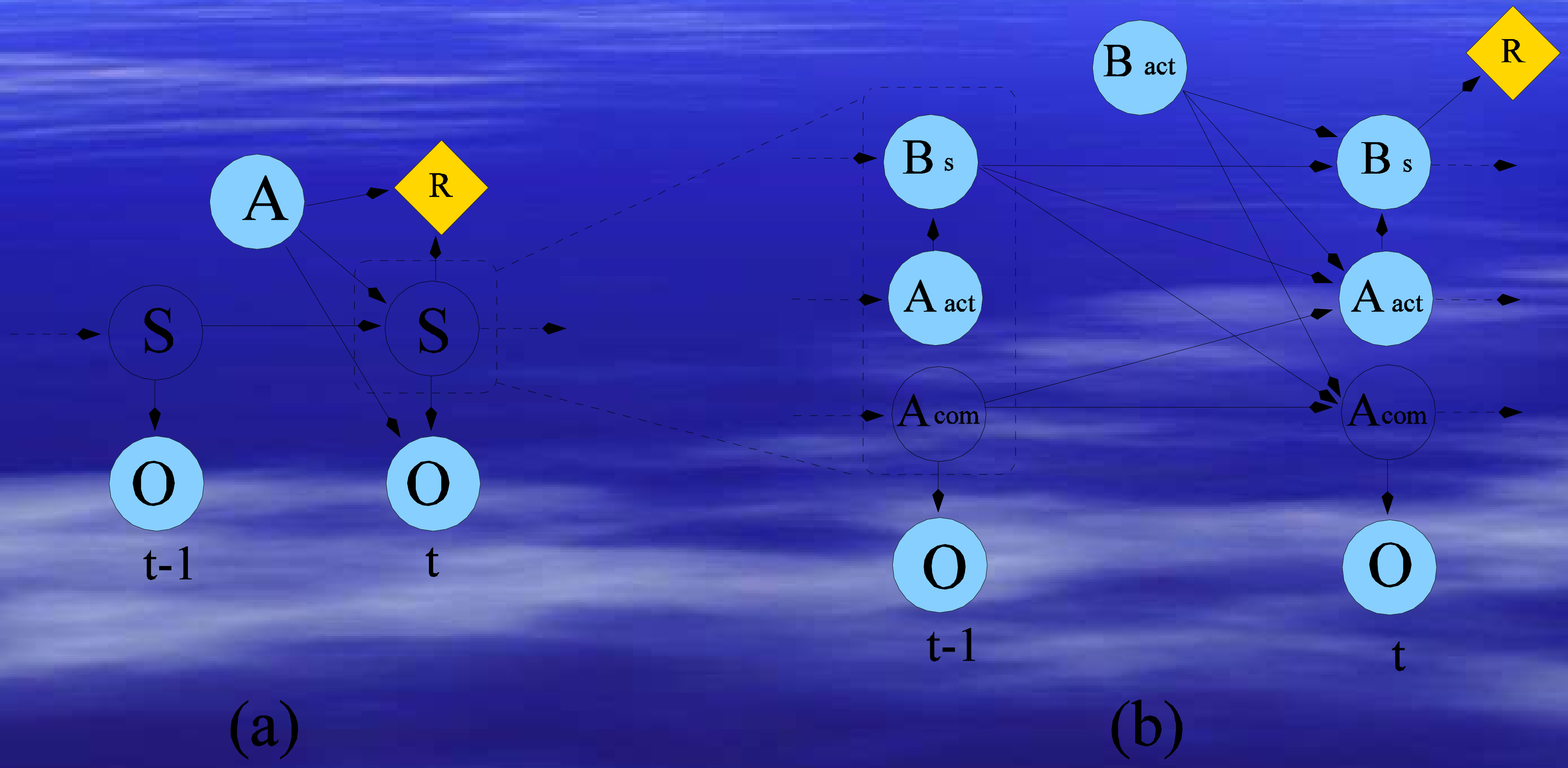
Partially Observable: a visual observation gives us uncertain information about the state of the world

Markov: State t depends only on State t-1

Decision Process: actions are determined by Rewards

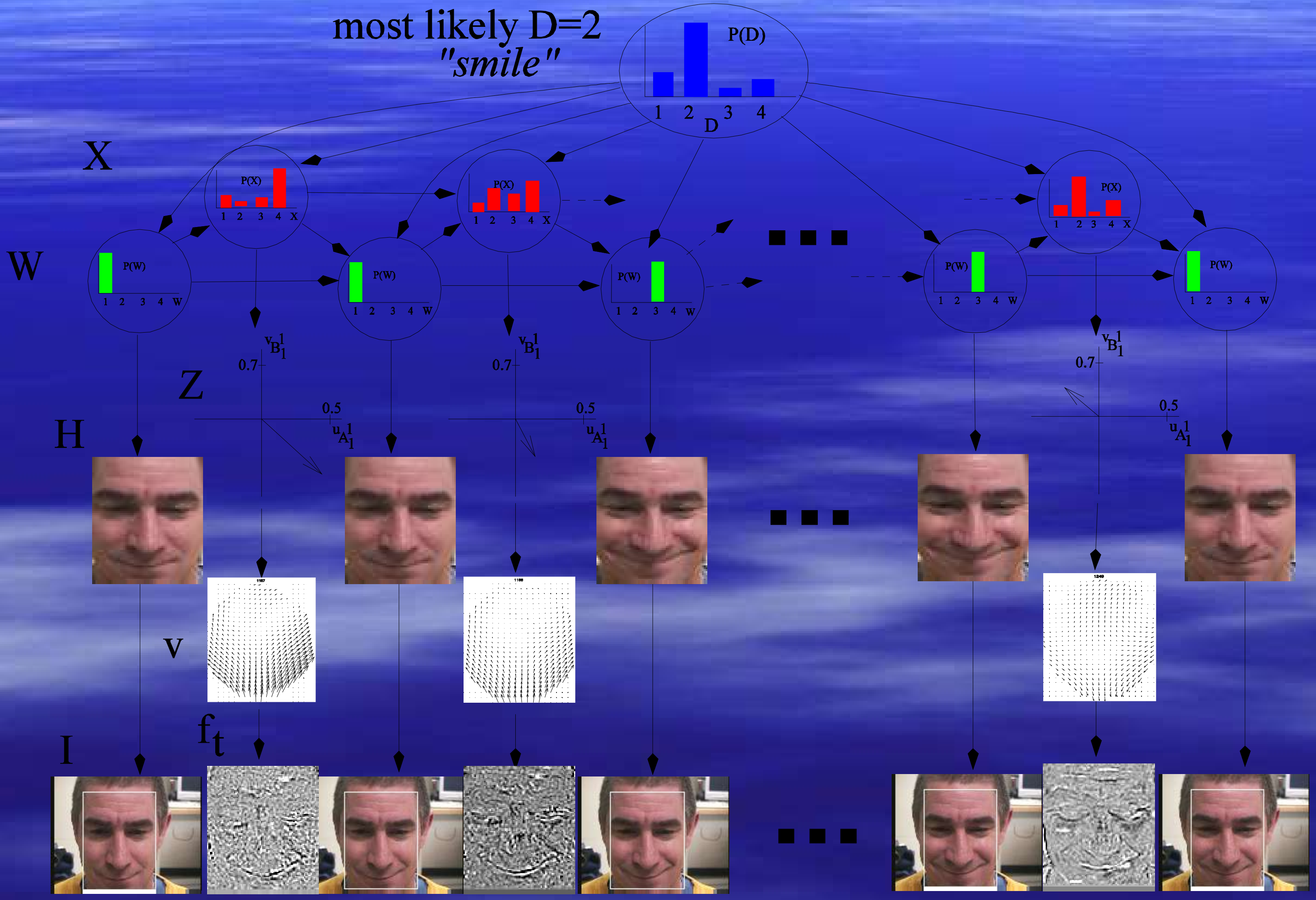


Coupled HMMs for Facial Expressions



Light blue circles: observables; Yellow diamonds: policies/utiities

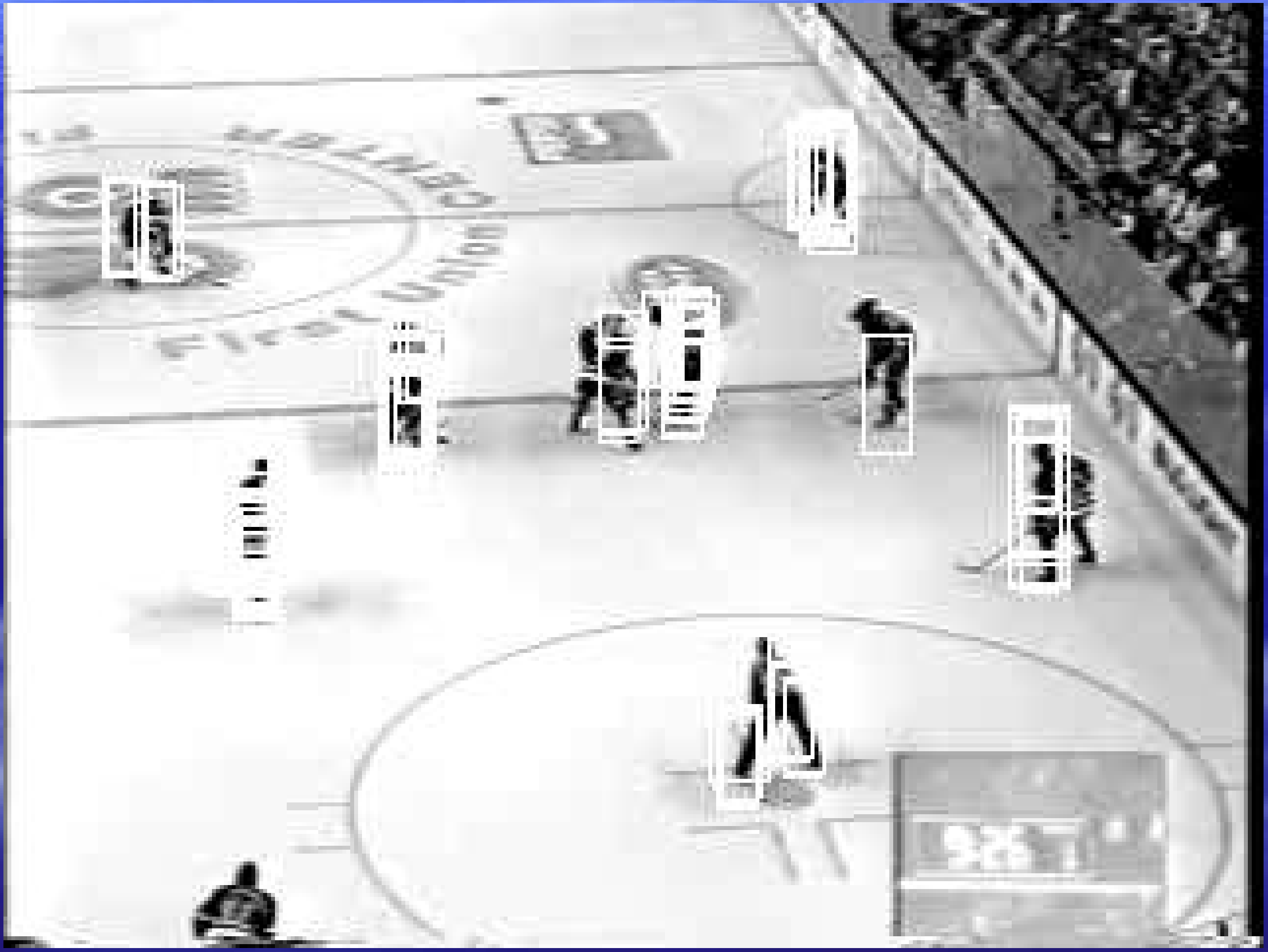
Smiling analysed by mixture of Coupled Hidden Markov Models (CHMMs)



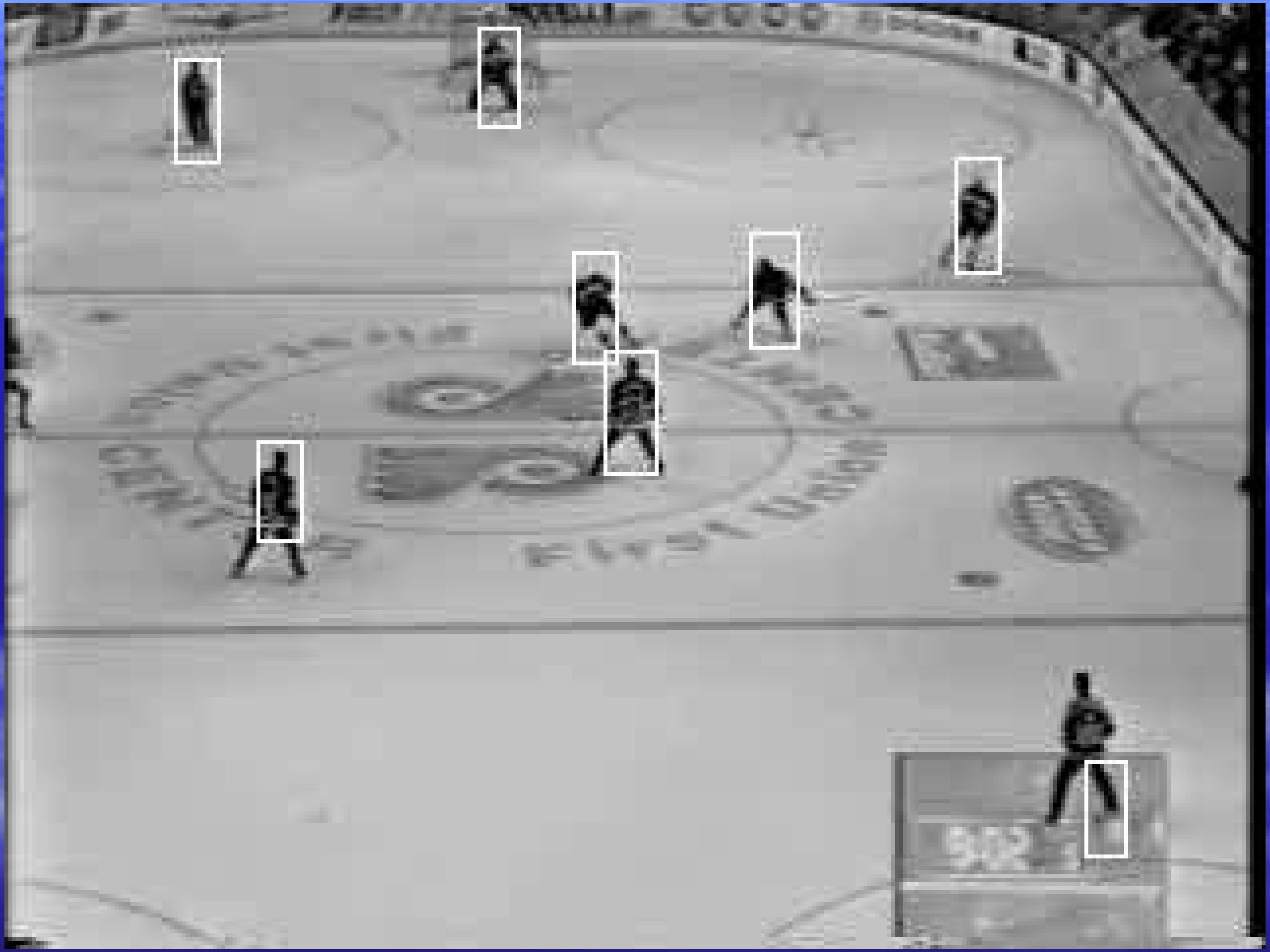
Trajectories: tracking

Result movie:

Tracking people

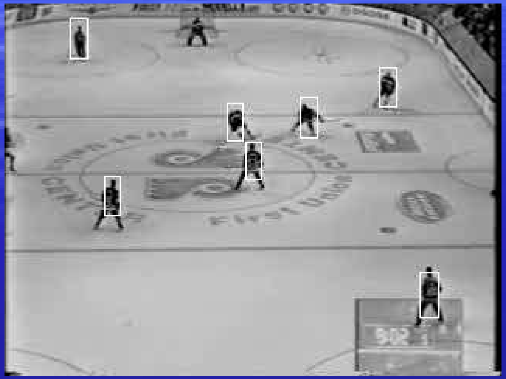


Tracking

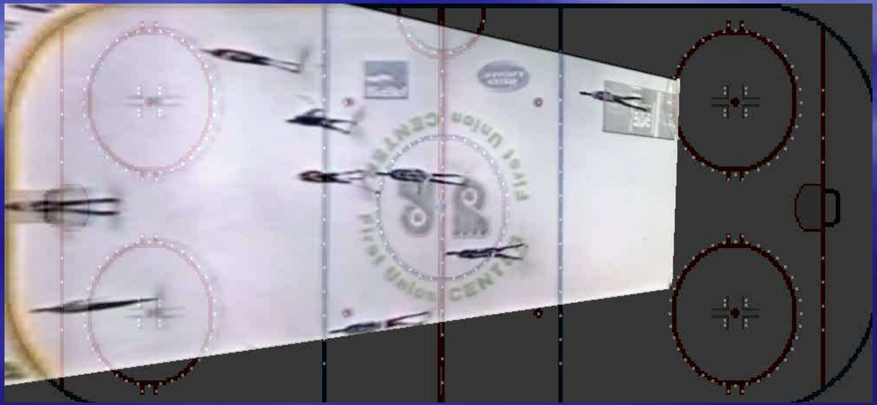


Result

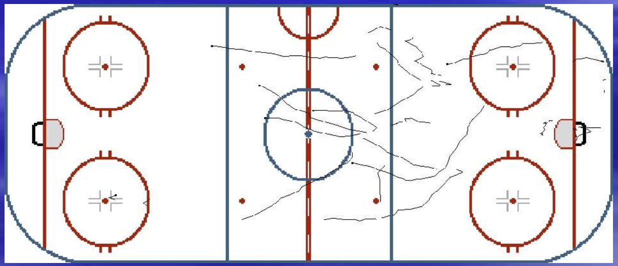
=



+

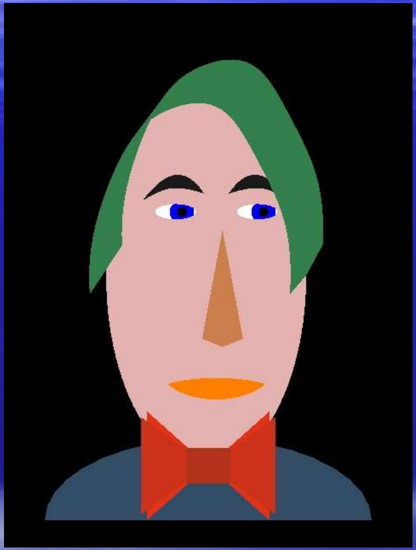


=

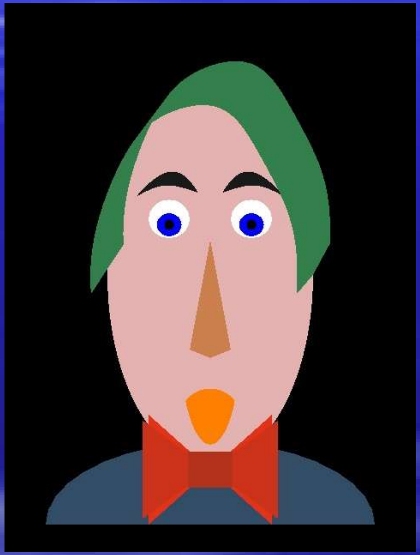


Understanding motion

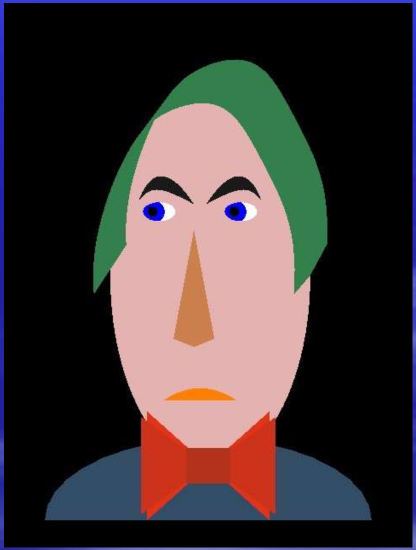
Face: displaying emotional state



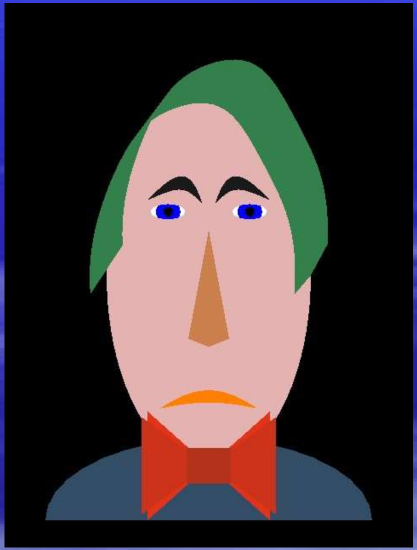
Neutral



Surprised



Angry



Sad

The face communicates the attitude of the robot to the interaction.

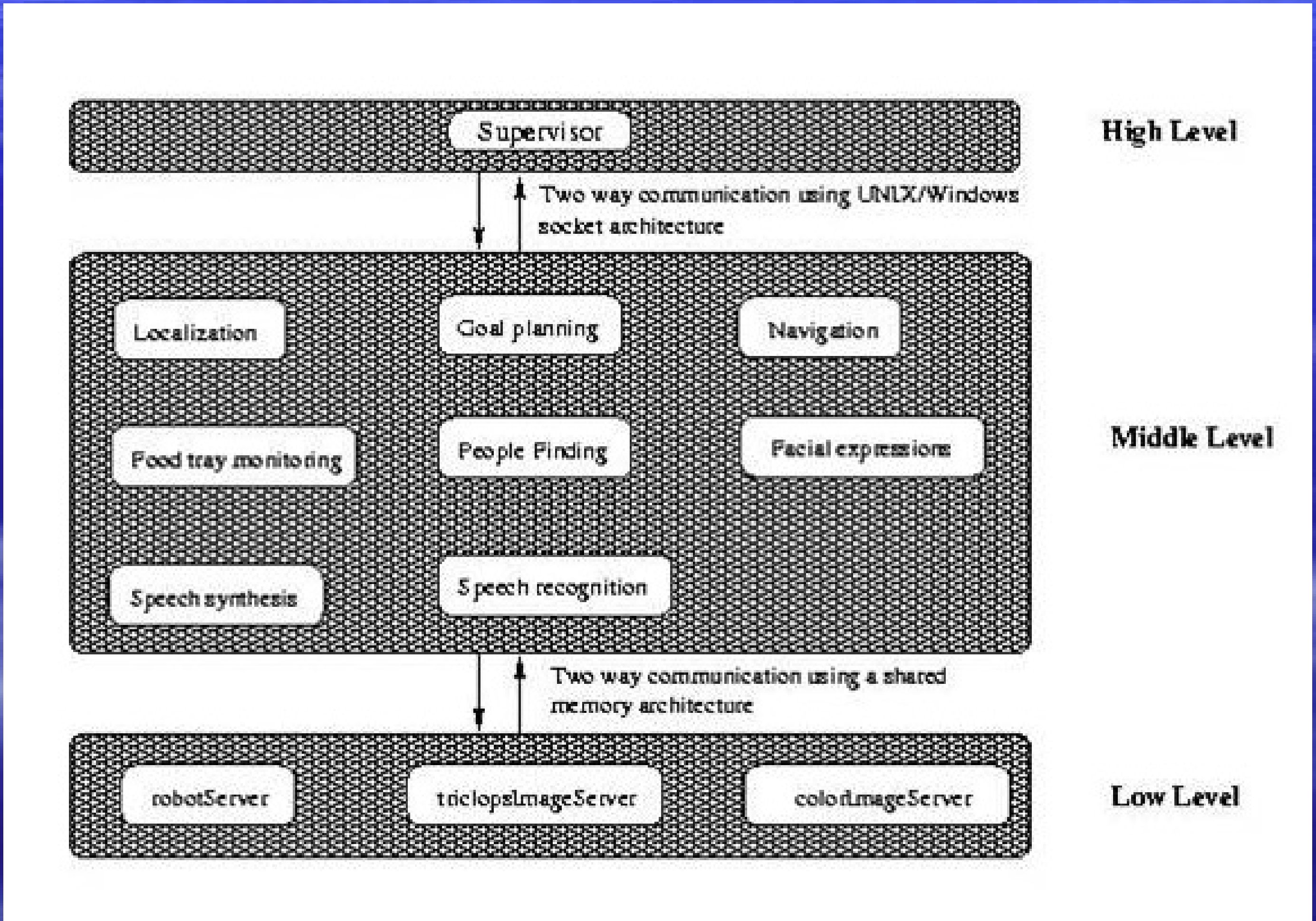
Robot Roles

- Given their capabilities, our robots can play different roles:

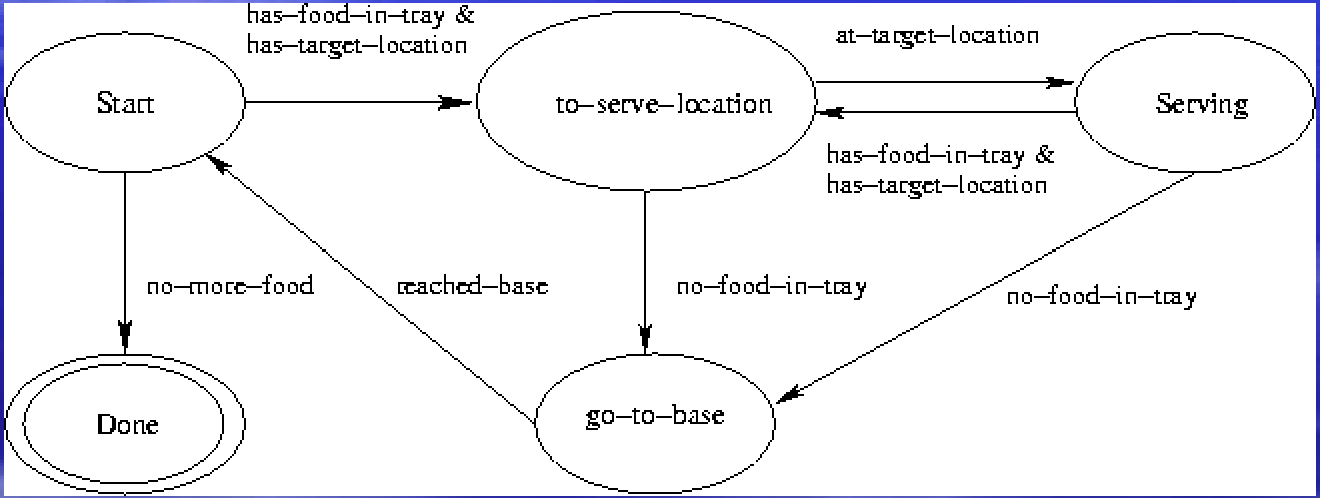
a role is a set of activities with a collection of goals

- Jose is a robot waiter
- Homer delivers personal messages

Control Architecture



Control Architecture (Cont.)



- Supervising Behavior decides what to do next
- There are 5 main states
- At each state more than one Behavior may be active

Alan Mackworth

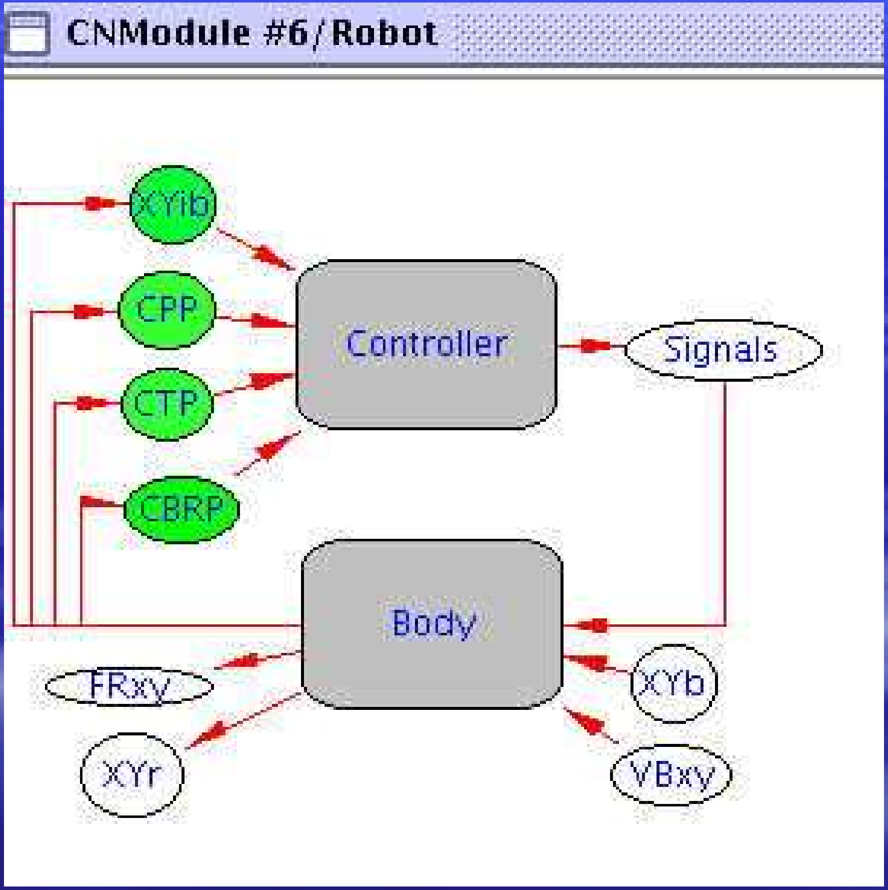
- Constraint-based agents: models, languages and systems
- Computational vision and robotics
- Multiagent systems including soccer-playing robots
- Specification, modeling and verification of hybrid dynamical systems

Constraint Nets: theory, tools, applications

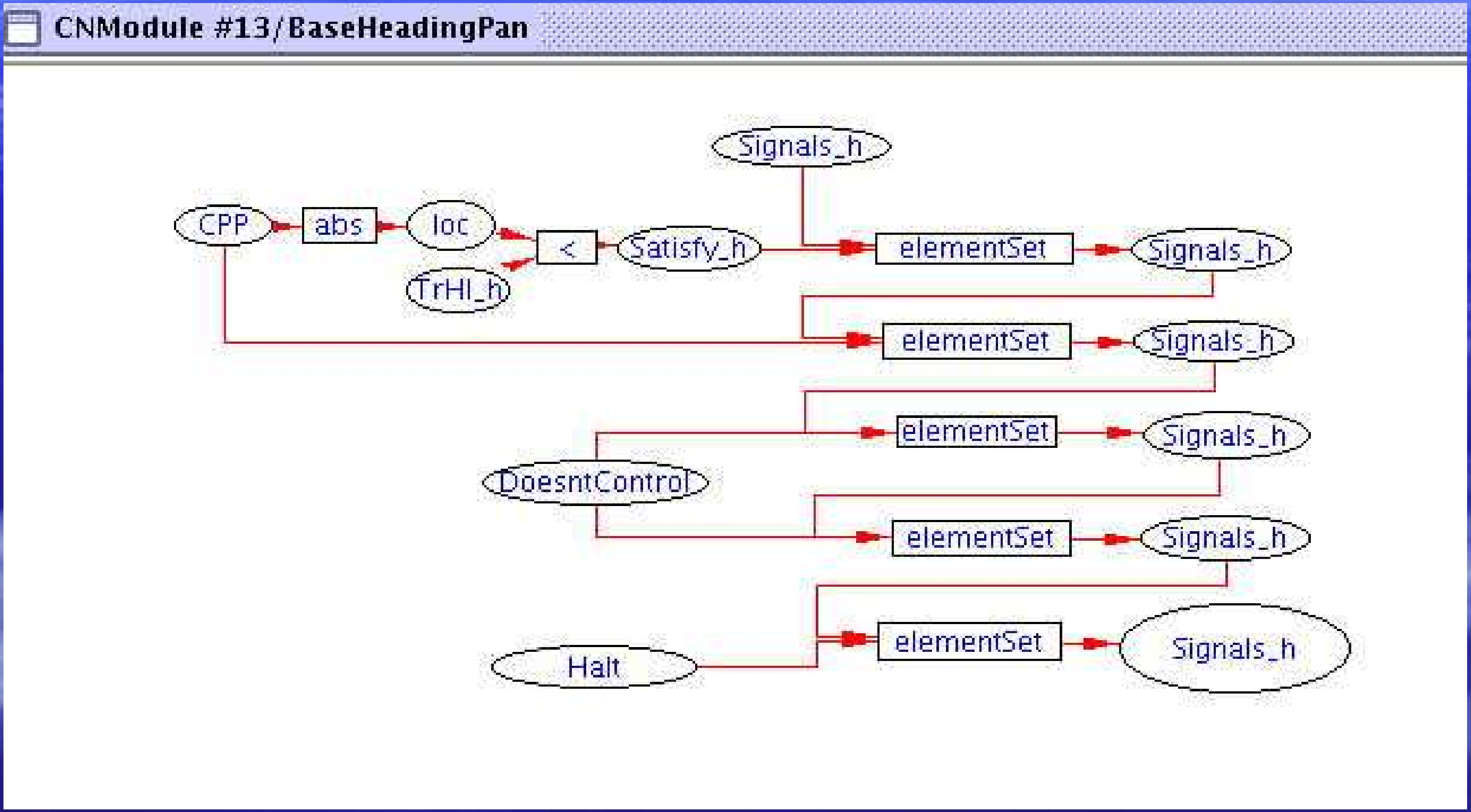
- Robert St-Aubin & Mackworth are designing and building Probabilistic Constraint Nets (PCN) for representing uncertainty in robotic systems.
- Song & Mackworth have designed and implemented CNJ, a visual programming environment for Constraint Nets, implemented in Java.

  The system includes a specification and implementation of CNML, an XML environment for Constraint Nets.
- CNJ has been used as a tool by Pinar Muyan to develop a constraint-based controller for a simple robot soccer player.

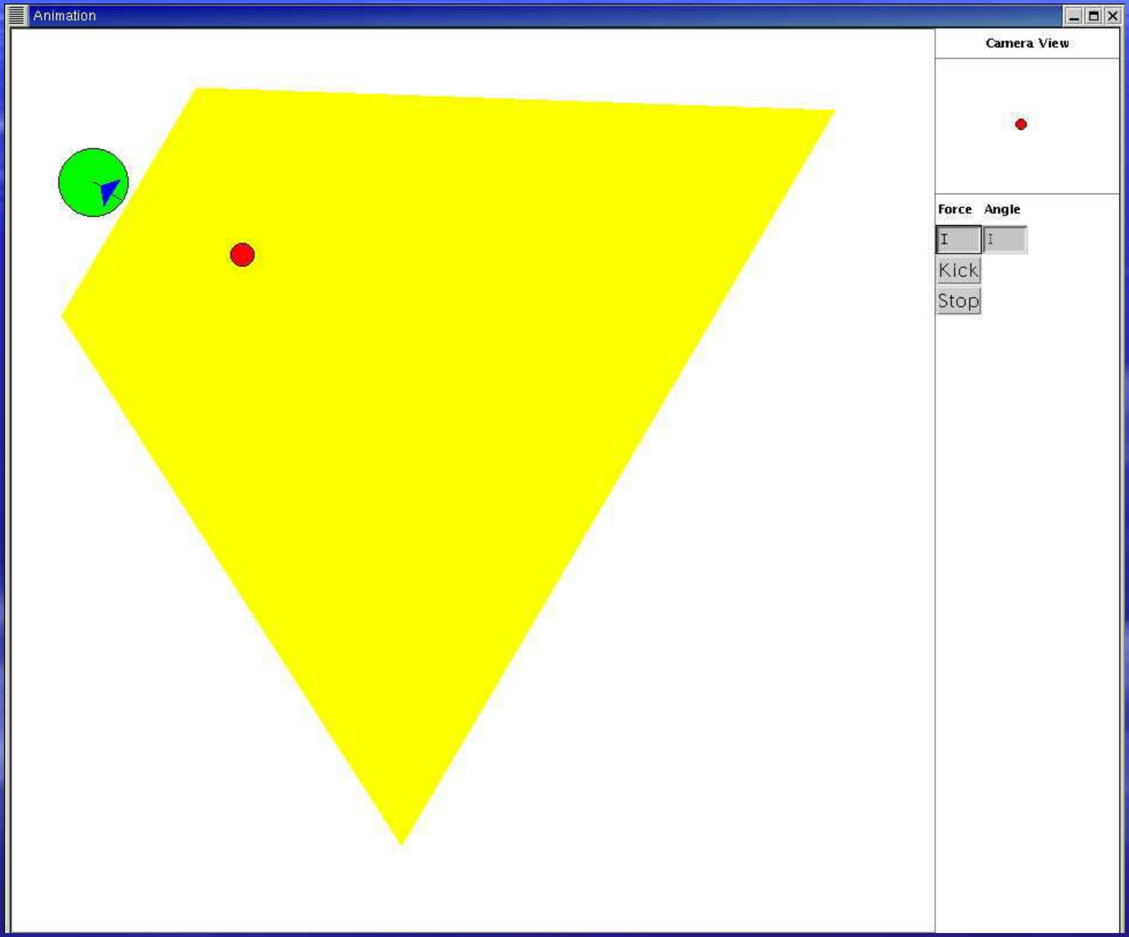
CNJ: Robot in pursuit of ball



CNJ: Controlling Rotation/Pan/Tilt



CNJ: Animation of controller



What’s Missing?

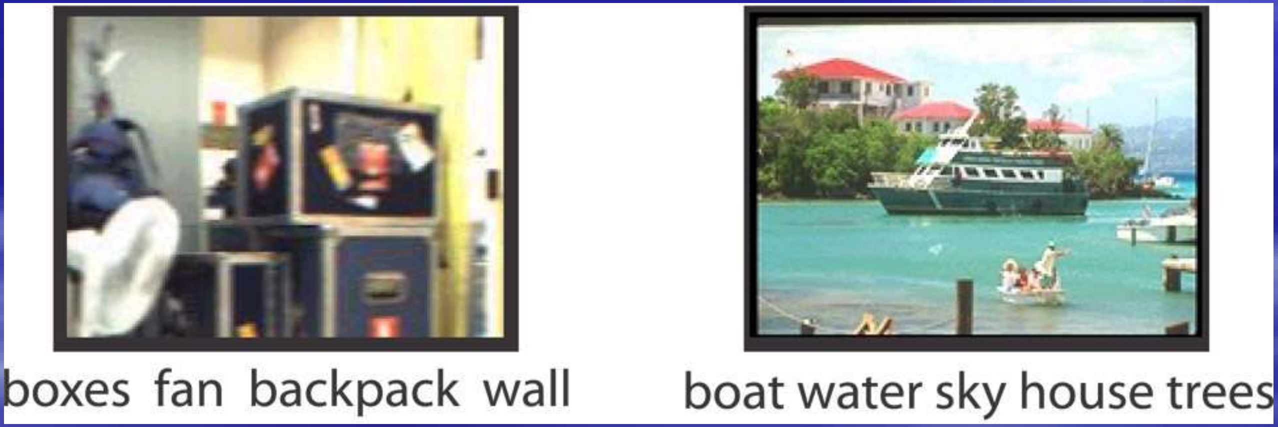
We want to tell the robot what’s important: where to be and what to do – cognitive robotics

We want to program the robot by describing a scenario: what’s in the world and what happens – how people and objects interact

The robot should communicate: generate speech to explain actions/situations and learn verbal descriptions

Associating words with objects in images

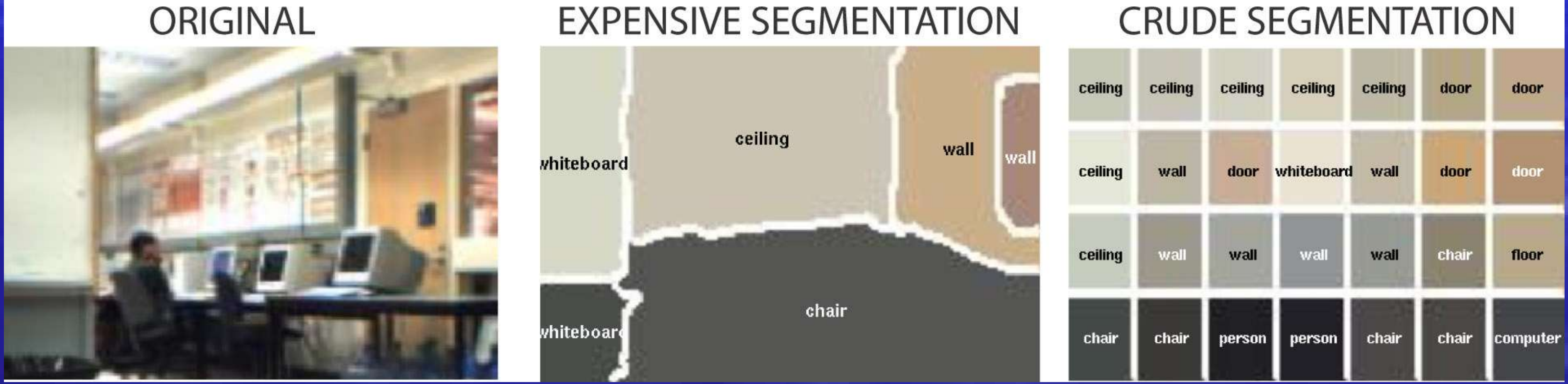
- statistical model for learning the probability that a word is associated with an object in a scene.
- learn these relationships without access to the correct associations between objects and words.

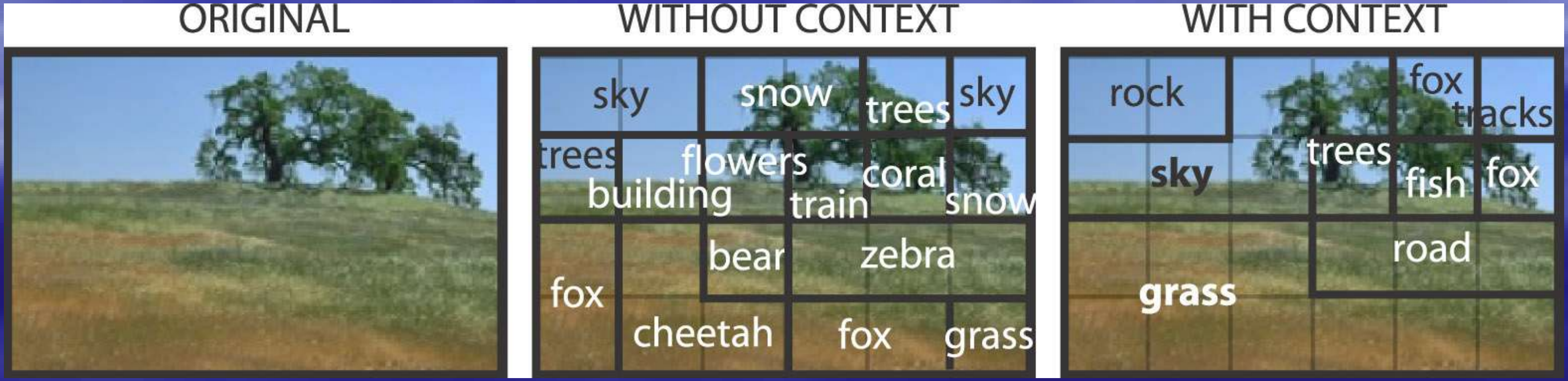


- a Bayesian scheme for automatic weighting of features (e.g., colour, texture, position) improves accuracy by preventing overfitting on irrelevant features.

Contextual Translation

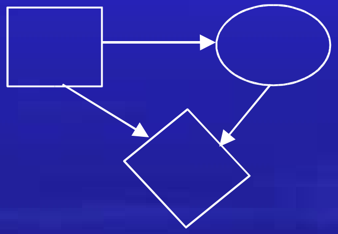
- Issues: image segmentation and how context helps


ORIGINAL
EXPENSIVE SEGMENTATION
whiteboard
ceiling
wall
wall
whiteboard
chair
CRUDE SEGMENTATION
ceiling ceiling ceiling ceiling ceiling door door
ceiling wall door whiteboard wall door door
ceiling wall wall wall wall chair floor
chair chair person person chair chair computer



Kevin Murphy

- Machine learning/ computational statistics
  - Probabilistic graphical models (PGMs)



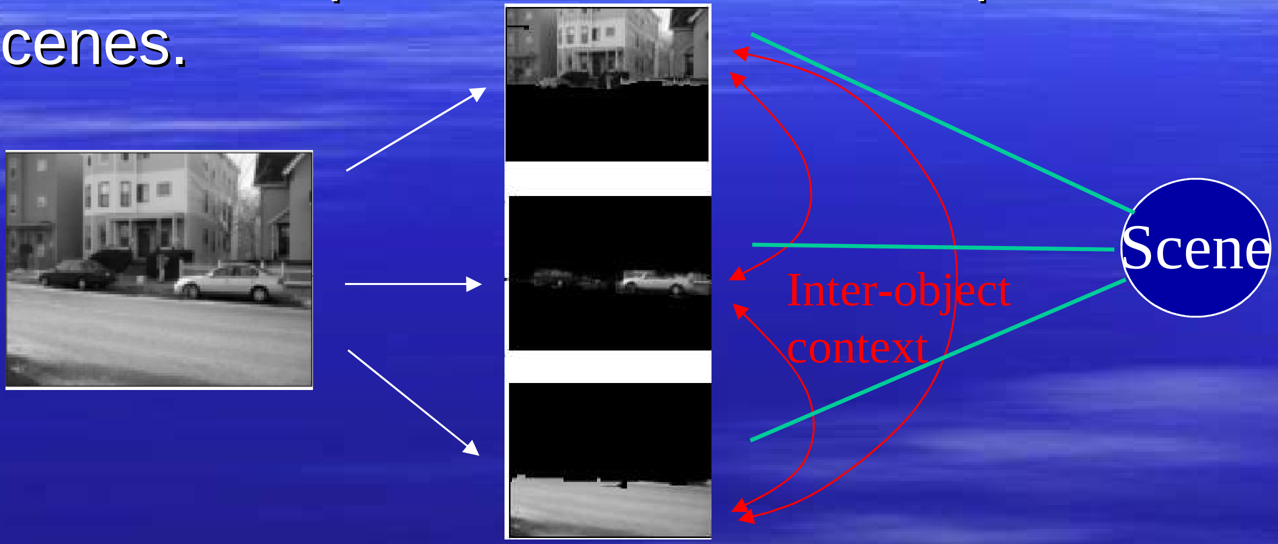
  - Combines graph theory and probability theory

  My focus:

  - Efficient (exact and approximate) inference algorithms
  - Flexible software toolkits (e.g., BNT)
- Applications to computer vision
  - Visual object detection and scene understanding

Visual object detection and image understanding

- My focus: model probabilistic relationships between objects and scenes.



Scene

Inter-object context

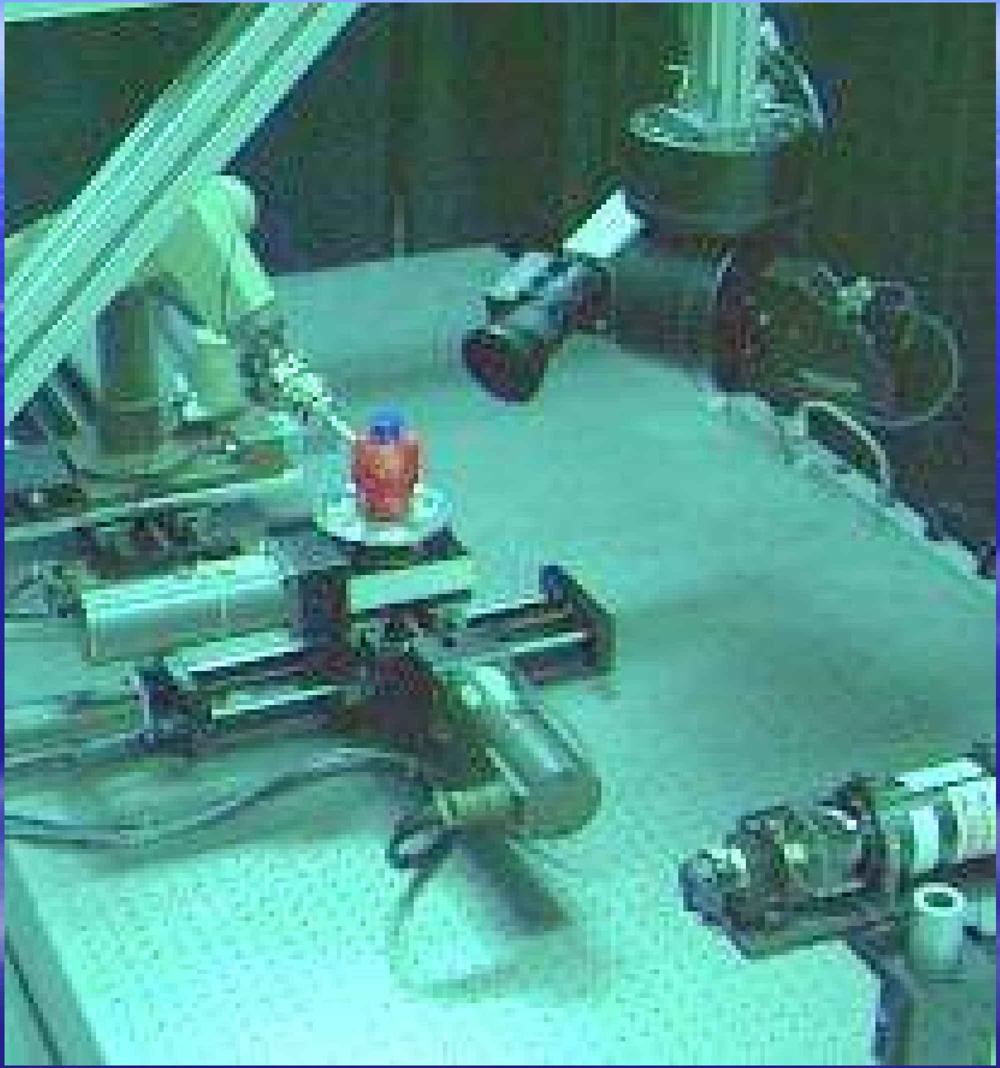
- Applications to wearable computing and mobile robotics.

Robert J. Woodham

- Focus: Computer interpretation of 3D shape and visual motion
- Objective: To understand how the measurement of visual motion can support high-level interpretation tasks related to object identity, non-visual physical properties and, for an object that is an intelligent agent, to actions and intentions.
- Strategy: Link the interpretation of motion and 3D shape *as early as possible* in visual processing.

Research interests

- Image databases and content-based image retrieval
- Remote sensing and geographic information systems (GIS)
- Connections to biological vision, especially colour vision



LCI-related Grad Courses

Term 1

- 502: AI I – Mackworth
- 505: Image Understanding 1 – Little
- 532c: Graphical Models – K Murphy

Term 2

- 525: Image Understanding 2 – Lowe
- 540:   Machine Learning – de Freitas

FIN

What were the successes and failures?

- Failure: Image understanding systems but currently urban image analysis, e.g. building detection an rec
- Successes
  - new algorithms: normalized cuts, graph methods for stereo/motion
  - Generative methods in general
  - Bayesian methods

What are the new developments in the last decade which give us hope to overcome past difficulties?

- Most substantial change has been the change in focus from GOFAIR to situated agents living in dynamic unstructured environments
- dynamics mean that the low-level (subsymbolic) elements of vision become time-critical HENCE real-time stereo and motion analysis; attention becomes central – we know now from studies of inattentional blindness that representation of the world is incomplete at best
- movement from engineered systems with knowledge bases
- movement from, e.g., detectors that are generic to learned systems that operate at high speed, e.g., Viola's face detector using boosting methods and fast implementation methods

Hopeful new developments in the last decade (cont.)

- Move from a focus on pure deduction to a balanced approach with a focus on proper accounting for random variables in vision systems
- lead to a concentration on probabilistic methods
- e.g., mixture models
- Bayesian priors are replacing explicit knowledge as a focus

What is the relation between explicxit and implicit knowledge in cognitive vision?

- What is implicit knowledge? Generic vs specific data?
- Knowledge about the structure of the world – the connections of dependence/independence relations
- Is explicit knowledge manipulable? Symbolic?
- What is the status of priors?

Motivation

- An agent needs a map to reason about actions
- Where am I? localization aids the agent to begin reasoning with its location in the map
- The map itself is an important artifact for use by others: stereo gives shape data, images
  - Virtualized reality by Kanade et al. (CMU)
  - Ikeuchi – Kamakura Great Buddha project

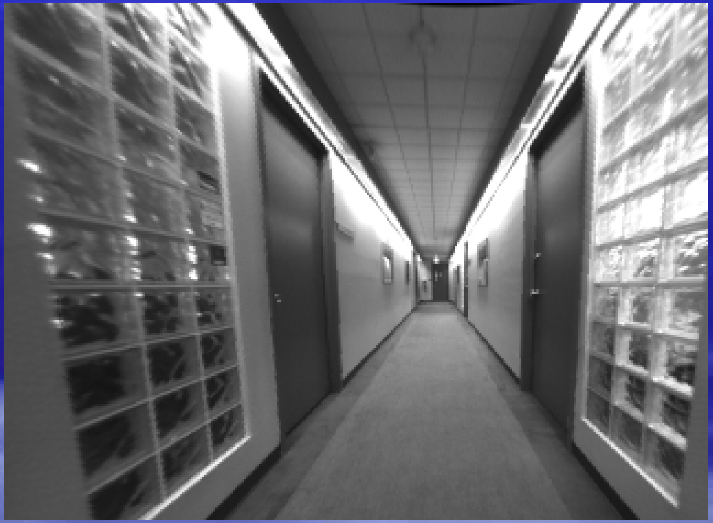
Navigation

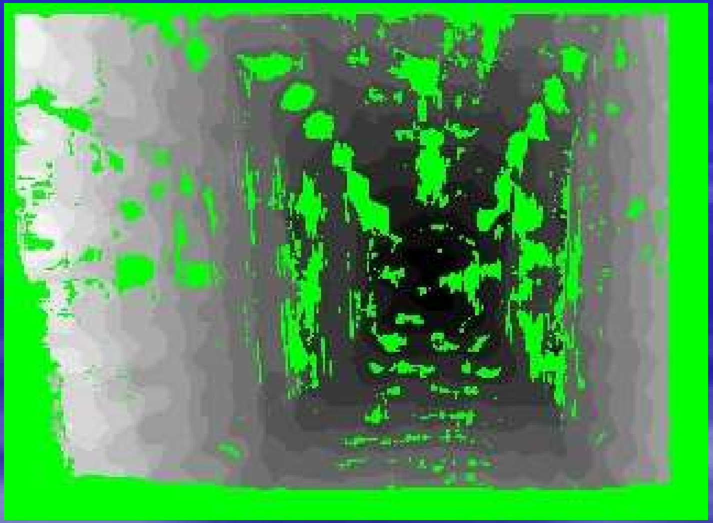
- Vision based
- Build occupancy grid map of static features in the environment (using stereo)
- Path planning on grid
- Detect dynamic obstacles using stereo and bump sensors
- Update map and re-plan

Patchlet Surface Representation

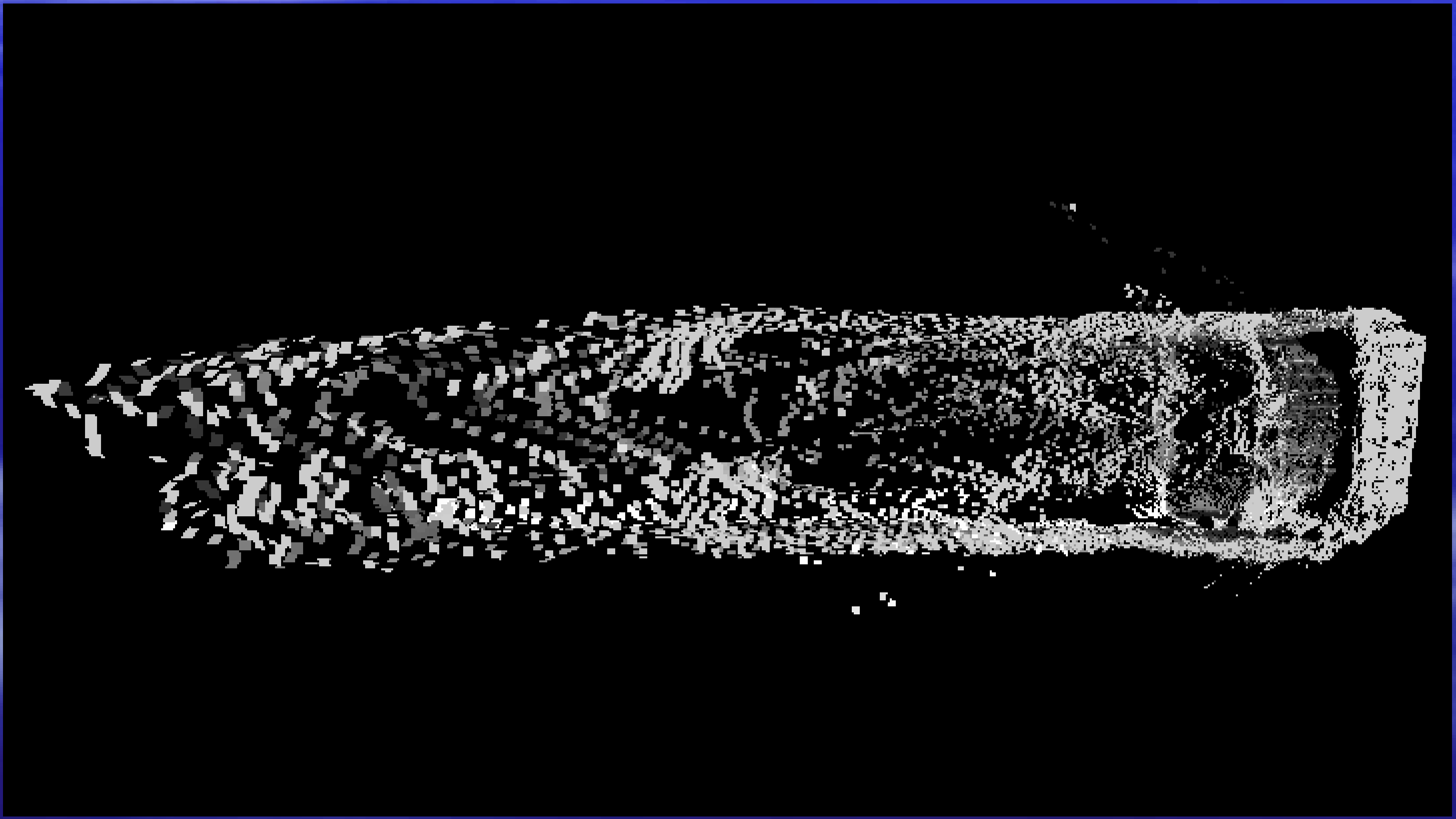
- Goal: to properly interpret the uncertainty of stereo measurements in surface reconstruction.
- The sensor elements considered are local patches in the stereo image that create patchlets.
- These patchlets are fit to a plane and the uncertainty of the plane in orientation and position is determined from the stereo 3d points.

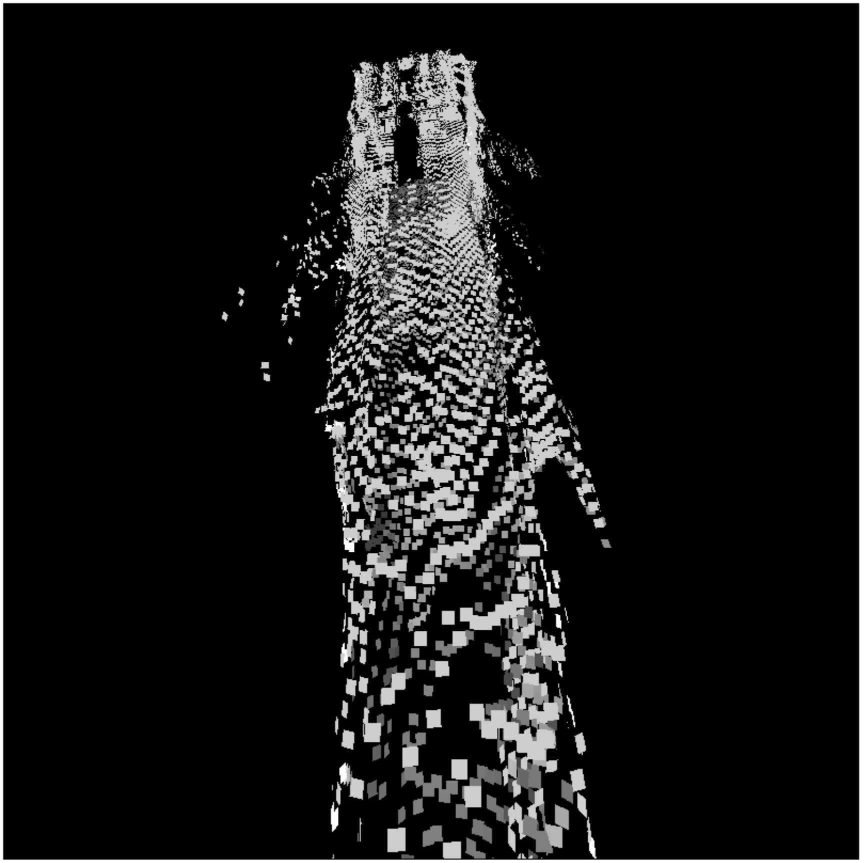
Brightness and Depth





Depth Pixels and Scale





HMMs

- HMMs and models of observable/nonobservable variables
- Observation model – mapping from unobservable states of the system to the sensor observations
- Typically Gaussian models of sensor noise are adopted but modern estimation methods, particularly Monte Carlo – Markov Chain (MCMC) methods enable more realistic and adaptable models of variation
- Hidden Markov Models (HMMs) describe transitions between unobservable states and relations between states and observables
- MDPs and POMDPs – methods for decision making using these models
- Belief networks, Bayesian networks -> dynamic Bayes nets

Action and Motion

- Recognizing people by their gait
- Identifying gestures and expressions
- Analyzing image sequences to identify activities of players from their trajectories – tracking and identification of context

Biometrics

- Recognition by visual or other means of the persons with whom the agent interacts
- The basic methods of measurement of shape, colour, motion, and so forth are common with analysis of gesture, and recognition of activity

Overview

We have developed a novel vision system that can recognize people by the way they walk. The system computes optical flow for an image sequence of a person walking, and then characterizes the shape of the motion with a set of sinusoidally-varying scalars. Feature vectors composed of the phases of the sinusoids are able to discriminate among people.

Input Sequence

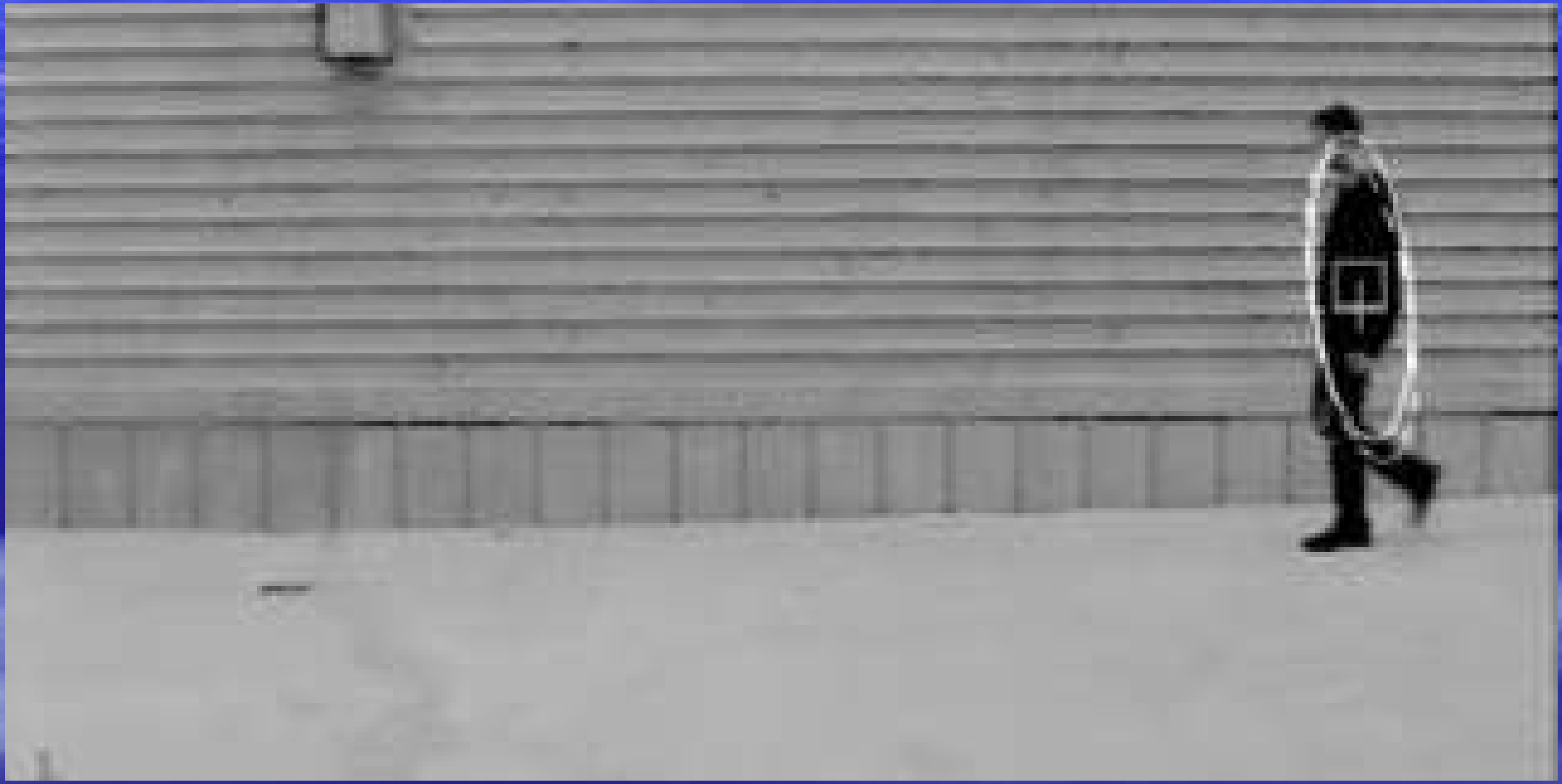
Optical Flow

Optical flow (Little, Bulthoff and Poggio): *n* frames of (*u*,*v*) data, where *u* is the *x* flow and *v* is the *y* flow.

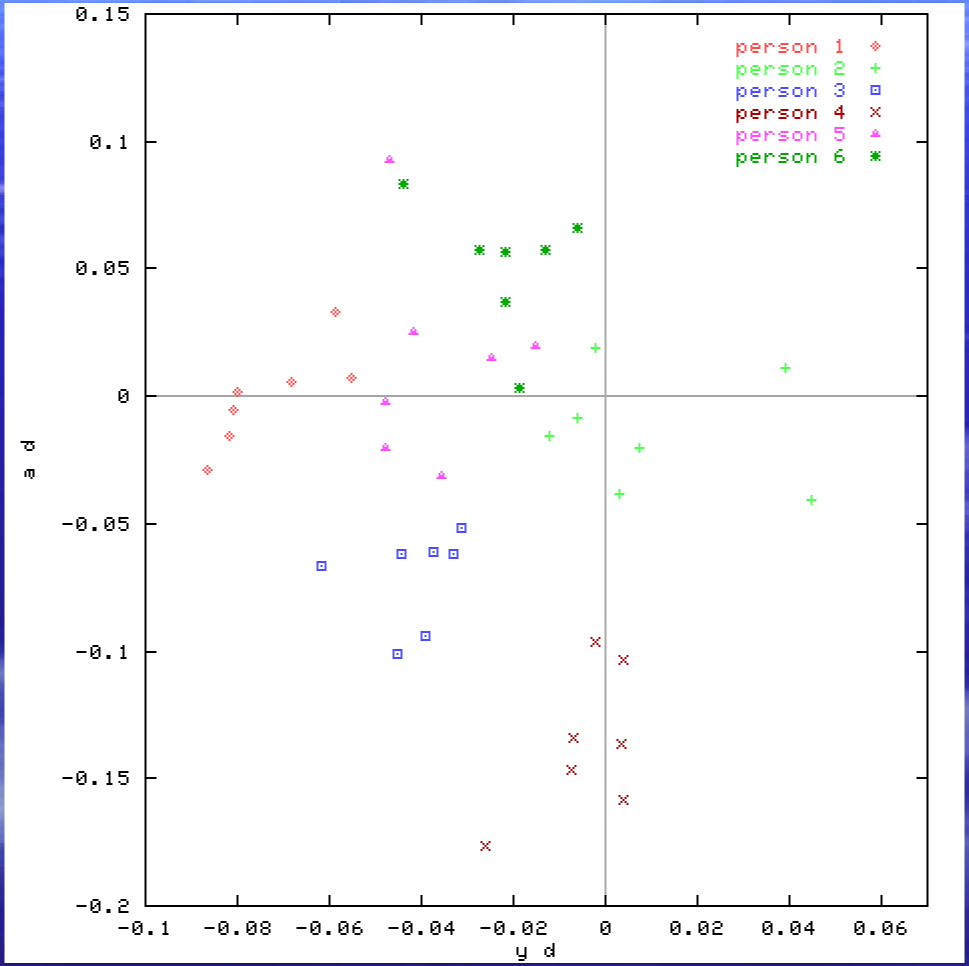
Points (white) where flow is non-zero.

Magnitude of flow

Fitting Ellipses to motion



Scatterplot



Separation is sufficient so that the k-nearest neighbours classification succeeds 96%

Surveillance and recognition

Jim Little

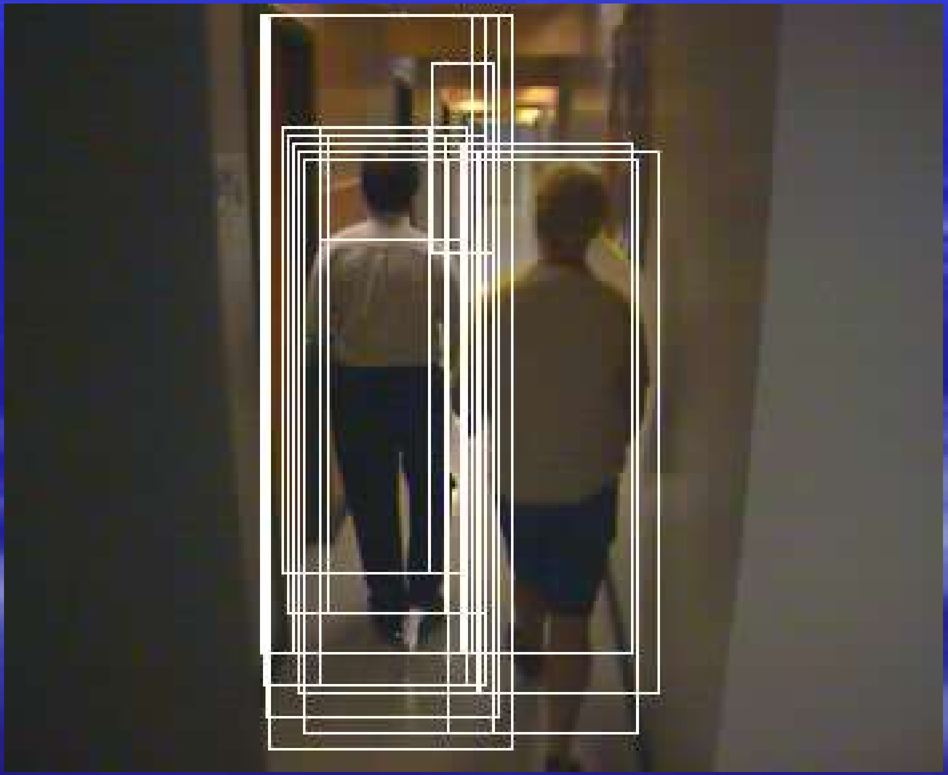
Work by Jim Clark and his students (McGill University)

Multiple Camera Area Surveillance Techniques

- To distinguish "normal" people, objects, and activities from anomalous ones, and alert a security agent in the case of anomalous conditions.
- Our current projects are:
  - Fast people detection in corridor images, to be used as input for activity recognition.
  - Similarity filter development, for knowing when a given scene has been seen before.
  - Context-based object and scene recognition algorithms.

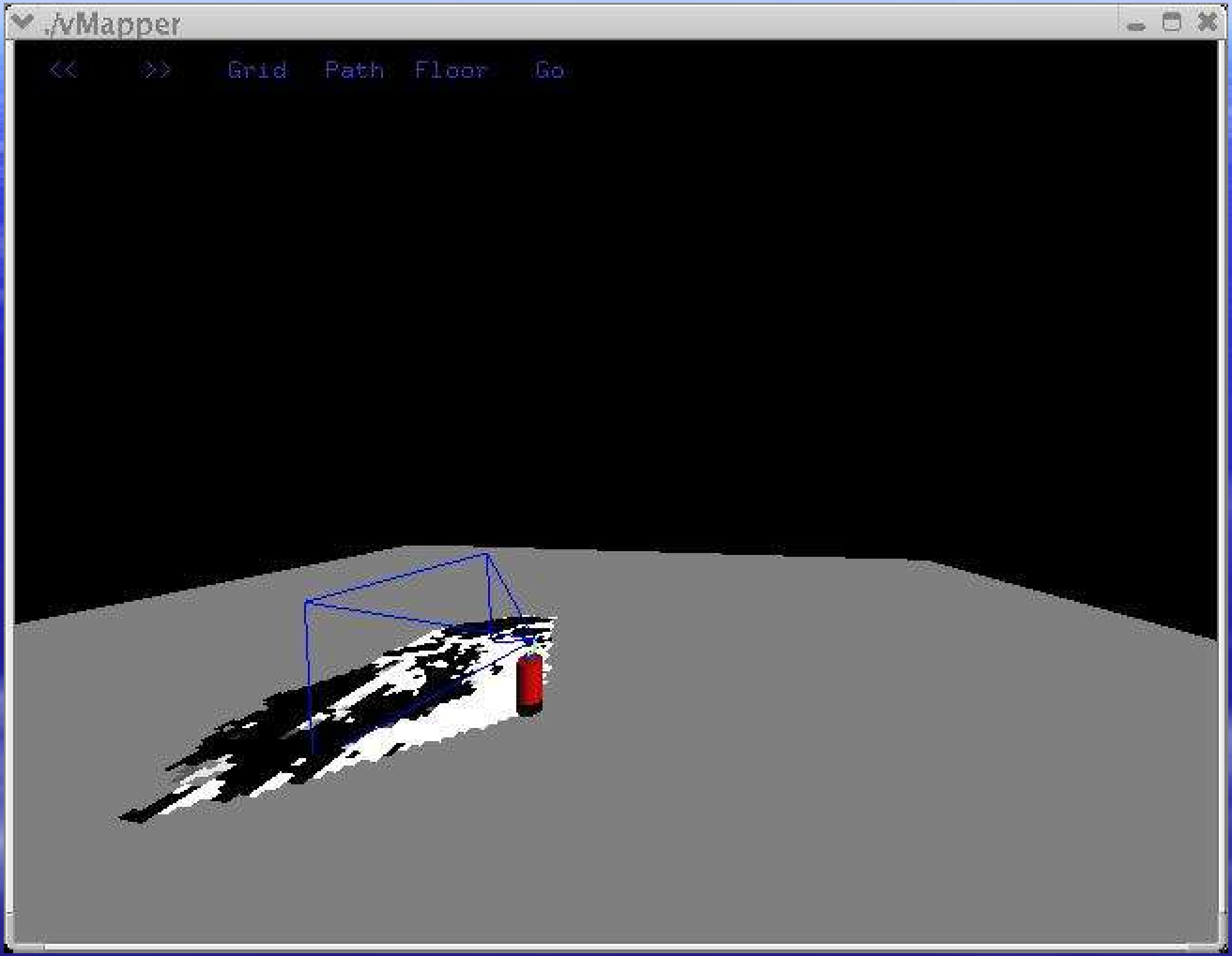
People Detection Algorithm

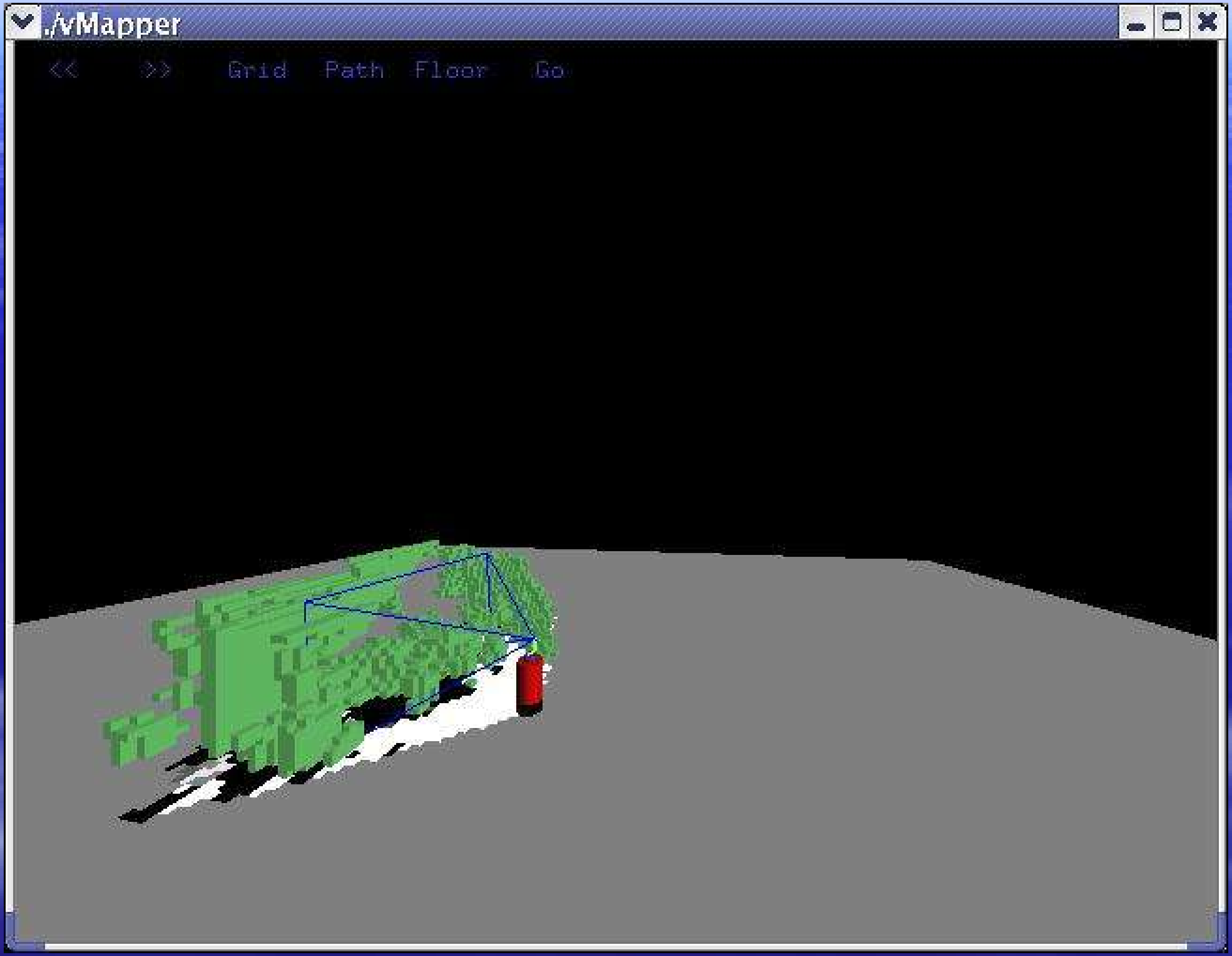
- uses JPEG encoded images from a network camera. A *large* set of image features is computed, *without* the need for complete JPEG decompression.
- A support-vector machine (SVM) is used to classify image regions into people or non-people regions.
- A related project is developing an FPGA hardware implementation.



What can be done in the next few years, in the next 5 - 7 years, what are long term goals?

- Return back to our roots in AI – reasoning is required to move past the most simple methods using priors, which themselves have little structure
- Reasoning allows one to specify structure
- How can we introduce knowledge bases into vision
- From a statistical model of images to the next level – association of objects in scenes ?
- Local, within contexts, learning of association – requires some reasonable form of
- Segmentation – identification of constituent units
- Rich description


./vMapper
<<
>>
Grid
Path
Floor
Go


./vMapper
<<
>>
Grid
Path
Floor
Go

Lip-Reading

- Tulips1 database
- 12 subjects - 4 words

Affine (2 ZPs):	**66%**
First 7 ZPs	**76%**
2,4,8,9,10,14,22:	**79%**

(96 sequences, 835 frames)