Active Visual Search:

Have I Learned Anything Since 1986?

John K. Tsotsos
Dept. of Computer Science & Engineering and
Centre for Vision Research
York University
Toronto, Canada

Active Perception: Not a New Idea

- ✧ Brentano 1874
  - *Act Psychology* - an act is a mental activity that affects percepts and images
- ✧ Barrow and Popplestone 1971
  - *...consider the object recognition program in its proper perspective, as a part of an integrated cognitive system. One of the simplest ways that such a system might interact with the environment is simply to shift its viewpoint, to walk round an object. In this way, more information may be gathered and ambiguities resolved. A further, more rewarding operation is to prod the object, thus measuring its range, detecting holes and concavities. Such activities involve planning, inductive generalization, and indeed, most of the capacities required by an intelligent machine.*

Continuing...

- ✧ Metzger 1974
  - provides list of reasons for active behavior
- ✧ Bajcsy 1985
  - intelligent control applied to the data acquisition process that depends on the current state of data interpretation including recognition.
- ✧ Aloimonos 1991
  - vision is the process of deriving purposive space-time descriptions
- ✧ Tsotsos 1992
  - comparing complexity of active vs passive perception; role within attentive vision

Why Active?

- To move fixation point/plane or to track motion
- To see a portion of the visual field otherwise hidden due to occlusion
  - manipulation
  - viewpoint change
- To see a larger portion of the surrounding visual world
  - exploration
- To compensate for spatial non-uniformity of a processing mechanism
  - foveation
- To increase spatial resolution or to focus
  - sensor zoom or observer motion
  - adjust camera depth of field, stereo vergence
- To disambiguate or to eliminate degenerate views
  - induced motion (kinetic depth)
  - lighting changes (photometric stereo)
  - viewpoint change
- To achieve a “pathognomonic” view
  - viewpoint change
- To complete a task
  - multiple fixations

Eliminate the need for Attention or Active Vision?

Much current research assumes the extent of the search space is seriously reduced before visual processing takes place, and most often even before the algorithms are designed:

- Fixed camera systems negate the need for selection of visual field
- Images out of their spatiotemporal context eliminate need for tracking
- Pre-segmentation eliminates the need to select a region of interest
- Clean backgrounds ameliorate the segmentation problem
- Assumptions about relevant features and the ranges of their values reduce their search ranges
- Knowledge of task domain negates the need to search a stored set of all domains
- Knowledge of which objects appear in scenes negates the need to search a stored set of all objects
- Knowledge of which events are of interest negates the need to search a stored set of all events

Visual Attention

(ICCV 1987, IJCV 1988, BBS 1990, IJCV 1992, AIJ 1995, CVIU 2005, JOV 2009...)

- ✧ Computational Complexity Explains the Attentional Bottleneck
  - formalizes and quantifies the nature of the problem and points to solutions
- ✧ How the Brain Cures the Complexity Curse
  - the general vision problem is not the one the brain is solving
- ✧ The Cure's Side-Effects
  - pyramid representations have many problems; but these problems constrain solutions
- ✧ Selective Tuning
  - a model of attention and vision where a generic architecture is dynamically tuned to optimize it for the current visual task
- ✧ Explanations and Predictions
  - ST has strong predictive power as evidenced by many experiments

Have proved 9 theorems that are central to the problem:

1: Unbounded (Bottom-Up) Visual Match (UVM) is NP-Complete, with time complexity an exponential function of *P* (the worst-case time complexity is $O(P\ 2^P))$ (median case also exponential).

2: Bounded (Task-Directed) Visual Match (BVM) has time complexity linear in *P* (the worst-case time complexity is $O(P\ ||G||))$. (median case also linear).

3: Unbounded Visual Search (UVS) is NP-Complete.

4: Bounded Visual Search (BVS) has time complexity linear in *P*.

5: Active Unbounded Visual Search Is NP-Complete.

6: Active Bounded Visual Search has time complexity linear in *P*.

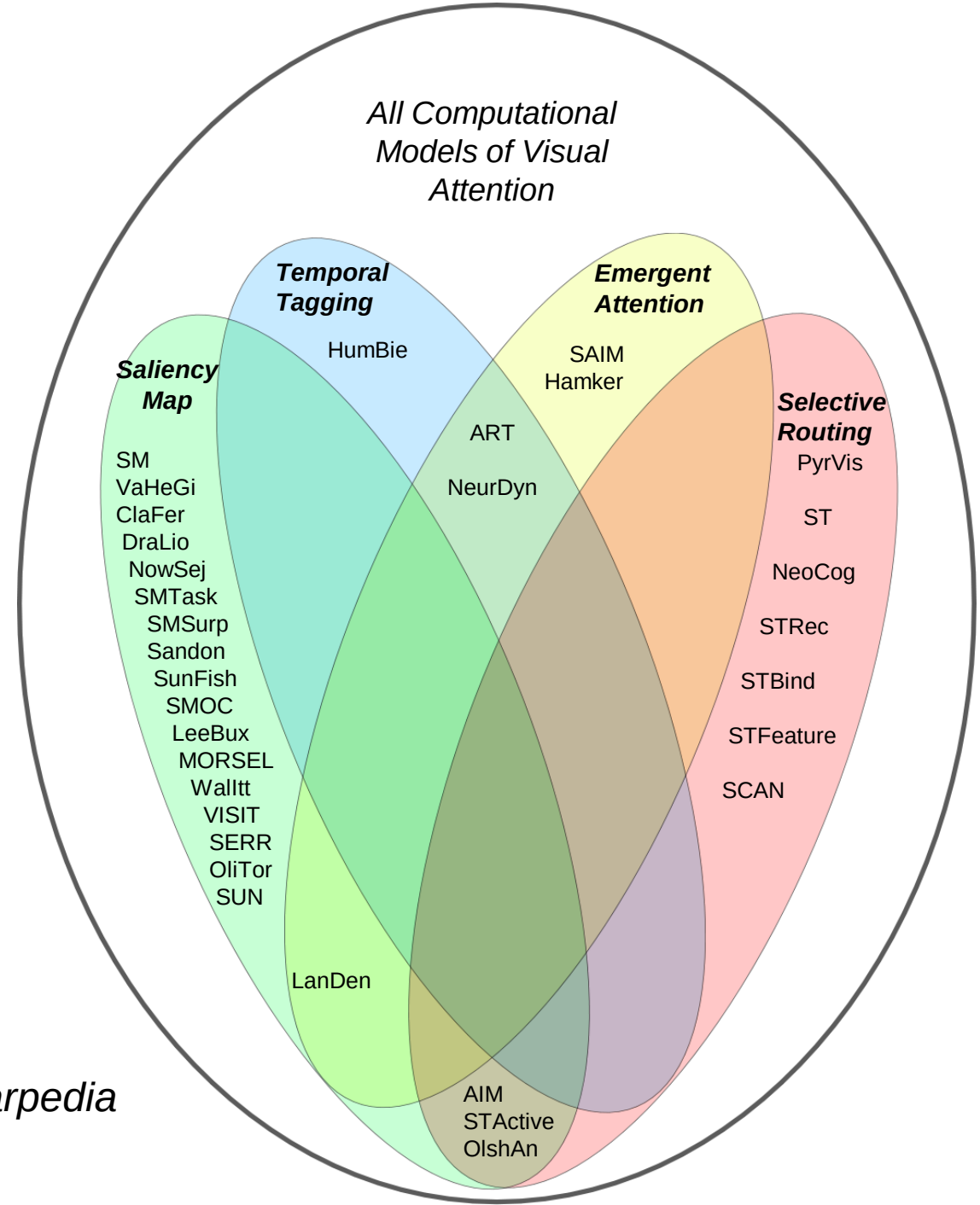
7: Unbounded Stimulus-Behavior Search is NP-hard.

8: Bounded Stimulus-Behavior Search has time complexity linear in *P*.

9: Object Search (sensor planning) is NP-hard.

Visual Attention

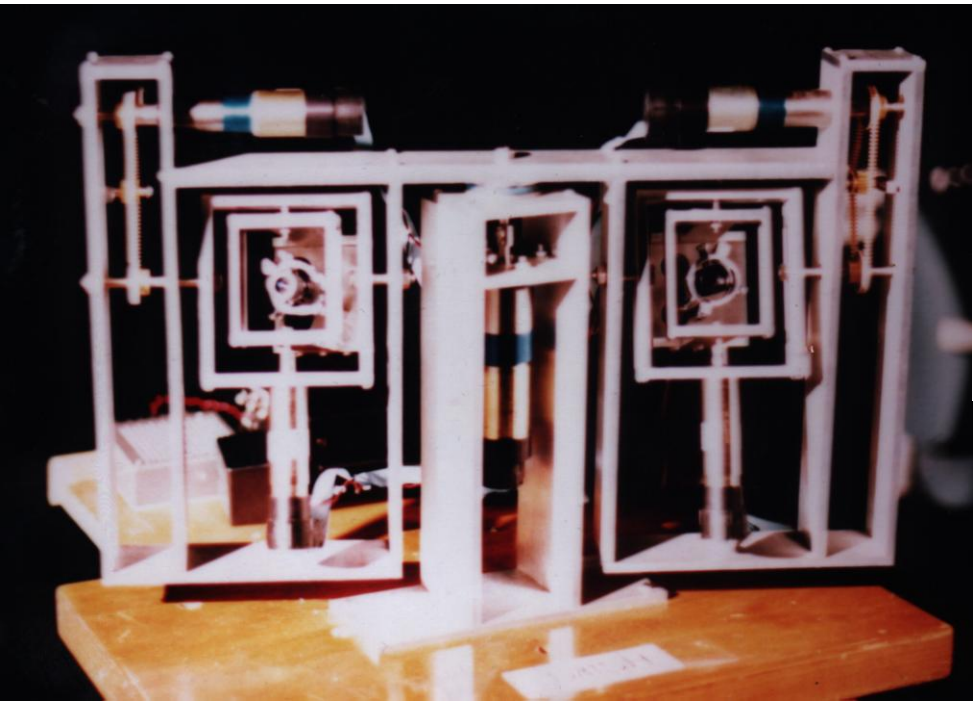
Modeling is in its infancy

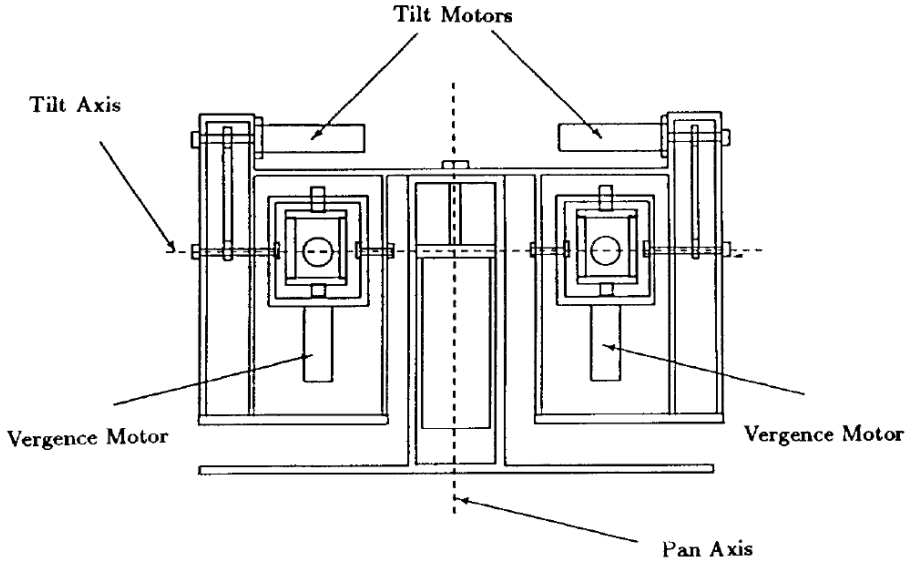


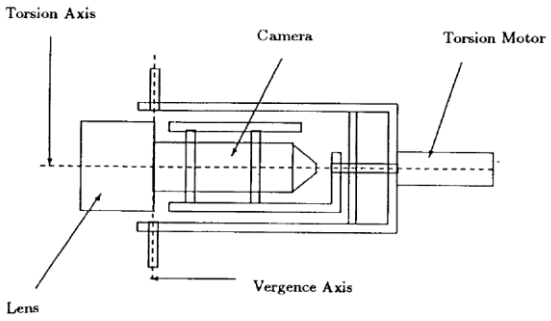
to appear in *Scholarpedia*

The TRISH Project (Milios et al. IJPRAI 1993)

- ✧ TRISH – Toronto IRIS Stereo Head – Version 1
  - 7 dof



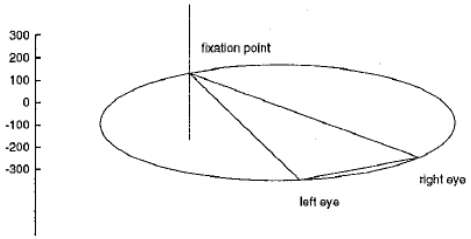


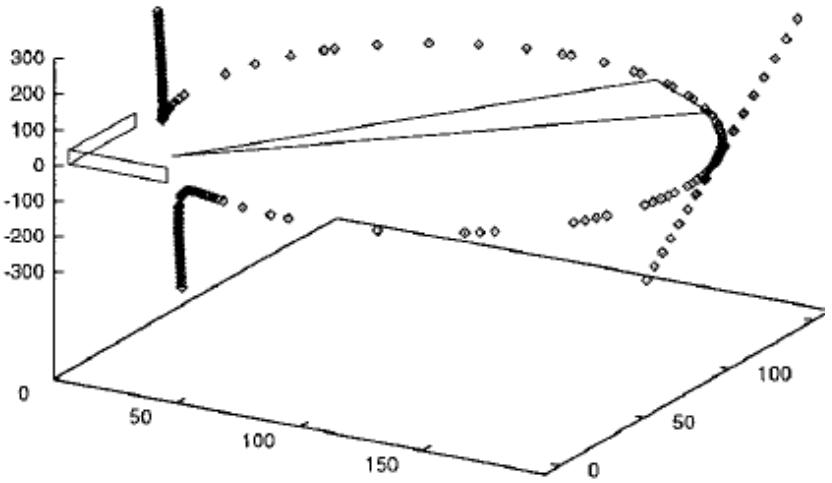


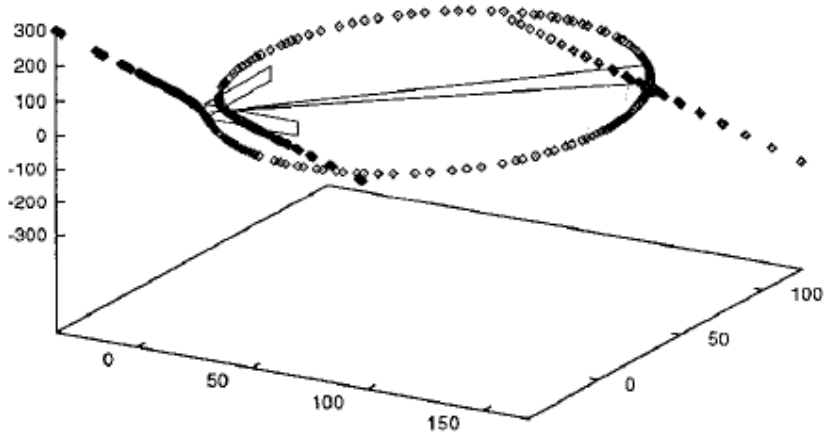
TRISH1 Performance

Jenkin & Tsotsos CVPR 1994, Jenkin et al. IAPR 1994

	Human performance			TRISH design		
	range (deg)	velocity (deg/second)	accuracy (deg)	range (deg)	velocity (deg/second)	accuracy (deg)
Pan	±90			±80	54	0.003
Tilt	±45	800	2	±45	54	0.003
Torsion	+20			±180	180	0.09
	Horoptor curve	800	2	±35	100	0.0019



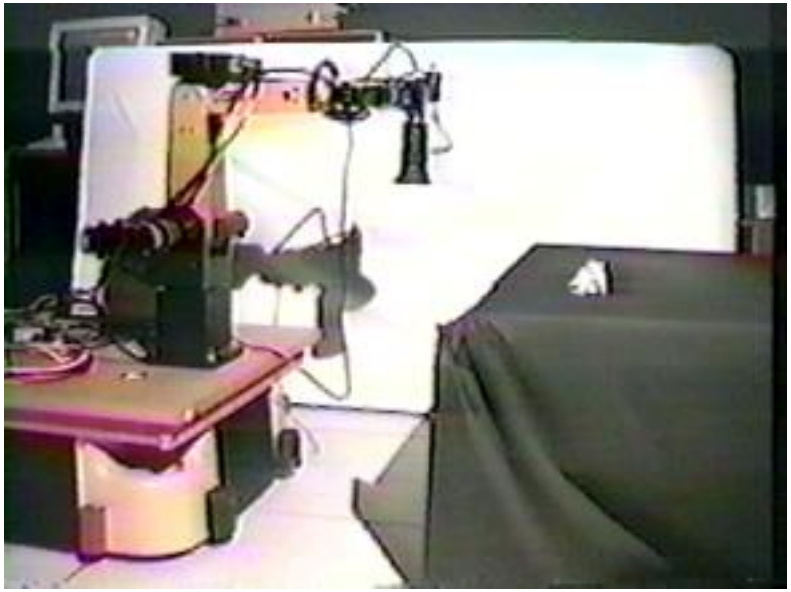




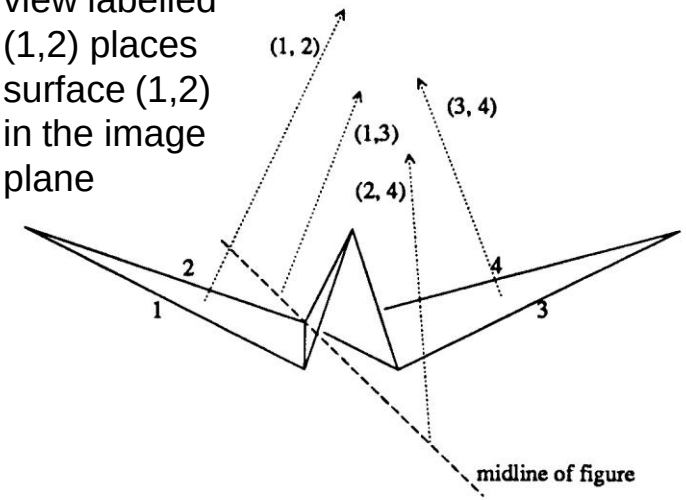
Active Search for Pathognomonic Views

Wilkes & Tsotsos CVPR 1992
Wilkes 1994

scene of origami objects

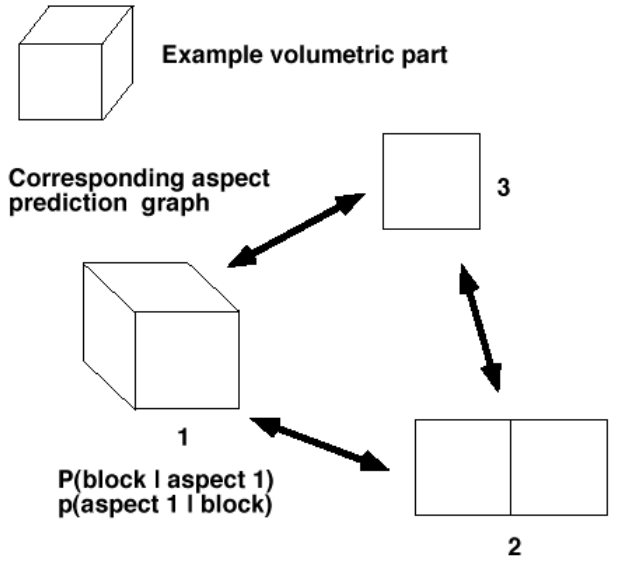


view labelled (1,2) places surface (1,2) in the image plane

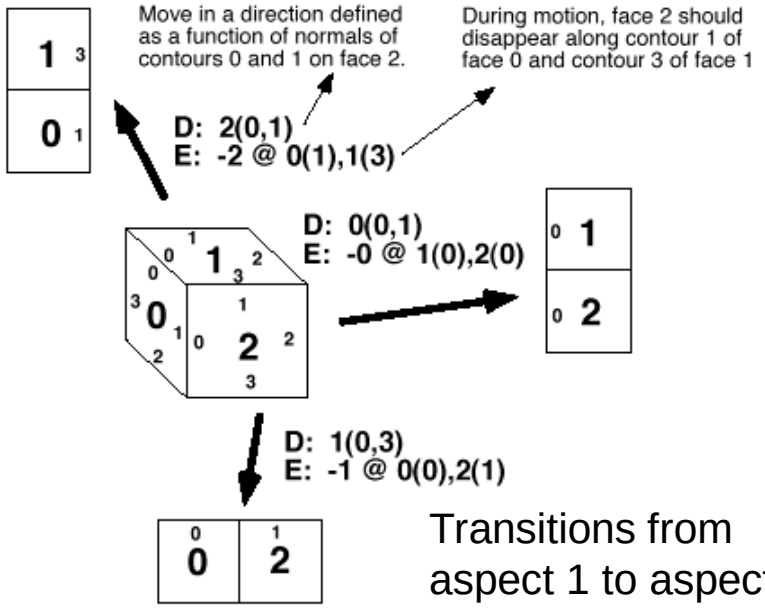


Active Search for Non-Degenerative Views

Dickinson et al. CVIU 1997; PAMI 1999; Wilkes et al. ICCV 1995

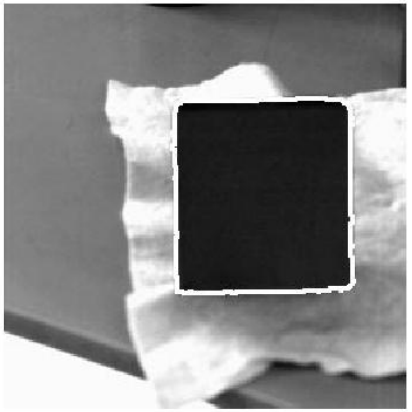


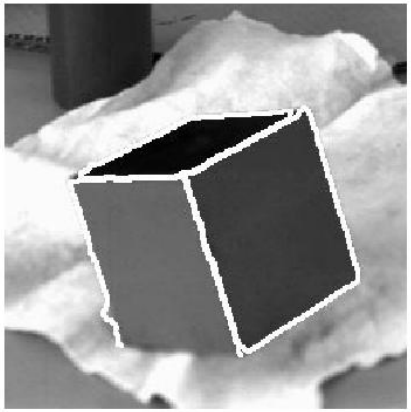
Aspect prediction graph for a block



Transitions from aspect 1 to aspect 2

an ambiguous (degenerate) view





moving to the least ambiguous aspect

The Object Search Problem

Ye & Tsotsos 1996; CVIU 1999; CIJ 2001; Shubina & Tsotsos CVIU 2010

Definition: Select a sequence of actions that will maximize the probability that a mobile robot with active sensing capabilities will find a given object in a partially unknown 3D environment with a given set of resources.

Find $F \subset O_\Omega$ which satisfies $T(F) \leq K$ and maximizes $P[F]$

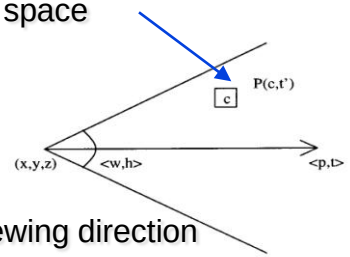
F	the ordered set of actions applied in the search
O_Ω	Ω is 3D region (union of space elements); O_Ω all possible actions that can be applied to Ω
T(F)	total time required to apply F
P[F]	probability of finding the target with F
K	a constant - maximum acceptable time

Problem is provably NP-hard

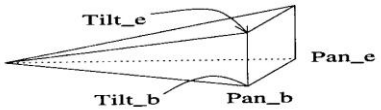
Since the problem is NP-hard, try a heuristic approach

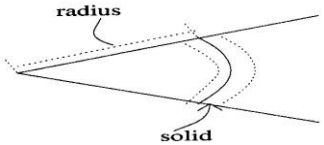
Representing the world and the actions

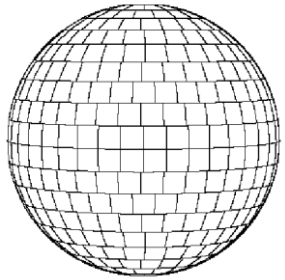
world is the union of cubes, each has an associated value, the probability of the target being found in that bit of space



viewing direction

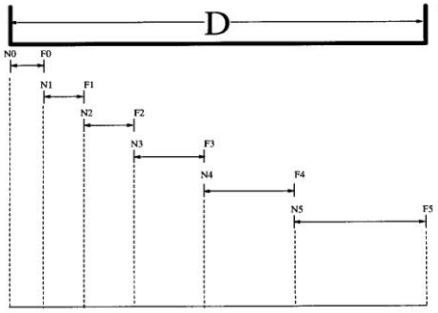


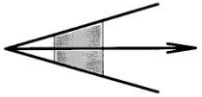




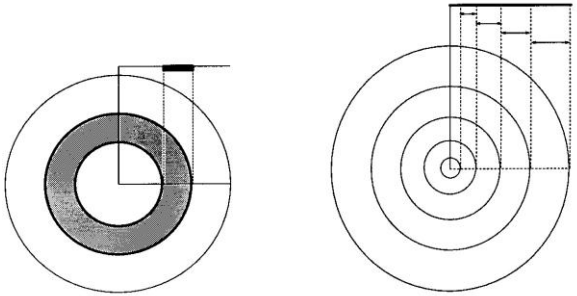
the sphere is partitioned into equally sized solid angles; stereo senses presence of a solid within that angle

sensed sphere

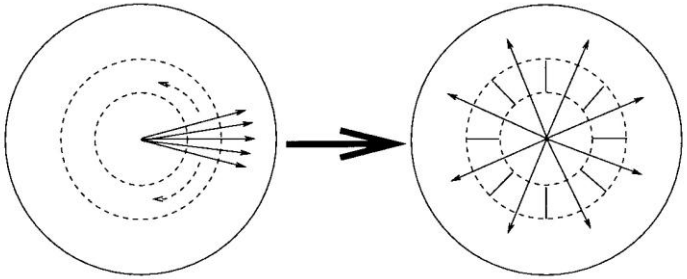




effective depth of field of camera is object dependent (detection function)



partition sphere into layers



partition possible viewing directions into non-overlapping sets

Where to Look Next

Assign uniform probabilities to all cubes in space (or prior knowledge)

Choose where to look next

- Calculate total target probability within each solid angle of the sensed sphere
- Angle with the largest value > Θ_{look} corresponds to the next "best" viewing direction
- Choose sensor settings and execute

Apply a detection algorithm in that direction

If Successful, end

Update target probability distribution (sum remains at 1)
Repeat until viewing angles exhausted

If viewing angles at current position exhausted, move to new position

Where to Move Next

Choose where to move next

- Determine accessible floor positions
- Compute expected sensed sphere for each potential new position Ψ_{pos}
- Choose the largest target probability > Θ_{move} within its corresponding sensed sphere

$$pos_{\tau+} = \arg\max_{pos} \sum_{c_i \in \Psi_{pos}} \mathbf{p}(c_i, \tau_{\mathbf{f}})$$

If position exists, move to new position

Otherwise fail

Determine where to look from new position

The Search Agent

Basic requirements:

- method determining depth
- method for detecting target
- control over sensor parameters
- control over mobility

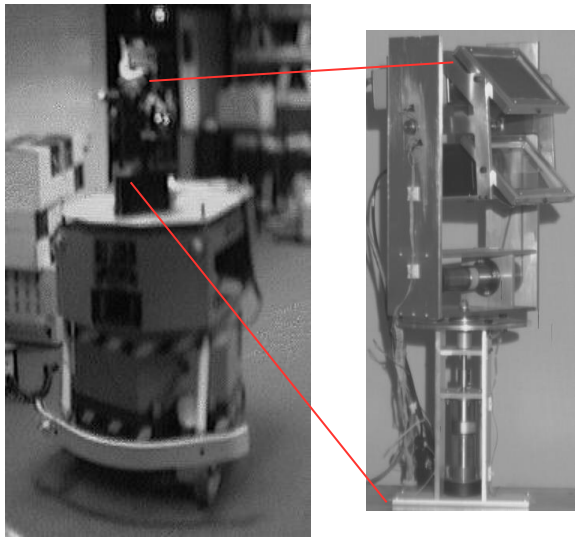
Have tested with three agents:

Shubina's platform

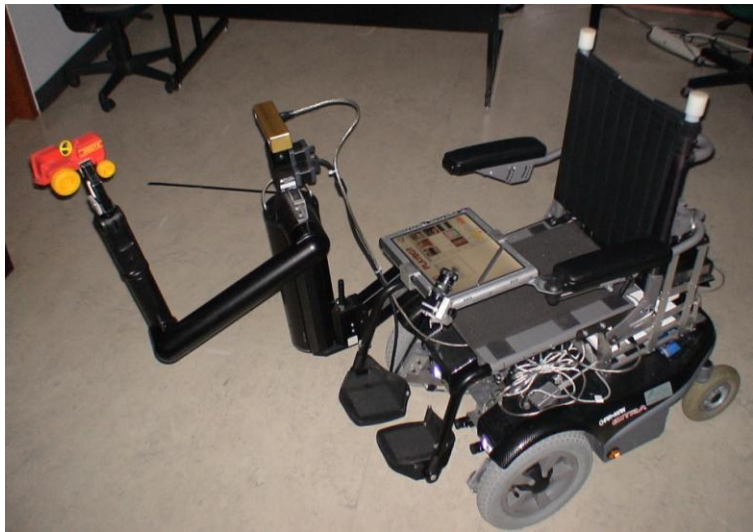
Pioneer 3 robot
Point Grey Research
Bumblebee
Directed Perception
pan-tilt unit
Triclops StereoVision SDK

Ye's platform

Cybermotion
Navmaster
Laser-Eye
sonar
video
infrared



Playbot platform



A Priori Search Knowledge

Among the possibilities, individually or in combination, are:

Type 1 No knowledge
- initial PDF uniform

Type 2 Indirect Search Knowledge (Garvey 1976; Wixson & Ballard 1994)
- intermediate target objects in spatial proximity to target

Type 3 Hints
- initial PDF highlights order of regions to try first

Type 4 Saliency Knowledge (distinctive target features in search region)
(Itti et al. 1998; Frintrop 2005)
- PDF modified by saliency computation on each acquired image

Type 5 Predictive Knowledge (Kelly 1971; Tsotsos 1980)
- prediction due to pre-processing or spatiotemporal constraints

In general, no simplifying assumptions about distributions are possible

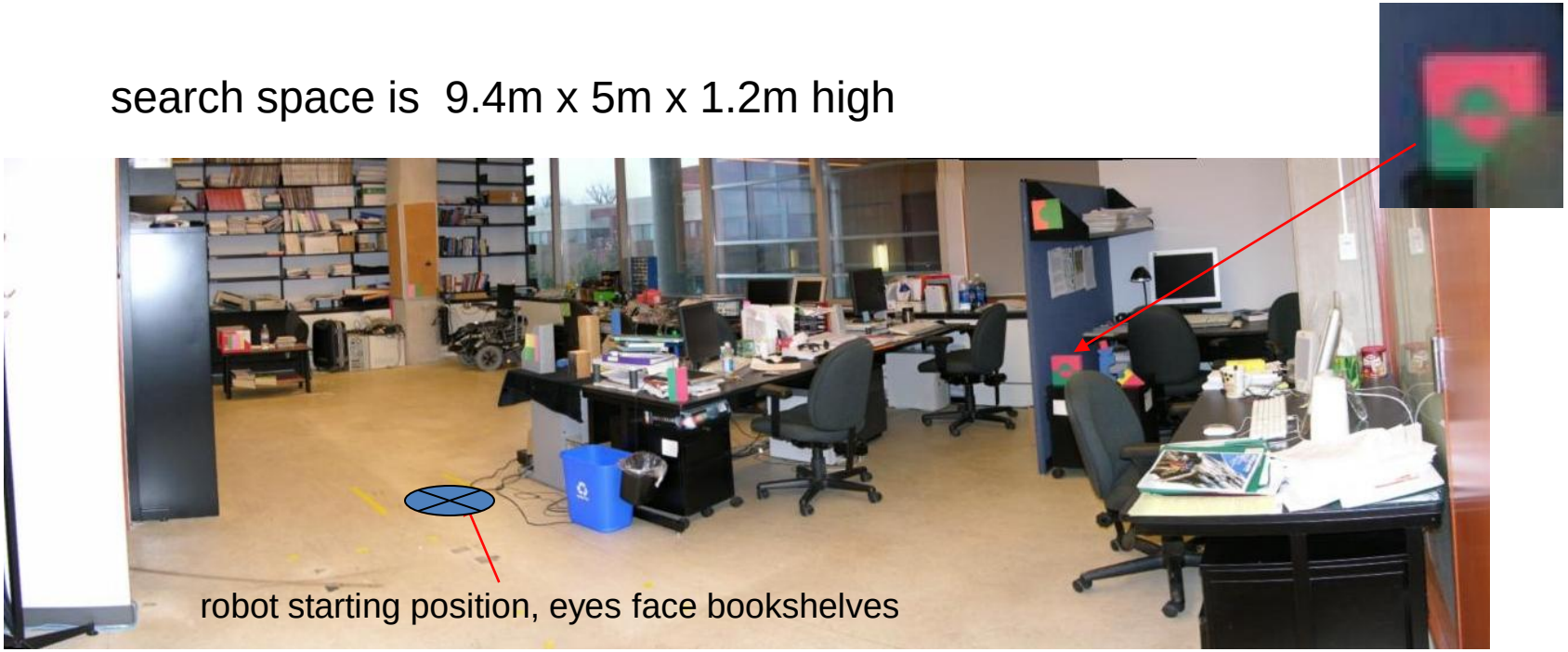
The Current Implementation

Simplifications:

- Type 1 a priori - uniform initial PDF
- tilt limit of 30°
- no focal length control
- location probabilities not viewpoint dependent
- assume recognizable face of object is visible from free space robot has to move in
- no path planner
- used detector of McLean & Tsotsos (2001, 2008) Normalized Grey-Scale Correlation in a Pyramid Image Representation

Simple Example of the Shubina-Ye-Tsotsos (SYT) Strategy

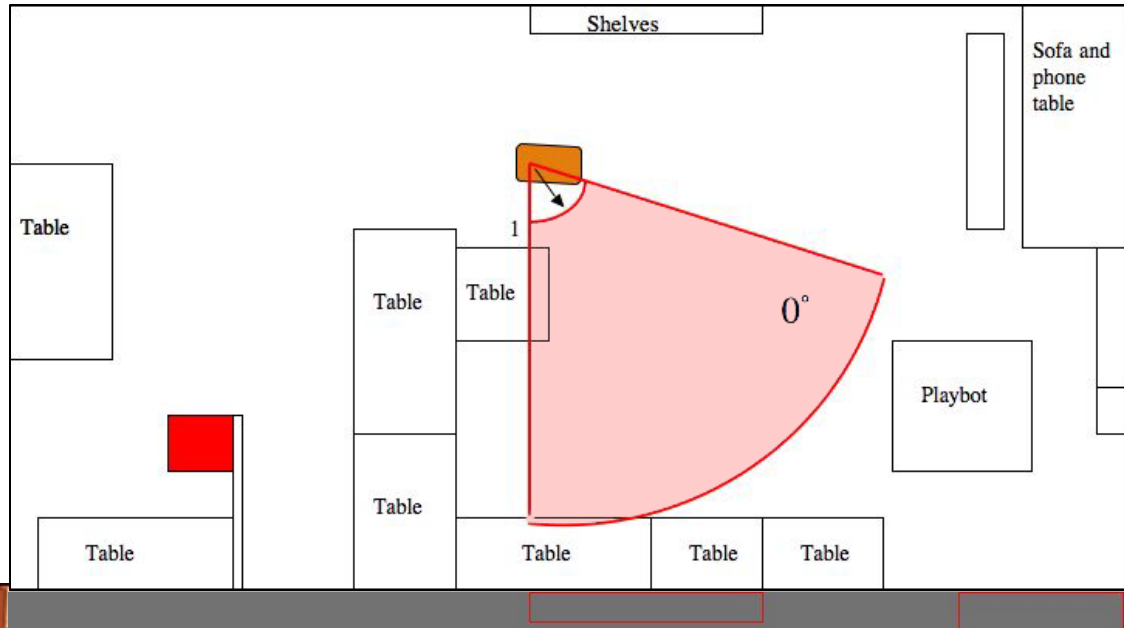
search space is 9.4m x 5m x 1.2m high



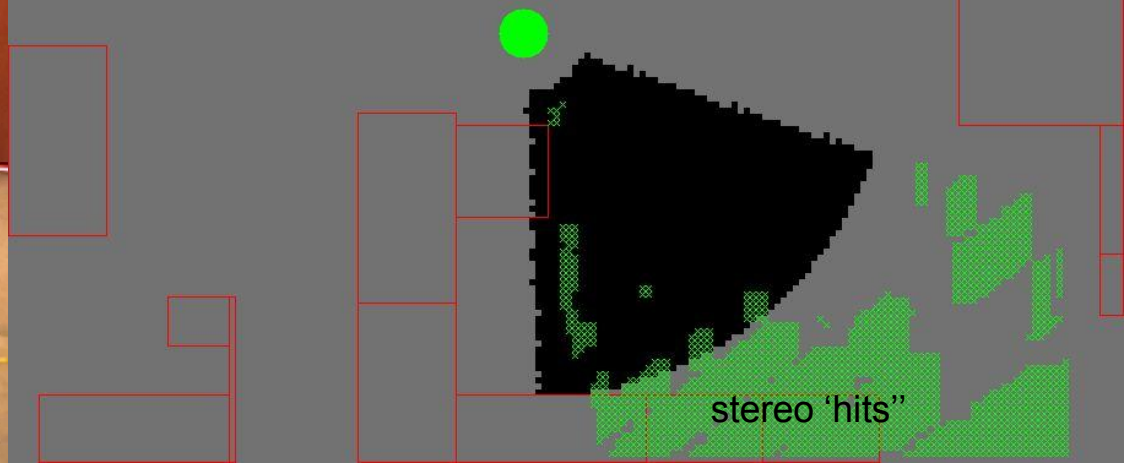
- total possible robot positions: 32 (1m x 1m tiles)
- total possible camera directions at any position: 17
  each covers 70°H x 50°V field of view, depth of field = .5m to 2.5
- total size of action set to choose from at any step: 544
  always includes ‘recognize’ and stereo
- total number of 125cm^3 occupancy grid positions: 451,200
  each codes 2 binary values + probability
- number of possible states = a big number!

Position 1, Sensing Action 1

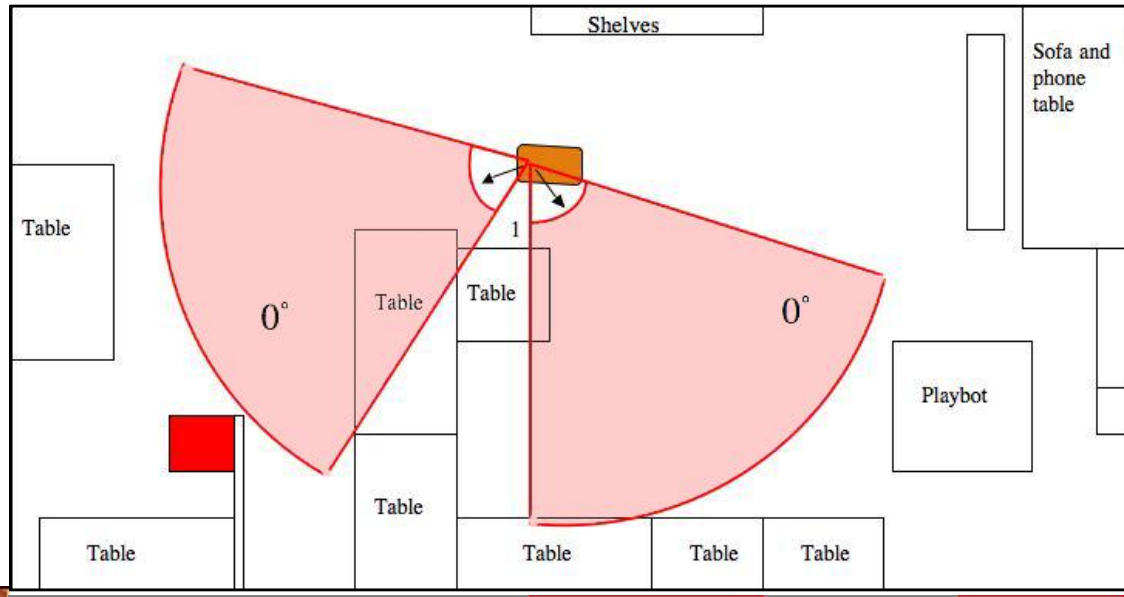
map of lab *not* known,
only exterior walls

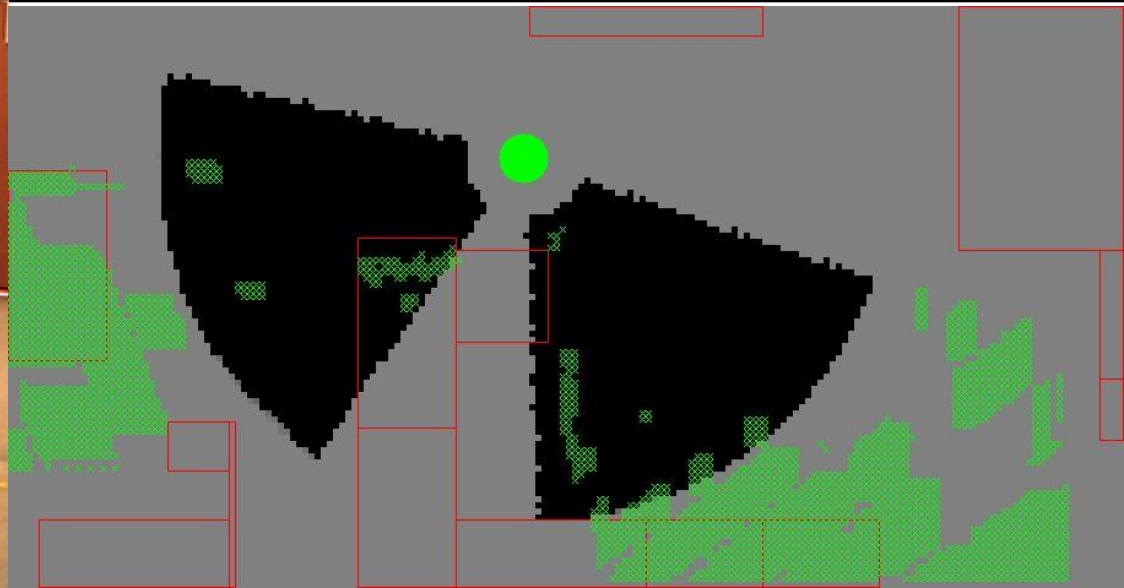


probability of target presence in horizontal plane

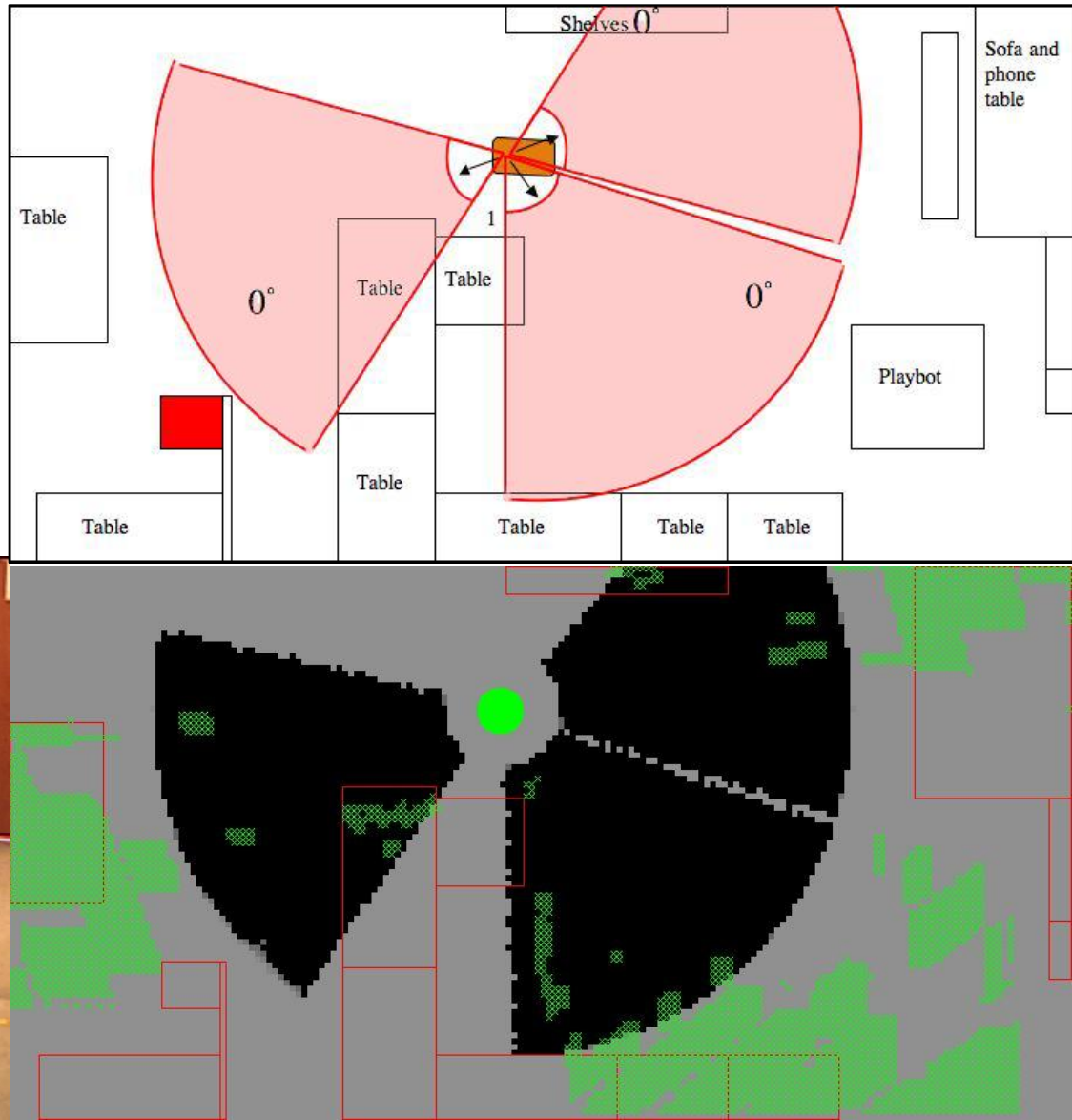


Position 1, Sensing Action 2

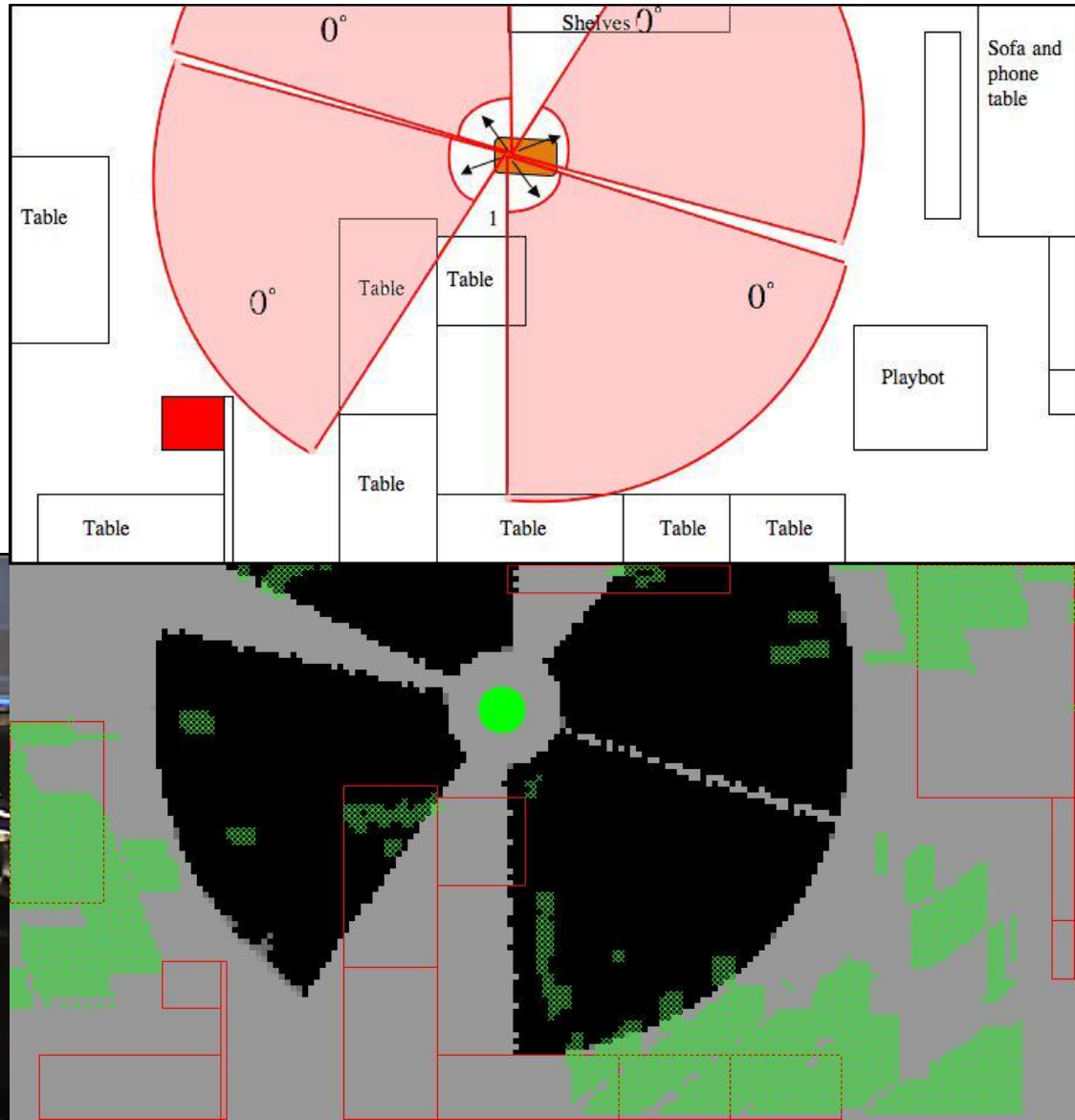




Position 1, Sensing Action 3



Position 1, Sensing Action 4

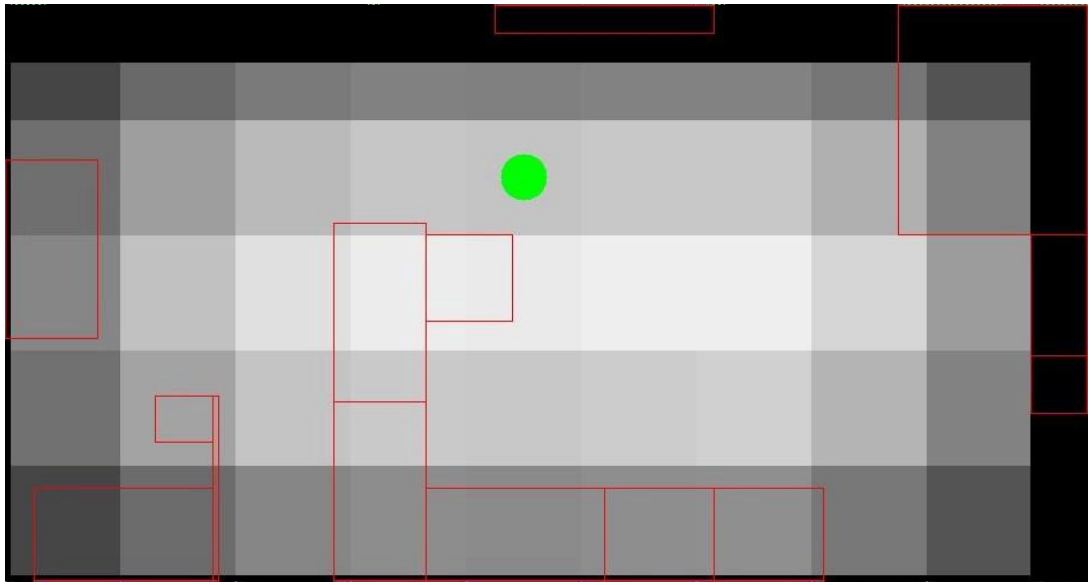


Map of where to move next after Position 1

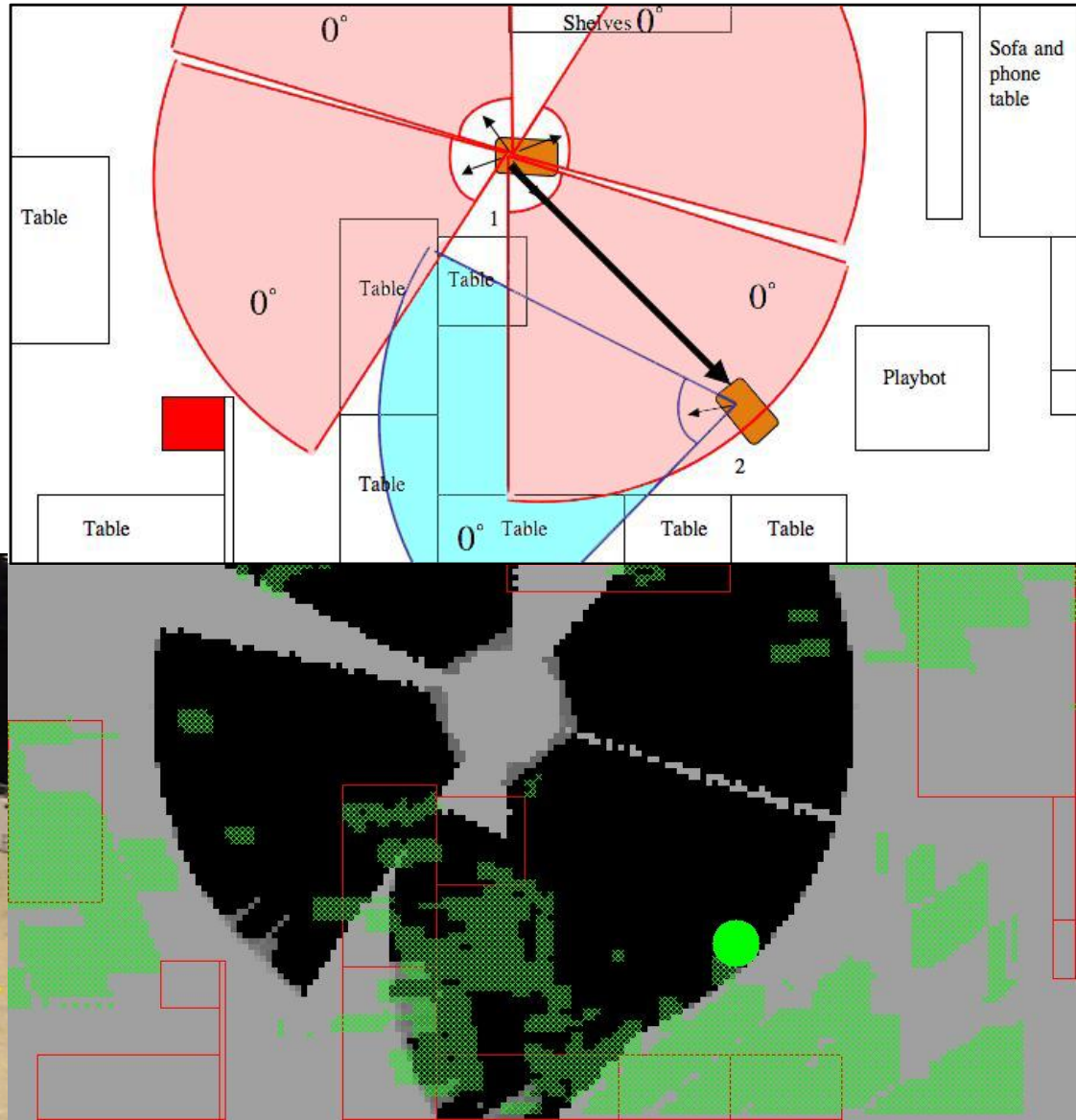
- grey level in each square shows sum of probabilities within sensed sphere from that location (white > black)

- crossed-out regions are inaccessible by robot (lack of map and planner)

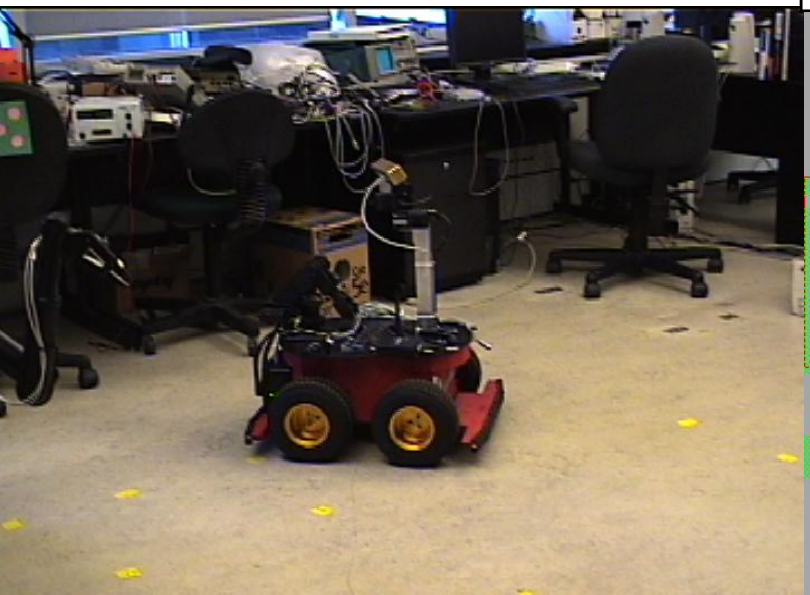
- shows how this map changes for the first 4 actions and before the first move of position

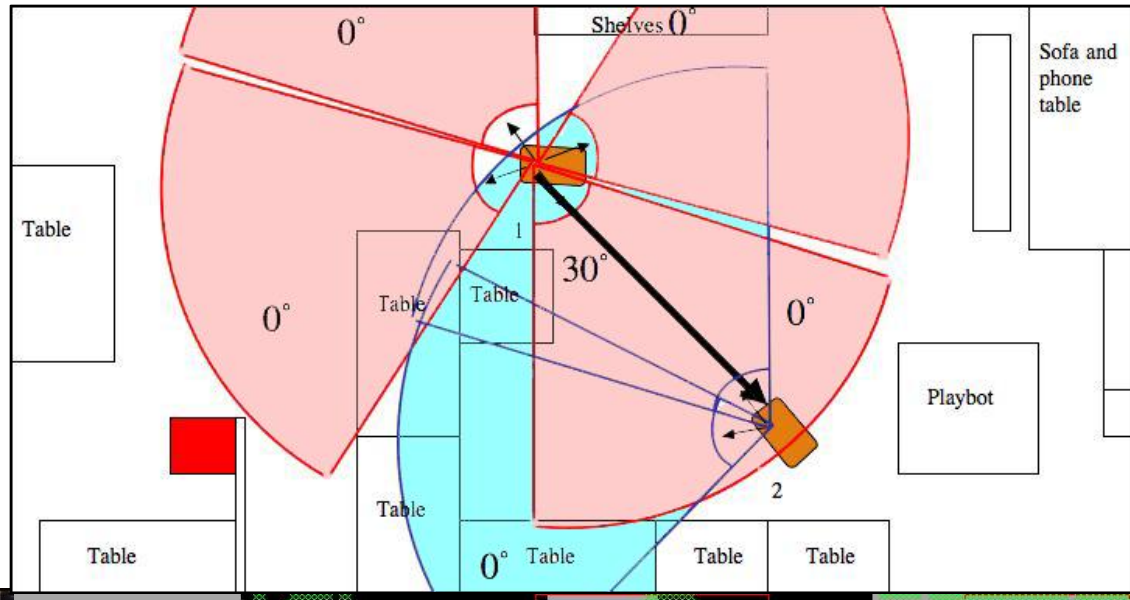


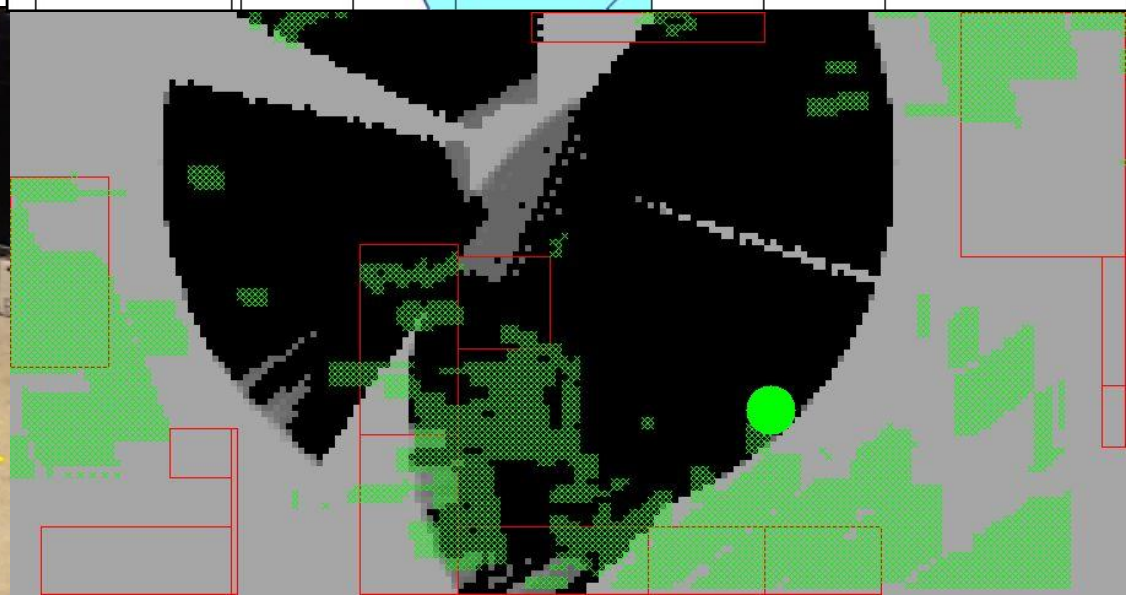
Position 2, Sensing Action 1



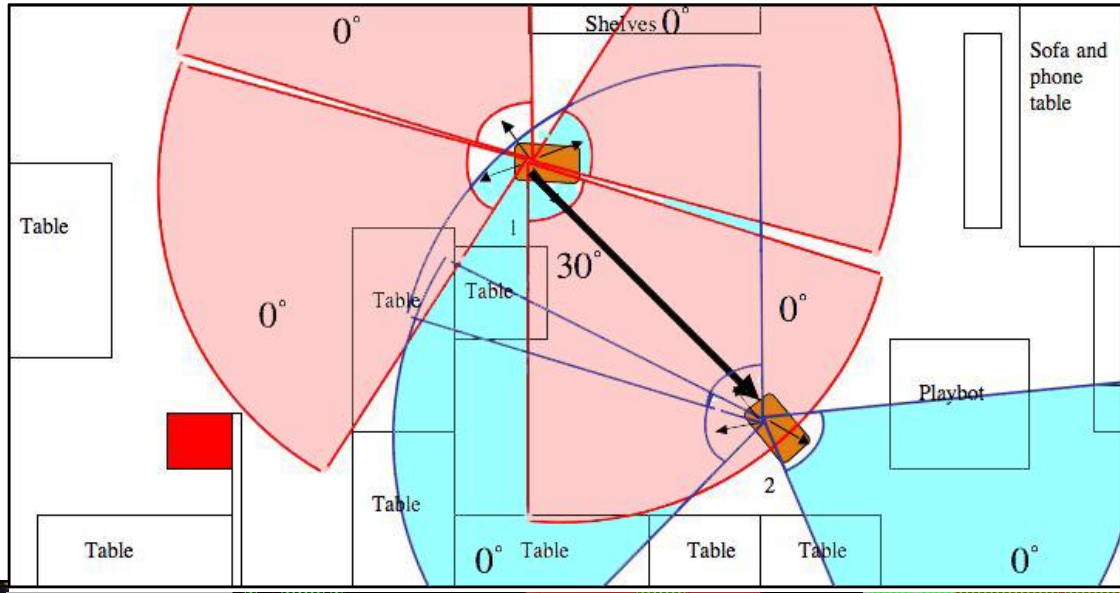
Position 2, Sensing Action 2

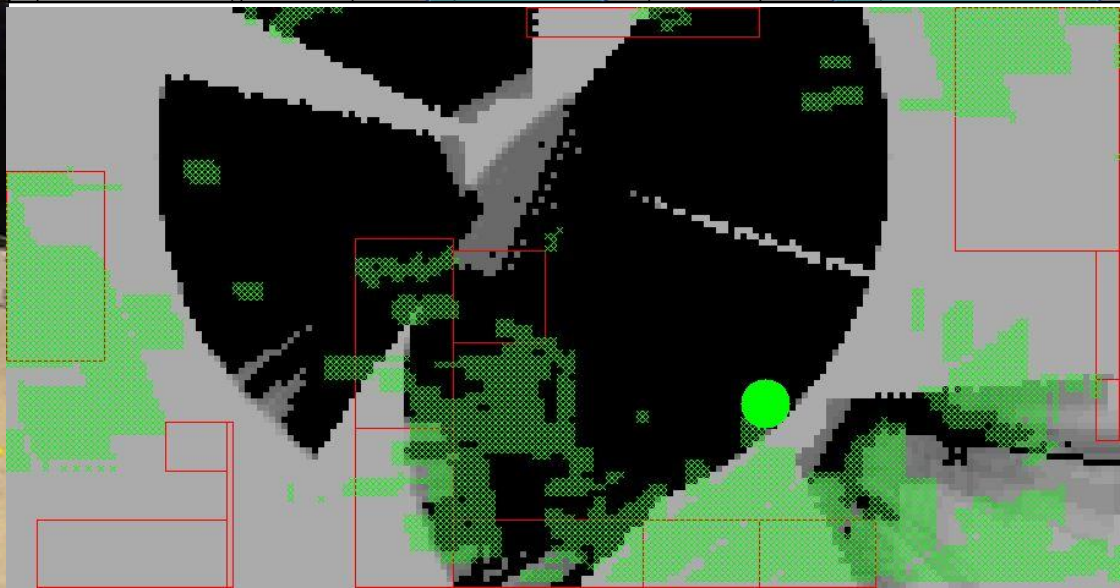




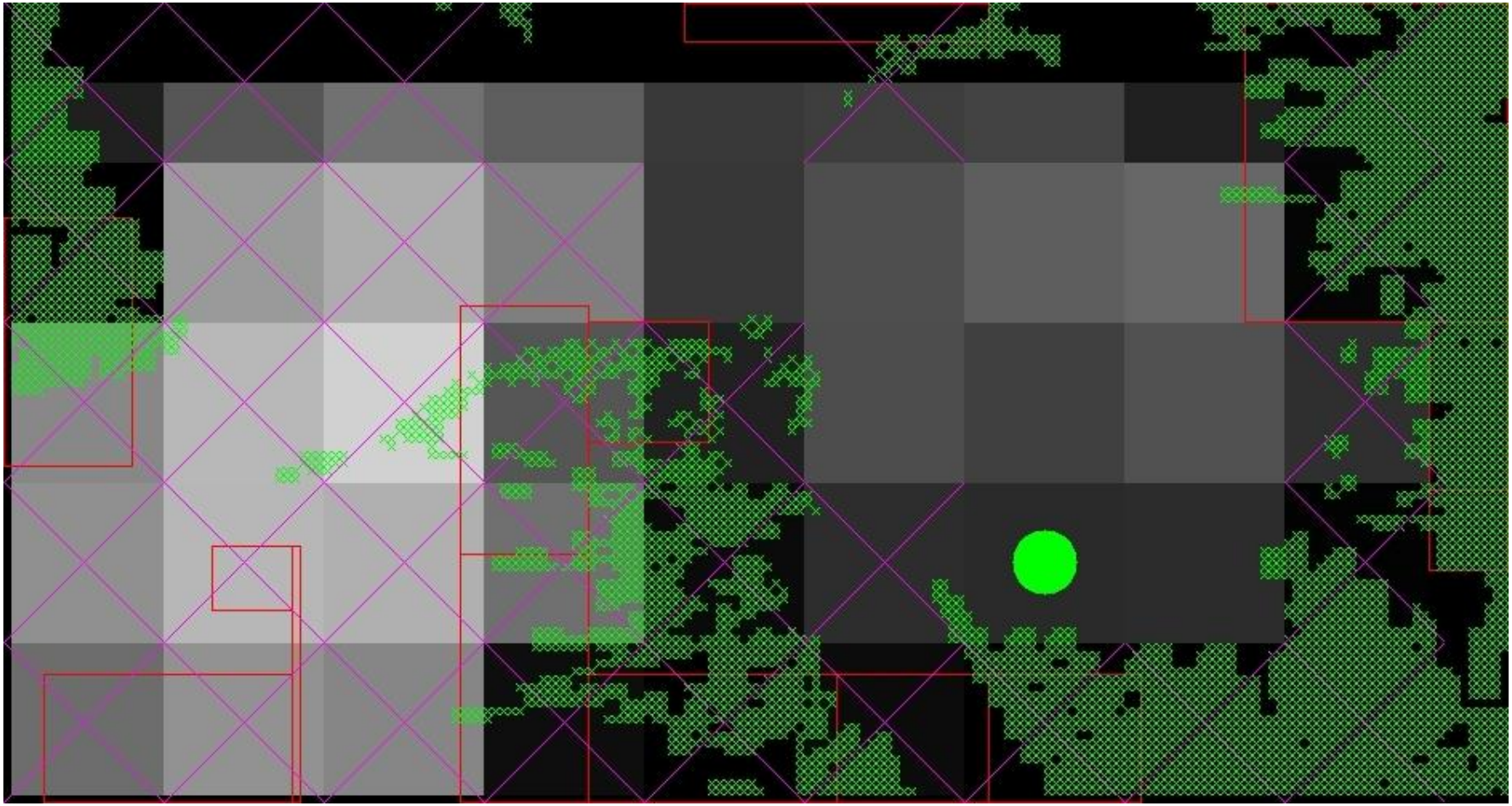


Position 2, Sensing Action 3

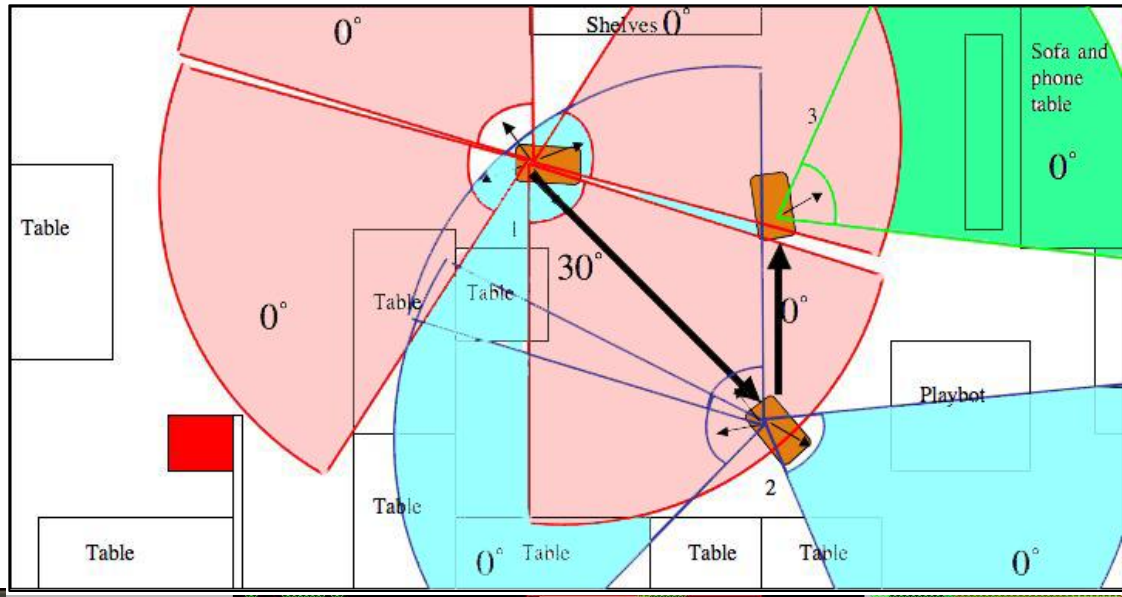


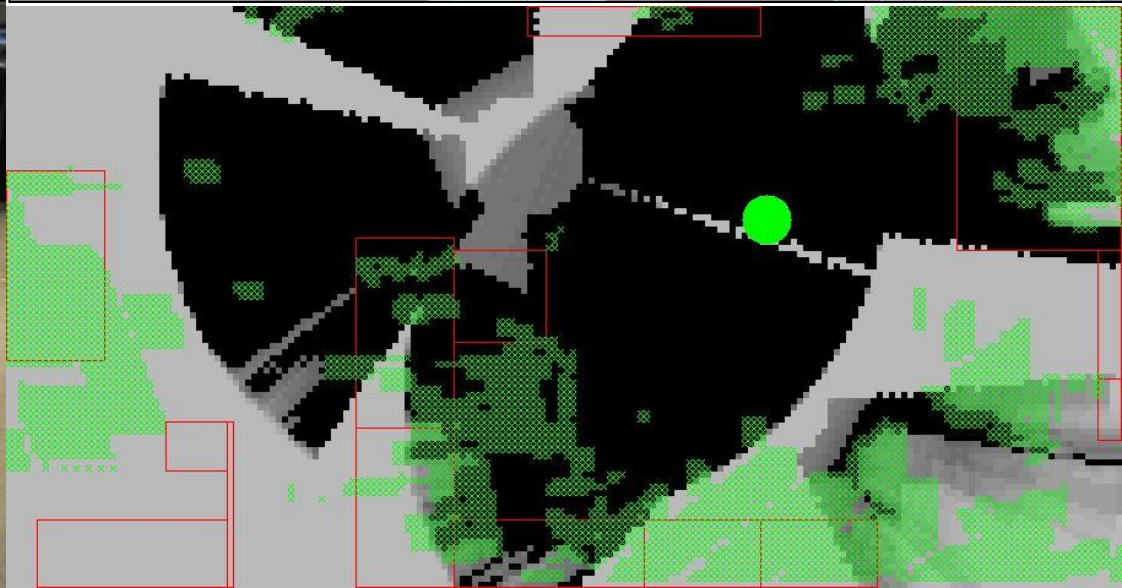


Map of where to move next after Position 2

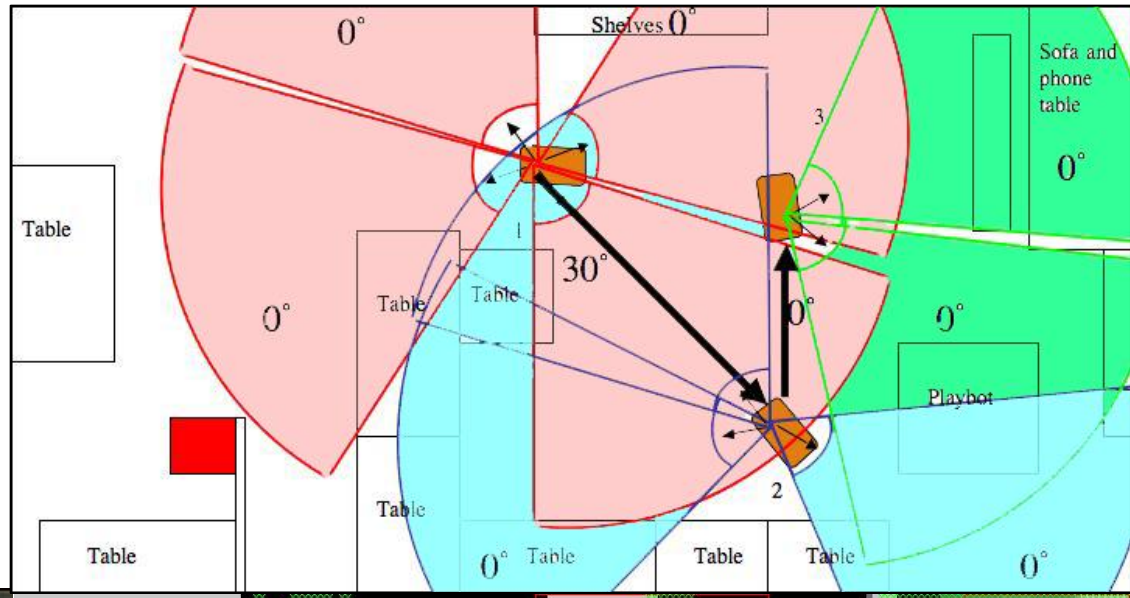


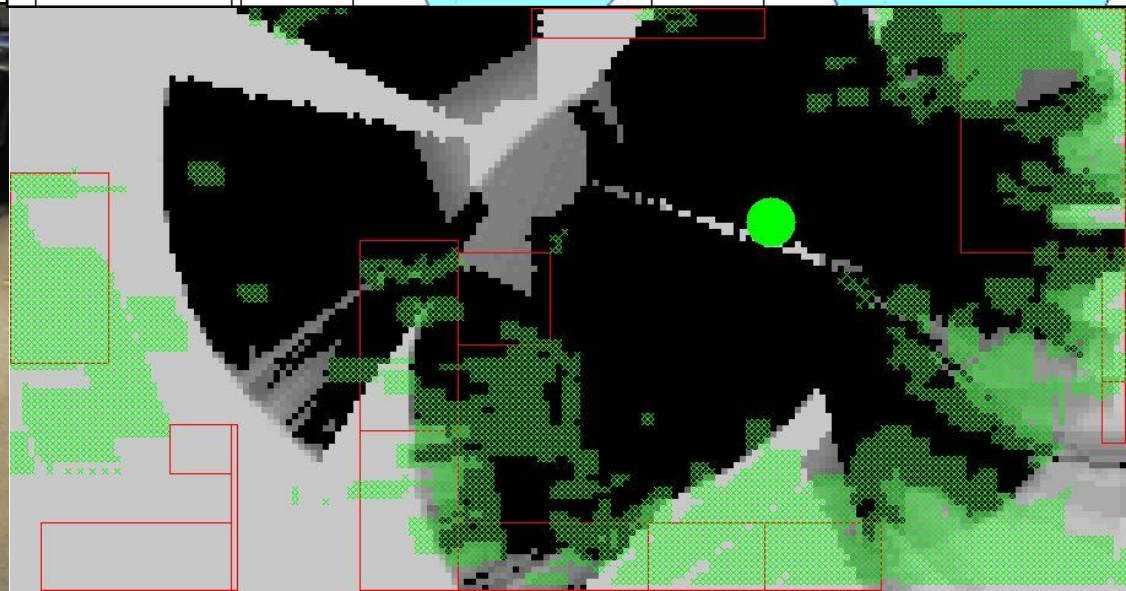
Position 3, Sensing Action 1



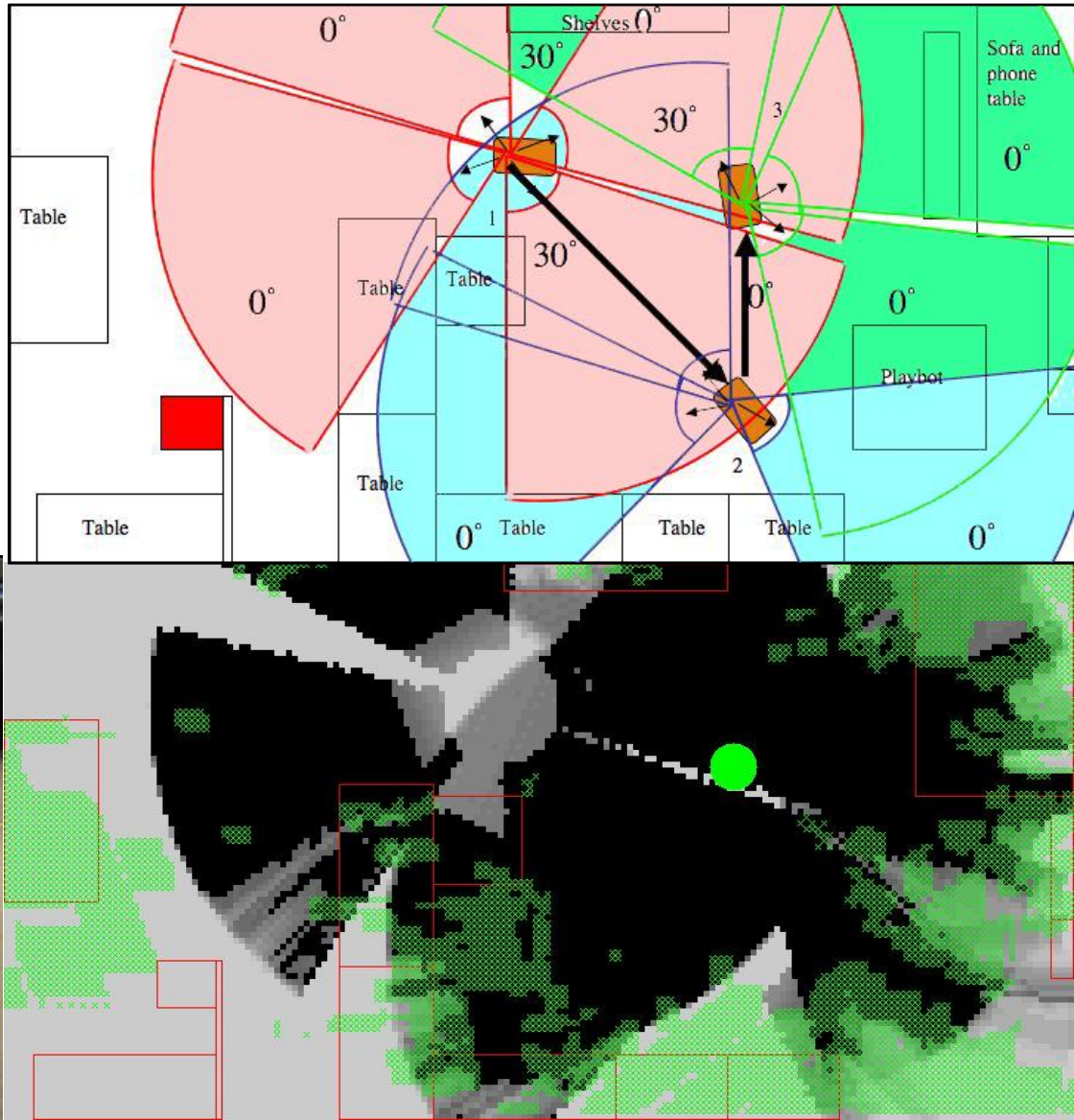


Position 3, Sensing Action 2

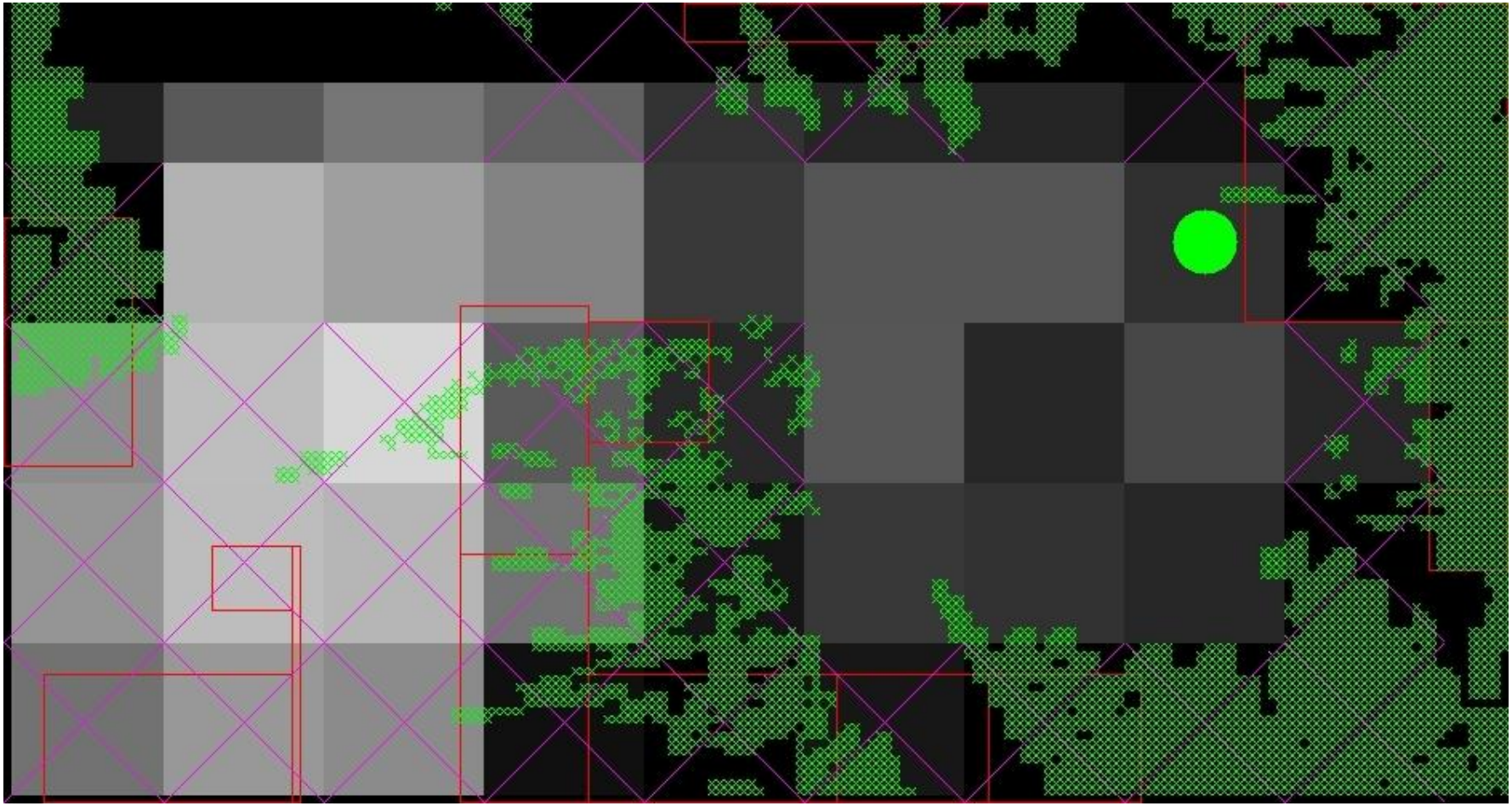




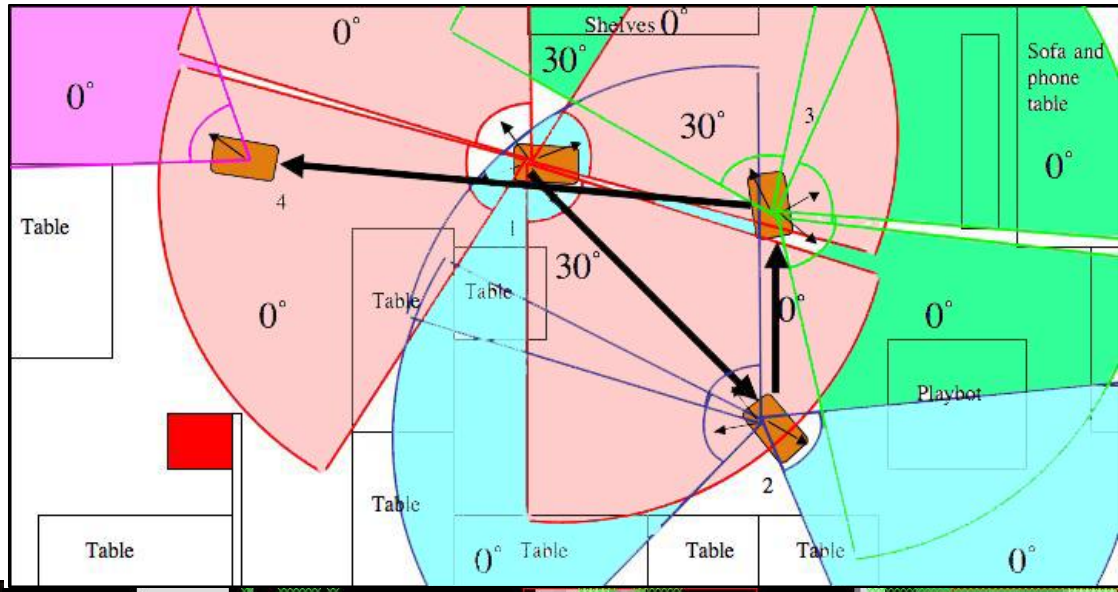
Position 3, Sensing Action 3

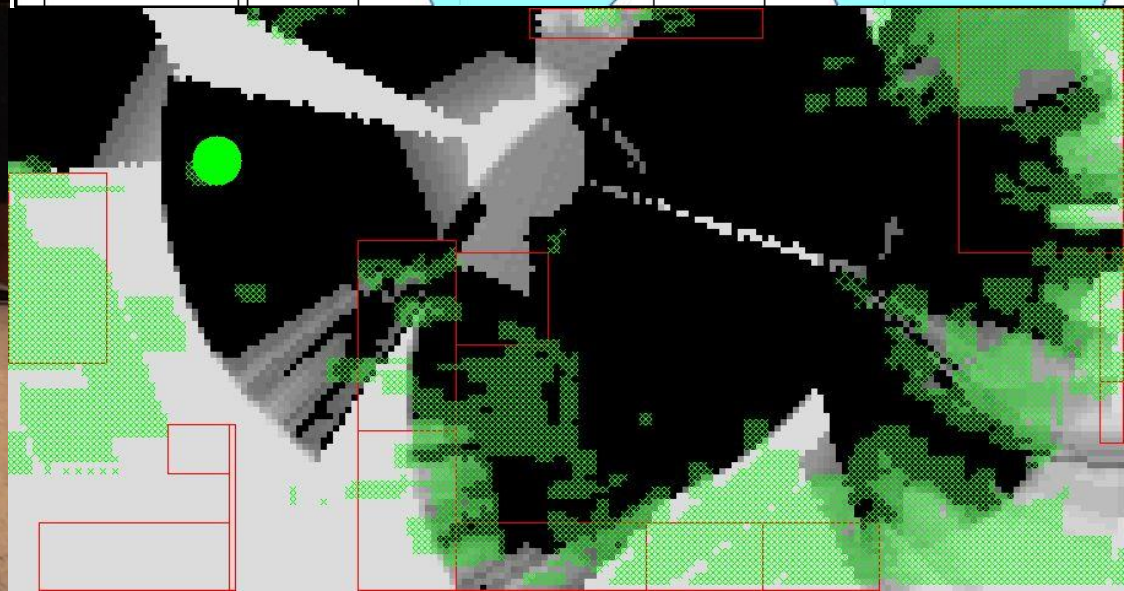


Map of where to move next after Position 3

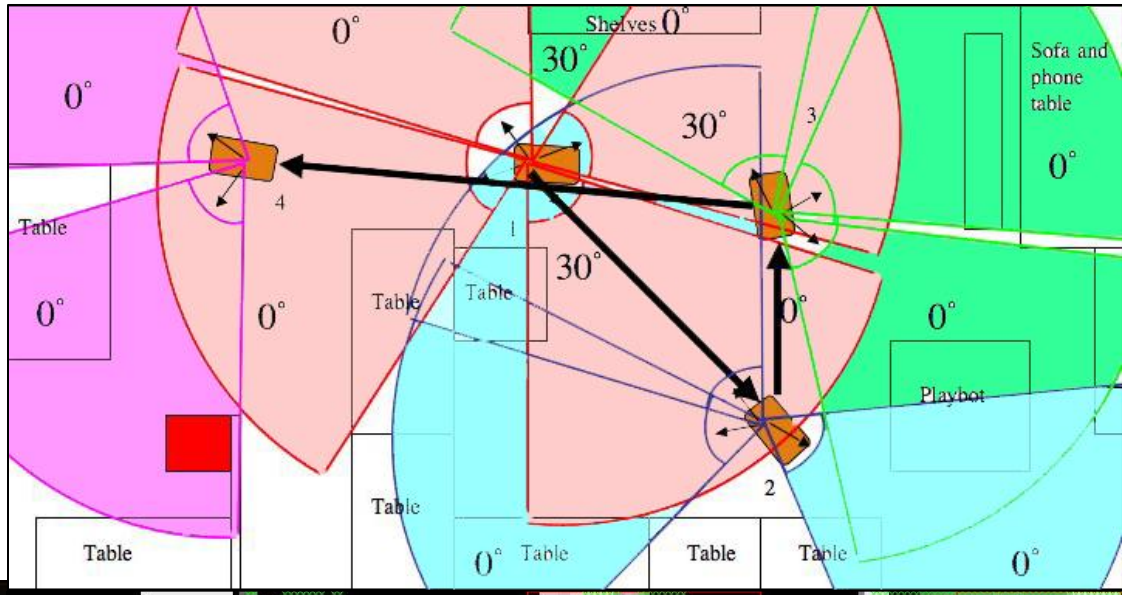


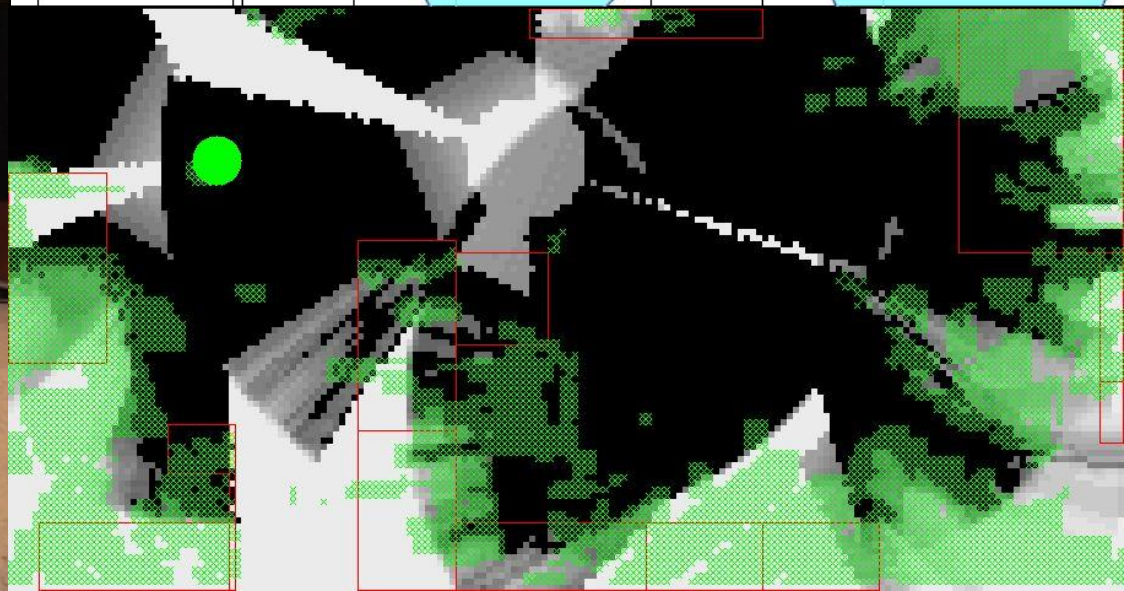
Position 4, Sensing Action 1



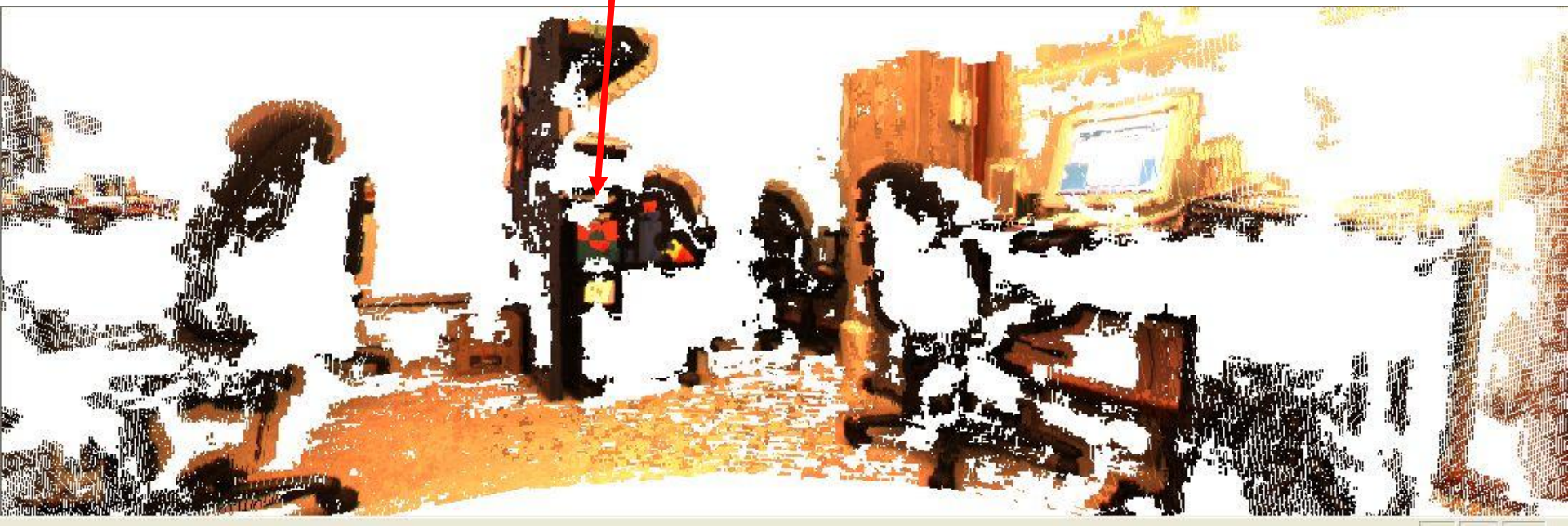


Position 4, Sensing Action 2





Sample Stereo Data



last viewpoint

Effect of Different Search Strategies

Are there differences in performance with different search strategies?

A. Choose the action (pan, tilt, x, y) with the largest detection probability.

$$\mathbf{f}_{\tau+} = \arg\max_{\mathbf{f}} \sum_{c_i \in \Psi_{\mathbf{f}}} \mathbf{p}(c_i, \tau_{\mathbf{f}})$$

B. Explore the current position first. Next position maximizes detection probability.

$$pos_{\tau+} = \arg\max_{pos} \sum_{c_i \in \Psi_{pos}} \mathbf{p}(c_i, \tau_{\mathbf{f}})$$

* the previous examples

C. Explore the current position first. Next position maximizes detection probability while minimizing the distance to the position.

$$pos_{\tau+} = \arg\max_{pos} \frac{\sum_{c_i \in \Psi_{pos}} \mathbf{p}(c_i, \tau_{\mathbf{f}})}{\mathbf{dist}(pos)}$$

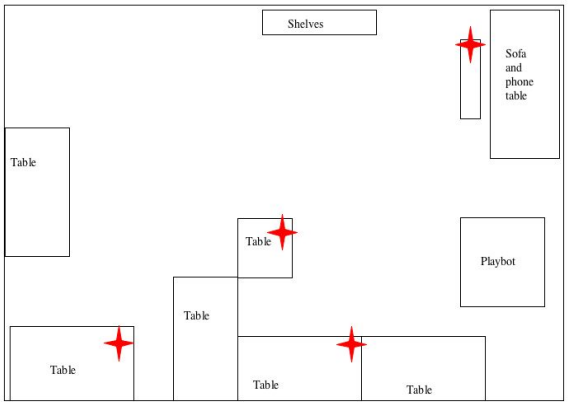
D. Explore the current position first. Next position maximizes the detection probability but relaxing distance minimization, controlled by Φ (Φ=1 used)

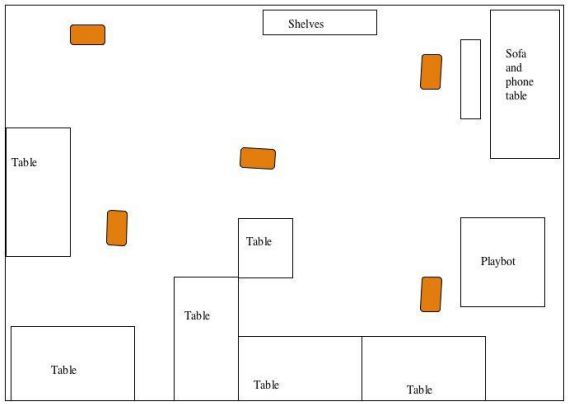
$$pos_{\tau+} = \arg\max_{pos} \sum_{c_i \in \Psi_{pos}} \mathbf{p}(c_i, \tau_{\mathbf{f}})\left(1 + \frac{\Phi}{\mathbf{dist}(pos)}\right)$$

The Experiment

- For each strategy, we did two types of experiments:
  - No prior knowledge
  - The search agent knows that the target object is on one of the tables and knows the location of the tables in the room.
- For each combination of prior knowledge and cost options, we ran 20 experiments.

4 target locations





5 robot starting locations

- Single target for all runs; detection function based on SIFT

Experimental Results: General

Total number of runs: 160

Total number of successful runs: 145

Causes for failures:

1. No localization (only has dead reckoning so it gets lost with respect to map)
2. Unreliable stereo (canned stereo algorithm sometimes sees imaginary obstacles and stops)

Experimental Results: Performance

No prior knowledge

Average per run (20 runs for each strategy)	*A*	*B*	*C*	*D*
Number of actions	**9.44**	**8.72**	**8.45**	**8.73**
Total time (min.)	**16.19**	**8.11**	**6.8**	**7.53**
Distance traveled (m.)	**22.39**	**8.77**	**3.85**	**8.03**

The target is on one of the tables

Average per run (20 runs for each strategy)	*A*	*B*	*C*	*D*
Number of actions	**5**	**5.5**	**4.9**	**4.93**
Total time (min.)	**10.44**	**5.9**	**3.8**	**4.33**
Distance traveled (m.)	**13.47**	**5.98**	**3.41**	**4.14**

Best Strategy:
Explore the current position first.
Choose the next position that maximizes the detection probability while minimizing the distance to the next position

Some knowledge is better than none
Need to minimize
time: C + knowledge
actions: A, C, D +knowledge
distance: C + knowledge

Doorway Behavior

Andreopoulos, A., Tsotsos, J.K., A Framework for Door Localization and Door Opening Using a Robotic Wheelchair for People Living with Mobility Impairments, RSS 2007 Manipulation Workshop: Sensing and Adapting to the Real World, Atlanta, Jun. 30, 2007.

on a humanoid robot

- collaboration with Honda Research Institute Europe
- the team: A. Andreopoulos, S. Hasler, H. Wersing, H.Janssen, J.Tsotsos, E. Körner

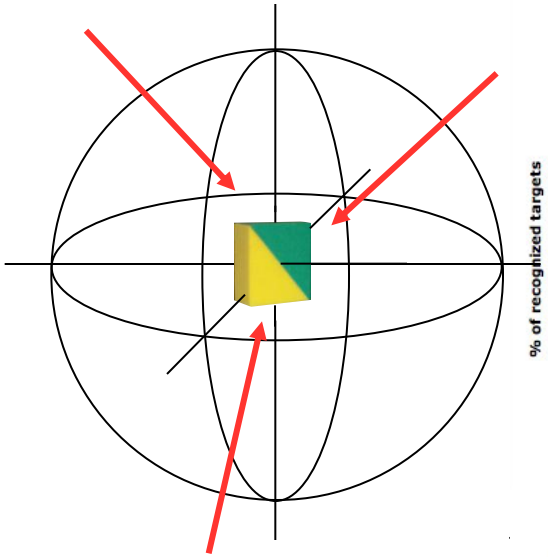
What would the Role of Learning be in Active Visual Search?

- ✧ Consider the elements of our algorithm:
  - where to look next
    - distribution of likely object positions (in scene, with respect to other objects – must be viewpoint invariant)
  - where to move next
    - distribution of likely paths
  - target detection
    - distribution of likely object poses
- ✧ Potentially Useful Elements outside our Algorithm:
  - most useful manipulations
  - acceptable variability in pose with respect to recognizer

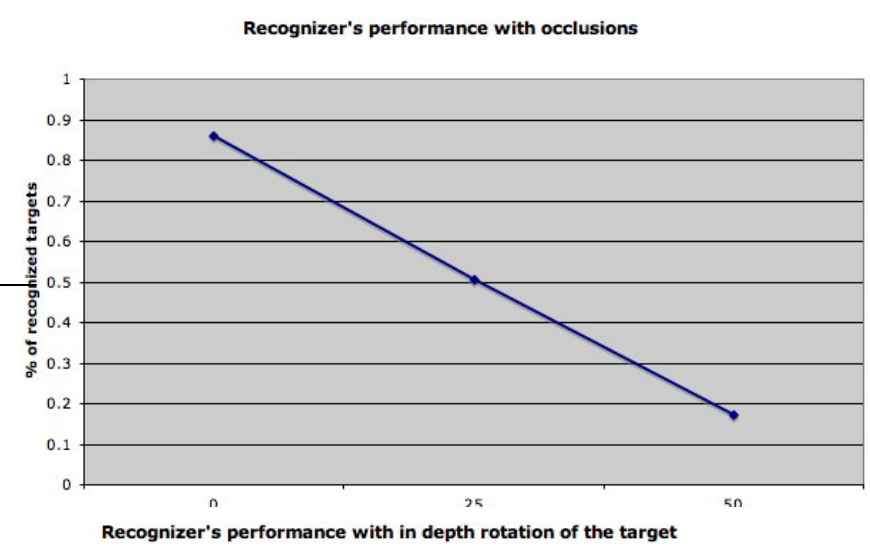
Object-Centred Detectability Function

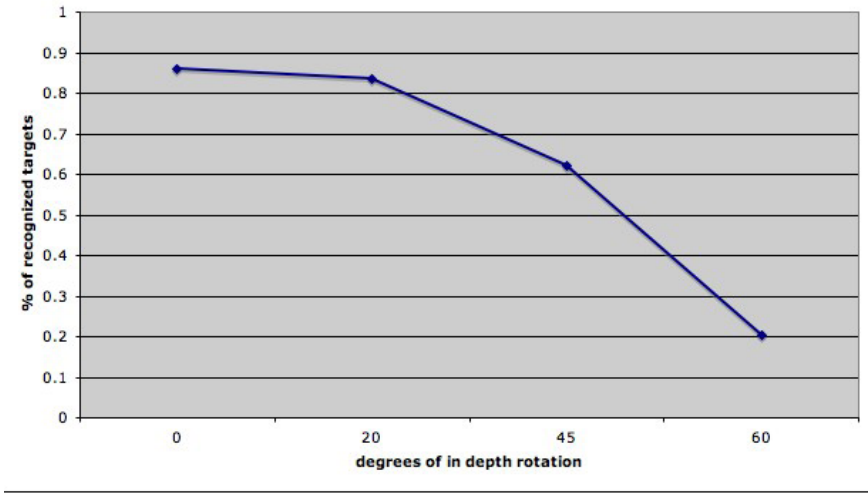
Detectability depends on:

- viewpoint
- image size
- distance
- scale
- 3D pose
- occlusion



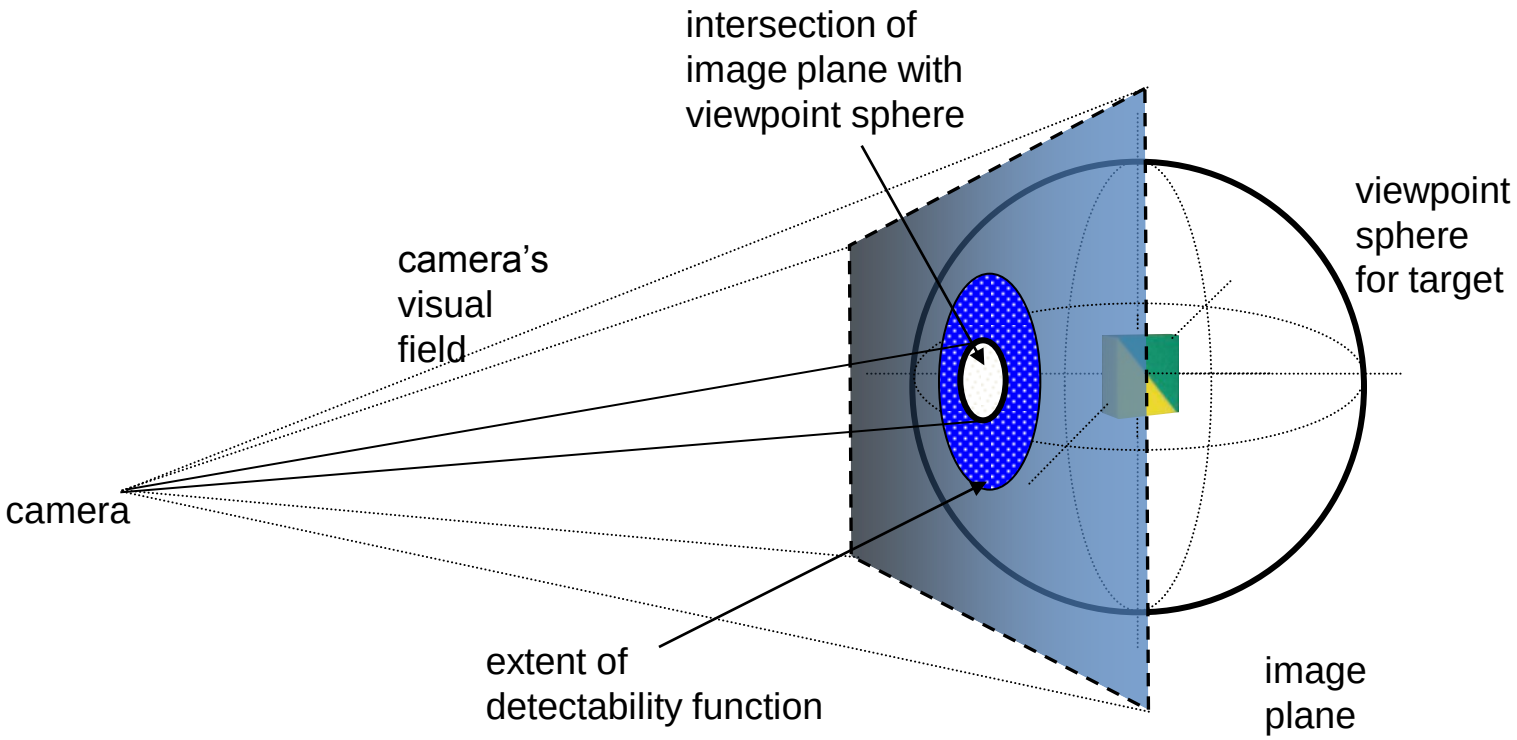
Determines positions/viewpoints from where the target can be recognized





Viewpoint-Dependent Visibility Function

Each fixation on a location at most tests a subset of viewpoints



A location is not fully ruled out until the full viewing sphere is inspected

(up to detectability function constraints for the target)

Task

- The importance of task and domain knowledge has been wildly under-considered

- Why? Marr (1982):

  The general trend in the computer vision community was to believe that recognition was so difficult that it required every possible kind of information. (p.35).....

  although some top-down information is sometimes used and necessary it is of only secondary importance .... evidence ... was willfully ignored by the computer vision community. (p. 100).

- Theoretical, empirical and brain evidence refutes this view: Tsotsos 1987, 1988, 1990, 1992, 1995; Parodi et al. 1998; Müller & Rabbit 1978; Posner et al. 1978+

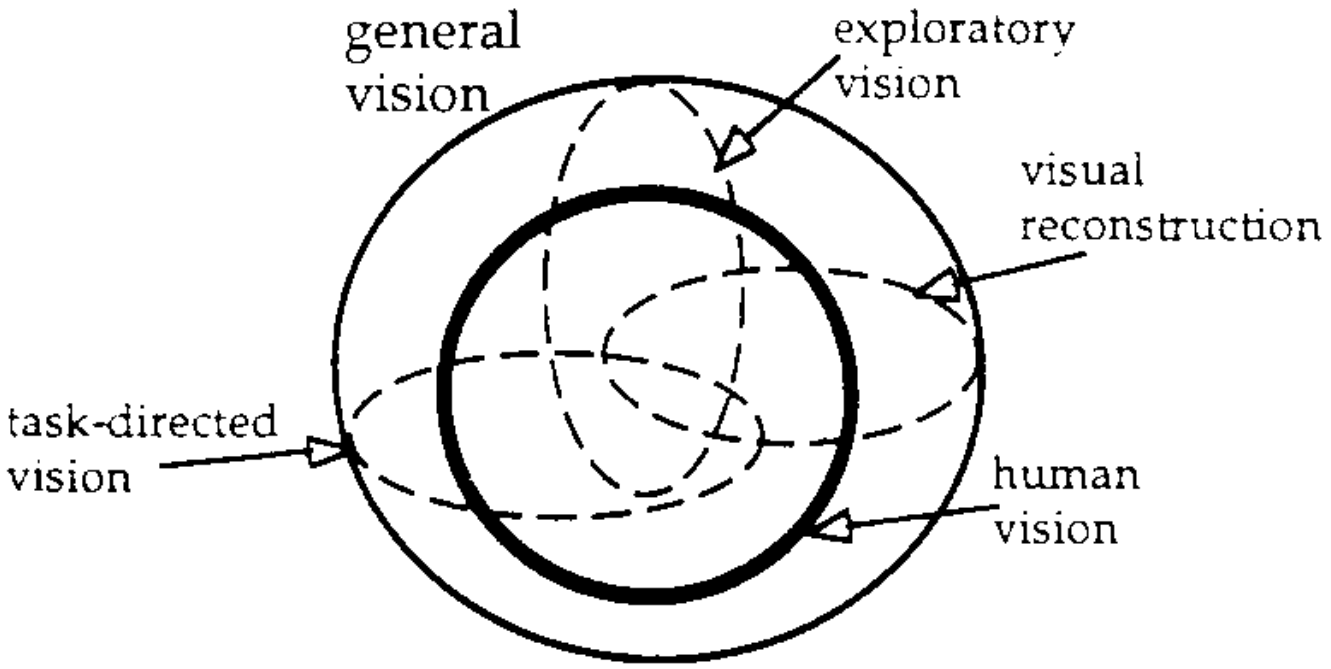
Two General Conclusions

- ✧ Attention is a set of mechanisms that help tune and control the search processes inherent in perception and cognition. Active vision is a subset of these.

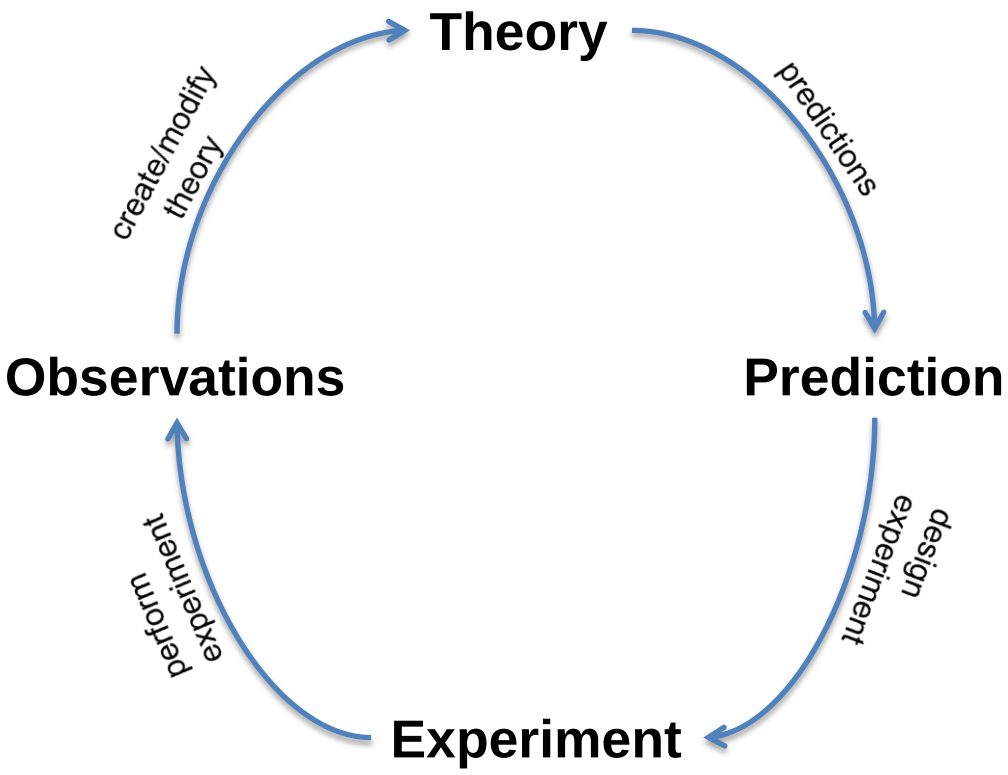
- ✧ Vision as Dynamic Tuning of a General Purpose Processor
  - – it can solve a particular class of vision problems very quickly
  - – it can be tuned dynamically to adapt its performance given the task and scene at hand for the remaining sub-classes of vision problems but at a cost of greater time to process.

Get back to trying to Understand Visually-Directed Behavior

- from Tsotsos, J.K., There is no one way to look at vision, CVGIP: Image Understanding 60:1, (Commentary on Tarr & Black), p320-322, 1994.



Trying to UNDERSTAND VISION is more fun than building a box with 1.37% better performance than the other guy...



for more....

Active Search for Pathognomonic Views

Wilkes, D., Tsotsos, J.K., Active Object Recognition, Proc. CVPR, 1992 , p136 - 141.

Wilkes, D., Active Object Recognition, February 1994 , PhD Dissertation, Dept. of Computer Science, University of Toronto.

Active Search for Non-Degenerate Views

Dickinson, S., Christensen, H., Tsotsos, J.K., Olofsson, G., Active object recognition integrating attention and viewpoint control, *CVIU* 67-3, p239 - 260, 1997.

Dickinson, S., Wilkes, D., Tsotsos, J.K., A computational model of view degeneracy, *IEEE PAMI* 21-8, p673 - 689, 1999.

3D Active Object Search

Ye, Y., Sensor Planning in 3D Object Search, January 1997, PhD Dissertation, Dept. of Computer Science, University of Toronto.

Ye, Y., Tsotsos, J.K., Sensor Planning for Object Search, *CVIU* 73-2, p145 - 168, 1999.

Ye, Y., Tsotsos, J.K., A Complexity Level Analysis of the Sensor Planning Task for Object Search, *Computational Intelligence* 17-4, p605 – 620, Nov. 2001.

Shubina, K., Sensor Planning for 3D Object Search, MSc Thesis, Dept. of Computer Science & Engineering, York University, Jan. 2007.

Tsotsos, J.K., Shubina, K., Attention and Visual Search: Active Robotic Vision Systems, Int. Conference on Computer Vision Systems, Bielefeld, Germany, March 21-24, 2007.

Shubina, K., Tsotsos, J.K. Visual Search for an Object in a 3D Environment using a Mobile Robot, *Computer Vision and Image Understanding*, 2010.