Active Vision

FOR

DUM..., er, DOCTORS IN THE MAKING

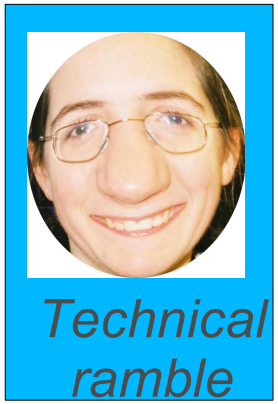
by Cathy Laporte

What is active vision ?

A typical active vision system:

- is built with a particular task in mind
- involves the acquisition of multiple visual measurements over time
- automatically fuses the task related evidence as it is acquired
- carefully selects the parameters of the next measurement

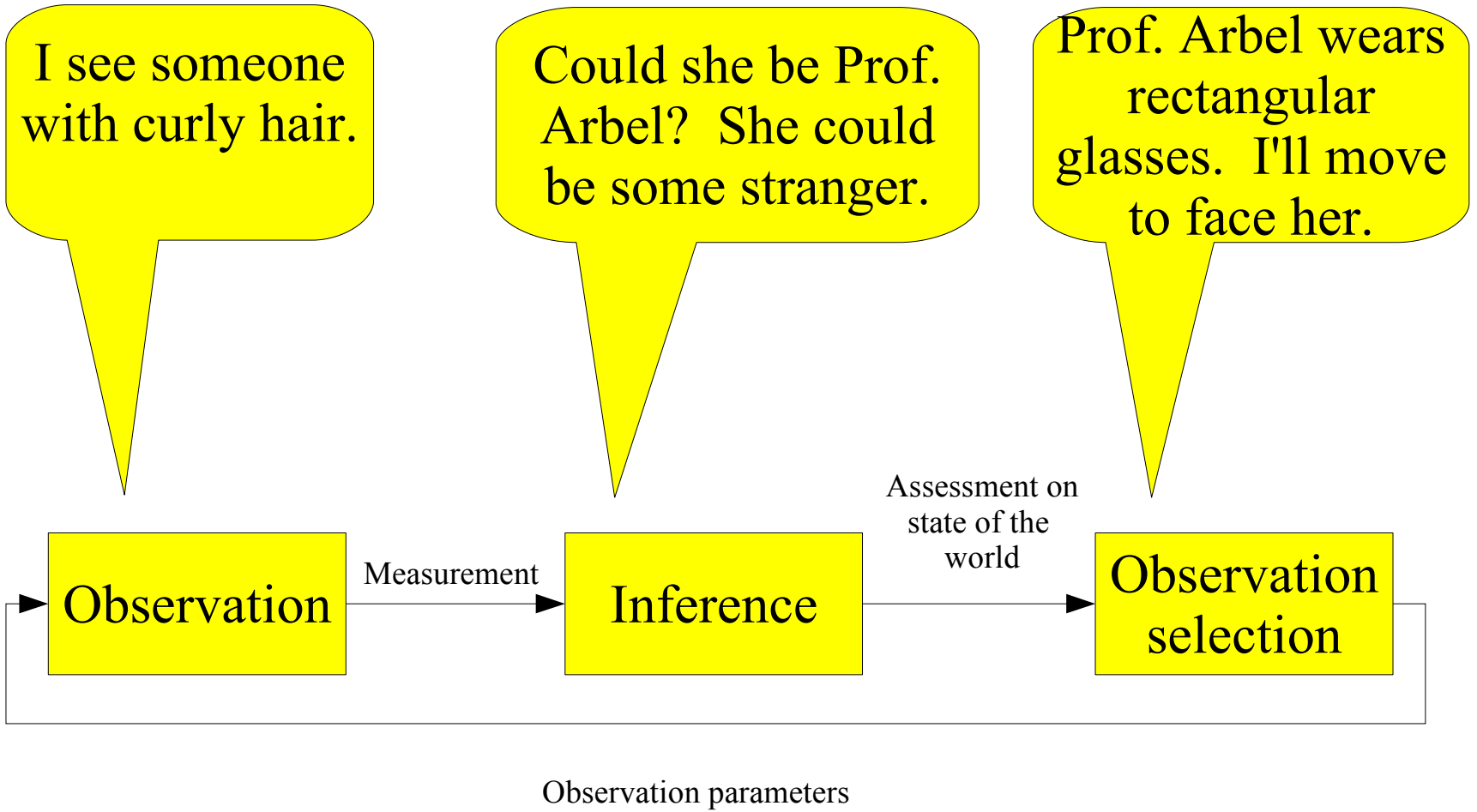
The last point is what differentiates an active vision system from a regular passive vision system.



Active vision v.s. active sensing

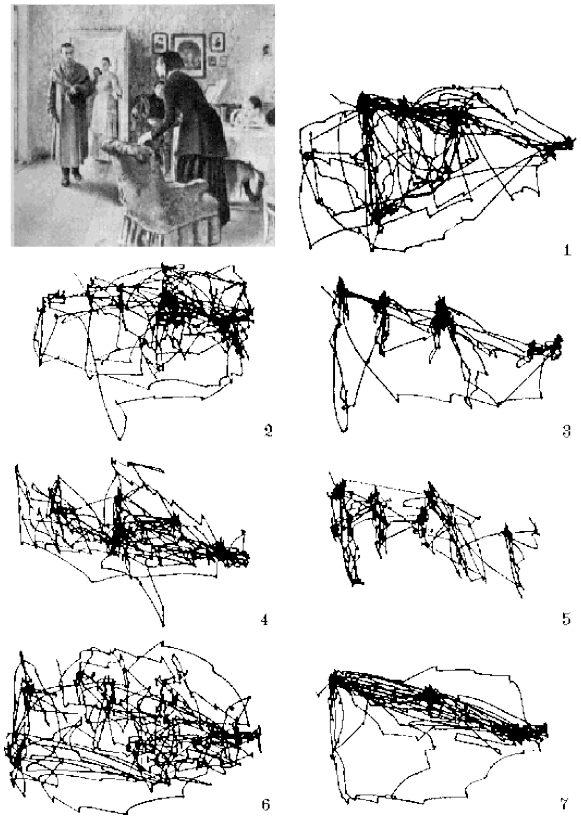
Active sensing refers to using an active sensor, a device which emits its own stimuli, like a laser range finder. It is very different from *active vision*, although active vision systems sometimes use active sensors.

The active vision loop



We do active vision, too!

Our brains do active vision at more or less conscious levels, by picking out which of the millions of stimuli in the world we focus our attention on (example from Yarbus, 1967).



1. Free examination
2. Estimate material circumstances of family
3. Give the ages of people
4. Hypothesise what family was doing before arrival of visitor
5. Remember clothes worn by people
6. Remember positions of objects and people in the room
7. Estimate how long the visitor had been away from the family

Biological roots of active vision

Manifestations of active vision behaviour:

- Pupil contraction/dilation depending on lighting conditions
- Unconscious attention shifts (micro-saccades)
- Conscious attention shifts (eye and head movements)
- Wearing different kinds of glasses for reading than for driving
- Etc.

Why active computer vision?

Active vision can be used to tune acquisition parameters (e.g. lighting, viewpoint, zoom) to optimise the performance of a given computer vision algorithm.

Example task: person recognition with an active vision system to tune the camera's zoom

picture from http://www.cim.mcgill.ca/~benoits/ICPR_2004_Cambridge/index.html

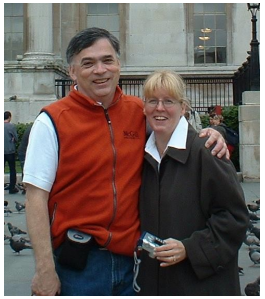
Chosen zoom setting depends on recognition algorithm

Algorithm 1: eigenfaces

Choose zoom so that faces are at the same scale as training images.

Algorithm 2: silhouette matching

Choose zoom so that the whole silhouette is in the image while occupying most of space.

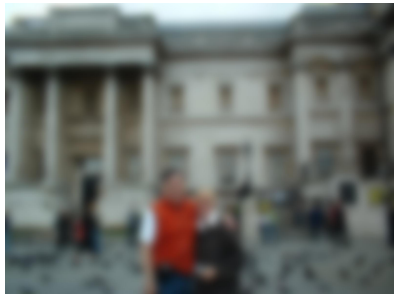


Why active computer vision?

Active vision can be used to tune the parameters of a general purpose algorithm to the requirements of a given task, thereby making its own trade-off adjustments.

Example algorithm: Gaussian smoothing with an active vision system to choose the standard deviation of the filter.

Task 1: People tracking
Choose standard deviation large to smooth out small moving objects.

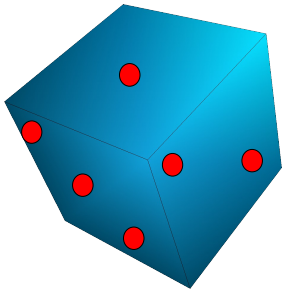


Task 2: bird counting
Choose standard deviation small to reduce noise but not birds.

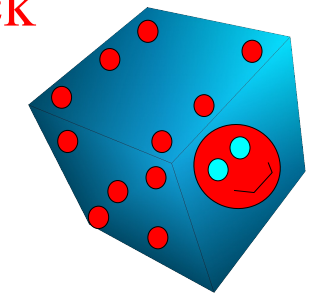
Why active computer vision?

Active vision allows selective processing of data where it matters most, avoiding resource consuming computations on irrelevant or non-informative data.

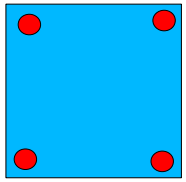
First observation:
Active vision is used to select the viewpoint for further observations

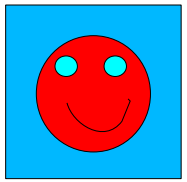


Task 1: object modeling
Best strategy is to pick which gives fullest coverage:

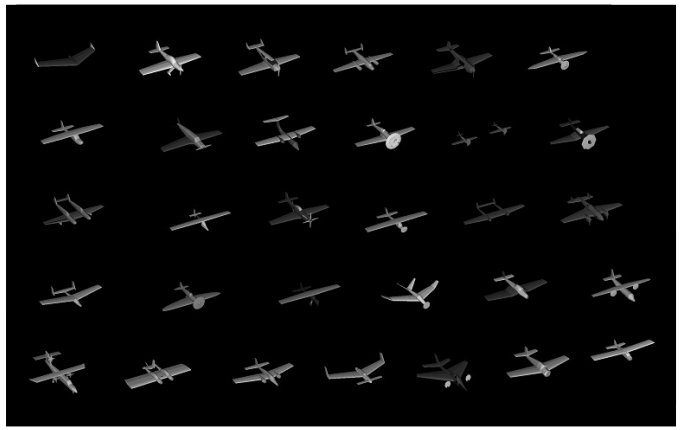


Task 2: object recognition
Best strategy is to pick the most distinctive view:





Case study: object recognition

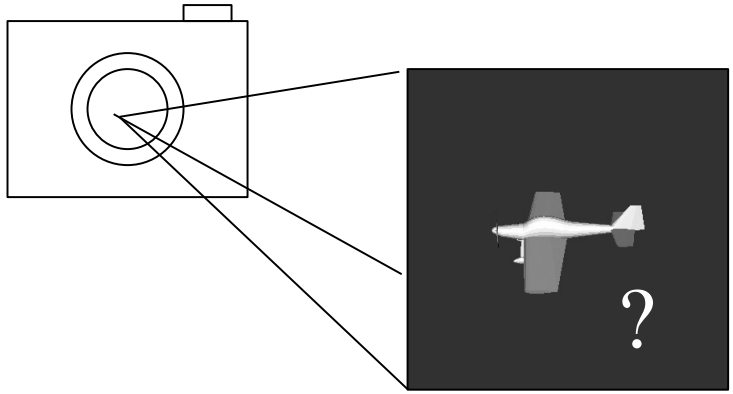


Models from http://www.rccad.com/Gallery Classic8.htm.

determine the identity of an unknown object belonging to this database, based on sensory measurement vector d.

The Problem

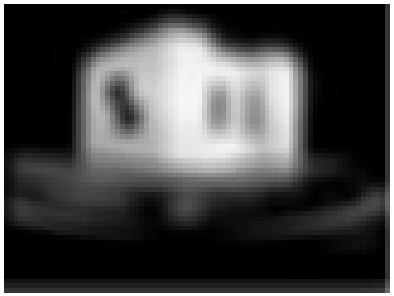
Given a set of objects $\{o_k\}$, $k = 1, \ldots, K$ described by some appearance model,

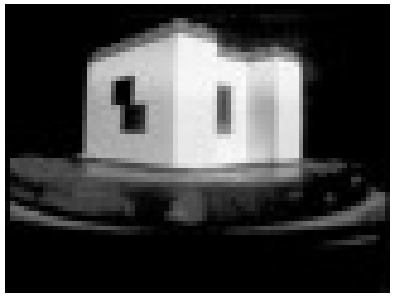


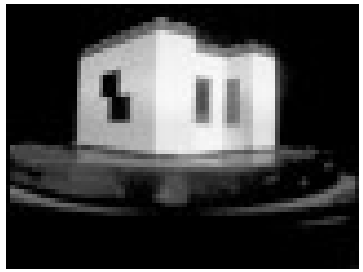
Why is this difficult?

The measurement process is subject to uncertainty

If I measure this, does it come from this? ... or that?

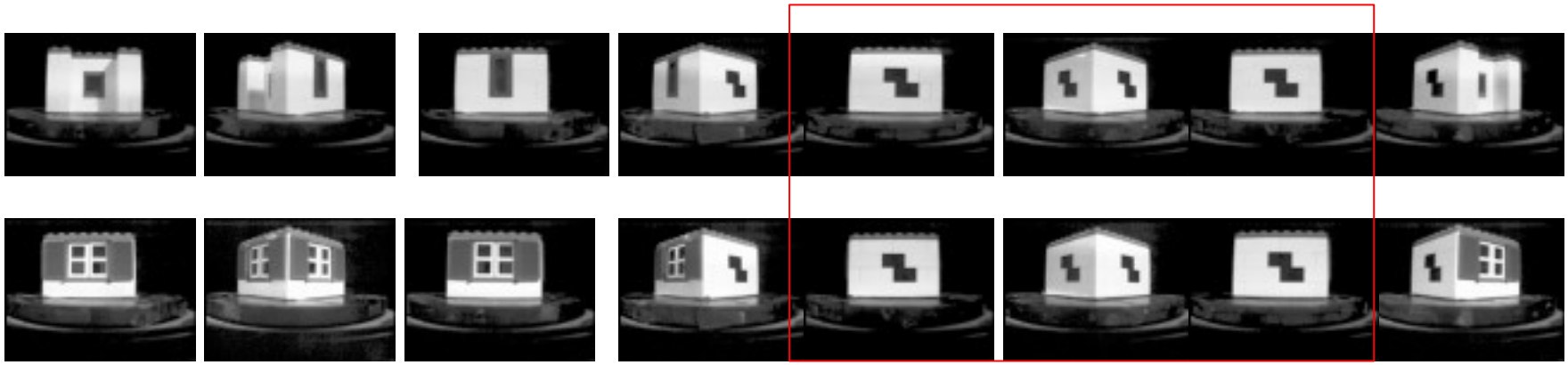






But it gets worse...

Different objects may look similar from certain points of view



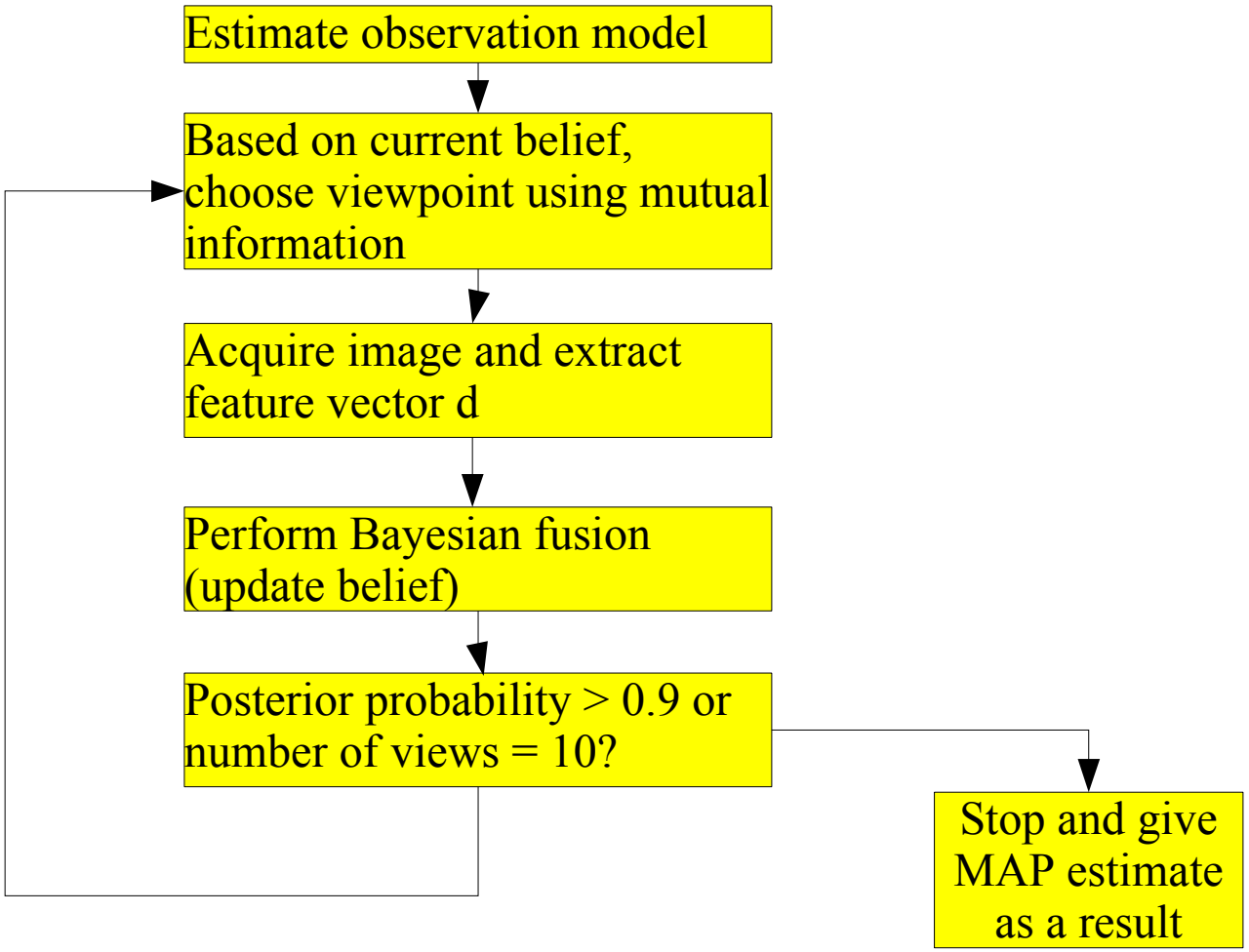
The active vision solution

- Combine measurements from several points of view instead of only one
- Carefully *choose* measurements such that they are useful to distinguish competing hypotheses

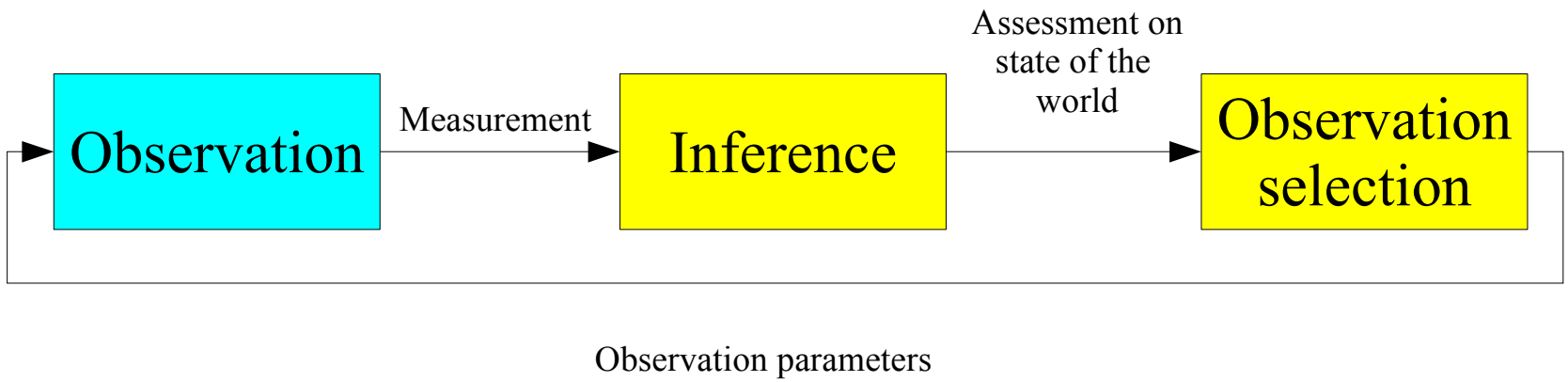
There are many ways to do this. Here, we will look at the particular approach proposed by

J. Denzler and C. M. Brown (2002), "Information theoretic sensor data selection for active object recognition and state estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:2, 145-157.

Denzler & Brown's approach



Modeling in active vision



The variables

The state vector x :
A vector representing the hidden state of the system (at least part of which we are trying to estimate). For the object recognition problem, x = o, the unknown label of the observed object.

The observation vector d:
A vector of data acquired by our sensor and feature extraction algorithms. Denzler & Brown use PCA features in their object recognition paper.

The control parameter vector a:
A vector describing the action leading to the current observation. For Denzler & Brown's object recognition, this is the (discretised) camera viewpoint.

The prior model

The **prior model** $p(x_1)$ describes our belief in initial state of the world.

In the object recognition example, this is usually an uninformative prior distribution, i.e. a discrete uniform distribution over all possible object labels.

The observation model

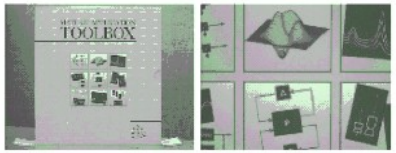
The **observation model** p(d|x, a)

• describes the relationship between observations and the state of the world depending on control parameters.

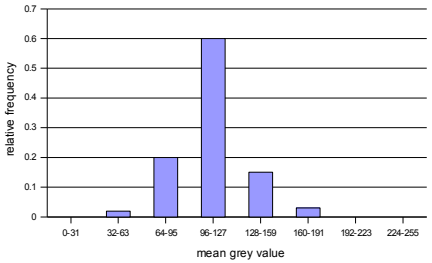
• can be derived from physics or from labeled training data

• can be discrete or continuous (Denzler & Brown look at both)

Observation model based on discrete features



A discrete feature: mean grey value of image, quantised to 8 possible bins.



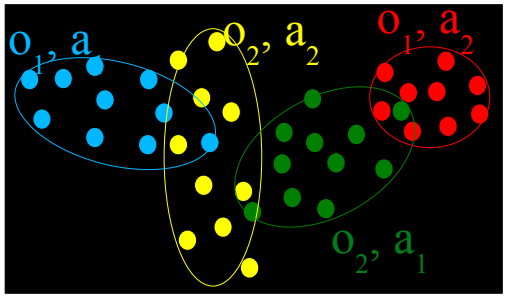
For an object o seen with zoom parameter a, p(d|o, a) takes the form of a discrete histogram, derived from mean grey values in training data.

The observation model (ctd)

Observation model based on continuous features

A low-dimensional continuous feature space can be derived with PCA.

For a certain object o and viewpoint a, a Gaussian p.d.f. can be fitted to training data in eigenspace and used as p(d|o, a).



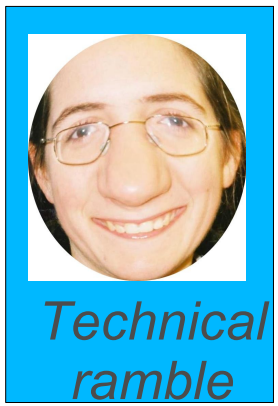
Continuous control parameters

Denzler & Brown's work assumes, somewhat artificially, that the control (viewpoint) parameters are discrete. How could we build an observation model with continuous a?

The state transition model

The **state transition model** $p(x_t | x_{t-1}, a_t)$ describes how the world changes based on its previous state and the control parameters.

In object recognition, the hidden state (object label) does not change. The state transition model takes the trivial form $p(o_t | o_{t-1}, a_t) = \delta(o)$.

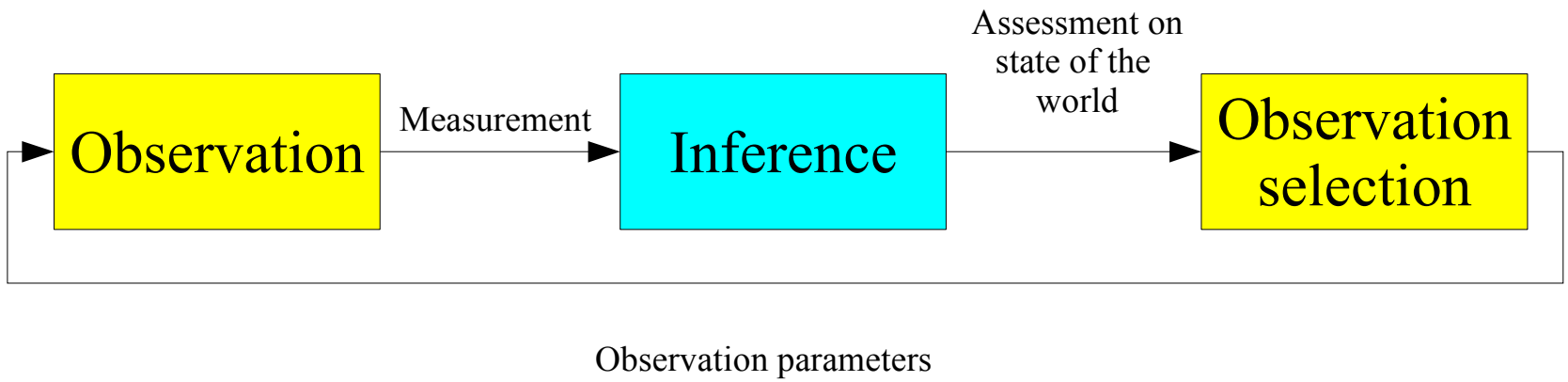


The Markov assumption

Problems like tracking require a non-trivial $p(x_t | x_{t-1}, a_t)$.

To keep things simple, a *Markovian* model is often chosen, meaning that x_t is completely independent of older states given x_{t-1} and a_t.

Data fusion in active vision



One observation

Problem: estimate current state x_1 from observation d_1 acquired using control parameters a_1.

Solution: from Bayes' rule, $p(x_1|d_1, a_1) \propto p(d_1|x_1, q_1)\, p(x_1)$

Nuisance variables

In object recognition, we might know about the effect of (not directly observable) lighting conditions on object appearance. How can we exploit this when estimating object identity?

Two (or more) observations

Problem: estimate $p(x_2|d_2,a_2,d_1,a_1)$

Solution for object recognition problem: Recall that $x_1 = x_2 = o$.

By Bayes' rule, $p(o|d_2,a_2,d_1,a_1) \propto p(d_2,d_1|o,a_1,a_2)p(o)$

Assuming that observations are independent given o and a,

$p(o|d_2,a_2,d_1,a_1) \propto p(d_2|o,a_1,a_2)\,p(d_1|o,a_1,a_2)\,p(o)$

Observation at time t only depends on control parameters at time t, so

$p(o|d_2,a_2,d_1,a_1) \propto p(d_2|o,a_2)\,p(d_1|o,a_1)\,p(o)$

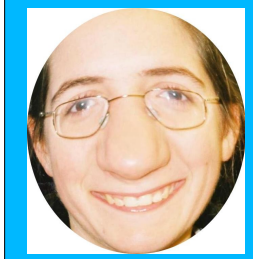
Rather conveniently, this gives us $p(o|d_2,a_2,d_1,a_1) \propto p(d_2|o,a_2)\,p(o|d_1,a_1)$

Generally, $p(o|d_1,\ldots,d_t,a_1,\ldots a_t) \propto p(d_t|o,a_t)\,p(o|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1})$

Data fusion in the general case

For problems with a dynamic state, the solution is more complicated:

$$p(x_t|d_1,\ldots,d_t,a_1,\ldots,a_t) \propto \int p(d_t|x_t,a_t)\,p(x_t|x_{t-1},a_t)\,p(x_{t-1}|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1})\,dx_{t-1}$$



Technical ramble

In general, data fusion is easier said/written than done

Solution: use approximate methods

- Kalman filters (Gaussian approximation)
- Particle filters
- MCMC sampling

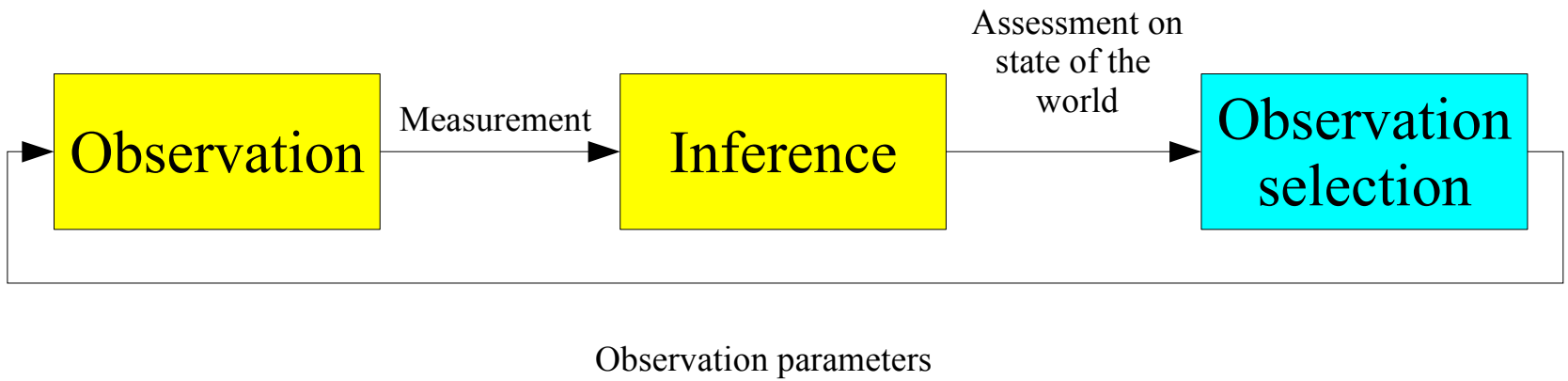
Food for thought

Deriving the general inference formula

Try deriving it from first principles (hint: use marginalisation). What assumptions are needed?

Also: derive the object recognition case from the general formula.

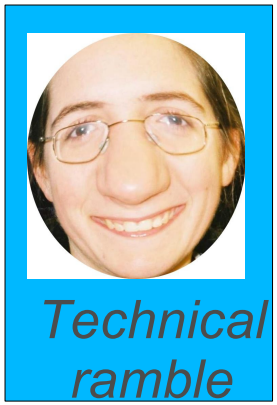
Decision making in active vision



Decision making in active vision

Goal: choose a_t to maximise a task related utility function U(x, d, a).

Things to consider in utility function:
- Informativeness of observation
- Danger involved in acquiring observation
- Computational requirements of processing observation
- Computational requirements of calculating cost function



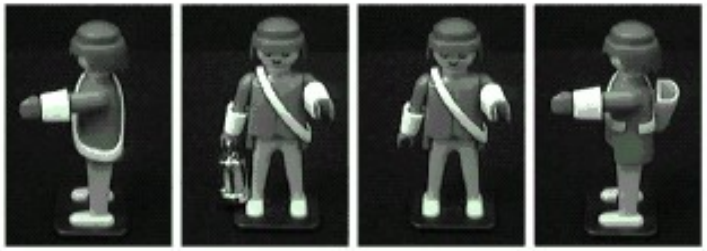
Uncertain decision making

Since we don't know x or d in advance, U(x,d,a) must be maximised over possible outcomes. For example,

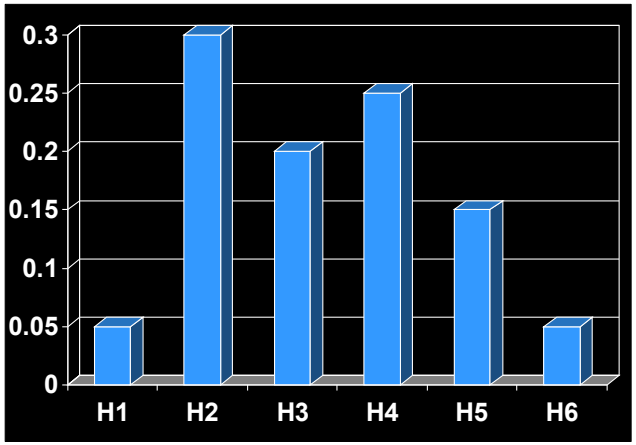
1. $a_t^* = arg\,max(E(U(x_t, d_t, a_t)))$ (most common)
2. $a_t^* = arg\,max(min(U(x_t, d_t, a_t)))$ (adverse environment)

What is a useful observation?

Object recognition example:

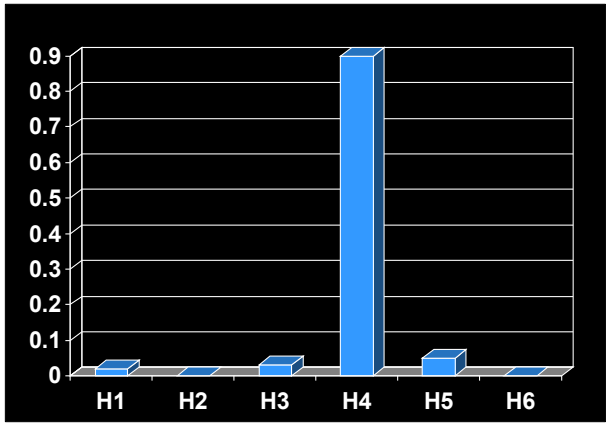


Prior distribution



Most useful observation

Potential posterior distribution



Mutual information criterion

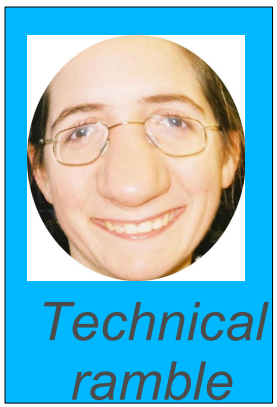
Idea: select a_t which is most informative about x on average.

Equivalently, select the observation that most reduces the entropy in the current distribution over x:

$$a_t^* = arg\,max\, H(x_t | d_{1,..}, d_{t-1}, a_{1,..}, a_{t-1}) - H(x_t | d_t, a_t)$$

The quantity to maximise is thus the mutual information:

$$a_t^* = arg\,max\, I(x_t; d_t | a_t)$$



Myopic v.s. global optimisation

The criterion presented here is *myopic* as it only maximises utility one step ahead. A *global* criterion would consider all possible *sequences* of future observations, solving a Partially Observable Markov Decision Problem (POMDP).

MI and active object recognition

Based on the commutativity of mutual information,

$$I(o;d_t|a_t)=H(o|a_t)-H(o|d_t,a_t)=H(d_t|a_t)-H(d_t|o,a_t)$$

Using the definition of conditional entropy,

$$I(o;d_t|a_t)=-\int_{d_t} p(d_t|a_t)\log p(d_t|a_t)\,dd_t$$
$$+\sum_{k=1}^{K}\int_{d_t} p(o_k|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1})\,p(d_t|o_k,a_t)\log p(d_t|o_k,a_t)\,dd_t$$

Applying Bayes' rule for marginalisation and combining relevant terms,

$$I(o;d_t|a_t)=\sum_{k=1}^{K}\int_{d_t} p(o_k|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1})\,p(d_t|o_k,a_t)\log\left(\frac{p(d_t|o_k,a_t)}{p(d_t|a_t)}\right)dd_t$$

Calculating mutual information

For continuous features,

$$I(o;d_t|a_t)=\sum_{k=1}^{K}\int_{d_t} p(o_k|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1})\,p(d_t|o_k,a_t)\log\left(\frac{p(d_t|o_k,a_t)}{p(d_t|a_t)}\right)dd_t$$

- The integral is impossible to compute analytically
- Numerical quadrature is intractable for 4 or so dimensions in d

Monte Carlo integration

If f(x) = g(x)p(x), where p(x) is a probability density function, then

$$F=\int_{-\infty}^{\infty} f(x)=\int_{-\infty}^{\infty} g(x)\,p(x)\,dx=E(g(x))$$

The integral can thus be approximated by drawing samples from p(x), and computing their average:

$$F=E(g(x))\approx\frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

Monte Carlo evaluation of MI

Denzler & Brown re-express the mutual information

$$I(o;d_t|a_t)=\sum_{k=1}^{K}\int_{d_t} p(o_k|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1})\,p(d_t|o_k,a_t)\log\left(\frac{p(d_t|o_k,a_t)}{p(d_t|a_t)}\right)dd_t$$

as expected values for Monte Carlo integration:

$$I(o;d_t|a_t)=E_{o|d_1,\ldots,d_{t-1},a_1,\ldots,a_{t-1}}\left(E_{d_t|o,a_t}\left(\log\left(\frac{p(d_t|o,a_t)}{p(d_t|a_t)}\right)\right)\right)$$

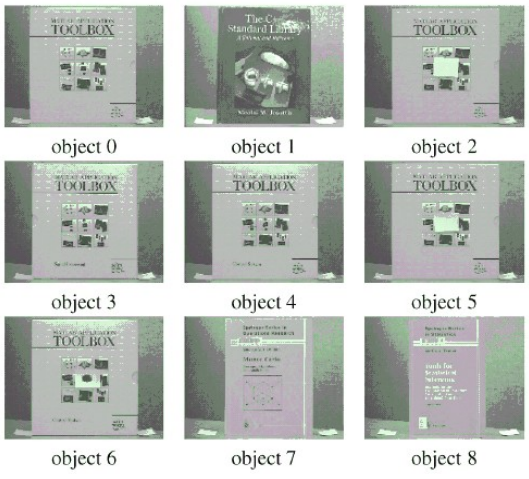
The approximation uses samples from p(d|o, a) and the posterior.

Just how nasty is mutual information?

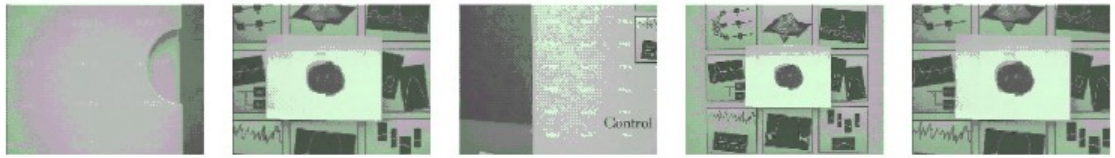
- p(d|o, a) is known from training
- p(o|d_1, ..., d_{t-1}, a_1, ..., a_{t-1}) is known from inference
- what about p(d_t|a_t)?

Active recognition example

Object database



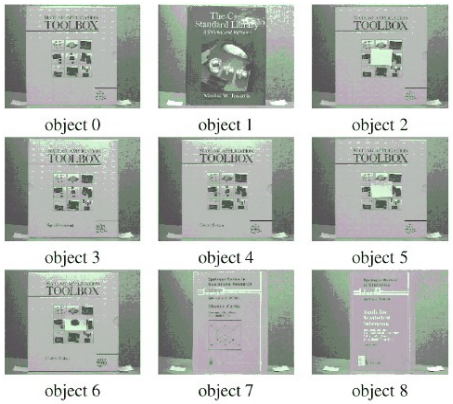
Observations of object no. 6 chosen by Denzler & Brown's algorithm



Evolution of posterior and entropy over time

view	o0	o1	o2	o3	o4	o5	o6	o7	o8	entr.
initial	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.95
view 1	0.251	0.000	0.251	0.000	0.073	0.251	0.095	0.078	0.000	0.72
view 2	0.523	0.000	0.034	0.000	0.113	0.014	0.256	0.050	0.000	0.50
view 3	0.125	0.000	0.069	0.000	0.237	0.027	0.542	0.000	0.000	0.53
view 4	0.003	0.000	0.092	0.000	0.000	0.043	0.861	0.000	0.000	0.22
view 5	0.003	0.000	0.005	0.000	0.000	0.000	0.990	0.000	0.000	0.03

The benefits of active vision



Compared to a random view selection approach, active viewpoint selection

- increases correct recognition rate when a limit is imposed on the number of views
- speeds up the recognition process

object	planned gaze control			random gaze control		
	rec. rate	mean no. views	mean max. prob.	rec. rate	mean no. views	mean max. prob.
o0	99.5	2.4	0.96	83.4	9.9	0.61
o1	100.0	1.0	1.00	99.6	1.2	1.00
o2	100.0	4.0	0.95	62.4	9.8	0.65
o3	100.0	2.3	0.96	76.0	9.8	0.64
o4	100.0	4.0	0.95	66.6	10.0	0.56
o5	99.2	3.5	0.97	68.2	9.9	0.57
o6	99.6	2.8	0.96	76.7	9.7	0.63
o7	100.0	1.7	0.98	100.0	2.5	0.97
o8	100.0	1.1	1.00	100.0	2.4	0.97
average	99.8	2.5	0.97	81.4	7.2	0.73

The good, the bad and the ugly

Good things about active vision:

- The framework can be independent of the type of data/features used.
- Feature extraction need not be elaborate: data fusion will cope.
- Combining multiple observations is a form of quality control.

Bad things about active vision:

- Active vision systems are often difficult to implement/test.
- Good active vision feedback loops require that models be unbiased.

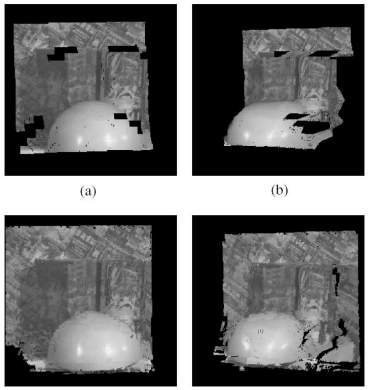
Ugly (well, challenging) questions about active vision

- While we're at it, what features are optimal?
- How to design a utility function that accounts for the cost of its (repeated) evaluation?
- How do we generalise our models for... ?

Other active vision applications

Scene reconstruction

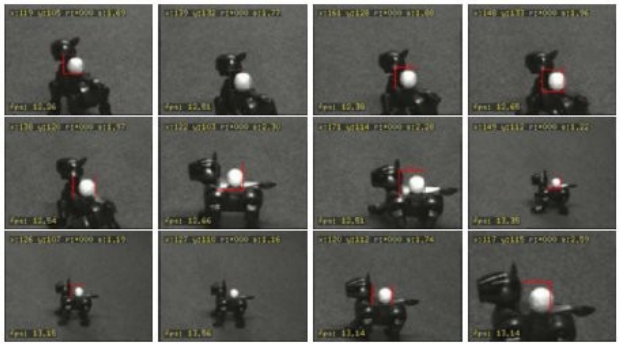
- Continuous state



Pictures from Lin et al., 2002.

Tracking

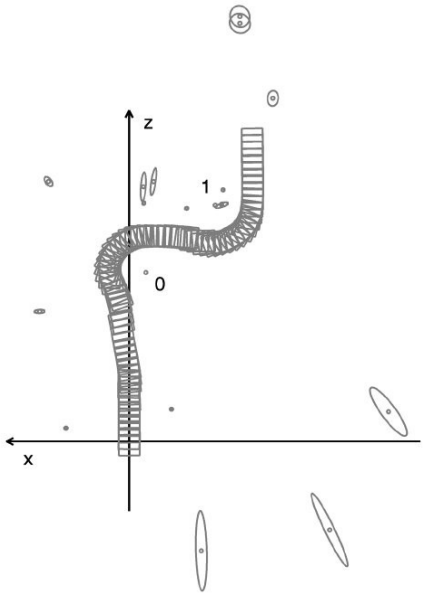
- Continuous state
- Dynamic state
- Possible adversity



Pictures from Denzler et al., 2003.

SLAM

- Continuous state
- Dynamic state
- Agent changes state



Picture from Davison and Murray, 2002.

Further reading

Introductory papers:

• J. Aloimonos, I. Weiss and A. Bandyopadhyay, “Active vision”, *International Journal of Computer Vision*, 1988.
• R. Bajcsy, “Active perception”, *Proceedings of the IEEE*, 1988.
• D. H. Ballard, “Animate vision”, *Artificial Intelligence*, 1991.

Attention and eye movements:

• J.K. Tsotsos, “Analyzing vision at the complexity level”, *Behavioral and Brain Sciences*, 1990.
• J.J. Clark, ``Spatial Attention and Latencies of Saccadic Eye Movements", *Vision Research*, 1999.

Active visual tracking

• J. Denzler, M. Zobel and H. Niemann, “Information theoretic focal length selection for real-time active 3D object tracking”, Proc. *ICCV*, 2003.
• A. J. Davison, “Active search for real time vision”, Proc. *ICCV*, 2005.

Further reading

Active object recognition

• S. A. Hutchinson and A. C. Kak, "Planning sensing strategies in a robot work cell with multi-sensor capabilities", *IEEE Trans. Robotics and Automation*, 1989.

• K. D. Gremban and K. Ikeuchi, "Planning multiple observations for object recognition", *International Journal of Computer Vision*, 1994.

• S. J. Dickinson, H. I. Christensen, J. K. Tsotsos and G. Olofsson, "Active object recognition integrating attention and viewpoint control", *Computer Vision and Image Understanding*, 1997.

• S. Kovačič, A. Leonardis and F. Pernuš, "Planning sequences of views for 3D object recognition and pose determination", *Pattern Recognition*, 1998.

• L. Paletta and A. Pinz, "Active object recognition by view integration and reinforcement learning", *Robotics and Autonomous Systems*, 2000.

• T. Arbel and F. P. Ferrie, "Entropy-based gaze planning", *Image and Vision Computing*, 2001.

• J. Denzler and C. M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002.

Further reading

Active scene reconstruction

• P. Whaite and F. P. Ferrie, “Autonomous exploration: driven by uncertainty”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997.

• E. Marchand and F. Chaumette, “An autonomous active vision system for complete and accurate 3D scene reconstruction”, *International Journal of Computer Vision*, 1999.

• C.-Y. Lin and S.-W. Shih and Y.-P. Hung and G. Y. Tang, “A new approach to automatic reconstruction of a 3-D world using active stereo vision”, *Computer Vision and Image Understanding*, 2002.

Other applications of active vision

• R. D. Rimey and C. M. Brown, “Control of selective perception using Bayes nets and decision theory”, *International Journal of Computer Vision*, 1994.

• H. Buxton and S. Gong, “Visual surveillance in a dynamic and uncertain world”, *Artificial Intelligence*, 1995.

• A. J. Davison and D. W. Murray, “Simultaneous localization and map-building using active vision”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002.