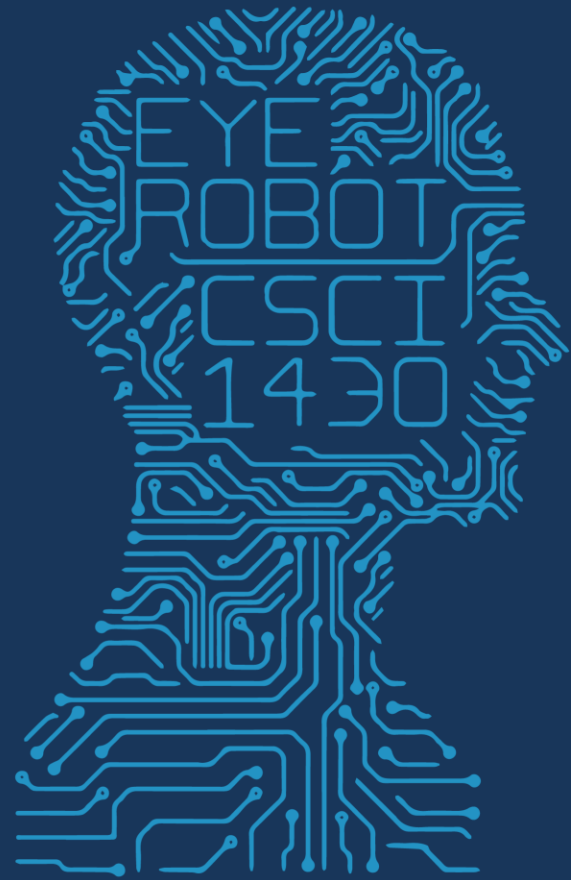




1950

FUTURE VISION



25 FEBRUARY 2019

COMPUTER VISION

TO COMPLETE YOUR REGISTRATION, PLEASE TELL US
WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:



NO YES

ANSWER QUICKLY—OUR SELF-DRIVING
CAR IS ALMOST AT THE INTERSECTION.

SO MUCH OF "AI" IS JUST FIGURING OUT WAYS
TO OFFLOAD WORK ONTO RANDOM STRANGERS.

TO COMPLETE YOUR REGISTRATION, PLEASE TELL US
WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:



NO YES

ANSWER QUICKLY—OUR SELF-DRIVING
CAR IS ALMOST AT THE INTERSECTION.

SO MUCH OF "AI" IS JUST FIGURING OUT WAYS
TO OFFLOAD WORK ONTO RANDOM STRANGERS.

Alt-text: "Crowdsourced steering" doesn't sound quite as appealing as "self driving".



Large-scale category recognition and Advanced feature encoding

Computer Vision
Many slides from James Hays

Scene Categorization

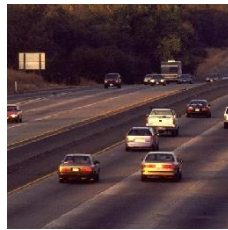
Oliva and Torralba, 2001



Coast



Forest



Highway



Inside
City



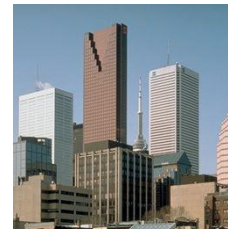
Mountain



Open
Country



Street



Tall
Building

Fei Fei and Perona, 2005

+



Bedroom



Kitchen



Living Room



Office



Suburb

Lazebnik, Schmid, and Ponce, 2006

+



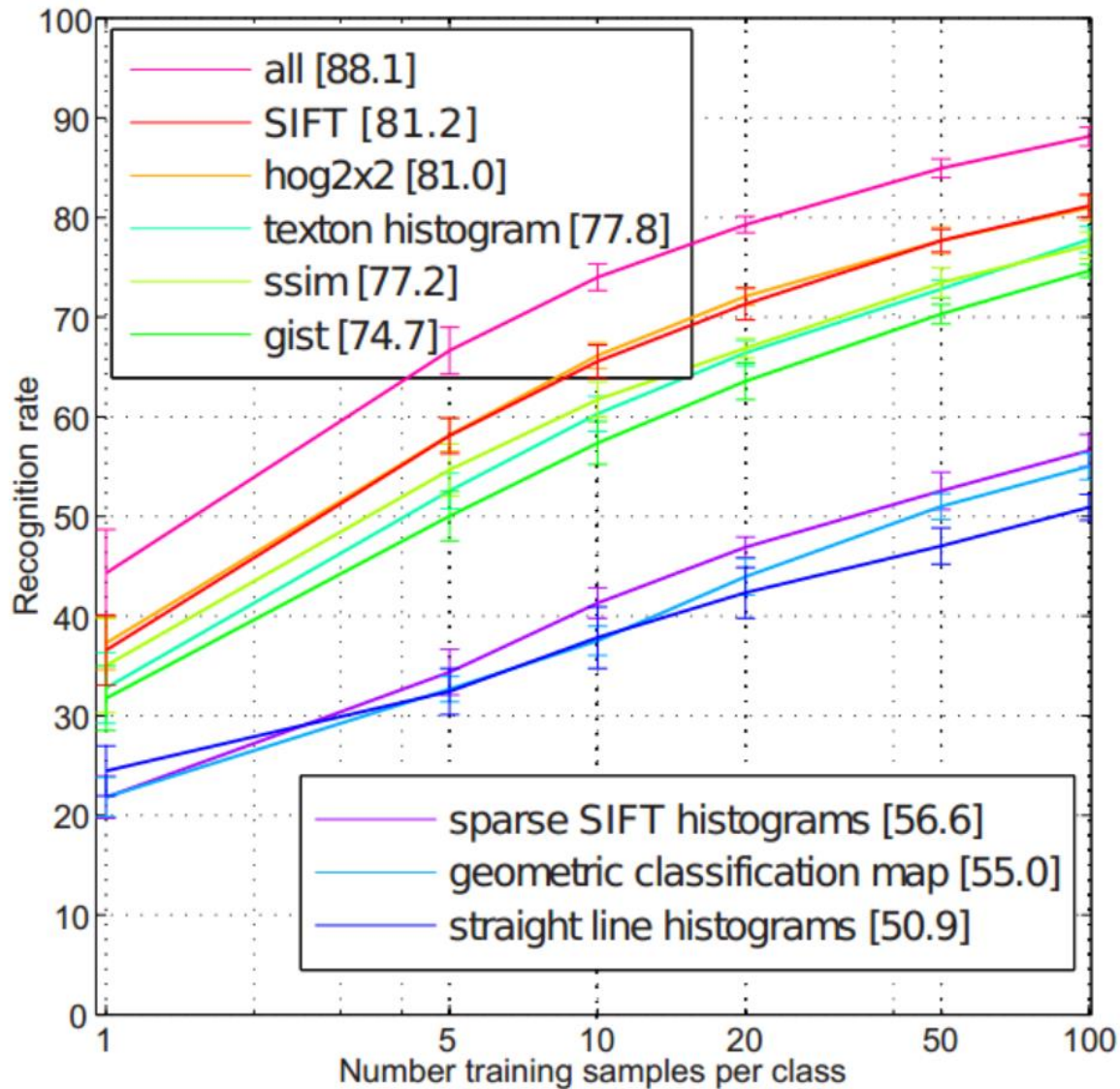
Industrial



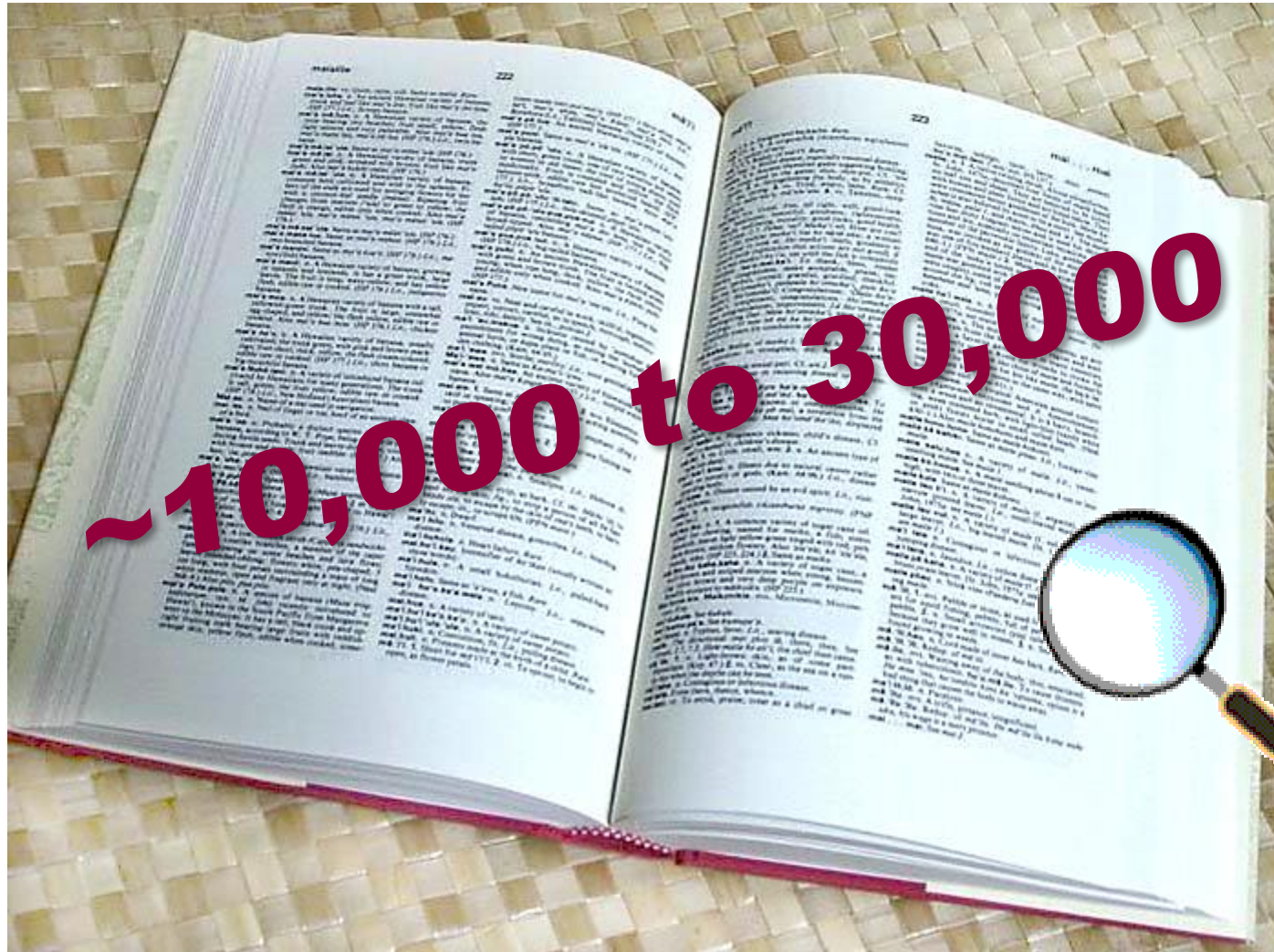
Store

15 Scene Database

15 Scene Recognition Rate



How many object categories are there?



OK, but how many places?

Biederman 1987

abbey



airplane cabin



airport terminal





apple orchard



assembly hall



bakery





car factory



cockpit



construction site





food court



interior car



lounge





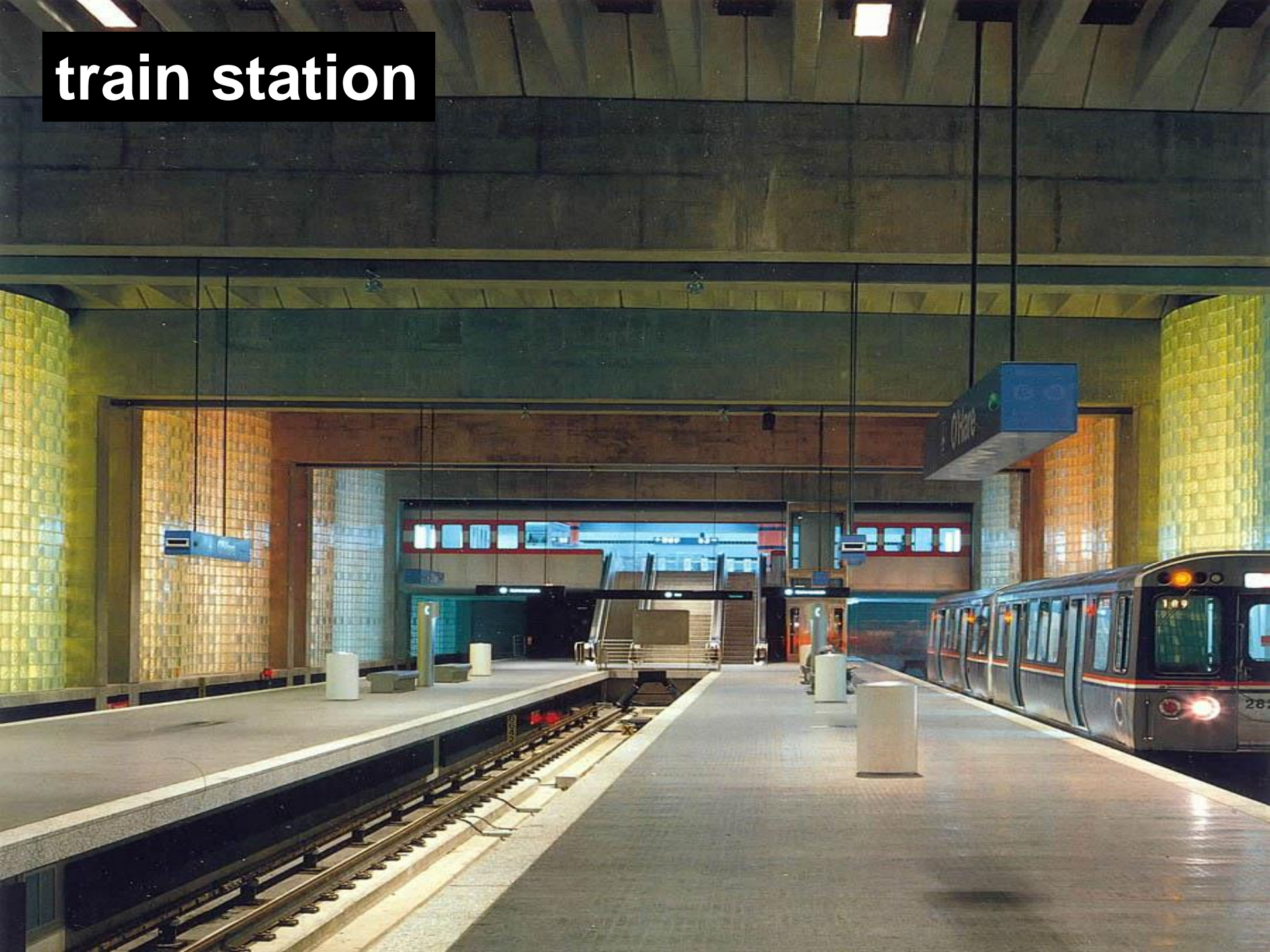
stadium



stream



train station





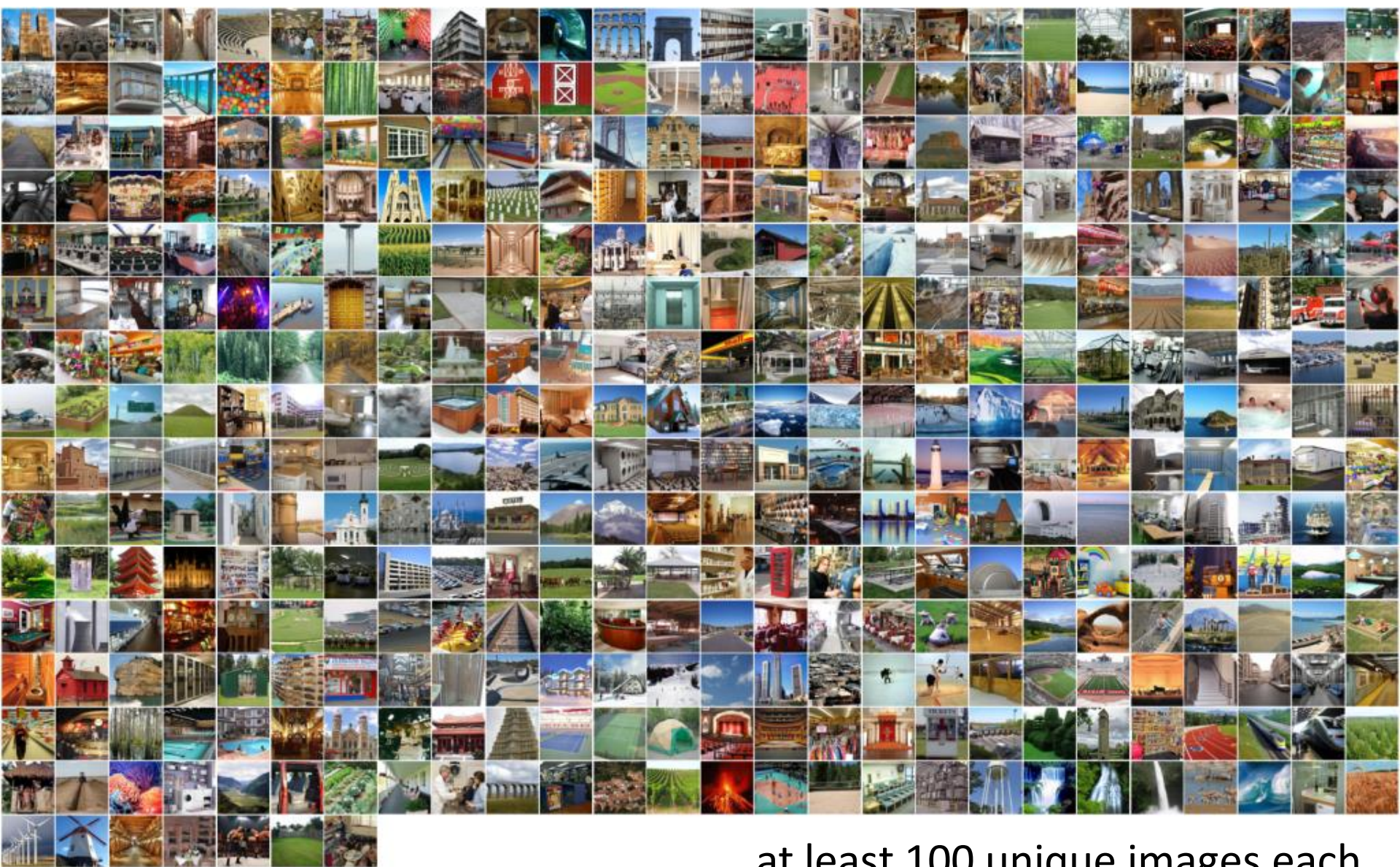


SUN Database – Xiao et al. CVPR 2010

130k images
899 categories



397 Well-sampled Categories



...at least 100 unique images each.

Evaluating Human Scene Classification



?

Accuracy

98%

90%

68%

bathroom(100%)



beauty salon(100%)



bedroom(100%)



bullring(100%)



playground(100%)



phone booth(100%)



greenhouse outdoor(100%)



podium outdoor(100%)



tennis court outdoor(100%)



wind farm(100%)



veterinarians office(100%)



riding arena(100%)



Scene category

Most confusing categories

Inn (0%)



Bayou (0%)



Basilica (0%)



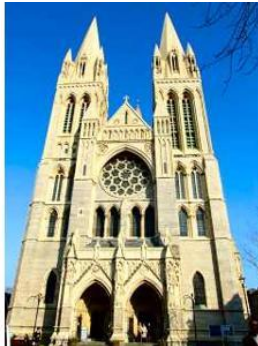
Restaurant patio (44%)



River (67%)



Cathedral(29%)



Chalet (19%)



Coast (8%)



Courthouse (21%)



Conclusion: humans can do it

- The SUN database is reasonably consistent and categories can be told apart by humans.
- With many very specific categories, humans get it right 2/3rds of the time *from experience and from exploring the label space.*

So, how do humans classify scenes?

How do we classify scenes?



Ceiling
Light
Door Door Door
Wall Door Door Wall Door
Floor

Ceiling
Lamp
mirror Painting mirror
wall
armchair Fireplace armchair
Coffee table

wall
painting
wall
Lamp
phone
alarm
Bed
Side-table
carpet

Different objects, different spatial layout

Which are the important elements?



cabinets ceiling cabinets
window window window
seat seat
seat seat
seat seat
seat seat

cabinets ceiling cabinets
window window window
seat seat
seat seat
seat seat
seat seat

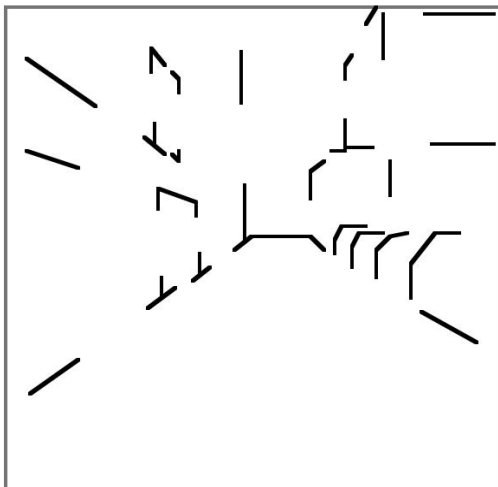
ceiling
wall column screen
seat seat
seat seat
seat seat seat seat
seat seat seat seat
seat seat seat seat

Similar objects, and similar spatial layout

Different lighting, different materials, different “stuff”

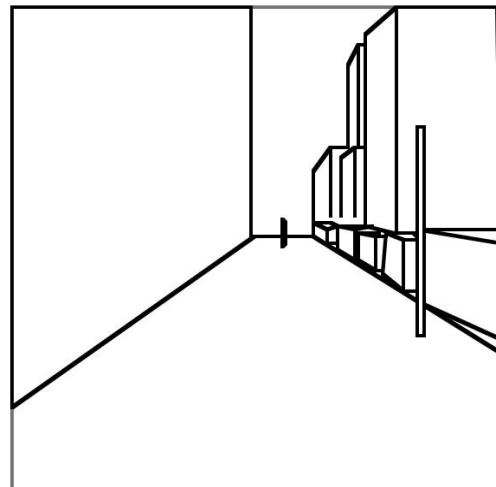
Scene emergent features

“Recognition via features that are not those of individual objects but “emerge” as objects are brought into relation to each other to form a scene.” – Biederman 81



Biederman, 1981

Suggestive edges and junctions



Biederman, 1981

Simple geometric forms



Bruner and Potter, 1969

Blobs



Oliva and Torralba, 2001

Textures

Global Image Descriptors

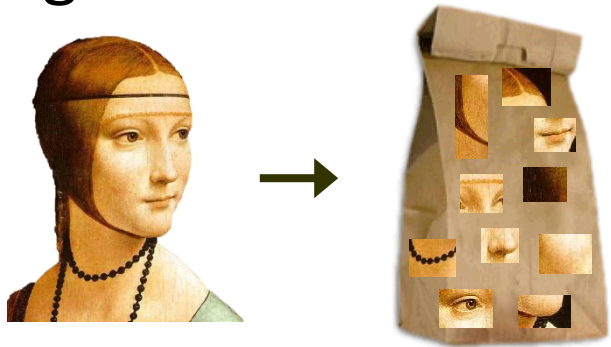
- Tiny images (Torralba et al, 2008)
- Color histograms
- Self-similarity (Shechtman and Irani, 2007)
- Geometric class layout (Hoiem et al, 2005)
- Geometry-specific histograms (Lalonde et al, 2007)
- Dense and Sparse SIFT histograms
- Berkeley texton histograms (Martin et al, 2001)
- HoG 2x2 spatial pyramids
- Gist scene descriptor (Oliva and Torralba, 2008)



Texture
Features

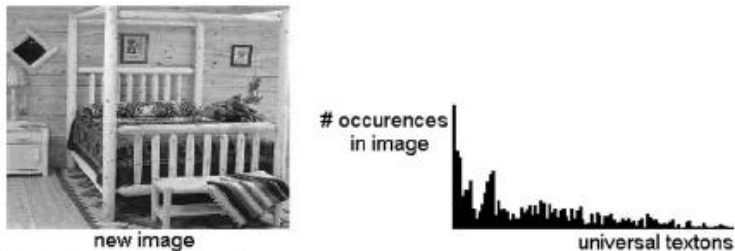
Global Texture Descriptors

Bag of words



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

Non-localized textons



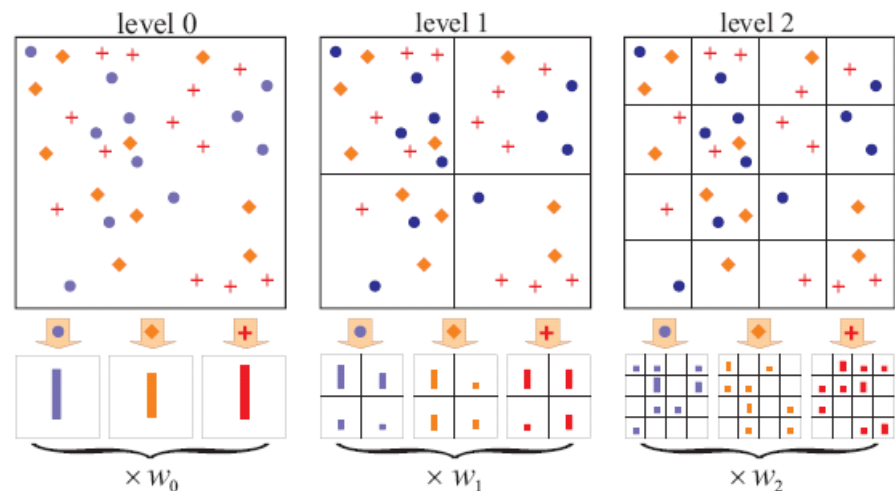
Walker, Malik. Vision Research 2004

...

Spatially organized textures



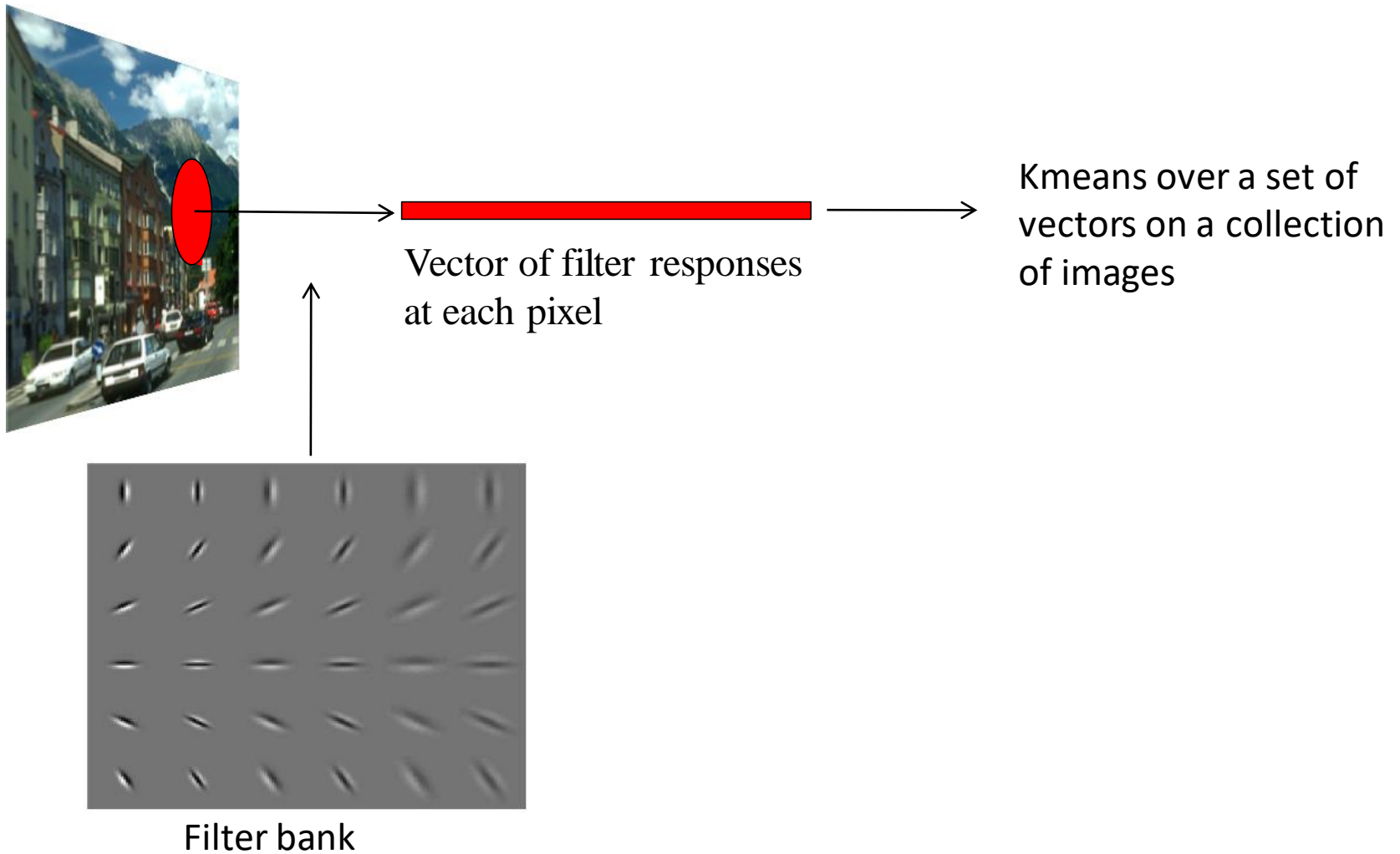
M. Gorkani, R. Picard, ICPR 1994
 A. Oliva, A. Torralba, IJCV 2001



S. Lazebnik, et al, CVPR 2006

...

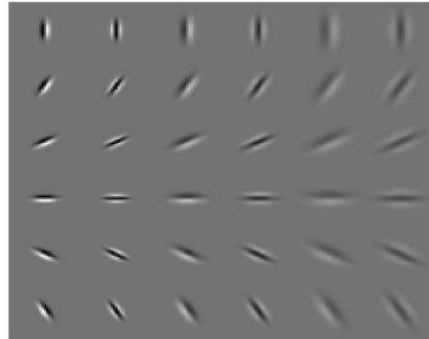
Textons



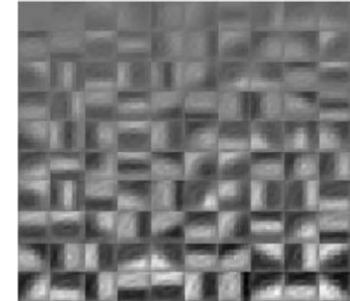
Textons



Filter bank



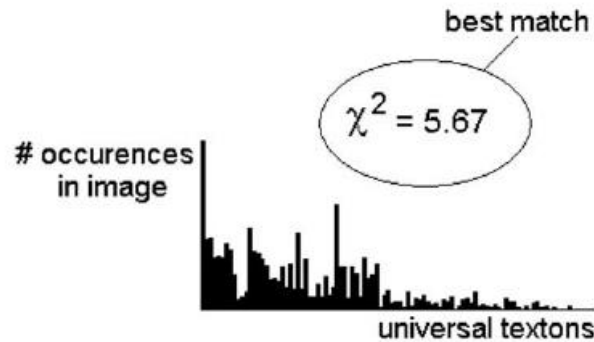
K-means (100 clusters)



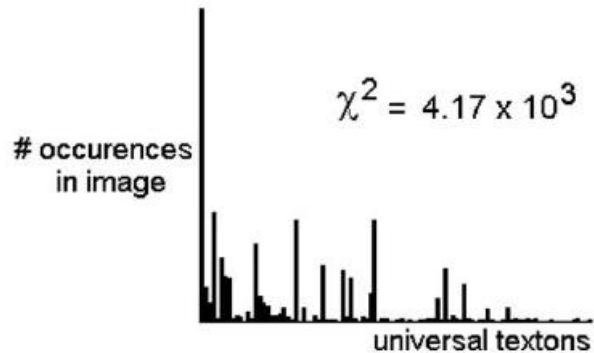
Malik, Belongie, Shi, Leung, 1999



label = bedroom



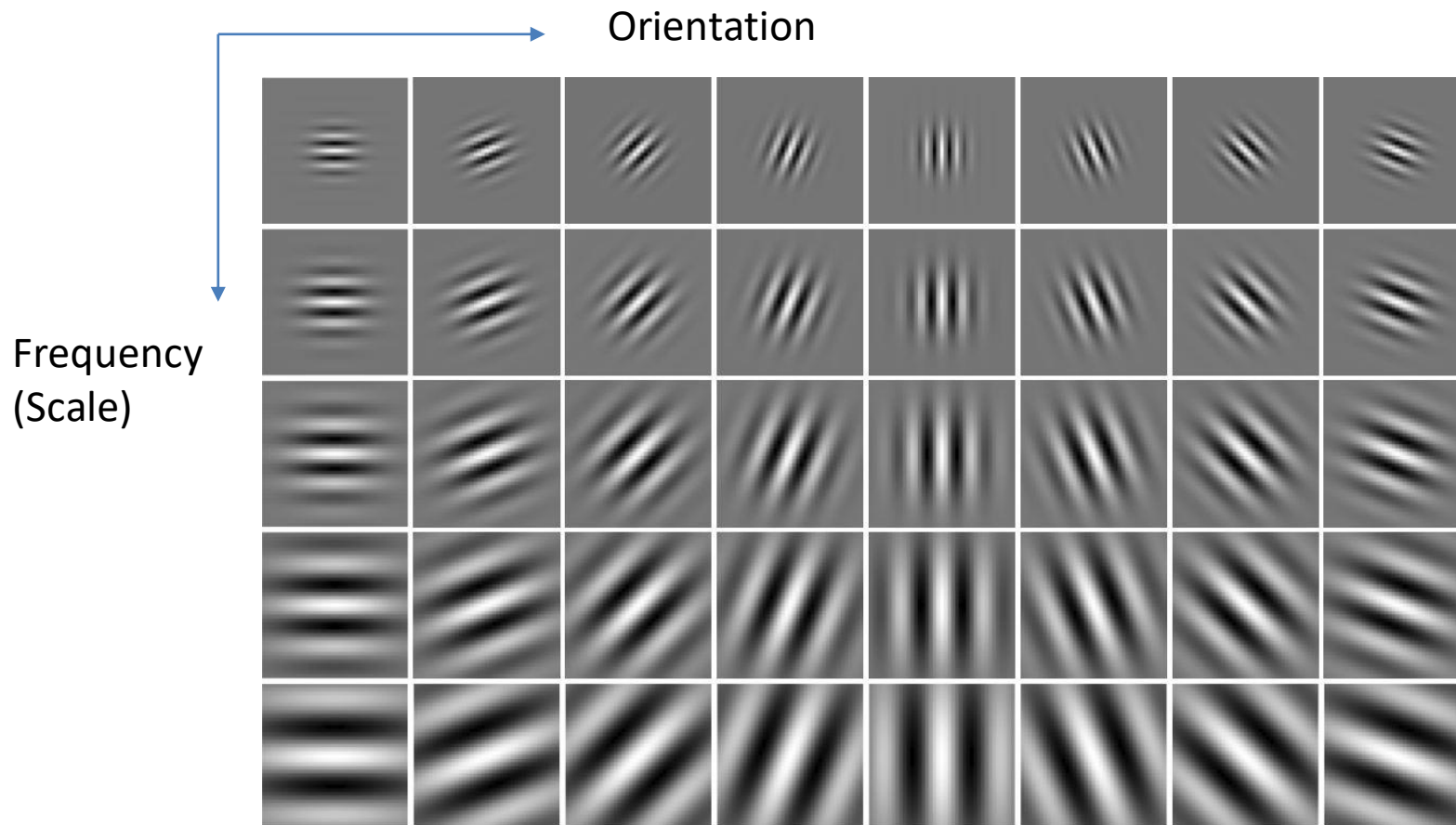
label = beach



Walker, Malik, 2004

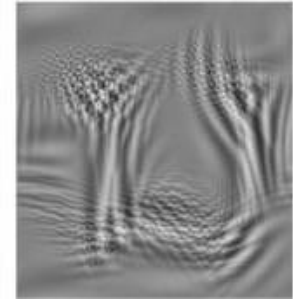
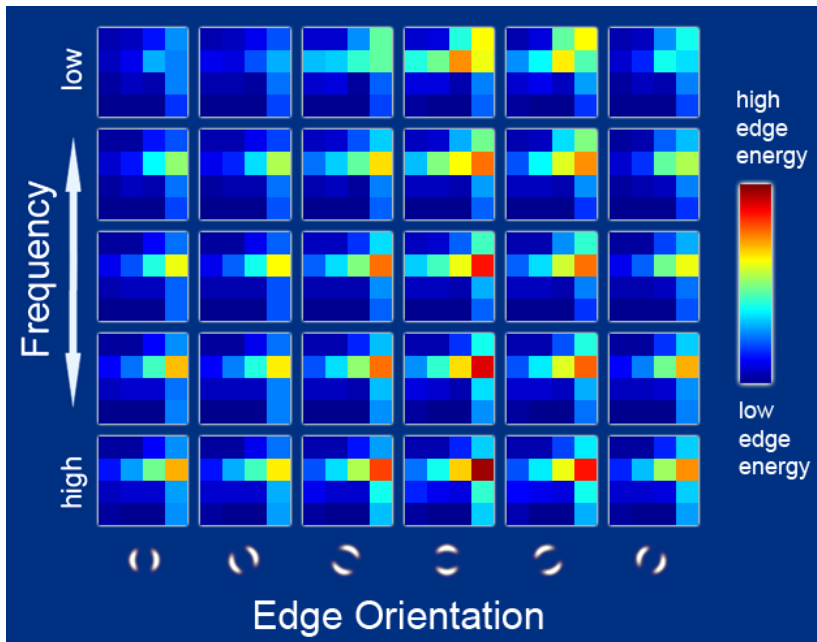
Gabor filter

- Sinusoid modulated by a Gaussian kernel



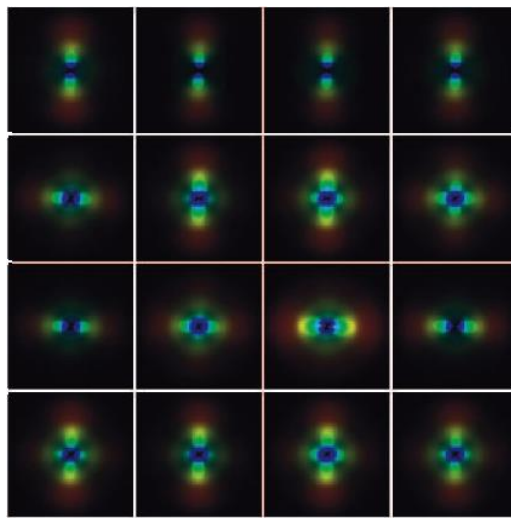
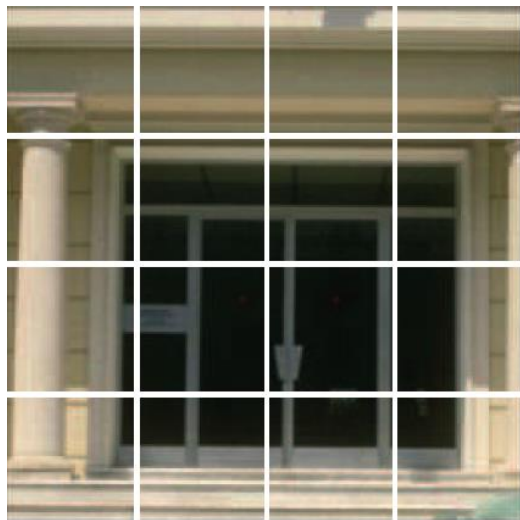
Global scene descriptors: GIST

- The “gist” of a scene: Oliva & Torralba (2001)



Gist descriptor

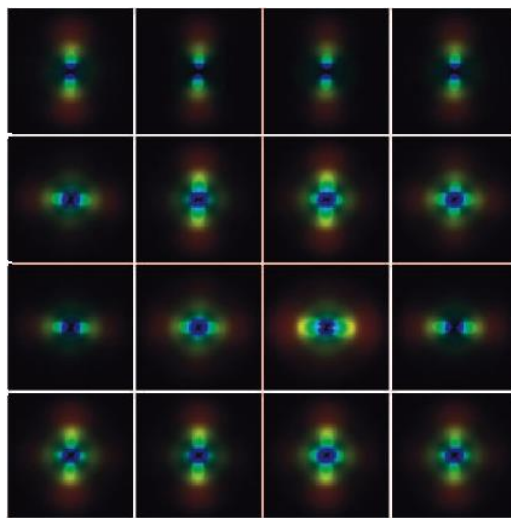
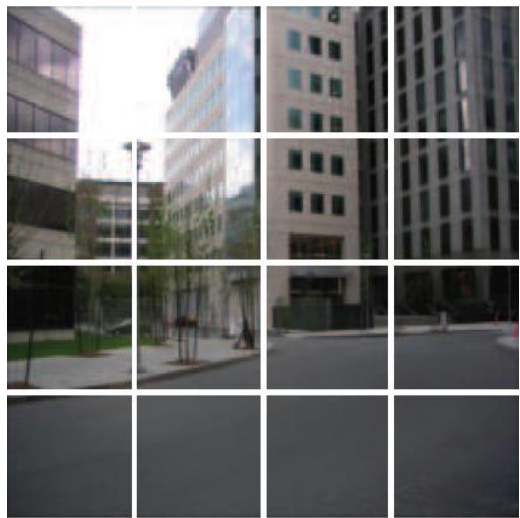
Oliva and Torralba, 2001



Apply oriented Gabor filters
over different scales.

Average filter energy per bin.

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



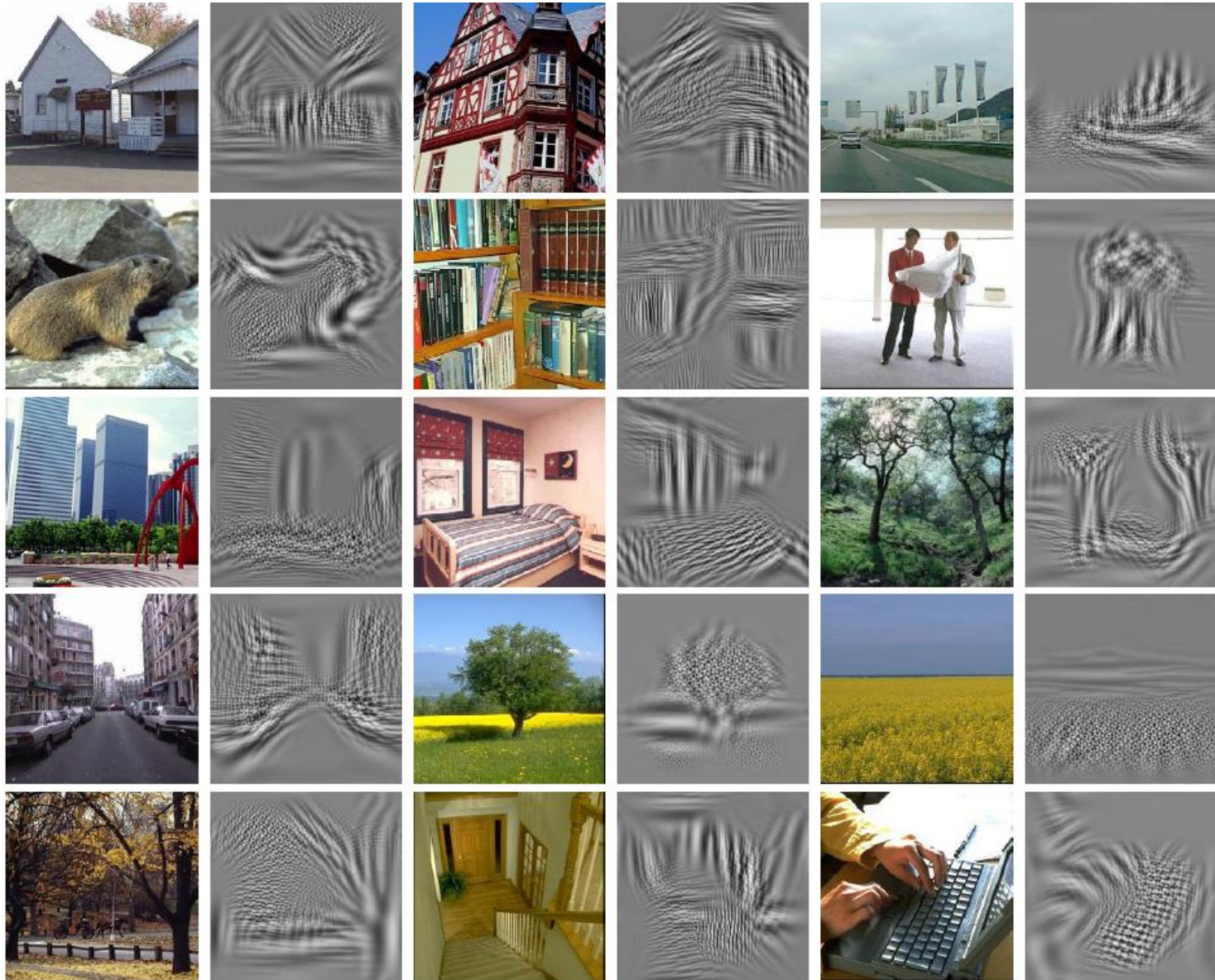
8 orientations

4 scales

x 16 bins

512 dimensions

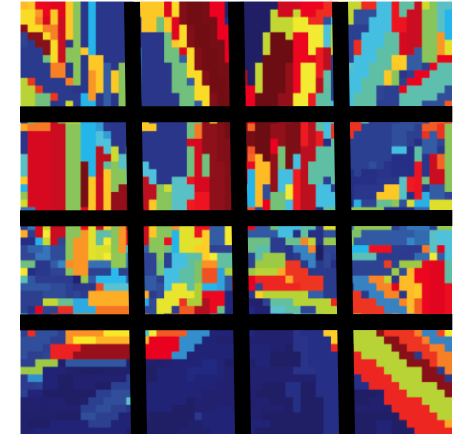
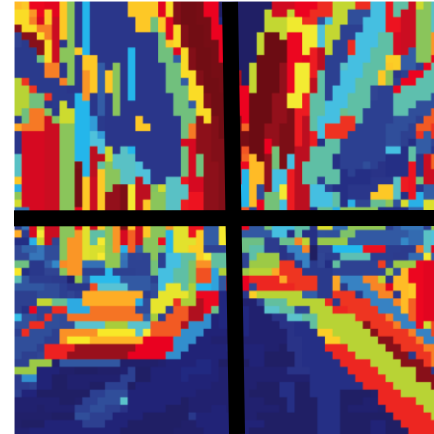
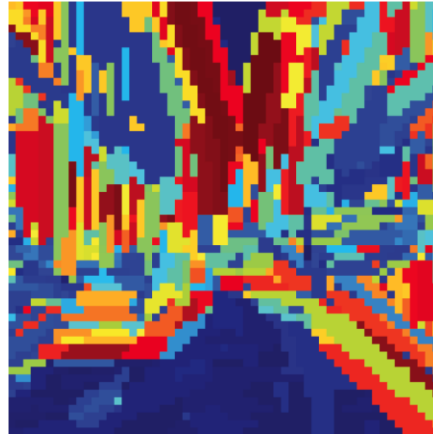
Example visual gists



Global features (I) ~ global features (I')

Bag of words & spatial pyramid matching

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

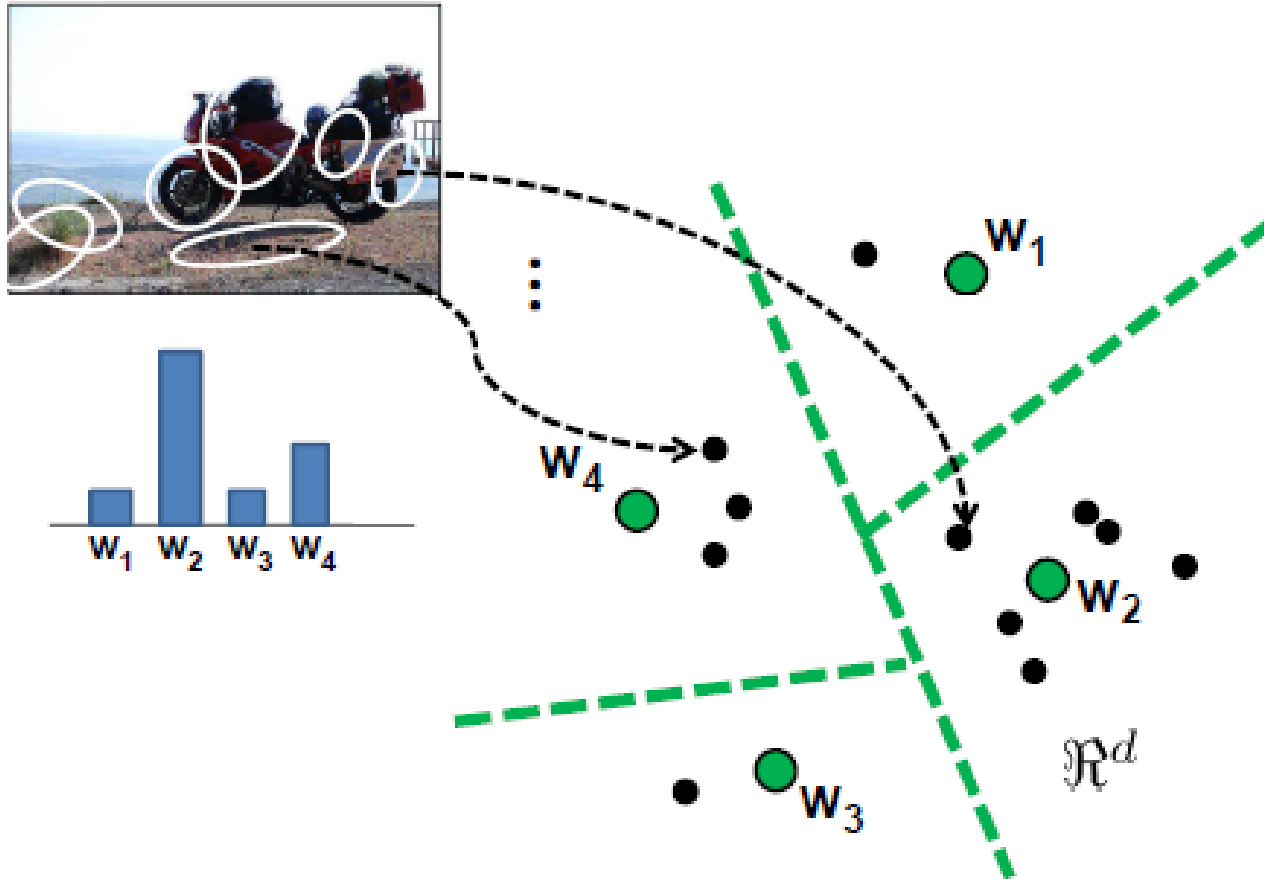


But any way to improve the quantization approach itself?

Better Bags of Visual Features

- More advanced quantization / encoding methods that are near the state-of-the-art in image classification and image retrieval.
 - Mixtures of Gaussians
 - Soft assignment (a.k.a. Kernel Codebook)
 - VLAD – Vectors of Locally-Aggregated Descriptors
- Deep learning has taken attention away from these methods...

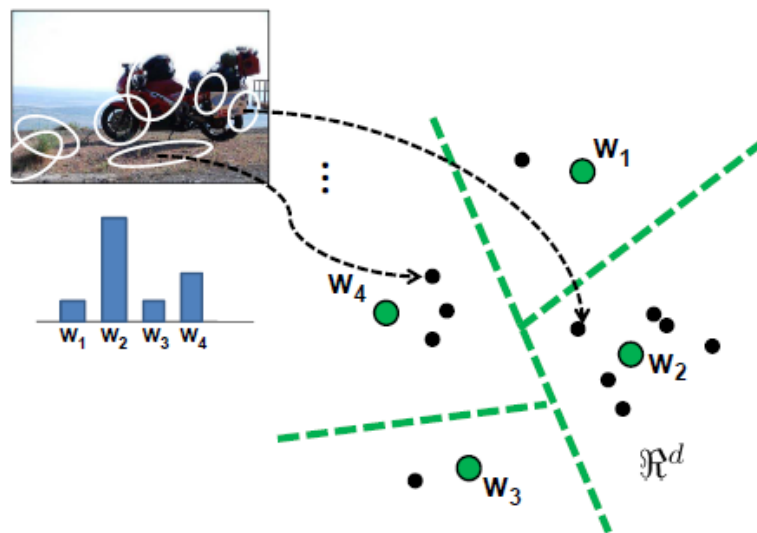
Standard K-means Bag of Words



Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**?

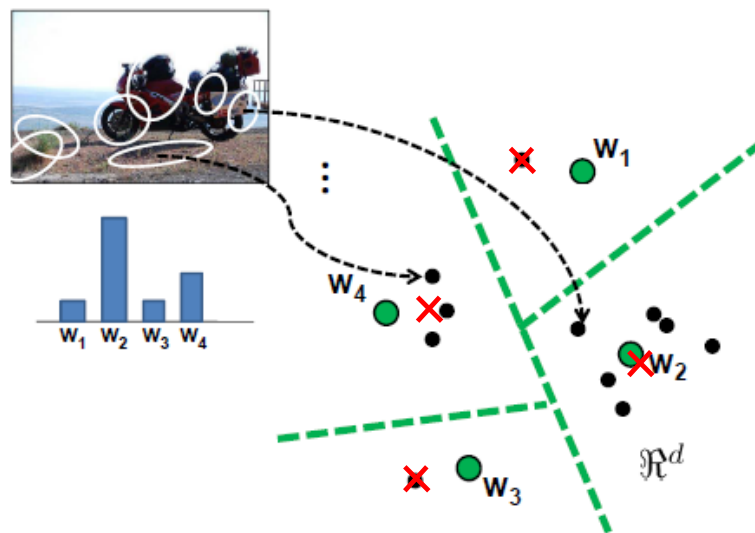


Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**? For instance:

- mean of local descriptors ✗



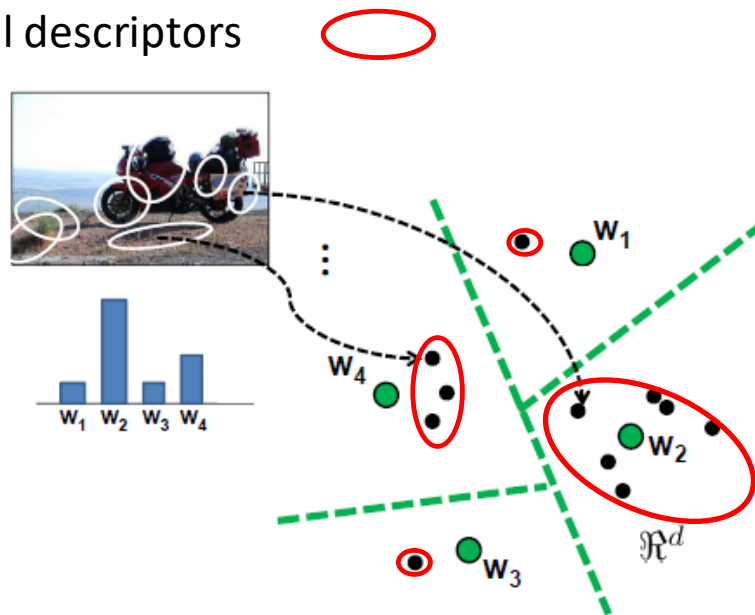
http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**? For instance:

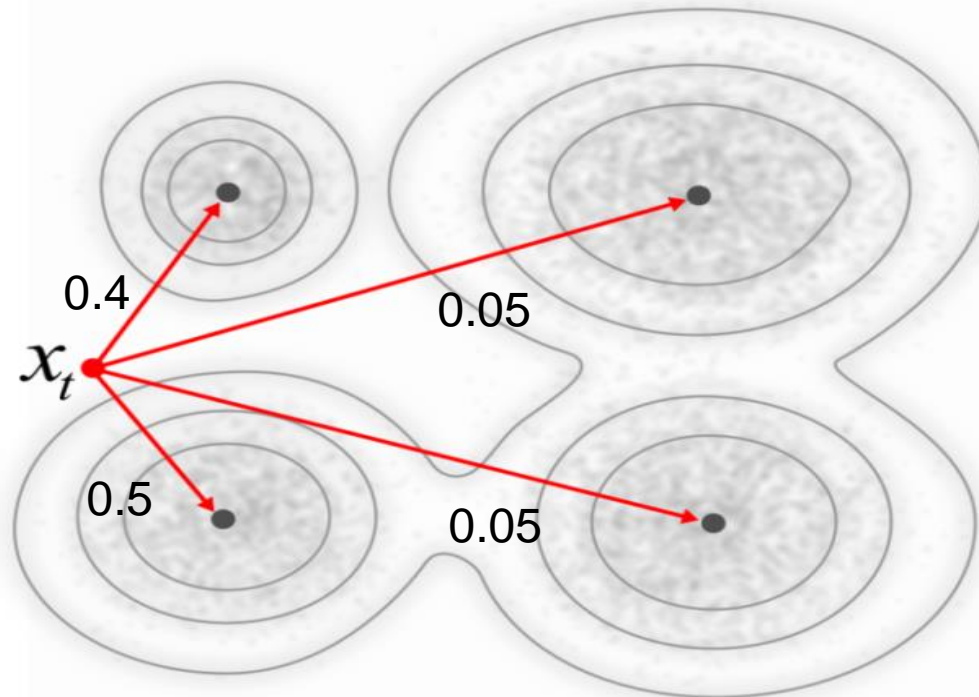
- mean of local descriptors
- (co)variance of local descriptors



http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

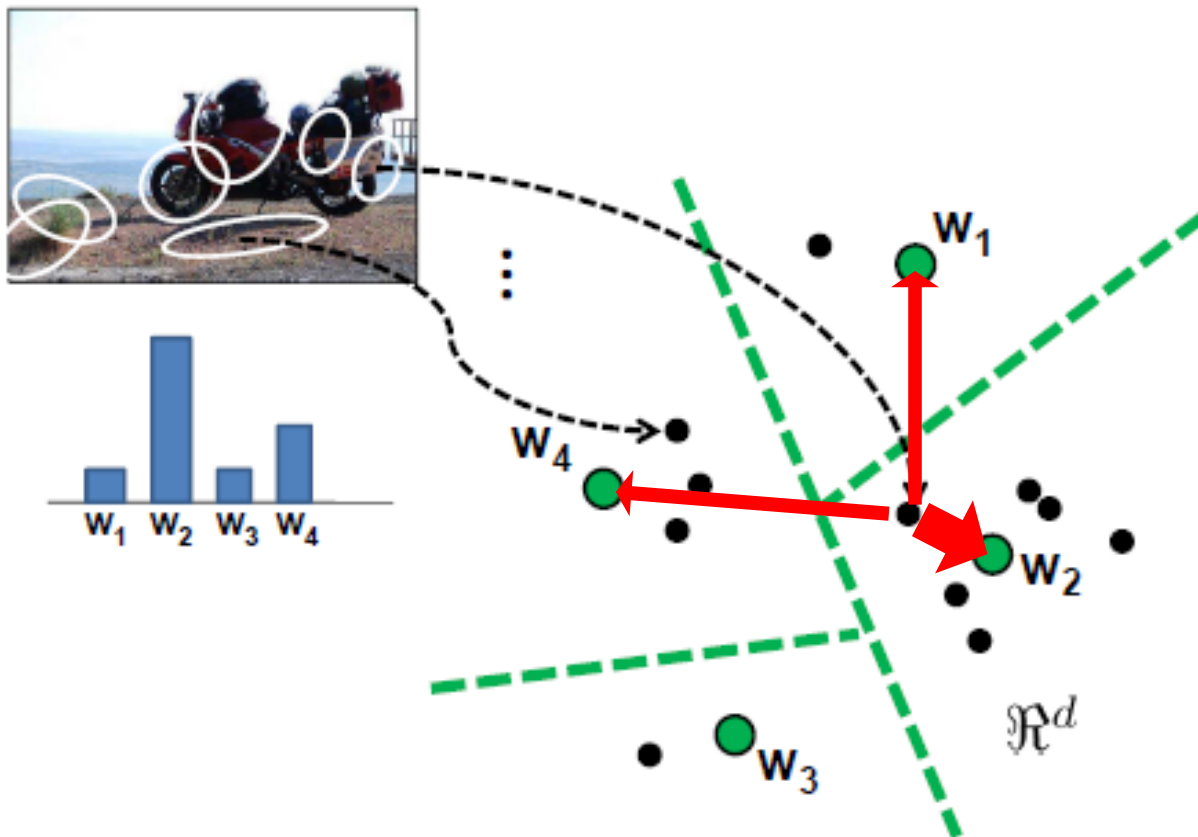
Gaussian Mixture Model (GMM)

- GMM can be thought of as “soft” k-means.
- Each component has a mean and a standard deviation along each direction (or full covariance)
- Can easily represent non-circular distributions



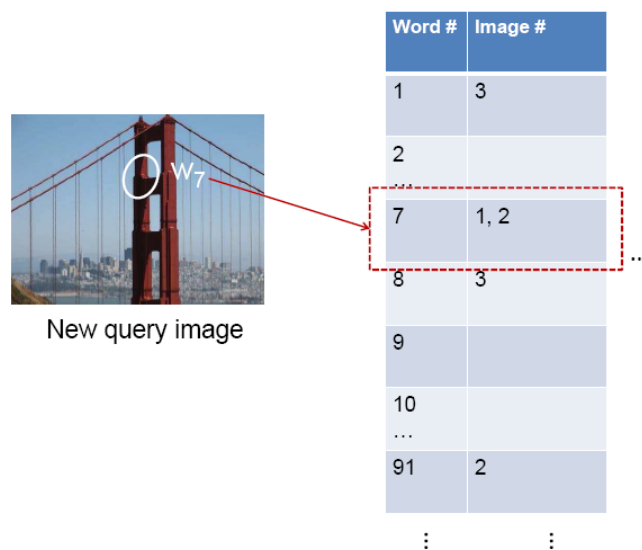
Simple case: Soft Assignment

- “Kernel codebook encoding” by Chatfield et al. 2011.
- Cast a set of proportional votes (weights) to n most similar clusters, rather than a single ‘hard’ vote.



Simple case: Soft Assignment

- “Kernel codebook encoding” by Chatfield et al. 2011.
- Cast a set of proportional votes (weights) to n most similar clusters, rather than a single ‘hard’ vote.
- This is fast and easy to implement, but it makes an inverted file index *less sparse*.

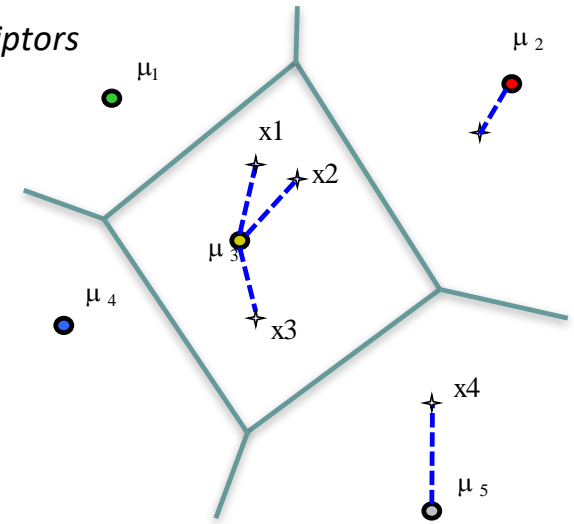


VLAD – Vectors of Locally-Aggregated Descriptors

Given a codebook $\{\mu_i, i = 1 \dots N\}$,
e.g. learned with K-means, and a set of
local descriptors $X = \{x_t, t = 1 \dots T\}$

① assign: $\text{NN}(x_t) = \arg \min_{\mu_i} \|x_t - \mu_i\|$

① assign descriptors



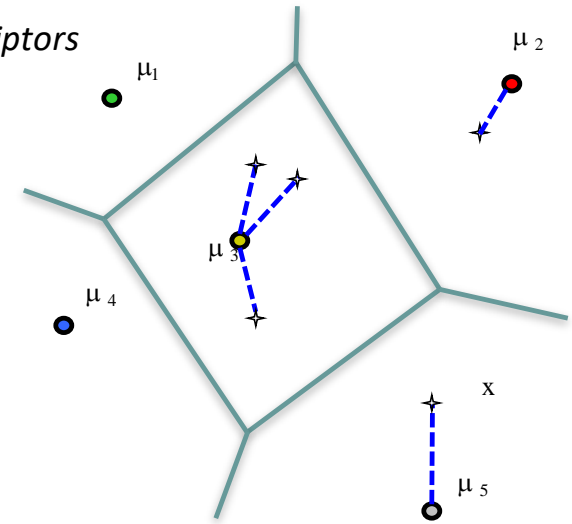
VLAD – Vectors of Locally-Aggregated Descriptors

Given a codebook $\{\mu_i, i = 1 \dots N\}$,
e.g. learned with K-means, and a set of
local descriptors $X = \{x_t, t = 1 \dots T\}$

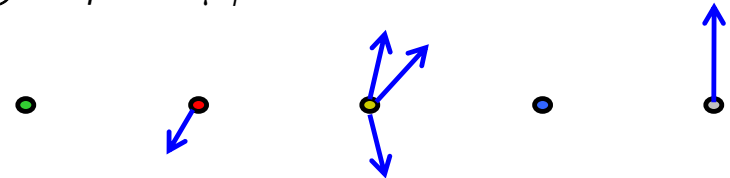
① assign: $\text{NN}(x_t) = \arg \min_{\mu_i} \|x_t - \mu_i\|$

②③ compute: $v_i = \sum_{x_t: \text{NN}(x_t) = \mu_i} x_t - \mu_i$

① assign descriptors



② compute $x - \mu_i$



VLAD – Vectors of Locally-Aggregated Descriptors

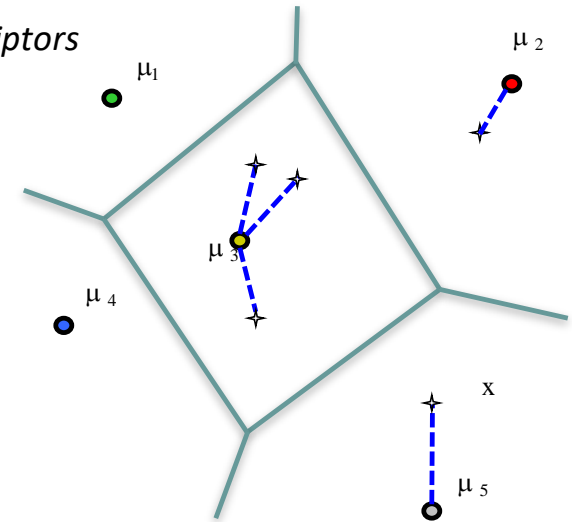
Given a codebook $\{\mu_i, i = 1 \dots N\}$,
e.g. learned with K-means, and a set of
local descriptors $X = \{x_t, t = 1 \dots T\}$

① assign: $\text{NN}(x_t) = \arg \min_{\mu_i} \|x_t - \mu_i\|$

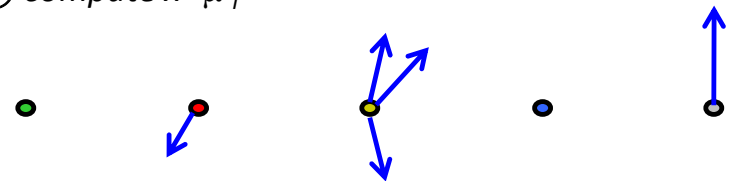
②③ compute: $v_i = \sum_{x_t: \text{NN}(x_t) = \mu_i} x_t - \mu_i$

- concatenate v_i 's + ℓ_2 normalize

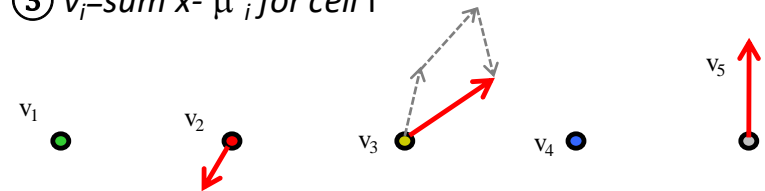
① assign descriptors



② compute $x - \mu_i$

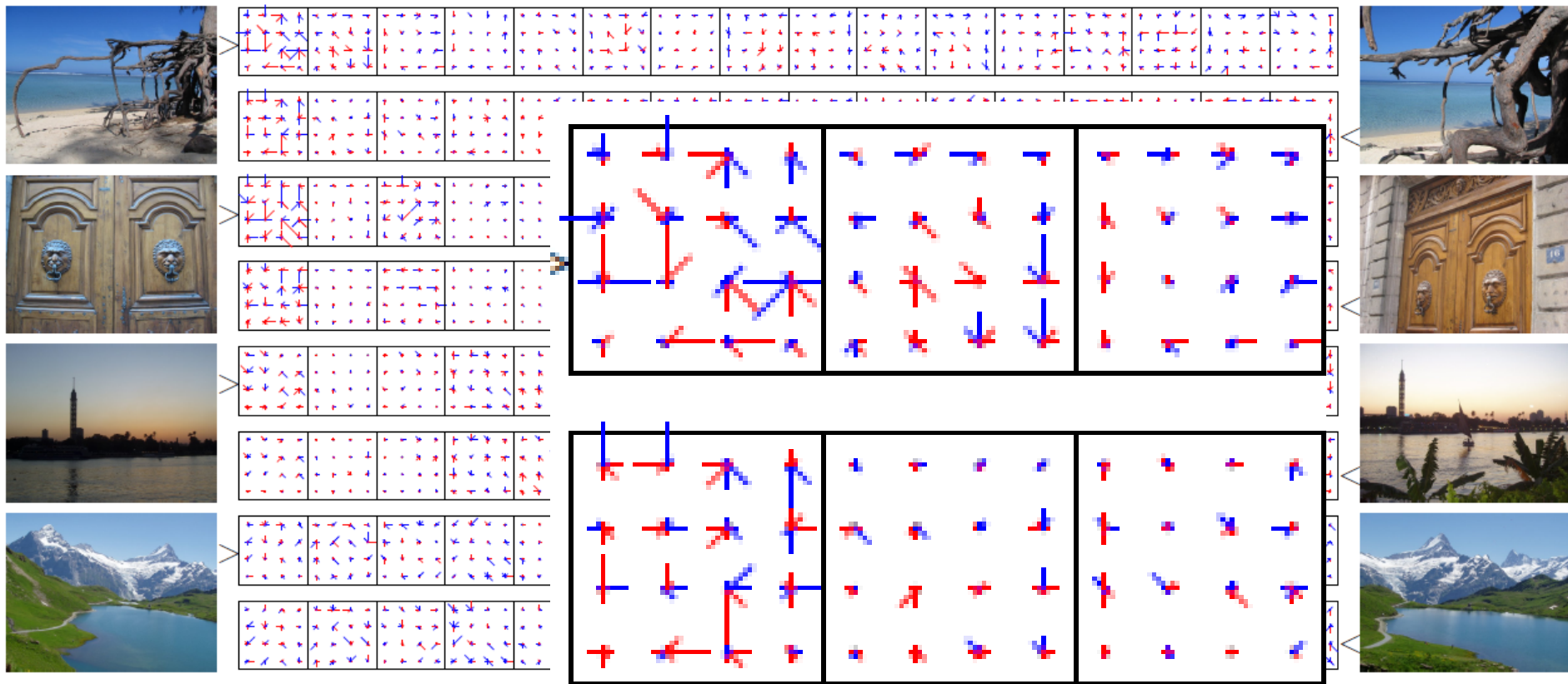


③ $v_i = \text{sum } x - \mu_j \text{ for cell } i$



A first example: the VLAD

A graphical representation of
$$v_i = \sum_{x_t: \text{NN}(x_t) = \mu_i} x_t - \mu_i$$



Jégou, Douze, Schmid and Pérez,
"Aggregating local descriptors into a compact image representation",
CVPR'10.

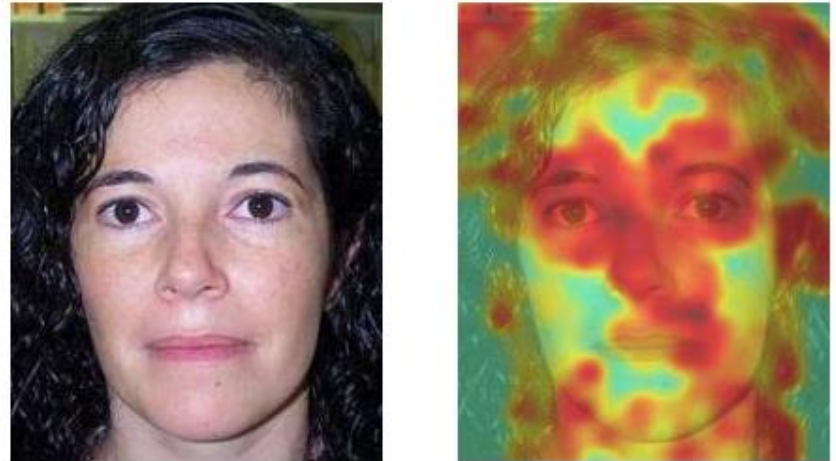
Why can't we train good recognition systems?

- Training Data
 - Huge issue, but not always a variable we control.
- Representation
 - Are the local features themselves lossy?
 - What about feature quantization?

What about skipping quantization completely?

In Defense of Nearest-Neighbor Based Image Classification
Boiman, Shechtman, Irani

Quantization inherently averages the parts which are *most discriminative* !!!



Quantization error of densely computed image descriptors (SIFT) using a large codebook (size 6,000) of Caltech-101. Red = high error; Blue = low error. The most informative descriptors (eye, nose, etc.) have the highest quantization error

What about NN image-to-image matching?

In Defense of Nearest-Neighbor Based Image Classification
Boiman, Shechtman, Irani

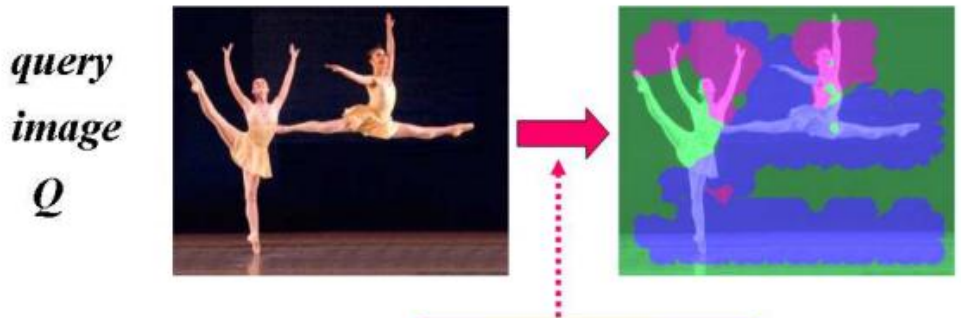


Image to class features NN: $KL(p_Q | p_C) = 8.35$

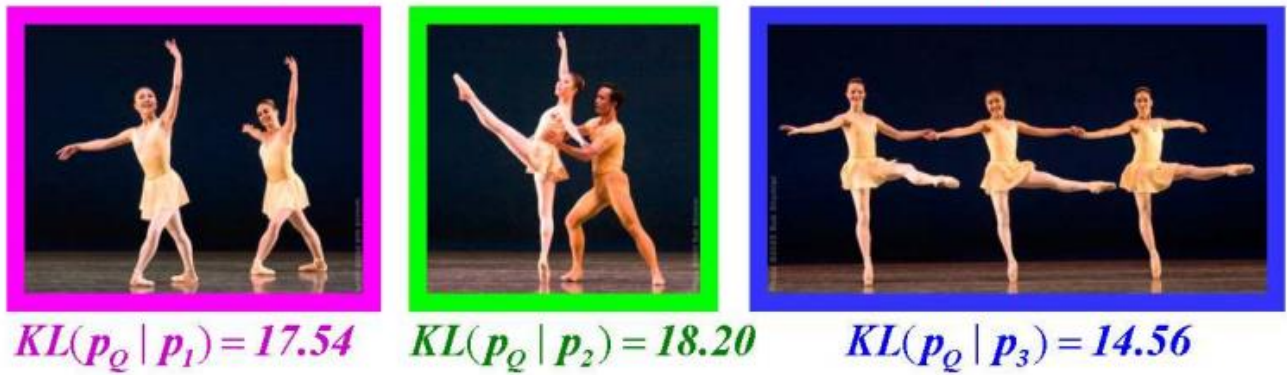
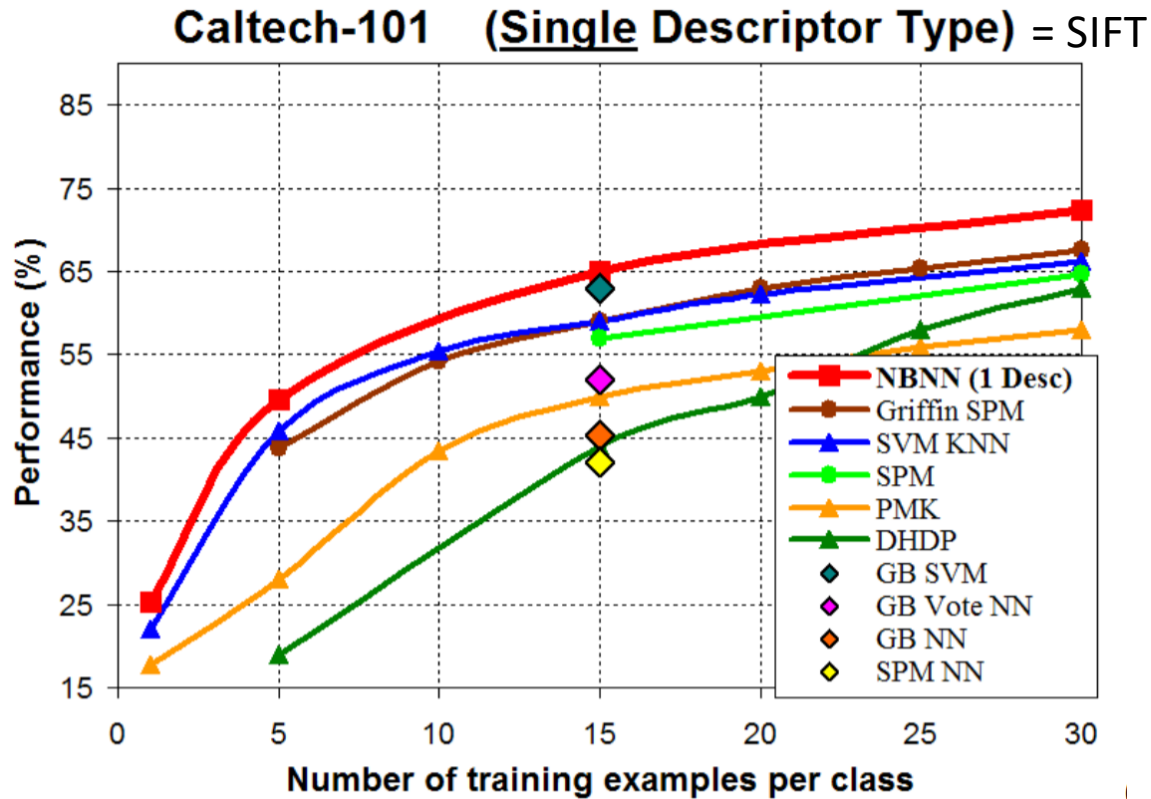


Image to image features NN

If I do both of these, NN can be a pretty good classifier!



In Defense of Nearest-Neighbor Based Image Classification
Boiman, Shechtman, Irani

Summary

- Methods to better characterize the distribution of visual words in an image:
 - Soft assignment (a.k.a. Kernel Codebook)
 - VLAD
 - No quantization

Learning Scene Categorization

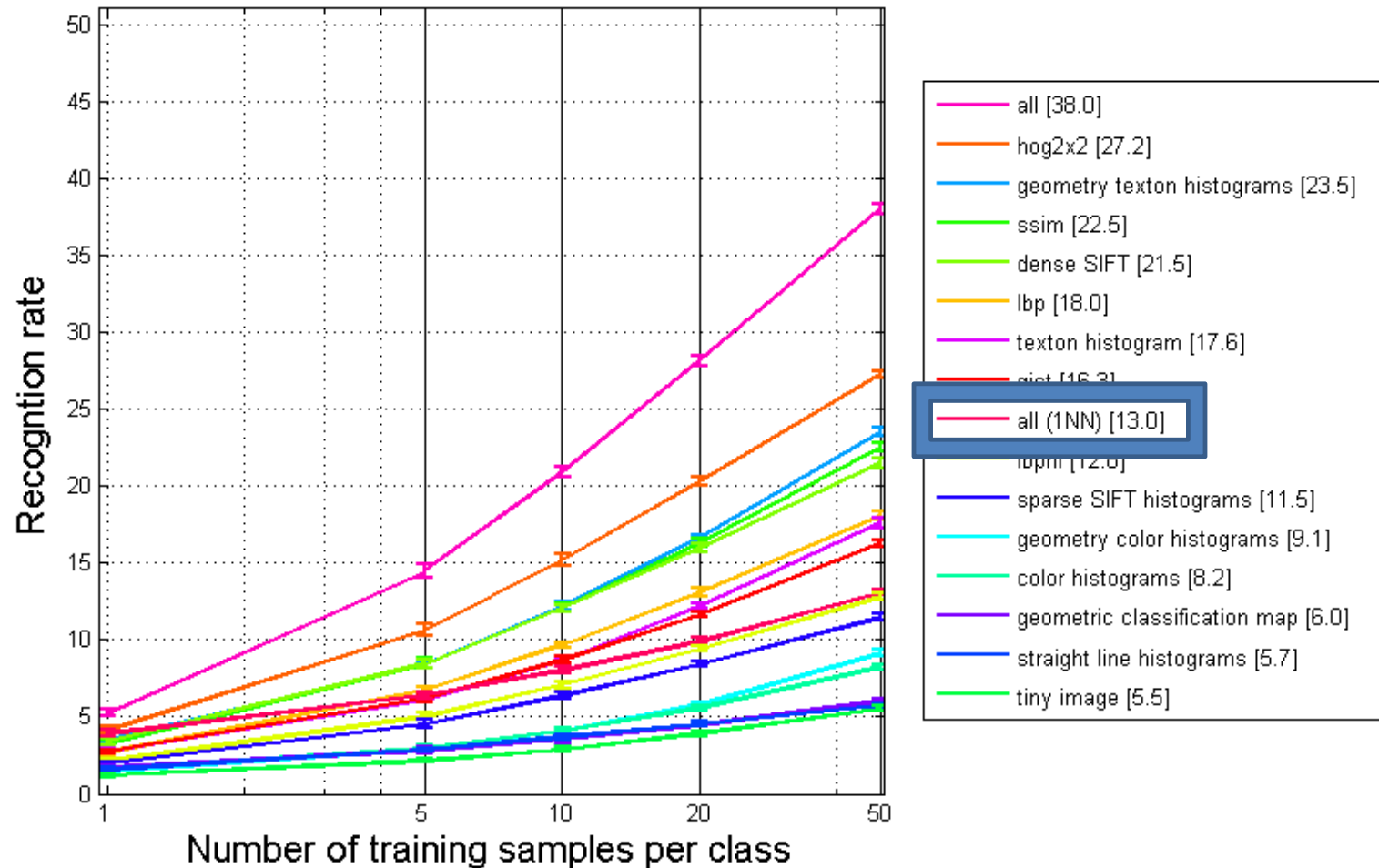


Forest path
Vs.
all

Living - room
Vs.
all

Feature Accuracy

Humans [68.5]



Classifier: 1-vs-all SVM with histogram intersection, chi squared, or RBF kernel.

A look into the results

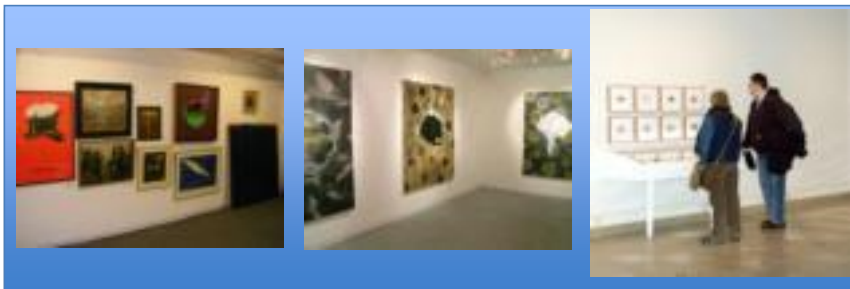
Airplane cabin (64%)



Van interior Discotheque Toyshop



Art gallery (38%)



Iceberg Hotel room Kitchenette



All the results available on the web

...

limousine interior
(95% vs 80%)



riding arena
(100% vs 90%)



sauna
(96% vs 95%)



skatepark
(96% vs 90%)



subway interior
(96% vs 80%)



**Humans good
Comp. good**

**Humans bad
Comp. bad**

**Human good
Comp. bad**

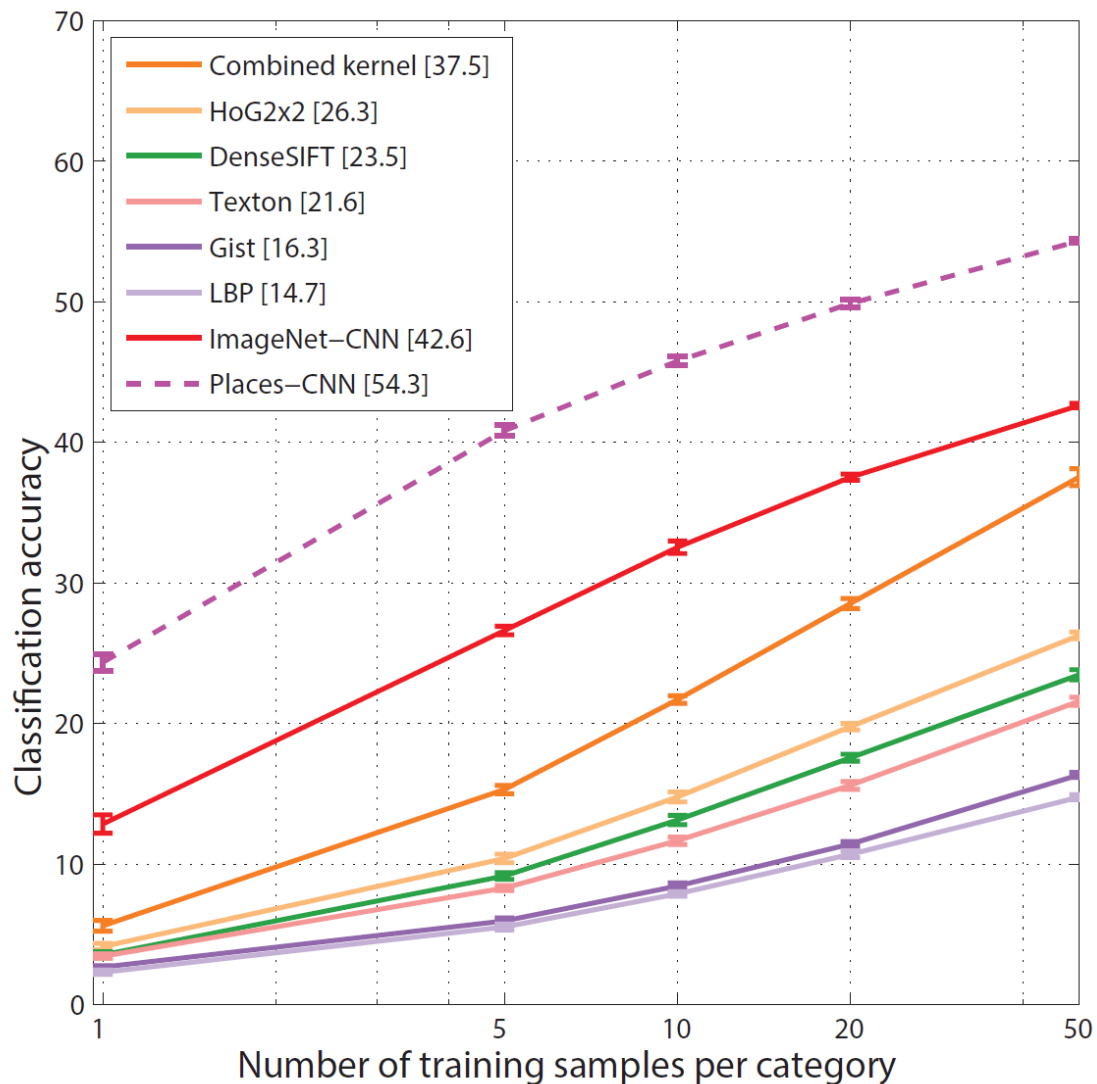
**Human bad
Comp. good**

How do we do better than 40%?

- Features from deep learning based on ImageNet allow us to reach 42%...

Not much better...

Benchmark on SUN397 Dataset



B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. "Learning Deep Features for Scene Recognition using Places Database." Advances in Neural Information Processing Systems 27 (NIPS), 2014