Statistics of Natural Image Categories

Authors: Antonio Torralba and Aude Oliva

Presented by:
Sebastian Scherer

Experiment

Please estimate the average depth from the camera viewpoint to all locations(pixels) in the next picture.

Next you will see a circle for 3s, then a picture for 1s.

Concentrate on the circle.

After that I will ask you about the average depth of the picture.

What would you estimate the mean depth of this picture was?

What is the mean depth of this picture?

What is the mean depth of this picture?

Problem

We want to determine global properties about an image.

However avoid

- Explicit segmentation
- No object recognition

Properties that are important for later stages of image processing are

- Scale
- Type

Outline

Global features: power spectra

Examples

Localized spectra

Applications of global features

- Naturalness/Openness classification
- Scene categorization
- Object recognition
- Depth estimation

Power spectra

Decompose the image using a discrete Fourier transform:

$$I(f_x, f_y) = \sum_{x,y=0}^{N-1} i(x, y)h(x, y)e^{-j2\pi(f_x x + f_y y)}$$

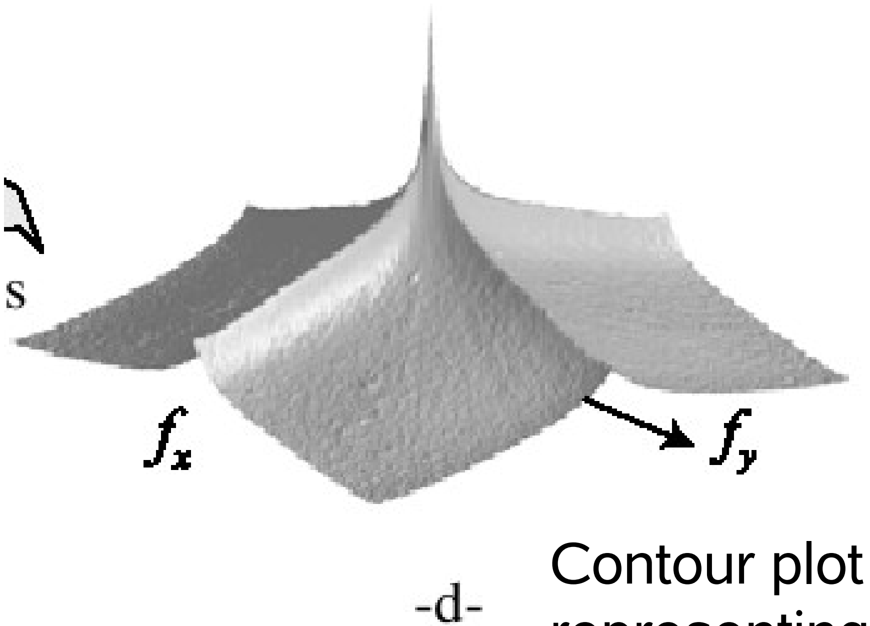
$$= A(f_x, f_y)\, e^{j\Phi(f_x, f_y)}$$

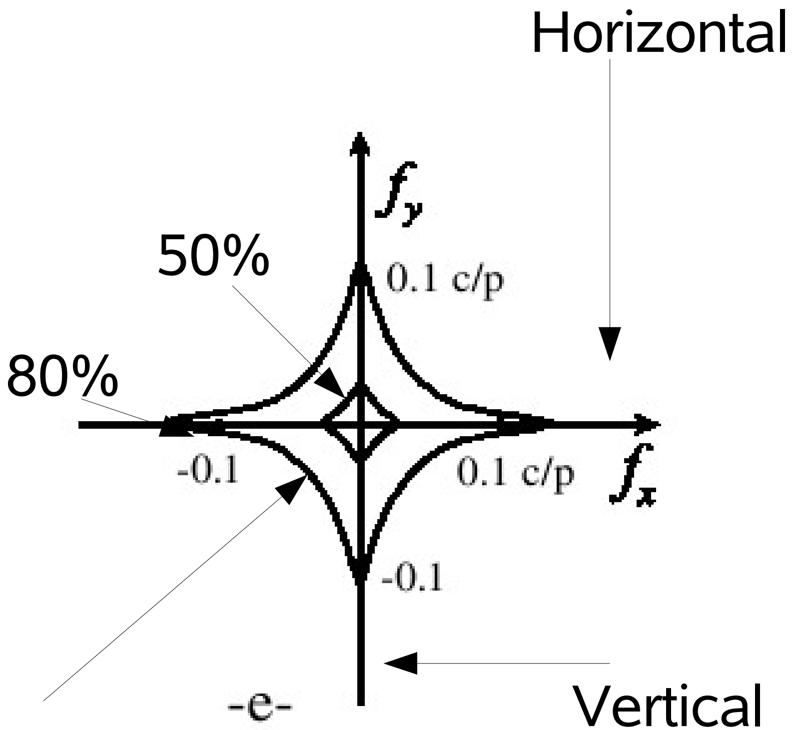
The power spectrum is then given by the amplitude and phase

$$A(f_x, f_y) = |I(f_x, f_y)|$$

$$\Phi(f_x, f_y)$$

Power spectrum plot

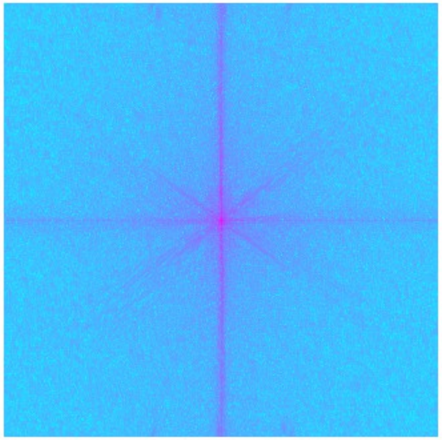


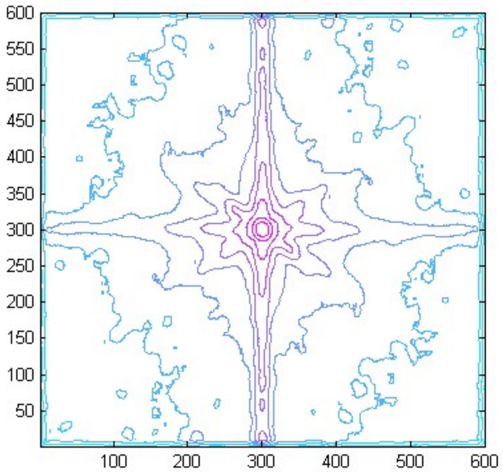


Contour plot representing a percentage of total energy of the spectrum

Computing and visualizing a spectrum in Matlab

- Computing the Spectrum (Matlab):
  - `Ifft = abs(fftshift(fft2(I,w,h)));`
- Visualization:
  - `imshow(log(Ifft)/max(max(log(Ifft))));`
  - `colormap(cool);`
- (From Jonathan Huang's slides)





Interesting fact: 1/f Spectra

Natural Image Spectra follow a power law!

$$A(f) \approx \frac{A_s(\theta)}{f^{2-\eta(\theta)}}$$

$A_s(\theta)$ is called the *Amplitude Scaling Factor*

$2-\eta(\theta)$ is the *Frequency Exponent*. η clusters around 0 for natural images.

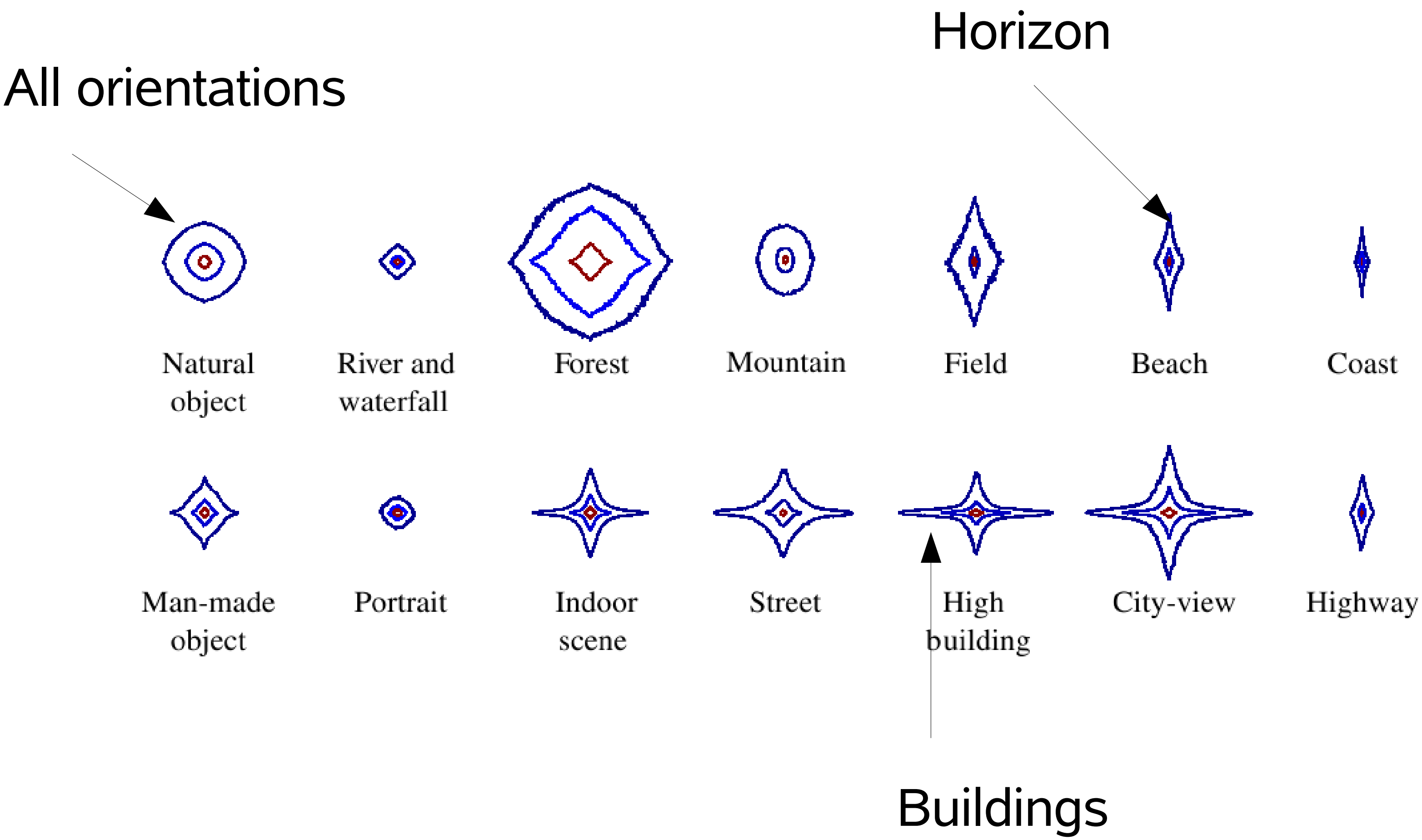
Any guesses on why this law holds?

Use the spectra of images in order to categorize a scene

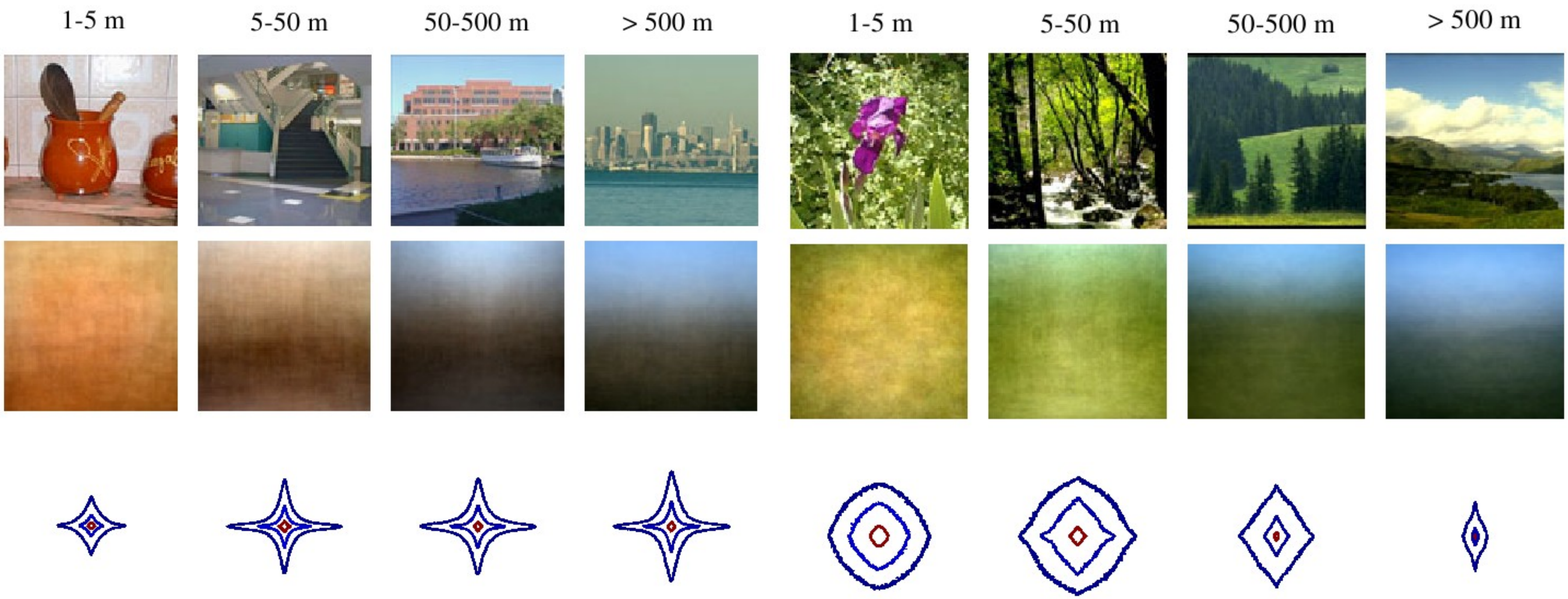
Why is this a good idea?

- Texture varies with scene scale:
  - Atmosphere is a low pass filter?
    - Sky
    - Mountains
  - vs
    - Leaves
    - Objects
- Phase varies with environment:
  - Man-made
  - Nature

Spectral signatures for different scenes



Scene scales



"The point of view that any given observer adopts on a specific scene is constrained by the volume of the scene."

What can one spectrum of an image not capture?

Generally we have a upright viewpoint.
The horizon is towards the top.

Non-stationary power spectra

Decompose the image using a DFT in local regions:

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

The localized power spectrum is then given by the amplitude and phase for a specific region

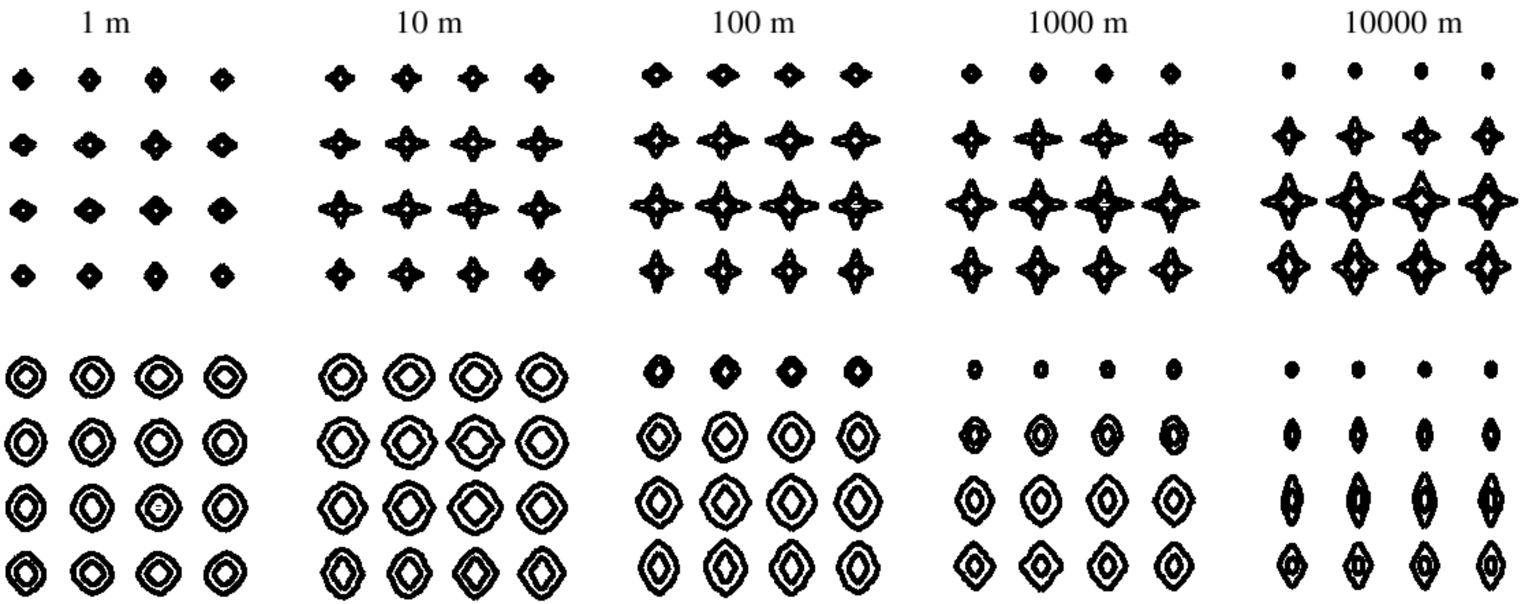
$$A(x, y, f_x, f_y)^2 = |I(x, y, f_x, f_y)|^2$$

In their case: 8x8 spatial locations

Non-stationary power spectra at different depth scales

Man-made

Natural



What can we do with the power spectra?

Replicate perception of humans along different scales

- Naturalness
- Openness

Semantic categorization

- Determine context of scene
- Apply specialized methods after context is determined

Object recognition

- Determine if an object exists in the scene
- Only presence no location
- Likely regions of objects

Depth estimation

- Estimate the mean depth of the scene
- Provides cue for object recognition

Naturalness vs. Openness

Projection of images on the second and third principal component.

Openness

Naturalness is represented by the third principal component

PCA

$\Gamma(k_x, k_y)$, The power spectrum

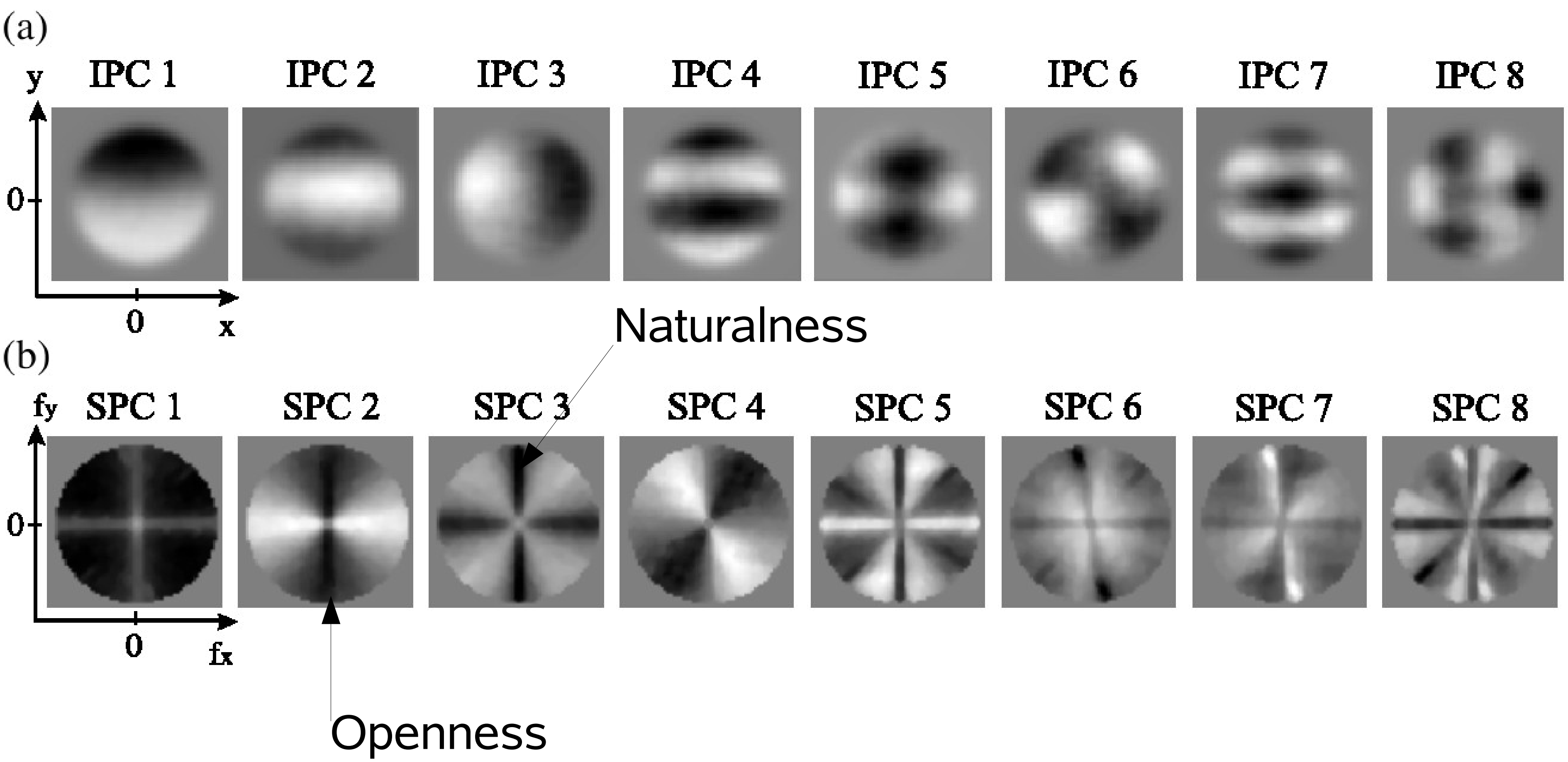
Normalization:

$$\Gamma'(k_x, k_y) = \Gamma(k_x, k_y)/\mathrm{std}[\Gamma(k_x, k_y)]$$
$$\mathrm{std}[\Gamma(k_x, k_y)] = \sqrt{E[(\Gamma(k_x, k_y) - E[\Gamma(k_x, k_y)])^2]}$$

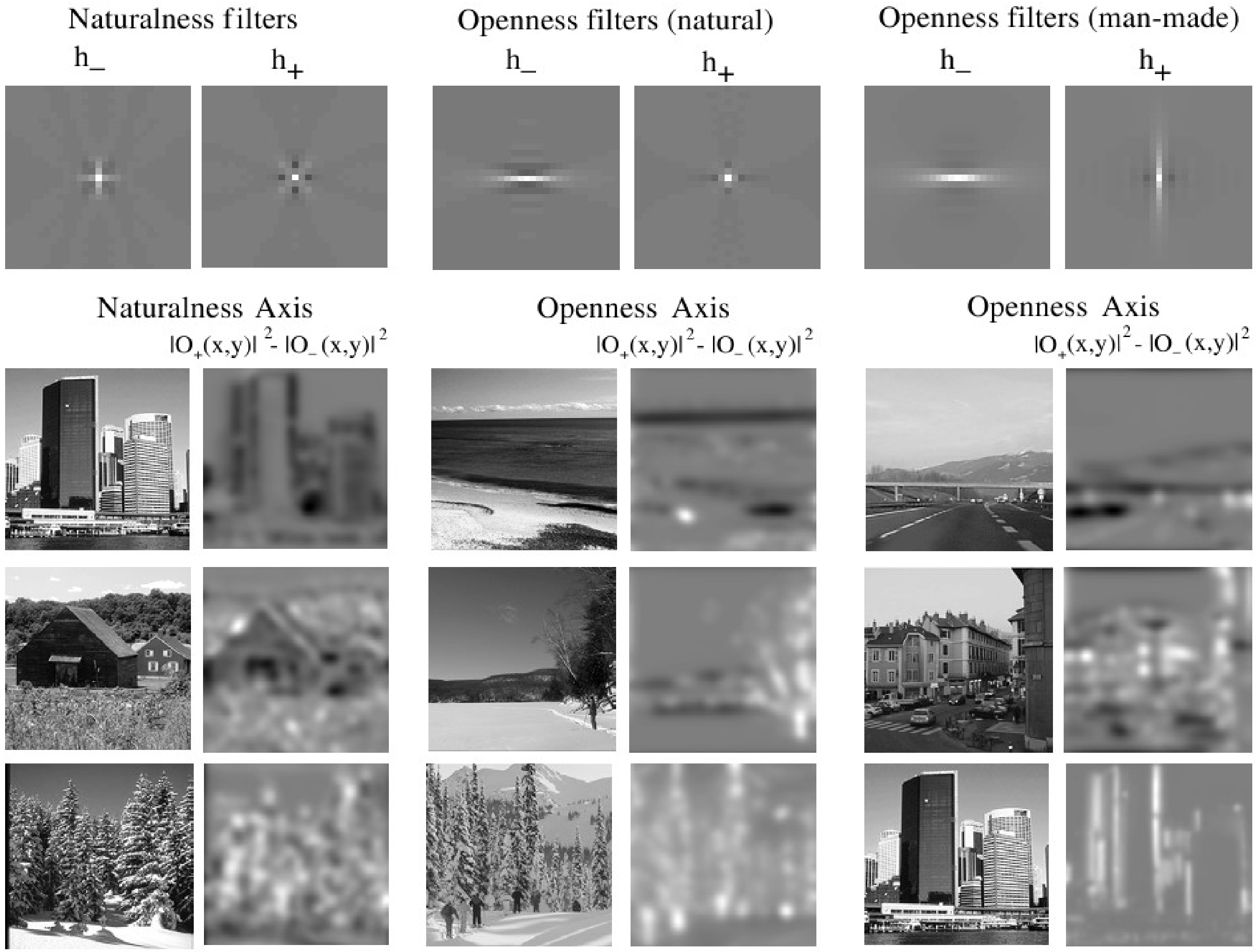
Perform PCA on the normalized power spectrum to get the spectral principal components (SPC).

$$\Gamma'_s(k_x, k_y) = \sum_{n=1}^{P} u_n \mathrm{SPC}_n(k_x, k_y)$$

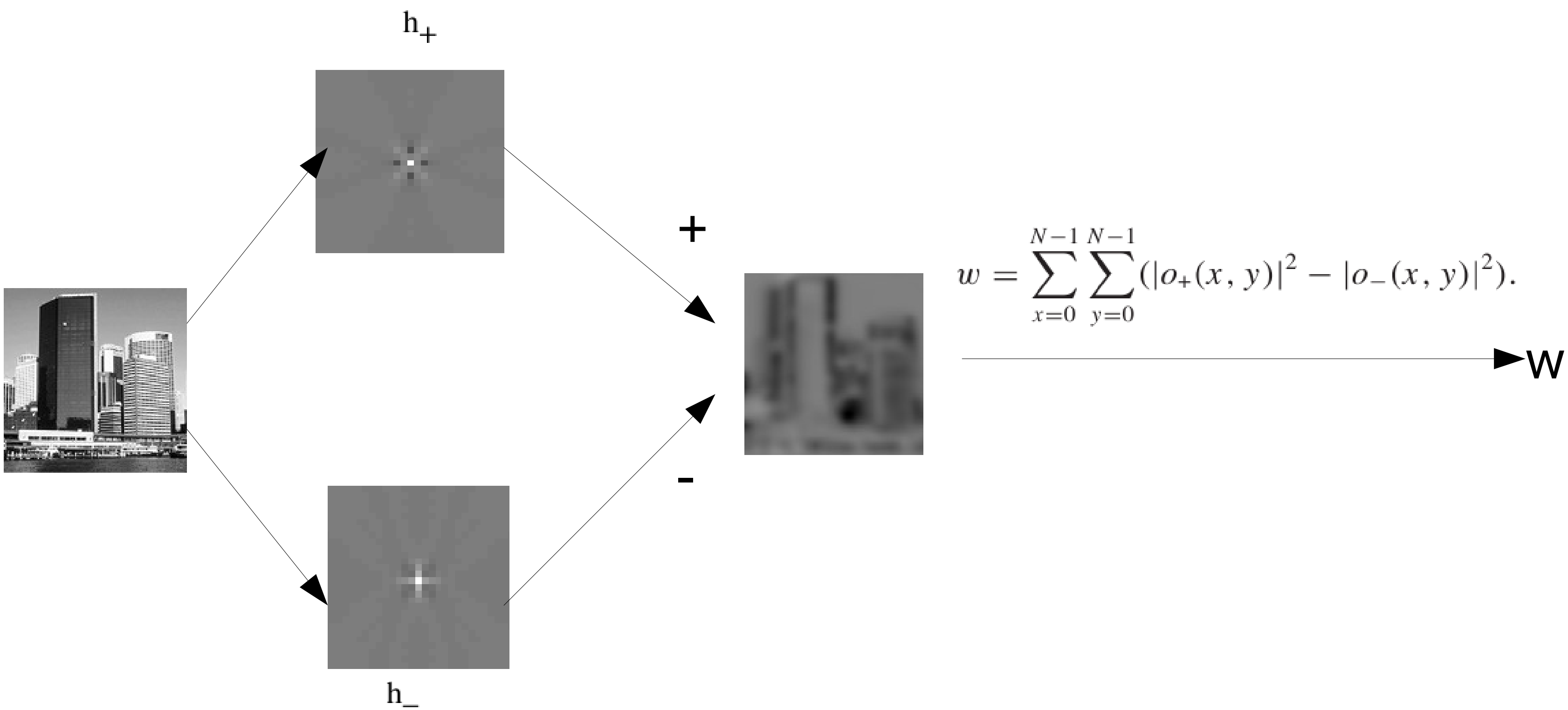
The first principal components of a set of images



Semantic categorization

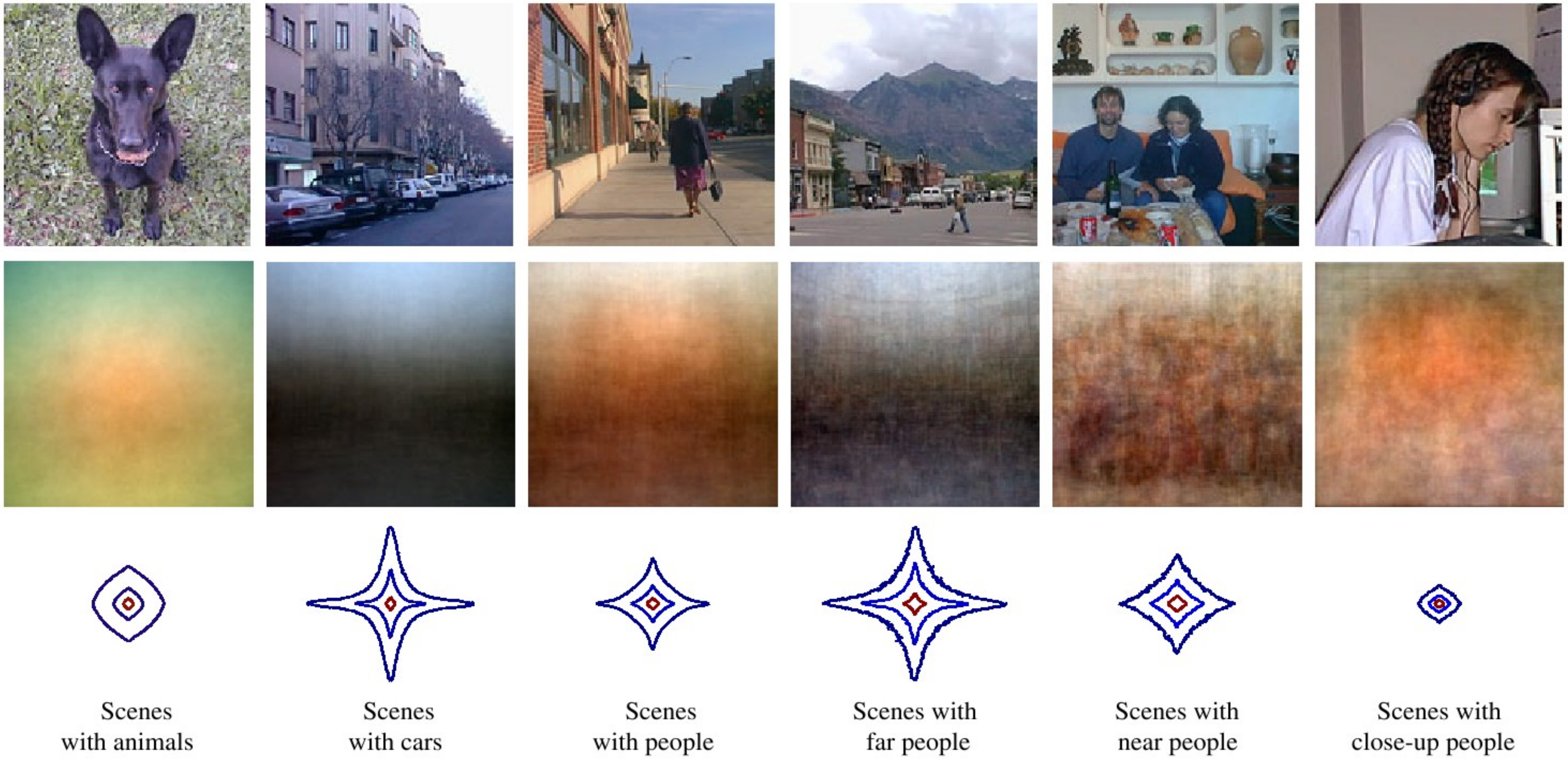


Semantic categorization calculation



$$w = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (|o_+(x, y)|^2 - |o_-(x, y)|^2).$$

Object recognition



Object recognition – Algorithm

During training phase learn $P(O|\vec{v}_C)$ from a set of annotated pictures.

O: object class, v_c: image statistics

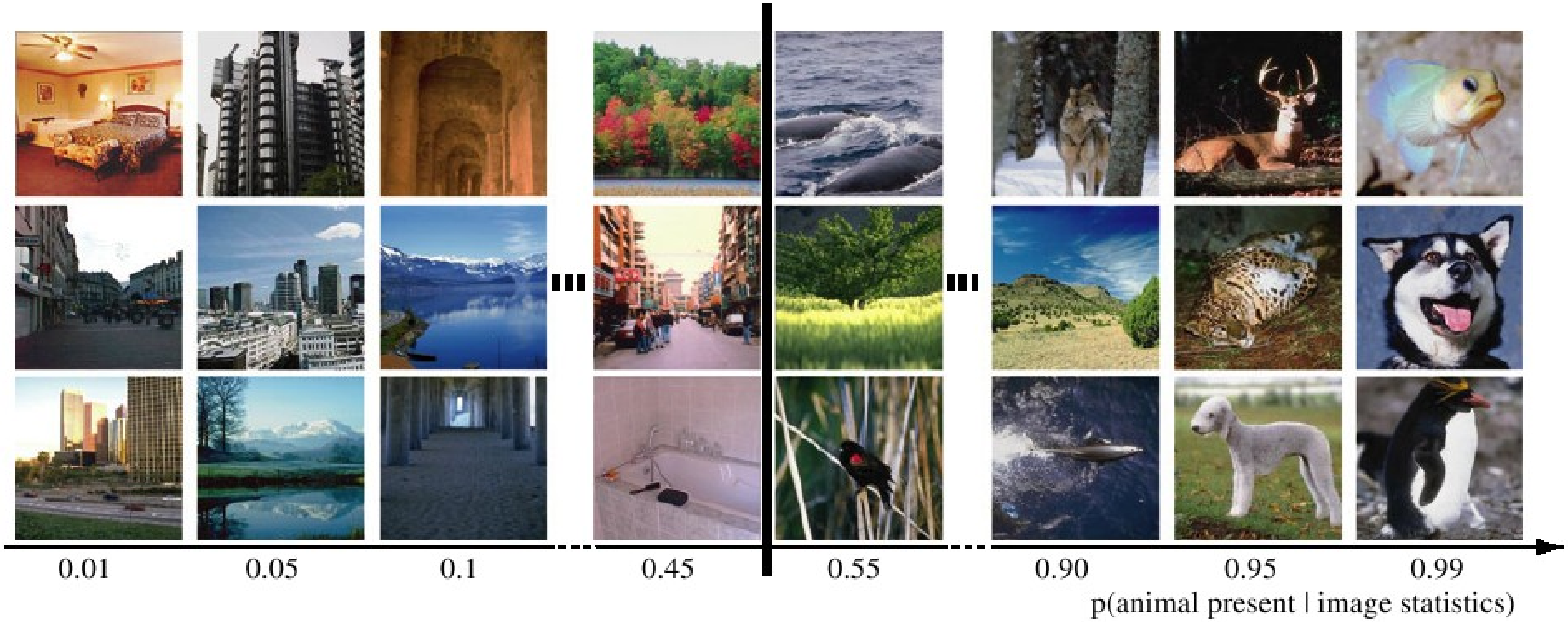
Bayes rule (without marginalization):

$$P(O|\vec{v}_C) = P(\vec{v}_C|O)P(O) + P(\vec{v}_C|\overline{O})P(\overline{O})$$

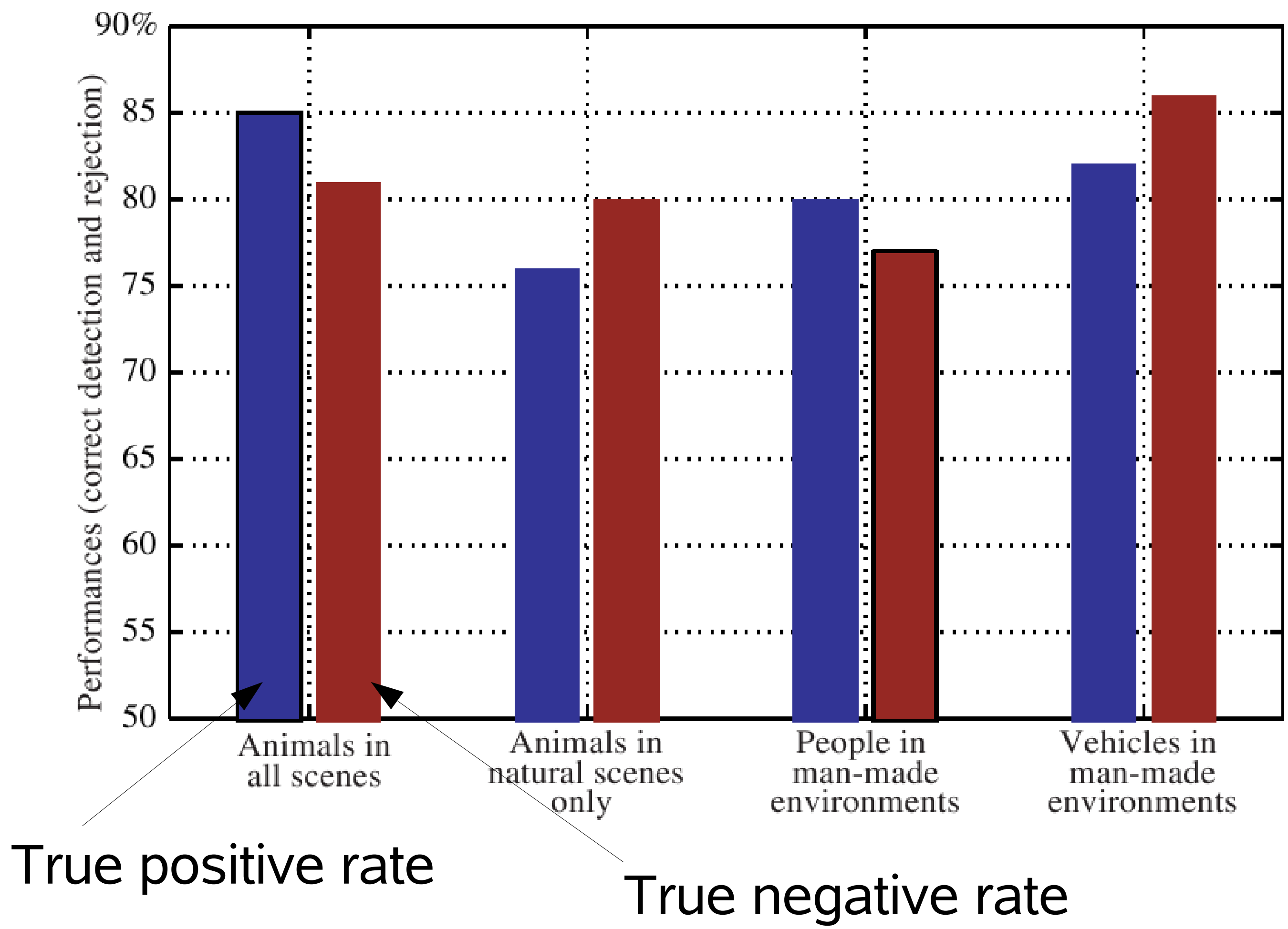
Equal prior is assumed: $P(\overline{O}) = P(O) = 1/2$

Estimate $P(\vec{v}_C|O)$ $P(\vec{v}_C|\overline{O})$ with a mixture of Gaussians from training set.

Object recognition – Results 1



Object recognition – Results 2



Adding spatial information

Split the picture in four equal regions.

Learn a mixture of Gaussians in order to determine the region where one is most likely to find an object.

$P(\vec{x}|\vec{v}_C)$ Given the spectral feature at four location what is the most likely position of a face.

$$P(\vec{x}, \vec{v}_C) = \sum_{i=1}^{M} b_i G(\vec{x}; \vec{x}_i, \boldsymbol{X}_i) G(\vec{v}_C; \vec{v}_i, \boldsymbol{V}_i)$$

Attention will be on the most likely region to find a face.

Regions of interest

Figure 15. Examples of images and the selected regions that are expected to contain faces based on contextual features. The regions are selected according to global image statistics and not to the actual presence of the object of interest.

90% of faces were within a region of
35% of the size of the image of the largest $P(x|v_c)$

Gist $p(X = x, y, s|G)$

Use the global features as a prior on the location of objects in a object detection and localization algorithm. Since *x* is dependent on many factors only learn *y* and *s*.

$$p(X, G) = \sum_q P(q)P(G|q)P(X|G,q) = \sum_q \pi(q)\mathcal{N}(G; \mu_q^{(1)}, \Sigma_q^{(1)})\mathcal{N}(X; W_q G + \mu_q^{(2)}, \Sigma_q^{(2)})$$

where π(q) are the mixing weights, W is the regression matrix, μ are mean vectors, and Σ are covariance matrices for cluster q.

Gist – Results

Fig. 8. Example of location priming for screens, keyboards, cars and people using global features. In each group, the image on the left is the input image, and the image on the right is the input image multiplied by the probability, given the gist, of the object being present at a given location, i.e., $I. * P(x|G(I))$.

Depth Estimation – Feature vector

Have a feature vector v.

v' consists of the downsampled energy vector

$$A_M^2(\mathbf{x}, k) = \left\{ |I(\mathbf{x}, k)|^2 \downarrow M \right\}.$$

k: wavelet index, x: location, M spatial resolution

Feature vector size: M^2 K

Apply PCA to reduce the dimensionality of v' to get v.

v is a L-dimensional vector obtained by projecting v' on the first L principal components.

=> v is size L.

Depth Estimation - Learning

Want to optimize this expression

$$\hat{D} = E[D \mid \mathbf{v}] = \int_{-\infty}^{\infty} D\, f_{D|v}(D \mid \mathbf{v})\, dD$$

$$f(D, \mathbf{v} \mid art) = \sum_{i=1}^{N_c} g(D \mid \mathbf{v}, c_i)\, g(\mathbf{v} \mid c_i),\ p(c_i)$$

D:depth, v: features, Nc: number of clusters, p(ci): cluster weight, g(v|ci): multivariate gaussian, g(D|v,ci):

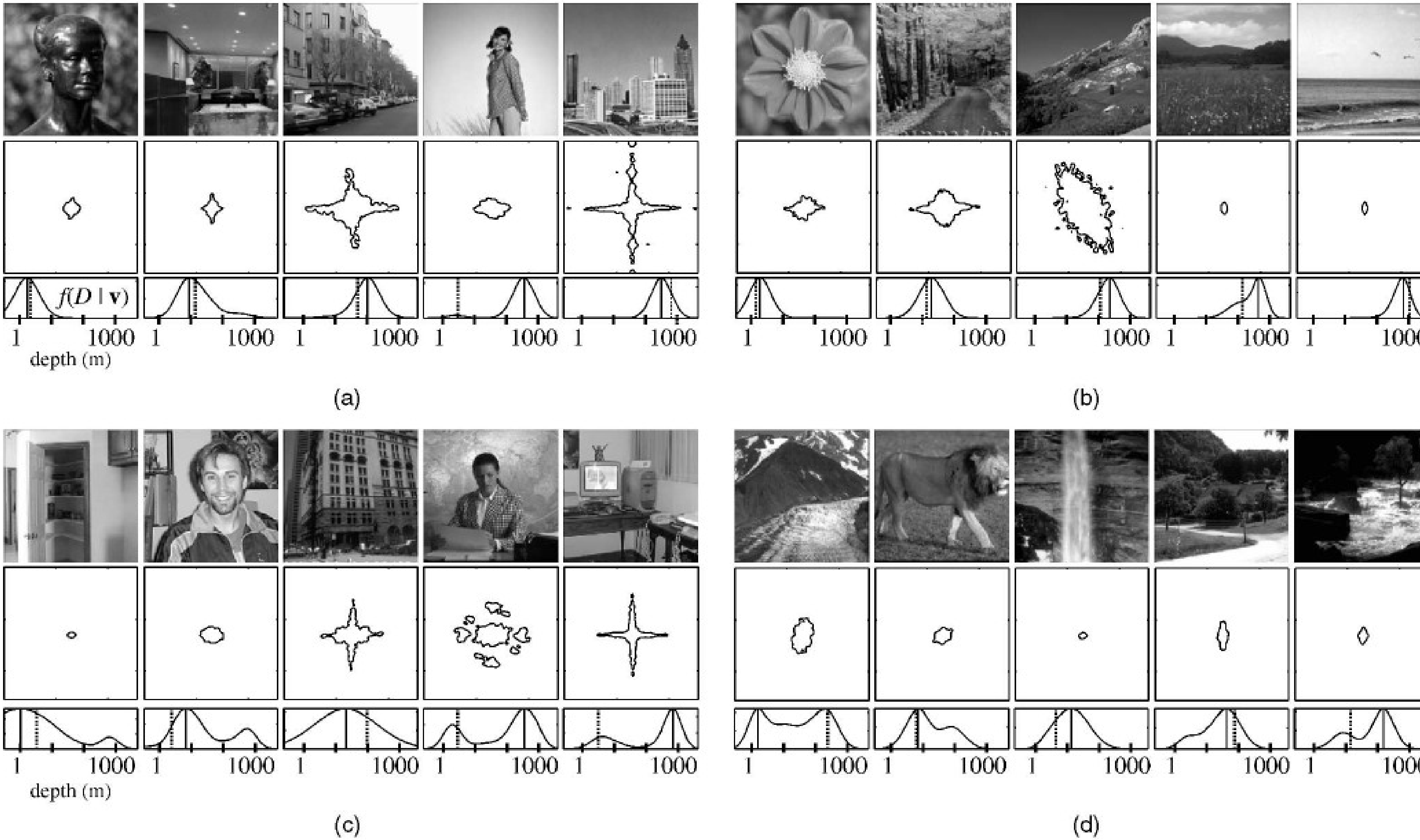
$$g(D \mid \mathbf{v}, c_i) = \frac{\exp\left[-\left(D - a_i - \mathbf{v}^T \vec{b}_i\right)^2 / 2\sigma_i^2\right]}{\sqrt{2\pi}\sigma_i}$$

Result is a mixture of linear regressions:

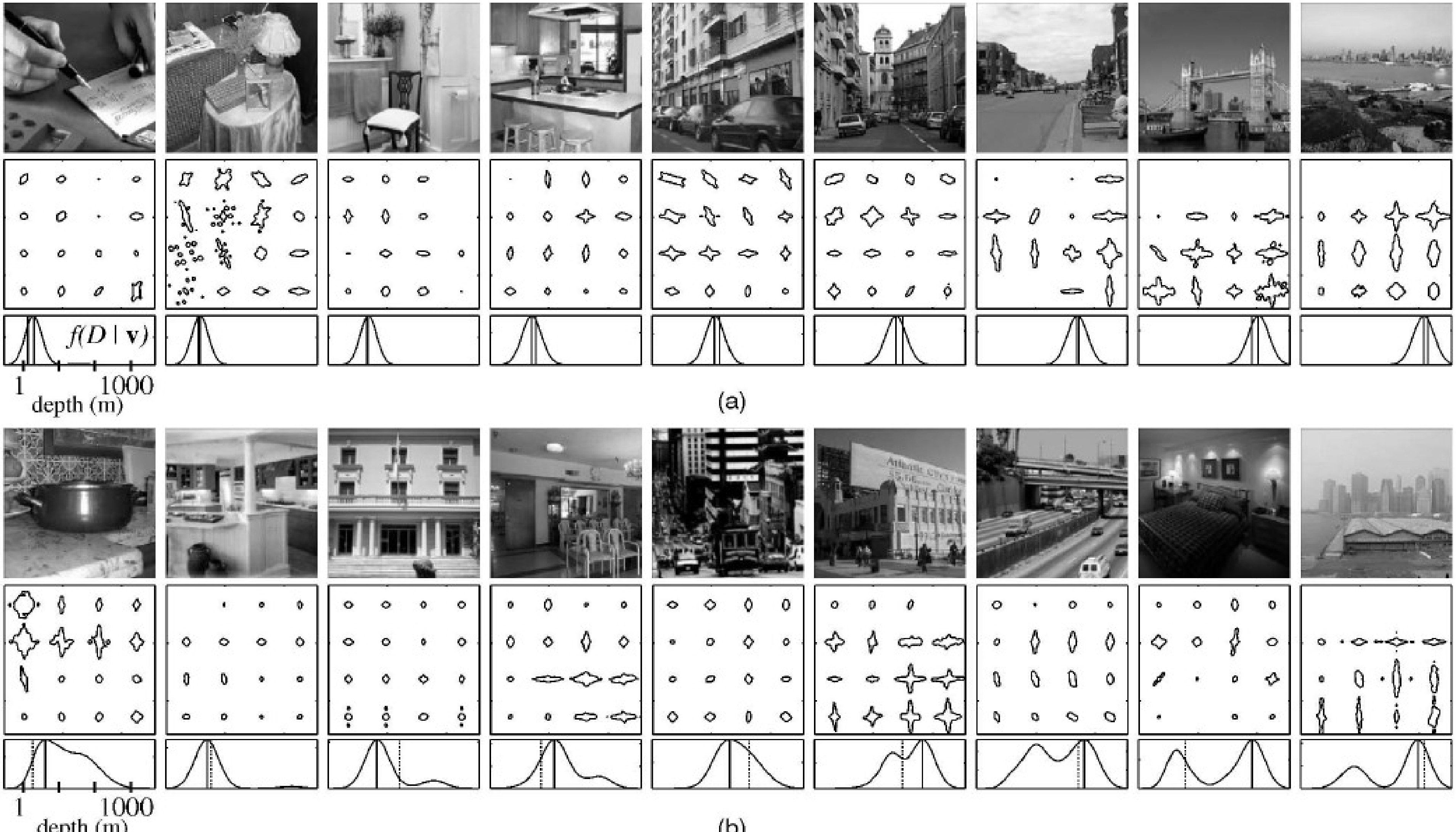
$$\hat{D} = \frac{\sum_{i=1}^{N_c} (a_i + \mathbf{v}^T \vec{b}_i)\, g(\mathbf{v} \mid c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v} \mid c_i)\, p(c_i)}$$

$$\sigma_D^2 = E[(\hat{D} - D)^2 | \mathbf{v}] = \frac{\sum_{i=1}^{N_c} \sigma_i^2\, g(\mathbf{v} \mid c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v} \mid c_i)\, p(c_i)}.$$

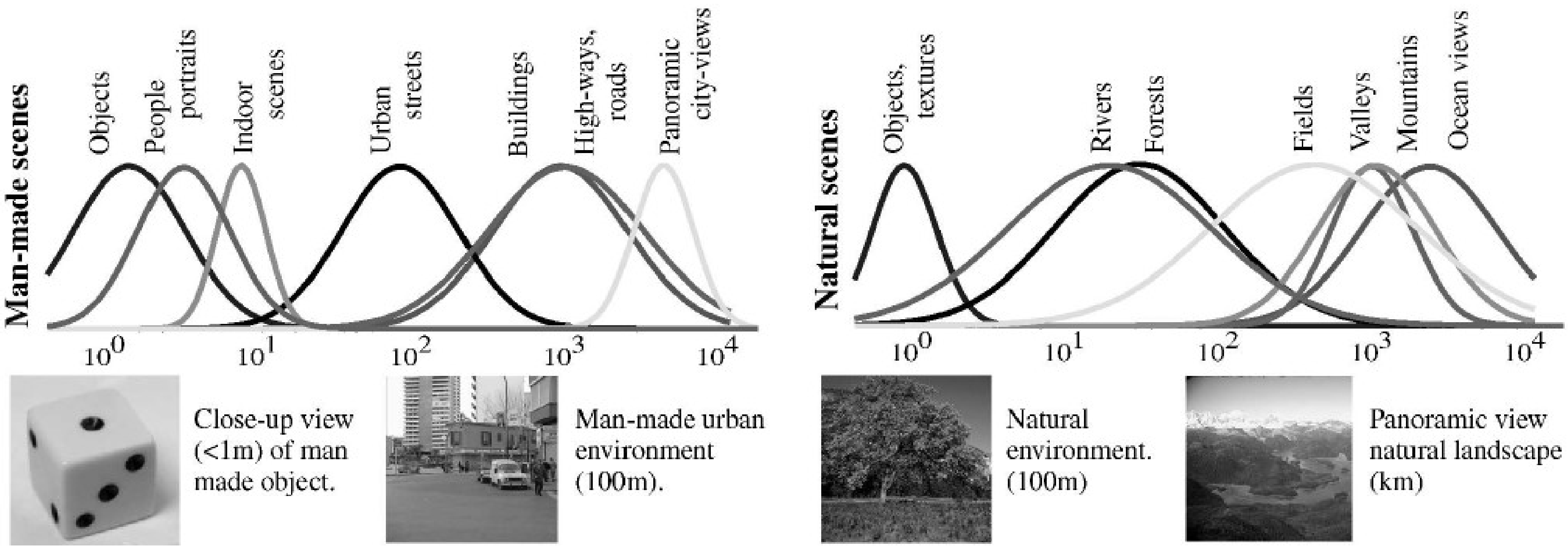
Depth Estimation – Global features



Depth Estimation – Localized features



Scene category from depth



Depth Estimation – Face Detection

Determine the size of an object as

$$\hat{S}_{obj} \simeq K_{obj} / \hat{D}^2$$

Now have approximately the right scale for object detection:

Discussion

Why do power spectra work so well?

Why is there such a large distinction between man-made and human? Are there possibly more distinct classes? Rural streets?

How do humans calculate *mean* depth when estimating the depth for training?