Scene Recognition

Adriana Kovashka
UTCS, PhD student

Problem

- Statement
  - Distinguish between different types of scenes
- Applications
  - Matching human perception
  - Understanding the environment
    - Indexing of images / video
    - Robotics
  - Graphics
    - In-painting

Background

- Definition of a scene
  - "[A] scene is mainly characterized as a place in which we can move" [Oliva 2001]
- Assumptions
  - Human categorization
- Approaches
  - Parsing of the scene – as a whole, or in parts

Coast [Oliva 2001]




Mountain [Oliva 2001]




Inside City [Oliva 2001]

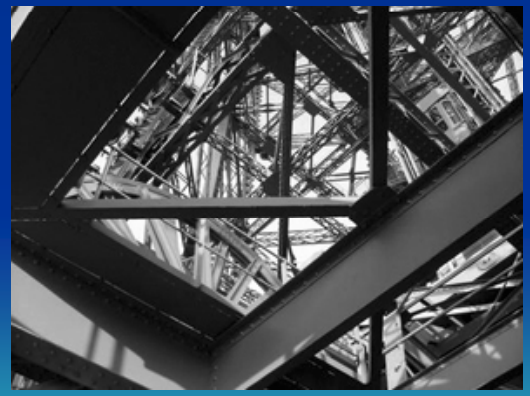


Street [Oliva 2001]



Kitchen [Lazebnik 2006]

Industrial [Lazebnik 2006]



Scene?

Scene?

Urban or natural?

Urban or natural?

Spatial Envelope [Oliva 2001]

- Inspiration from human perception
  - Naturalness, openness, roughness
  - Expansion, ruggedness
- Holistic, no recognition of objects
- Three levels
  - “cars and people” vs. “street” vs. “urban environment”

Spatial Envelope (cont'd)

- Scene modeling
  - Discrete Fourier Transform

$$I(f_x, f_y) = \sum_{x,y=0}^{N-1} i(x, y)h(x, y)e^{-j\,2\pi(f_x x + f_y y)}$$
$$= A(f_x, f_y)\, e^{j\Phi(f_x, f_y)}$$

[Oliva 2001]

  - Windowed Fourier Transform

$$I(x, y, f_x, f_y)$$
$$= \sum_{x',y'=0}^{N-1} i(x', y')\, h_r(x' - x, y' - y)e^{-j\,2\pi(f_x x' + f_y y')}$$

[Oliva 2001]

Spatial Envelope (cont'd)

– Principal Components Analysis

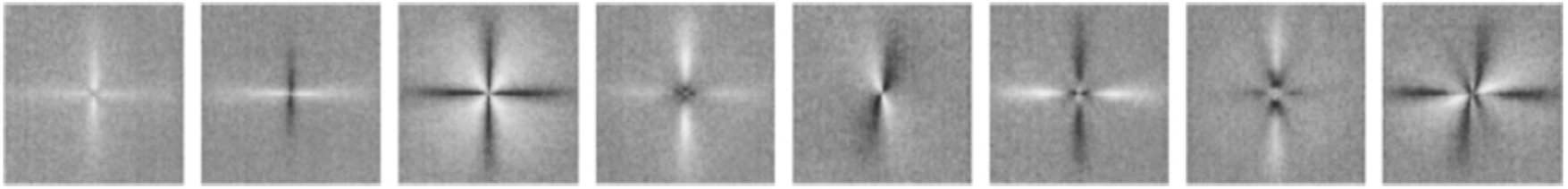


Figure 2. The first eight principal components for energy spectra of real-world scenes. The frequency $f_x = f_y = 0$ is located at the center of each image.

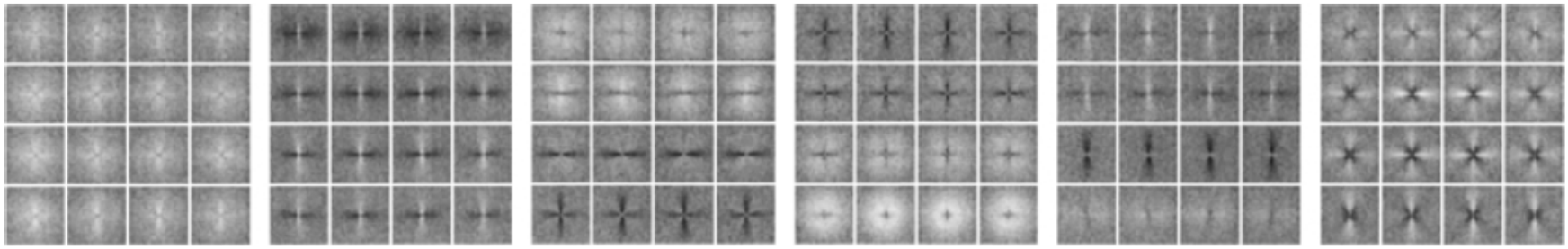


Figure 3. The first six principal components of the spectrogram of real-world scenes. The spectrogram is sampled at 4×4 spatial location for a better visualization. Each subimage corresponds to the local energy spectrum at the corresponding spatial location.

[Oliva 2001]

Spatial Envelope (cont'd)

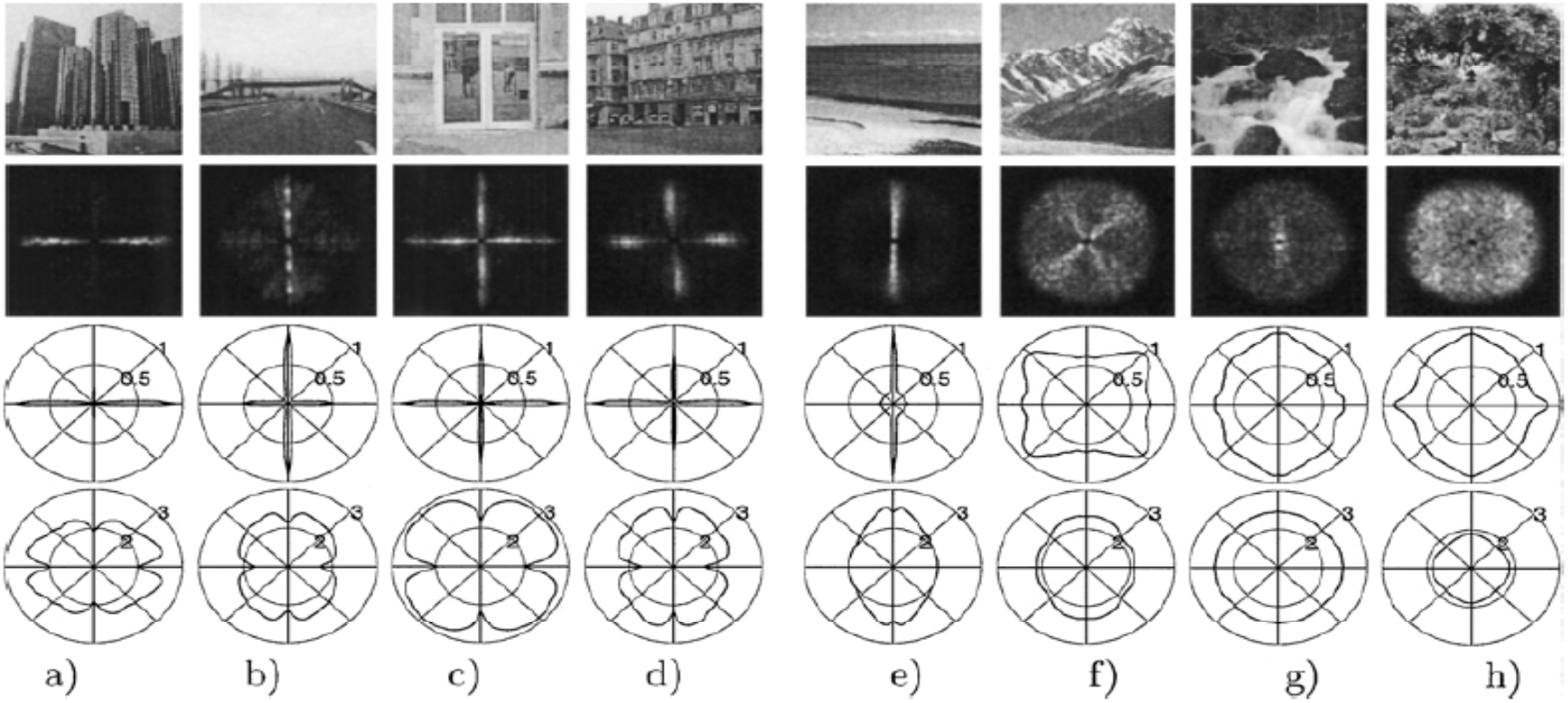


Figure 5. Examples of scenes from different categories, their respective energy spectrum (energy spectra have been multiplied by f^2 in order to enhance the visibility of high spatial frequencies) and the spectral signatures of their category: function $\Gamma_s(\theta)$ and the bottom line shows the function $\alpha_s(\theta)$ in a polar diagram. From a) to h), scenes illustrate the categories: tall building, highway, urban close-up views, city center, coast, mountain, natural close up views and forests.

[Oliva 2001]

Spatial Envelope (cont'd)

- Properties of the spatial envelope
  - Discriminant spectral template (DST)
    - Relates spectral components to properties of the spatial envelope
    - Parameter *d* learned through matching of feature vectors and property values
  - Windowed discriminant spectral template (WDST)

Spatial Envelope (cont'd)

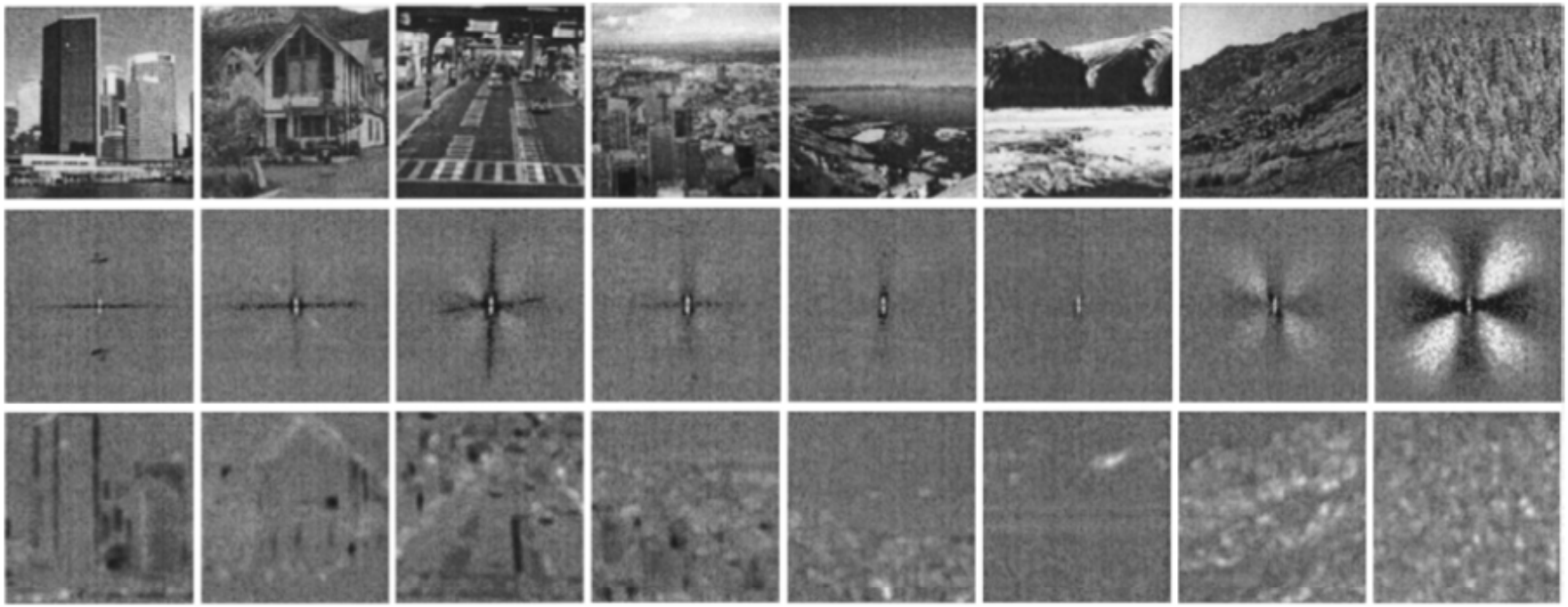


Figure 9. From top to bottom: Samples of images selected at random ordered along the naturalness axis, from man-made environments (left) to natural landscapes (right); their energy spectra multiplied by the DST; the opponent energy image (we have suppressed the effect of the Hanning window for clarity). Natural and man-made components are respectively represented by white and black edges.

[Oliva 2001]

Spatial Envelope (cont'd)

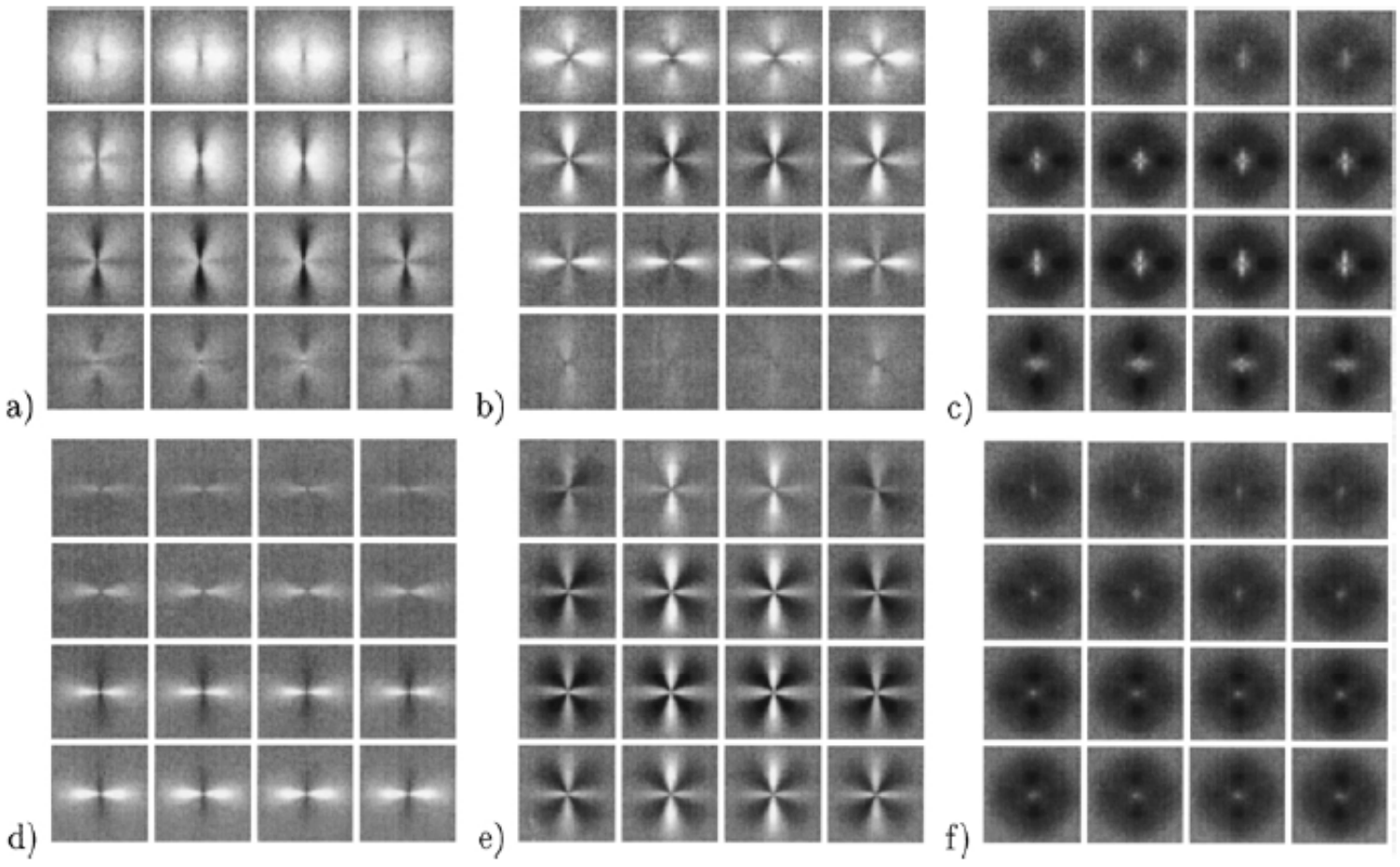


Figure 11. Discriminant spectral templates $WDST(x, y, f_x, f_y)$ with $N_L = 30$. For natural scenes: a) openness, b) ruggedness and c) roughness. For man-made scenes: d) openness, e) expansion and f) roughness.

[Oliva 2001]

Spatial Envelope (cont'd)

- Results
  - Scene properties

Table 2. Correlation between orderings of natural scenes made by observers and the two templates for each spatial envelope property.

	Openness	Ruggedness	Roughness
DST	$m = 0.82$	0.73	0.82
WDST	$m = 0.88$	0.79	0.86
Agreement	0.92	0.82	0.87

Agreement measures the concordance between subjects.

[Oliva 2001]

Spatial Envelope (cont'd)

[Oliva 2001]

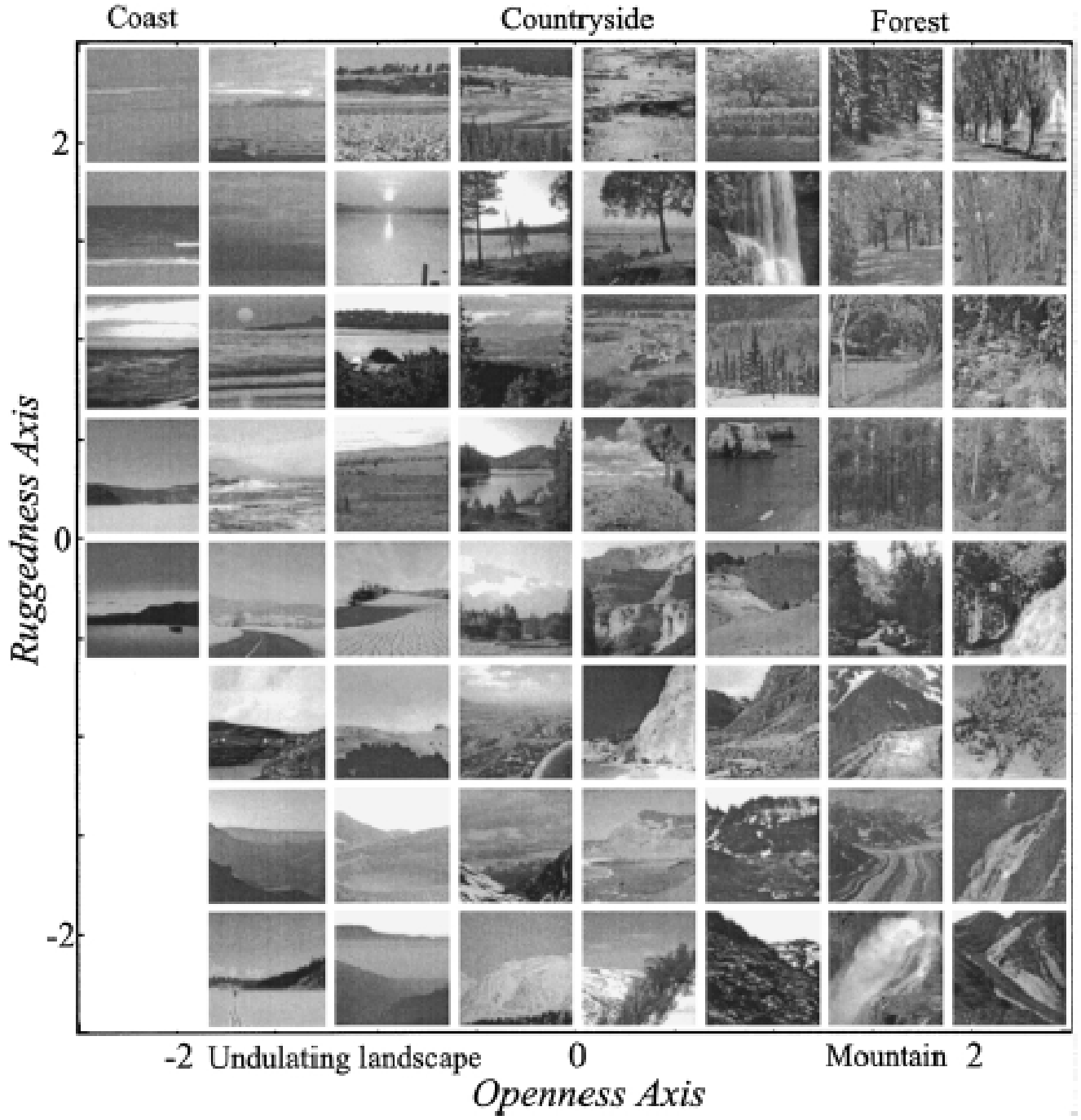


Figure 15. Organization of natural scenes according to the openness and ruggedness properties estimated by the WDSTs.

Spatial Envelope (cont'd)

[Oliva 2001]

Figure 17. Examples of man-made scenes (target) with four neighbors sharing similar spatial envelope, estimated with the DST and the WDST procedures. The bottom example is an error.

Spatial Envelope (cont'd)

- Classification
  - K-nn
  - 4 out of 7 neighbors picked by humans

[Oliva 2001]

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

Spatial Envelope (cont'd)

- Strengths
  - Higher-level descriptions
  - Low dimensionality
  - Invariance to object composition
  - Weak local information
- Weaknesses
  - Significant number of human labels


[Oliva 2001]

Spatial Pyramid [Lazebnik 2006]

- Global, locally orderless
- Bag-of-features
- Extension of Pyramid Match Kernel in 2-d
- Regular clustering of features

Spatial Pyramid (cont'd)

$$\kappa^L(X,Y) = \mathcal{I}^L + \sum_{\ell=0}^{L-1} \frac{1}{2^{L-\ell}} \left(\mathcal{I}^\ell - \mathcal{I}^{\ell+1}\right)$$

$$= \frac{1}{2^L}\mathcal{I}^0 + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}} \mathcal{I}^\ell \,.$$

[Grauman 2005] as quoted in [Lazebnik 2006]

$$K^L(X,Y) = \sum_{m=1}^{M} \kappa^L(X_m, Y_m)$$

[Lazebnik 2006]

Spatial Pyramid (cont'd)

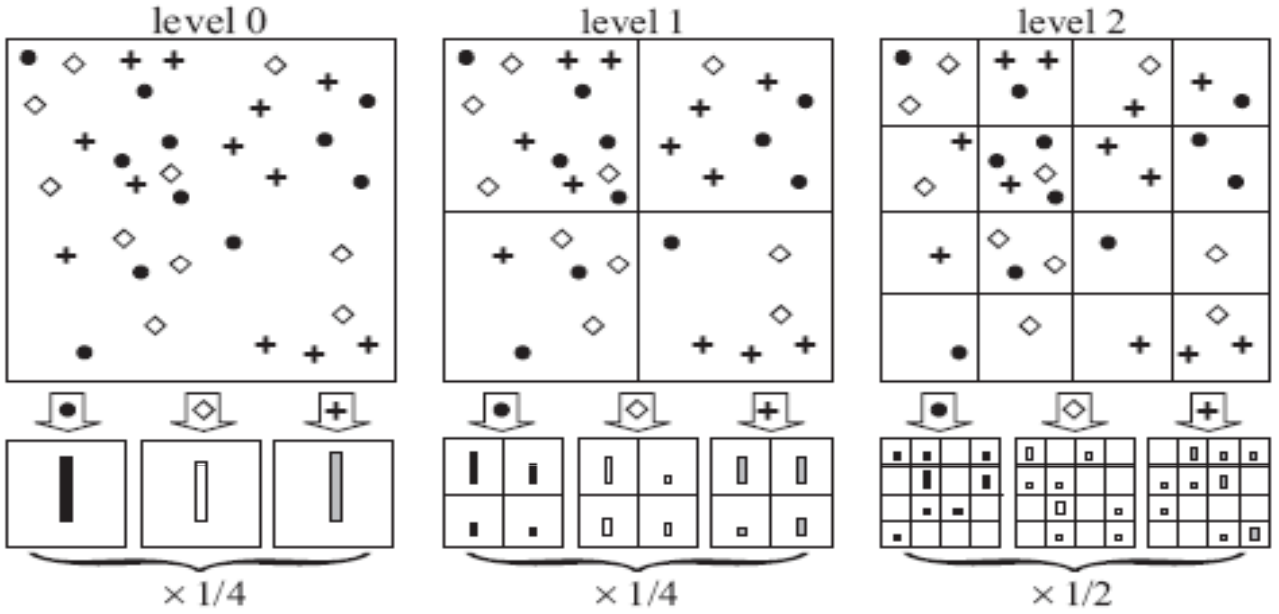
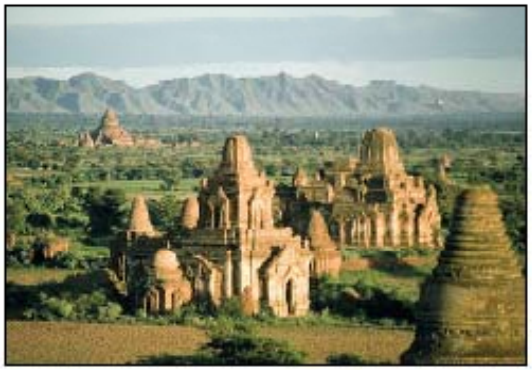


Figure 1. Toy example of constructing a three-level pyramid. The image has three feature types, indicated by circles, diamonds, and crosses. At the top, we subdivide the image at three different levels of resolution. Next, for each level of resolution and each channel, we count the features that fall in each spatial bin. Finally, we weight each spatial histogram according to eq. (3).

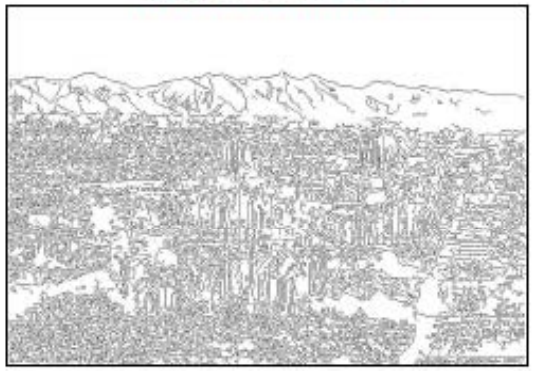
[Lazebnik 2006]

Spatial Pyramid (cont'd)

Feature extraction



Weak features



Edge points at 2 scales and 8 orientations (vocabulary size 16)

Strong features

SIFT descriptors of 16x16 patches sampled on a regular grid, quantized to form visual vocabulary (size 200, 400)

[Lazebnik 2006]

Spatial Pyramid (cont'd)

- Results
  - SVM classification
  - Scene recognition

	Weak features ($M = 16$)		Strong features ($M = 200$)		Strong features ($M = 400$)	
L	Single-level	Pyramid	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6		74.8 ±0.3	
1 (2×2)	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 (4×4)	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 (8×8)	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.

[Lazebnik 2006]

65.2% for [Fei-Fei 2005]

Spatial Pyramid (cont'd)

[Lazebnik 2006]

Figure 4. Retrieval from the scene category database. The query images are on the left, and the eight images giving the highest values of the spatial pyramid kernel (for $L = 2, M = 200$) are on the right. The actual class of incorrectly retrieved images is listed below them.

Spatial Pyramid (cont'd)

– Object recognition

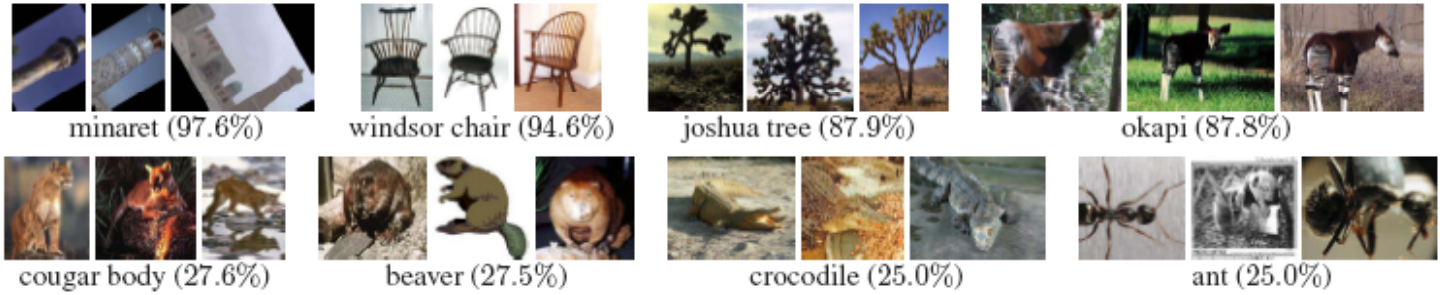


Figure 5. Caltech-101 results. Top: some classes on which our method ($L = 2, M = 200$) achieved high performance. Bottom: some classes on which our method performed poorly.

	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.

[Lazebnik 2006]

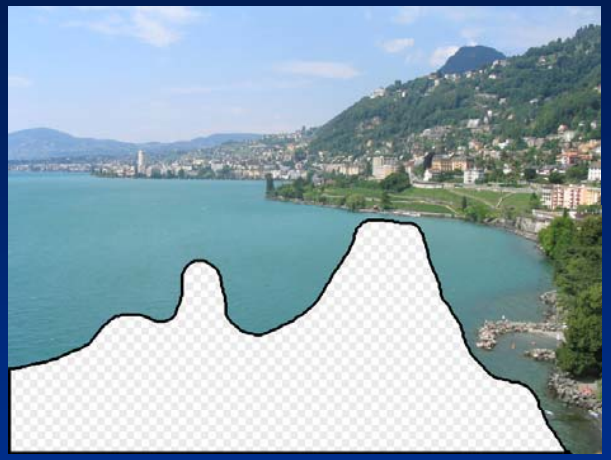
Spatial Pyramid (cont'd)

- Strengths
  - Reasonable dimensionality
  - "Locally orderless"
  - Dense representation
  - "Robust to failures at individual levels"
- Weaknesses
  - No invariability to composition of image
  - Not robust to clutter

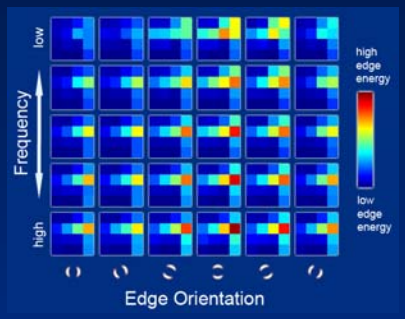
[Hays 2007]

Scene Completion

http://graphics.cs.cmu.edu/projects/scene-completion/



Input image



Scene Descriptor

Image Collection

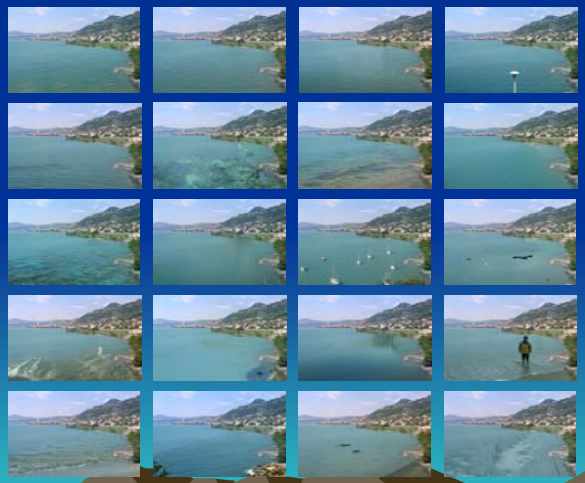
...

200 matches

Hays and Efros, SIGGRAPH 2007

...

Context matching + blending



20 completions


[Oliva 2001]

Topic Models [Fei-Fei 2005]

- Bayesian hierarchical model
- Intermediate representations
- Bag-of-features
  - 4 ways to extract regions
  - 2 types of features

Topic Models (cont'd)

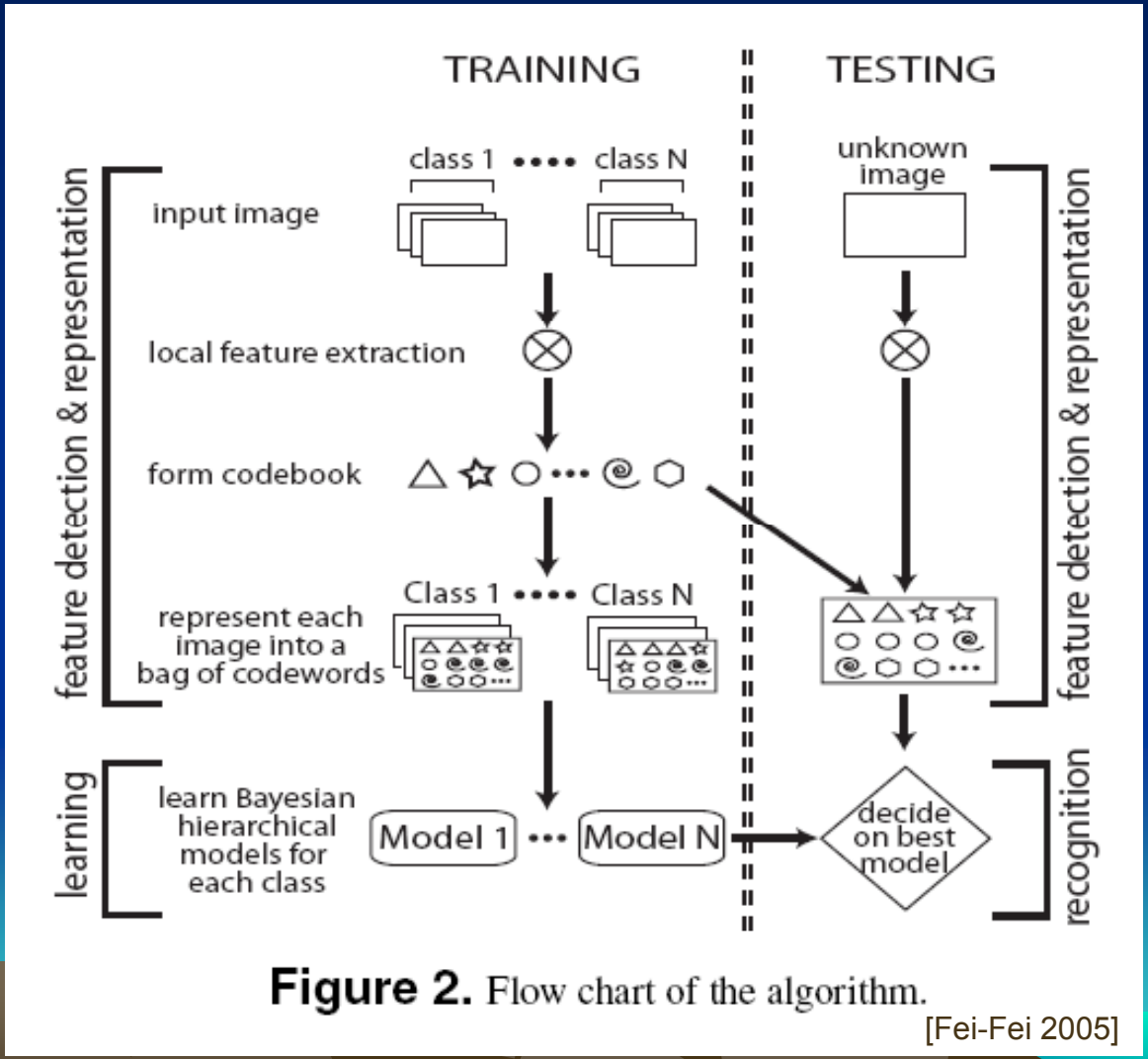
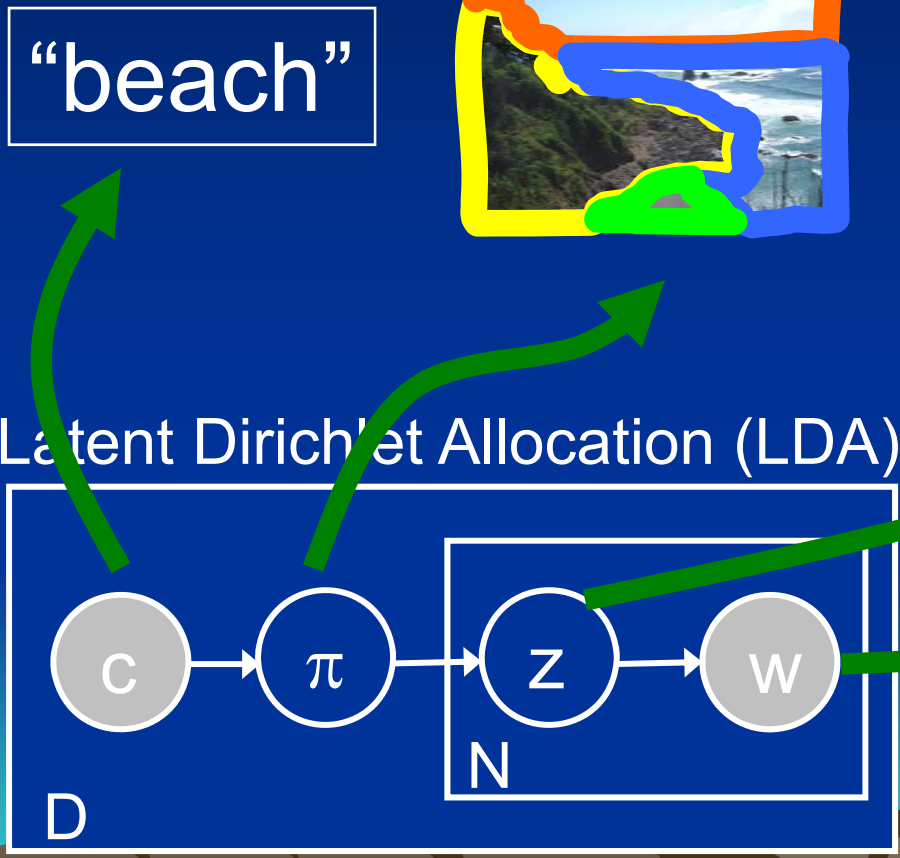


Figure 2. Flow chart of the algorithm.

[Fei-Fei 2005]

Hierarchical Bayesian text models

[Fei-Fei 2005]



Fei-Fei et al. ICCV 2005

Topic Models (cont'd)

- η – distribution of class labels
- θ – parameter (estimated by EM)
- c – class label
- π – distribution of themes for image
- z – theme
- x – patch
- β – parameter (estimated by EM)

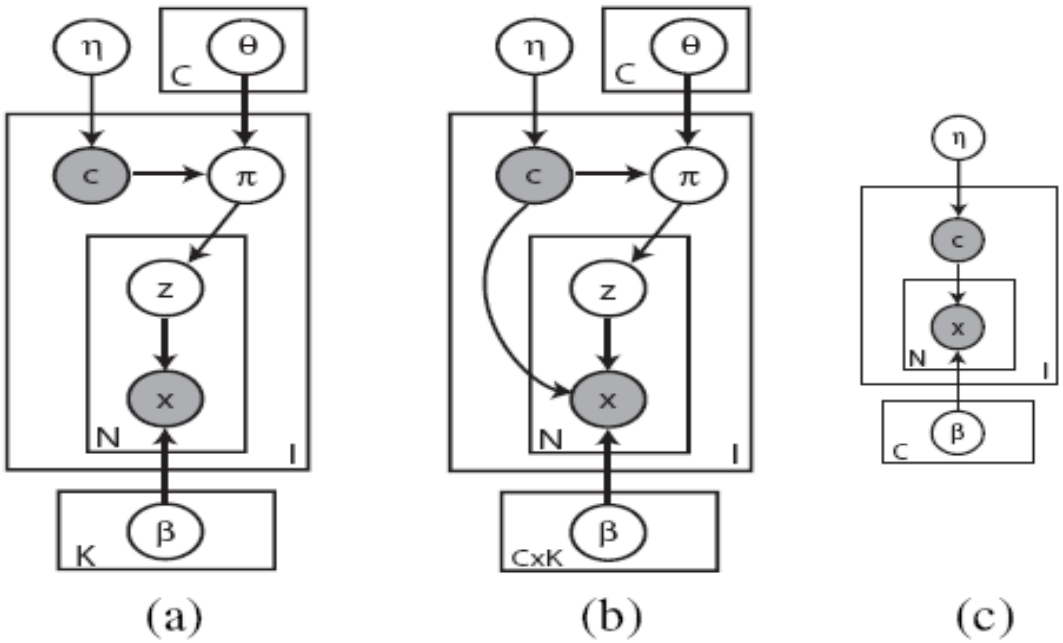


Figure 3. (a) Theme Model 1 for scene categorization that shares both the intermediate level themes as well as feature level codewords. (b) Theme Model 2 for scene categorization that shares only the feature level codewords; (c) Traditional texton model [5, 16].

[Fei-Fei 2005]

Topic Models (cont'd)

- Codebook
  - 174 codewords

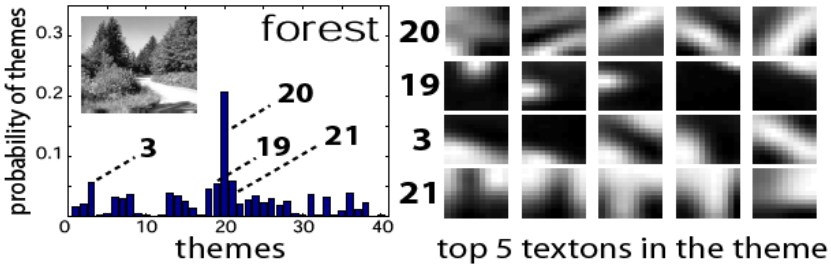


Figure 8. Example of themes for the forest category. **Left Panel** The distribution of all 40 themes. **Right Panel** The 5 most likely codewords for each of the 4 dominant themes in the category. Notice that codewords within a theme are visibly consistent. The "foliage" (#20, 3) and "tree branch" (#19) themes appear to emerge automatically from the data.

[Fei-Fei 2005]

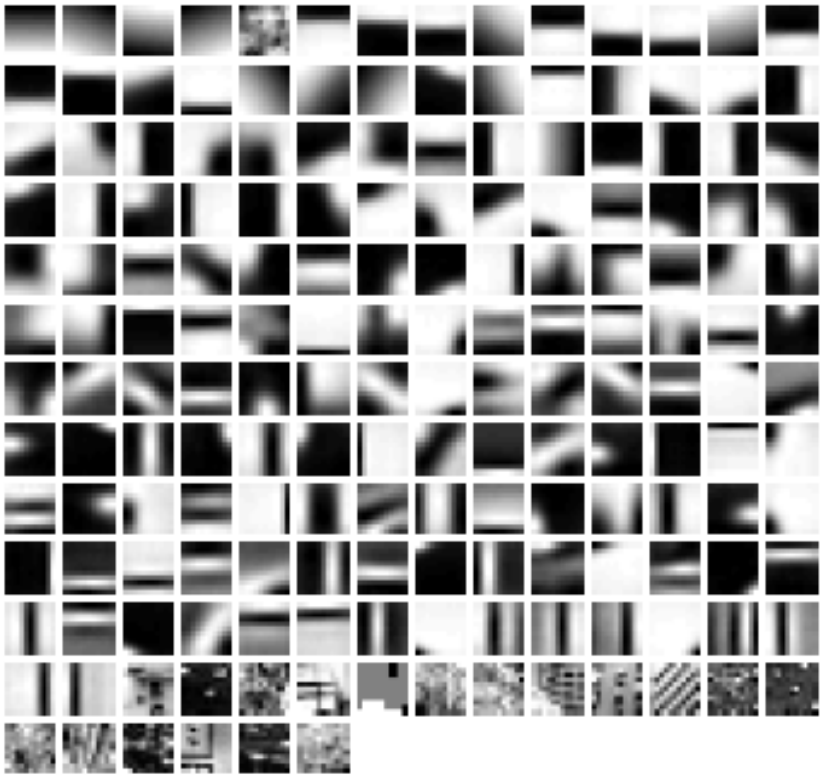
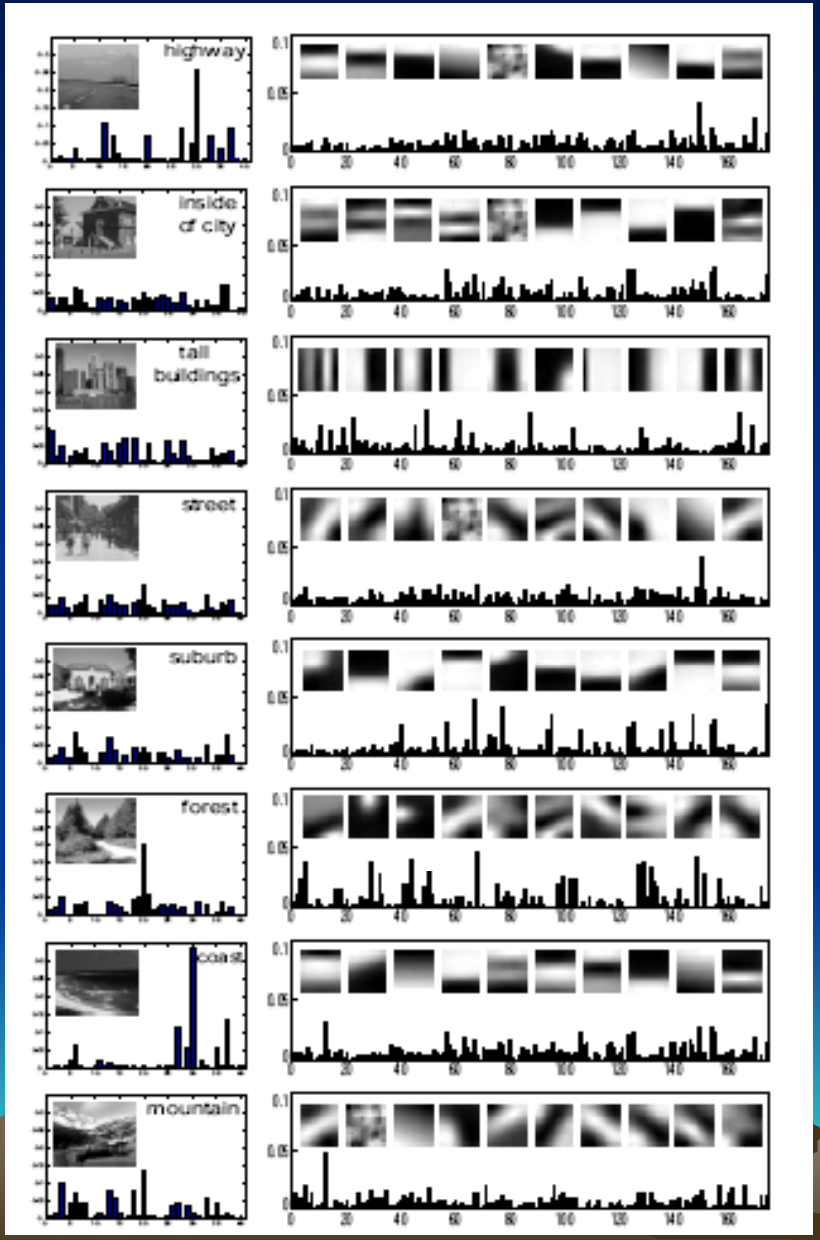


Figure 4. A codebook obtained from 650 training examples from all 13 categories (50 images from each category). Image patches are detected by a sliding grid and random sampling of scales. The codewords are sorted in descending order according to the size of its membership. Interestingly most of the codewords appear to represent simple orientations and illumination patterns, similar to the ones that the early human visual system responds to.

Topic Models (cont'd)



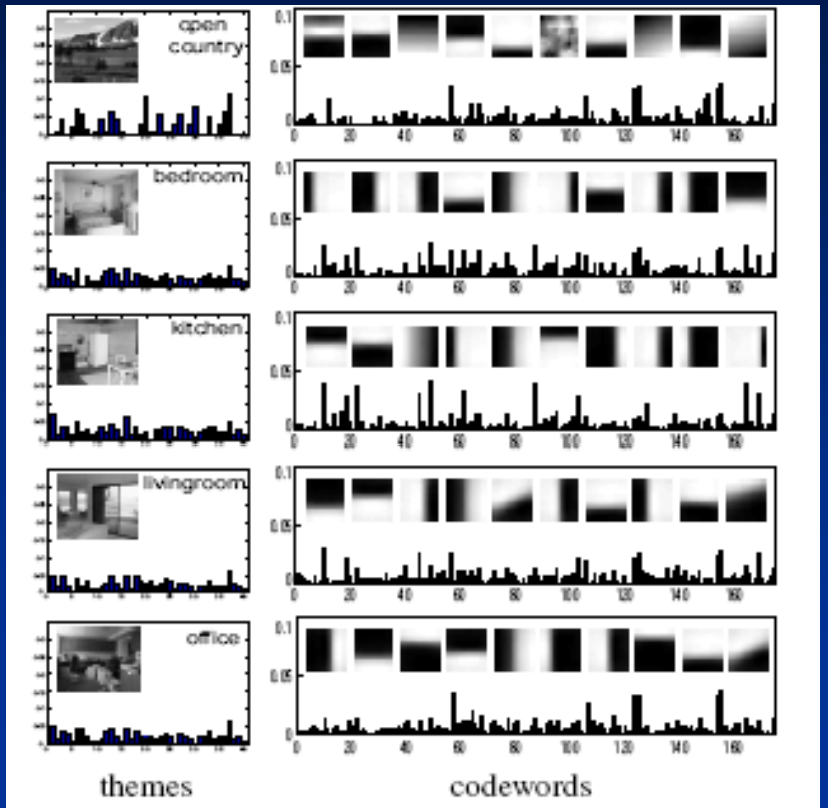
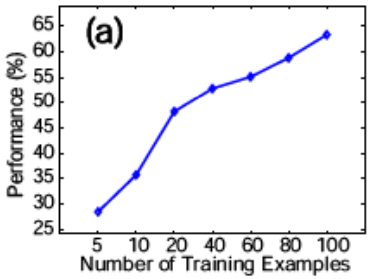


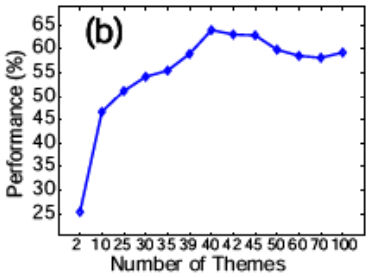
Figure 5. Internal structure of the models learnt for each category. Each row represents one category. The left panel shows the distribution of the 40 intermediate themes. The right panel shows the distribution of codewords as well as the appearance of 10 codewords selected from the top 20 most likely codewords for this category model.

[Fei-Fei 2005]

Topic Models (cont'd)

- Results





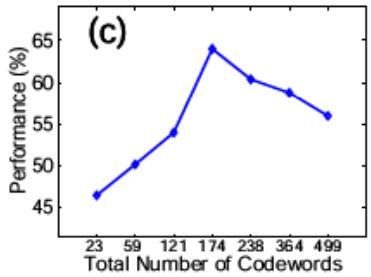


Figure 10. (a) Number of training examples vs. performance. (b) Number of themes vs. performance. (c) Number of codewords vs. performance. All performances are quotes from the mean of the confusion table.

	# of categ.	training # per categ.	training requirements	perf. (%)
Theme Model 1	13	100	unsupervised	76
[17]	6	$\sim$ 100	human annotation of 9 semantic concepts for 60, 000 patches	77
[9]	8	250 $\sim$ 300	human annotation of 6 properties for thousands of scenes	89

Table 2. Comparison of our algorithm with other methods. The average confusion table performances are for the 4 comparable categories (forest, mountain, open country and coast) in all methods. We use roughly 1/3 of the number of training examples and no human supervision than [9]. Fig.10(a) indicates that given more training examples, our model has the potential of achieve higher performances.

[Fei-Fei 2005]

Topic Models (cont'd)

- Strengths
  - Unsupervised
  - Invariant to composition
- Weaknesses
  - No geometry
  - Matches of themes to categories
  - No correspondence to semantic categories

Comparison

- Global vs. local
  - Spatial Envelope, Spatial Pyramid
  - Topic Models
- Viewpoint / location biases vs. invariability
  - Spatial Pyramid
  - Topic Models, Spatial Envelope

Comparison (cont'd)

- Intermediate representations
  - Spatial Envelope, Topic Models
- Supervision vs. no supervision
  - Spatial Envelope
  - Topic Models, Spatial Pyramid
- Object recognition?

Discussion

- Object recognition vs. scene recognition
  - Global approaches
    - Spatial Pyramid, scenes vs. objects results
  - Bag-of-features
- Use of scene recognition
- Ambiguous scenes
- Human recognition of scenes
  - Importance

References

[Fei-Fei 2005] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. CVPR 2005.

[Grauman 2005] K. Grauman and T. Darrell. The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. ICCV 2005.

[Hays 2007] J. Hays and A.A. Efros. Scene completion using millions of photographs. SIGGRAPH 2007.

[Lazebnik 2006] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. CVPR 2006.

[Oliva 2001] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. IJCV 2001.