Computational Eye Glass

Literature Survey-Scene Recognition

Venkatesh N
University of Massachusetts Amherst

Courtesy: Antonio Torralba

Agenda

- **Scene Recognition**

  -Problem definitions

  - Challenges

  - Literature

- **Fixations**

  - Problem Definition

  - Literature

Memory Confusion:
The details of some objects are forgotten

You have seen these pictures

You were tested with these pictures

Problem Definition

A scene is a view of a real-world environment that contains multiples surfaces and objects, organized in a meaningful way

Distinction Between Objects and Scene.

Indoor vs Outdoor Scene

Indoor:



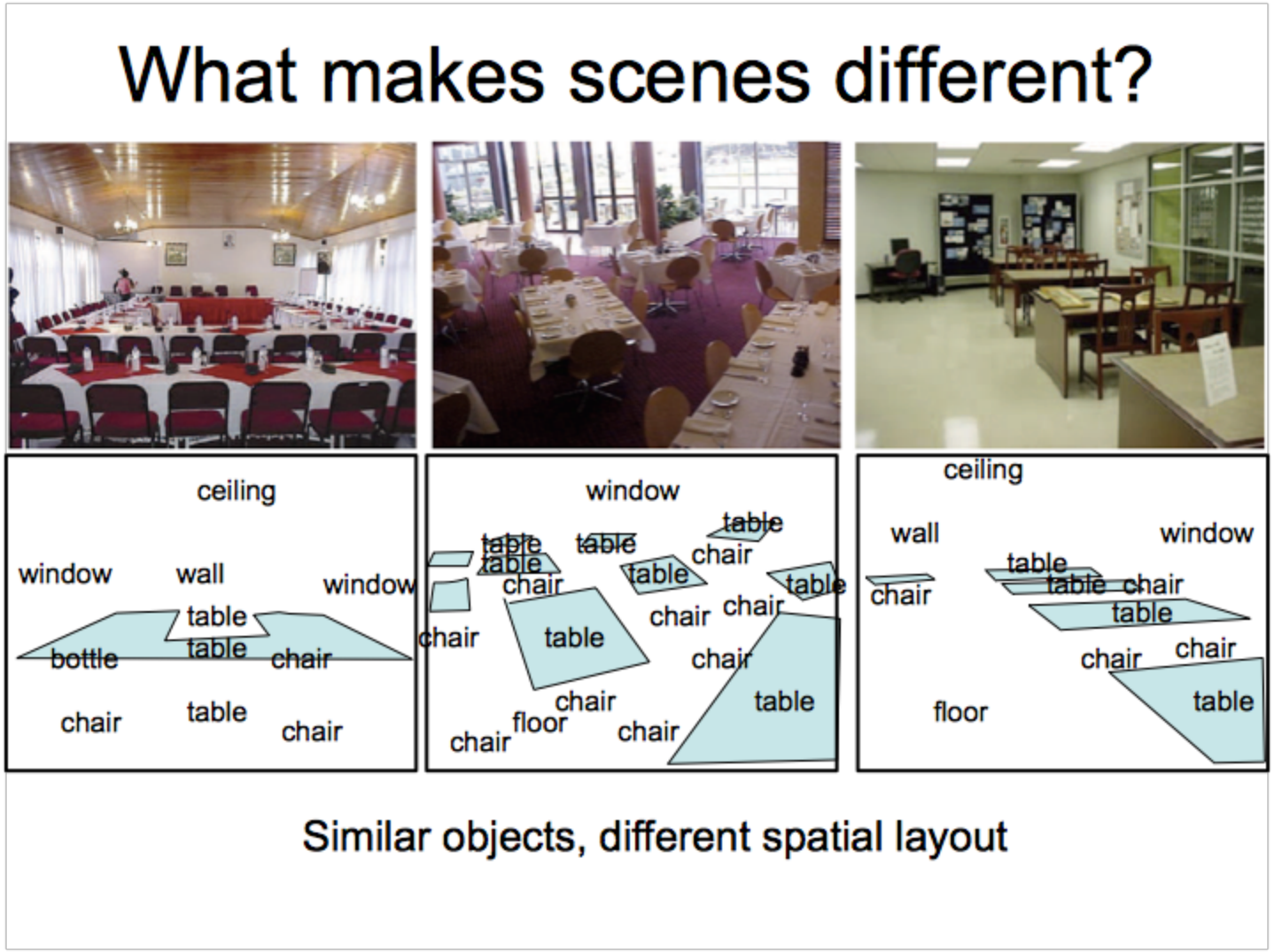
Outdoor:



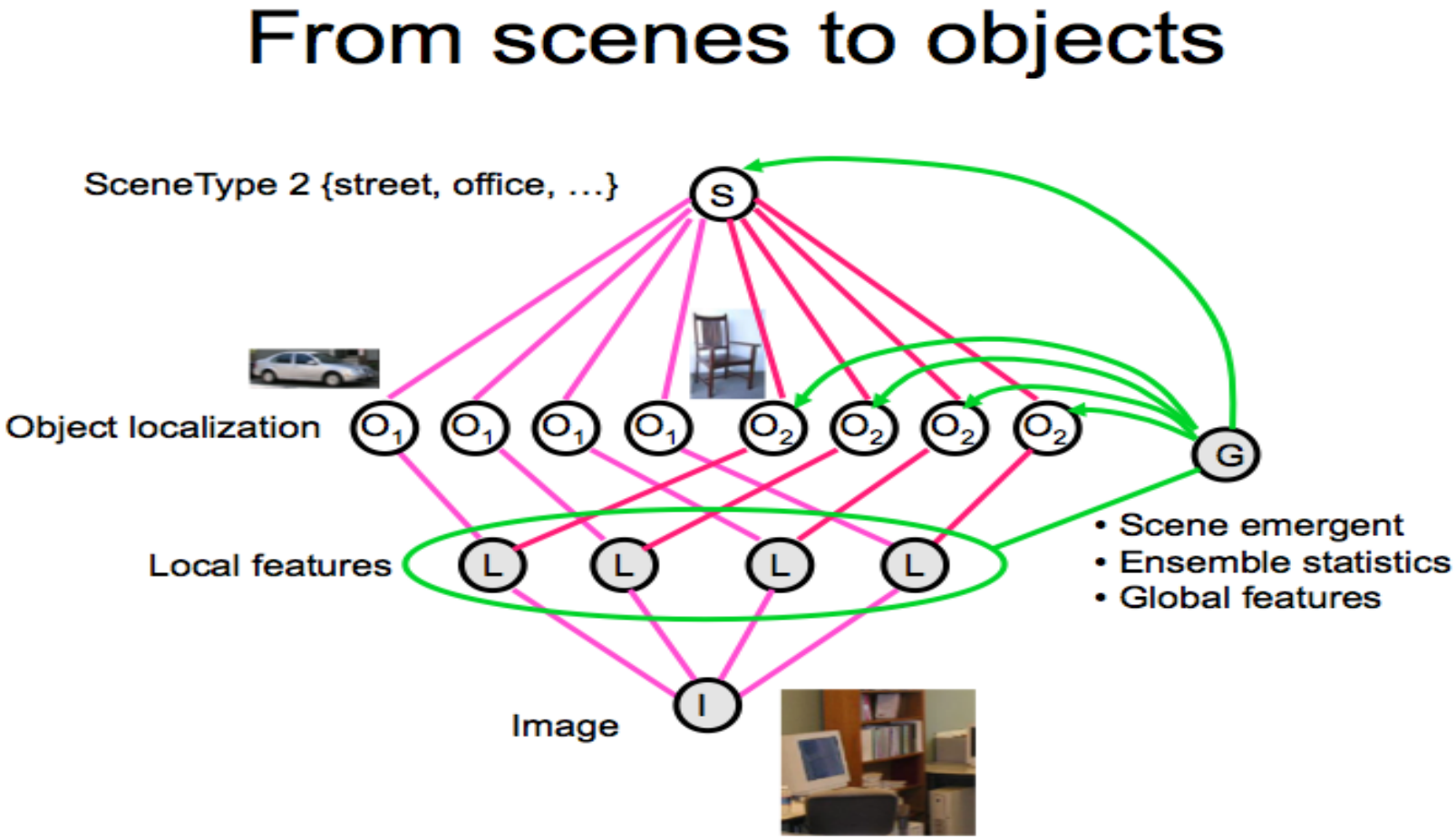
Challenges

- Its a hard task(Similar Structures of object but they are different)
- Illumination
- Limited Resources
- Real-time performance
- Viewing angle
- Intra class variations are high

Literature


What makes scenes different?
ceiling
window
wall
window
table
table
bottle
chair
chair
table
chair
window
table
table
table
table
chair
chair
chair
table
chair
chair
table
chair
chair
floor
chair
chair
table
ceiling
wall
window
table
chair
table
chair
table
chair
chair
floor
table
Similar objects, different spatial layout

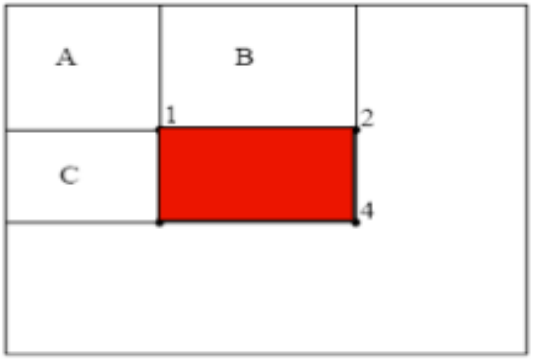
Literature


From scenes to objects
SceneType 2 {street, office, …}
S
Object localization
O1
O1
O1
O1
O2
O2
O2
O2
G
Local features
L
L
L
L
• Scene emergent
• Ensemble statistics
• Global features
I
Image

Literature

Haar-like filters and cascades

Viola and Jones, ICCV 2001

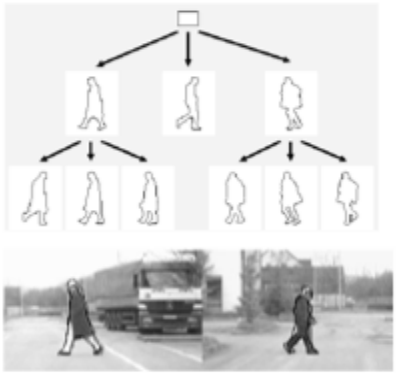


The average intensity in the block is computed with four sums independently of the block size.

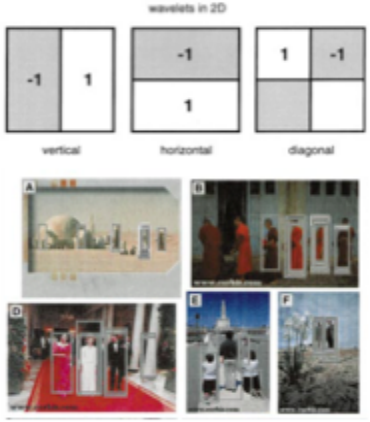
Also Fleuret and Geman, 2001

Literature

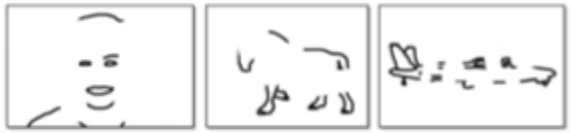
Generic objects: Edge based descriptors



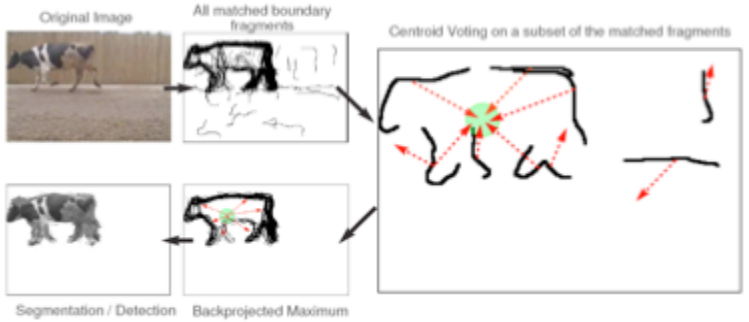
Gavrila, Philomin, ICCV 1999



Papageorgiou & Poggio (2000)



J. Shotton, A. Blake, R. Cipolla. PAMI 2008.

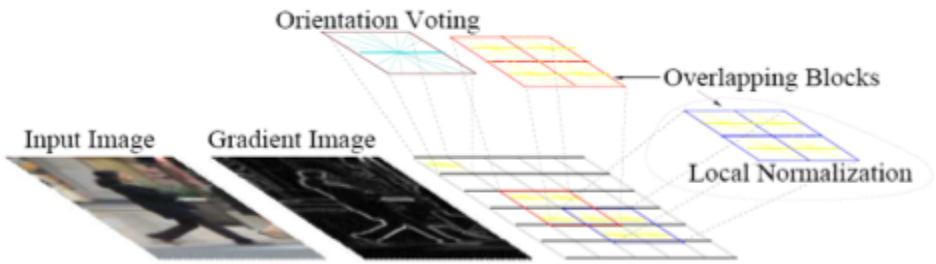


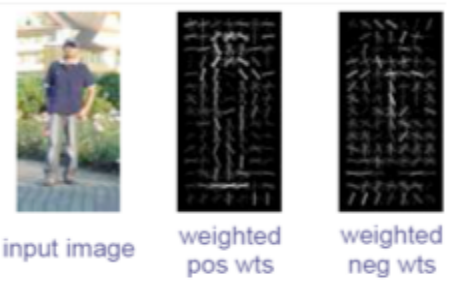
Opelt, Pinz, Zisserman, ECCV 2006

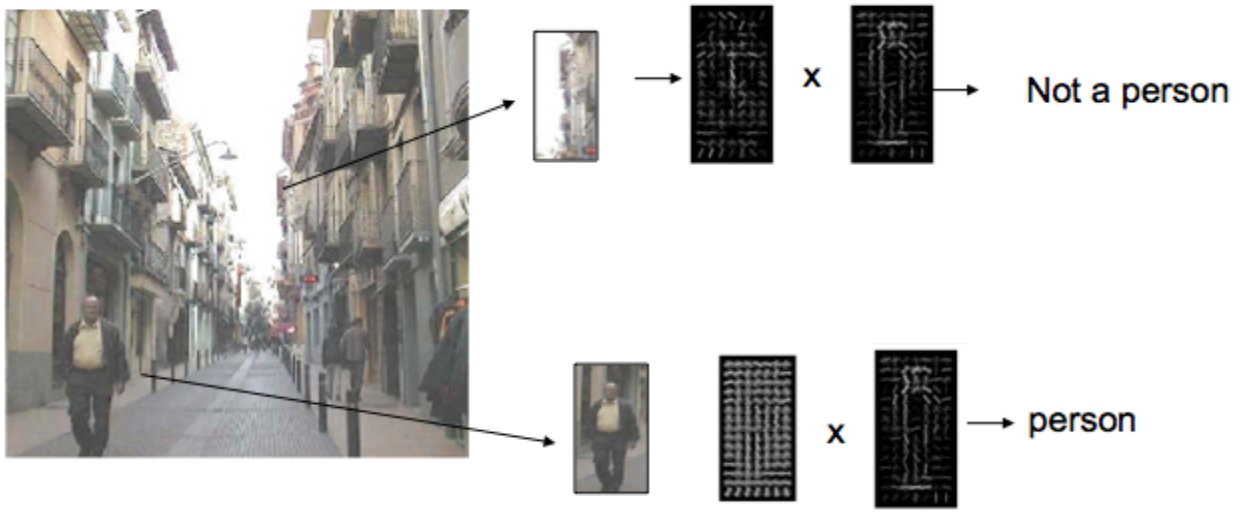
Literature

Histograms of oriented gradients

Dalal & Trigs, 2006



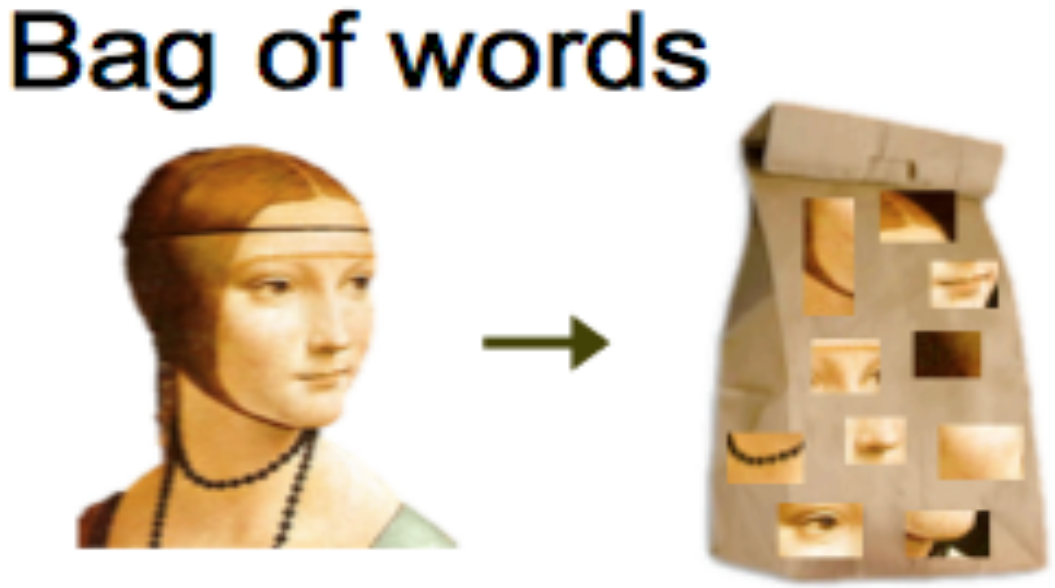




Global image descriptors

- Bag of Words
- Gist
- Textons
- Spatial Pyramid Matching

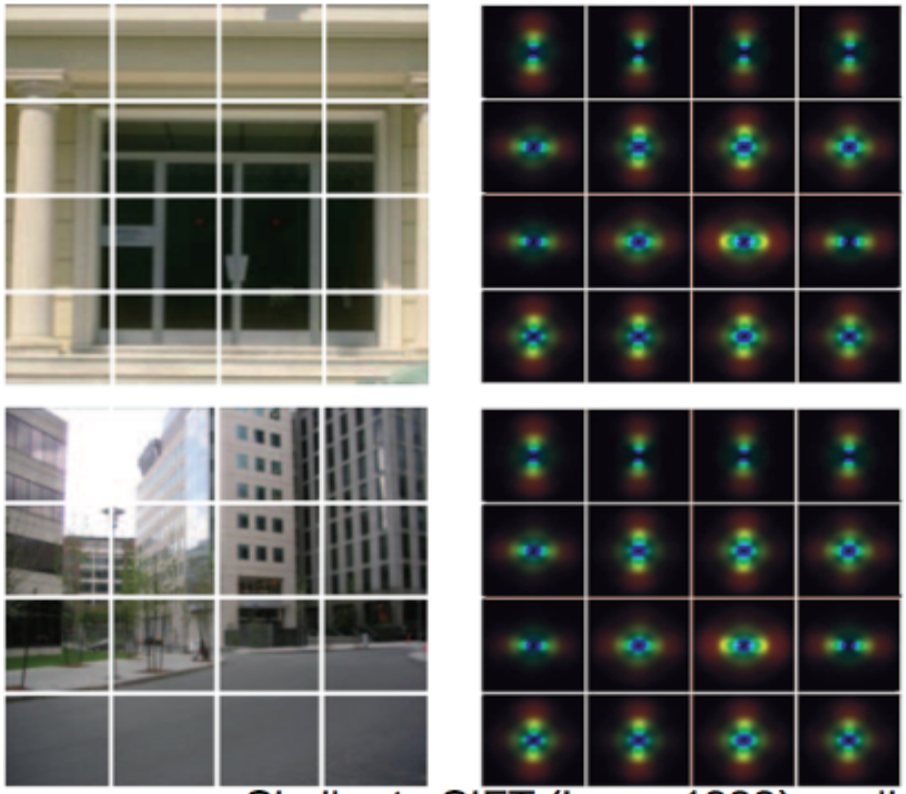
Bag of Words


Bag of words
Sivic et. al., ICCV 2005
Fei-Fei and Perona, CVPR 2005

Global image descriptors

Gist descriptor

Oliva and Torralba, 2001



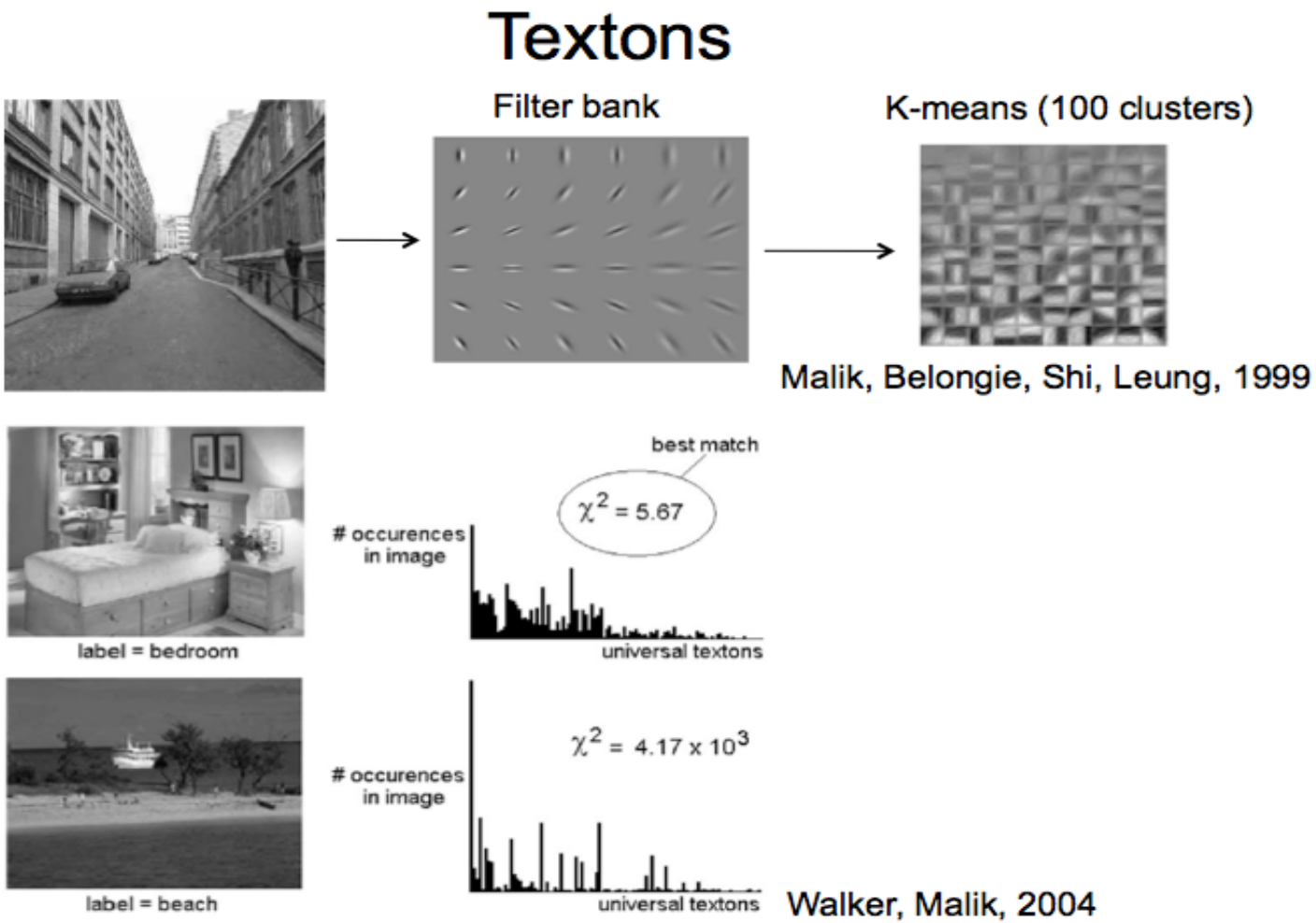
- Apply oriented Gabor filters over different scales
- Average filter energy in each bin

8	orientations
4	scales
x 16	bins
512	dimensions

Similar to SIFT (Lowe 1999) applied to the entire image

M. Gorkani, R. Picard, ICPR 1994; Walker, Malik. Vision Research 2004; Vogel et al. 2004; Fei-Fei and Perona, CVPR 2005; S. Lazebnik, et al, CVPR 2006; …

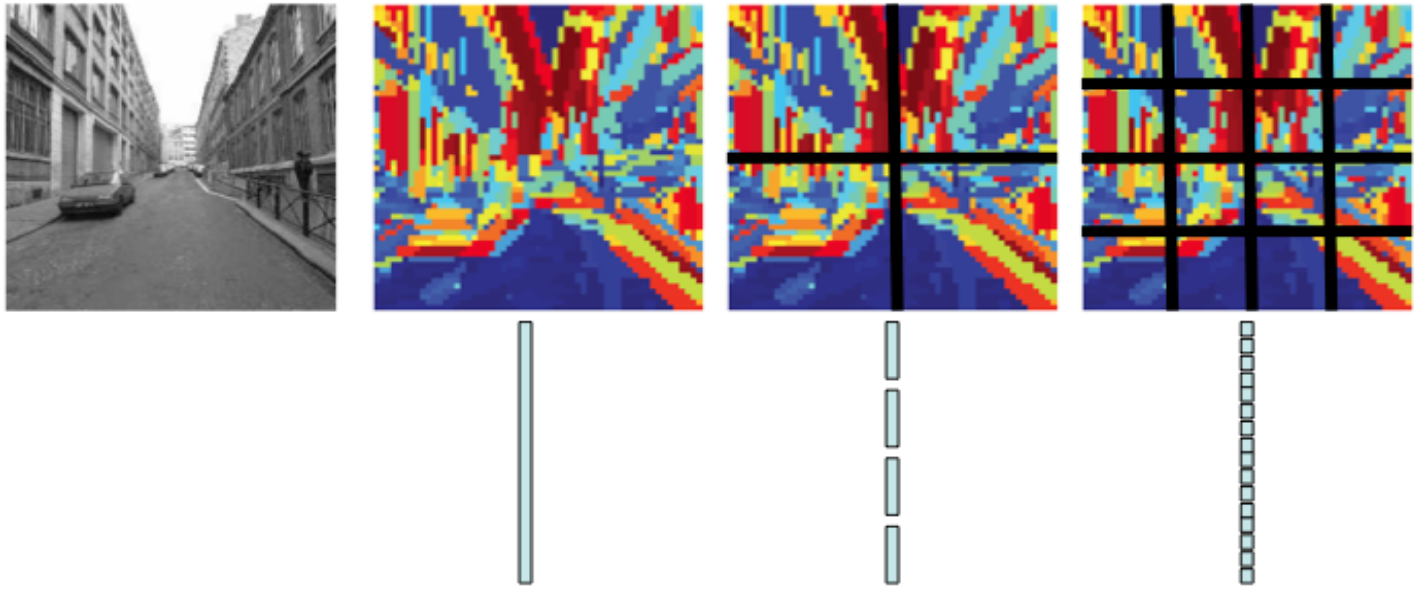
Global image descriptors



Global image descriptors

Bag of words & spatial pyramid matching

Sivic, Zisserman, 2003. Visual words = Kmeans of SIFT descriptors



S. Lazebnik, et al, CVPR 2006

Performance

The 15-scenes benchmark

Oliva & Torralba, 2001
Fei Fei & Perona, 2005
Lazebnik, et al 2006

Office

Skyscrapers

Suburb

Building facade

Coast

Forest

Bedroom

Living room

Industrial

Street

Highway

Mountain

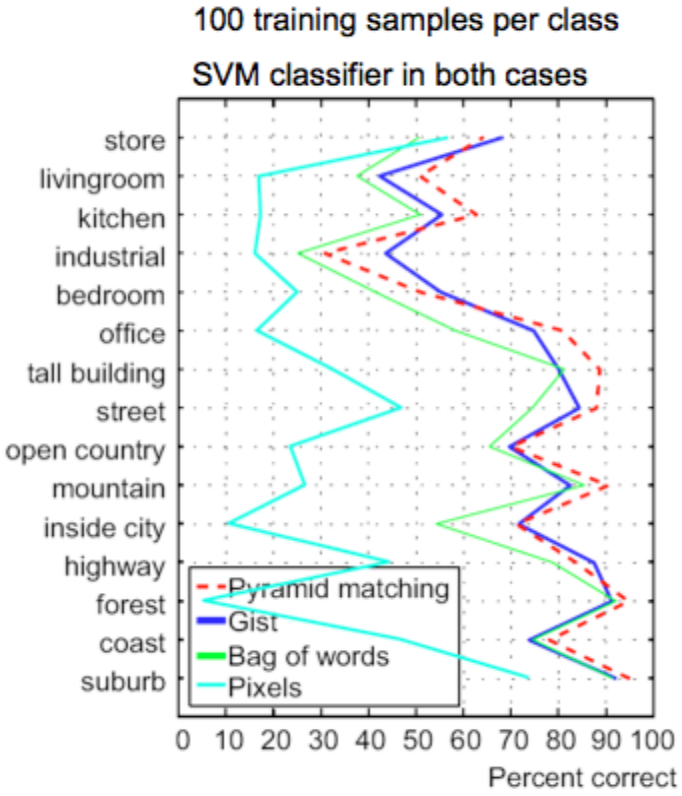
Open country

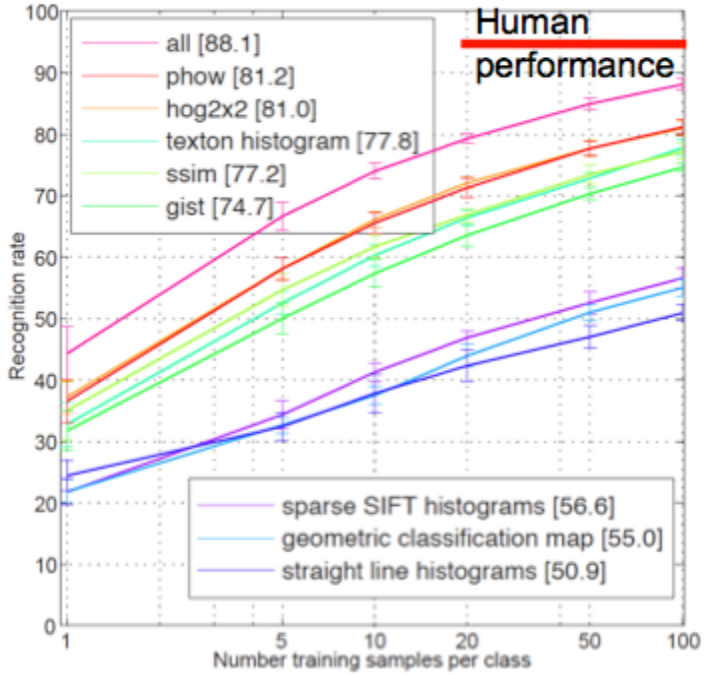
Kitchen

Store

Performance

Scene recognition





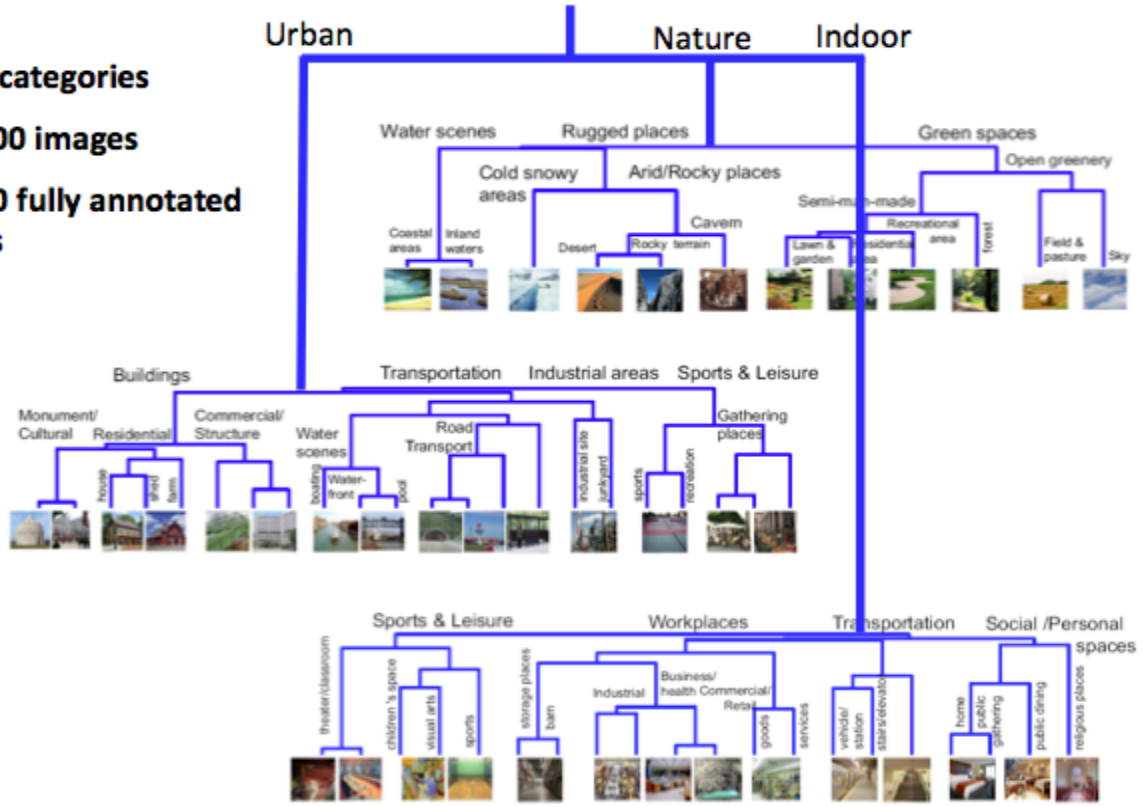
Performance

Large Scale Scene Recognition

~1,000 categories

>130,000 images

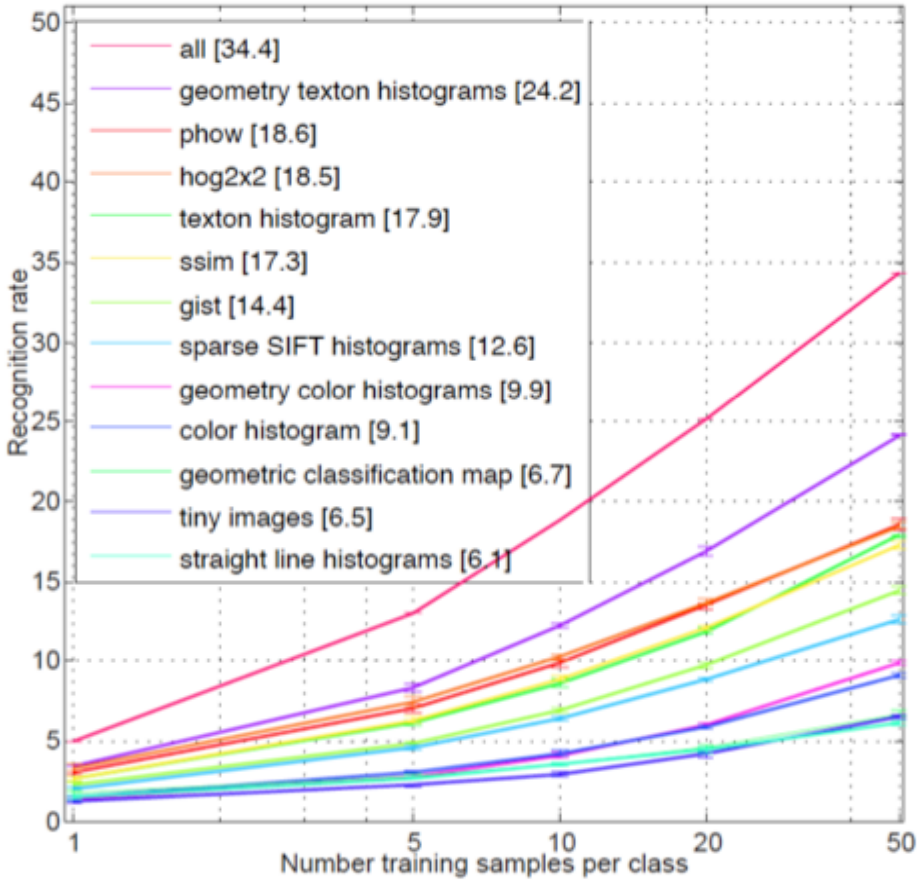
>12,000 fully annotated images



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010

Performance

Performance with 400 categories



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010

Importance of Context

(a) Isolated object

(b) Object in context

(c) Low-res Object

Literature

Context-based vision system for place and object recognition [Torralba et al CVPR 2003]

Goal: Scene and object recognition based on object.

Main Contribution: Proposed a model for place recognition and identifying the category of the location.

Input data: 120x160 at 4 images per second

Feature and Model

- Features

Local representation (L)

-Wavelet decomposition

-N = 24 (6 orientations, 4 scales)

Global representation

-Pyramid approach

-PCA to reduce to 80 dimensions

- Model

HMM Model

-Hidden states = location (63 values)

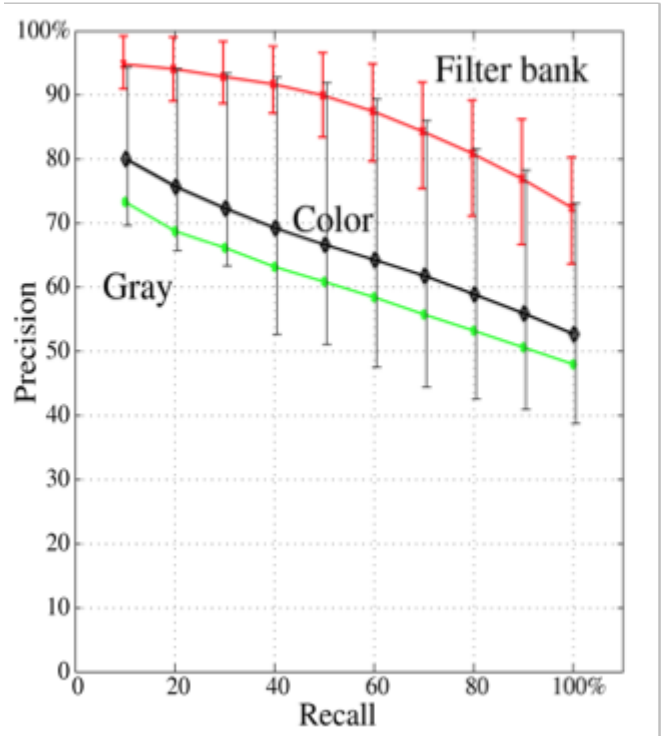
-Observations = R80 features

-Transition model encodes topology of environment

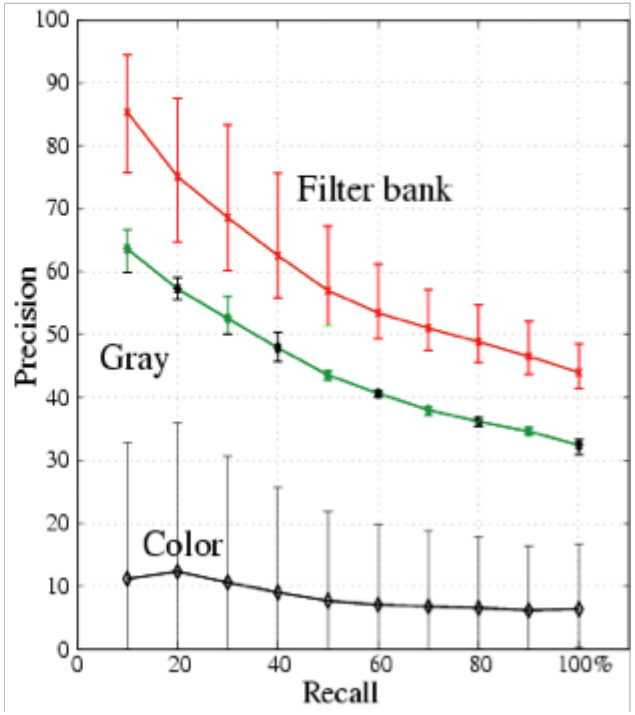
-Observation model is a mixture of Gaussians (100 views per place)

Results

Recognition

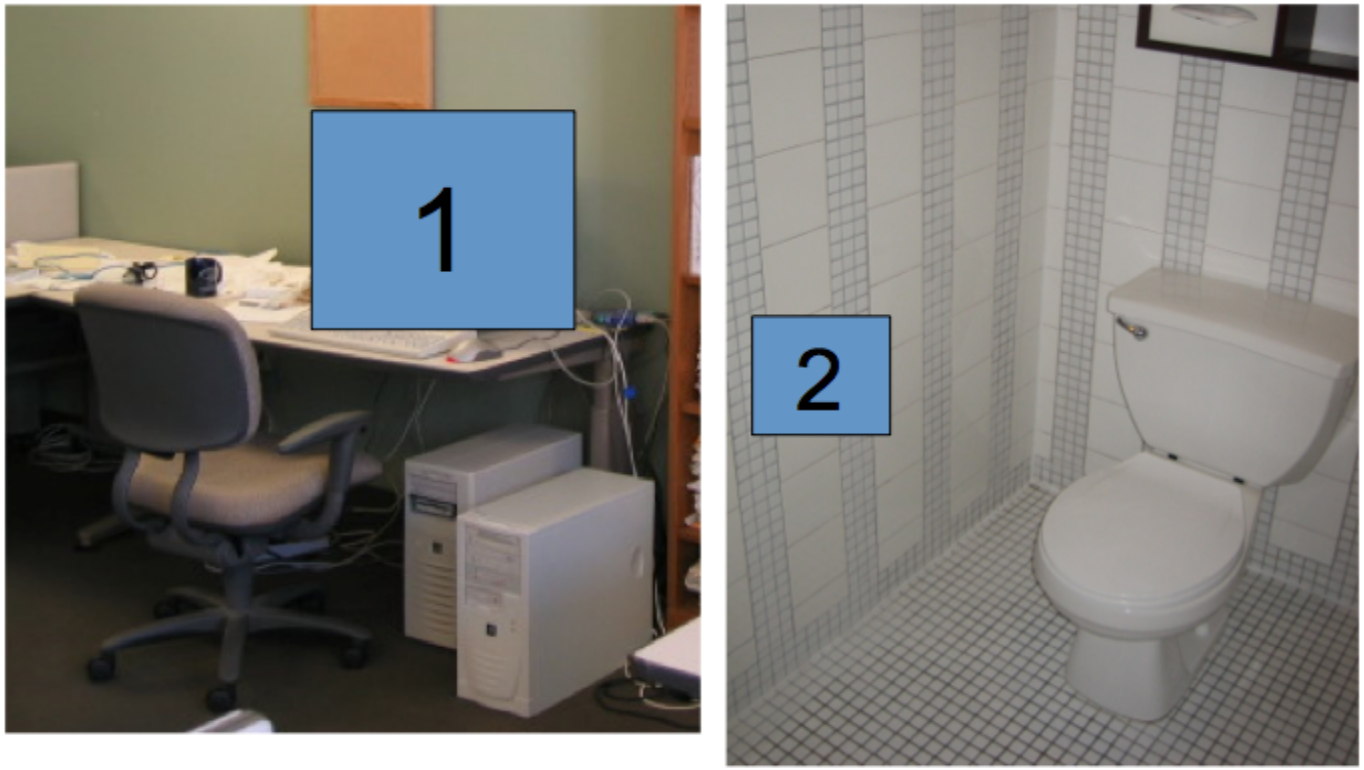


Categorization



Object-Object Relationships

What are the hidden objects?



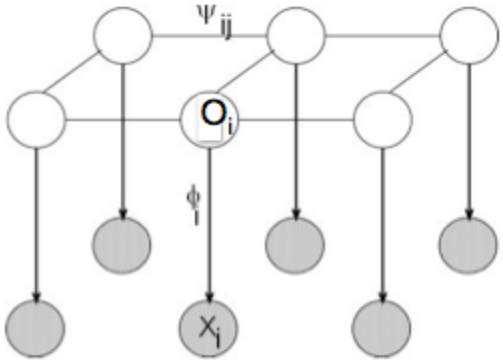
Object-Object Relationships

What are the hidden objects?

Object-Object Relationships

Pixel labeling using MRFs

Enforce consistency between neighboring labels, and between labels and pixels

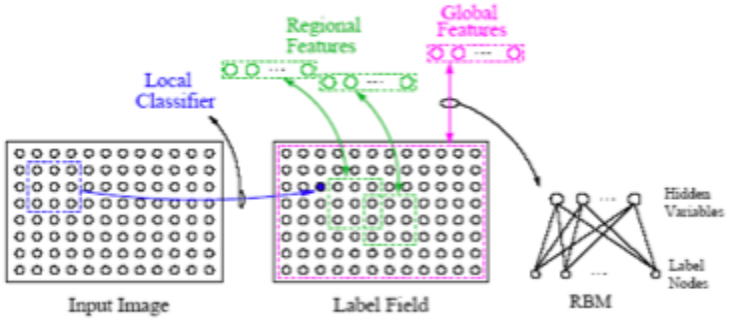


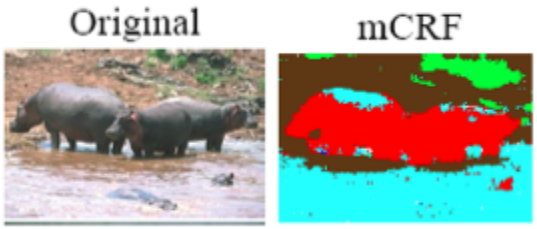
Carbonetto, de Freitas & Barnard, ECCV'04

Object-Object Relationships

Object-Object Relationships

Use latent variables to induce long distance correlations between labels in a Conditional Random Field (CRF)





He, Zemel & Carreira-Perpinan (04)

Object-Object Relationships

Slide by Fei-fei

L-J Li & L. Fei-Fei, ICCV 2007

Literature

Families of recognition algorithms

Bag of words models

Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

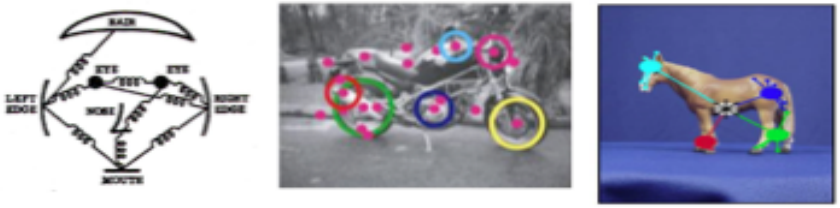
Voting models

Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schneiderman, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching
Deformable models

Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

Constellation models



Fischler and Elschlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

Rigid template models

Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006

Fixations (Where Humans look?)

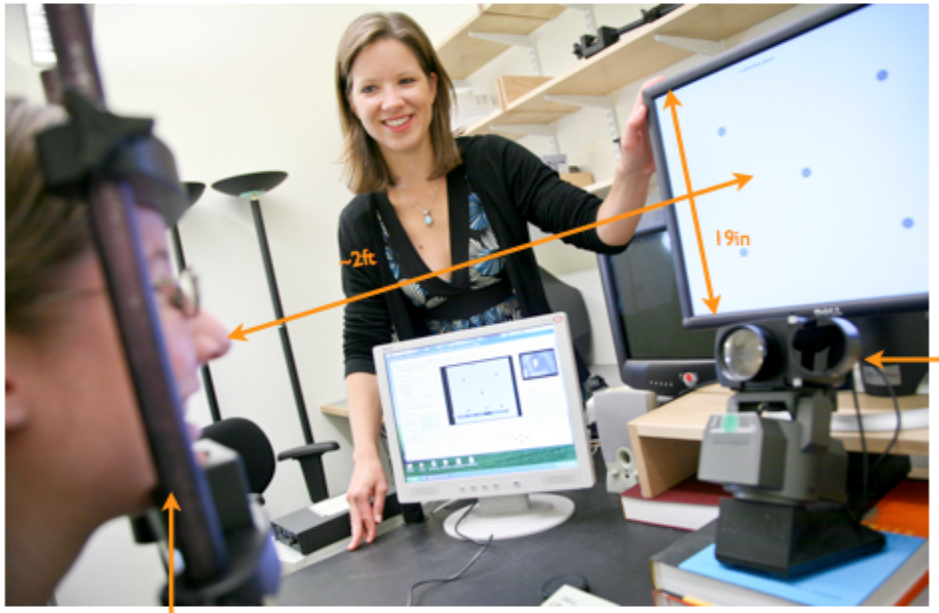
Judd et al ICCV 2009

Fixations

Judd et al ICCV 2009

Fixations

Collect eye tracking data



15 users on 1003 images

Judd et al ICCV 2009

Fixations

Features

- Low level
  illuminance, orientation, color
- Mid level?
  vanishing point, horizon line
- High level
  face detection, object detection

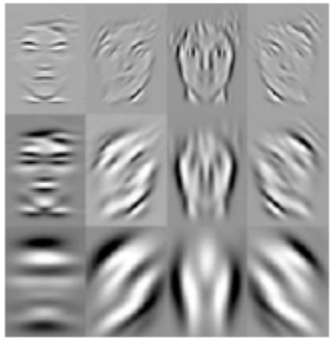
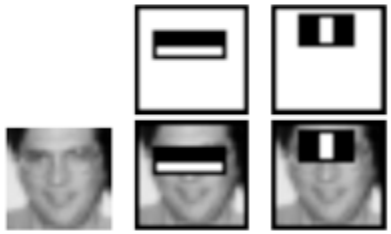


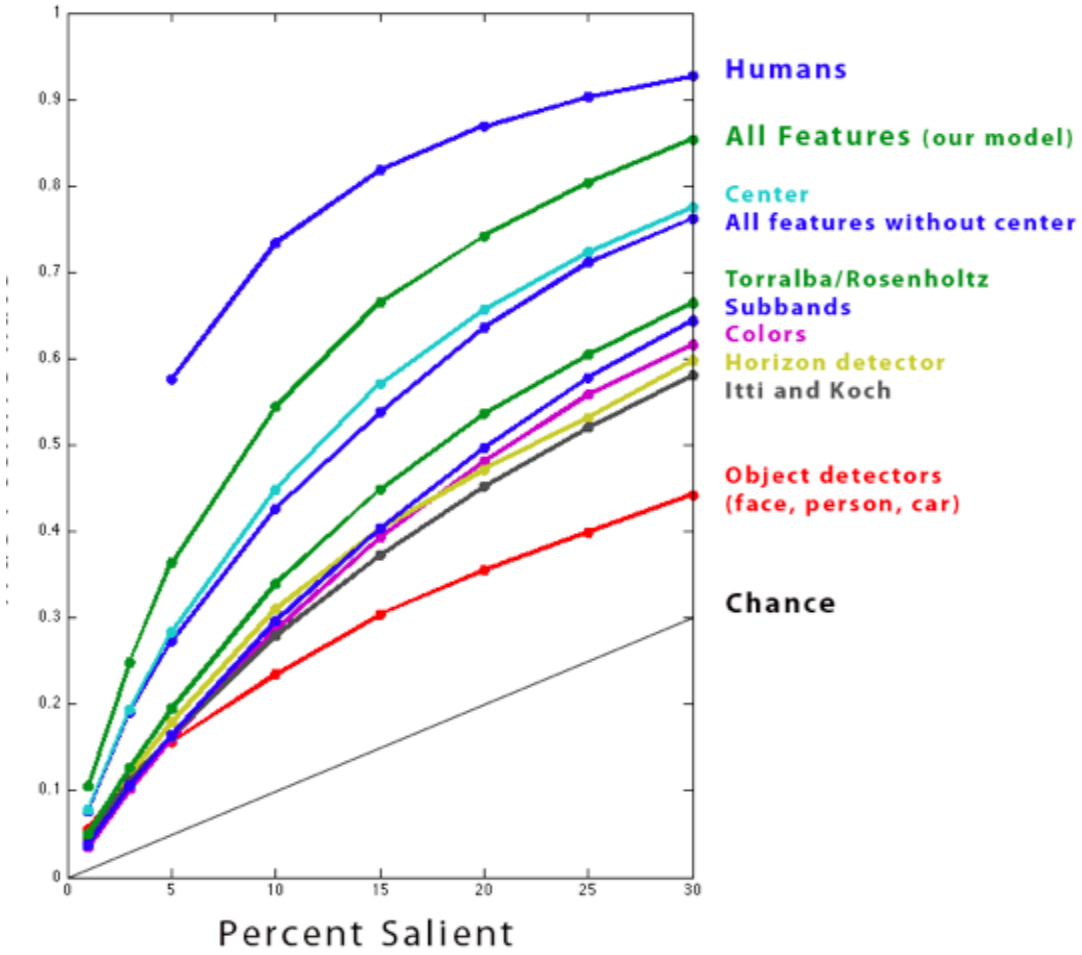
Image filtered with Difference-of-Gaussian(DoG) filters

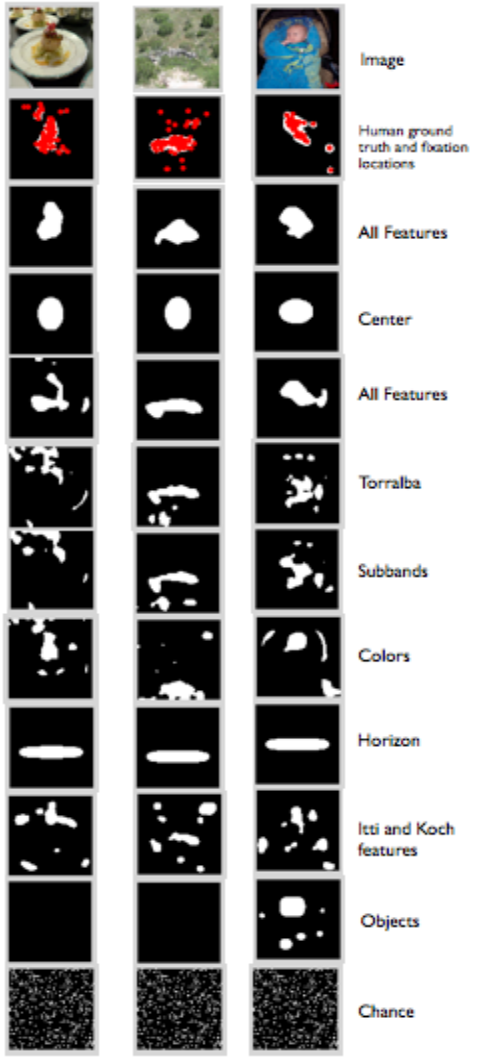


Viola Jones Face detector

Judd et al ICCV 2009

Fixations





Judd et al ICCV 2009

Planned Work

- Simultaneous Capture of images from CMU cam and Stonyman cam.

- Extracting GIST features and performing SVM classification of scenes.

- Validate the performance in both the scenario (Is Stonyman cam sufficient for scene detection?)