Global Scene Representations

Tilke Judd

Papers

- Oliva and Torralba [2001]
- Fei Fei and Perona [2005]
- Labzebnik, Schmid and Ponce [2006]

Commonalities

- Goal: Recognize natural scene categories
- Extract features on images and learn models
- Test on database of scenes
- in general, accuracy or generality improves

Past theories

- Scene recognition based on
  - edges, surfaces, details
  - successive decision layers of increasing complexity
  - object recognition

But now...

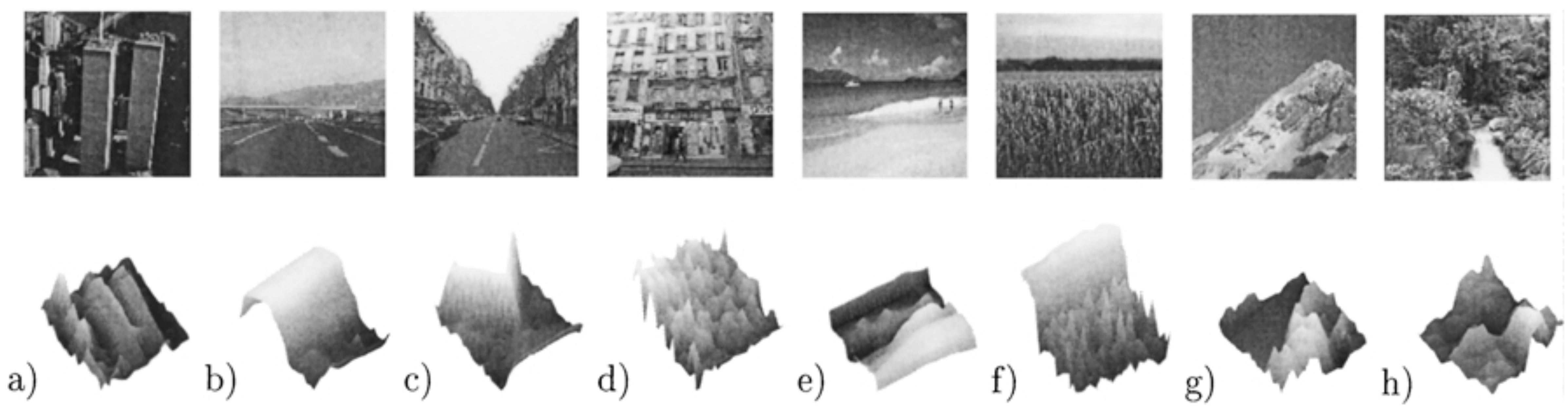
- Scene recognition may be initiated by low resolution global configuration
  - enough information about meaning of scene in < 200ms [Potter 1975]
  - understanding driven from arrangements of simple forms or “geons” [Biederman 1987]
  - spatial relationship between blobs of specific size and aspect ratios [Schyns and Oliva 1994, 1997]

Modeling the Shape of the Scene: A Hollistic Representation of the Spatial Envelope

Aude Oliva and Antonio Torralba 2001

Shape of a scene

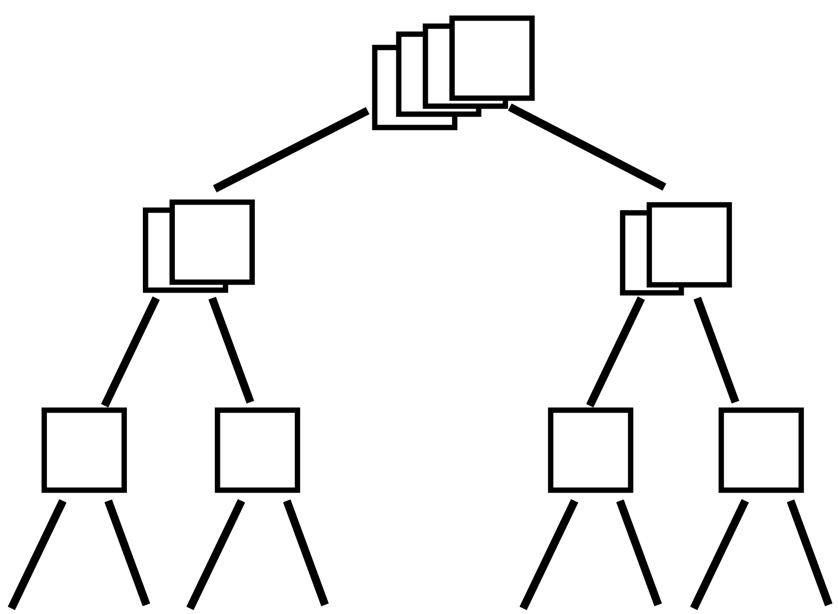
- Pose a scene as a SHAPE instead of a collection of objects
- Show scenes of same category have similar shape or spatial structure



[Image from Oliva and Torralba 2001]

Spatial Envelope

- Design experiment to identify meaningful *dimensions* of scene structure
- Split 81 pictures into groups then describe them



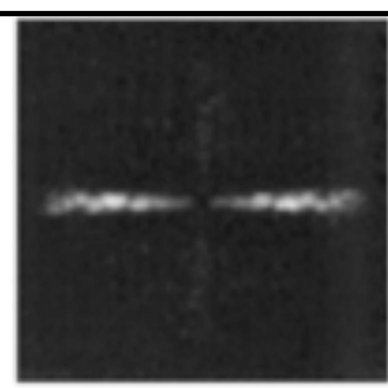
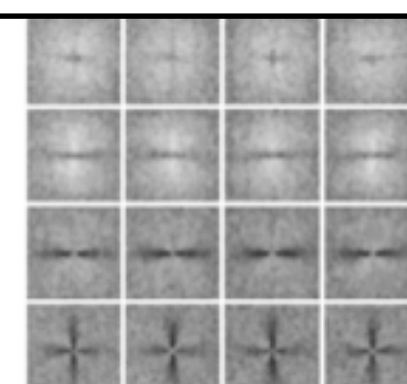
Used words like
“man-made” vs “natural”
“open” vs “closed”

Spatial Envelope

- 5 Spatial Envelope Properties
  - Degree of *Naturalness*
  - Degree of *Openness*
  - Degree of *Roughness*
  - Degree of *Expansion*
  - Degree of *Ruggedness*
- Goal: to show these 5 qualities adequate to get high level description of scene

Modeling Spatial Envelope

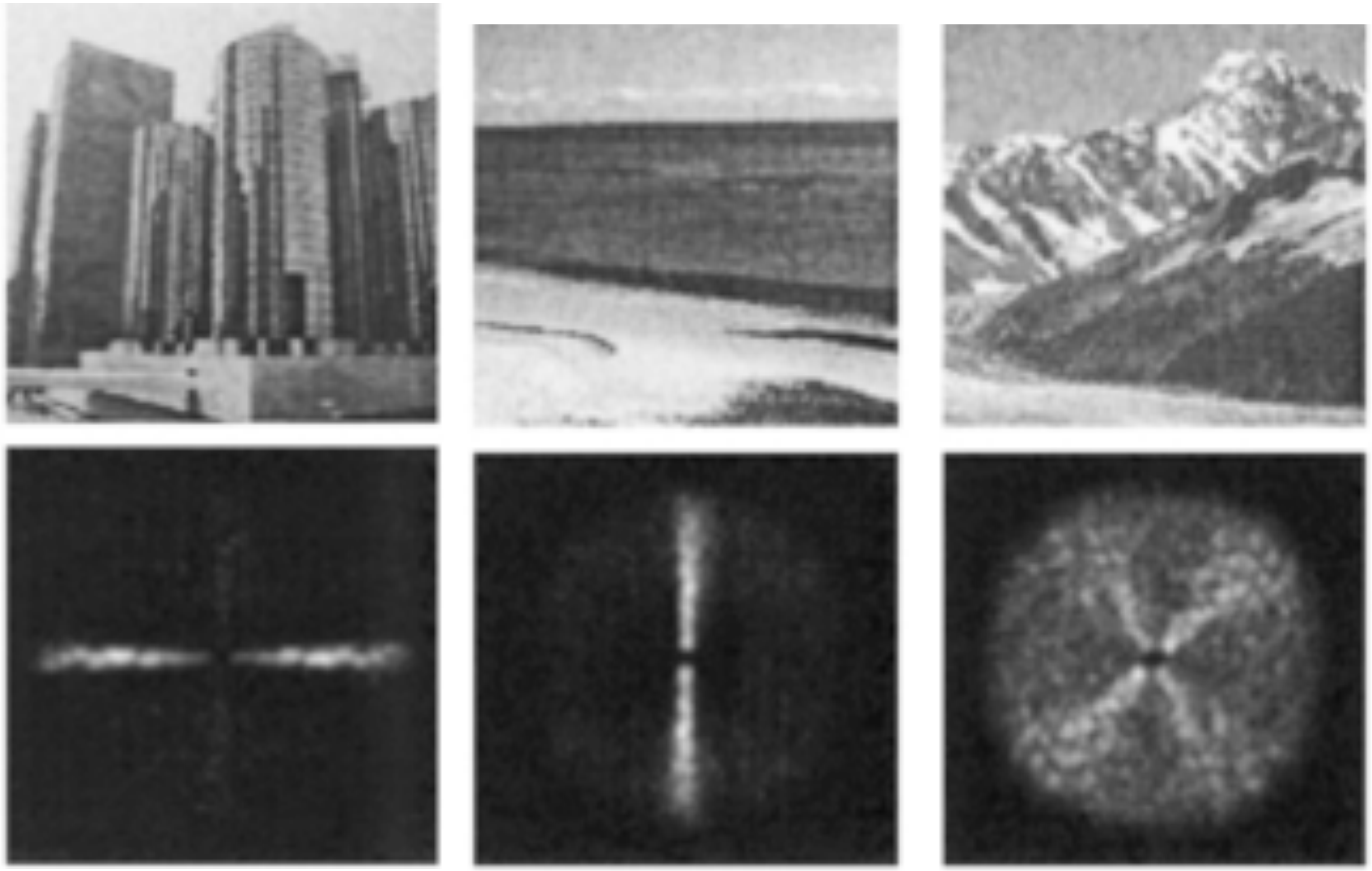
- Introduce 2nd order statistics based on Discrete Fourier Transform

Energy Spectrum	**Spectrogram**
squared magnitude of FT = distribution of the signal's energy among different spatial frequencies	spatial distribution of spectral information
DFT	Windowed DFT
unlocalized dominant structure	structural info in spatial arrangement
	
good results	more accurate

Both are high dimensional representation of scene

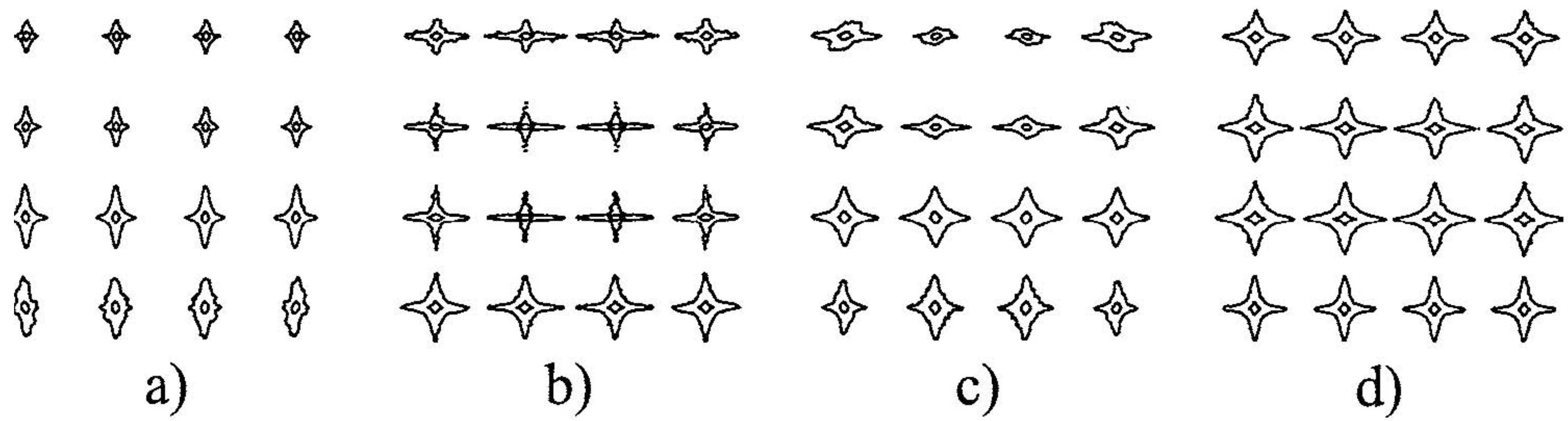
Reduced by PCA to set of orthogonal functions with decorrelated coefficients

Energy Spectrum



Mean Spectrogram

- Structural aspects are modeled by energy spectrum and spectrogram



Man made open | urban vertical | perspective view of streets | far view of city center buildings

Mean spectrogram from hundreds of same category

[Image from Oliva and Torralba 2001]

Learning

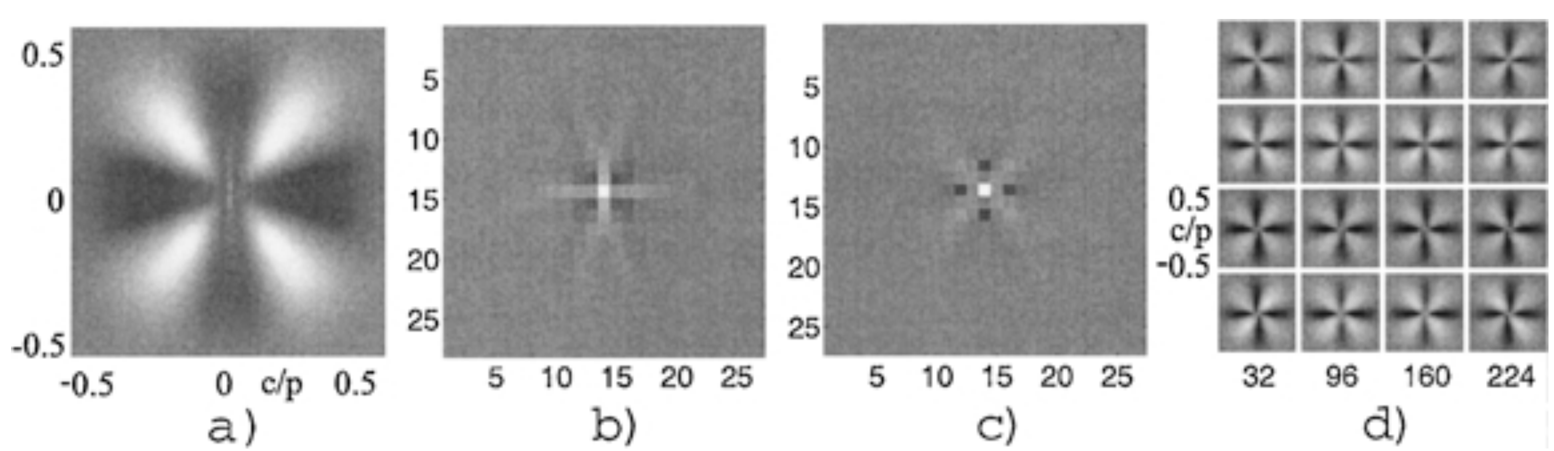
- How can Spatial Envelope propertie s be estimated by global spectral features v?
- Simple linear regression
  - 500 images placed on axis of desired property
  - used for learning regression model parameters d

$$\hat{s} = \mathbf{v}^T \mathbf{d} = \sum^{N_G} v_i \, d_i$$
$$= \iint A(f_x, f_y)^2 \, DST(f_x, f_y) \, df_x \, df_y$$

  - s = amplitude spectrum * Discriminant Spectral Template (DST)
  - Use regression for continuous features and binary features

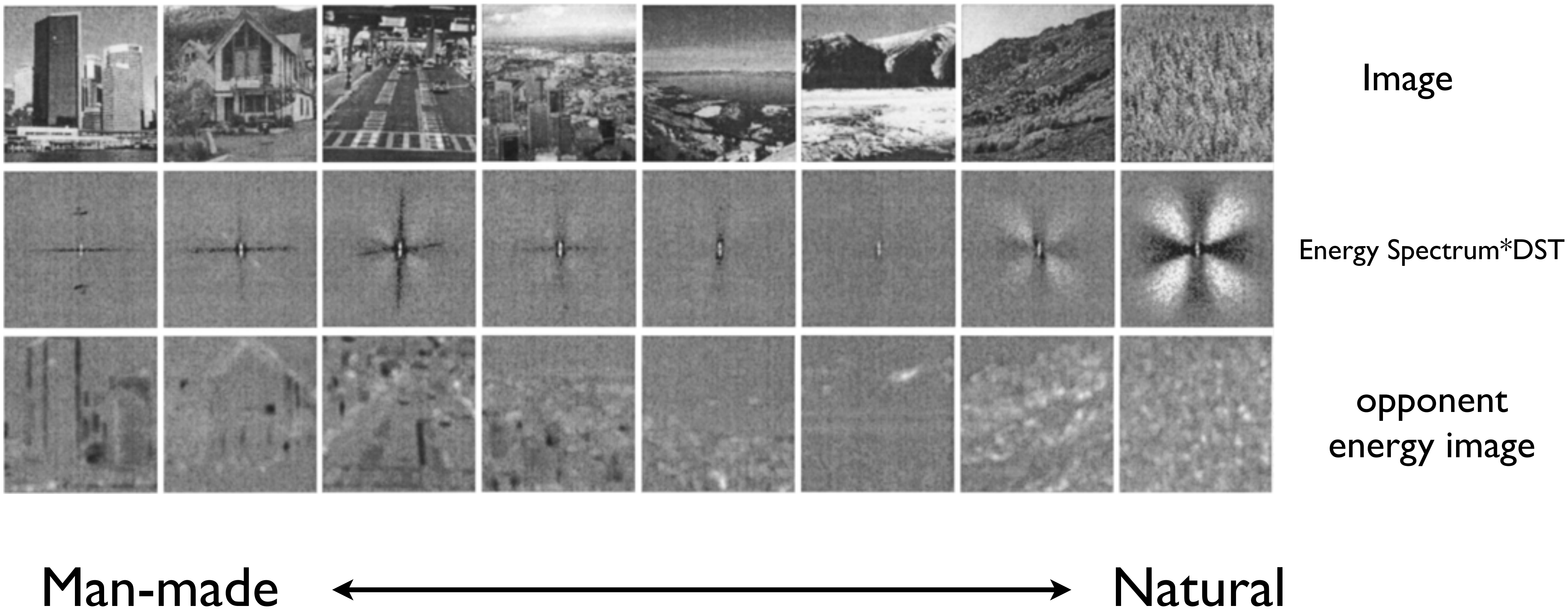
DST

- show how spectral components of energy spectrum should be weighted
- example: natural vs man-made
  - white: high degree of naturalness at low diagonal frequencies
  - black: low degree of naturalness at H and V frequencies



DST WDST

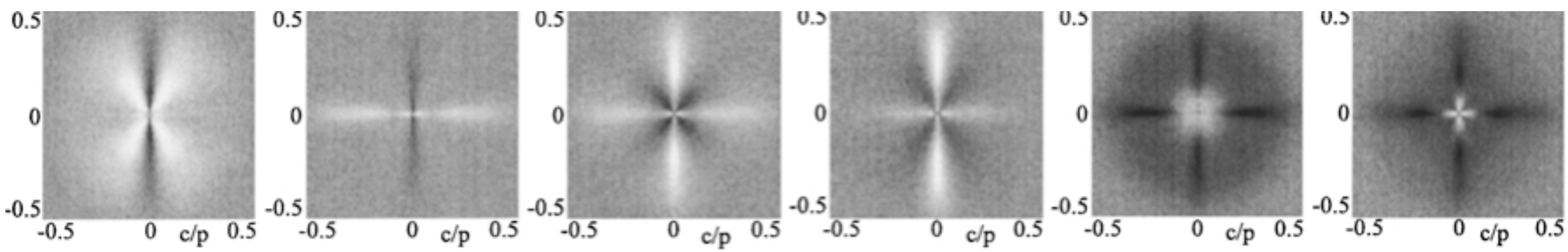
Naturalness



Value of naturalness = sum (Energy Spectra * DST)

Leads to 93.5% correct classification of 5000 test scenes

DST for other properties



Natural openness

Man-made openness

Natural ruggedness

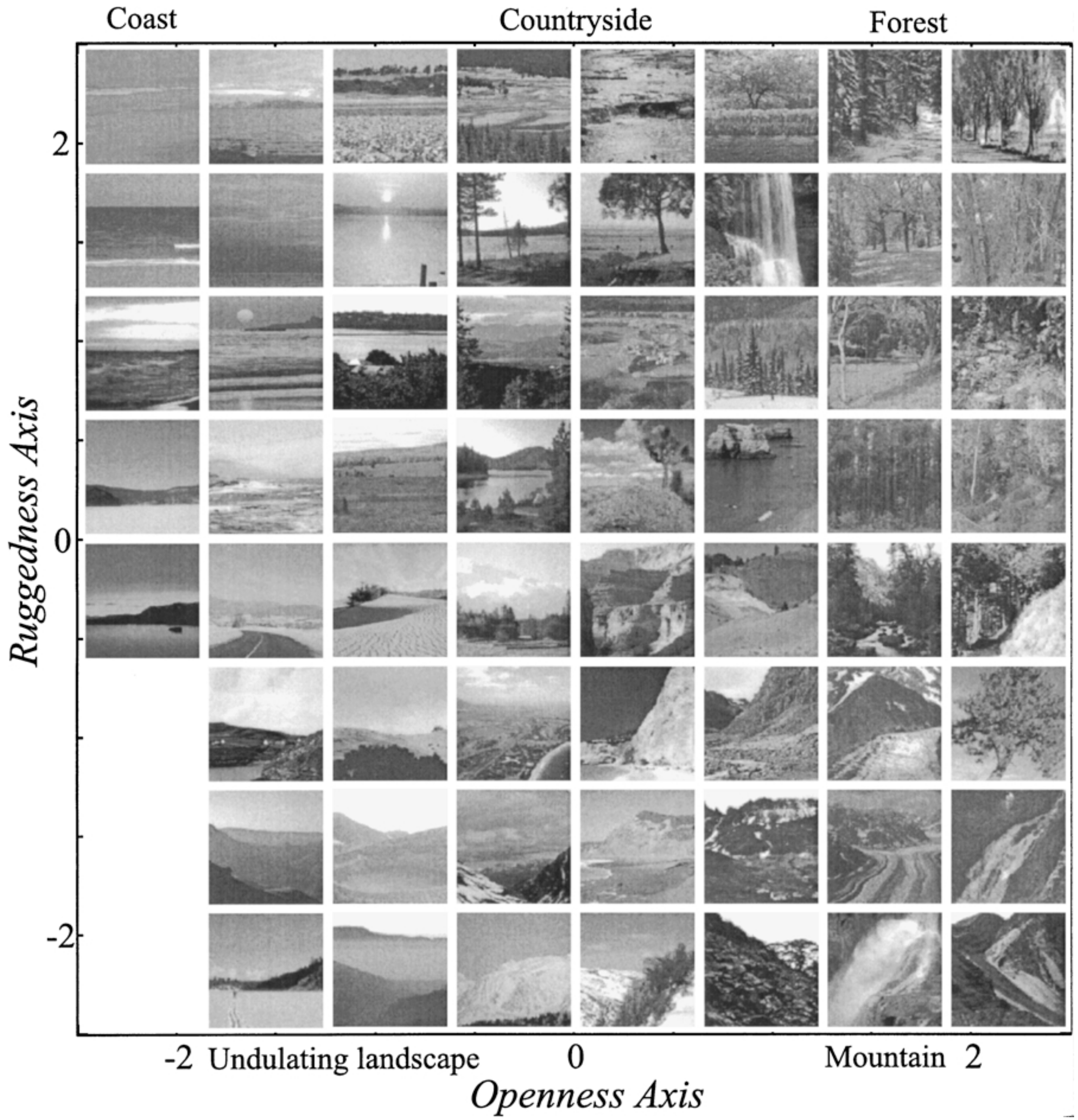
Man-made expansion

...

Categories

- Have *spectral energy model* for *spatial envelope features*
- Now need mapping of *spatial envelope features* to *categories*

Categories



Shows set of images projected into 2D space corresponding to openness and ruggedness

Scenes close in the space have similar category membership

Categories

- Projected typical exemplars of categories (coasts, mountains, tall buildings etc) into spatial envelope space to make database
- classification performed by K nearest neighbors classifier:
    - given new scene picture K-NN looks for K nearest neighbors of image within the labeled training dataset
    - these correspond to images with closest spatial envelope properties
    - category comes from most represented category of k images

Accuracy

Table 4. Confusion matrix (in percent) between typical scenes of coasts, countryside (fields, valleys, hills, rolling countryside), enclosed forests and mountains ($N = 1500$).

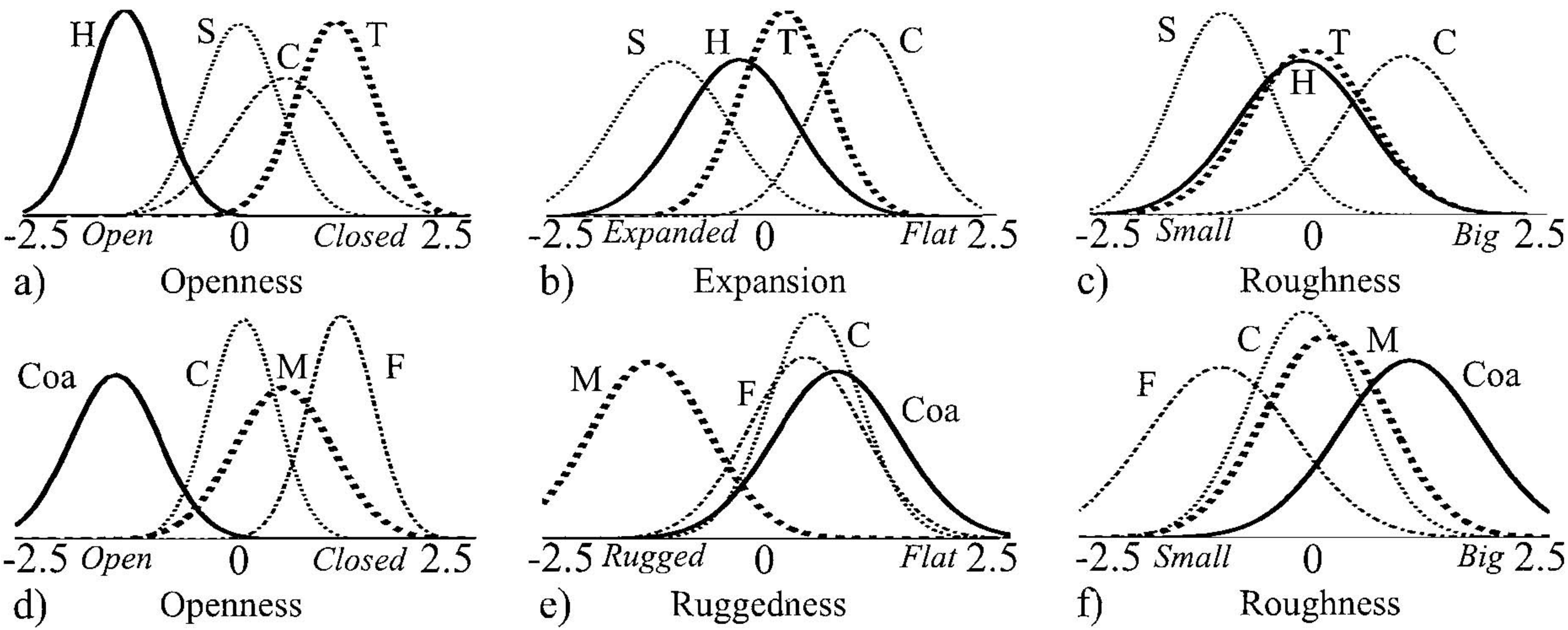
	Coast	Country	Forest	Mountain
Coast	88.6	8.9	1.2	1.3
Country	9.8	85.2	3.7	1.3
Forest	0.4	3.6	91.5	4.5
Mountain	0.4	4.6	3.8	91.2

Table 5. Confusion matrix (in percent) for the classification between highways, city center streets, city center close views, and tall buildings/skyscrapers ($N = 1400$ images).

	Highway	Street	Close-up	Tall building
Highway	91.6	4.8	2.7	0.9
Street	4.7	89.6	1.8	3.9
Close-up	2.5	2.3	87.8	7.4
Tall building	0.1	3.4	8.5	88

Classification is on average 89% with WDST (86% with DST)

Accuracy



H - Highway
S - Street
C - Coast
T - Tall buildings

different categories lie on different locations of the spatial envelope axes

Summary

- find semantically meaningful spatial envelope properties
- show spatial properties strongly correlated with second order statistics DST and spatial arrangement of structures WDST
- spatial properties can be used to infer scene category

Summary

- find semantically meaningful spatial envelope properties
- show spatial properties strongly correlated with second order statistics DST and spatial arrangement of structures WDST
- spatial properties can be used to infer scene category

A Bayesian Heirarchical Model for Learning Natural Scene Categories

Li Fei Fei and Pietro Perona 2005

Overview

- Goal: Recognize natural scene categories
- Insight: use intermediate representation before classifying scenes
  - labeled wrt global or local properties
  - Oliva and Torralba - spatial envelope properties hand labeled by human observers
- Problem with human labeling: hours of manual labor and suboptimal labeling
- Contribution: unsupervised learning of themes

Overview

- Inspiration: work on Texture models
  - first learn dictionary of textons
  - each category of texture captures a specific distribution of textons
  - intermediate themes ~ texture descriptions
- Approach: local regions clustered into themes, then into categories. Probability distribution learnt automatically, bypassing human annotation

Baysian Model

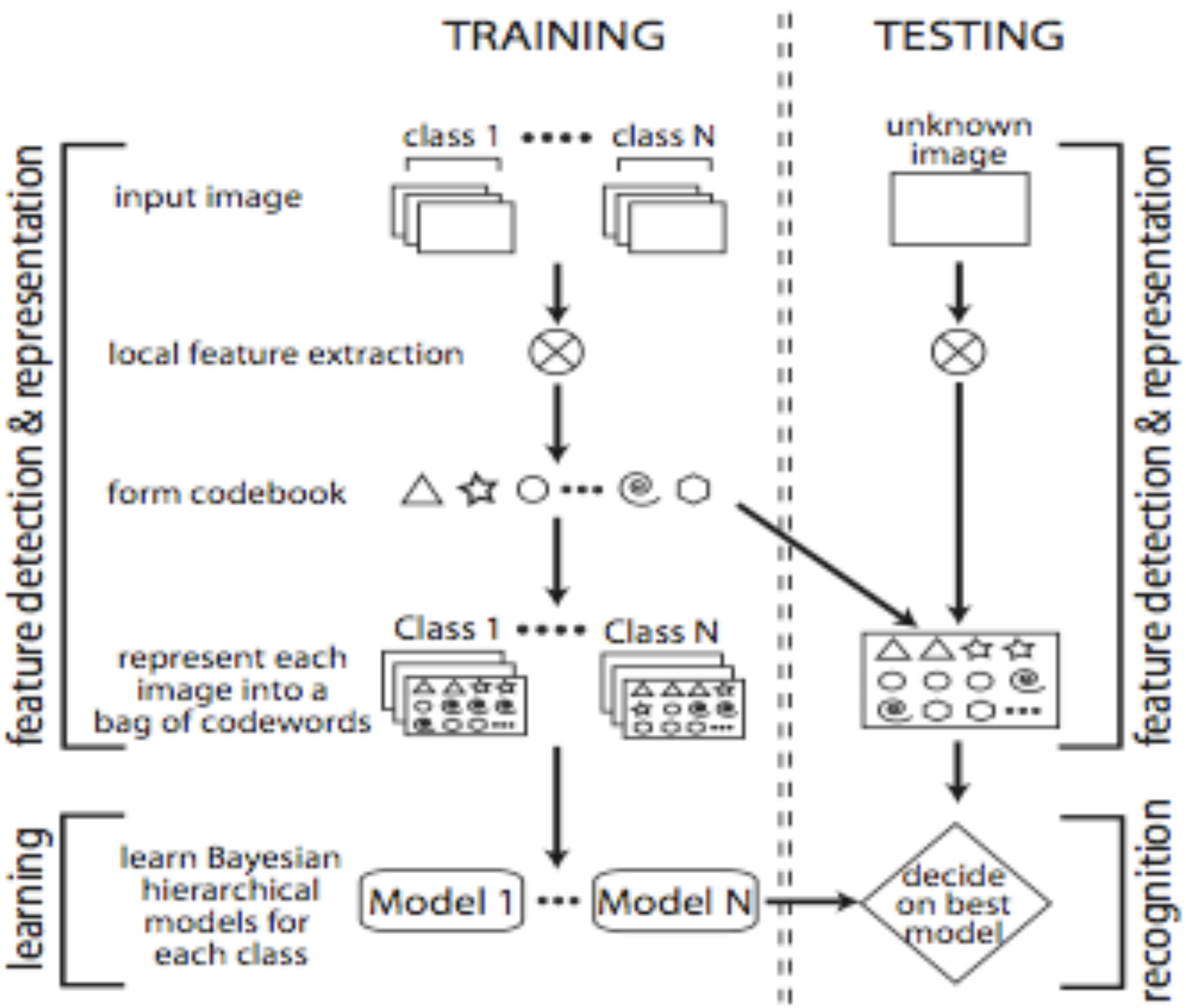


Figure 2. Flow chart of the algorithm.

Learn Baysian Model - requires learning joint probability of unknown variables

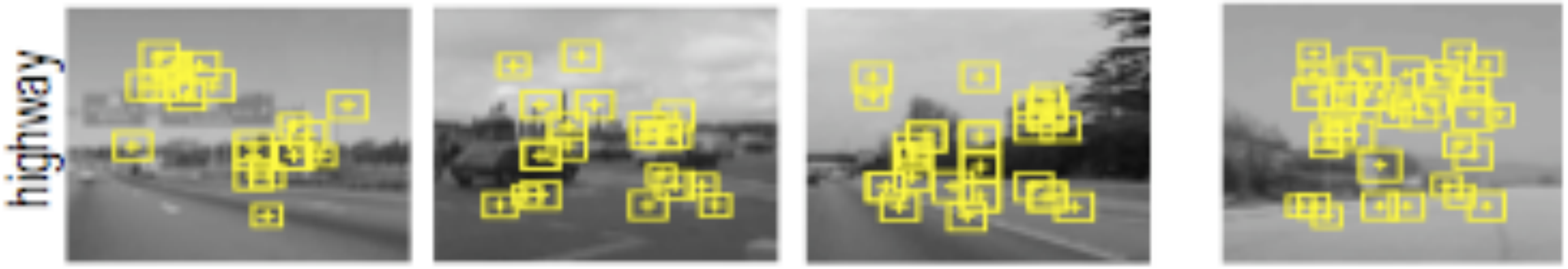
for new image, compute probability of each category given learned parameters

label is the category that gives the largest likelihood of the image

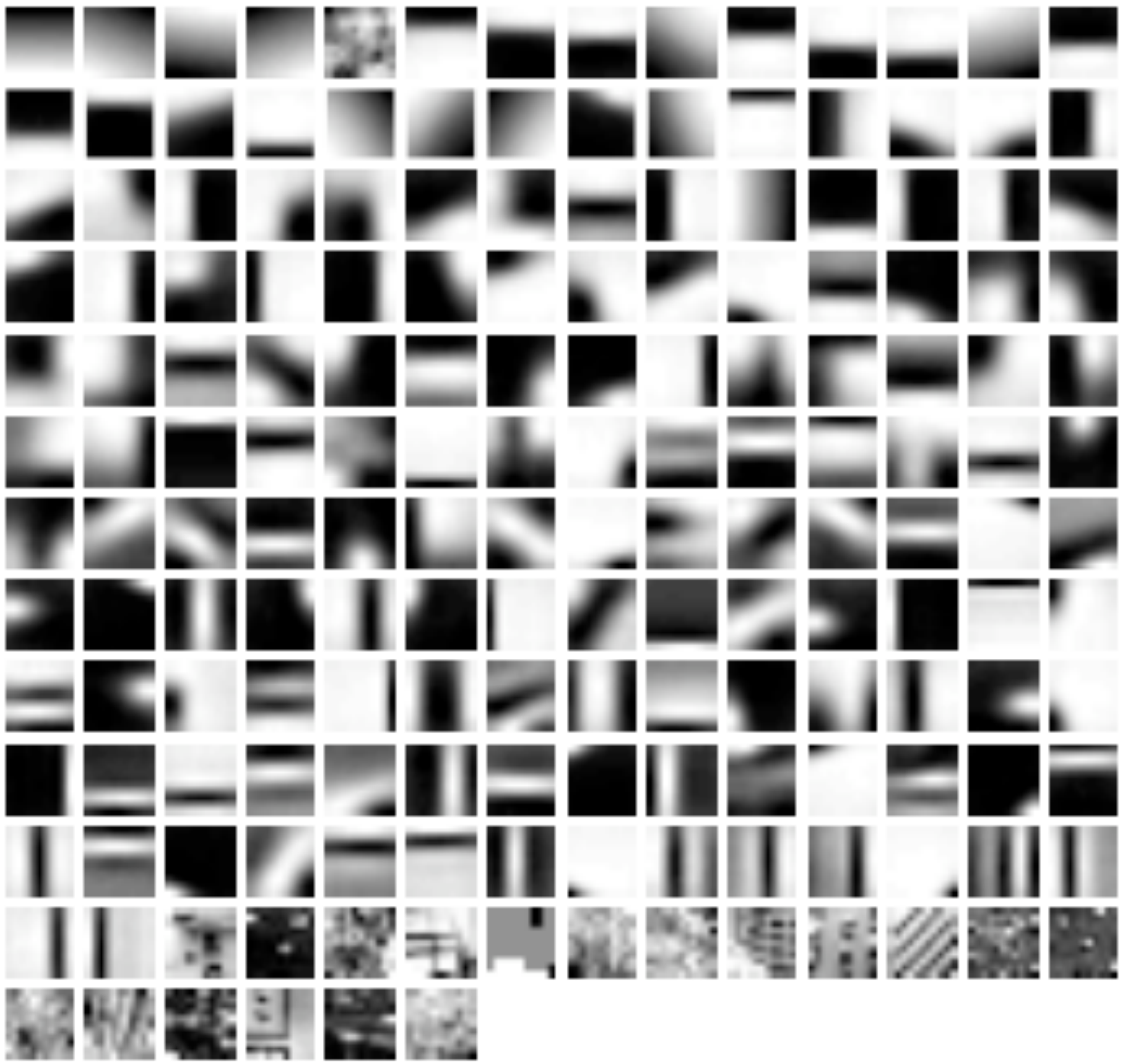
lots more math in the paper

Features

- previous model used global features (frequencies, edges, color histograms)
- They use LOCAL REGIONS
- Tried 4 ways of extracting patches
- Evenly sampled dense grid spaced 10x10 randomly sized patch between 10-30pxls



Codebook



Codewords obtained from 650 training examples

learn codebook through k-means clustering. codewords are center of cluster

best results when using 174 codewords

Shown in descending order according to size of membership.

correspond to simple orientations, illumination patterns similar to ones that early human visual system responds to.

Testing

- Oliva and Torralba dataset with 5 new categories = 13 category dataset
- Model trained on 100 images of each category (10 mins to train all 13)
- New image labeled with category that gives highest likelihood probability

Results

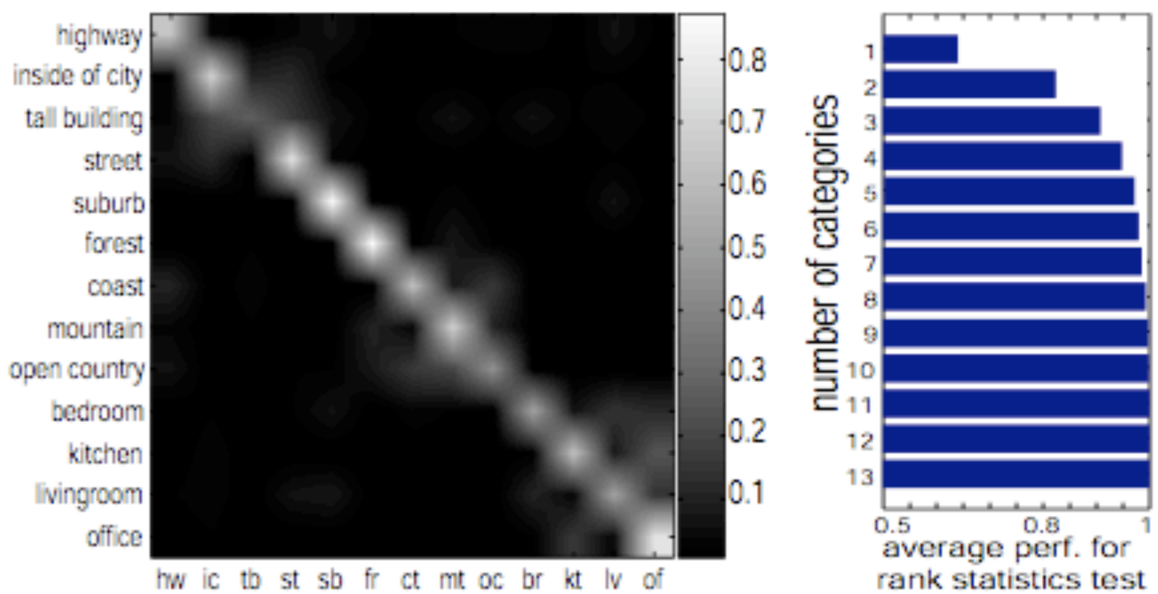


Figure 7. Left Panel. Confusion table of Theme Model 1 using 100 training and 50 test examples from each category, the grid detector and patch based representation. The average performance is 64.0%. **Right Panel.** Rank statistics of the confusion table, which shows the probability of a test scene correctly belong to one of the top N most probable categories. N ranges from 1 to 13.

Perfect confusion table would be straight diagonal

Chance would be 7.7% recognition

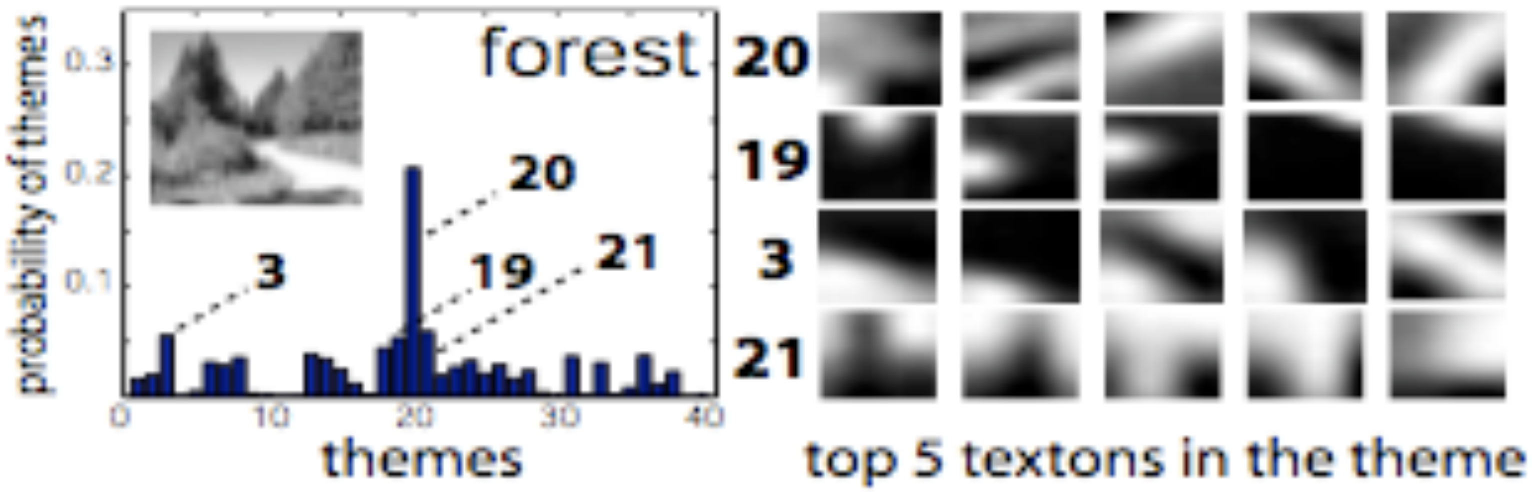
Results average 64% recognition

Recognition in top two choices 82%

Highest block of errors on indoor scenes

Results

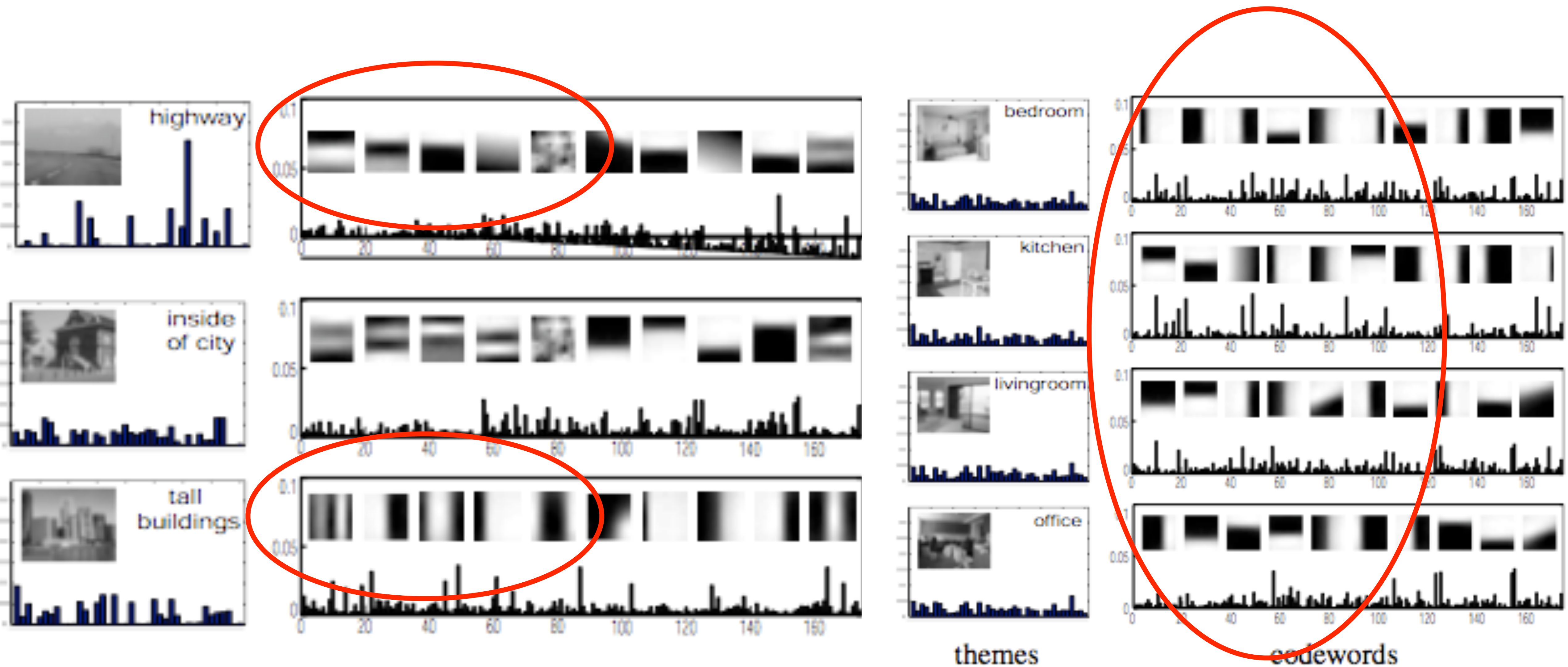
A look at the internal structure



Shows themes that are learned and corresponding codewords

Some themes have semantic meaning:
foliage (20, 3) and
branch (19)

Results



Indoor scenes

Summary

- Automatically learn intermediate codewords and themes using Baysian Model with no human annotation
- Obtain 64% accuracy of categorization on 15 category database, 74% accuracy on 4 categories

Big Picture so far

	Oliva and Torralba [2001]	**FeiFei and Perona [2005]**
# of categories	8	13
# of intermediate themes	6 Spatial Envelope Properties	40 Themes
training # per category	250-300	100
training requirements	human annotation of 6 properties for thousands images	unsupervised
performance	89%	76%
kind of features	global statistics (energy spectra & spectrogram)	Local patches

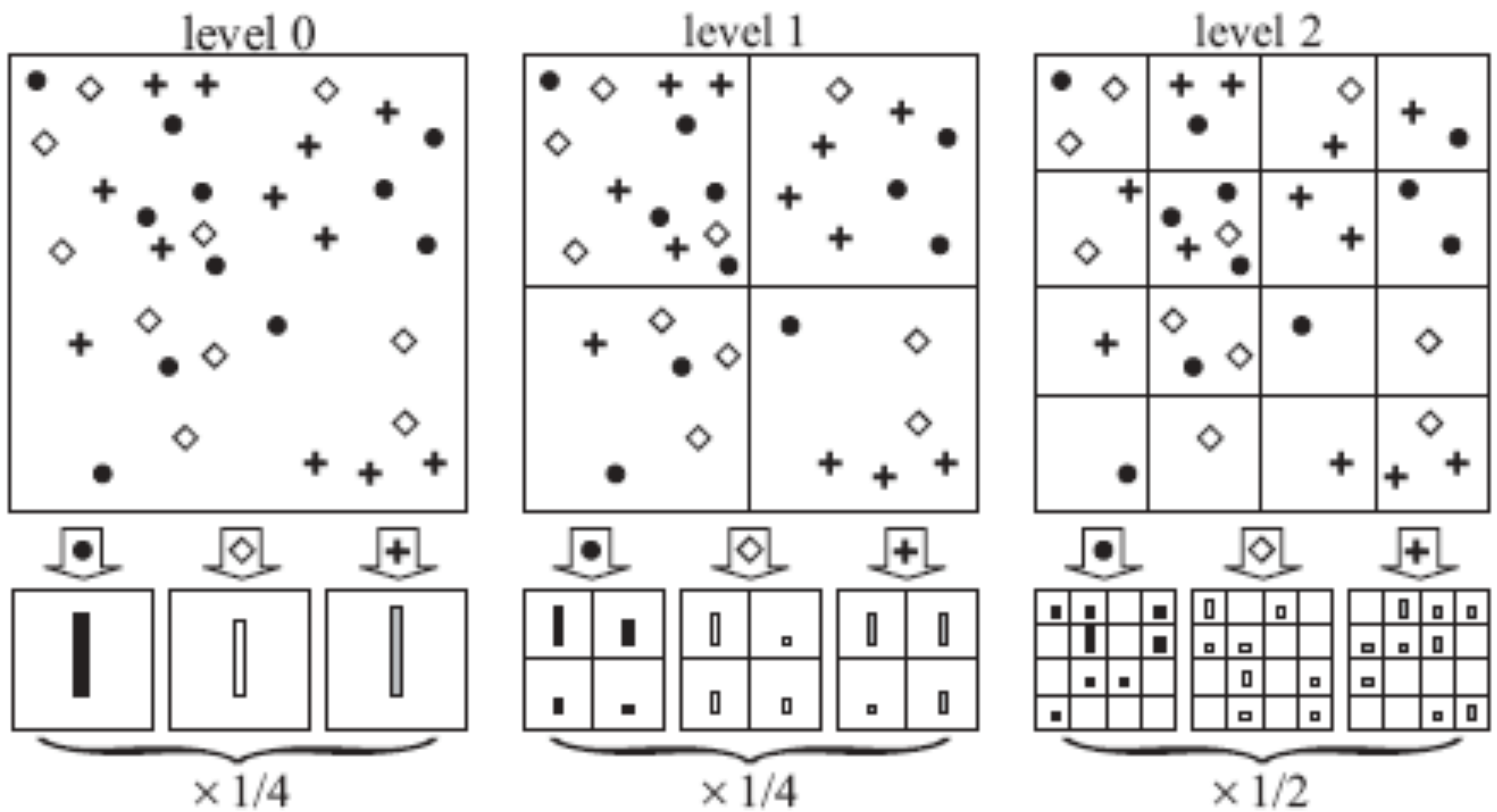
Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories

Labzebnik, Schmid, Ponce 2006

Overview

- Goal:
  Recognize photographs as a scene (forest, ocean) or as containing an object (bike, person)
- Previous methods:
  - Bag of features (disregard spatial information)
  - Generative part models and geometric correspondence (computational expensive)
- Novel Approach:
  - repeatedly subdivide image
  - compute histograms of local features over subregions
  - Adapted from Pyramid Matching [Grauman and Darrell]

Spatial Pyramid Matching



Constructing a 3-level pyramid.

- Subdivide image at three levels of resolution.
- For each level and each feature channel, count # features in each bin.
- The spatial histogram is a weighted sum of these values.
- Weight of match at each level is inversely proportional to size of bin
    penalize matches in larger cells
    highly weight matches in smaller cells

Features

- “weak” features
  - oriented edge points at 2 scales 8 orientations.
  - similar to gist
- “strong” features
  - SIFT descriptors of 16x16 patches over dense grid
  - cluster patches to form M=200 or M=400 large visual vocabulary

Testing

- 15 Category dataset - Scenes
  [Oliva &Torralba and FeiFei and Perona]
- Caltech 101 - objects
- Graz - objects

Results on Scenes

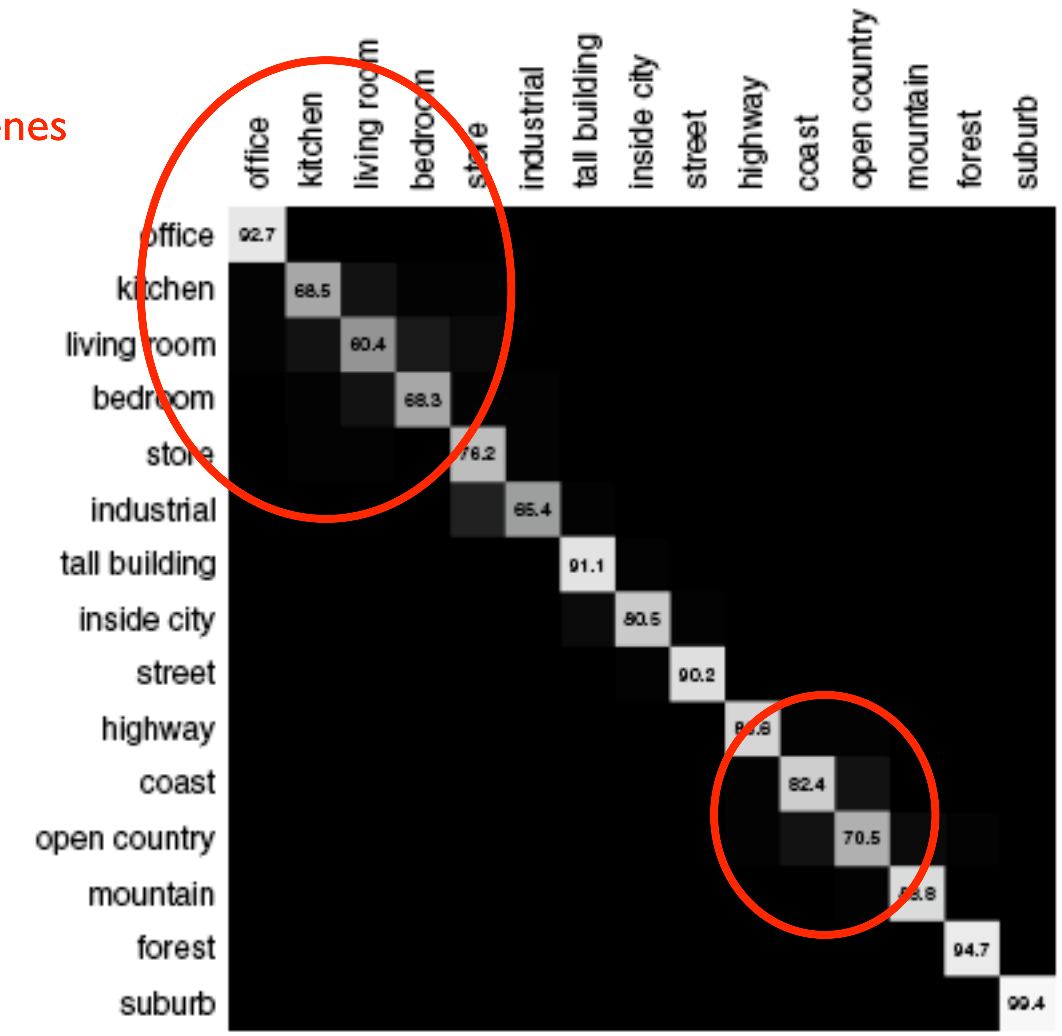
	Weak features ($M = 16$)		Strong features ($M = 200$)		Strong features ($M = 400$)	
L	Single-level	Pyramid	Single-level	Pyramid	Single-level	Pyramid
$0\ (1 \times 1)$	45.3 ± 0.5		72.2 ± 0.6		74.8 ± 0.3	
$1\ (2 \times 2)$	53.6 ± 0.3	56.2 ± 0.6	77.9 ± 0.6	79.0 ± 0.5	78.8 ± 0.4	80.1 ± 0.5
$2\ (4 \times 4)$	61.7 ± 0.6	64.7 ± 0.7	79.4 ± 0.3	**81.1** ± 0.3	79.7 ± 0.5	**81.4** ± 0.5
$3\ (8 \times 8)$	63.3 ± 0.8	**66.8** ± 0.6	77.2 ± 0.4	80.7 ± 0.3	77.2 ± 0.5	81.1 ± 0.6

Table 1. Classification results for the scene category database (see text). The highest results for each kind of feature are shown in bold.

- What does chart show?
- Multilevel pyramid setup better than single level
- For strong features, single level performance goes down from L=2 to L=3. Pyramid too finely subdivided. Even so, pyramid scheme stays same.
- Advantage: Pyramid combines multiple resolutions in principled fashion -- robust to failures at individual levels
- Strong features better than weak. But M=200 similar to M=400. Pyramid scheme more important than large vocabulary.

Results on Scenes

indoor scenes



coast and open country

Figure 3. Confusion table for the scene category dataset. Average classification rates for individual classes are listed along the diagonal. The entry in the ith row and jth column is the percentage of images from class i that were misidentified as class j.

Results on Scenes

Retrieval from the scene category database

Spatial pyramid scheme successful at finding
major elements, "blobs", directionality of lines

Also preserves high frequency detail (see kitchen)

Results on Caltech 101

Will this method work on OBJECTS?

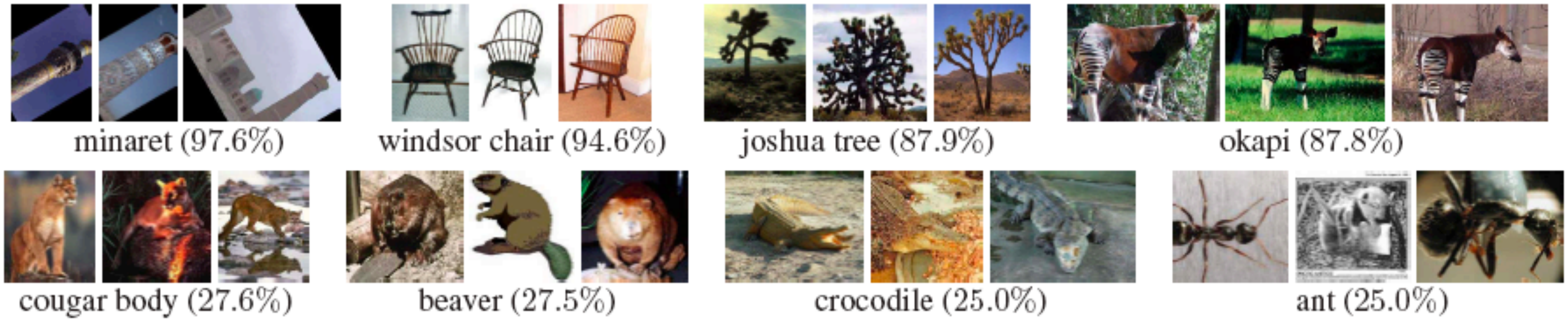


Figure 5. Caltech-101 results. Top: some classes on which our method ($L = 2, M = 200$) achieved high performance. Bottom: some classes on which our method performed poorly.

	Weak features		Strong features (200)	
L	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

Table 2. Classification results for the Caltech-101 database.

This outperforms orderless methods and geometric correspondence methods

Results on Graz

Will this method work on OBJECTS with lots of clutter?

bike person background

Figure 6. The Graz database.

Has images of bikes, persons, and backgrounds.

Images vary greatly within one category

Heavy clutter and pose changes

Class	$L = 0$	$L = 2$	Opelt [14]	Zhang [25]
Bikes	82.4 ±2.0	86.3 ±2.5	86.5	92.0
People	79.5 ±2.3	82.3 ±3.1	80.8	88.0

Table 4. Results of our method ($M = 200$) for the Graz database and comparison with two existing methods.

Summary

- Approach: repeatedly subdivide image and computing histograms of image features over subregions.
- Shown good results on 3 datasets
- simple global construction

Big Picture

	Oliva and Torralba [2001]	FeiFei and Perona [2005]	Labzebnik et al.[2006]
# of categories	8	13	15
# of intermediate themes	6 Spatial Envelope Properties	40 Themes	M=200 strong feature clusters
training # per category	250-300	100	NA?
training requirements	human annotation of 6 properties for thousands images	unsupervised	unsupervised?
performance	89%	76%	81% (on all 15 cat.)
kind of features	global statistics (energy spectra & spectrogram)	Local patches	"weak" oriented filters "strong" SIFT features
what is novel	can use global features for recognition	human annotation not needed	spatial pyramid scheme robust to different resolutions * Add object detection

Conclusion

- Results underscore surprising power of global statistics for scene categorization and even object recognition
- Can be used as “context modules” within larger object recognition systems