

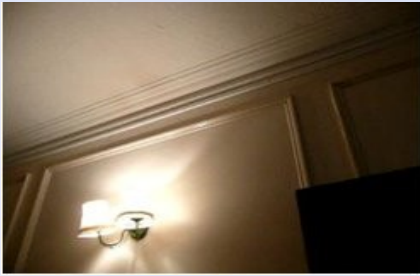
Learning textron models for real-time scene context

Alex Flint, Ian Reid, David Murray

Active Vision Laboratory

Oxford University

Motivation



Motivation



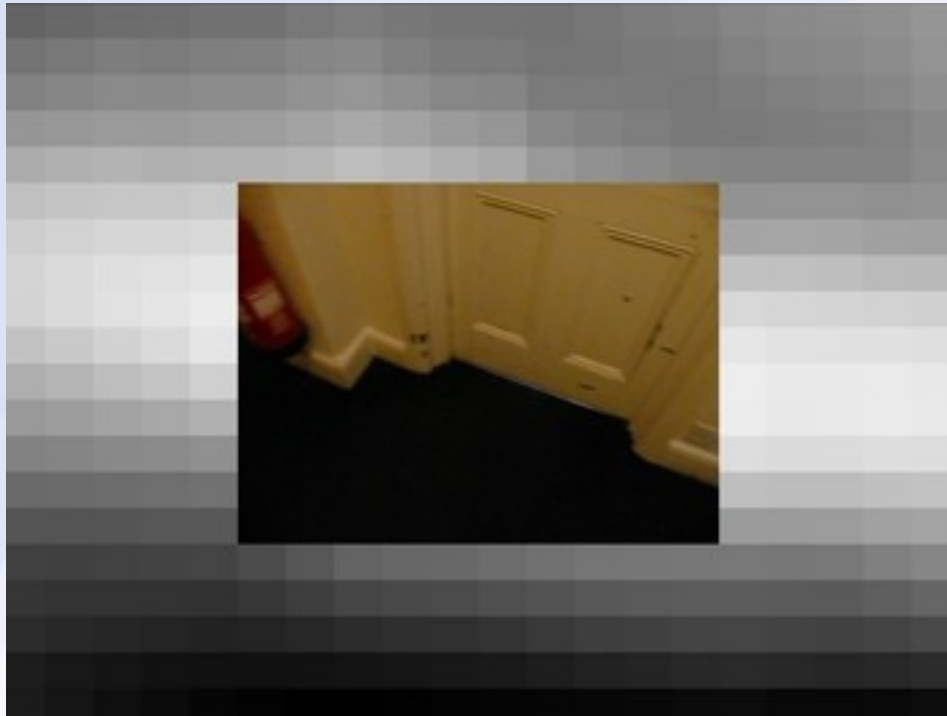
What type of room are we in?

What direction are we looking in?

Where should we look to find object X?

Motivation

Example: in which direction should we look for a fire extinguisher?

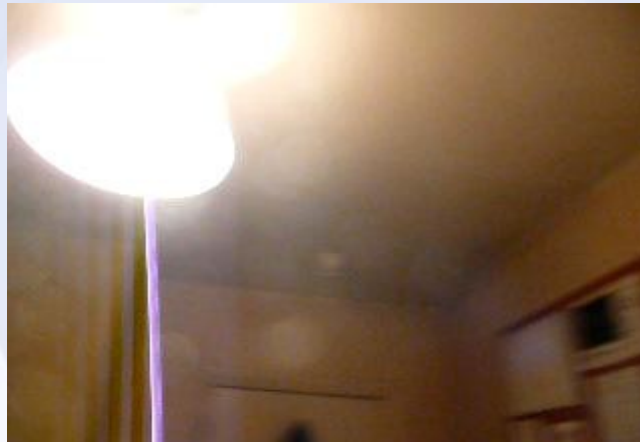


Motivation

Context for ego-centric vision

Use global image evidence to interpret local image features

Particularly important when data quality is low and difficult to interpret locally

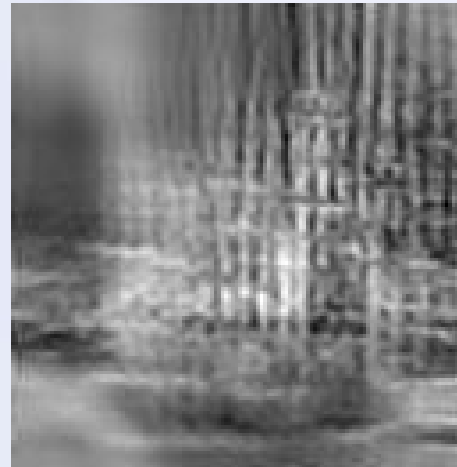
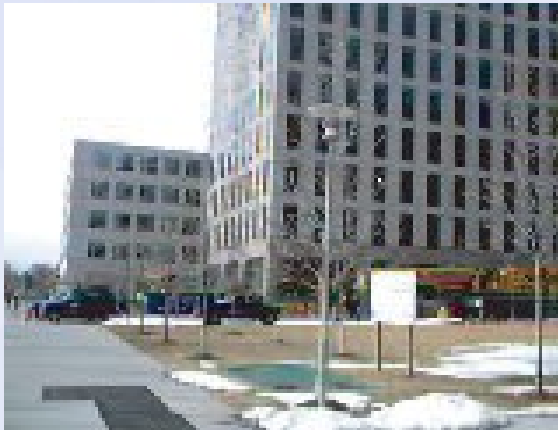


Previous Work

1. Global image features (“*gist*”)

Torralba et al, IJCV '03, ICCV '05

Characterize context by mean filter responses across grid cells



Previous Work

2. Recover structure explicitly

Hoiem, Efros, Hebert, ICCV '05, IJCV '07

Recover orientation of major surfaces

Filter object detections in unlikely positions



Previous Work

3. Correlate object locations with nearby texture

Heitz and Koller, ECCV '08

Identify superpixel characteristics that correlate with object locations

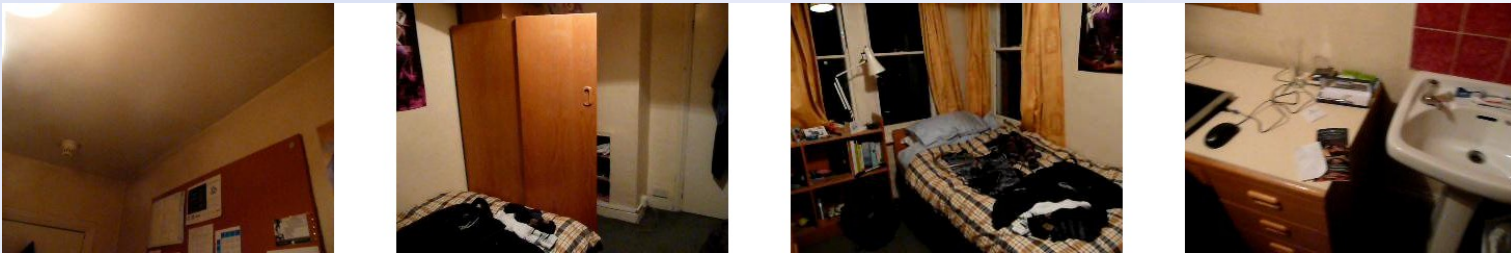


The Problem

1. Recognise rooms

Not exactly “place recognition”

Not exactly “scene category recognition”



The Problem

1. Recognise rooms

Not exactly “place recognition”

Not exactly “scene category recognition”

Previous methods are mostly inappropriate
for real-time egocentric applications

The Problem

1. Recognise rooms

Not exactly “place recognition”

Not exactly “scene category recognition”

Outline of our approach:

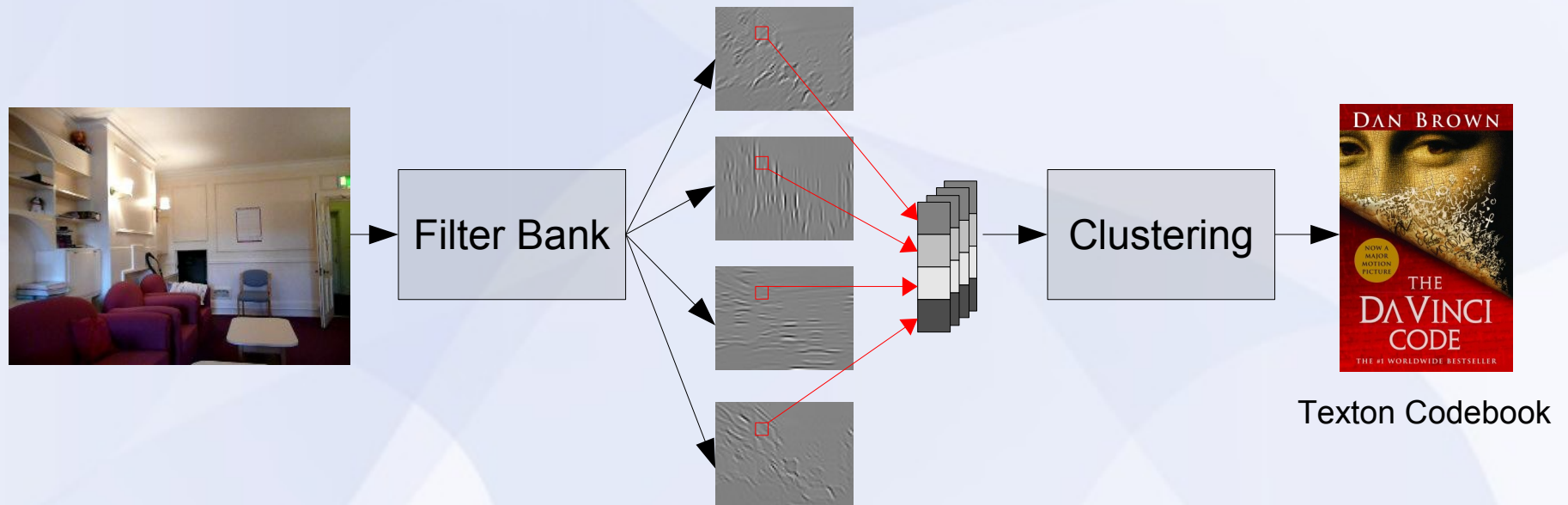
Identify textons

Reason from second order histogram

From Pixels to Textons

We follow Malik, ICCV '99

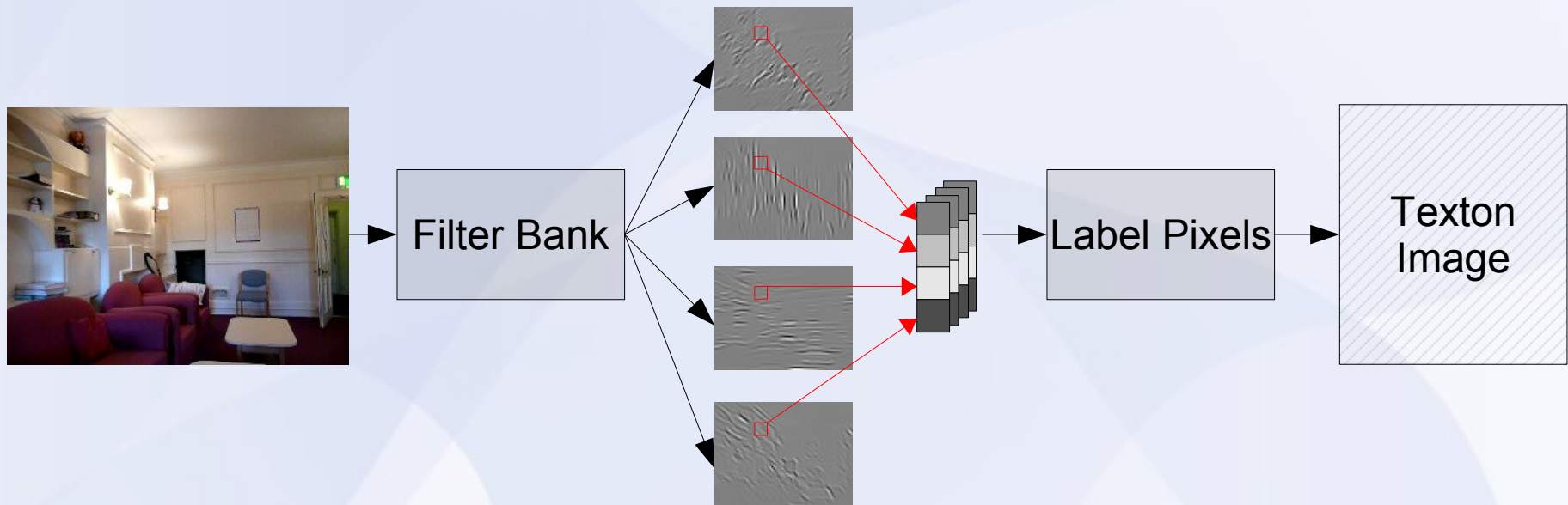
Offline: Cluster features to generate codebook



From Pixels to Textons

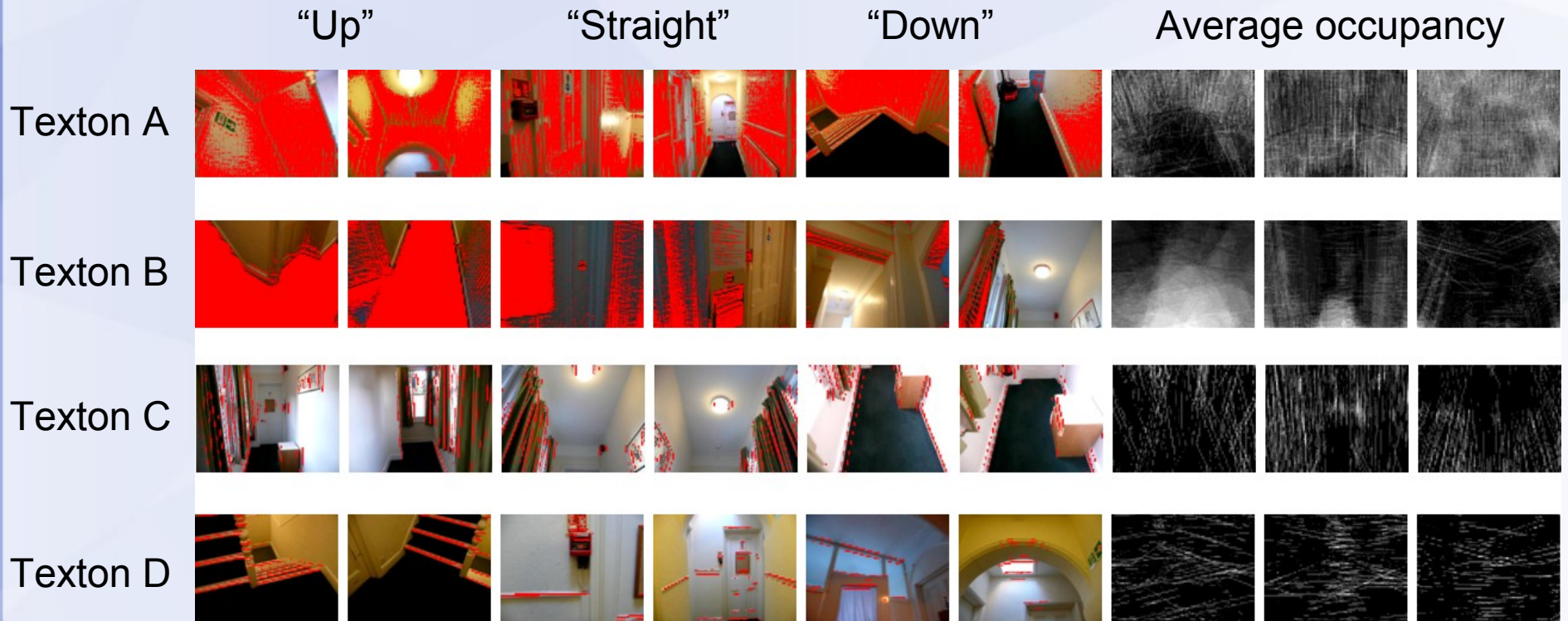
We follow Malik, ICCV '99

Online: Assign incoming pixels to nearest texton



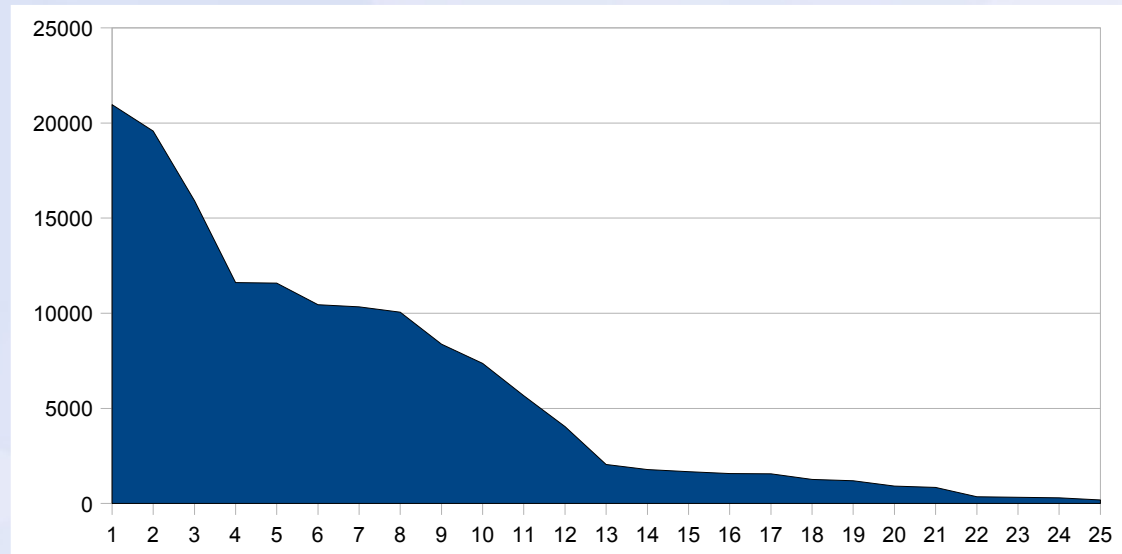
Textons and Context

Textons correlate with view orientation



Textons and Context

Textons select salient information



Untextured

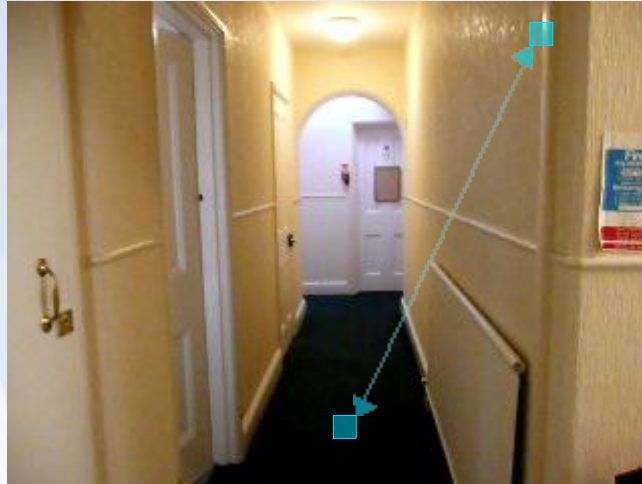
Edges and bars

Junctions,
line end points, etc

The Model

$$p(I | c) = \prod_{i=0}^N \prod_{j=0}^N p(t_i \ t_j \ s_{i,j} | c) \quad (1)$$

Texton labels Displacement Scene Class



The Model

$$p(I | c) = \prod_{i=0}^N \prod_{j=0}^N p(t_i \ t_j \ s_{i,j} | c) \quad (1)$$

Texton labels Displacement Scene Class

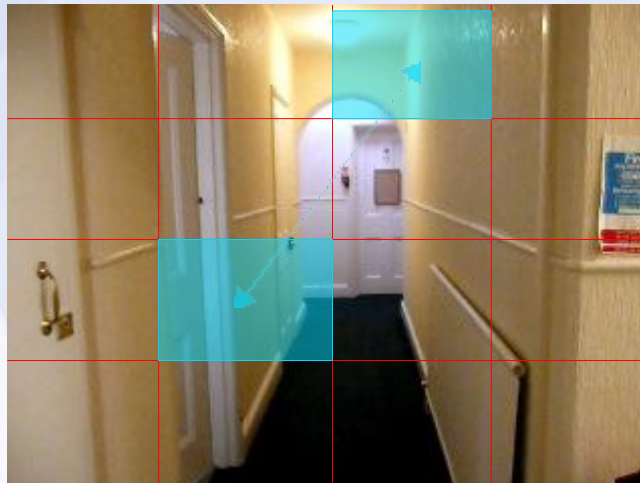
For an image with N pixels there are N² terms in (1)

The Model

Optimization: Aggregate texton counts over an $M \times M$ grid

$$\log p(I | c) = \sum_{a=0}^{M^2} \sum_{b=0}^{M^2} \sum_{i=0}^K \sum_{j=0}^K n_i^a n_j^b \log p(t_i t_j s_{a,b} | c) \quad (2)$$

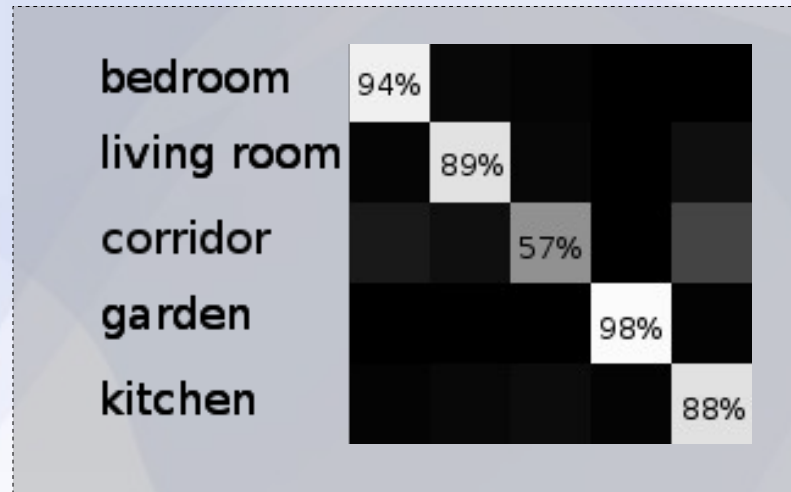
Sum over grid cell pairs Sum over texton pairs Number of texton i in cell a



$O(M^4 K^2)$

Room Recognition

| # train frames | Our system | Torralba <i>et al.</i> | KNN |
|----------------|------------|------------------------|-----|
| 103 | 81% | — | 45% |
| 230 | 83% | 62% | 52% |
| 565 | 85% | 70% | 55% |



Room Recognition



██████████ bedroom *

██████████ living room

██████████ corridor

██ garden

██████████ kitchen



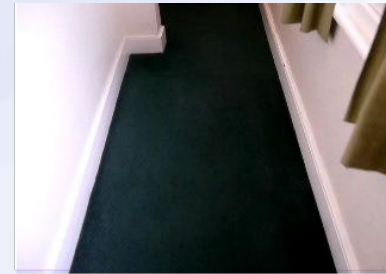
██████████ bedroom *

██████████ living room

██████████ corridor

██ garden

██████████ kitchen



██████████ bedroom

██████████ living room

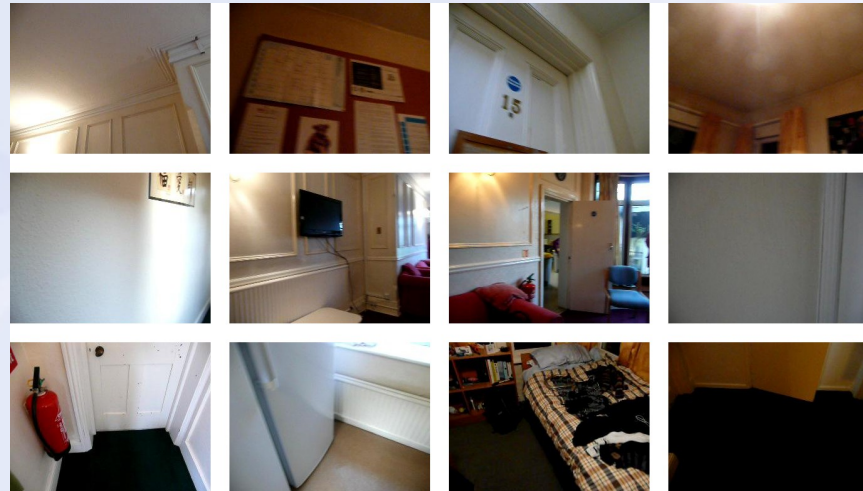
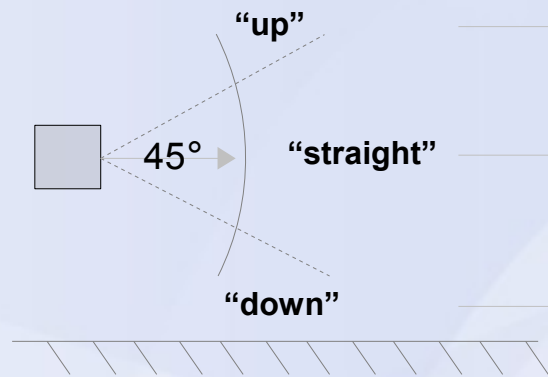
██████████ corridor *

██████████ garden

██████████ kitchen

Orientation Classification

2. Recovering coarse camera orientation



Same classification problem as before

Orientation Classification

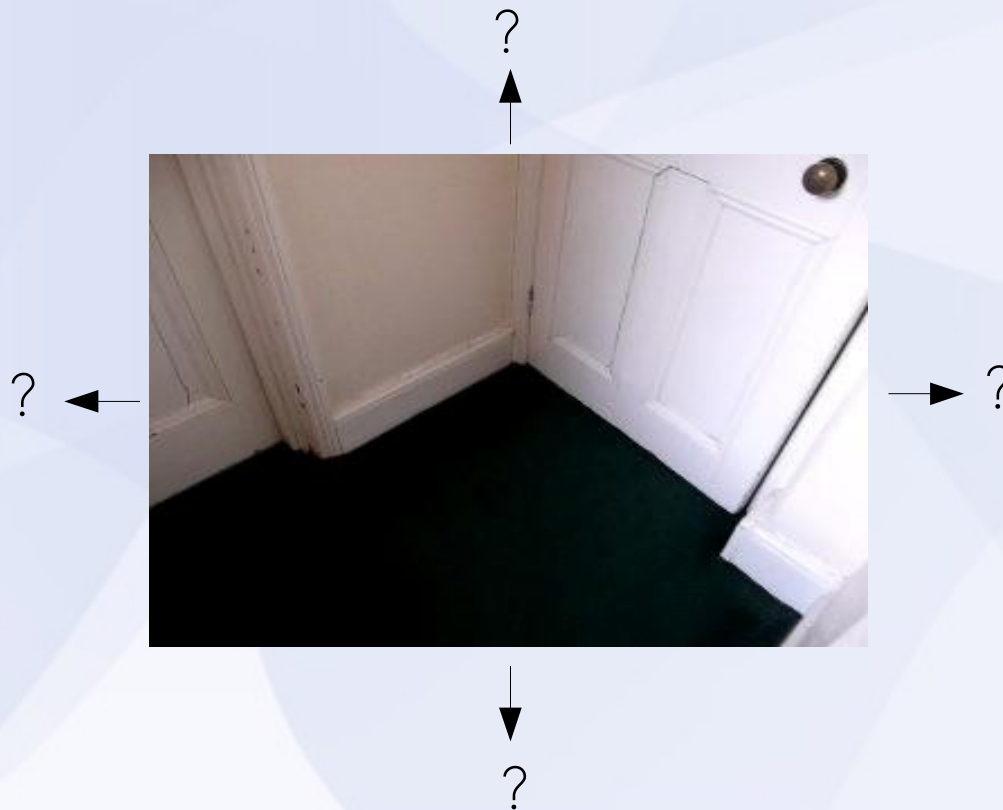
Results

| # train frames | Our system | Torralba <i>et al.</i> | KNN |
|----------------|------------|------------------------|-----|
| 88 | 70% | 61% | 59% |
| 728 | 79% | 63% | 59% |

| | | | |
|----------|-----|-----|-----|
| up | 90% | | |
| straight | | 72% | |
| down | | | 72% |

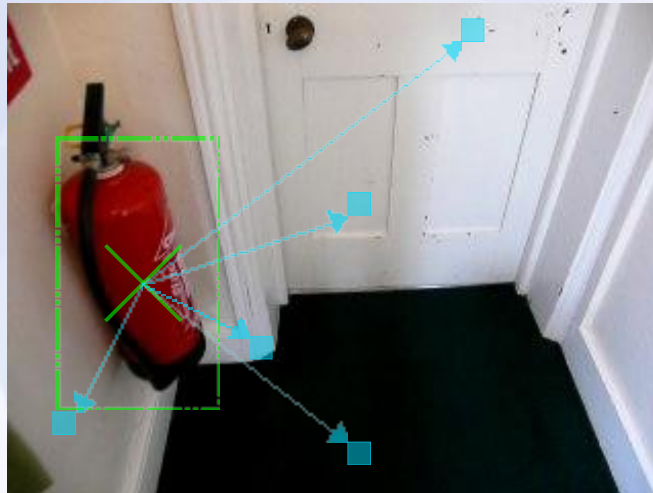
Active Search

3. Given the current frame, how should I move the camera to find object X?



Active Search

3. Given the current frame, how should I move the camera to find object X?

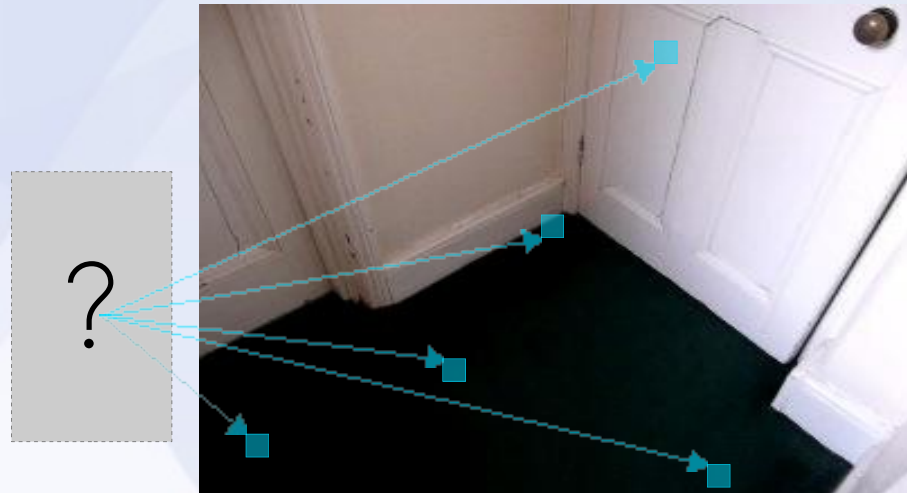


Active Search

$$\log p(\mathbf{x} \mid D) = \sum \log p(\mathbf{x} \mid t_i \mathbf{y}_i) - \sum \log p(t_i \mathbf{y}_i) \quad (3)$$

Object location Texton label Texton location

Reason about locations outside the image



Active Search

Training data: 71 frames containing fire extinguishers, annotated with object centre

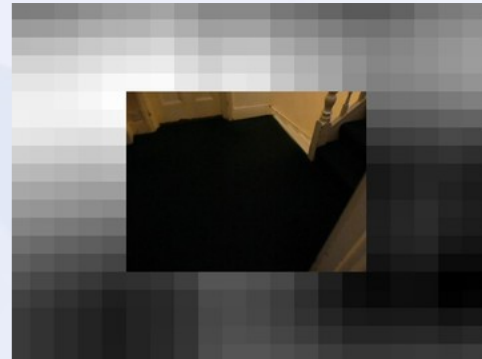
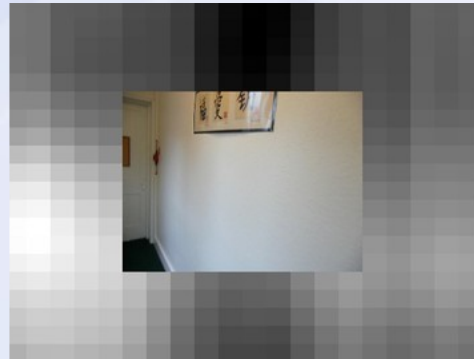
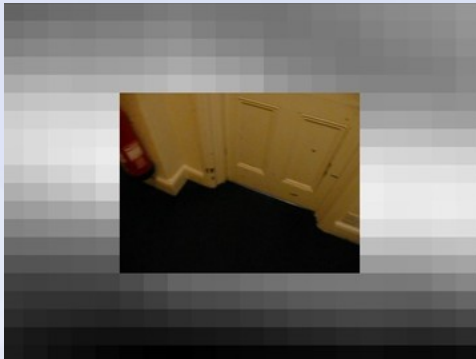
Build joint object/texton/displacement histogram

Evaluation:

Compute marginal for position of fire extinguisher

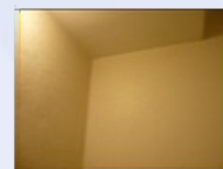
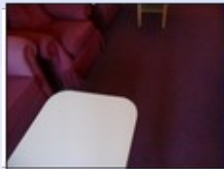
Active Search

Marginal for position of fire extinguishers



Active Search

Marginal for y coordinate of fire extinguishers



Conclusion

Preliminary experiments indicate that textons are a promising basis for inferring scene context

Particularly suited to ego-centric applications because

- Real-time performance

- Robust to uninformative views of a scene

Outperforms *gist* descriptor for this type of problem

Future work: further experiments!