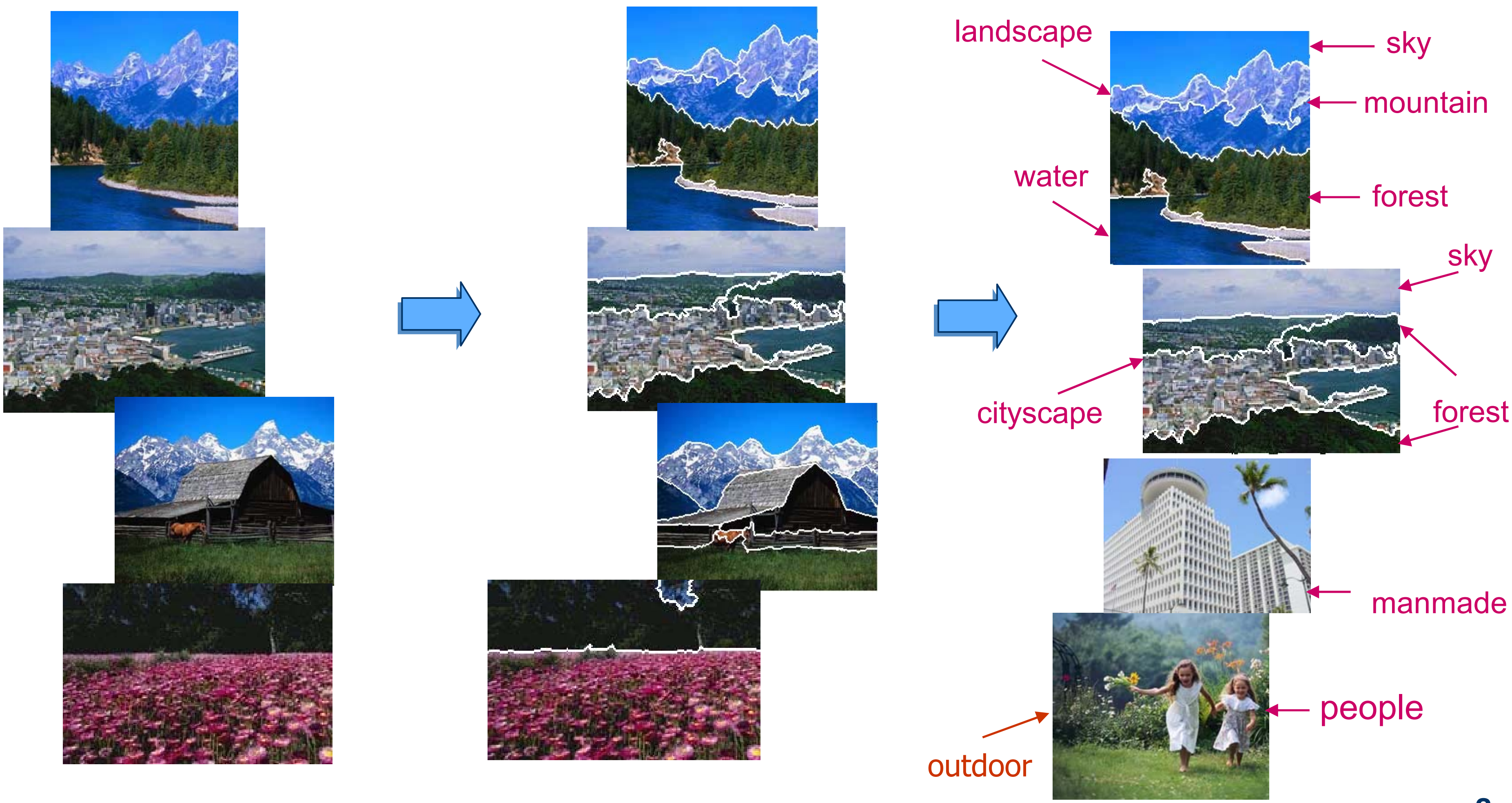
Perceptual Image Segmentation, Background Subtraction, and Semantic Classification

Aggelos K. Katsaggelos
Professor
Northwestern University
Department of EECS
aggk@eecs.northwestern.edu

Problem

Images | "Ideal" Segmentations | Semantic Categories



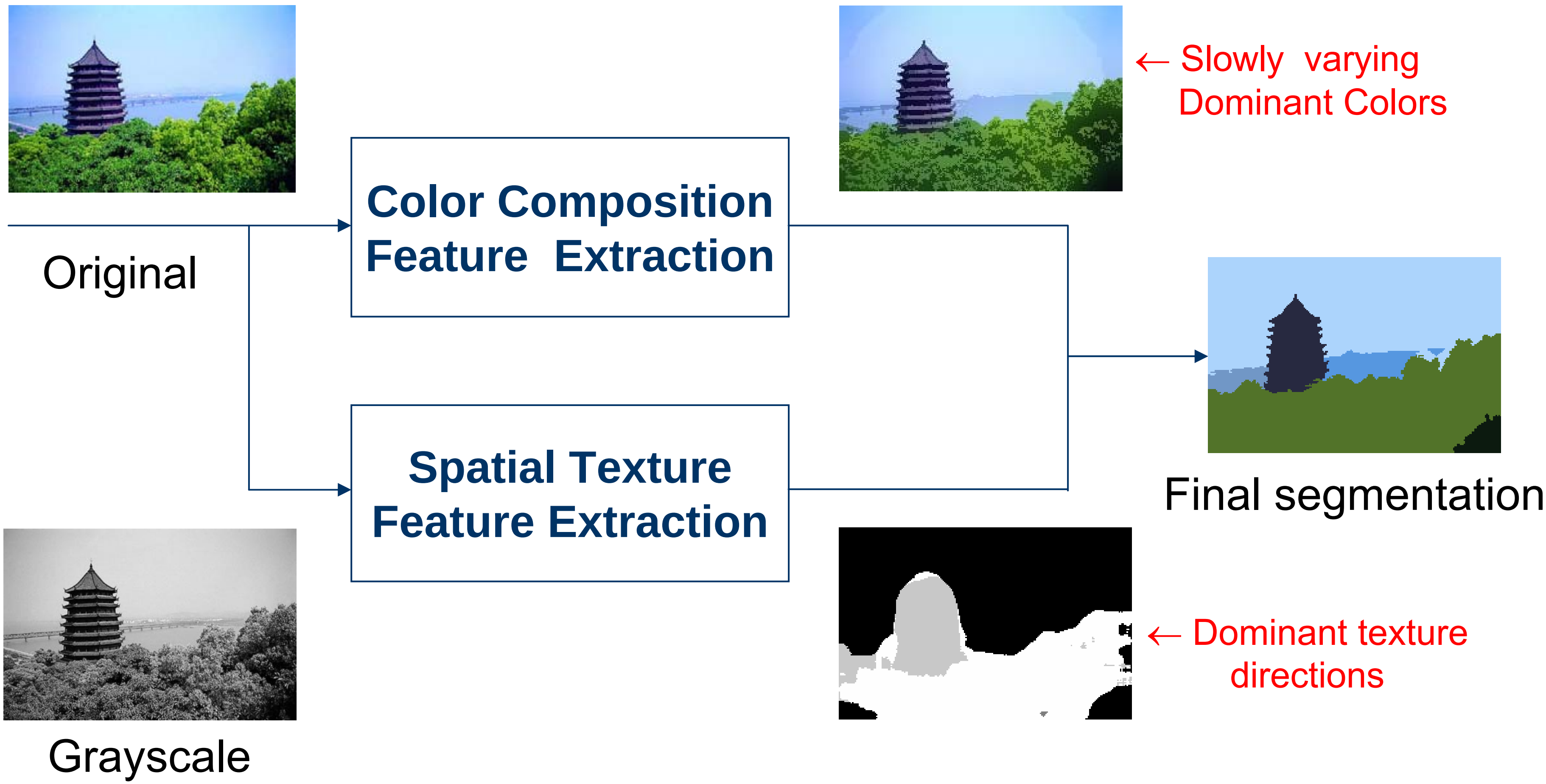
Segmentation Approaches

- Histogram Thresholding
- Clustering
- Edge-based Techniques
- Region Growing
- Split-and-Merge
- Watershed
- Model-Based Approaches

Natural Textures

- Combine color composition, spatial characteristics
- Non-uniform statistical characteristics (lighting, perspective)
- Perceptually uniform
- Need spatially adaptive features
- Small number of parameters

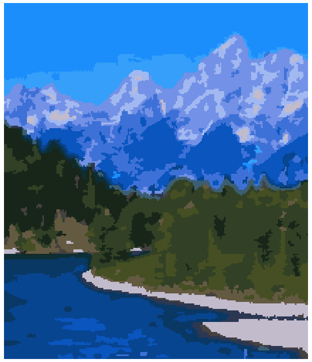
Adaptive Perceptual Color-Texture Segmentation



Color Composition Features

- Dominant Colors
  - Human eye cannot simultaneously perceive a large number of colors
  - Efficient representation
  - Easier to capture invariant properties of object appearance
  - Applied to image classification [Ma'97, Mojsilovic'00]
- Current Approaches
  - K-means (VQ) [LBG'80]
  - Mean-shift [Comaniciu-Meer'97]

  Assumption: constant dominant colors
- Spatially Adaptive Dominant Colors
  - Capture spatially varying image characteristics
  - Use ACA [pappas'92]



Color Composition Features

- Constant Dominant Colors:

$$f_c = \{ (c_i, p_i), \ i = 0, \ldots, n \}$$

c_i: color

p_i: percentage

- Spatially Adaptive Dominant Colors:

$$f_c(s, N_s) = \{ (c_i, p_i), \ i = 0, \ldots, n \}$$

- ACA adapts to local characteristics.
- Dominant colors relatively constant in small neighborhood; but change as we move across the image.

Adaptive Clustering Algorithm (ACA)

- K-means clustering (LBG)
  - Based on image histogram
  - No spatial constraints
  - Each cluster is characterized by constant intensity
- Add spatial constraints
  - Region model: Markov/Gibbs random field
- Make it adaptive
  - Cluster centers spatially varying
  - Texture model: spatially varying mean + WGN
- MAP estimates of segmentation x given observation y

$$p(x \mid y) \propto p(y \mid x)\, p(x)$$

ACA

- K-means minimizes

$$\sum_s (y_s - \mu^{x_s})^2$$

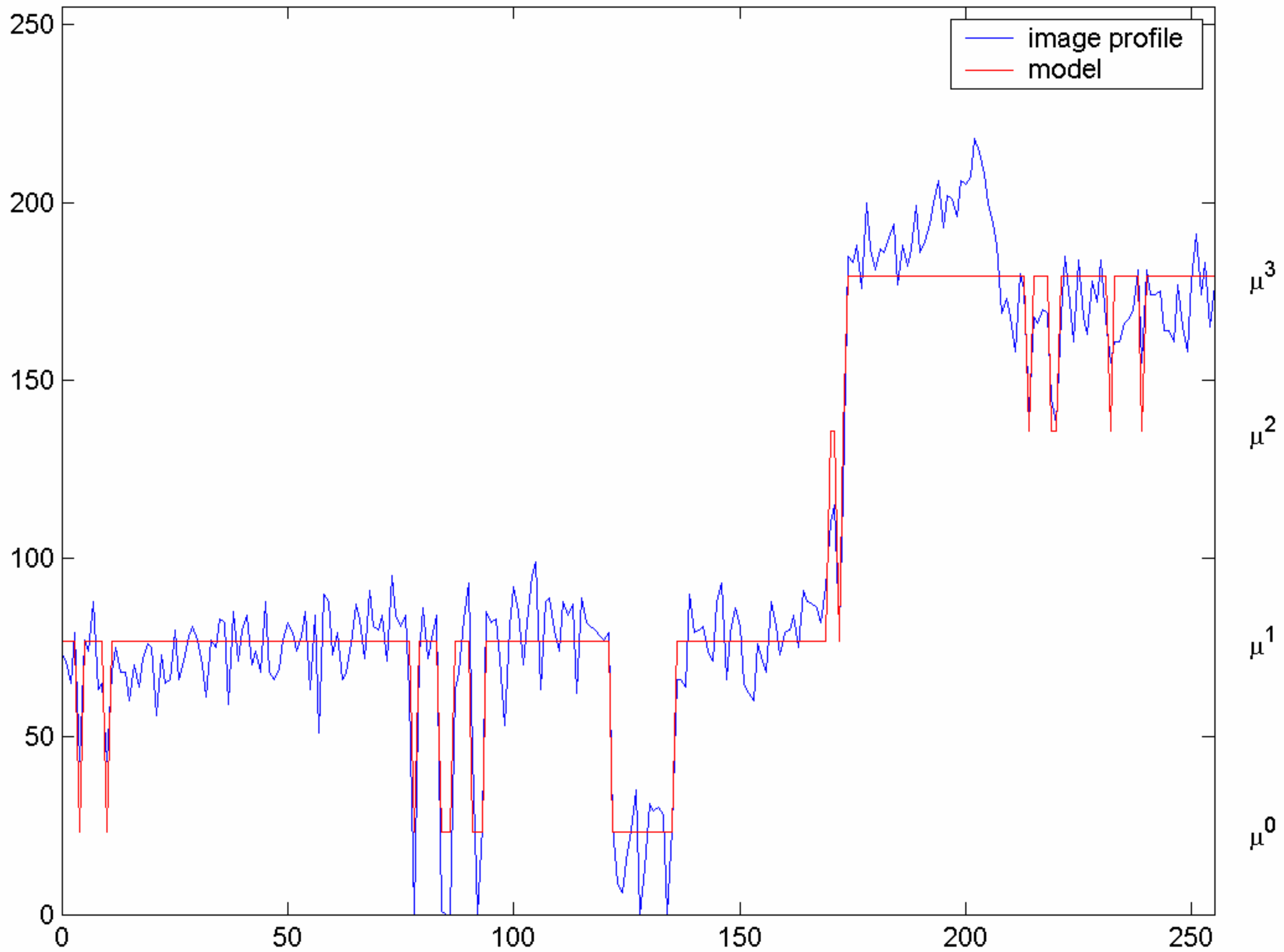
- Adaptive clustering maximizes

$$p(x \mid y) \propto \exp\left\{-\sum_s \frac{1}{2\sigma^2}(y_s - \mu_s^{x_s})^2 - \sum_C V_C(x)\right\}$$

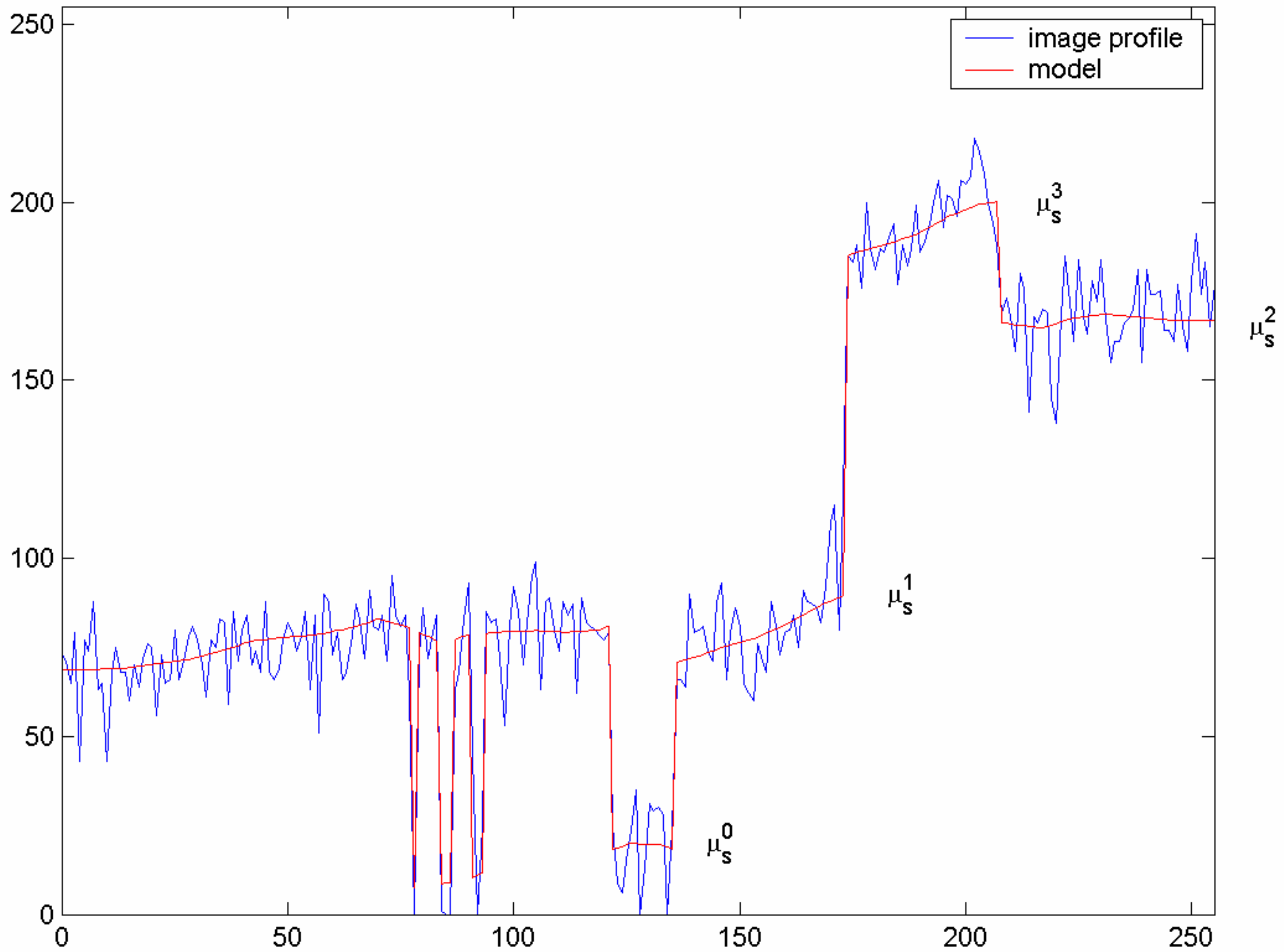
- Or, minimizes

$$\sum_s \frac{1}{2\sigma^2}(y_s - \mu_s^{x_s})^2 + \sum_C V_C(x)$$

K-means Clustering



ACA: Model (15x15)



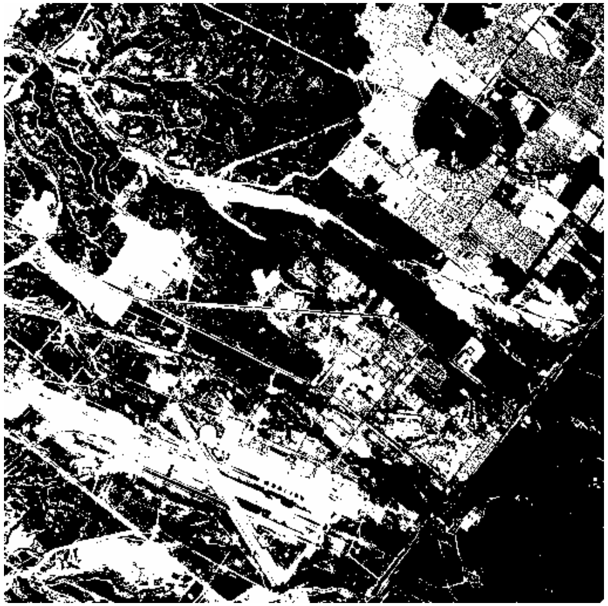
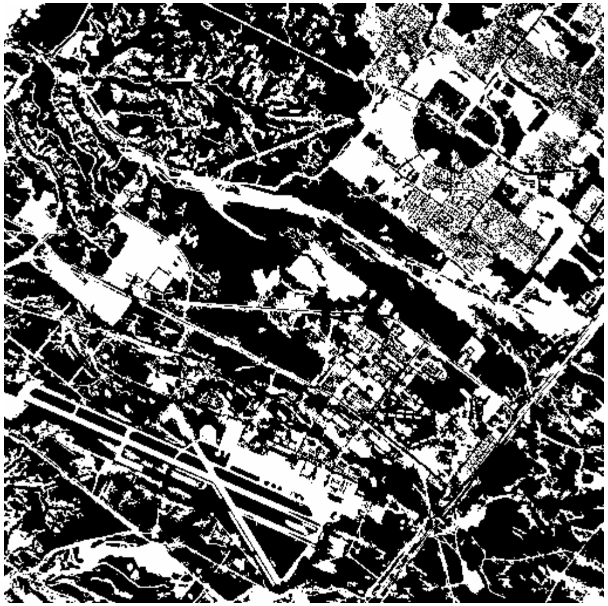
Adaptive Clustering Algorithm

Original Image

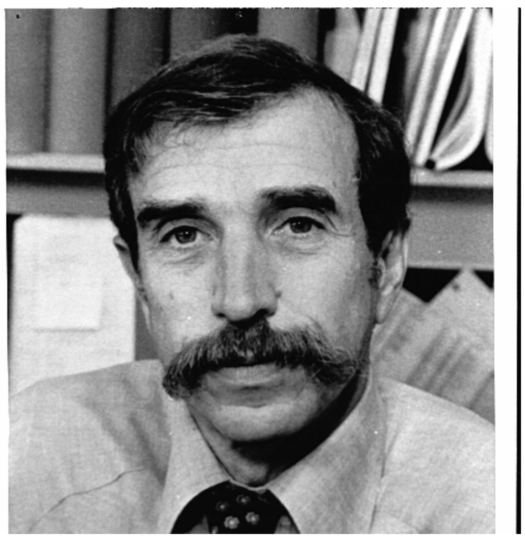
K-means Class Labels

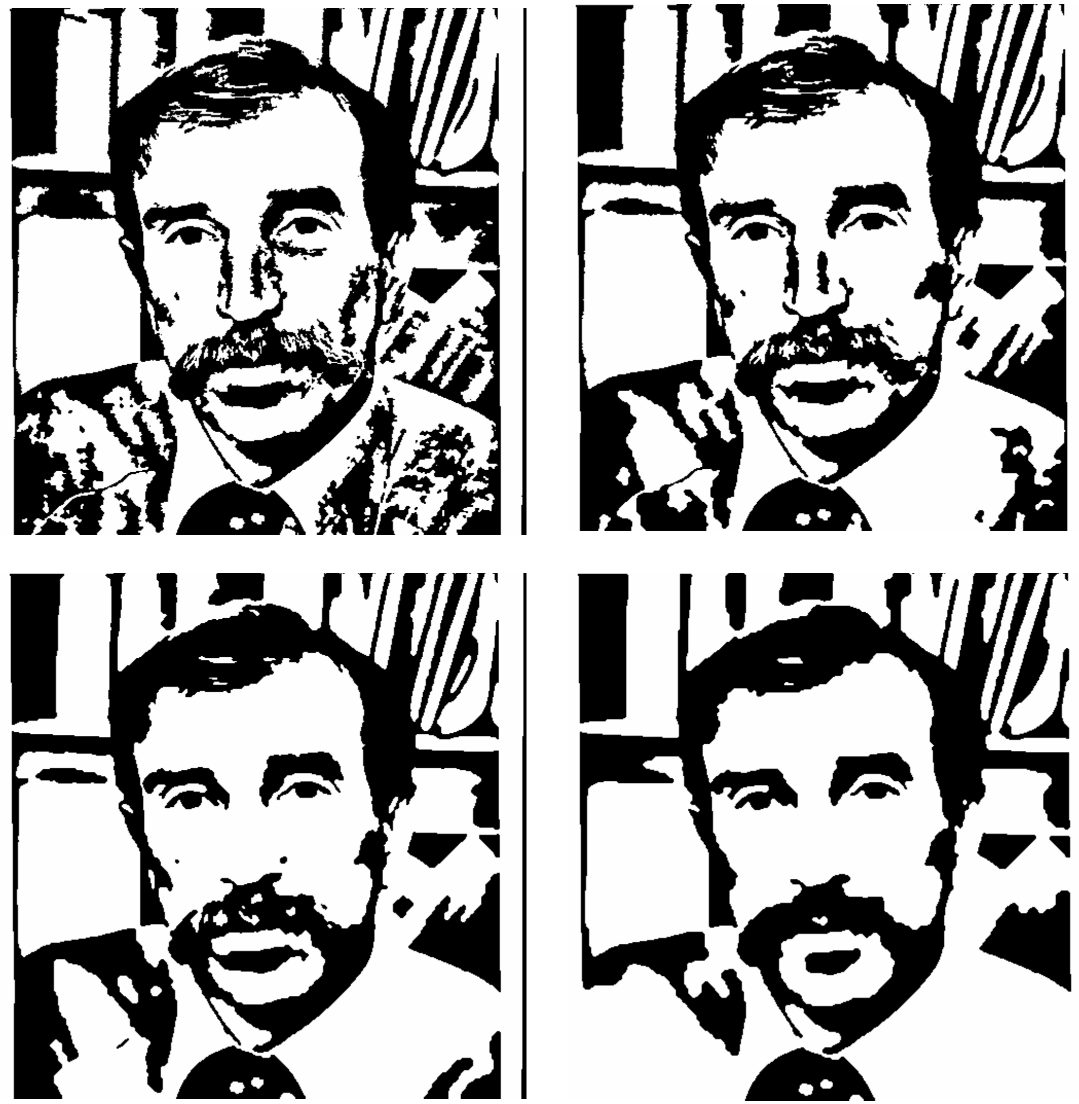
ACA Class Labels

K-means vs. ACA

ACA

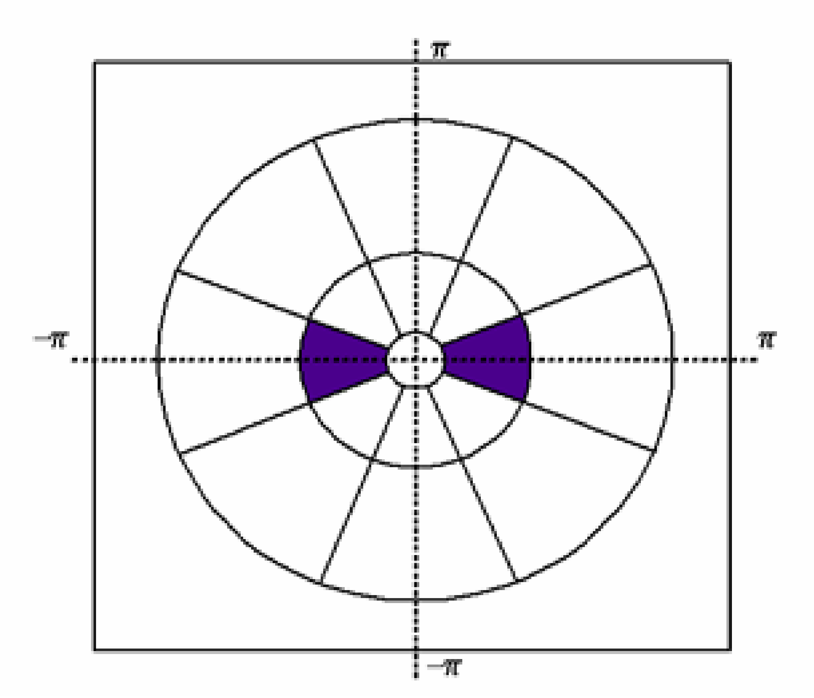


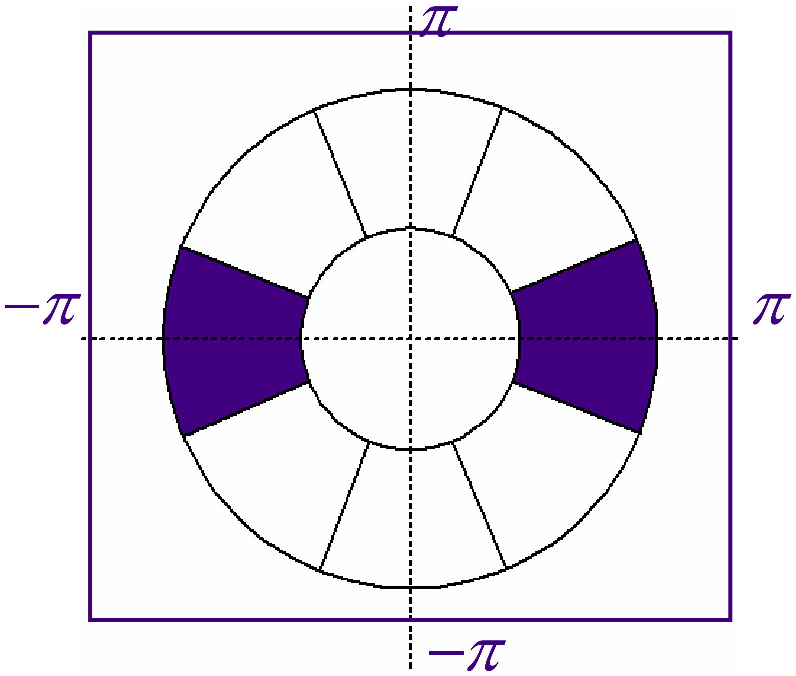


Spatial Texture Features

- Grayscale image component (vs. achromatic pattern map)
- Multiscale frequency decomposition
  - DWT (9/7 Daubechies)
  - Steerable filters [Freeman-Adelson'91]
  - Gabor filters [Daugman'86]
- Energy of subband coefficients is sparse
  - Use local median energy

Steerable Pyramid Decomposition





Ideal spectrum

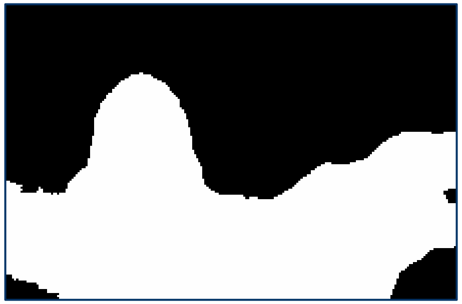
2-level decomposition

Ideal spectrum

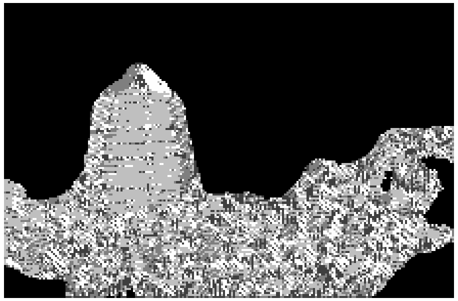
1-level decomposition

Spatial Texture Feature Computation

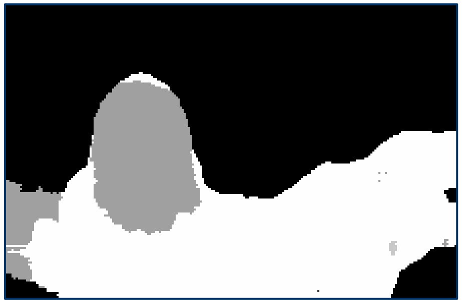
- At each pixel, compute
  - S_{max} = Maximum of 4 subband responses
  - S_i = Index of maximum coefficients
- Smooth vs. non-smooth classification
  - Local median energy of S_{max}
  - 2-level K-means
  - Use threshold provided by subjective test
- Non-smooth region classification
  - Construct local histogram of S_i
  - "Complex" if no dominant orientation
  - Otherwise classify according to dominant orientation as "horizontal," "vertical," "+45," "-45."



Smooth vs. non-smooth



S_i indices



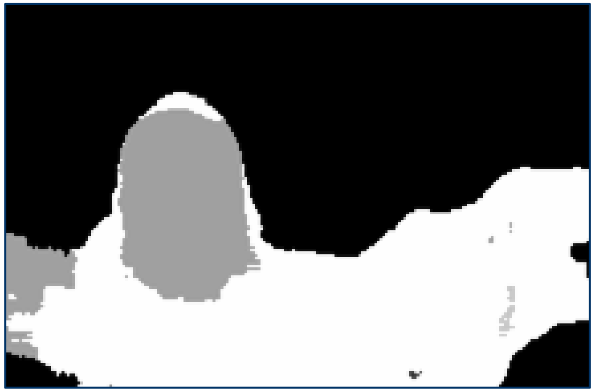
Texture classes

Multi-scale Texture Classification

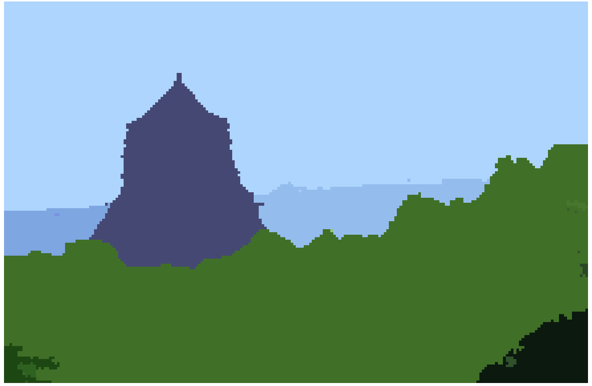
- Apply texture classification at each scale
- Combine texture classes from different scales based on the following rules:
  - "smooth": "smooth" at all scales
  - "Vertical," "Horizontal," "+45^o," "-45^o": consistent texture classification across all scales. Note: "complex" or "smooth" is consistent with any single direction
  - "complex": none of above satisfied

Segmentation

Color composition



Spatial texture

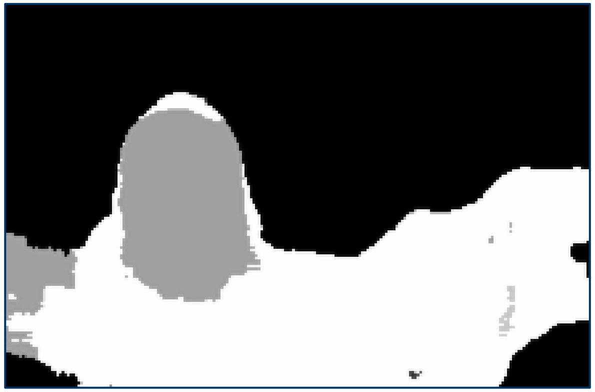


Crude segmentation

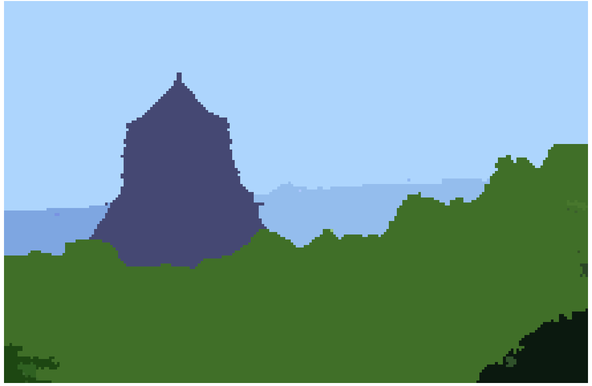
Final segmentation

Segmentation

Color composition



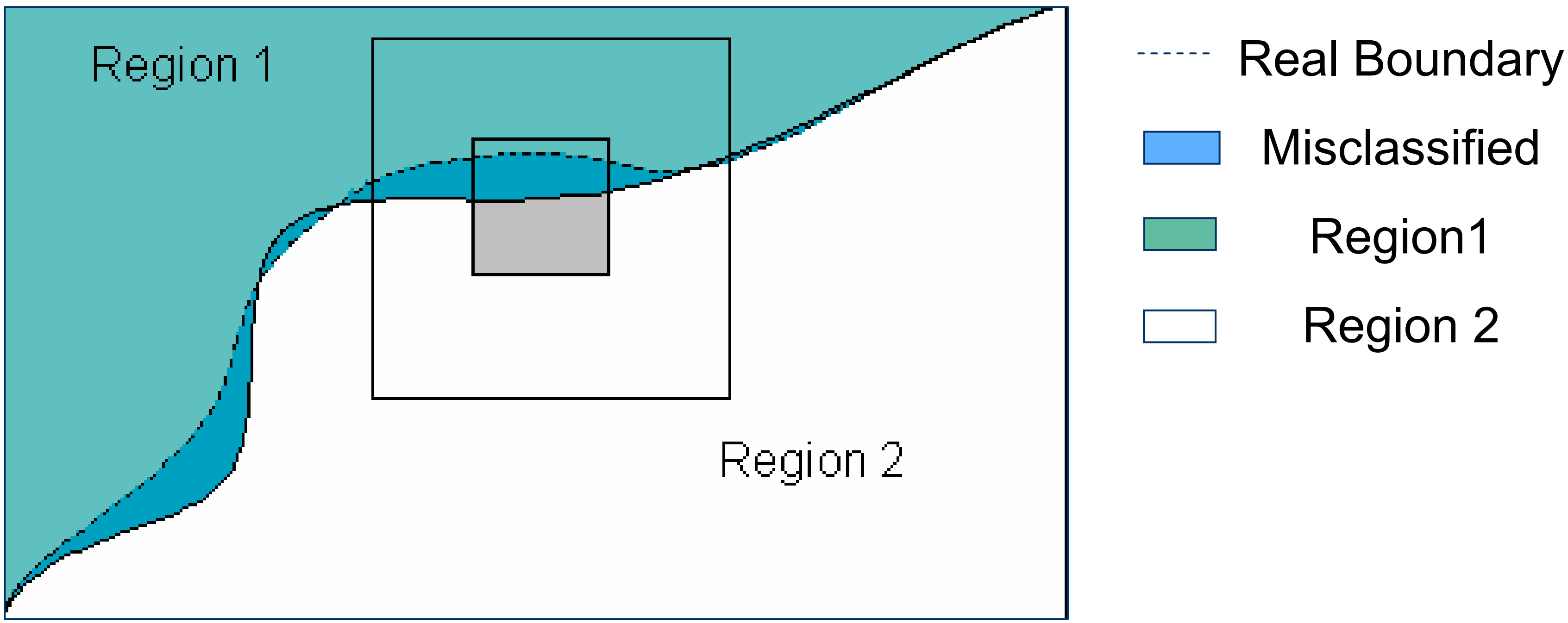
Spatial texture



Crude segmentation

Final segmentation

Iterative Border Refinement



Color features in inner window represent local features

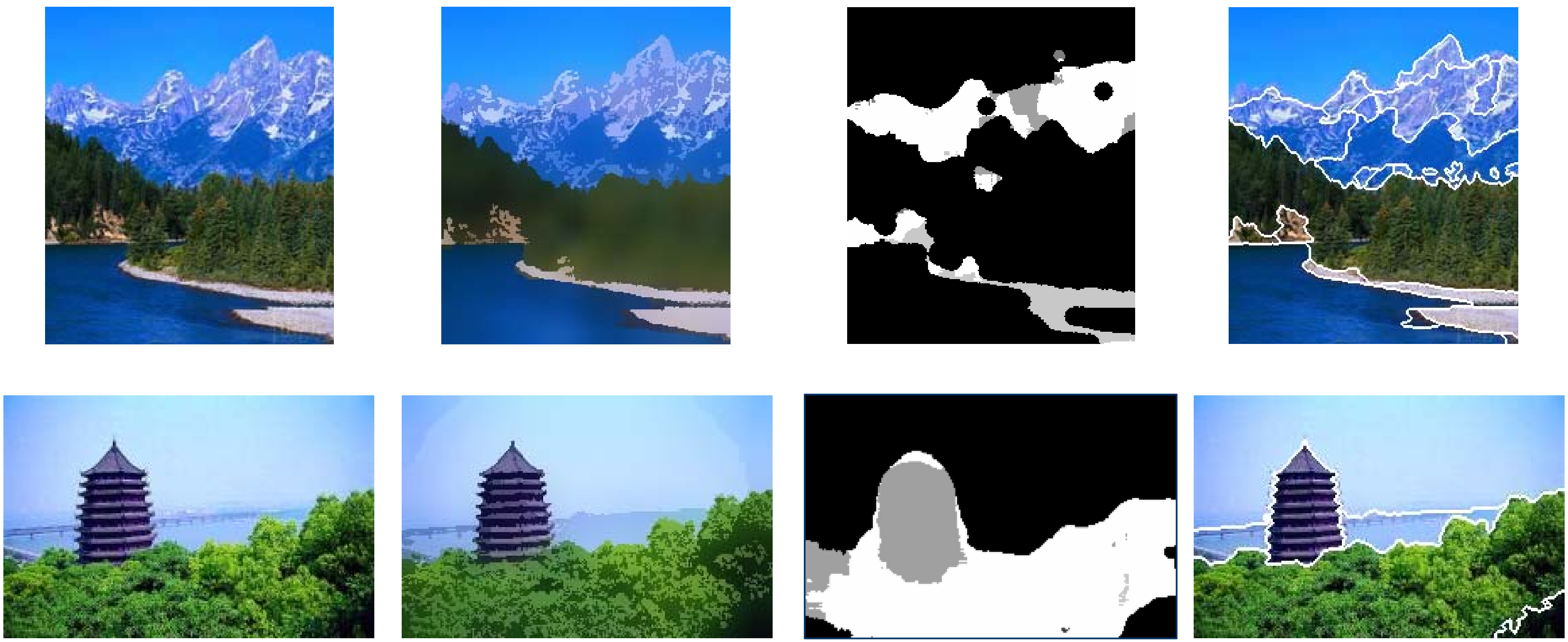
Color features in outer window represent region-wide characteristics

Window pairs used: {35/11, 21/9, 11/5, 11/3}

Results with steerable filters

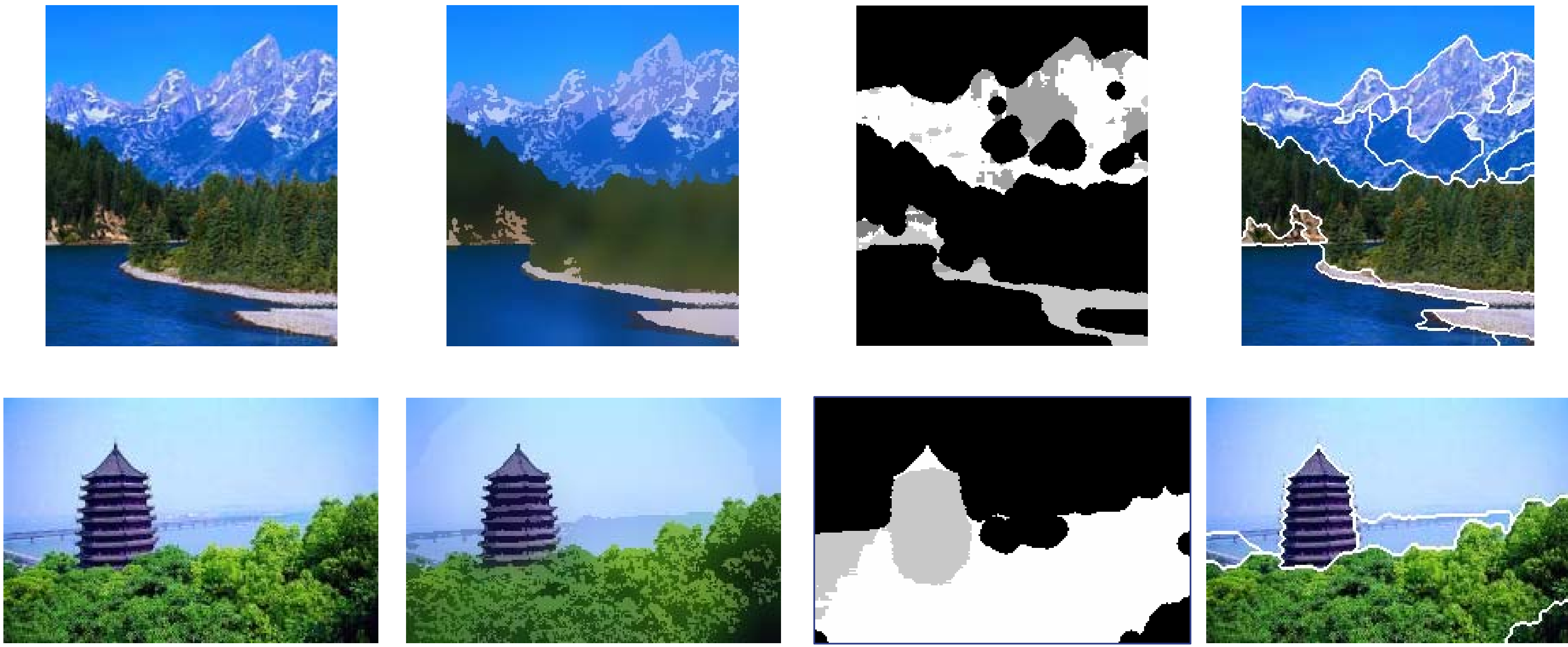
without Perceptual Tuning

Original ACA Texture Classes Segmentation



Results with steerable filters
with Perceptual Tuning

Original ACA Texture Classes Segmentation



Segmentation Results











Spatiotemporal Algorithm for Joint Video Segmentation and Foreground Detection

Background Subtraction

- Extracting moving (foreground) objects
- Building a background model
- Adaptation to changes in the scene
- Robustness
- Accuracy for applications like tracking

Video Segmentation

- Provides higher-level semantic representation compared to traditional pixel-based representation
  - Object-based Video coding (MPEG4)
  - Content extraction for indexing, retrieval (MPEG7)
- Goals
  - Complete object-based representation
  - Combination of video segmentation and foreground/ background separation

Important Issues in Background Subtraction

- Dynamic Background (sky, leaf, branch,light,specularity)
- Gradual Illumination Changes (Time of the day)
- Sudden Illumination Changes (Light switch, clouds)
- Sleeping person: Foreground object becomes completely still
- Waking person: Background object starts moving
- Shadows
- Bootstrapping (Initialization)

Approaches

Filtering
- MA
-Wallflower
- Kalman

Probabilistic Methods
- Parametric
- Nonparametric

Basic Methods

- Adjacent Frame Difference
- (Running) Average of Frames
- Wallflower
- Eigenbackgrounds
  - Images of motionless backgrounds
  - Principal Component Analysis
  - Difference between the projection and current frame is foreground
  - Exploits spatial correlation using covariance matrix

Unimodal: Pfinder

- Model the background pixel intensities by one Gaussian
- Update the Gaussian statistics with time
- Low complexity, low memory
- Good for unimodal backgrounds
  - Small lightning changes
  - Nearly stationary background

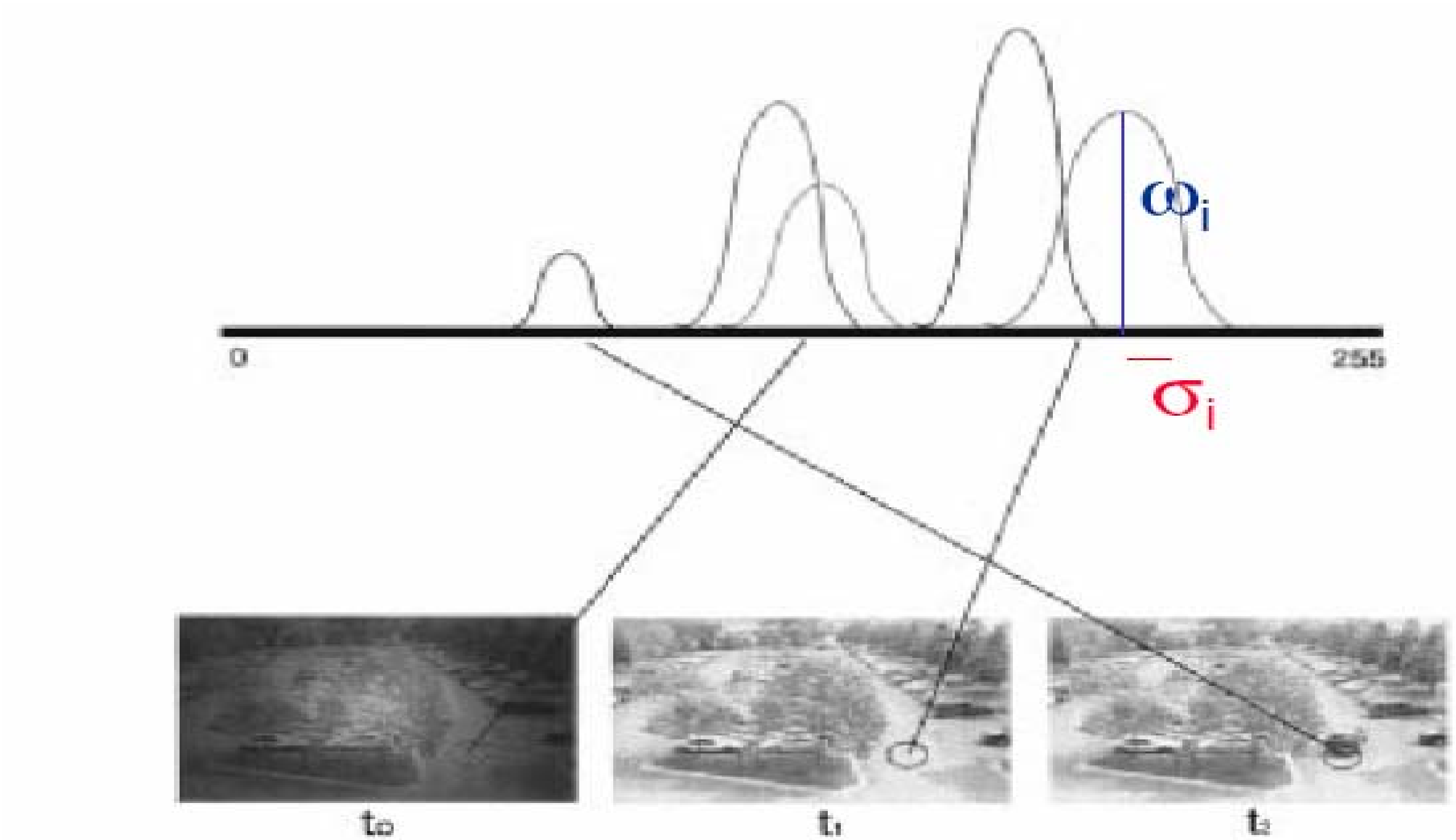
Mixture of Gaussians (MoG)

- Stauffer & Grimson 2000: Model the pixel intensity values by a mixture of Gaussians
- Complex time-varying multimodal backgrounds

$$P(y_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta\left(y_t, \mu_{i,t}, \Sigma_{i,t}\right)$$

- Adaptation – AR filtering with new data
- Relabeling of Gaussians

Mixture of Gaussians (MoG)



MoG: Relabeling of Gaussians

- Order distributions (ω/σ)
- Background / Foreground distribution decision

$$B = argmin_b \left(\sum_{k=1}^{b} \omega_k > T \right),$$

- T: measure of minimum portion of data accounted by background
- High T: multimodal background

MoG: Adaptation

- Every new pixel value is checked for a "match"
  - Start with the most likely distribution (highest ω/σ)
  - Pixel value within 2.5σ of a distribution
- Update and normalize weights
- Update match

$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha M_{k,t}$$

$$M_{k,t} = \begin{cases} 1 & \text{if match} \\ 0 & \text{if no match} \end{cases}$$

$$\mu_t = (1-\rho)\mu_{t-1} + \rho y_t$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho (y_t - \mu_t)^T (y_t - \mu_t)$$

$$\rho = \alpha \eta (y_t | \mu_k, \sigma_k)$$

Parametric Methods

- Advantages:
  - Fast
  - High adaptation to background changes
  - Fast initialization
- Disadvantages:
  - No spatial constraints (Post processing may be needed, especially in outdoor scenes)
  - Vulnerable to global changes in short-time

Kernel Density Estimation (KDE)

- Elgammal *et al.* '00

$$\hat{p}(y) = \sum_{i=1}^{N} \alpha_i K_\sigma (y - z_i)$$

- Does not assume specific shape for density
- Smoothed histogram: For high N, it converges to true density function
- Use Gaussian for K_σ
- Background pdf is estimated using N recent pixel values
- Adapt by adding new samples and dropping old ones

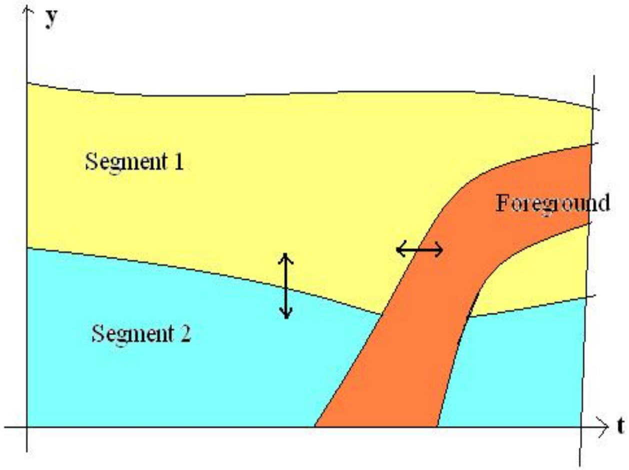
Nonparametric Methods

- Advantages:
  - Any probability distribution
  - Some spatial constraints
- Disadvantages:
  - High memory requirement
  - Slow
  - Initialization phase

Spatial Information

- Use spatial information to improve accuracy and robustness of foreground detection
  - Exploit spatial correlations
  - Spatiotemporal probabilistic model for pixel intensities
- Related prior work: 3-D ACA (adaptive clustering algorithm) [Hinds & Pappas'95]
  - Spatiotemporal MRF/GRF constraints
  - Spatiotemporally varying region intensities

Spatiotemporal Segmentation (3D-ACA)



- 3D-ACA can be used to detect foreground
  - New regions labeled as foreground
- Computationally expensive
- Temporally insensitive
  - Treats foreground/background boundary
    background boundaries
- Need more sensitivity for foreground segment detection
- More variation in spatial than temporal dimension
  - Still image vs. video coding
  - Inter vs. intra coding

Joint Spatiotemporal Segmentation and Background Subtraction

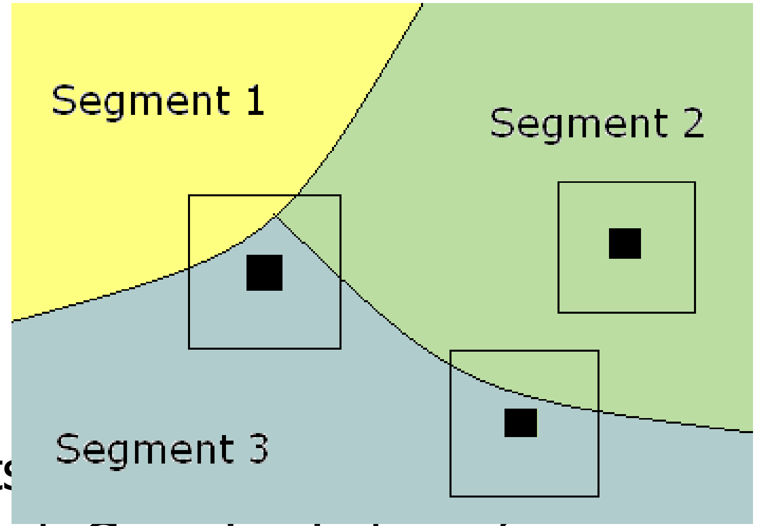
- Combine background subtraction with segmentation
  - Assume single stationary camera
  - Assume no foreground objects in the first few frames
- Initialize (first few frames) with 3D-ACA
- Use MRF constraints only in spatial dimension
  - Eliminate temporal MRF constraints for increased sensitivity
  - Spatial continuity
- Use spatiotemporal background model for background intensities
  - Spatiotemporally varying region intensities
  - Fidelity to data

Temporal Modeling

- Pixel distribution modeled by K spatiotemporal Gaussians

$$p(\mathbf{y}) \propto \sum_{x} p(\mathbf{y}|\mathbf{x}) \qquad P(y_{s,t}) = \frac{1}{K}\sum_{i=1}^{K} \eta\left(y_{s,t}; \mu_{i,s,t}, \Sigma_{i,s,t}\right)$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$



- All regi … (no weights …
- Compute local mean and variance for each Gaussian in base frame

Foreground Detection

- Pixel intensity compared with
  - K background distributions
  - (Any existing) Foreground distribution
- In case of no match, pixel is assigned to foreground
- Once new foreground object is encountered, build new foreground distribution (single Gaussian)
  - Single Gaussian is sufficient in case of small lightning changes and small texture difference
- Calculate local mean and variance (spatiotemporally, as for background regions)

Adaptation

- After labeling, compute the local statistics

$$\hat{\mu}_{i,s,t} = \sum_{x_{i,s,t}=i} y_{s,t}$$

- Apply a low-pass filter with exponential weighting

$$\mu_{i,s,t} = (1-\alpha)\mu_{i,s,t-1} + \alpha\hat{\mu}_{i,s,t}$$

$$\Sigma_{i,s,t} = (1-\alpha)\Sigma_{i,s,t-1} + \alpha(\hat{\mu}_{i,s,t} - \mu_{i,s,t})^T(\hat{\mu}_{i,s,t} - \mu_{i,s,t})$$

Properties

- Insensitive to learning parameters
  - Spatially smoothed data instead of raw
- Increased sensitivity over 3D-ACA
- High accuracy
- Medium Complexity (Real-time)
- Spatial MRF constraints necessary for stability

Video Segmentation

- Spatial MRF necessary for preserving continuity in background regions

Example A: Algorithm

Base Frame

Original

Labeling

Result

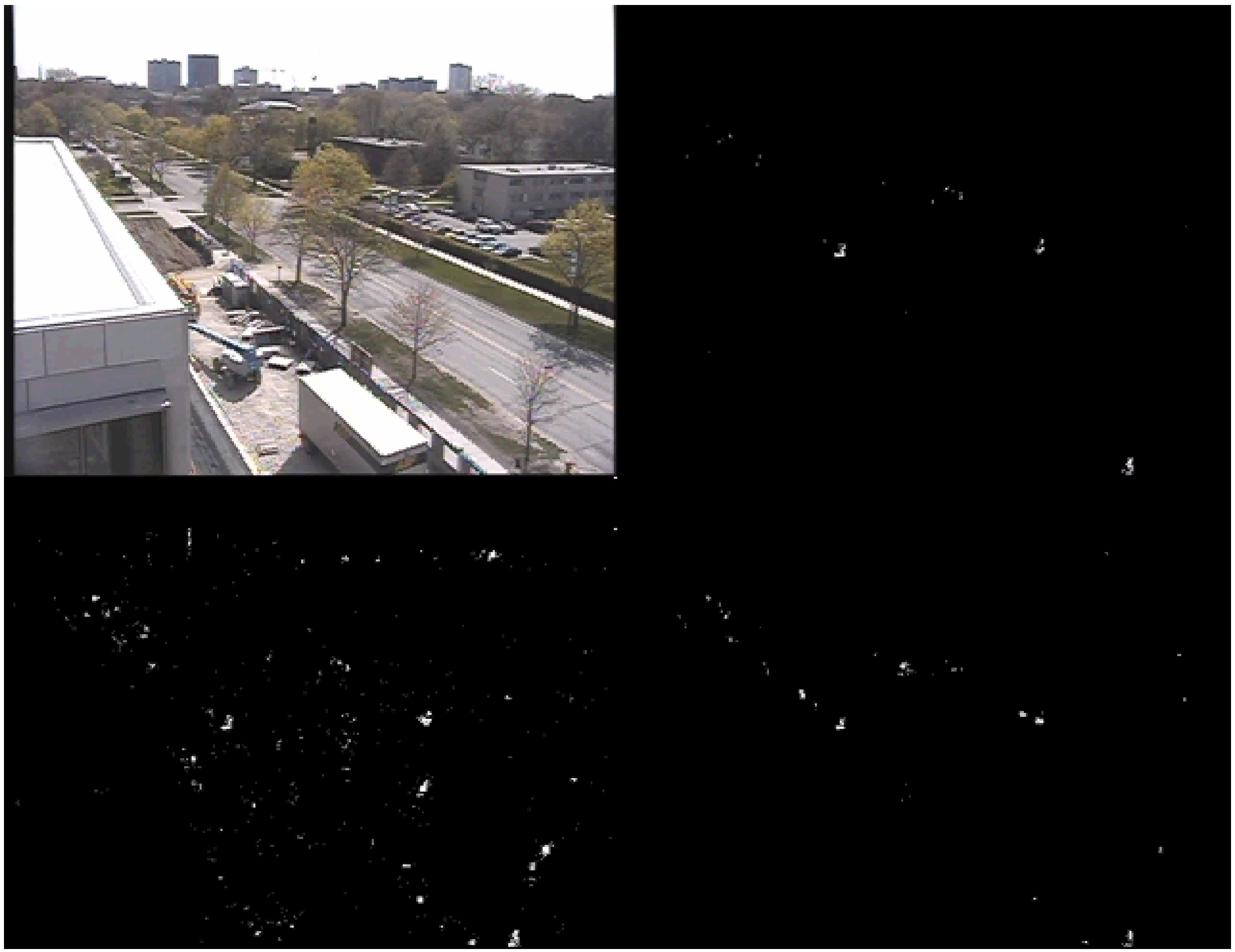
Example B: Hall Monitor

Original KDE

MoG Proposed

Example C: Ford Webcam

Original KDE



MoG Proposed

Example A: Hall Monitor

Original Sequence

Segmentation

Foreground Detection

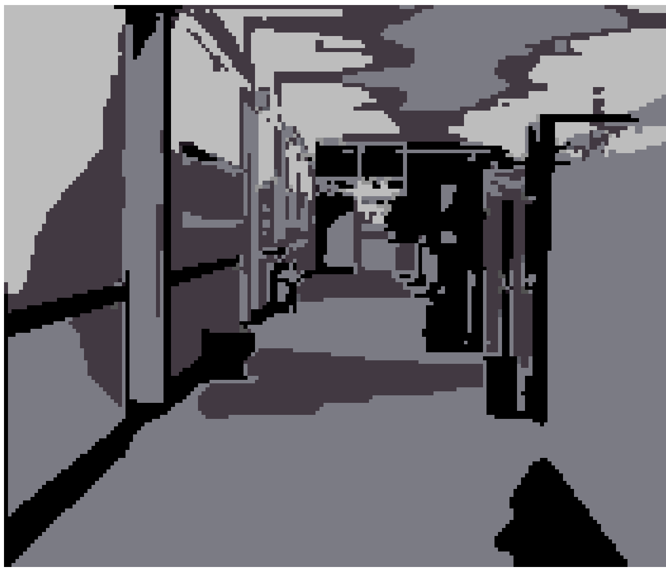
Example B: Ford Webcam

Original Sequence

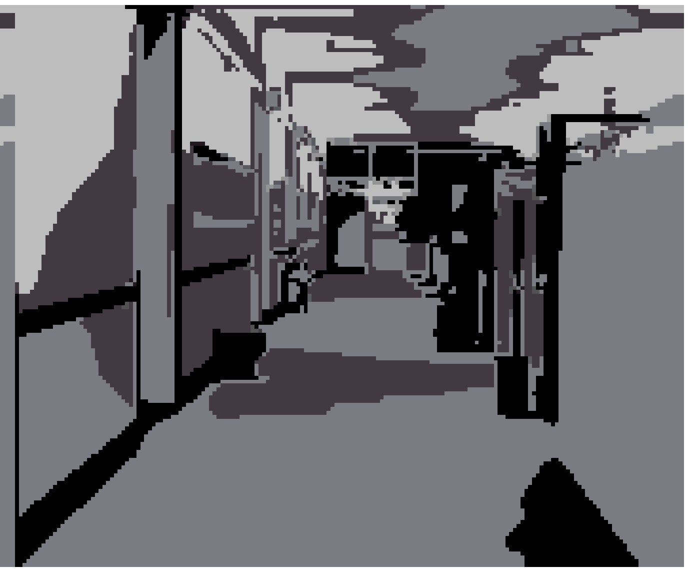
Segmentation

Foreground Detection

Example C:
Proposed vs 3D-ACA



Proposed



3D-ACA

MoG vs. KDE vs. Proposed

Low complexity	High complexity	Medium complexity
Low memory	High memory	Low memory
Very sensitive learning	Insensitive learning	Insensitive learning
Adaptation rate ?	Fast adaptation	Fast adaptation
Short initialization	Very long initialization	Short initialization
Low selectivity	High selectivity	High selectivity
High noise	Low noise	Low noise

Semantic Information Extraction

- Motivation
  - Proliferation of image and video acquisition devices (digital still and video cameras, image and video phones, PDAs)
  - World rich in digital visual content
  - Large personal repositories (consumer market)
  - Increasing processing capabilities
- Goal: Intelligent content management
  - Semantic labeling
  - Content organization
  - Efficient retrieval

Challenges

- What are the important semantic categories?
- How to link the low-level features to semantically important categories?

Semantic Categories

- Recent perceptual experiments by Mojsilovic and Rogowitz identified important semantic categories that humans use for image classification

- Conjecture: Semantic categories can be derived from combinations of low-level image features

Bridging the Semantic Gap

High level — Semantics

Medium level — Perceptually Uniform Segments

Low level — Primitives

Use segment descriptors and statistical techniques to relate segments (first) and scenes (later) to semantic categories/labels

Incorporate knowledge of human perception and image characteristics into feature extraction and algorithm design

Semantic Information Extraction
(at Segment level)

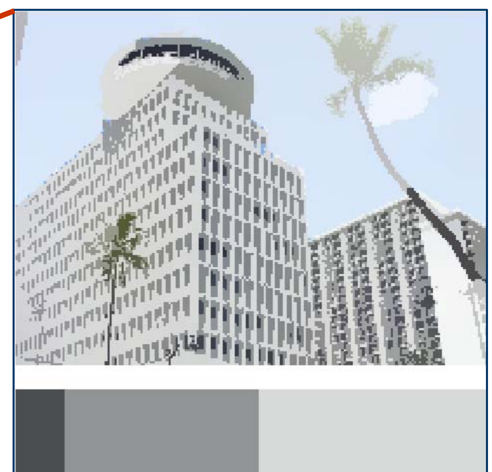
original

Dominant Colors (ACA)

segment 1

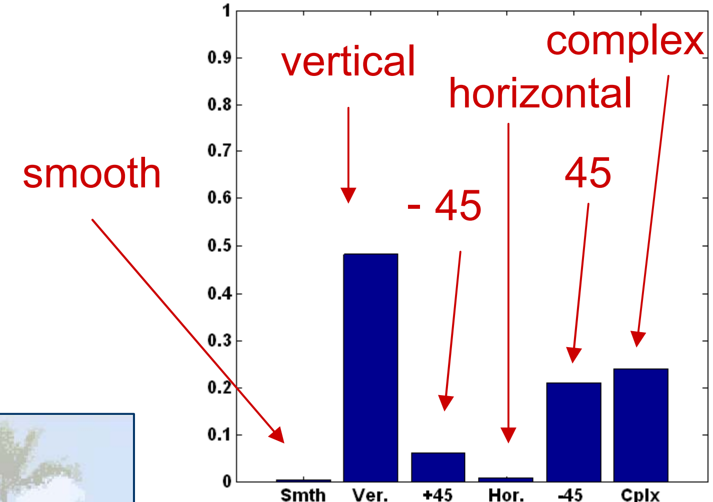
segment 2

segment 3



Dominant Colors & Percentages

quantize



Spatial Texture

Plus: Location
Shape
Size

Color Naming Syntax

Hue primary	Hue secondary	Lightness	Saturation	Achromatic
red orange brown yellow green blue purple pink beige magenta olive	reddish brownish yellowish greenish bluish purplish pinkish	grayish moderate medium strong vivid	blackish very-dark dark medium light very-light whitish	black gray white
	267 quantization points (NBS, Mojsilovic'02)			

Eleven Colors That Are Almost Never Confused (Boynton'89)

Labels

(consistent with NIST TRECVID 2003 development set)

Segment

Man Made

- Building
- Bridge
- Cityscape
- Car
- Boat
- Airplane
- Pavement
- Other Man Made

Natural

Vegetation

- Flower
- Grass
- Woods/Bushes
- Forest

Sky

- Day-sky
- Night-sky
- Sun
- Clouds
- Sunrise/Sunset

Landform

- Snow
- Mountain
- Ground

Water

People

- Face
- Person
- Crowd

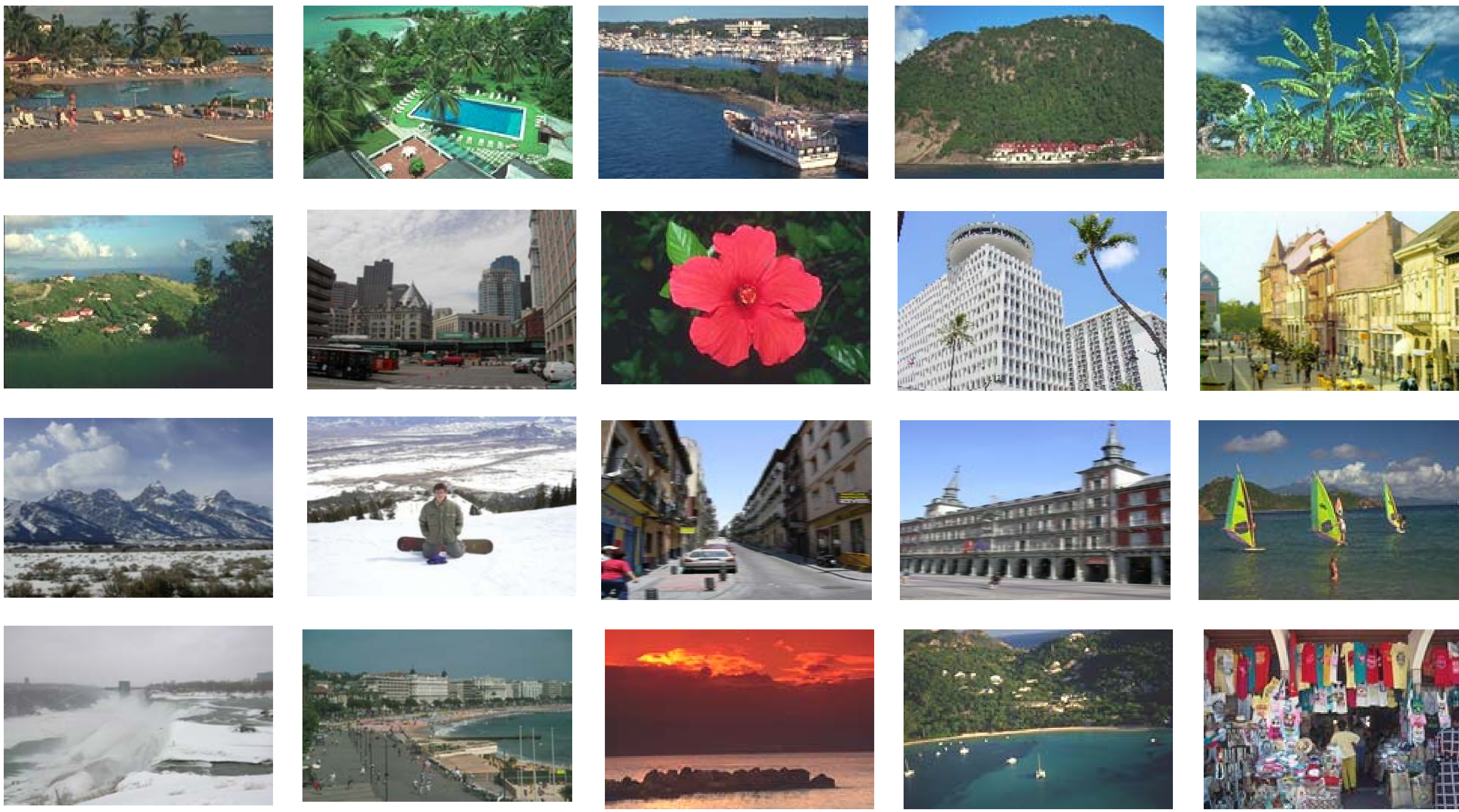
[Animal]

Scene

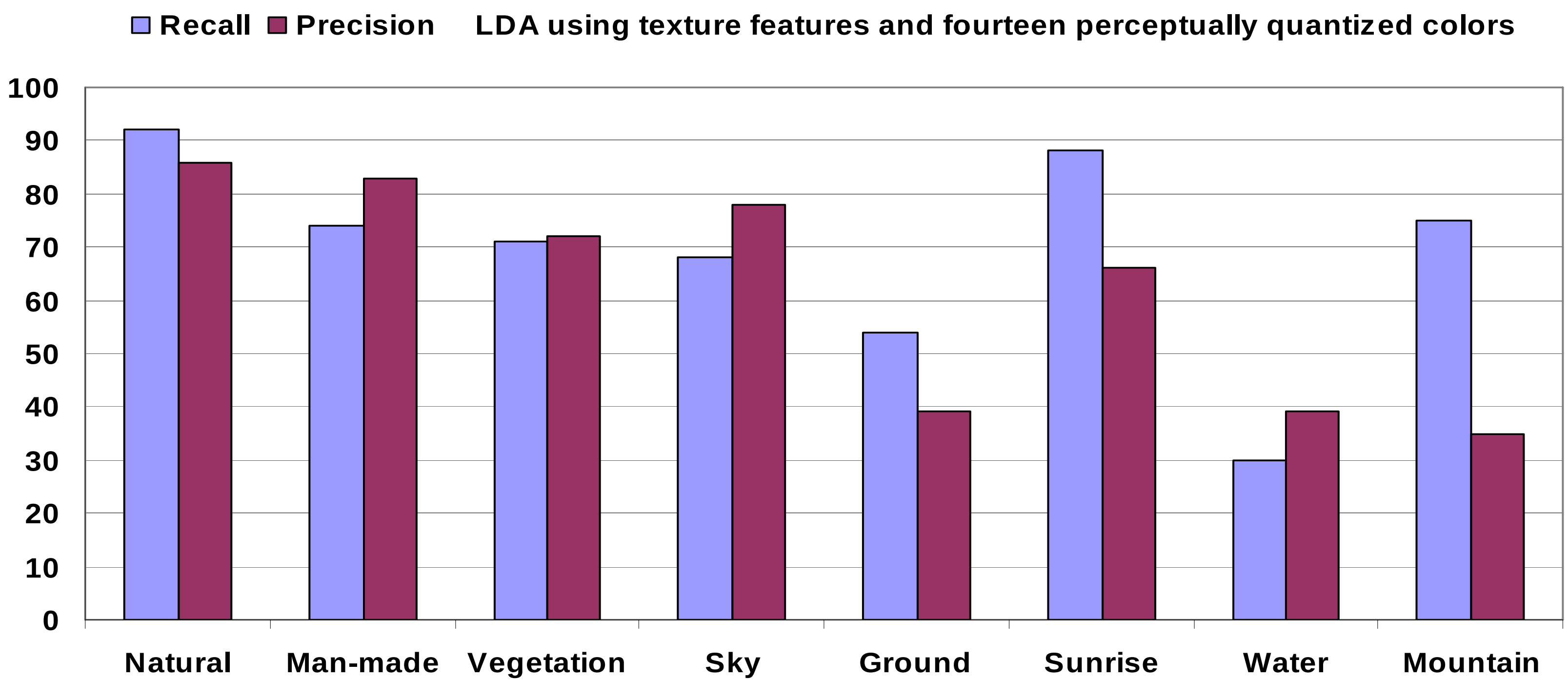
Indoor **Outdoor:** Street, skyline, beach, garden, night scene, day scene …

Database (Training, Testing)

9000 Labeled segments
2500 Images (Corel Stock Photo, Berkeley, other)



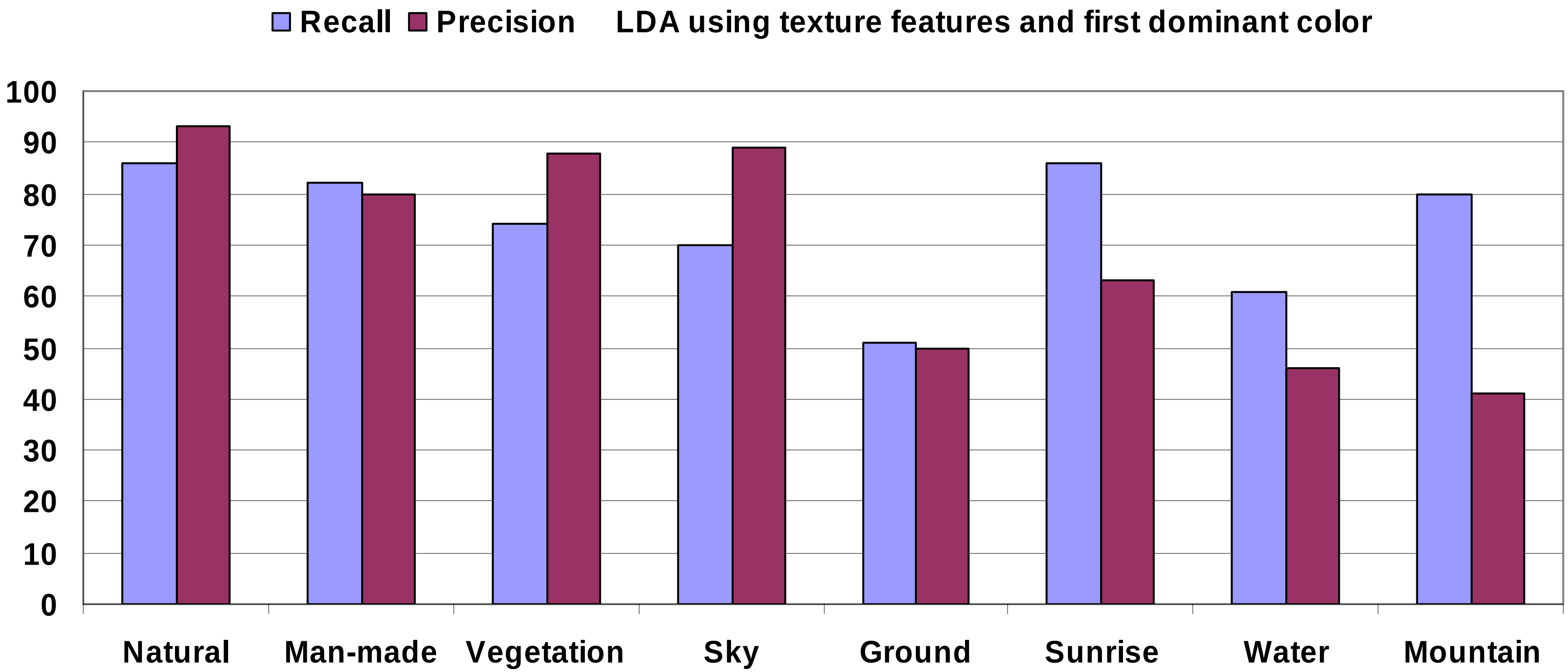
Results



$$\text{Recall} = \frac{\text{number of correctly classified segments}}{\text{total number of relevant segments}}$$

$$\text{Precision} = \frac{\text{number of correctly classified segments}}{\text{total number assigned to a label}}$$

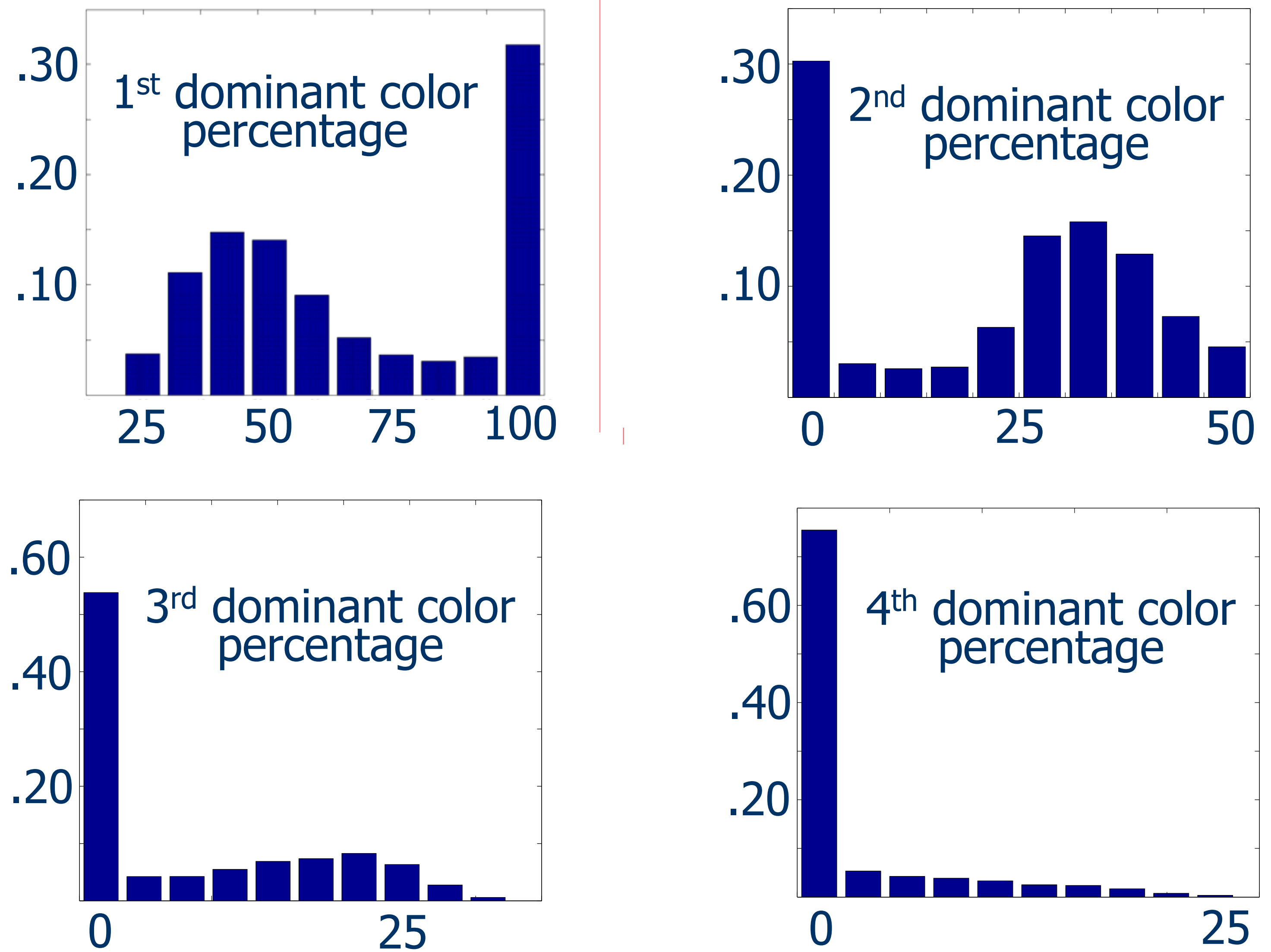
Results



$$\text{Recall} = \frac{\text{number of correctly classified segments}}{\text{total number of relevant segments}}$$

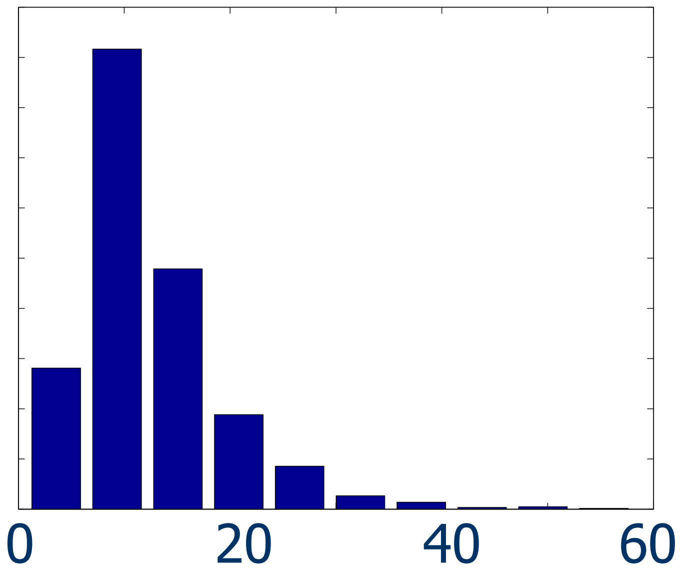
$$\text{Precision} = \frac{\text{number of correctly classified segments}}{\text{total number assigned to a label}}$$

Statistics of Dominant Colors



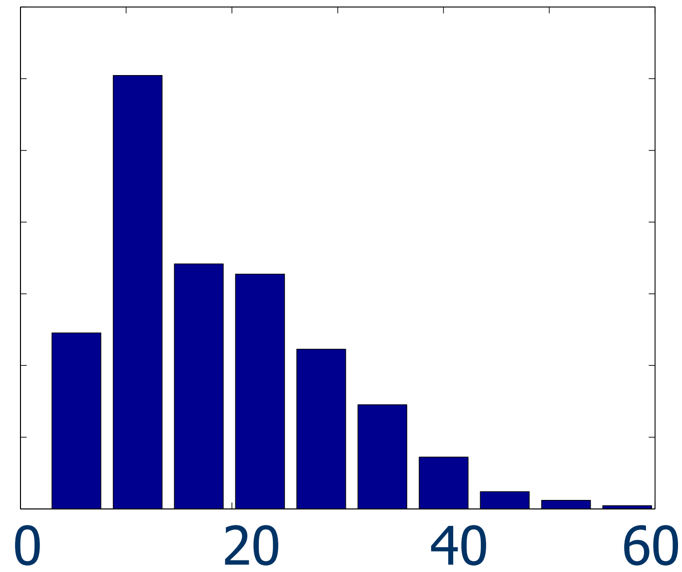
Statistics of Dominant Colors

Distance between 1st and 2nd dominant color



L*a*b distance

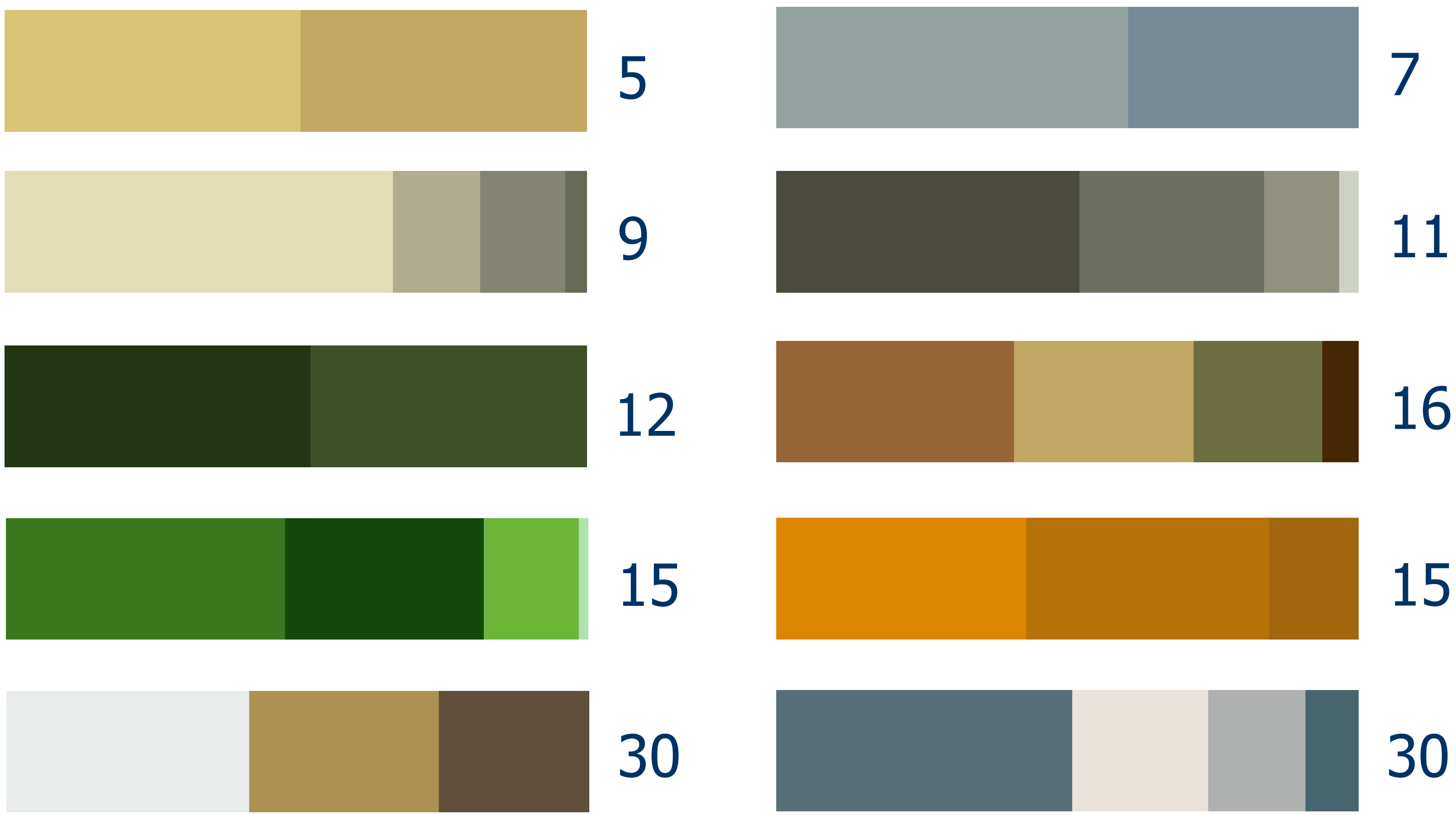
Distance between 1st and 3rd dominant color



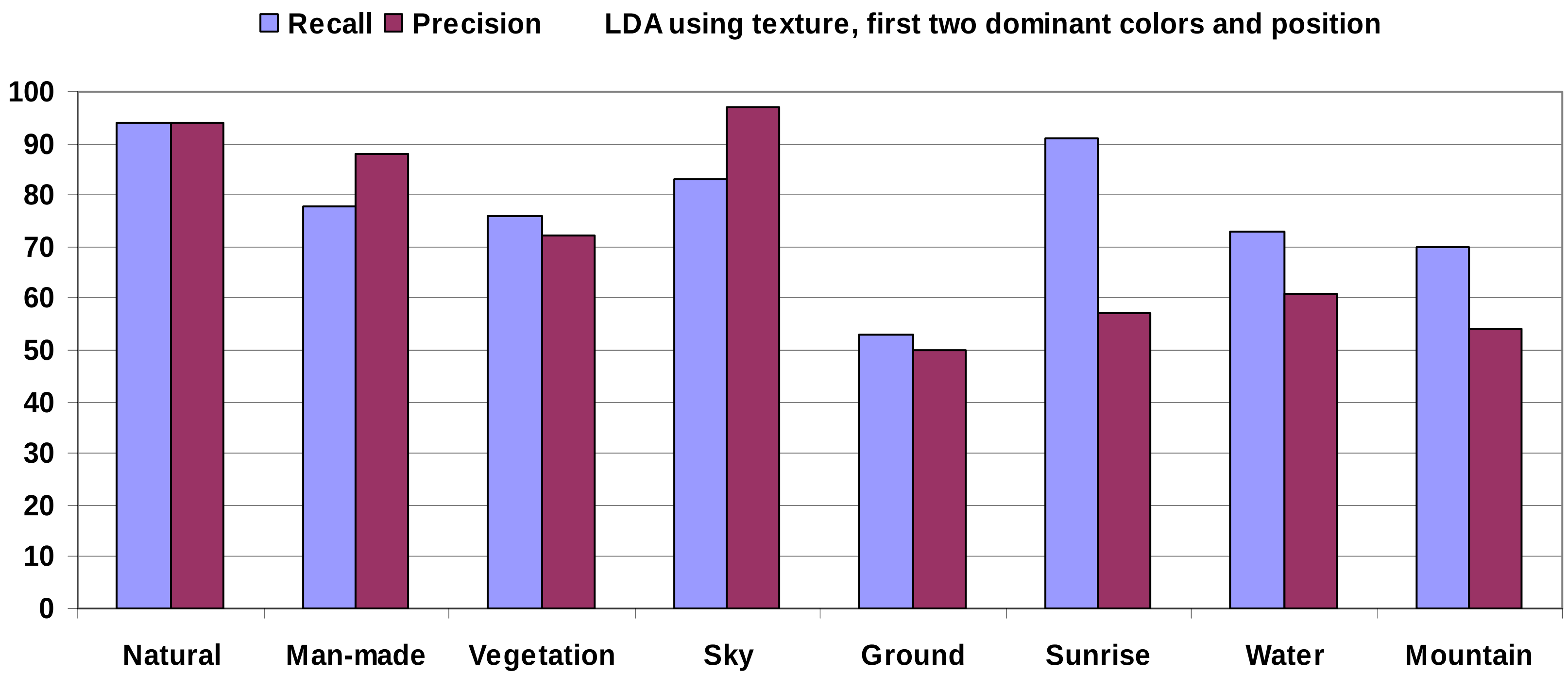
L*a*b distance

Statistics of Dominant Colors

L*a*b* distances between first and second dominant color:



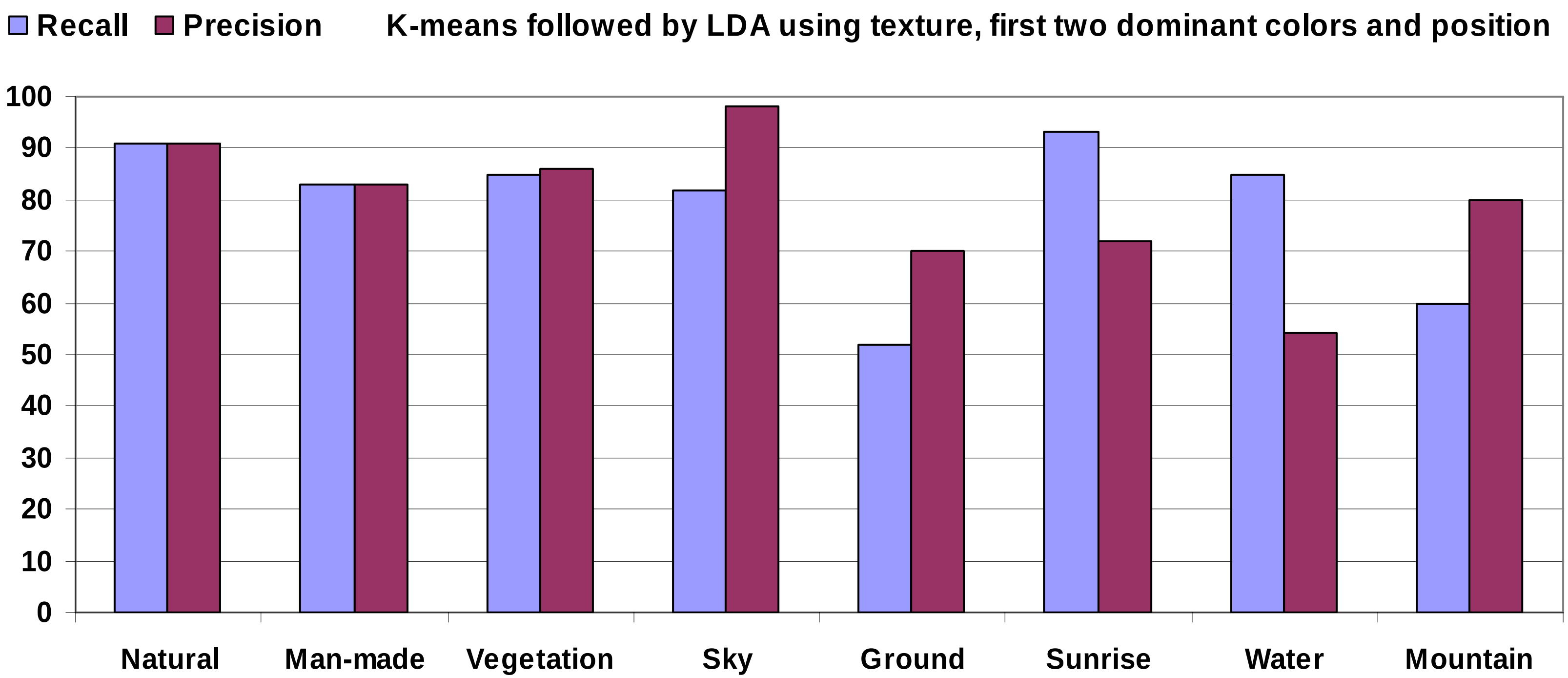
Results



$$\text{Recall} = \frac{\text{number of correctly classified segments}}{\text{total number of relevant segments}}$$

$$\text{Precision} = \frac{\text{number of correctly classified segments}}{\text{total number assigned to a label}}$$

Results



$$\text{Recall} = \frac{\text{number of correctly classified segments}}{\text{total number of relevant segments}}$$

$$\text{Precision} = \frac{\text{number of correctly classified segments}}{\text{total number assigned to a label}}$$

Publications

- J. Chen and T. N. Pappas, “Experimental determination of visual color and texture statistics for image segmentation,” *Human Vision and Electronic Imaging X,* Proc. SPIE Vol. 5666, pp. 227 - 236, Jan. 2005.
- J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz, “Adaptive perceptual color-texture image segmentation,” *IEEE Trans. Image Processing*, vol. 14, pp. 1524--1536, Oct. 2005.
- T.N. Pappas, J. Chen, and D. Depalov, “Learning perception,” *OE Magazine,* vol. 5, pp. 18 - 20, Oct. 2005.
- D. Depalov, T. N. Pappas, D. Li, and B. Gandhi, “Perceptually based techniques for semantic image classification and retrieval,” *Human Vision and Electronic Imaging XI*, Proc. SPIE Vol. 6057, (San Jose, CA), Jan. 2006.
- D. Depalov, T. N. Pappas, D. Li, and B. Gandhi, “A perceptual approach for semantic image retrieval,” *Proc. ICASSP-06,* (Toulouse, France), May 2006. To appear.
- D. Depalov, T. N. Pappas, D. Li, and B. Gandhi, “Perceptual feature selection for semantic image classification,” Proc. Int. Conf. Image Processing (ICIP-06), (Atlanta, GA), Oct. 2006. Submitted.