Cognitive Vision for Cognitive Systems

Barbara Caputo, Marco Fornoni
Idiap Research Institute
http://www.idiap.ch/~bcaputo
http://www.idiap.ch/~mfornoni
bcaputo@idiap.ch
mfornoni@idiap.ch

Useful Info

- **56 hours course** (28 teaching, 28 laboratory)
- **4 credits**
- **Topics**
  - *Scene Recognition and Understanding*
  - *Object Recognition and Categorization*
  - *Action Recognition and Understanding*
  - *Life Long Learning of Concepts*

Useful Info

- **web-page course:http://www.idiap.ch/ftp/courses/EE-700/CogVisCogSys.html**
- **how to reach me/Marco: email ({bcaputo,mfornoni}@idiap.ch)**
- **Exam:**
  - *Report on laboratory experiences, with discussion*
  - *Oral presentation of research paper*
  - *Date: ?????*

- **Exam: Report on laboratory experiences**
  - *For each topic, there will be a corresponding laboratory experience*
  - *It will consist of replicating the experiments of a seminal paper in the field, on the same data presented in the paper and on different data collections (mandatory)*
  - *For the mandatory part of the work, we provide software and data, you develop the tools for the analysis of the experimental results*

- **Exam: Report on laboratory experiences**
  - *Optional: more exciting, research-like stuff (will require some coding)*
  - *Once all the experiences are done, you write a report with one chapter for each experience, and you send it to bcaputo@idiap.ch*
  - *Minimum for passing the exam: all experiences done and well reported, plus at least for one experience some optional work done*
  - *No special requirements on length, template, etc*
  - *To be submitted at the very latest 15 days before the day of the exam!!*

- **Exam: Oral Presentation of Research Paper**
    - *For each topic, I will present the most recent trends in the research field, i.e. papers presented during the last 6-9 months at the top conferences in the field (acceptance rate 40-20%)*
    - *Between the papers presented in this lecture, you pick one by sending me an email (first come, first serve)*
    - *The day of the exam you make a 30m presentation of the paper, putting it into the context of what was discussed during lectures*
    - *Exam consists of: (1) doing lab experiences and reporting on them (2) discussion of the lab experience report (3) 30m presentation of paper chosen by you*

Scene Recognition (continued)

Some useful thoughts

- We easily (= quickly) distinguish between indoor and outdoor scenes

Some useful thoughts

- We are able to identify easily (= quickly) few landmark objects in a scene

Some useful thoughts

- We expect to find some objects only in certain parts of the scene

Human visual perception

- ***What do we remember and what do we forget when we recall a scene?***
  - *WE DO REMEMBER:* the gist of a scene, 4-5-landmark objects and their spatial configuration
  - *WE DO NOT REMEMBER:* all the objects in the scene, mid- to fine details

J. M. Wolfe. *Visual memory: what do you know about what you saw?* Current Biology, 1998, 8: R303-R304

Computer Vision

- Most of work on ***outdoor*** place recognition, only recently (2009) first attempts on indoor place recognition
- Gist of a scene = holistic representation
- Applications: image retrieval, context priming

A. Oliva, A. Torralba. *Modeling the shape of the scene: a holistic representation of the spatial envelope.* International Journal of Computer Vision, 42(3), 145-175, 2001

Towards indoor scene recognition

A.Quattoni, A.Torralba. *Recognizing indoor scenes.* Proc International Conference on Computer Vision and Pattern Recognition, 2009

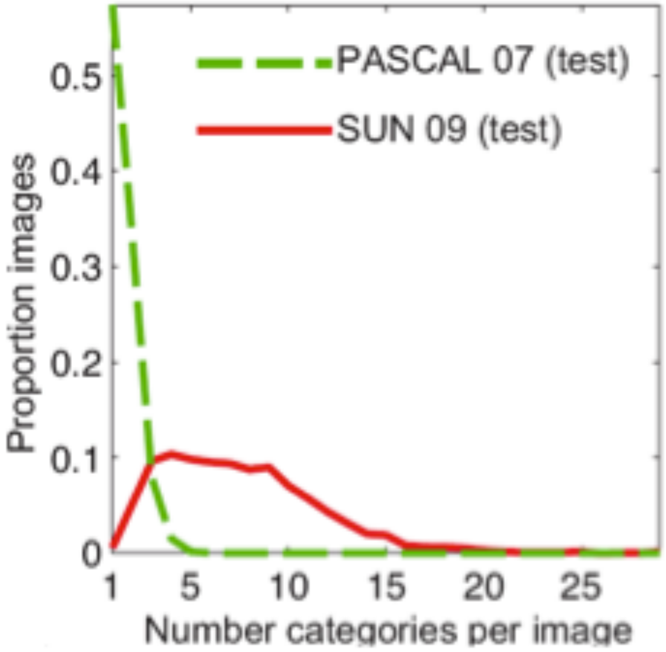
- *Contribution I:* experimental evaluation of several methods for outdoor recognition on Lazebnik et al 2006 database, outlining current limitations
- *Contribution 2:* a database of 67 indoor categories, publicly available
- *Contribution 3:* a new computational model for tackling the indoor scene recognition problem

But are 67 scenes enough?

J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba. *SUN database: large scale scene recognition from Abbey to Zoo.* Proc International Conference on Computer Vision and Pattern Recognition, 2010

- *Contribution l:* the largest existing database of visual scenes
- *Contribution 2:* annotation at the level of scenes and objects
- *Contribution 3:* baseline given in terms of algorithmic and human performance

SUN database

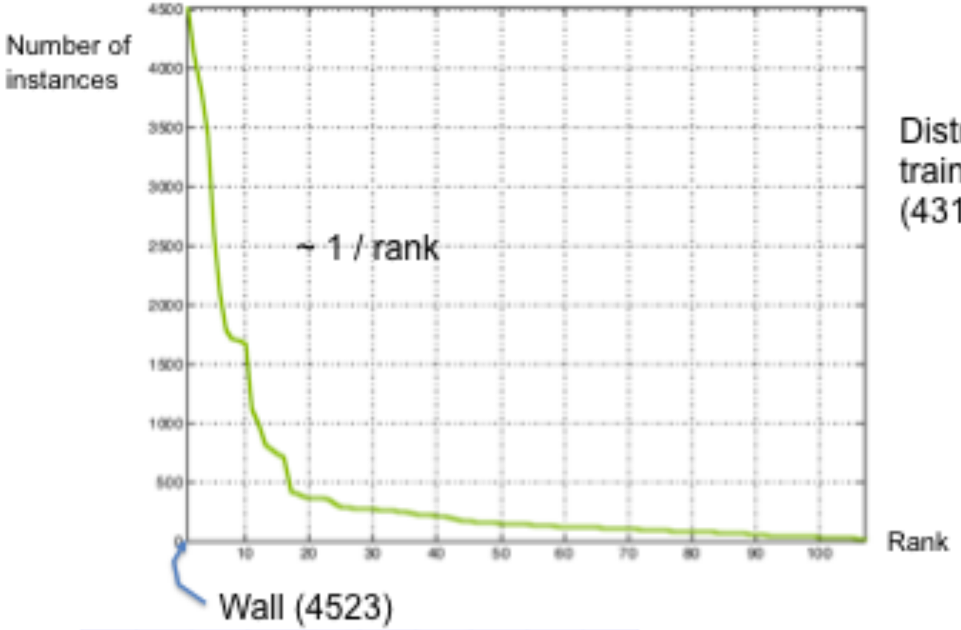


12,000 annotated images

107 object categories

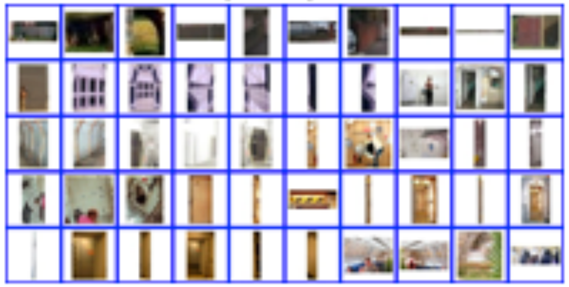
152,000 annotated object instances

Distribution of objects in scenes

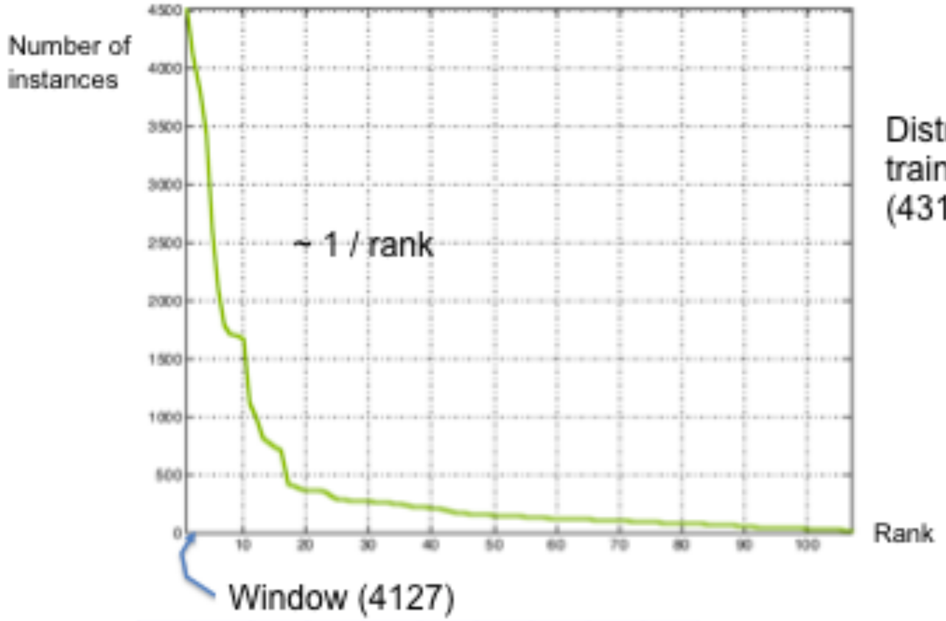


Wall (4523)

Distribution from training set (4317 images)

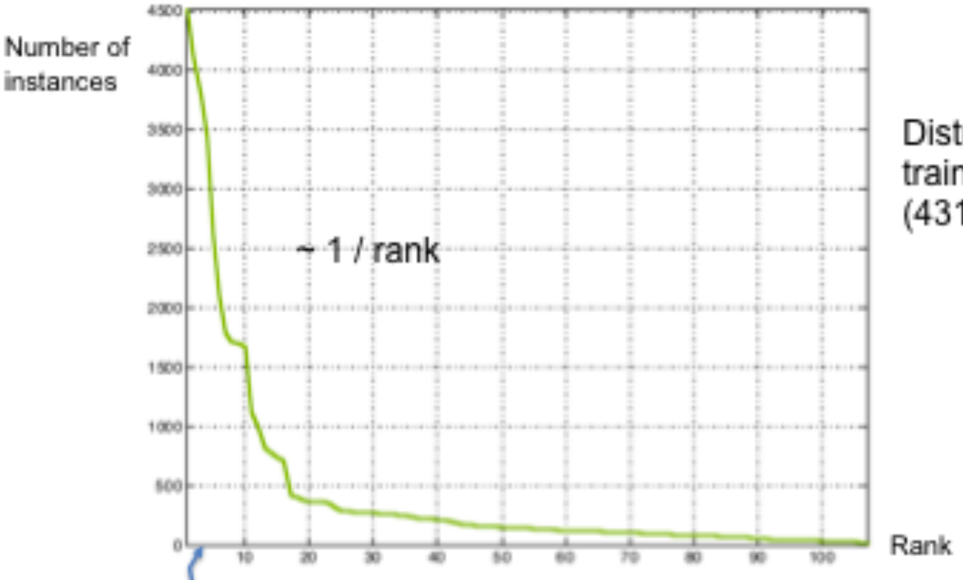


Distribution of objects in scenes



Distribution from training set (4317 images)

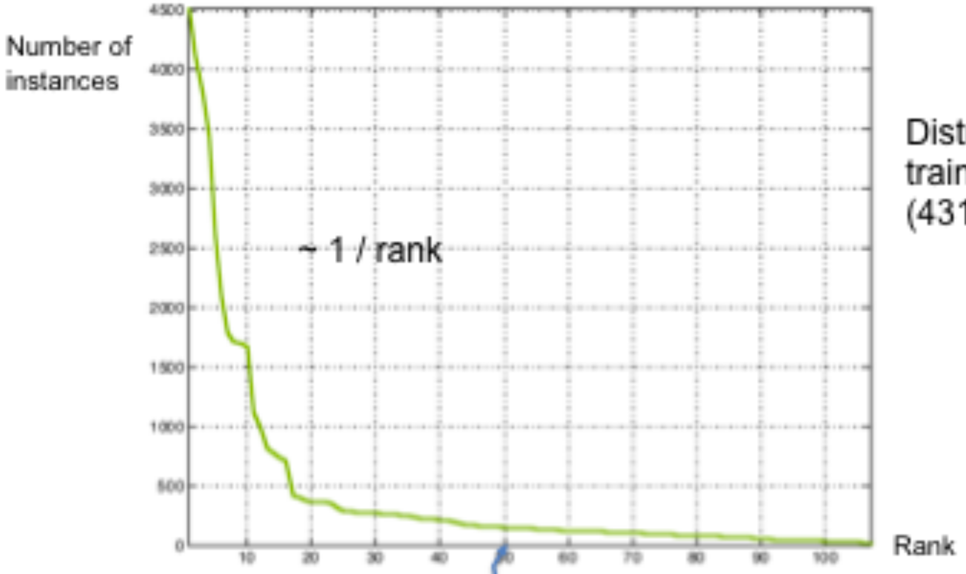
Distribution of objects in scenes



Distribution from training set (4317 images)

Building (3804)

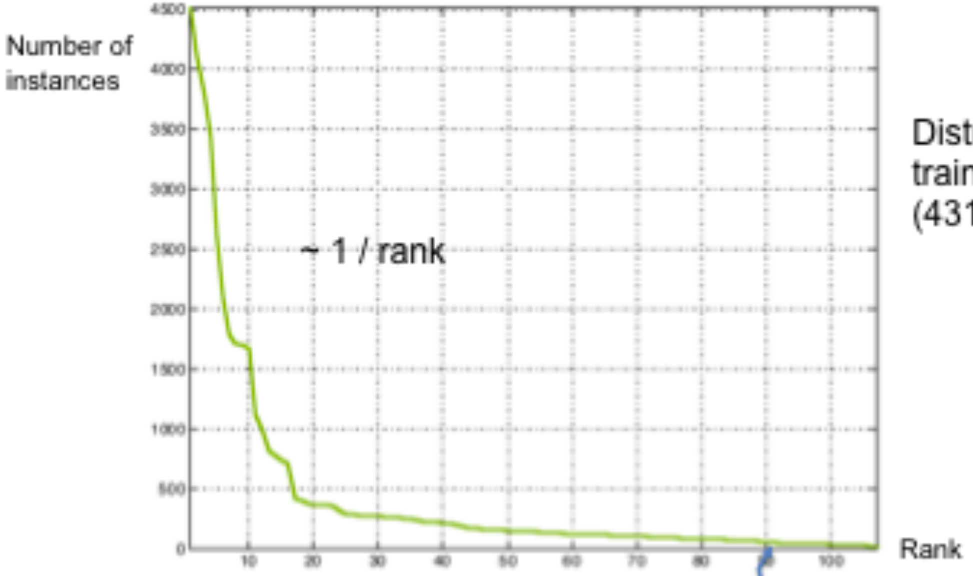
Distribution of objects in scenes



Distribution from training set (4317 images)

Sofa (158)

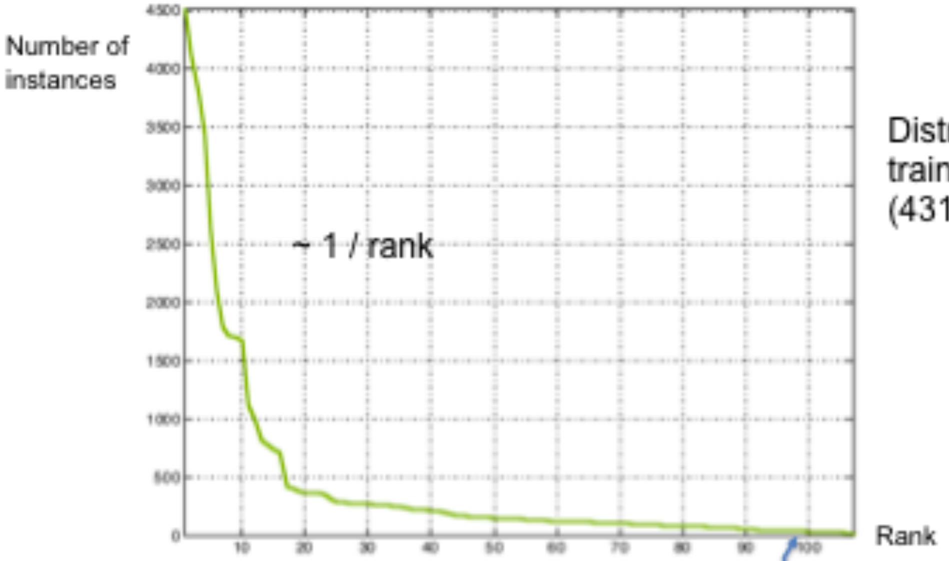
Distribution of objects in scenes



Distribution from training set (4317 images)

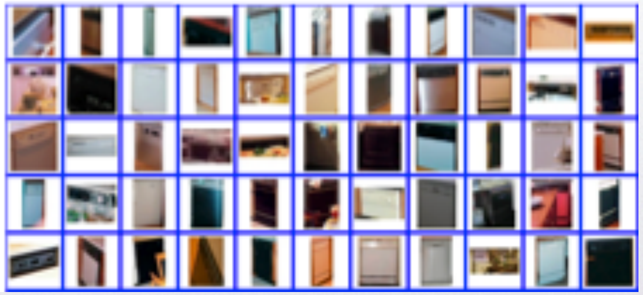
Headstone (62)

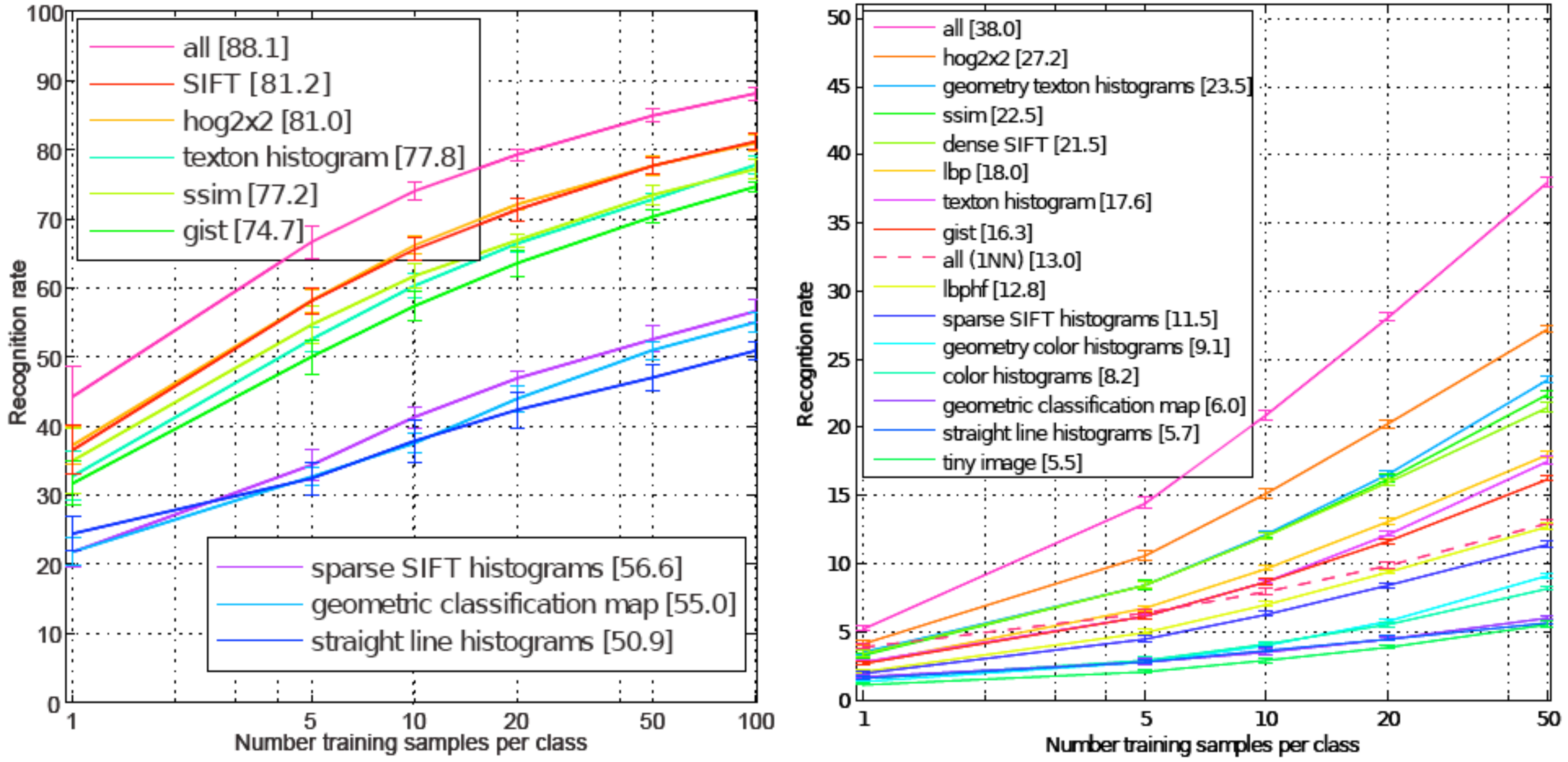
Distribution of objects in scenes



Distribution from training set (4317 images)

Dishwasher (44)





(a) 15 scene dataset

(b) SUN database

car interior frontseat
(91% vs 85%)

limousine interior
(95% vs 80%)

riding arena
(100% vs 90%)

sauna
(96% vs 95%)

skatepark
(96% vs 90%)

subway interior
(96% vs 80%)

volleyball court indoor
(95% vs 80%)

abbey
(0% vs 0%)

hunting lodge outdoor
(11% vs 5%)

inn outdoor
(0% vs 0%)

lecture room
(6% vs 5%)

library outdoor
(10% vs 5%)

monastery outdoor
(5% vs 5%)

synagogue indoor
(6% vs 5%)

bedroom
(100% vs 10%)

hospital room
(96% vs 10%)

gas station
(100% vs 15%)

balcony exterior
(87% vs 5%)

corral
(90% vs 10%)

gymnasium indoor
(100% vs 20%)

dam
(95% vs 15%)

sandbar
(5% vs 75%)

oast house
(30% vs 85%)

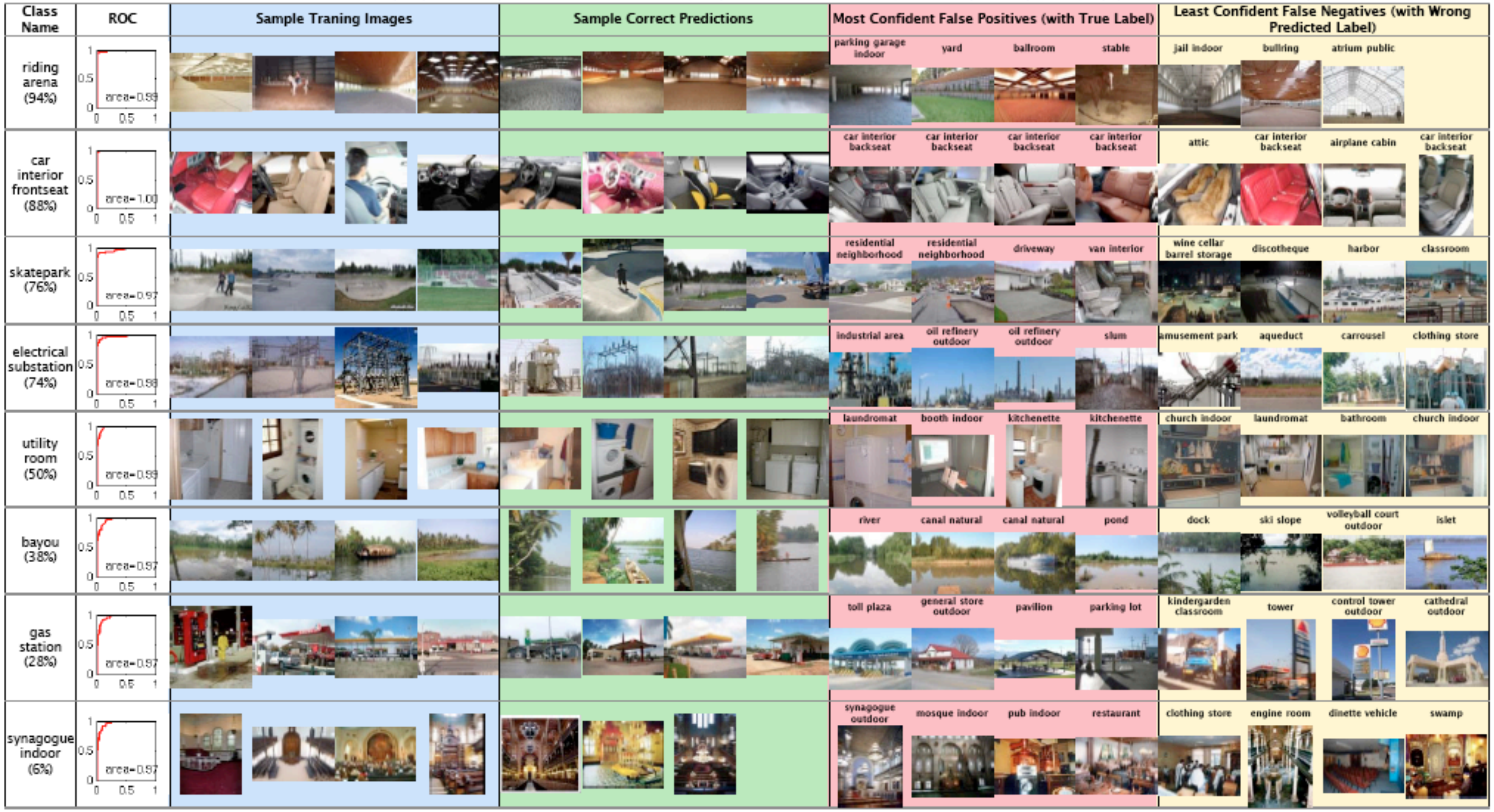
apse indoor
(0% vs 55%)

stadium baseball
(8% vs 55%)

landfill
(23% vs 65%)

medina
(24% vs 65%)

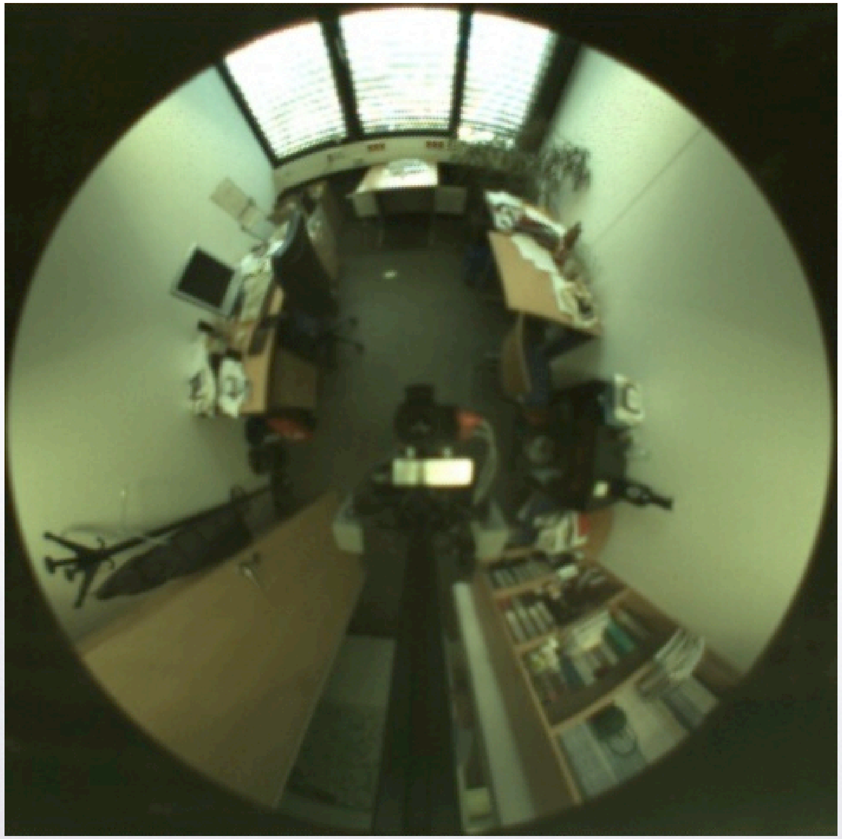
bayou
(0% vs 40%)

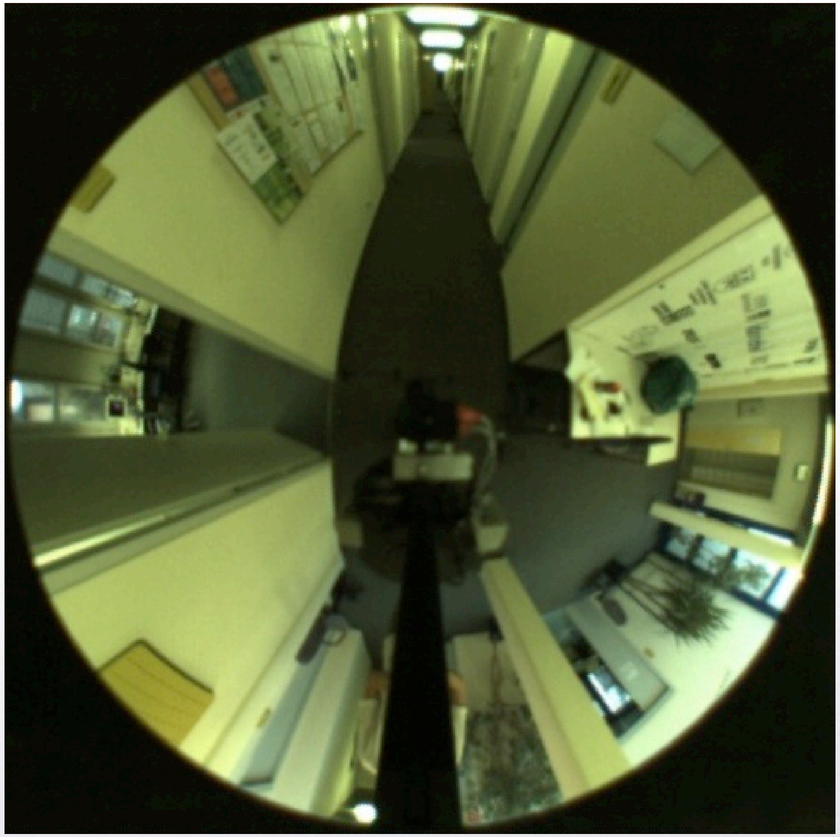


Class Name	ROC	Sample Traning Images	Sample Correct Predictions	Most Confident False Positives (with True Label)	Least Confident False Negatives (with Wrong Predicted Label)
riding arena (94%)	area=0.99			parking garage indoor yard ballroom stable	jail indoor bullring atrium public
car interior frontseat (88%)	area=1.00			car interior backseat car interior backseat car interior backseat car interior backseat	attic car interior backseat airplane cabin car interior backseat
skatepark (76%)	area=0.97			residential neighborhood residential neighborhood driveway van interior	wine cellar barrel storage discotheque harbor classroom
electrical substation (74%)	area=0.98			industrial area oil refinery outdoor oil refinery outdoor slum	amusement park aqueduct carrousel clothing store
utility room (50%)	area=0.99			laundromat booth indoor kitchenette kitchenette	church indoor laundromat bathroom church indoor
bayou (38%)	area=0.97			river canal natural canal natural pond	dock ski slope volleyball court outdoor islet
gas station (28%)	area=0.97			toll plaza general store outdoor pavilion parking lot	kindergarden classroom tower control tower outdoor cathedral outdoor
synagogue indoor (6%)	area=0.97			synagogue outdoor mosque indoor pub indoor restaurant	clothing store engine room dinette vehicle swamp

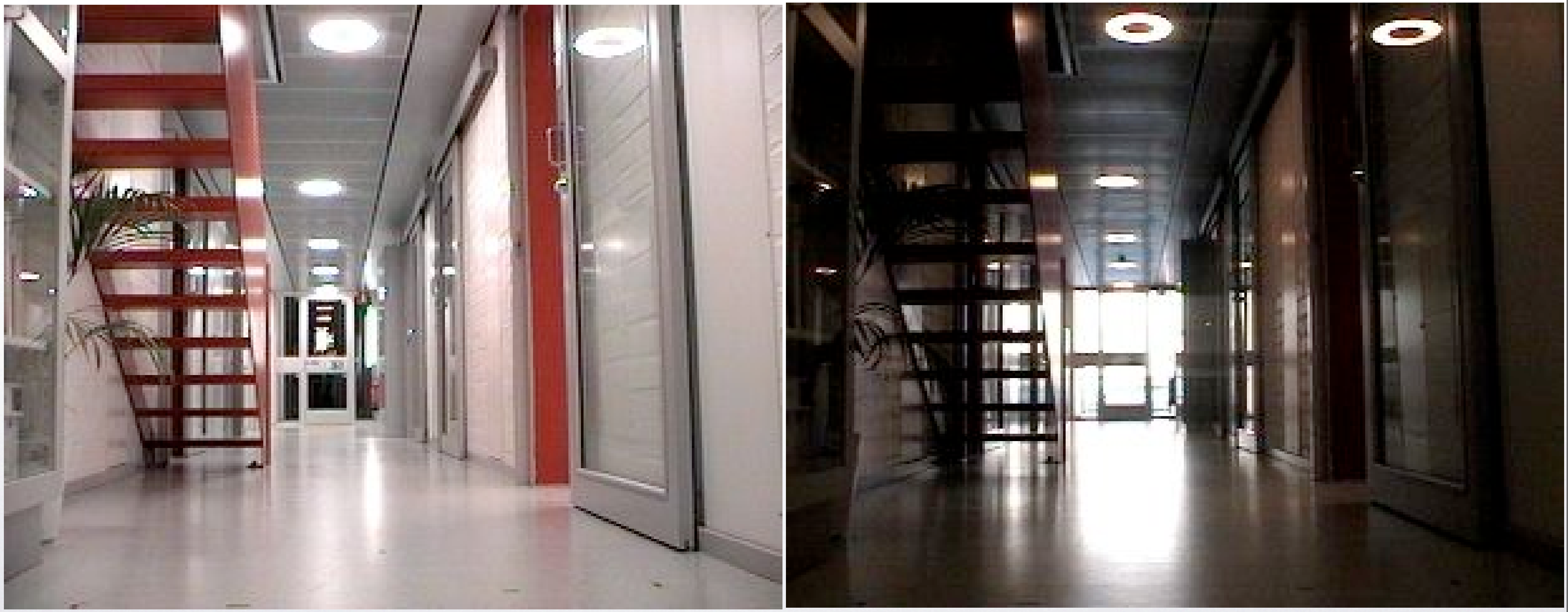
Scene Recognition --the Robot's Perspective

What do you see?



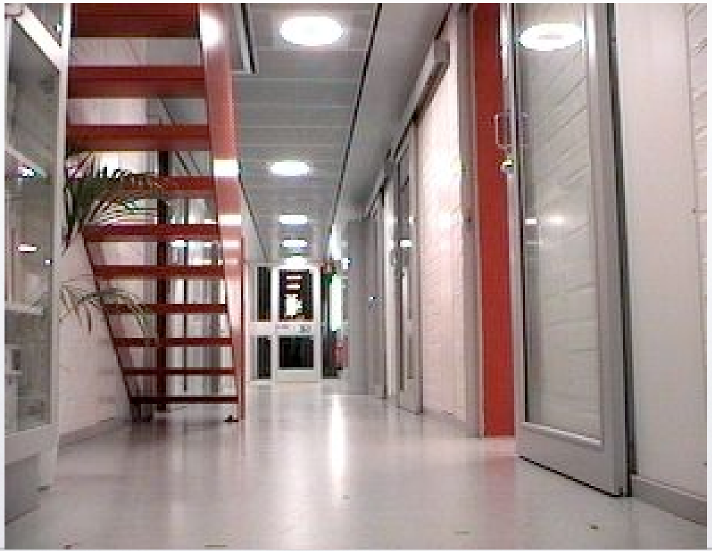


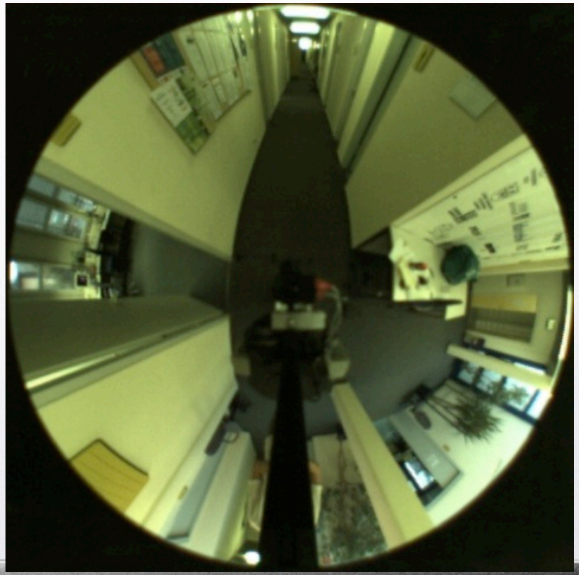
What do you see?



Some useful thoughts

- The embodiment (= where the camera is positioned) and the perceptual capabilities (= type of camera) determines what the robot sees of a scene





Some useful thoughts

- The robot does not know what is informative and what is not, therefore it acquires everything

Why it is useful?

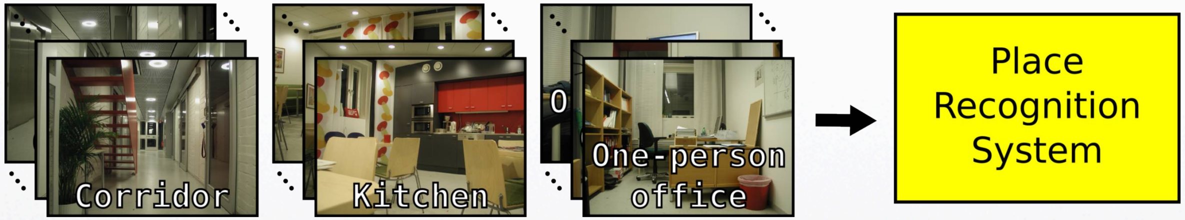
- Build a multi-layer representation of space and use it to navigate/interact in it

Step 1: place recognition

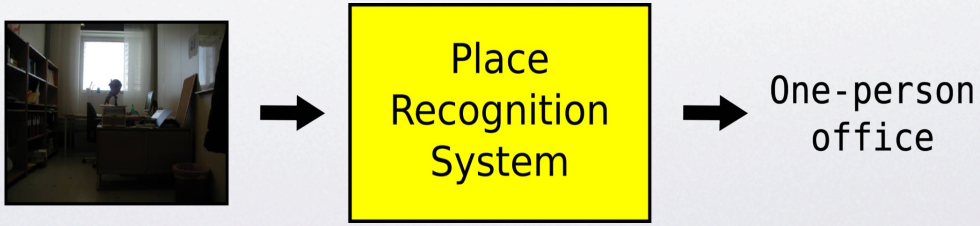
Place Recognition System [A. Pronobis, et al. IROS'06]

Fully supervised, appearance-based system capable of recognizing a indoor environment on the based of their visual appearance. We used global and local features as input of an SVM.

- Learning (Training)



- Recognition

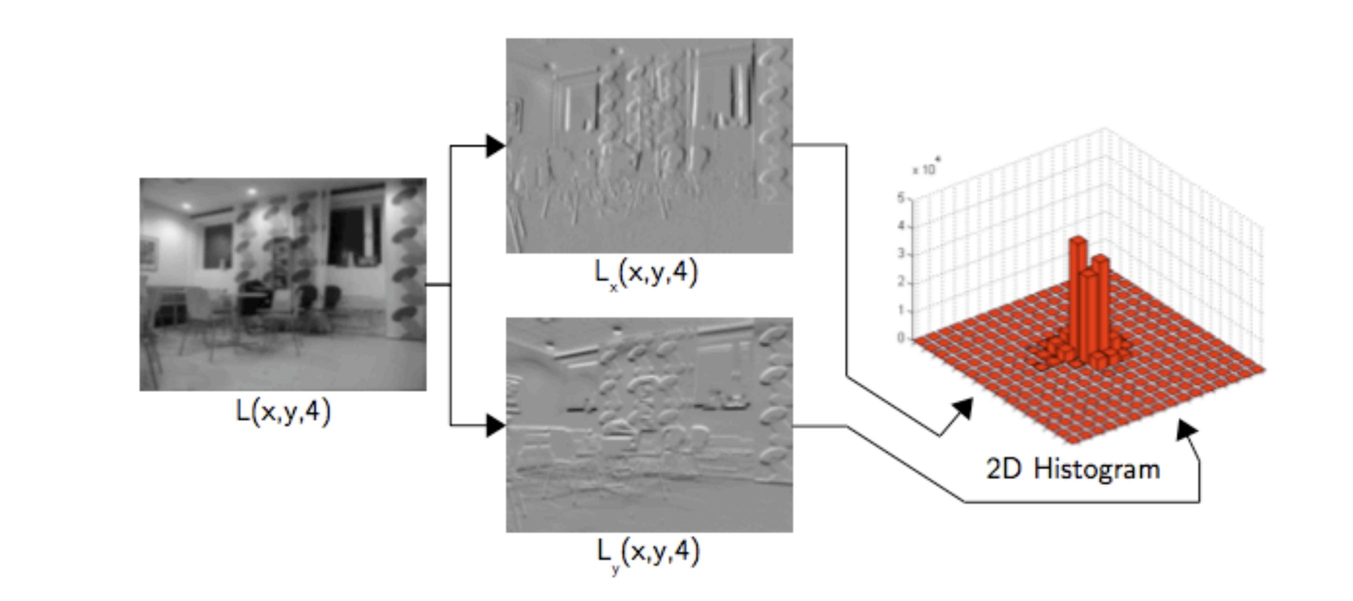


Place Recognition System

- Feature Extraction
    - CRFH: High Dimensional Composed Receptive Receptive Field Histogram [Linde and Lindeberg, ICPR’04]
    - SIFT [Lowe, ICCV’99]
- Classifier: Support Vector Machines
    - Good generalization properties

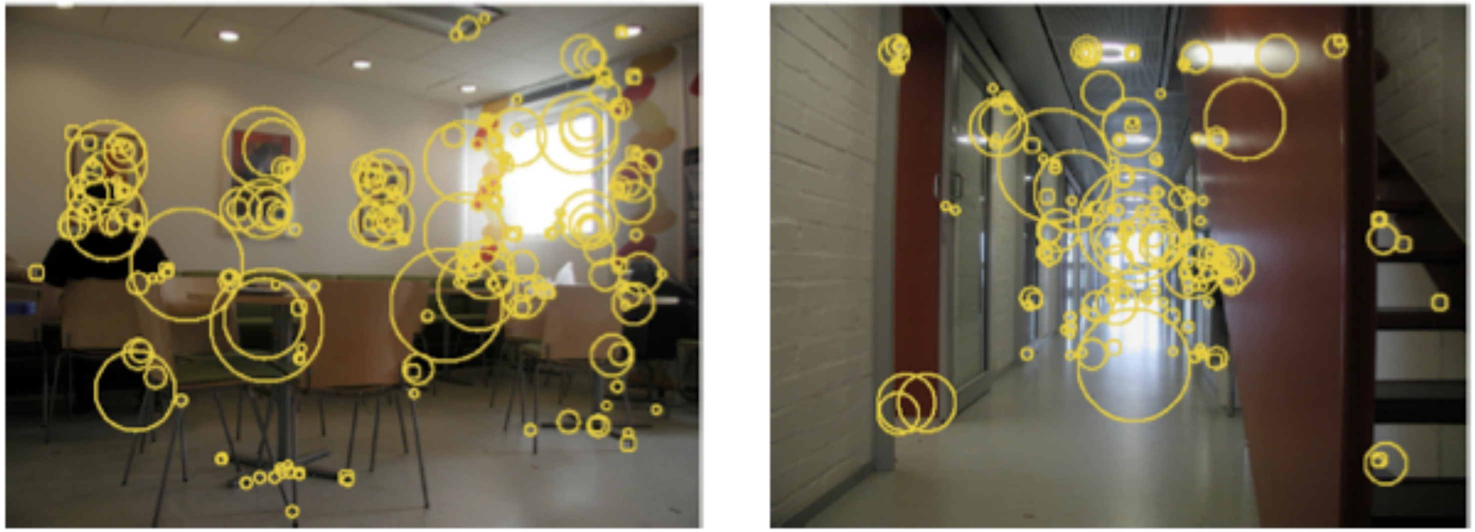
Place Recognition System

- Feature Extraction
  - CRFH: High Dimensional Composed Receptive Receptive Field Histogram [Linde and Lindeberg, ICPR'04]



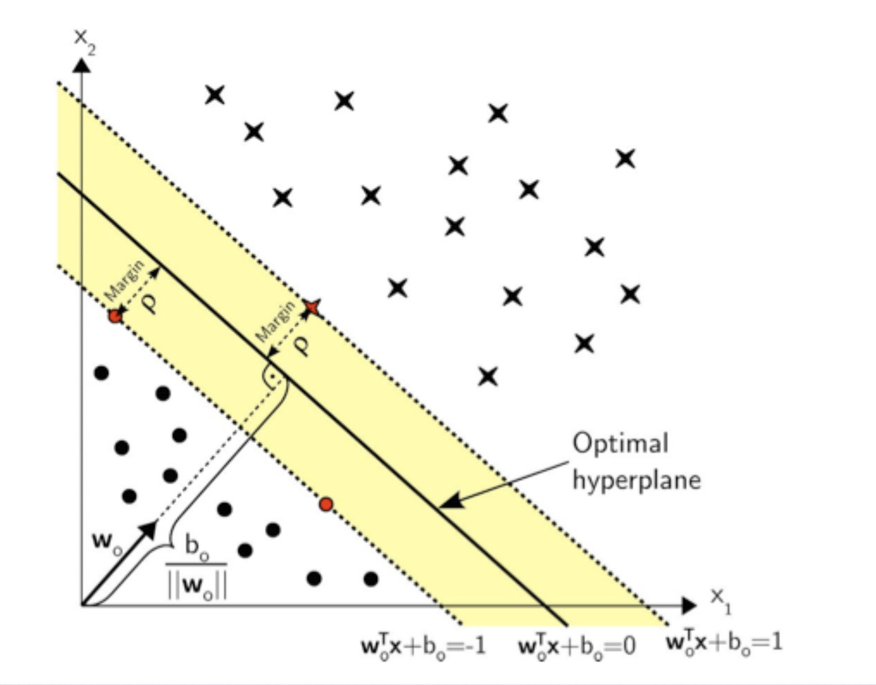
Place Recognition System

- Feature Extraction
  - SIFT [Lowe, ICCV'99]

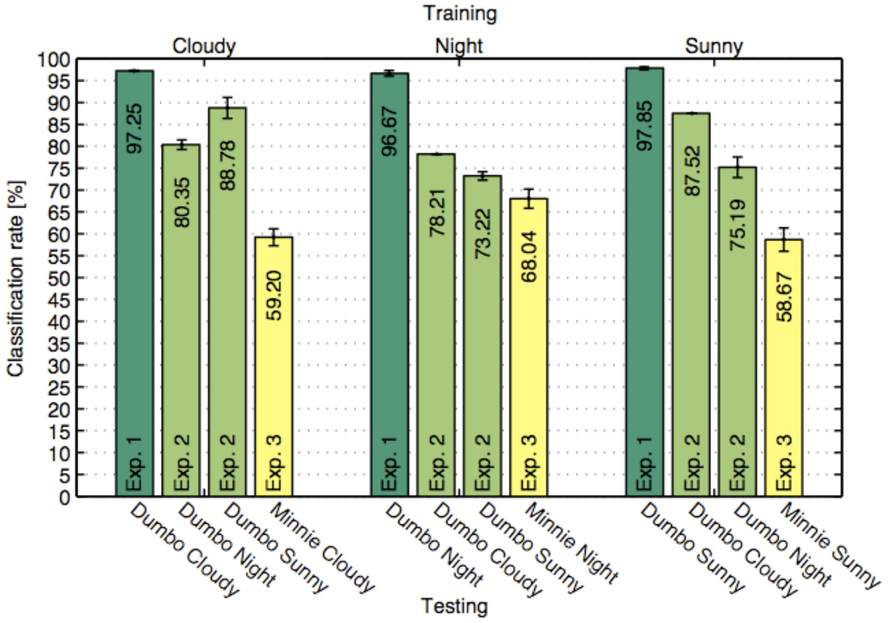


Place Recognition System

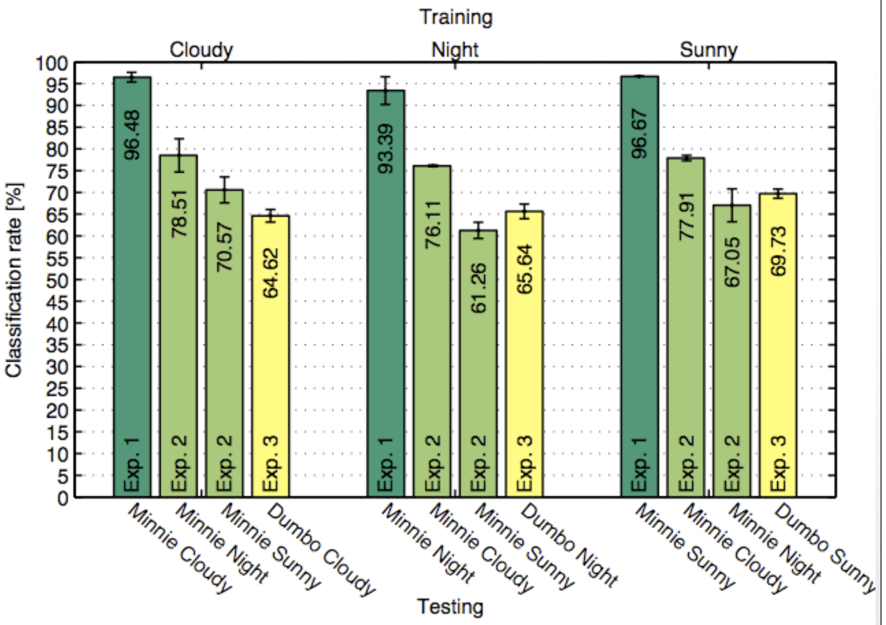
- Classifier: Support Vector Machines



Results

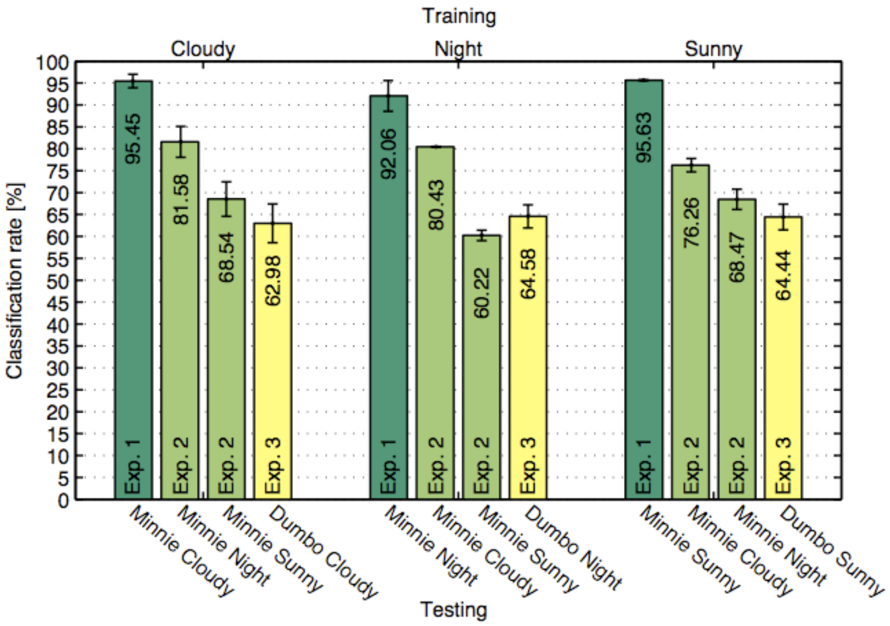


(c) Training on global features (*CRFH*) extracted from images acquired with *Dumbo*.

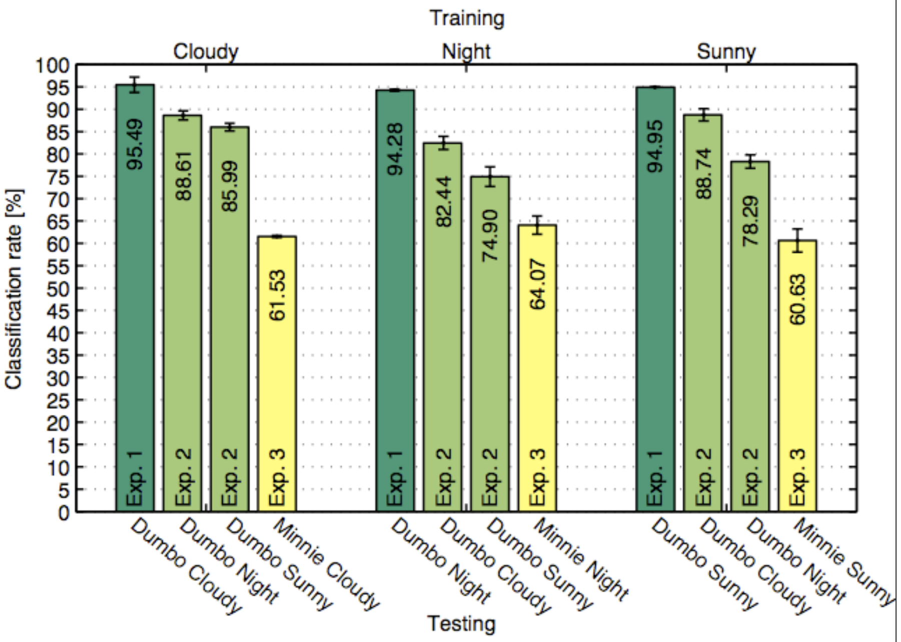


(a) Training on global features (*CRFH*) extracted from images acquired with *Minnie*.

Results



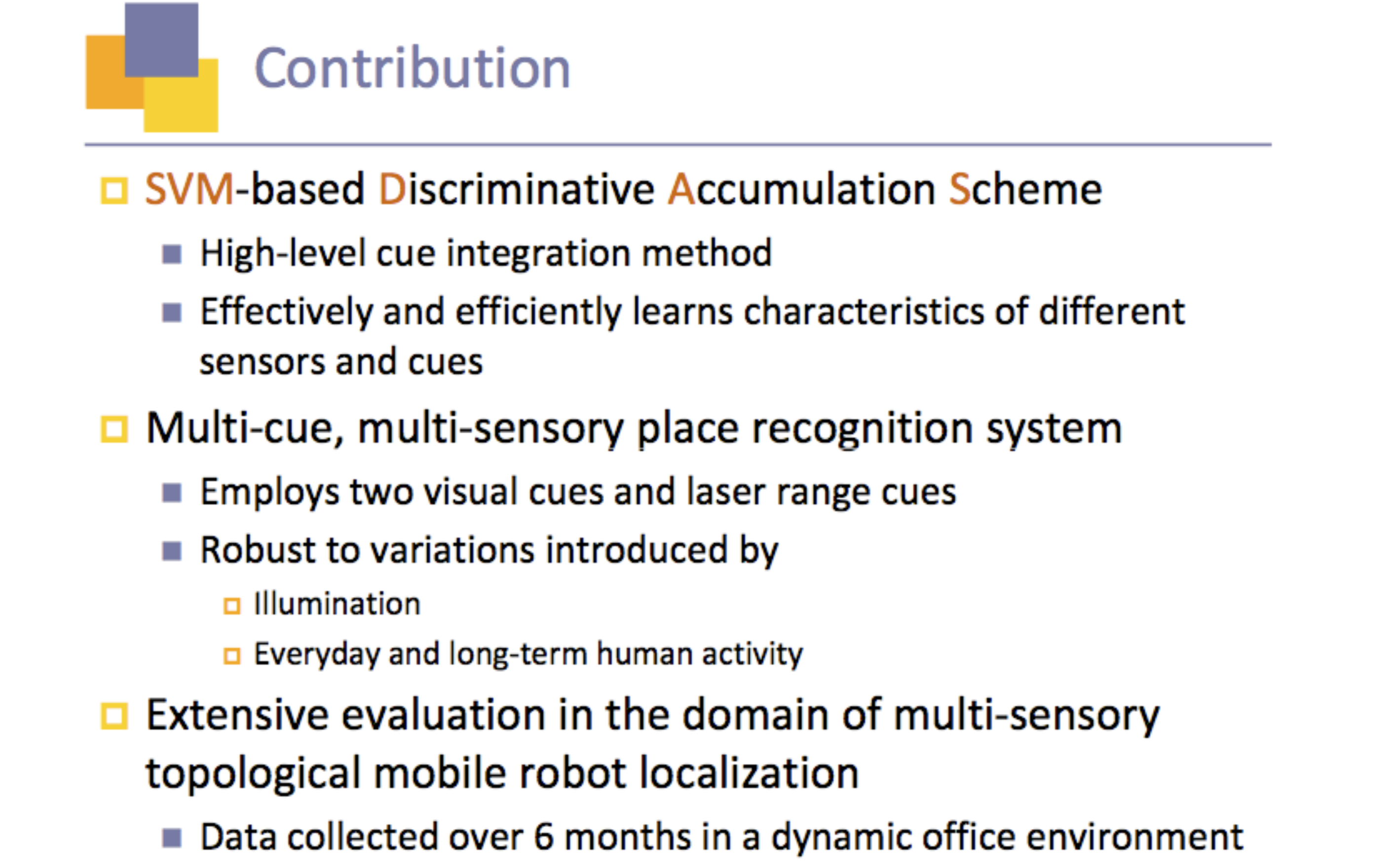
(b) Training on local features (*SIFT*) extracted from images acquired with *Minnie*.



(d) Training on local features (*SIFT*) extracted from images acquired with *Dumbo*.

15 min break!

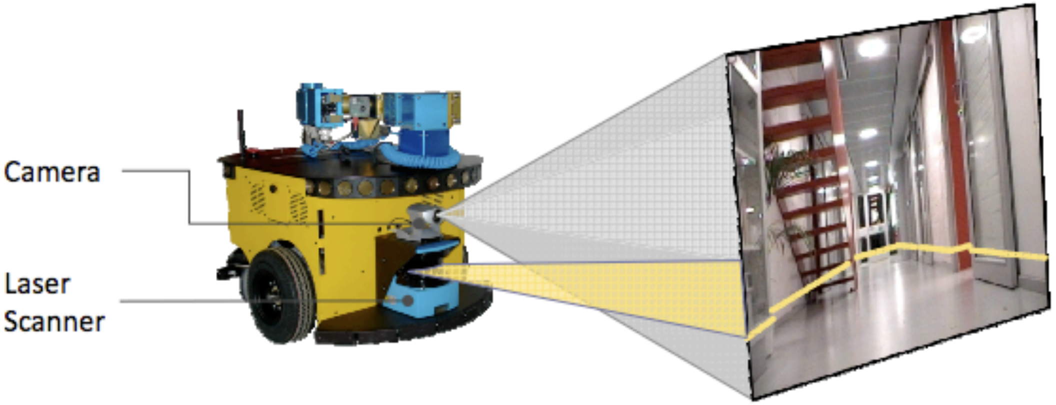
A. Pronobis, O. Martinez-Monoz, B. Caputo, P. Jensfelt. *Multi-modal semantic place classification*. IJRR, 29 (2-3): 298-320, 2010.



Contribution

- □ SVM-based Discriminative Accumulation Scheme
  - High-level cue integration method
  - Effectively and efficiently learns characteristics of different sensors and cues
- □ Multi-cue, multi-sensory place recognition system
  - Employs two visual cues and laser range cues
  - Robust to variations introduced by
    - □ Illumination
    - □ Everyday and long-term human activity
- □ Extensive evaluation in the domain of multi-sensory topological mobile robot localization
  - Data collected over 6 months in a dynamic office environment

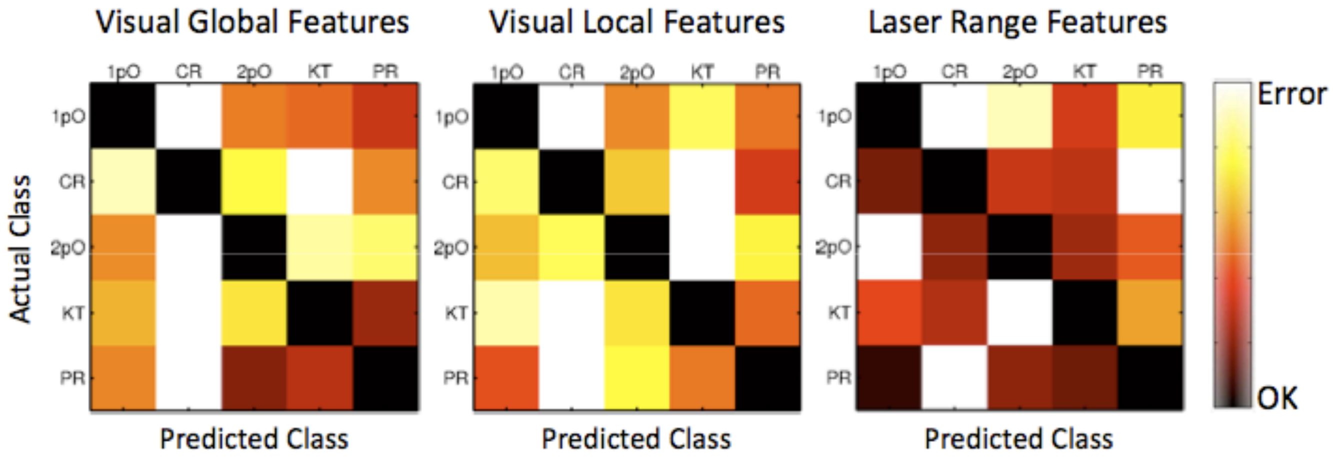
Motivation
Multimodal Cue Integration



- Range sensors
  - Pros
    - Robust to visual variations
    - Data easy to process
  - Cons
    - Suffers from perceptual aliasing
    - Purely metric information
- Visual sensors
  - Pros
    - Rich and descriptive
    - Source of semantic information
  - Cons
    - Noisy
    - More data to process

Motivation
Multi-cue Place Recognition

- Distribution of errors made by single cue systems



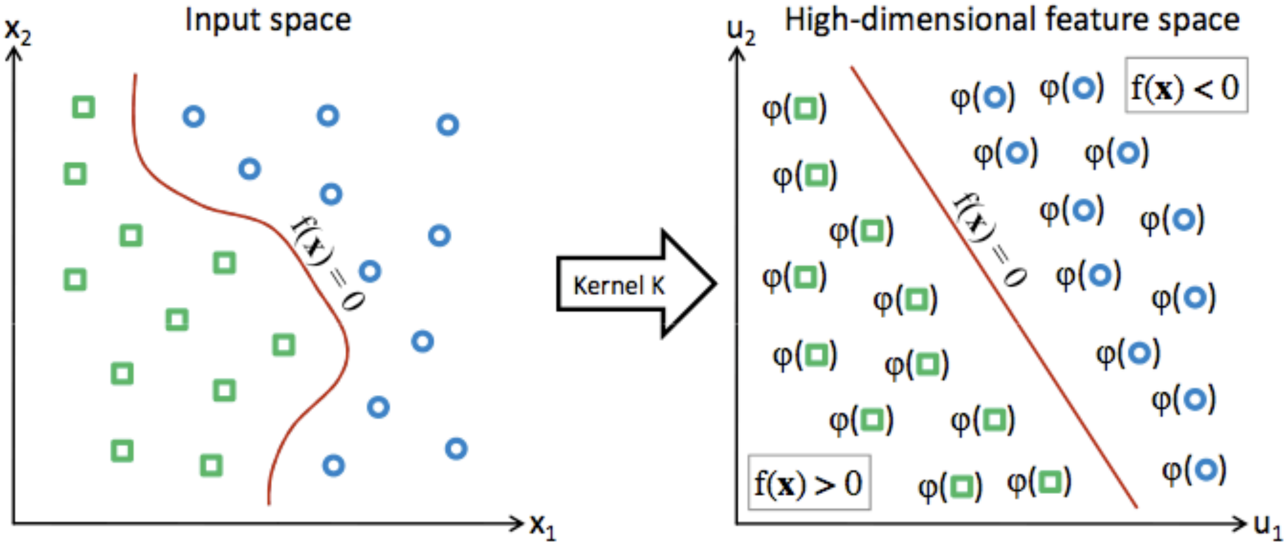
Visual Global Features

Visual Local Features

Laser Range Features

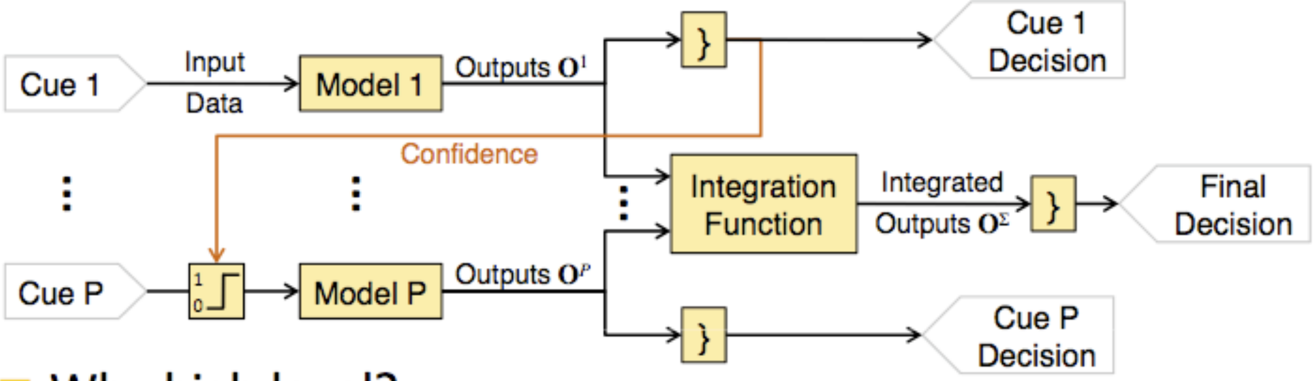
- How can we use multiple cues effectively?
- Can we learn these different patterns?
- Can we do it efficiently?

Support Vector Machines [Cristianini&Taylor'99]



- Discriminant function: $f(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i y_i \mathrm{K}(\mathbf{x}_i, \mathbf{x}) + b$
- Multi-class extensions: one-vs-one, one-vs-all, modified one-vs-all [Pronobis & Caputo '07]

SVM-DAS
High Level Integration



Cue 1

Input Data

Model 1

Outputs $\mathbf{O}^1$

}

Cue 1 Decision

Confidence

⋮

Integration Function

Integrated Outputs $\mathbf{O}^\Sigma$

Final Decision

Cue P

1 0

Model P

Outputs $\mathbf{O}^P$

Cue P Decision

- Why high level?
  - Cues are treated independently
    - Models adapted to characteristics of each cue
    - Misleading cues do not affect the others
  - Problem is divided into sub-problems
  - Not all cues must always be present
    e.g. Confidence-based Cue Integration [Pronobis&Caputo'07]

SVM-DAS
Integration Function

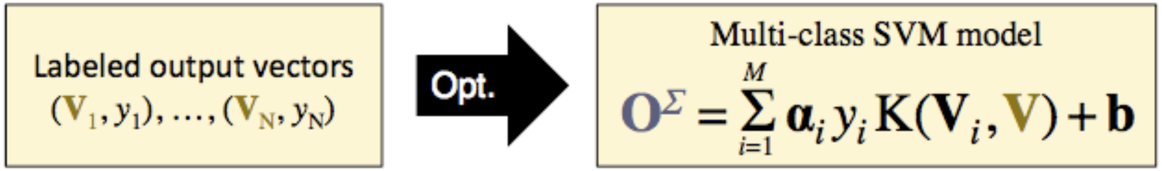
- Simple linear accumulation (G-DAS, [Pronobis&Caputo'07])

$$\mathbf{O}^{\Sigma} = a_1 {*} \mathbf{O}^1 + a_2 {*} \mathbf{O}^2 + \ldots + a_P {*} \mathbf{O}^P$$

Integrated output vector

Output vector for cue no. P

- SVM-DAS
  - All outputs in one vector $\mathbf{V} = [\mathbf{O}^1, \mathbf{O}^2, \ldots, \mathbf{O}^P]^{\mathrm{T}}$
  - Multi-class SVM trained on labeled output vectors



Labeled output vectors $(\mathbf{V}_1, y_1), \ldots, (\mathbf{V}_N, y_N)$

Opt.

Multi-class SVM model

$$\mathbf{O}^{\Sigma} = \sum_{i=1}^{M} \boldsymbol{\alpha}_i y_i \mathrm{K}(\mathbf{V}_i, \mathbf{V}) + \mathbf{b}$$

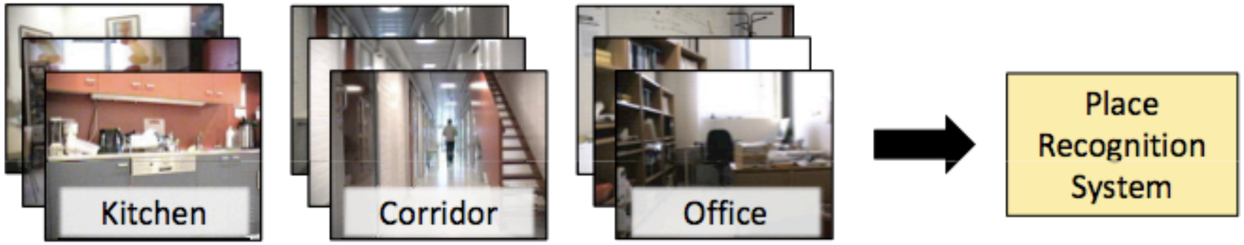
  - Kernel determines the complexity (linear, non-linear)
  - Final decision as in standard multi-class SVM

SVM-DAS vs. G-DAS

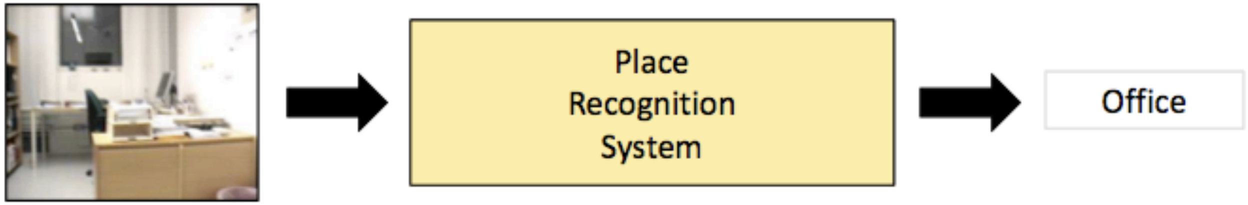
- G-DAS
  - Simple, linear function
  - Single weight for all outputs
  - Parameters found by extensive search
  - Integrates outputs of models of the same type
- SVM-DAS
  - Complex (non-linear) function
  - Each output treated separately
  - Model inferred from training data by optimization algorithm
  - Able to integrate outputs of different types of models
    - Can give correct results even if all single cues are wrong

The Place Recognition System Overview

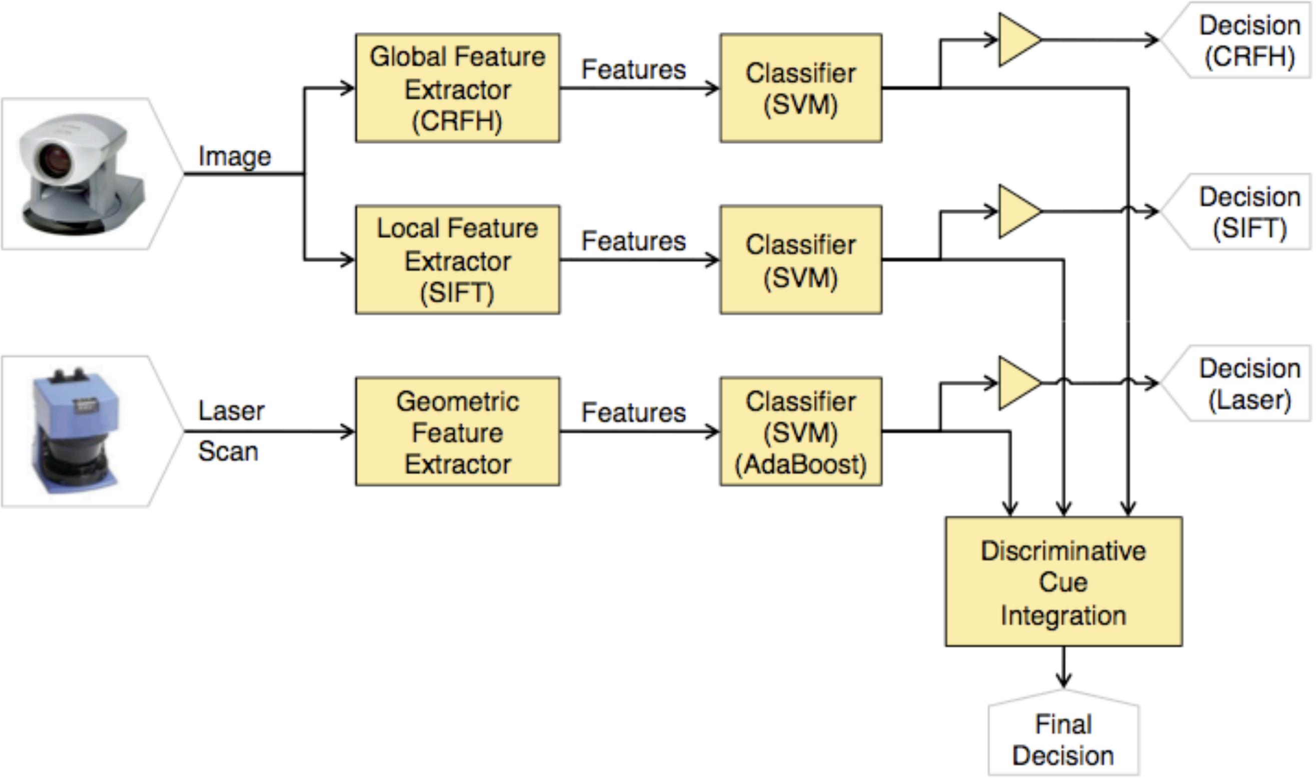
- Fully supervised approach [Pronobis et al. ’06 ’07]
- Training:



- Recognition:



The Place Recognition System Architecture



Image

Laser Scan

Global Feature Extractor (CRFH)

Local Feature Extractor (SIFT)

Geometric Feature Extractor

Features

Features

Features

Classifier (SVM)

Classifier (SVM)

Classifier (SVM) (AdaBoost)

Decision (CRFH)

Decision (SIFT)

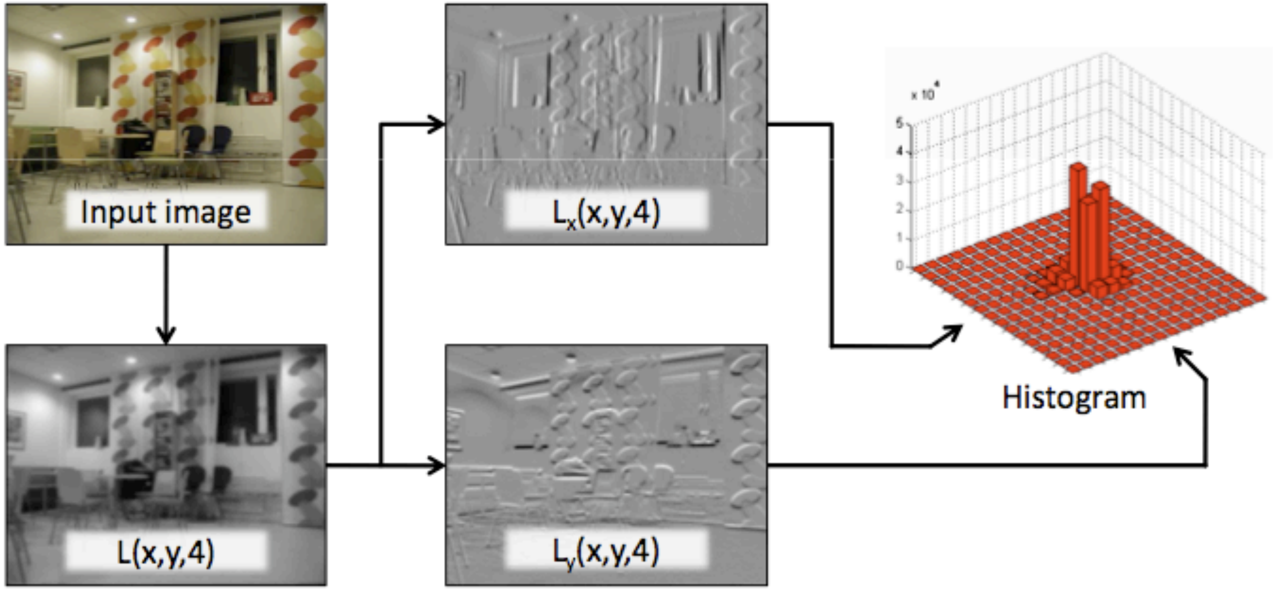
Decision (Laser)

Discriminative Cue Integration

Final Decision

The Place Recognition System
Global Visual Features

- High dimensional Composed Receptive Field Histograms (CRFH) [Linde & Lideberg ‘04]



Input image

$L(x,y,4)$

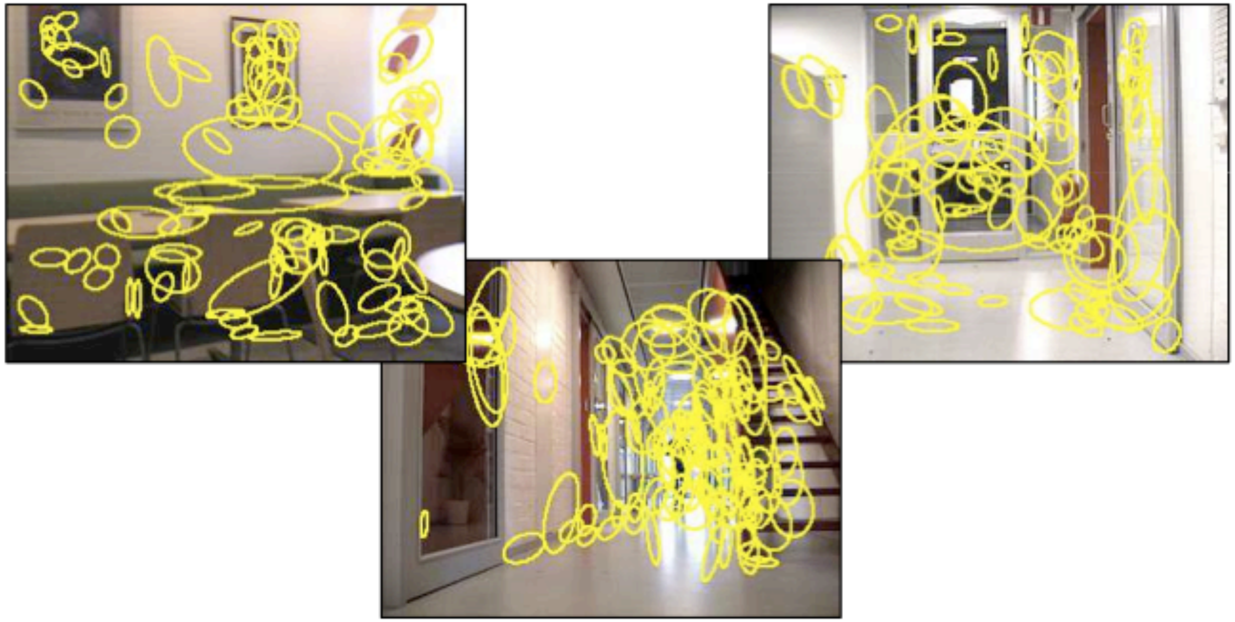
$L_x(x,y,4)$

$L_y(x,y,4)$

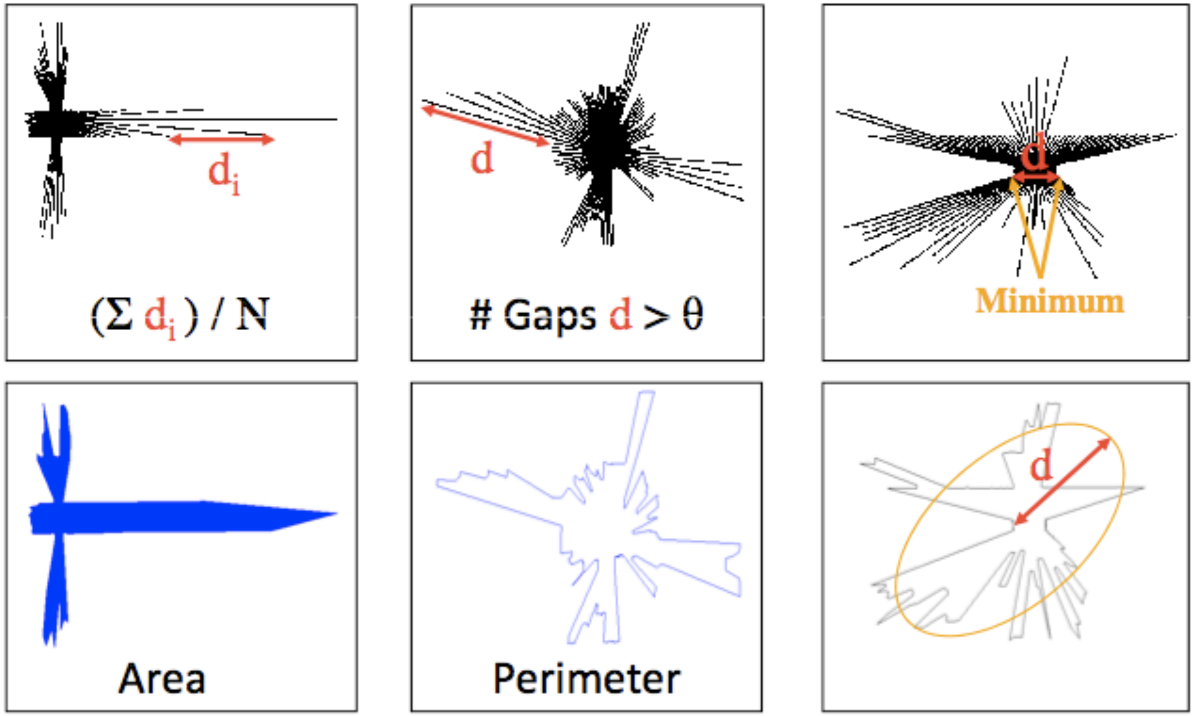
Histogram

The Place Recognition System
Local Visual Features

- Affine, scale-invariant DoG interest-point detector [Rothganger *et al.* ‘06] and SIFT descriptor [Lowe ‘04]



The Place Recognition System
Geometrical Laser-based Features



[Martínez Mozos *et al.* ‘07] with AdaBoost

Experimental Setup
The IDOL2 Database

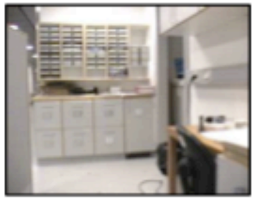
- Five rooms of different funtionality

One-person office

Corridor

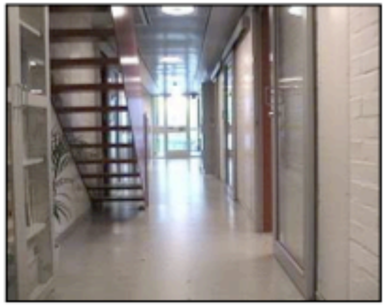
Two-persons office

Kitchen

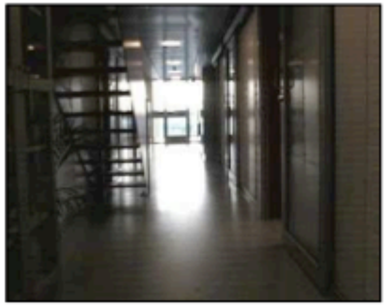


Printer area

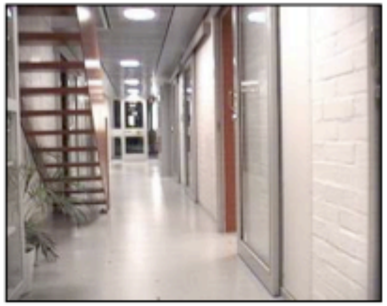
- Three illumination settings over three weeks



Cloudy



Sunny



Night

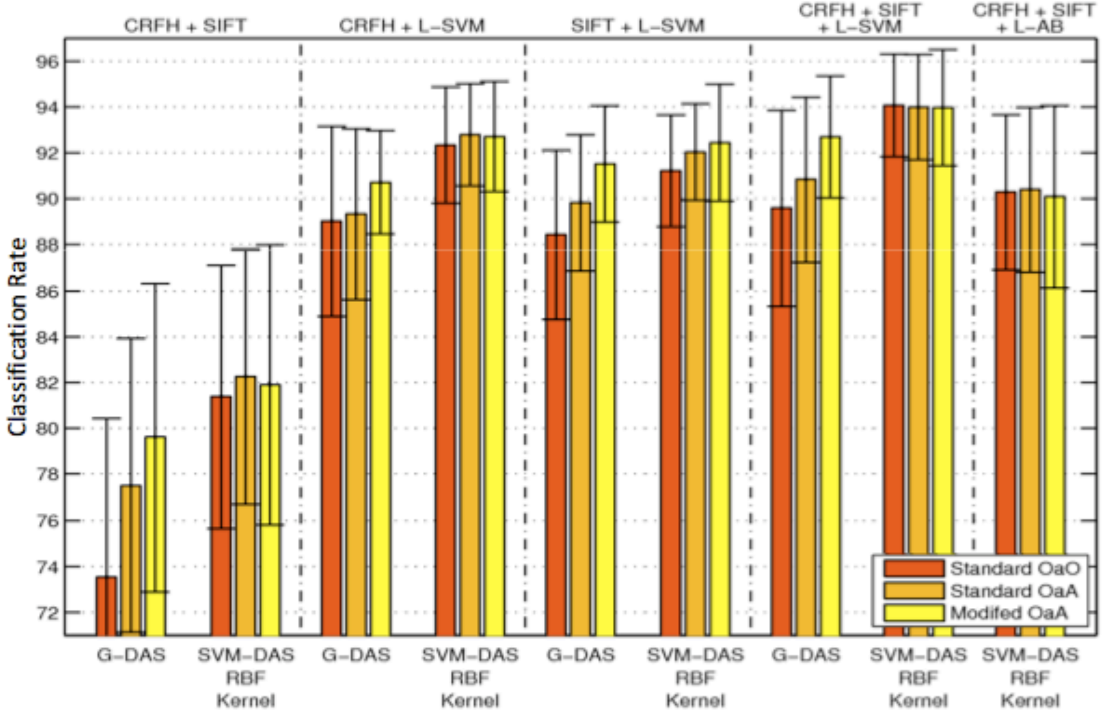
- Repeated after 6 months

Experimental Procedure

- Four sets of experiments
  - Exp. 1 – Stable illumination, close in time
  - Exp. 2 – Varying illumination, close in time
  - Exp. 3 – Stable illumination, distant in time
  - Exp. 4 – Varying illumination, distant in time
- Each set evaluates
  - Four single-cue models
    - SVM model trained on CRFH
    - SVM model trained on SIFT
    - SVM model trained on laser range features (L-SVM)
    - AdaBoost model trained on laser range features (L-AB)
  - Both cue integration schemes (G-DAS, SVM-DAS)

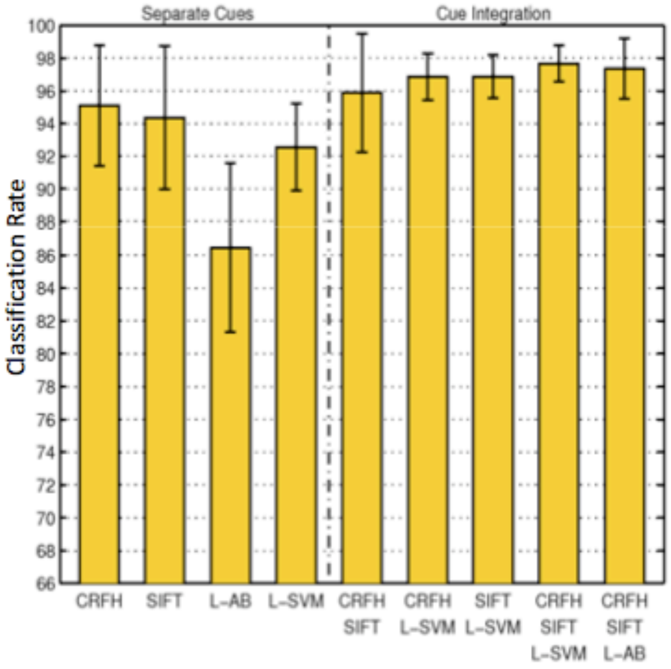
Results
Comparison of Cue Integration Methods

- Varying illumination, distant in time

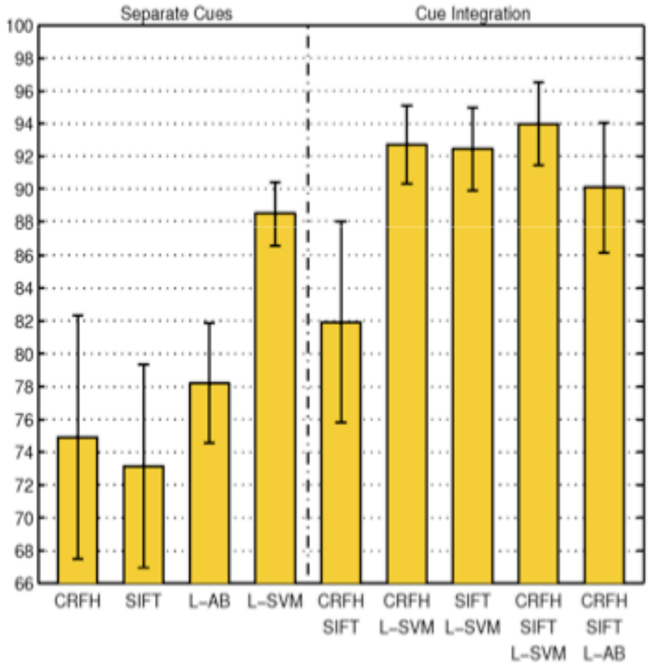


Results
Single Cue VS Multiple Cues

- □ Similar ill., close in time

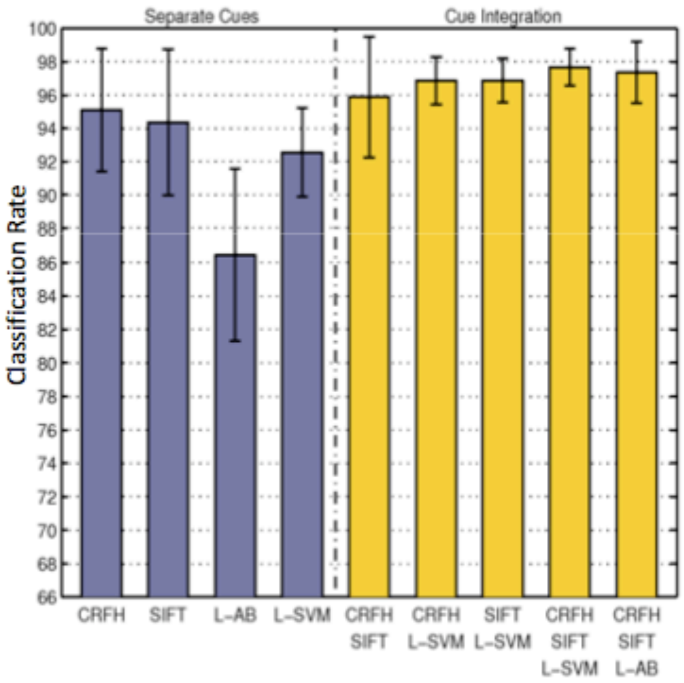


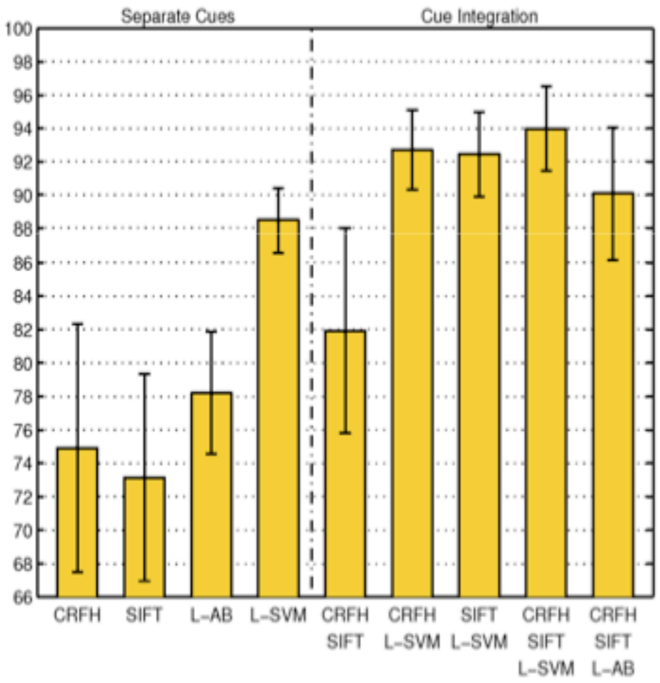
- □ Varying ill., distant in time



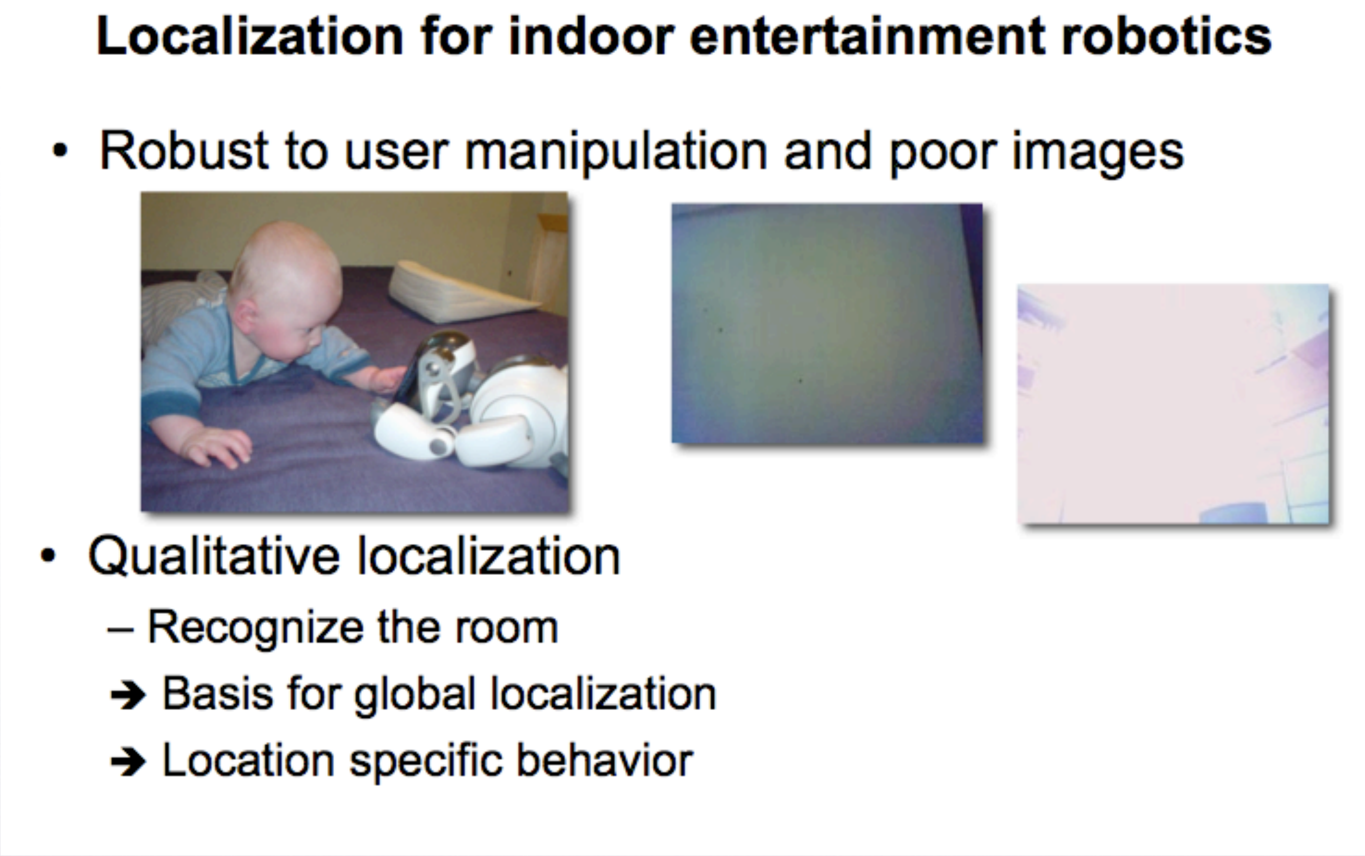
Results
Single Cue VS Multiple Cues

- Similar ill., close in time
- Varying ill., distant in time





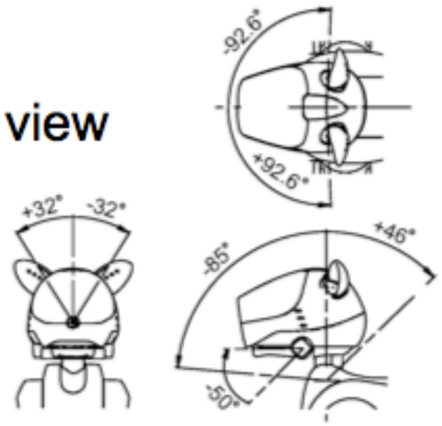
D. Filliat. *A visual bag of words method for interactive qualitative localization and mapping*. Proc ICRA 2007.



Localization for indoor entertainment robotics

- Robust to user manipulation and poor images
- Qualitative localization
  - – Recognize the room
  - ➔ Basis for global localization
  - ➔ Location specific behavior

- Vision only, standard camera
  - Affordable sensor, no panoramic view
  - ➔ Search for information



- No temporal coherence
  - User manipulation of the robot
  - No position tracking
  - ➔ "One shot" localization

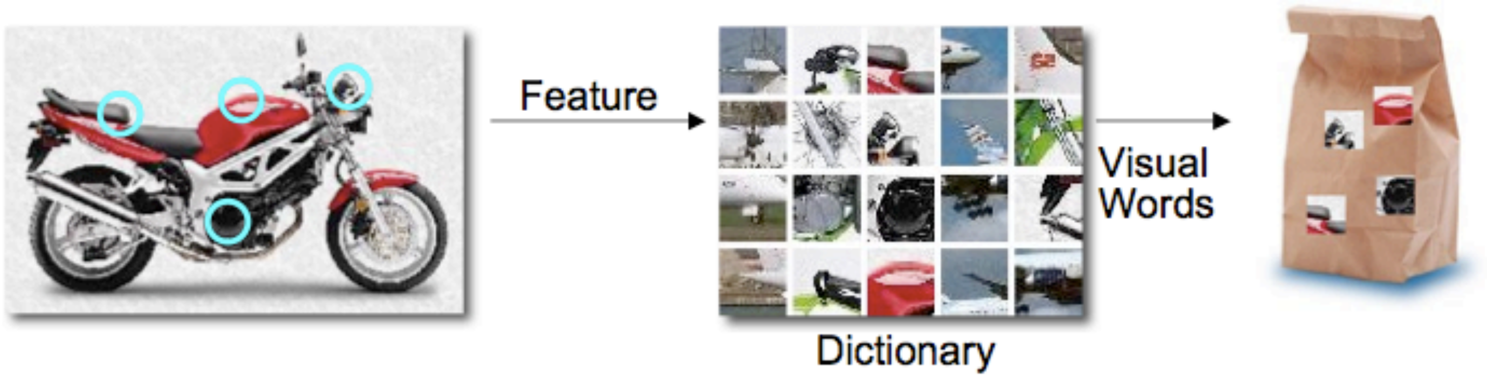
Goal : recognize room from images

➔ Image categorization

- Map-learning
  - Not a separate process (SLAM)
  - With discontinuous user supervision

Goal : Infer category from image *(Csurka et al. 2004)*

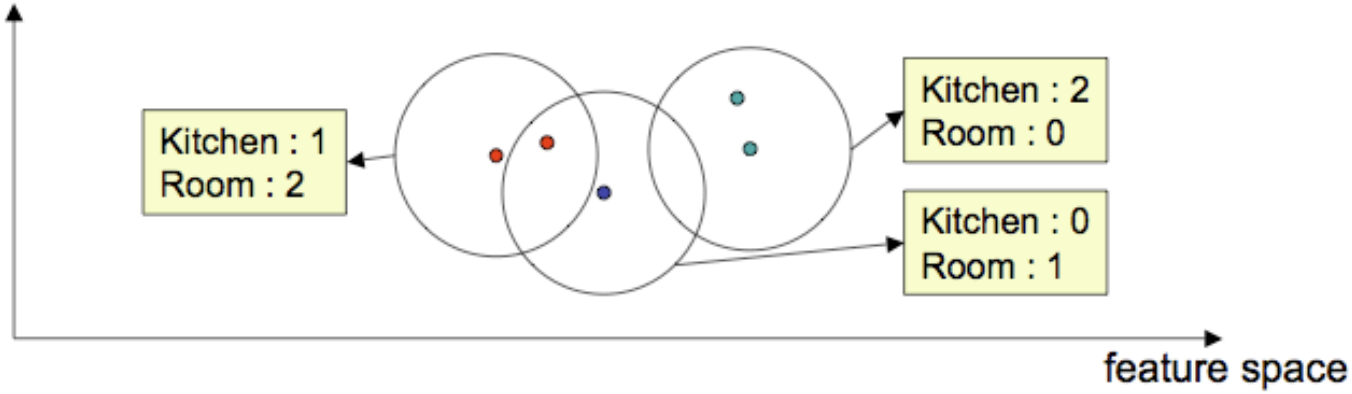
Image representation : set of unordered local “words”



Dictionary : quantization of feature space **OFFLINE**

Categorization : classifier built on bag of words **OFFLINE**

- Incremental training
  - Dictionary construction : incremental nearest neighbor



  - Classifier training :
    - Process new examples
    - Add new categories

→ *voting method*

- Problem structure :

Some images belong to several categories

All images taken from a position are in the same category

ACTIVE LOCALIZATION

Take only informative images

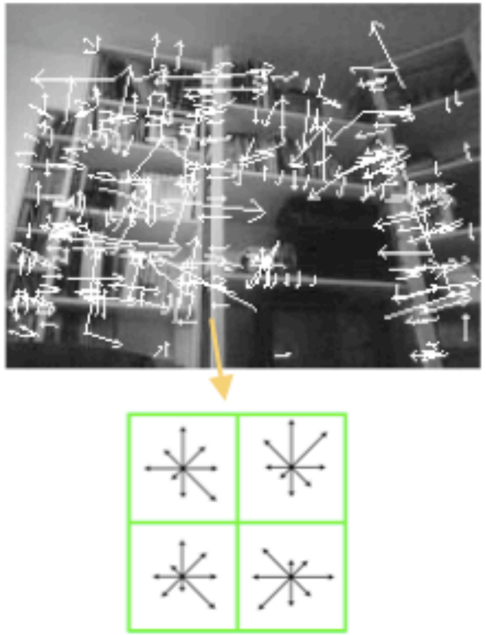
Take new images until localization is confident

- Discontinuous user supervision
  - Active learning : learn when errors are reported
    - ➔ less training data
    - ➔ long term stability

- Feature used
  - Depend on the environment
  - Multiple feature integration through the voting method
    - ➔ shape (SIFT), color (H hist), texture (V hist)

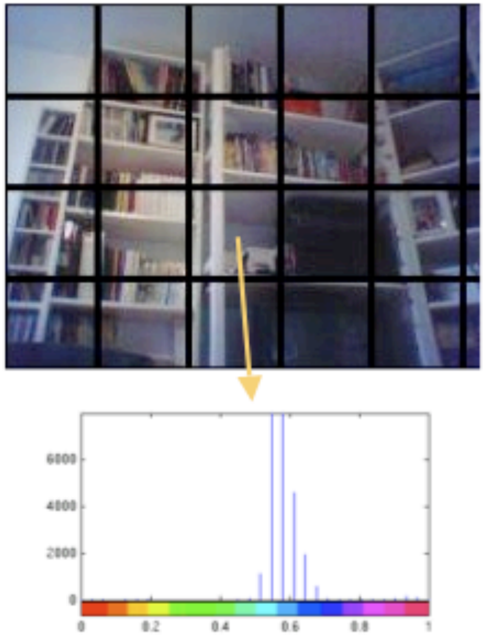
- Features

SIFT



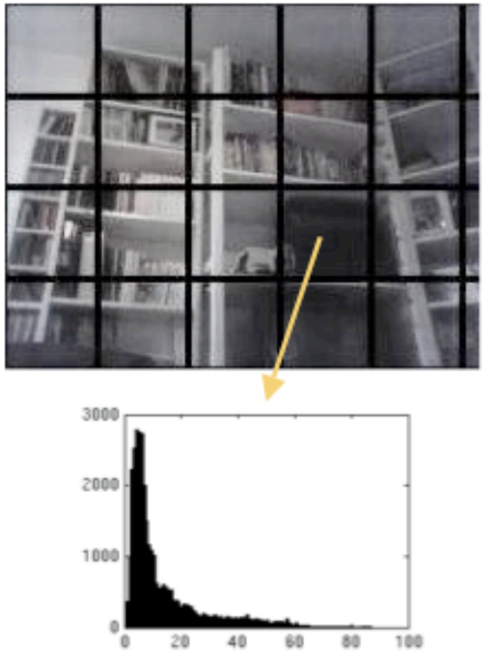
Dim 128

H histograms



Dim 16

V histograms

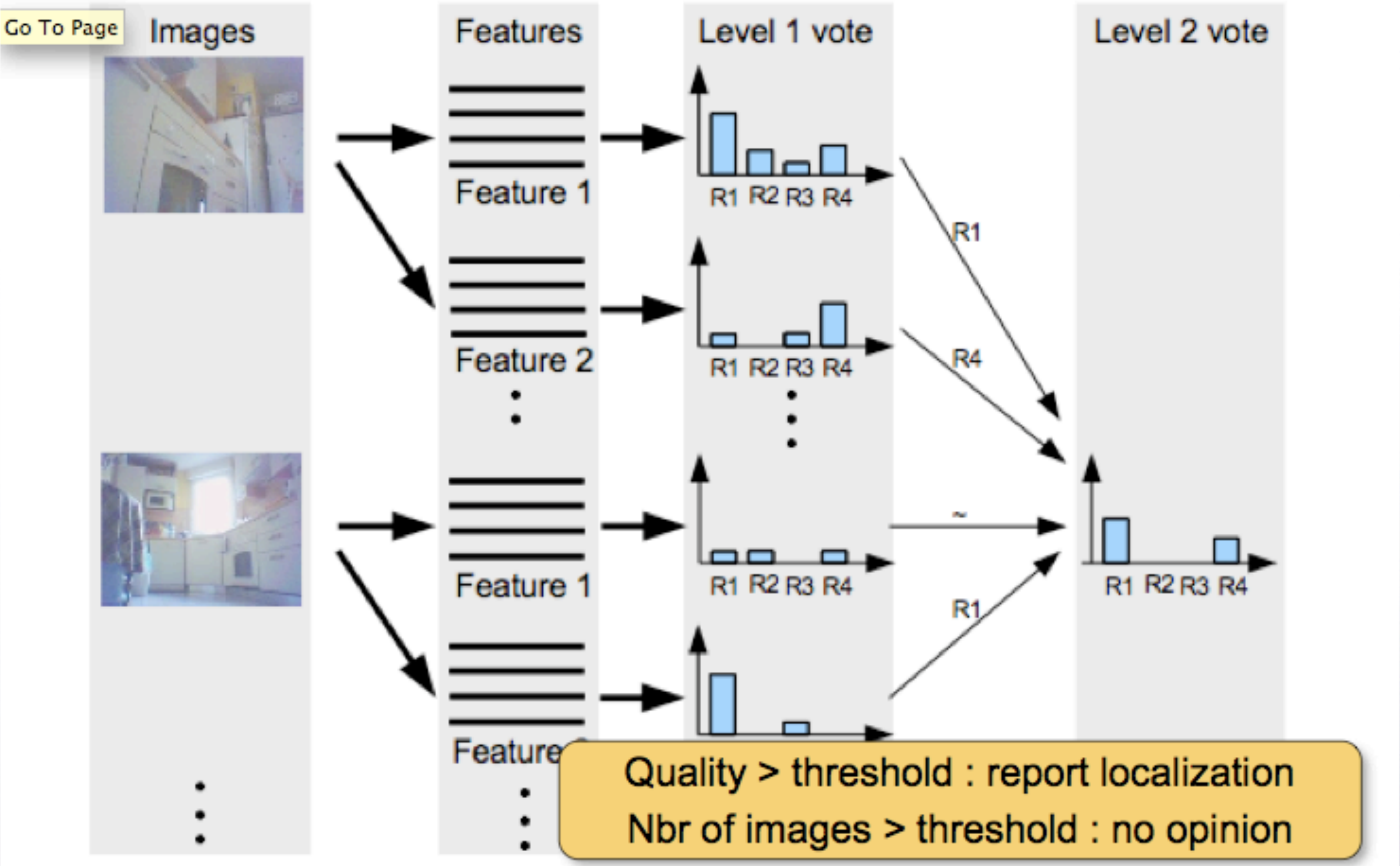


Dim 16

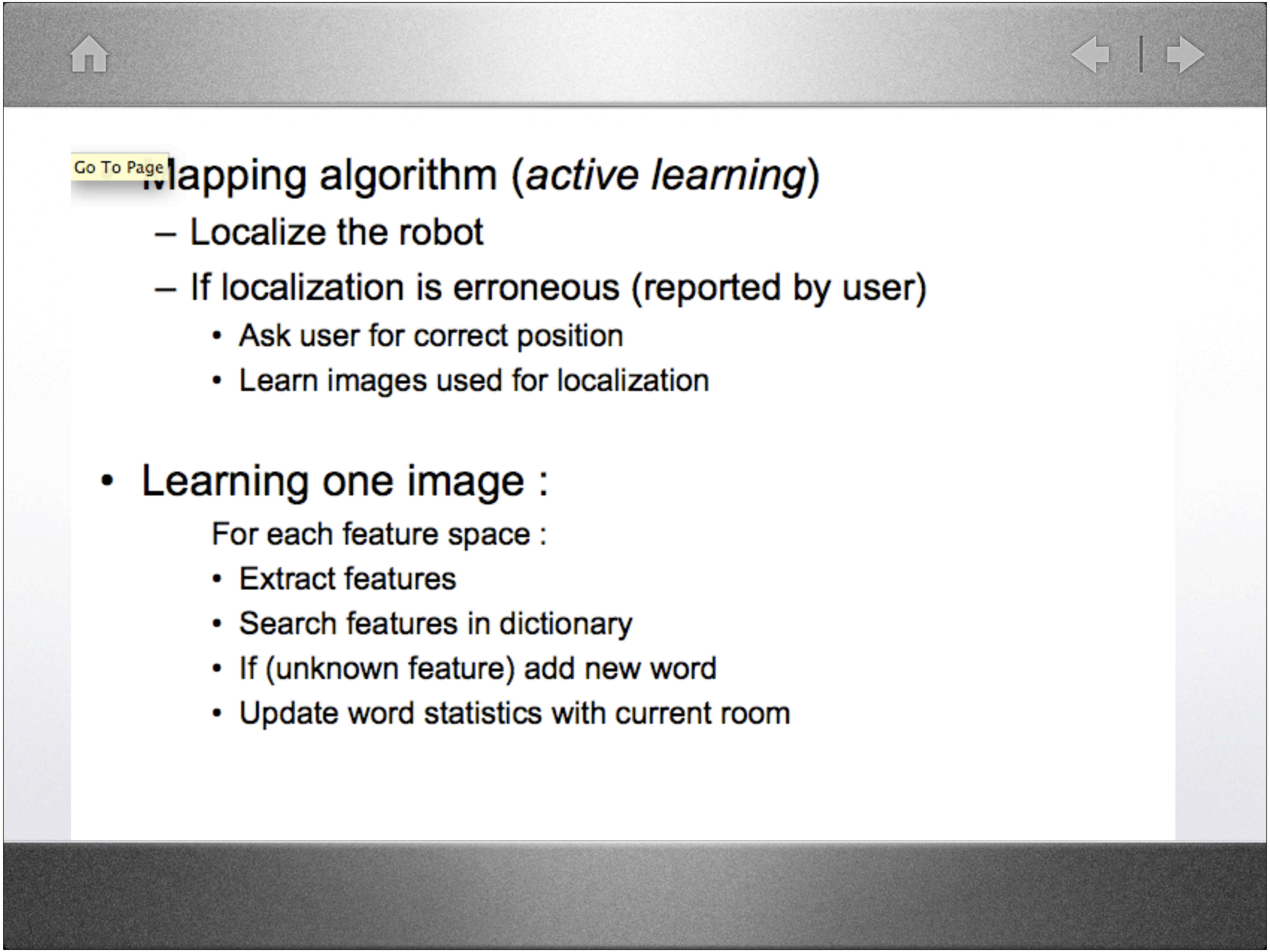
χ^2 distance $$\|H_1 - H_2\|^2 = \sum_i \frac{(H_{1,i} - H_{2,i})^2}{H_{1,i} + H_{2,i}}$$

- Map :
  - Dictionary for each feature space
  - For each word : number of times seen in each room
- Active localization :
  - 2 level voting scheme
  - First level : select informative images
  - Second level : estimate need for new information

$$quality = \frac{n_{Winner} - n_{Second}}{\sum_i n_i}$$



Quality > threshold : report localization
Nbr of images > threshold : no opinion



- Mapping algorithm (*active learning*)
  - Localize the robot
  - If localization is erroneous (reported by user)
    - Ask user for correct position
    - Learn images used for localization

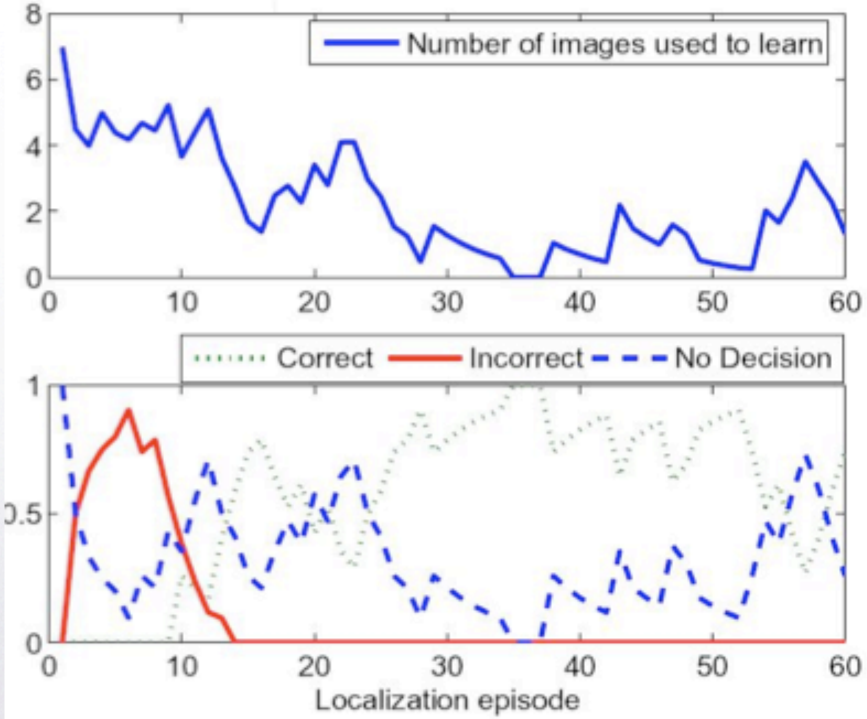
- Learning one image :

  For each feature space :

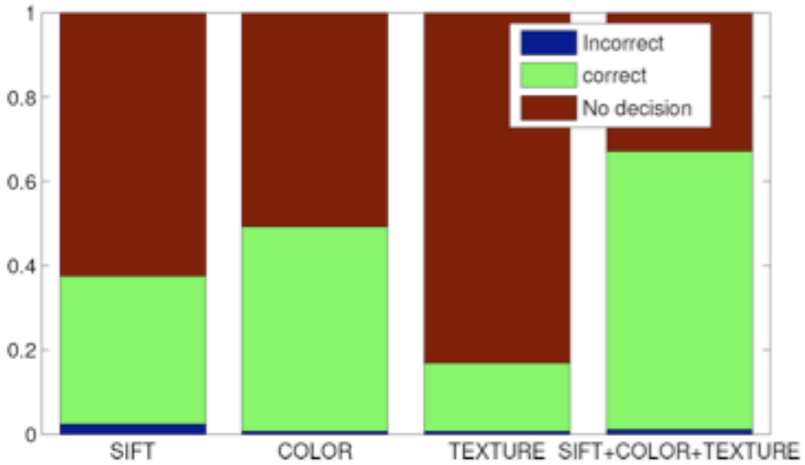
  - Extract features
  - Search features in dictionary
  - If (unknown feature) add new word
  - Update word statistics with current room

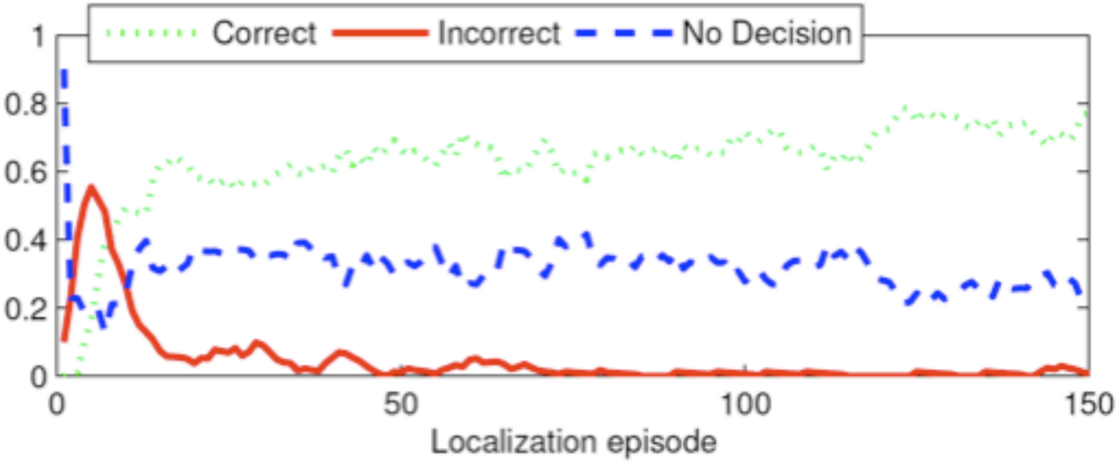
Go To Page

On-line Learning



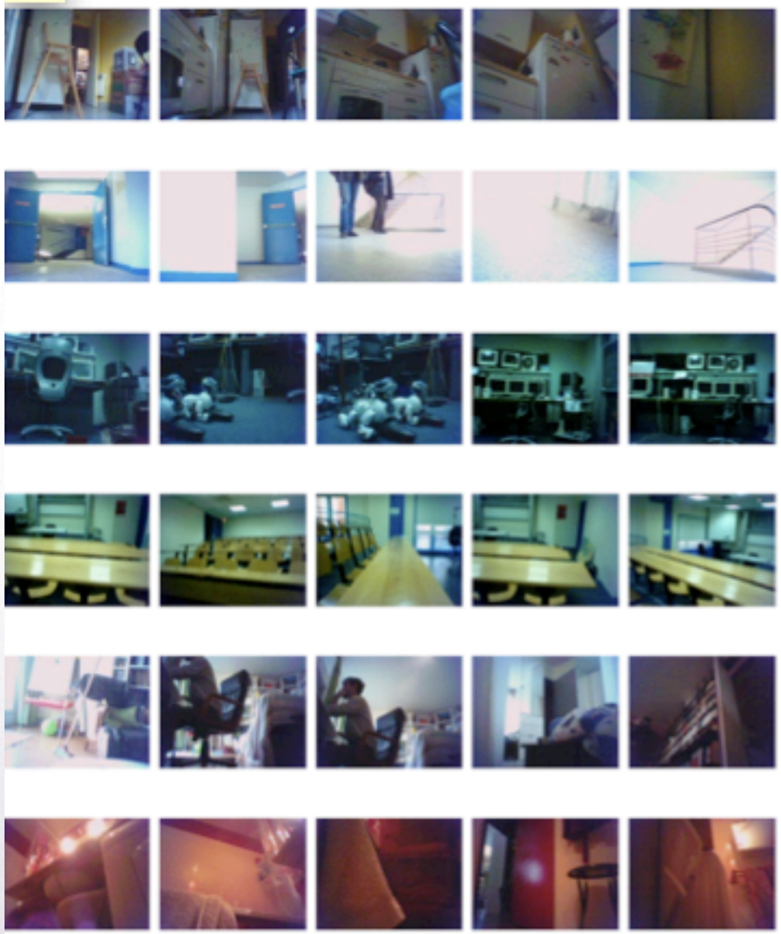
- Database results (5000 images in 10 rooms)
  - Random environments (3 - 7 rooms)



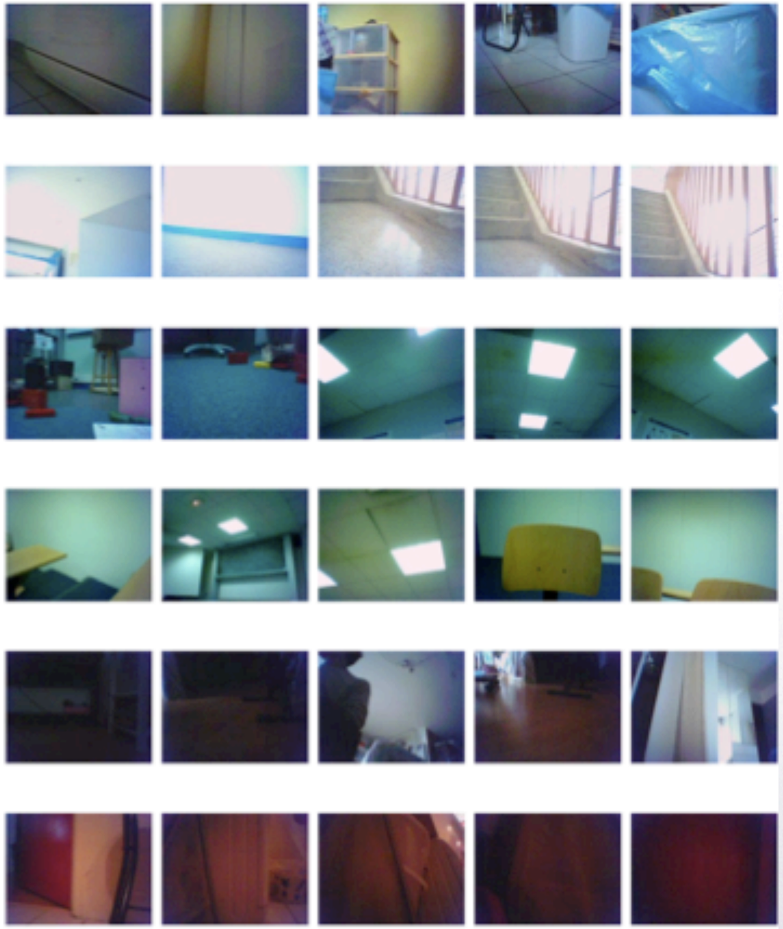


Back

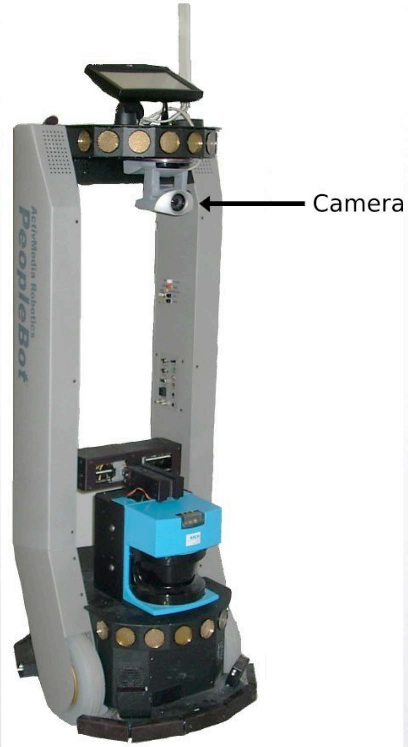
High quality



Low quality



L. Jie, A. Pronobis, B. Caputo, P. Jensfelt. *Incremental learning for place recognition in dynamic environments*. Proc IROS 2007.

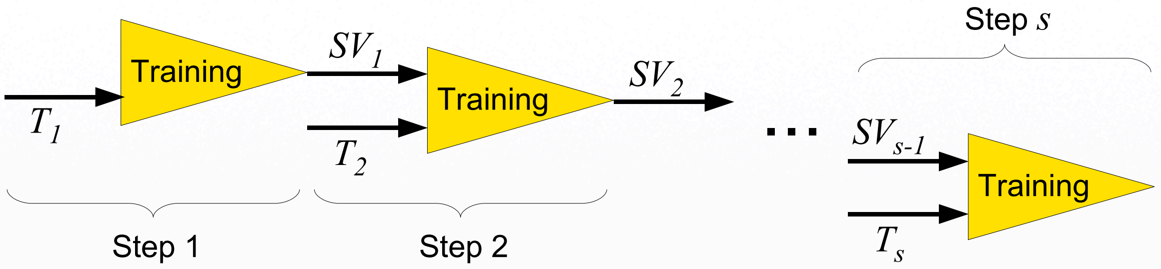


Minnie

Minnie

SVM Incremental Learning Extensions

- Fixed-partition technique [Syed. et al. IJCAI'99]



- Error-driven technique [Domeniconi et al. ICDM'01]
- Memory-controlled Incremental SVM [Pronobis & Caputo, ICVW06]

Memory-controlled Incremental SVM [Pronobis&Caputo, ICVW06]

- SVM Reduction Algorithm [Downs. et al. JMLR'02]

Discover the linear relationship between support vectors and discard those support vectors which are linearly dependent.

$$f(x)=\sum_{i=1}^{r}\alpha_i y_i K(x,x_i)+\sum_{j=r+1}^{n}\alpha_j y_j \sum_{i=1}^{r} c_{ij} K(x,x_i)+b$$

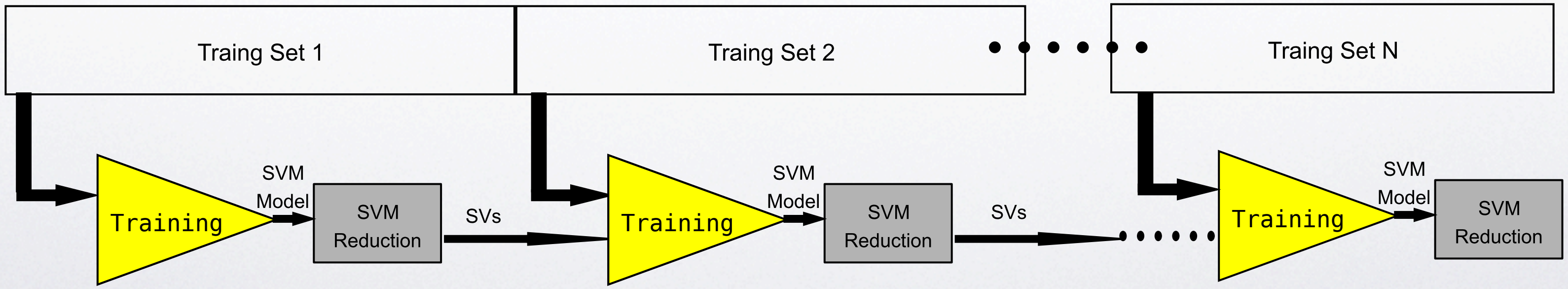
$$f(x)=\sum_{i=1}^{r}\tilde{\alpha}_i y_i K(x_i,x)+b \qquad \tilde{\alpha}_i=\alpha_i\left(1+\sum_{j=r+1}^{n}\frac{\alpha_j y_j c_{ij}}{\alpha_i y_i}\right)$$

n-r kernel evaluation and support vectors to store

Memory-controlled Incremental SVM [Pronobis&Caputo, ICVW06]

- SVM Reduction Algorithm [Downs. et al. JMLR'02]
- Incremental Extension

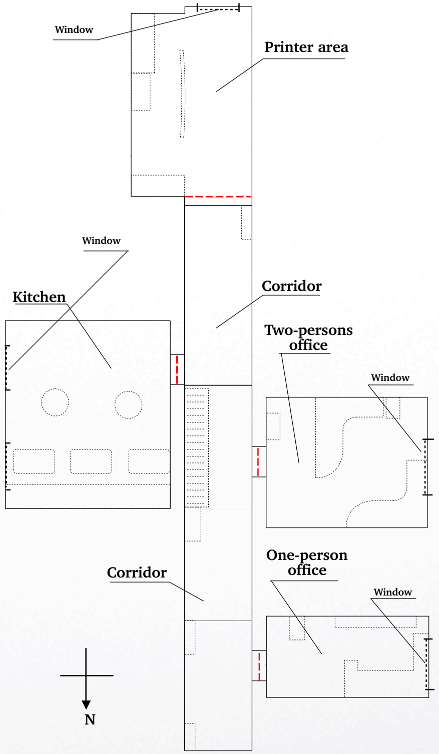
  Combine the reduction algorithm with the incremental techniques, and apply the reduction scheme at each incremental step.



The IDOL Database

Available at http://cogvis.nada.kth.se/IDOL

The database contains 24 image sequences acquired using two robot platforms under three different illumination conditions (sunny, cloudy and night), across a span time of six months. The acquisition was performed at an indoor laboratory environment, consisting of five rooms with different functionality.



One-person office

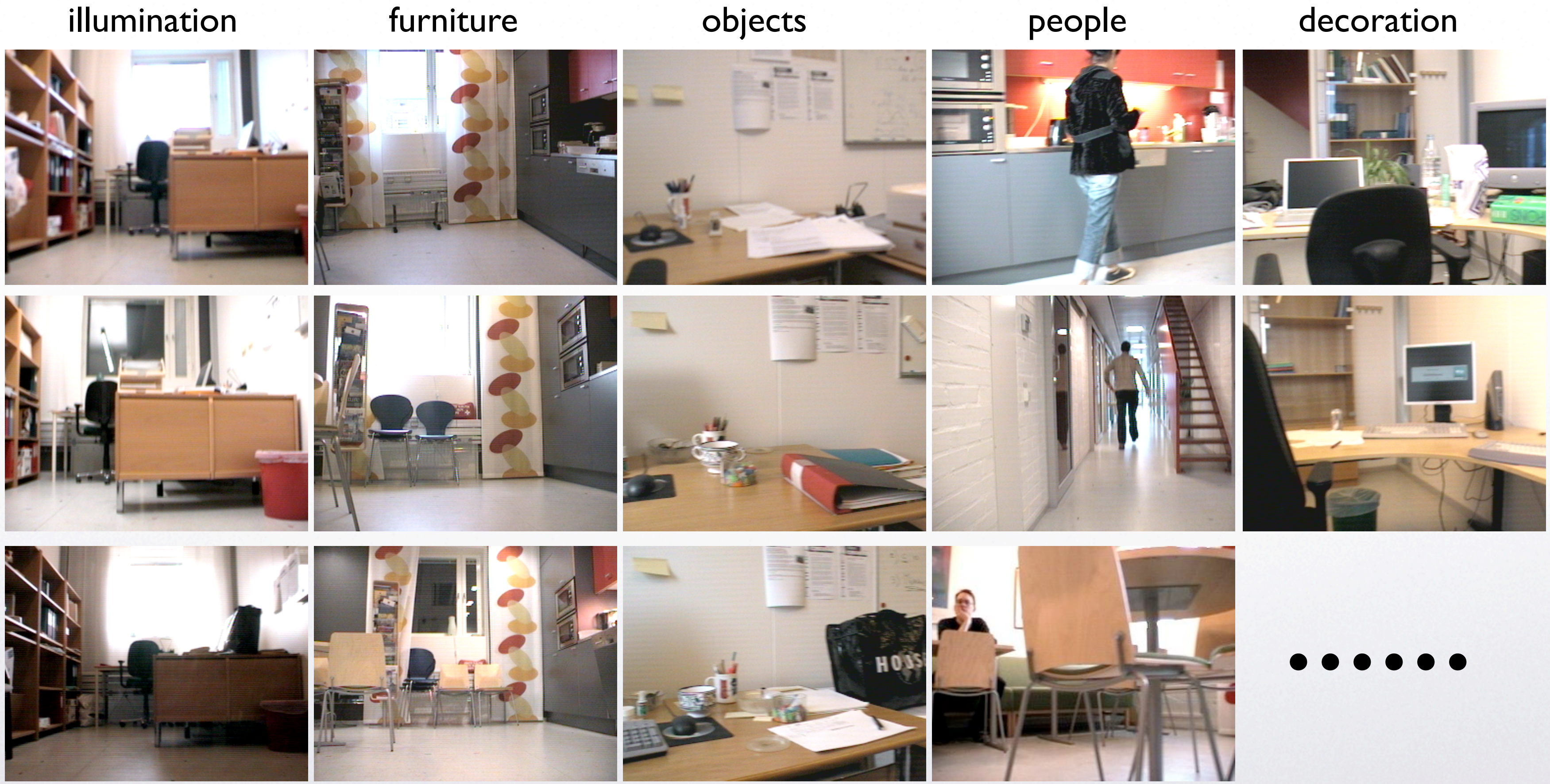
Corridor

Two-persons office

Kitchen

Printer Area

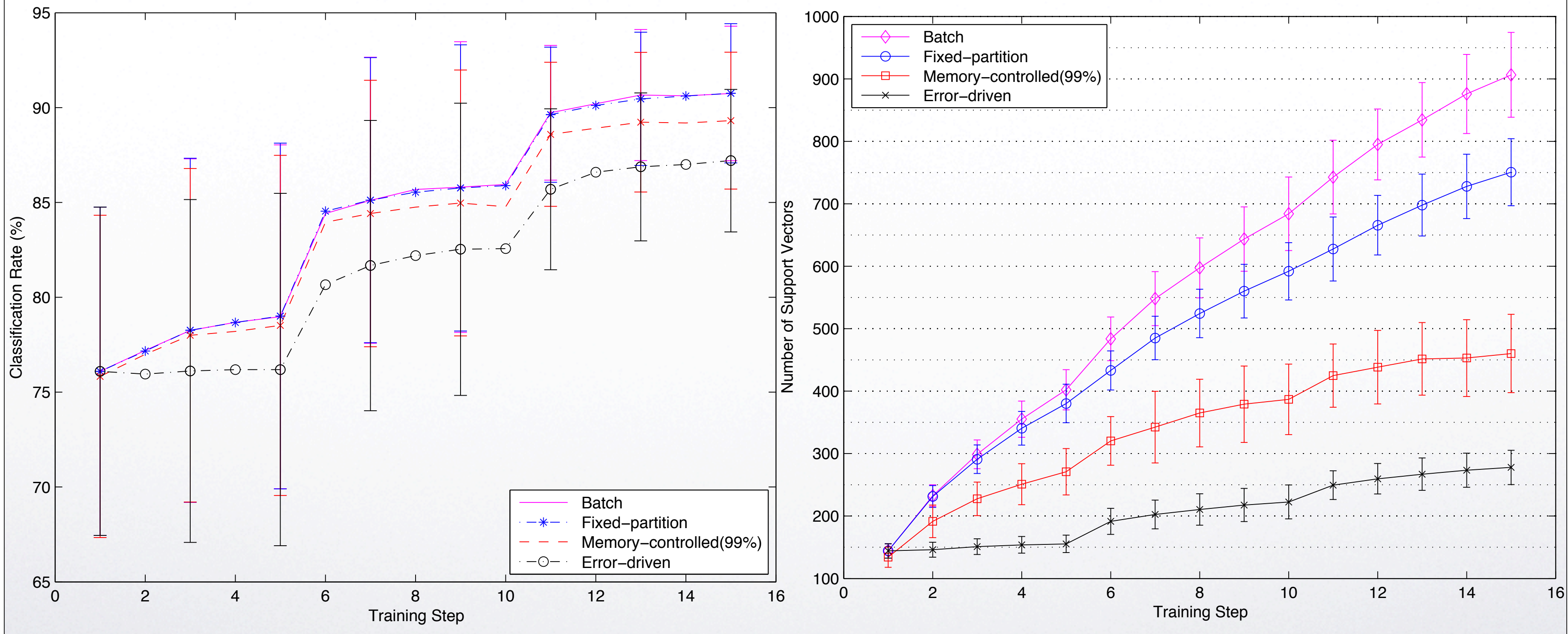
Environment Variations Captured in IDOL



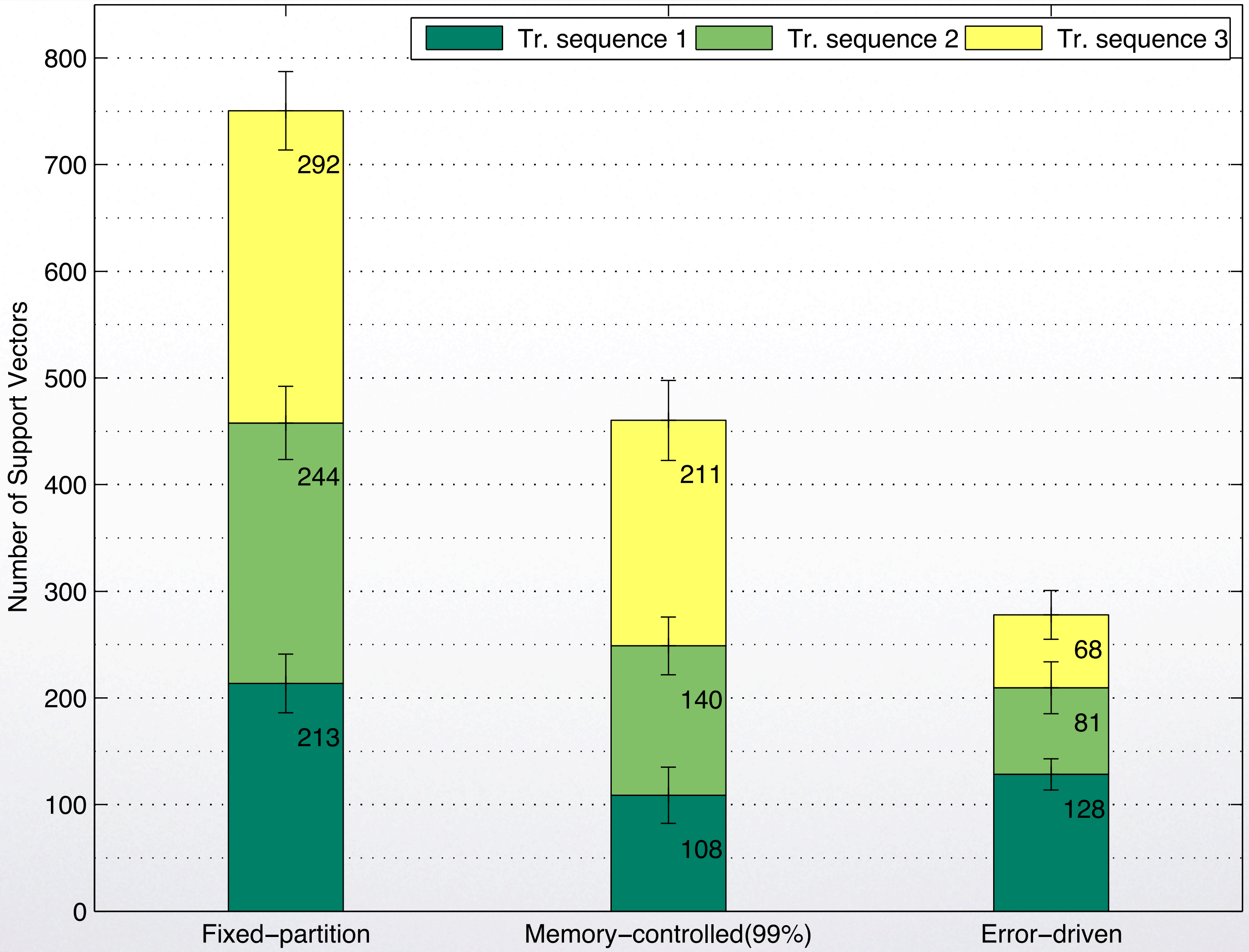
illumination furniture objects people decoration

......

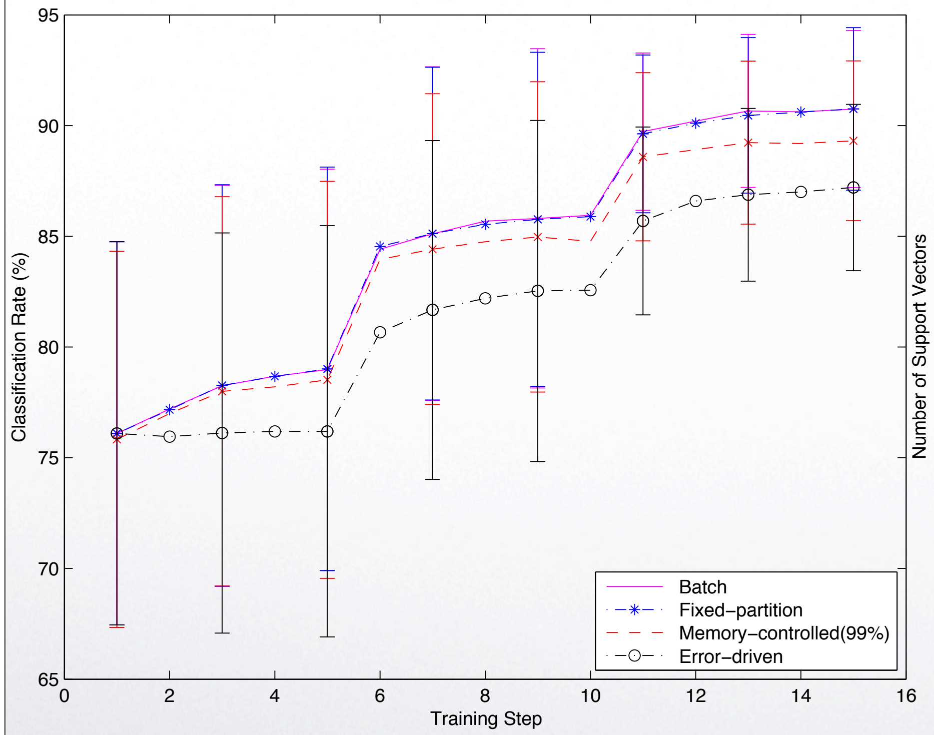
Scenario 1 : Results A

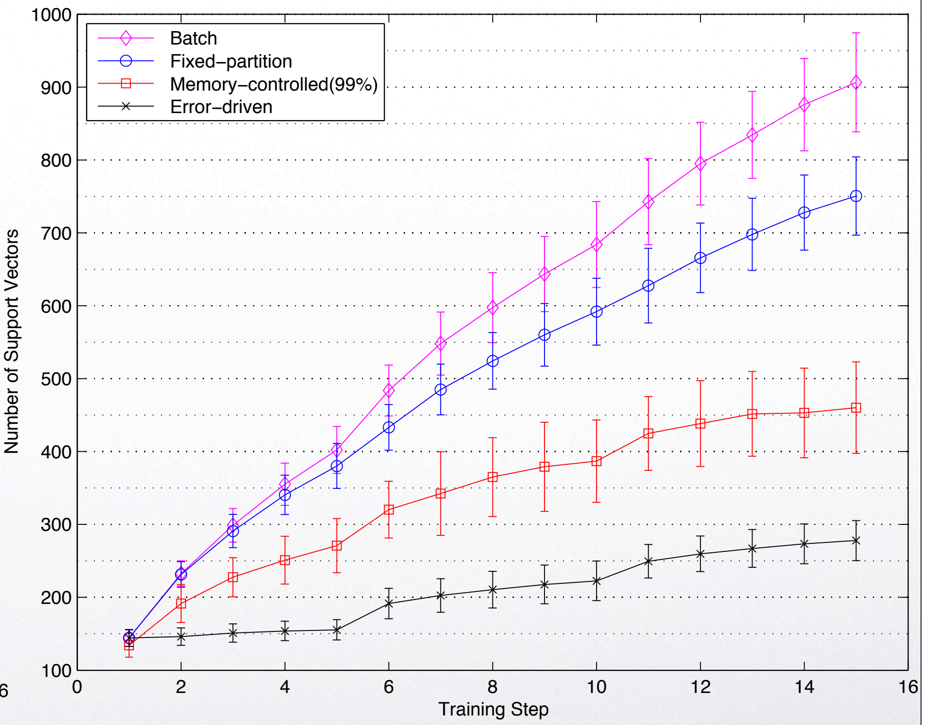


Scenario 1 : Results B

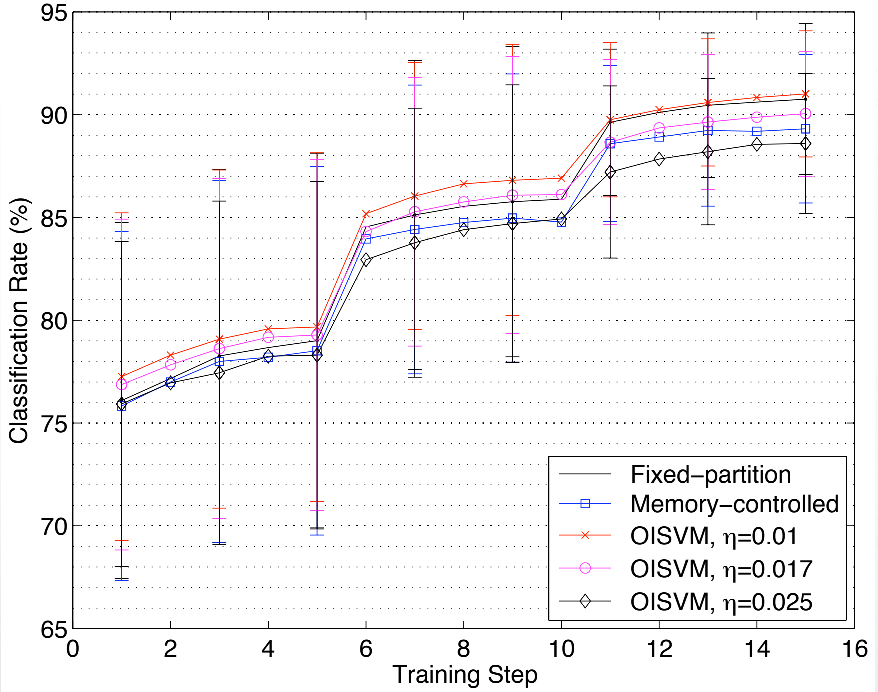


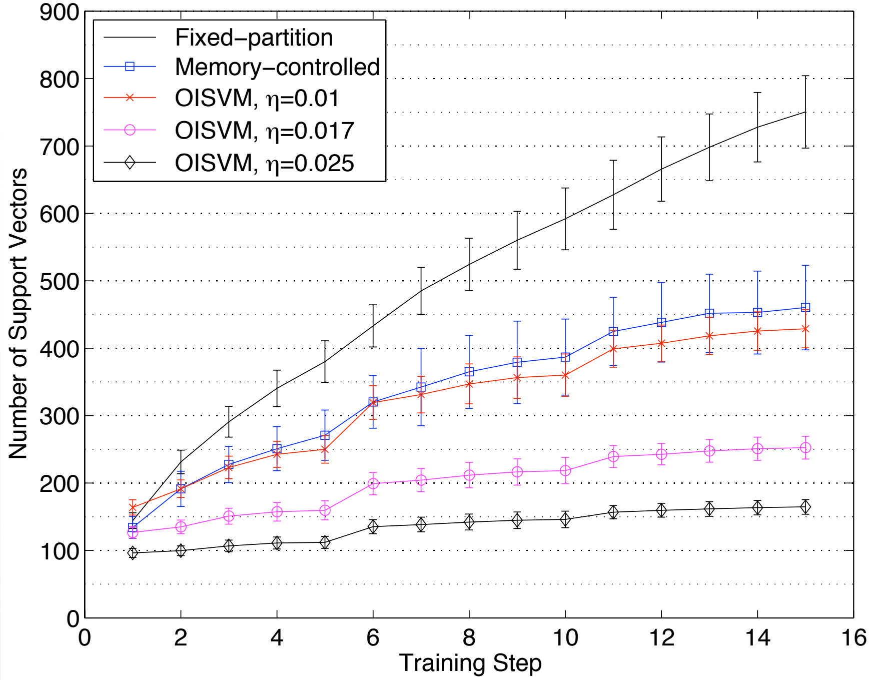
Open Problem: memory is not guaranteed to be bounded!





(Partial) Solution: check linear independence before updating the solution [Orabona et al, BMVC07]





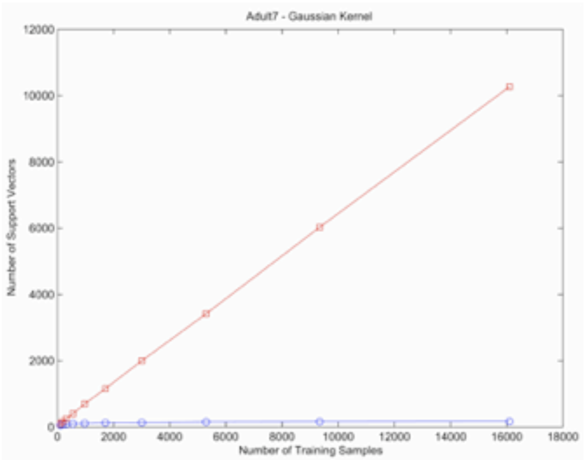
15 min break!

F. Orabona, C. Castellini, B. Caputo, J. Luo, G. Sandini. Online incremental support vector machines for place recognition. Proc BMVC 2007.

- Follows the L. Jie et al IROS 2007, and focuses on how to bound the memory growth without any compromise on performance
- *Contribution*: online SVM with bounded memory growth in the test model

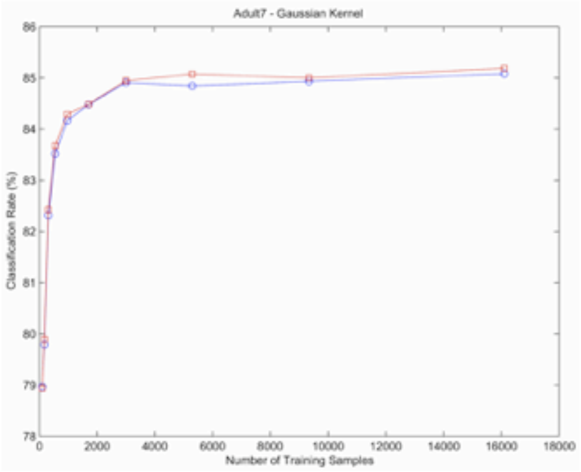
Our approach

- Modify the SVM to
  - Learn incrementally from the samples
  - Produce a solution that is bounded in memory
  - Retain as much as possible the good performances



Our approach

- Modify the SVM to
  - Learn incrementally from the samples
  - Produce a solution that is bounded in memory
  - Retain as much as possible the good performances



More mathematically…

- Given two set of samples we find a separating hyperplane $f(\mathbf{x})=\mathbf{w}\cdot\Phi(\mathbf{x})+b$ solving a constrained optimization problem

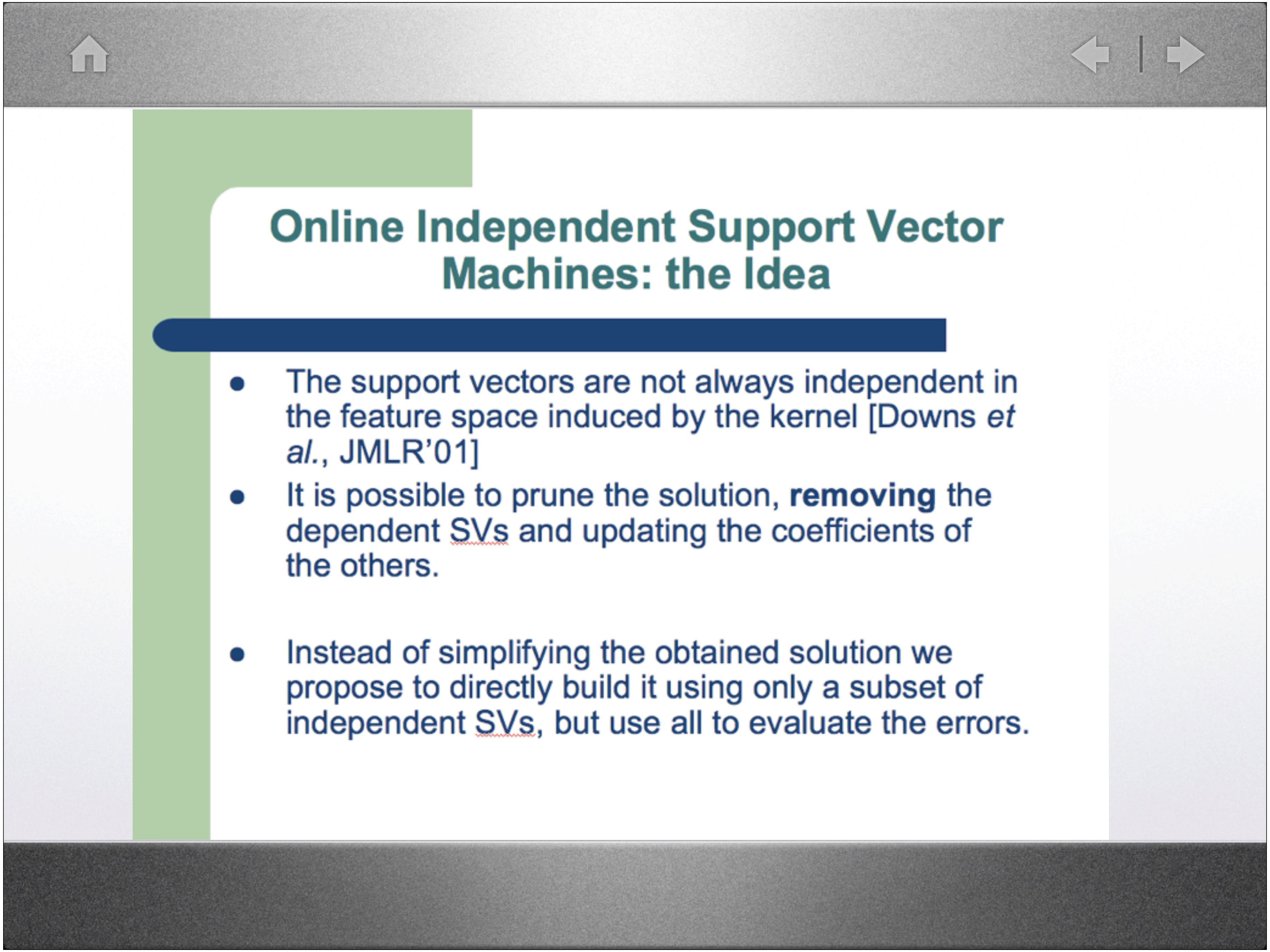
$$\min_{\mathbf{w}}\left(\|\mathbf{w}\|^2+C\sum_{i=1}^{l}L(\xi_i)\right)$$

- The solution is always written as

$$f(\mathbf{x})=\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x},\mathbf{x}_i)$$

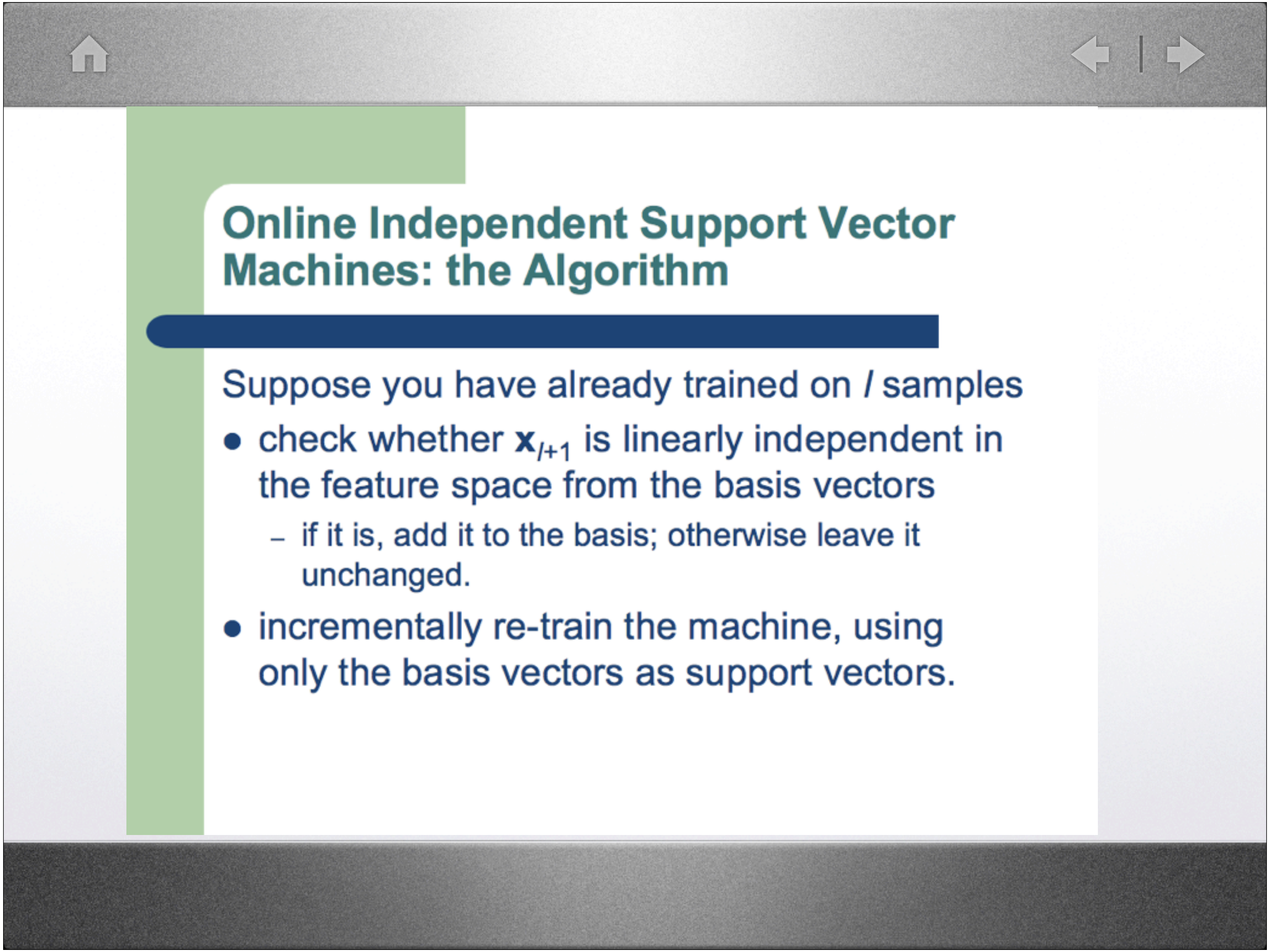
- Those samples for which the coefficients α_i are non-zero are called Support Vectors.

- Number of SVs goes to infinity -> testing time goes to infinity!!!



Online Independent Support Vector Machines: the Idea

- The support vectors are not always independent in the feature space induced by the kernel [Downs *et al.*, JMLR'01]
- It is possible to prune the solution, **removing** the dependent SVs and updating the coefficients of the others.
- Instead of simplifying the obtained solution we propose to directly build it using only a subset of independent SVs, but use all to evaluate the errors.



Online Independent Support Vector Machines: the Algorithm

Suppose you have already trained on l samples

- check whether $\mathbf{x}_{l+1}$ is linearly independent in the feature space from the basis vectors
  - if it is, add it to the basis; otherwise leave it unchanged.
- incrementally re-train the machine, using only the basis vectors as support vectors.

Linear independence check

- How to check to independence in the induced space?

$$\Delta = \min_{\mathbf{d}} \left\| \sum_{j \in \mathrm{B}} d_j \phi(\mathbf{x}_j) - \phi(\mathbf{x}_{l+1}) \right\|^2 =$$

$$= \min_{\mathbf{d}} \left(\mathbf{d}^T \mathbf{K}_{\mathrm{BB}} \mathbf{d} - 2\mathbf{d}^T \mathbf{k} + K(\mathbf{x}_{l+1}, \mathbf{x}_{l+1}) \right) =$$

$$= K(\mathbf{x}_{l+1}, \mathbf{x}_{l+1}) - \mathbf{k}^T \mathbf{K}_{\mathrm{BB}}^{-1} \mathbf{k} \le \eta$$

- Δ=0 means that $\mathbf{x}_{l+1}$ is dependent to the others vectors in set B
- It is possible to demonstrate that if η is greater than zero the number of SVs is finite.

Incremental update

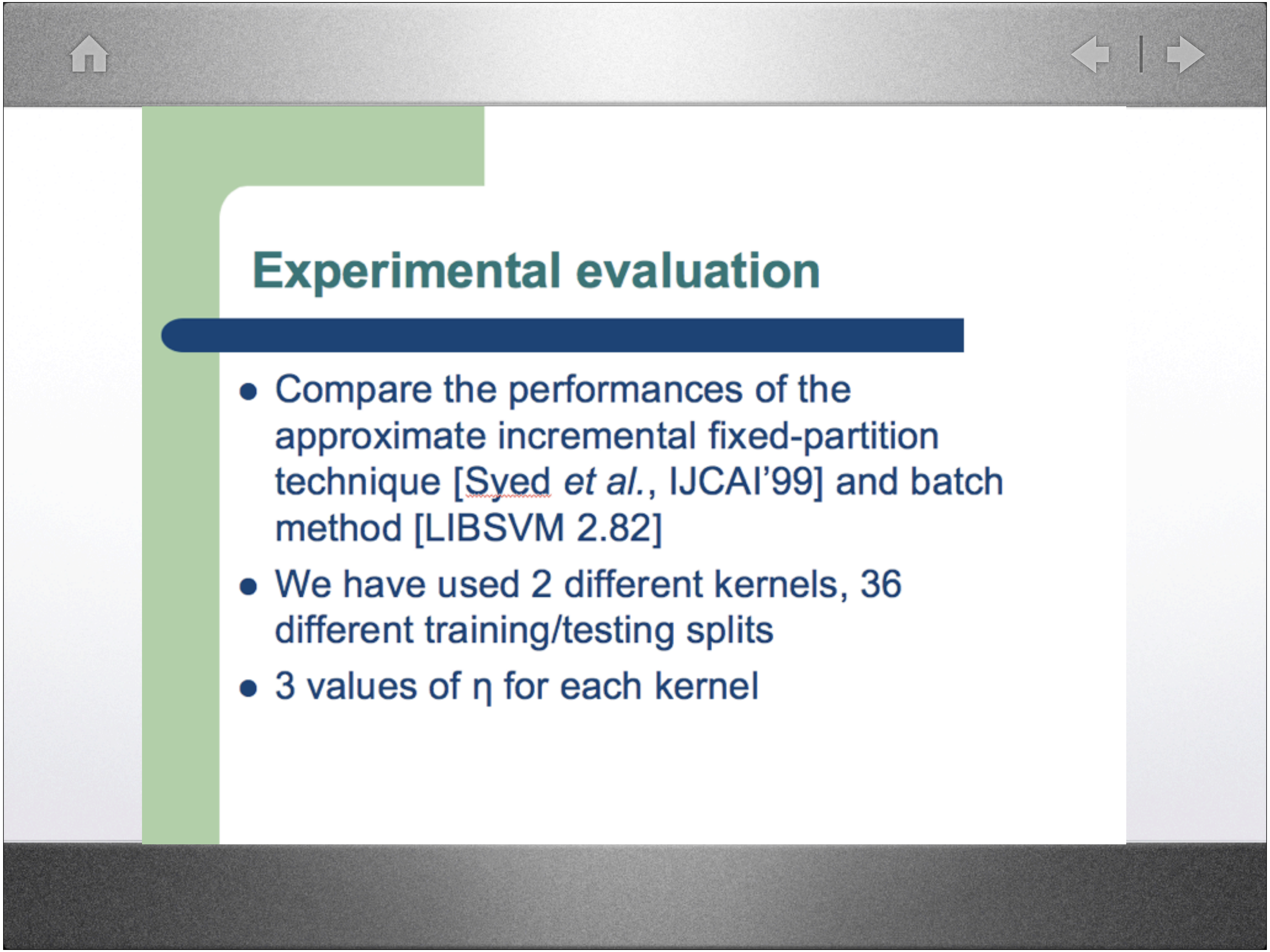
[Keerthi *et al.*, JMLR'06]

$$\min_{\hat{\mathbf{a}}}\left(\frac{1}{2}\hat{\mathbf{a}}^T\mathbf{K}_{DD}\hat{\mathbf{a}}+\frac{1}{2}C\sum_{i=1}^{l}\max\left(0,1-y_i\mathbf{K}_{iD}\hat{\mathbf{a}}\right)^2\right)$$

1) let $I=\{i:1-y_io_i>0\}$ where $o_i=\mathbf{K}_{iB}\hat{\mathbf{a}}$ and $\hat{\mathbf{a}}$ is the vector of optimal coefficients with l training samples; if I has not changed, stop.

2) otherwise, let the new $\hat{\mathbf{a}}$ be $\hat{\mathbf{a}}$ $-\gamma \mathrm{P}^{-1}\mathrm{g}$, where $\mathbf{P}=\mathbf{K}_{BB}+C\mathbf{K}_{BI}\mathbf{K}_{BI}^T$ and $\mathbf{g}=\mathbf{K}_{BB}\hat{\mathbf{a}}-C\mathbf{K}_B\left(y_I-o_I\right)$.

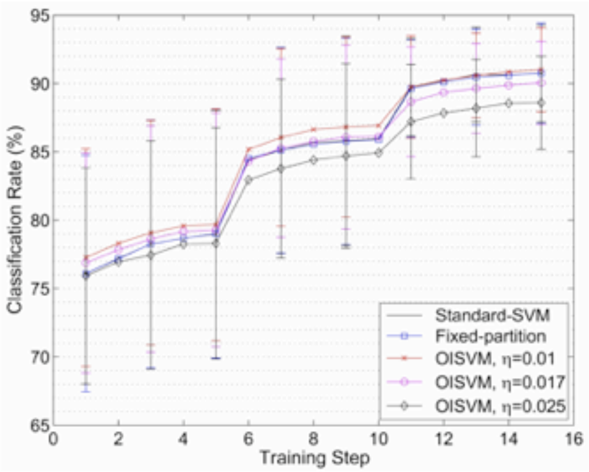
3) go back to Step 1.

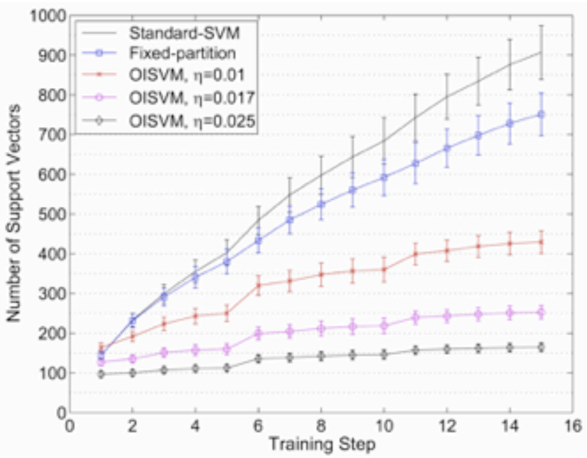


Experimental evaluation

- Compare the performances of the approximate incremental fixed-partition technique [Syed *et al.*, IJCAI'99] and batch method [LIBSVM 2.82]
- We have used 2 different kernels, 36 different training/testing splits
- 3 values of η for each kernel

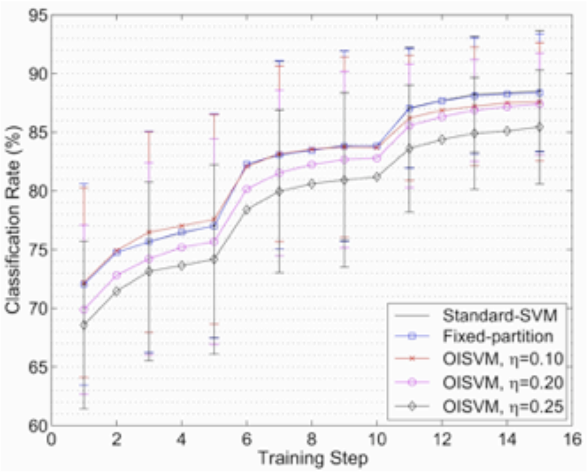
Results (CRFH – Chi^2 Kernel)

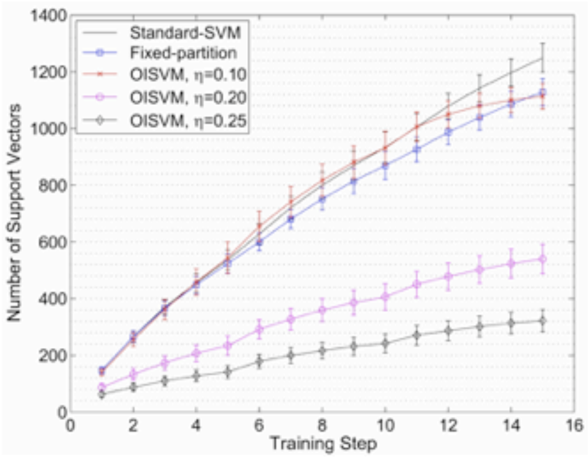




For η = 0.017 and 0.025 at the final incremental step, the number of SVs step is 3-4.5 times less of that of the fixed-partition method and 3.5-5.5 times of that of the standard batch method.

Results (SIFT – Local Kernel)

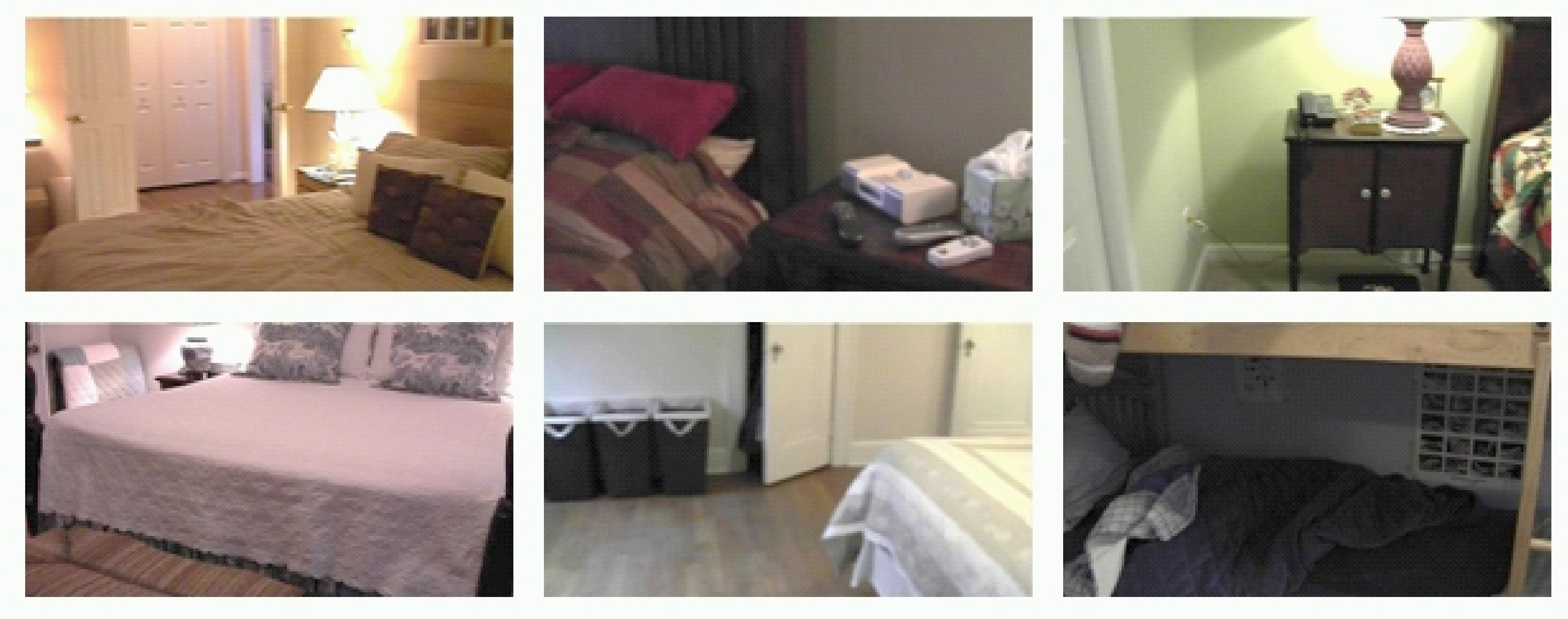




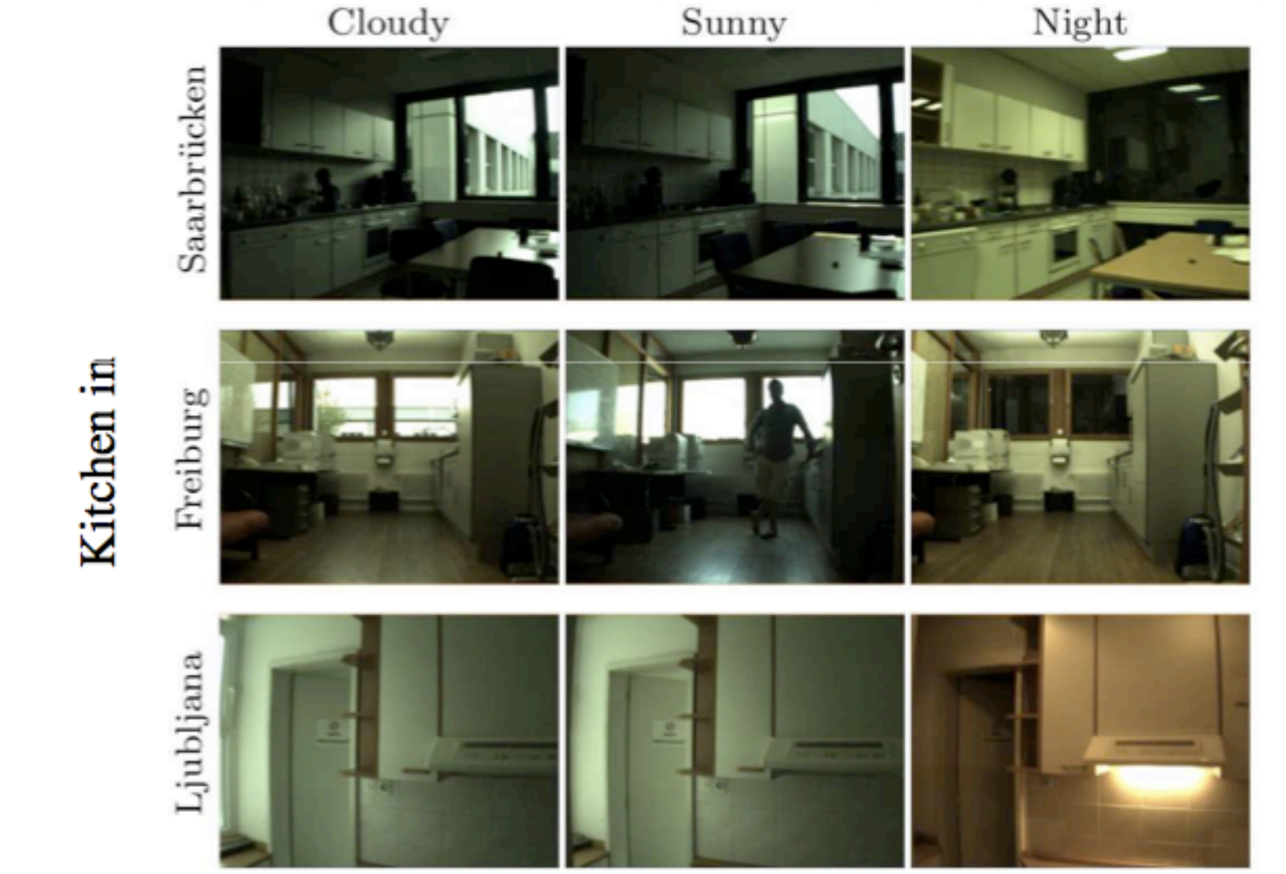
For η = 0.20 and 0.25 the size at the final incremental step, the speedups are respectively 2.3 and 2.1

What about recognizing Places?

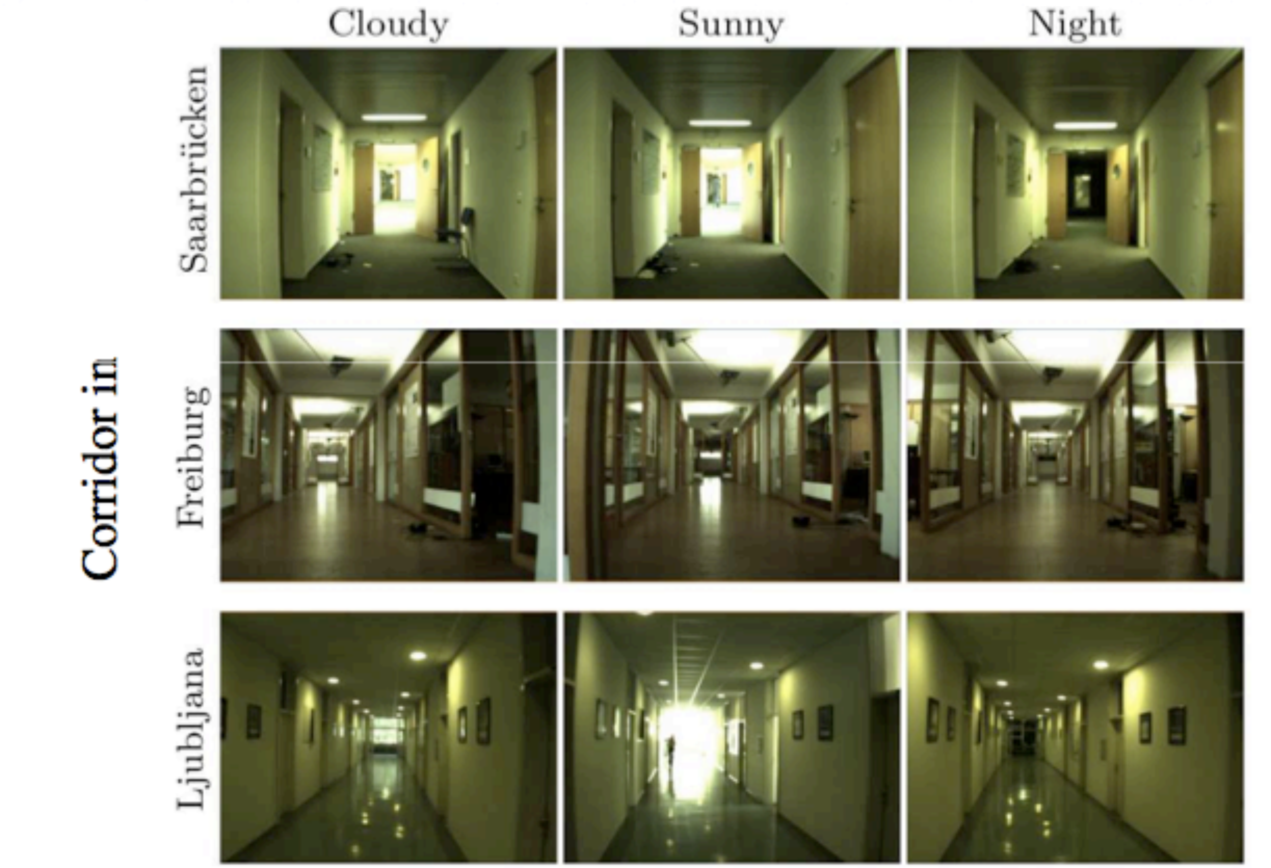
What about recognizing Places?

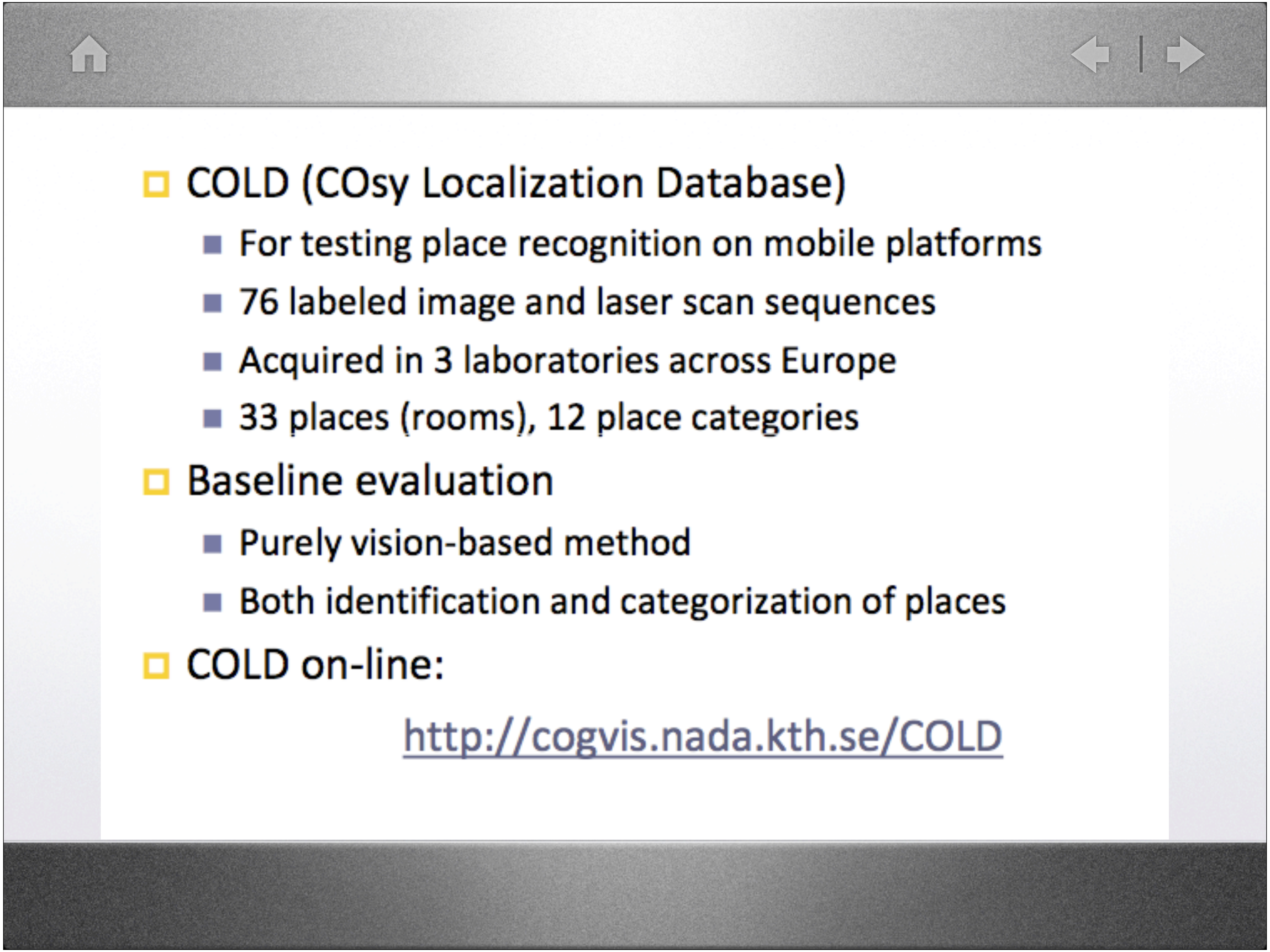


Place Recognition: Office Scenes



Place Recognition: Office Scenes





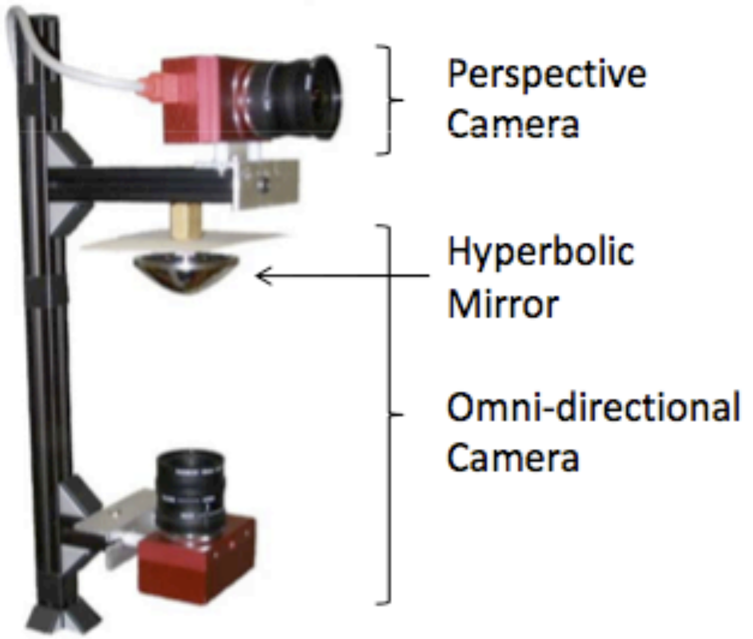
- COLD (COsy Localization Database)
  - For testing place recognition on mobile platforms
  - 76 labeled image and laser scan sequences
  - Acquired in 3 laboratories across Europe
  - 33 places (rooms), 12 place categories
- Baseline evaluation
  - Purely vision-based method
  - Both identification and categorization of places
- COLD on-line:

  http://cogvis.nada.kth.se/COLD

- Three sub-databases:
  - COLD-Ljubljana, COLD-Saarbrücken, COLD-Freiburg

- Acquisition setup
  - The same camera setup
  - Mounted on different robots
  - Images synchronized
  - Resolution 640x480
  - Laser range data available

Text

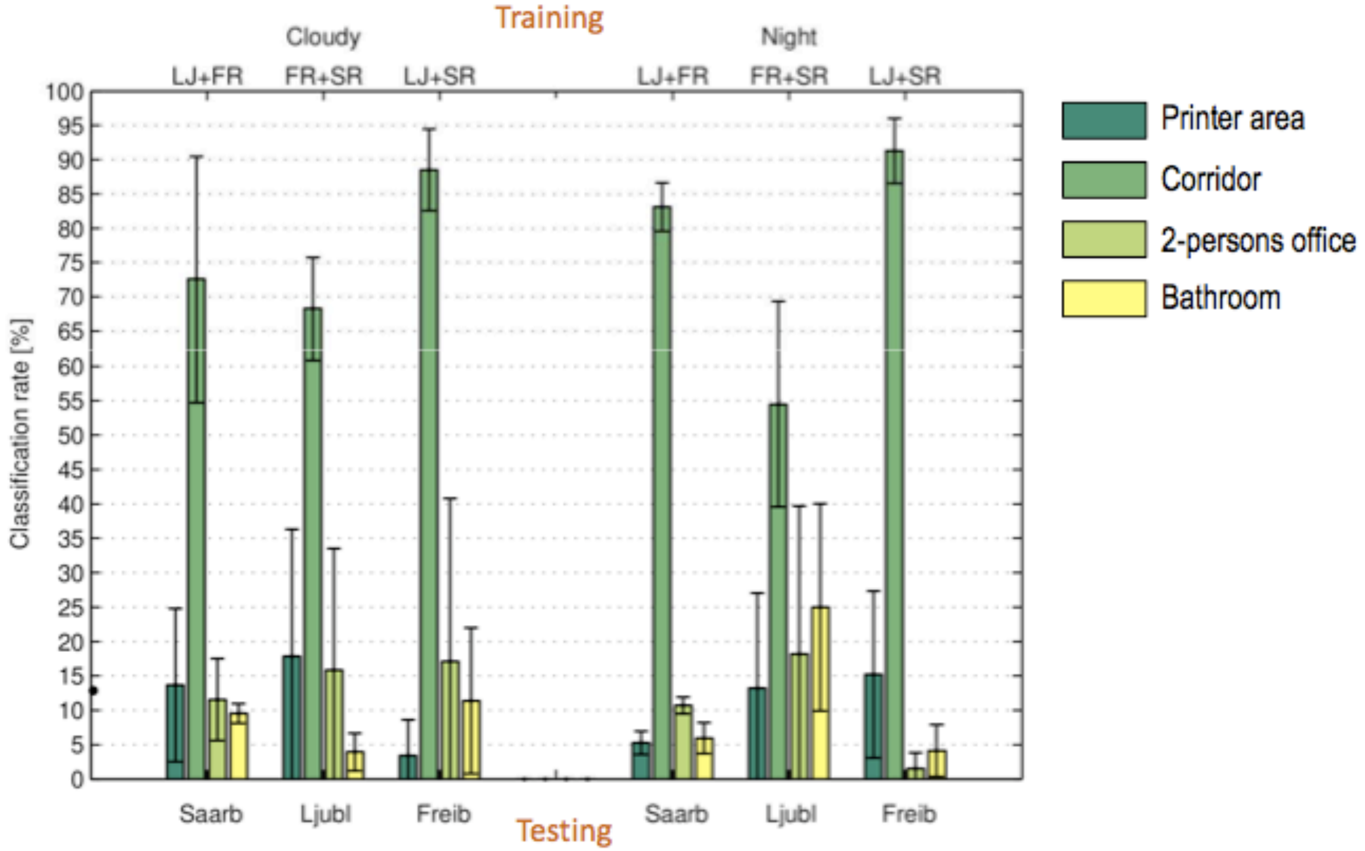


Place Recognition: Office Scenes

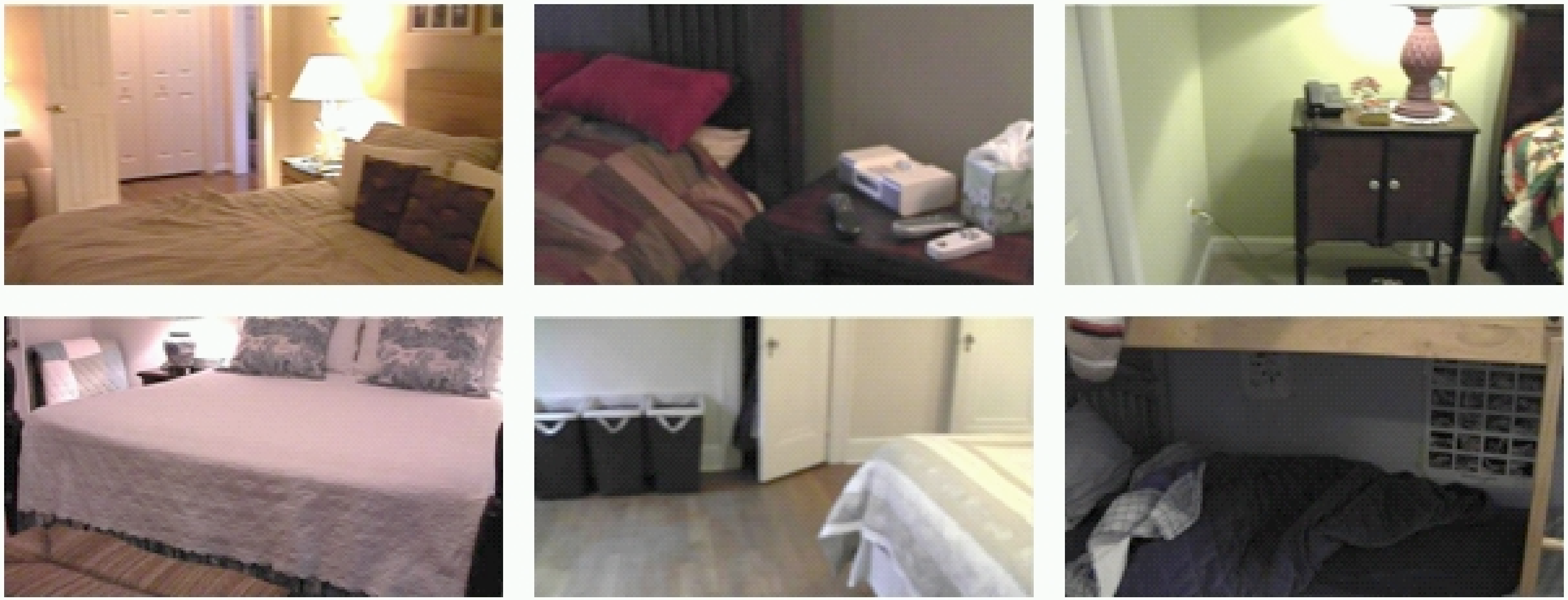
ActivMedia PeopleBot at Saarbrücken
ActivMedia Pioneer-3 at Freiburg
iRobot ATRV-Mini at Ljubljana

M. Ullah, A. Pronobis, B. Caputo, J. Luo, O. Jensfelt, H. Christensen. *Towards robust place classification for robot localization.* Proc International Conference on Robots and Automation, 2008

Place Recognition: Office Scenes

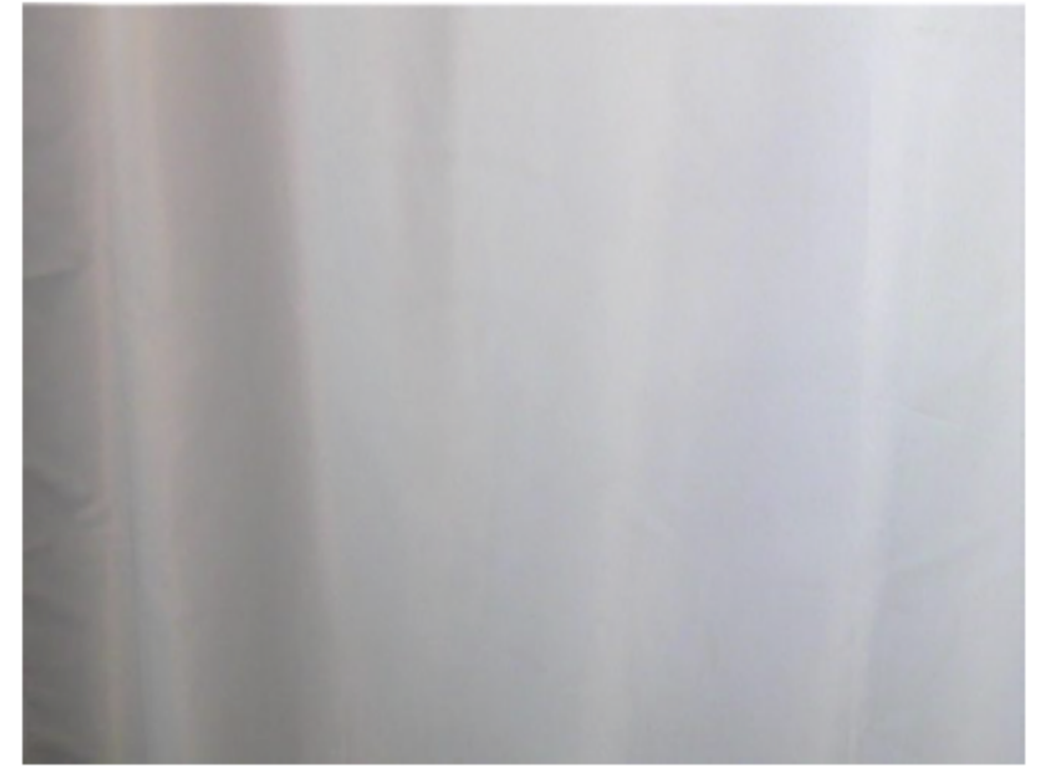


Place Recognition: Home Scenes

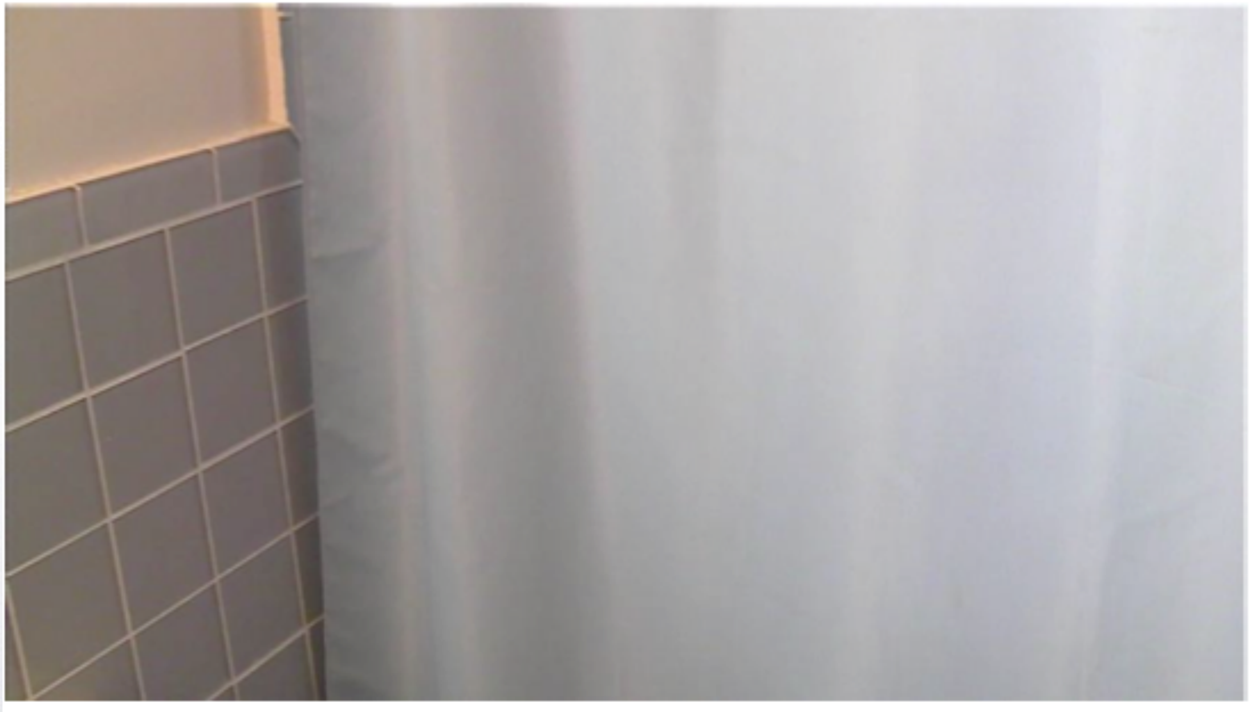


J. Wu, H. Christensen, J. Rehg. *Visual place categorization: problem, dataset, and algorithm.* Proc IROS2009

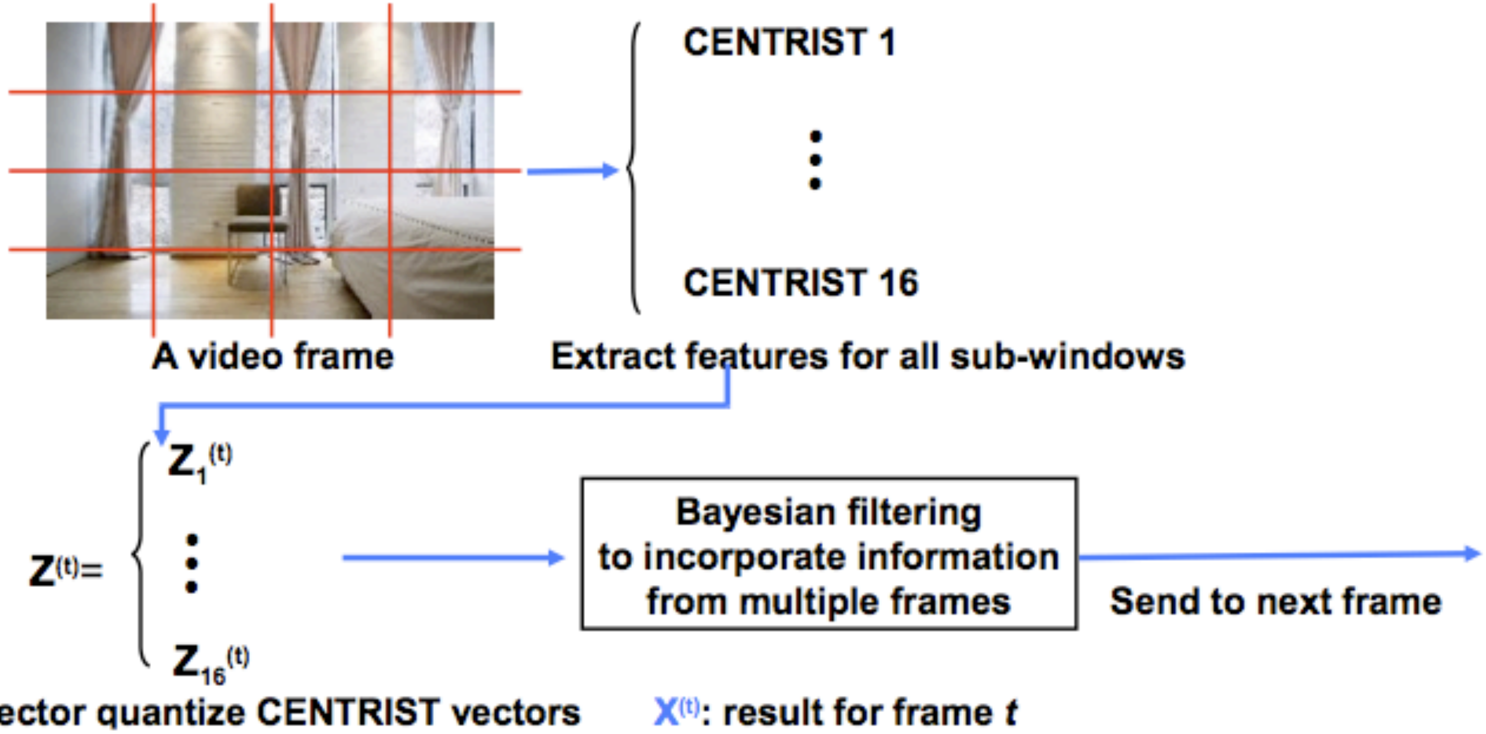
Place Recognition: Home Scenes



Place Recognition: Home Scenes



Place Recognition: Home Scenes



A video frame

CENTRIST 1

⋮

CENTRIST 16

Extract features for all sub-windows

$Z^{(t)}= \left\{ Z_1^{(t)}, \vdots, Z_{16}^{(t)} \right.$

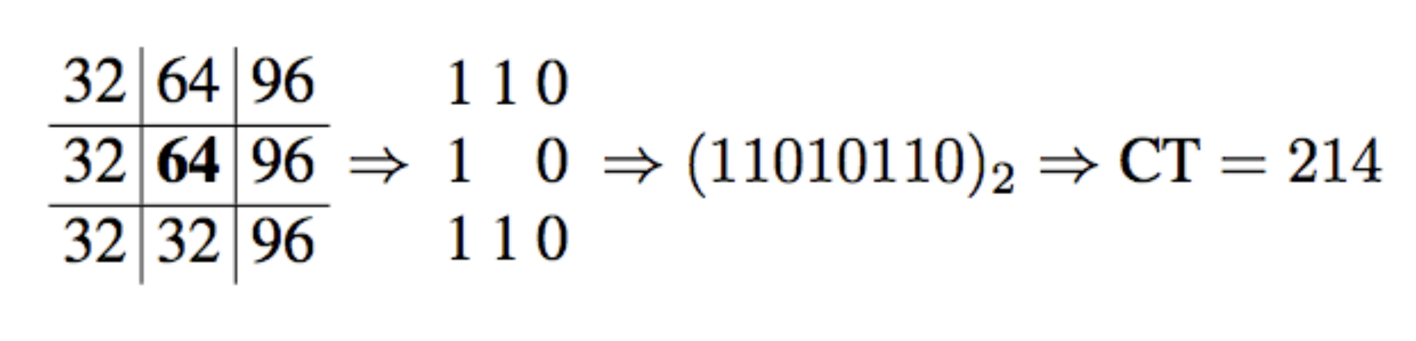
Vector quantize CENTRIST vectors

Bayesian filtering to incorporate information from multiple frames

Send to next frame

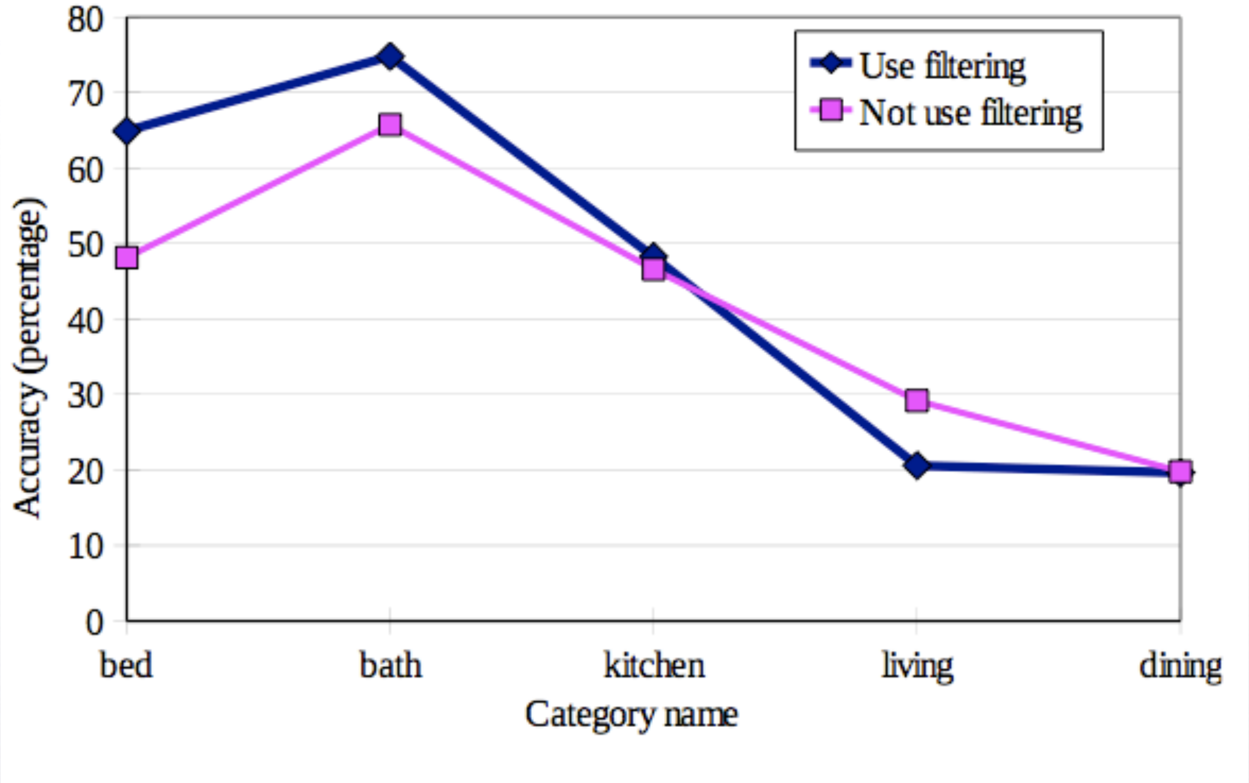
$X^{(t)}$: result for frame *t*

- Census transform compares the intensity value of a pixel with its eight neighboring pixels
- If the center pixel is >= one of its neighbors, a bit 1 is set in the corresponding location/0 otherwise
- Bit representation then converted to an integer [0.255]



$$\begin{array}{c|c|c} 32 & 64 & 96 \\ \hline 32 & \mathbf{64} & 96 \\ \hline 32 & 32 & 96 \end{array} \Rightarrow \begin{array}{ccc} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{array} \Rightarrow (11010110)_2 \Rightarrow \mathrm{CT} = 214$$

Place Recognition: Home Scenes



Take Home Message

- Robots need semantic visual information to describe where they are
- Most of images acquired in a room by a robot are non informative --this makes the problem harder
- preliminary attempts to build place recognition systems seem to work fine; place categorization much more challenging

that’s all folks!