Object Class Recognition using Images of Abstract Regions

Yi Li, Jeff A. Bilmes, and Linda G. Shapiro

Department of Computer Science and Engineering

Department of Electrical Engineering

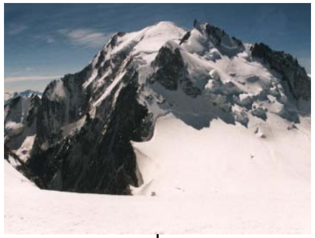
University of Washington

Problem Statement

Given: Some images and their corresponding descriptions

{trees, grass, cherry trees}

{cheetah, trunk}



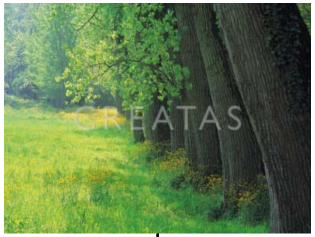
{mountains, sky}

{beach, sky, trees, water}

…

To solve: What object classes are present in new images

?



?

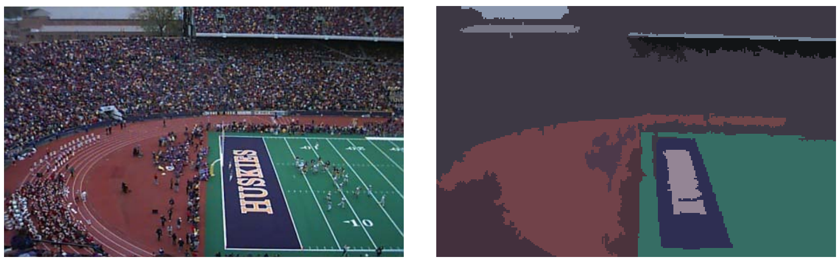
?

?

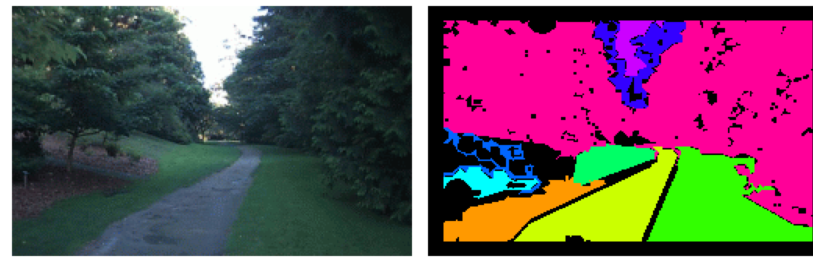
…

Image Features for Object Recognition

- Color

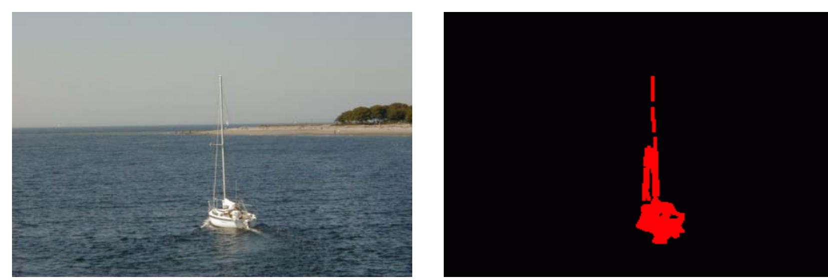


- Texture



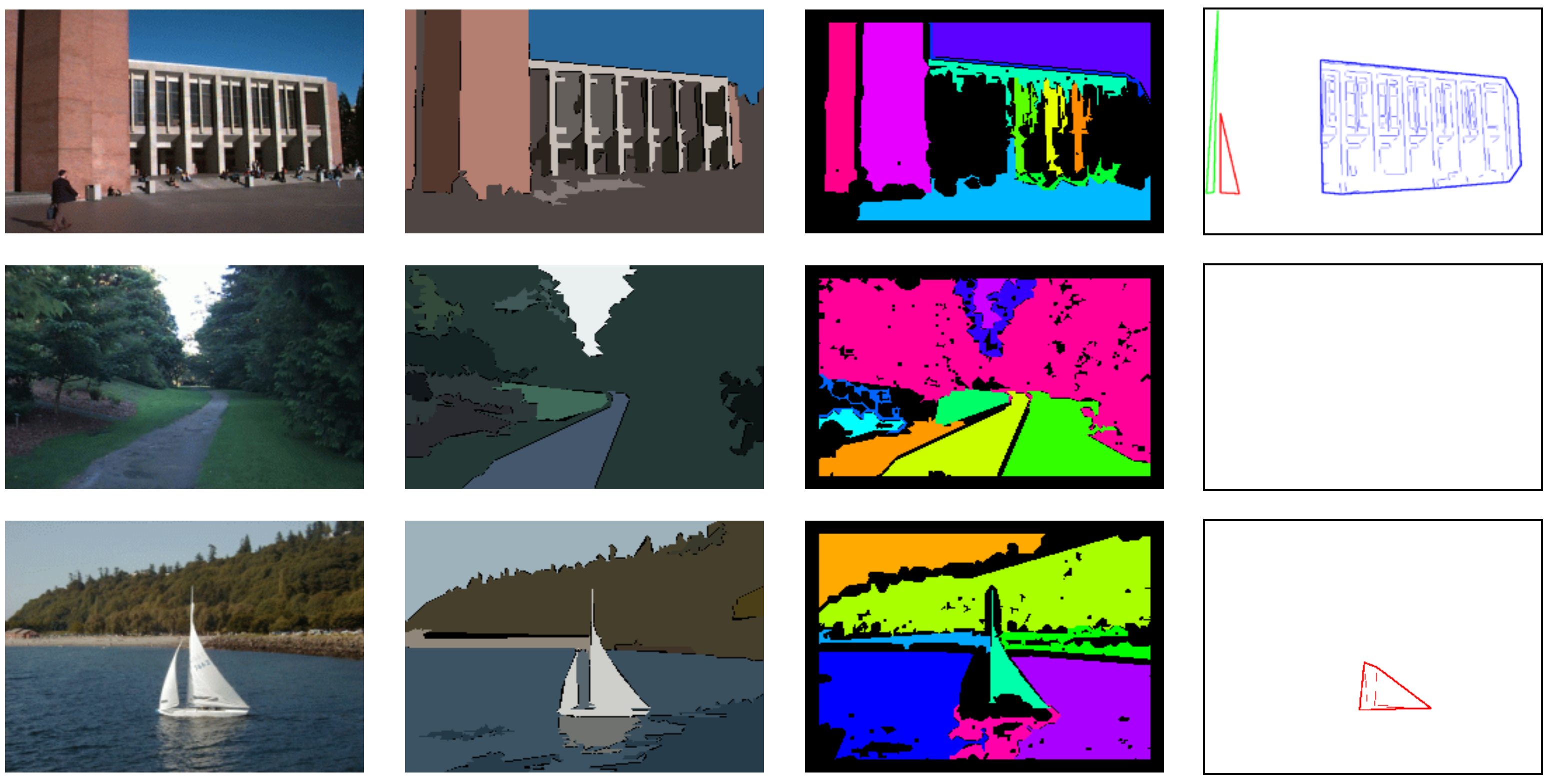
- Structure

- Context

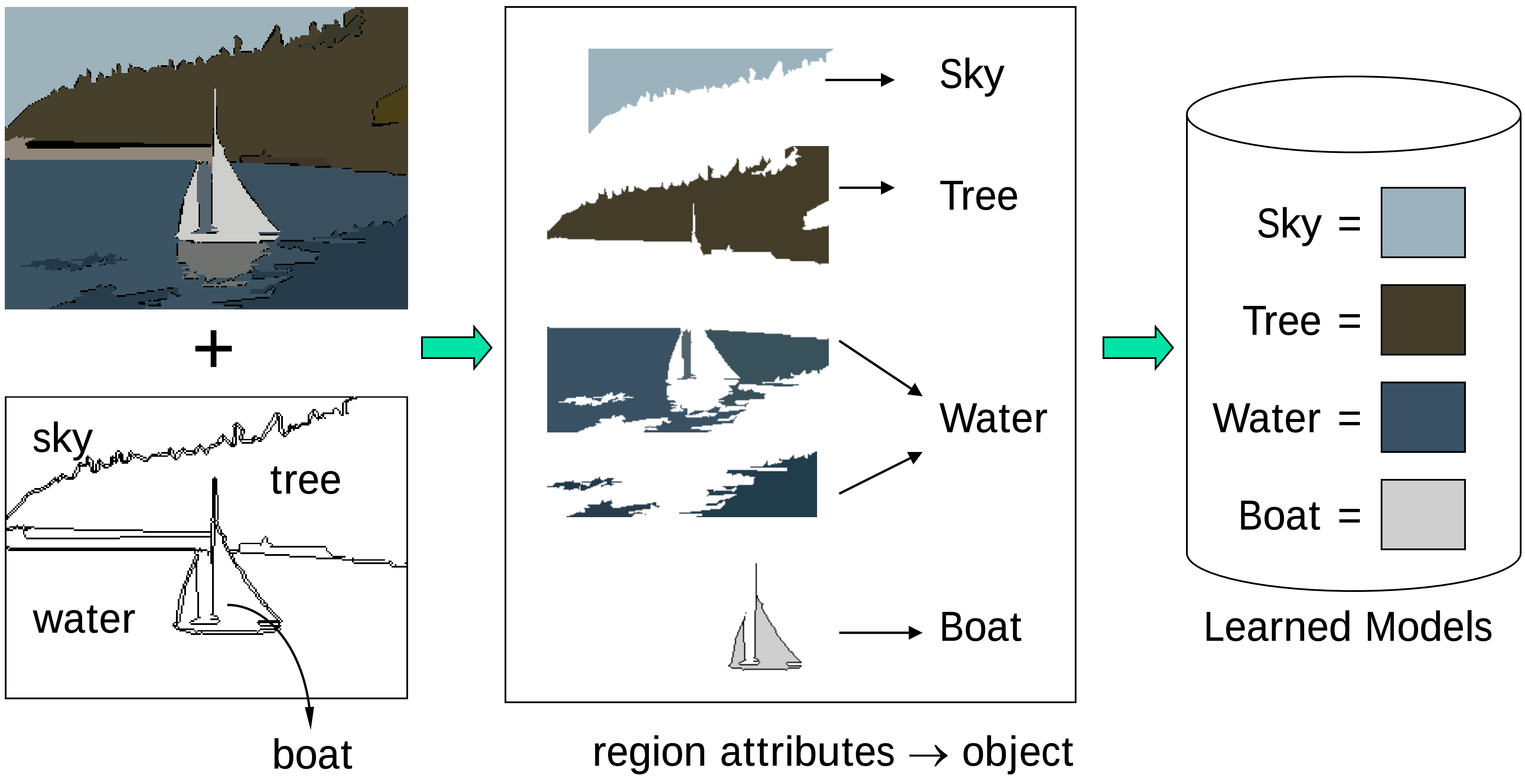


Abstract Regions

Original Images | Color Regions | Texture Regions | Line Clusters



Object Model Learning (Ideal)



Our Scenario: Abstract Regions

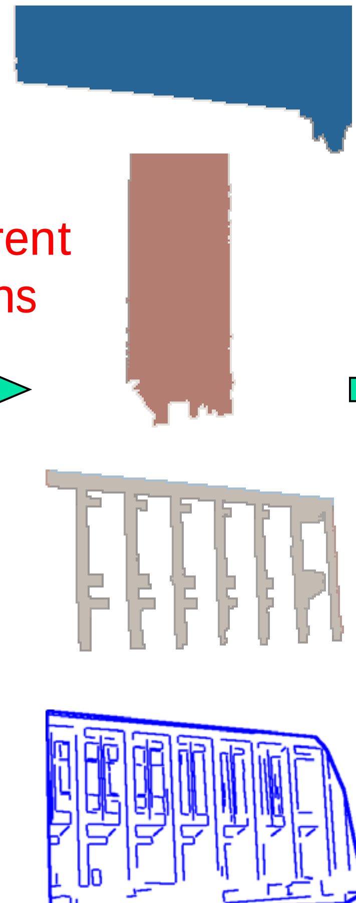
Multiple segmentations whose regions are not labeled; a list of labels is provided for each training image.

image

labels

{sky, building}

various different segmentations



region attributes from several different types of regions

Object Model Learning

Assumptions

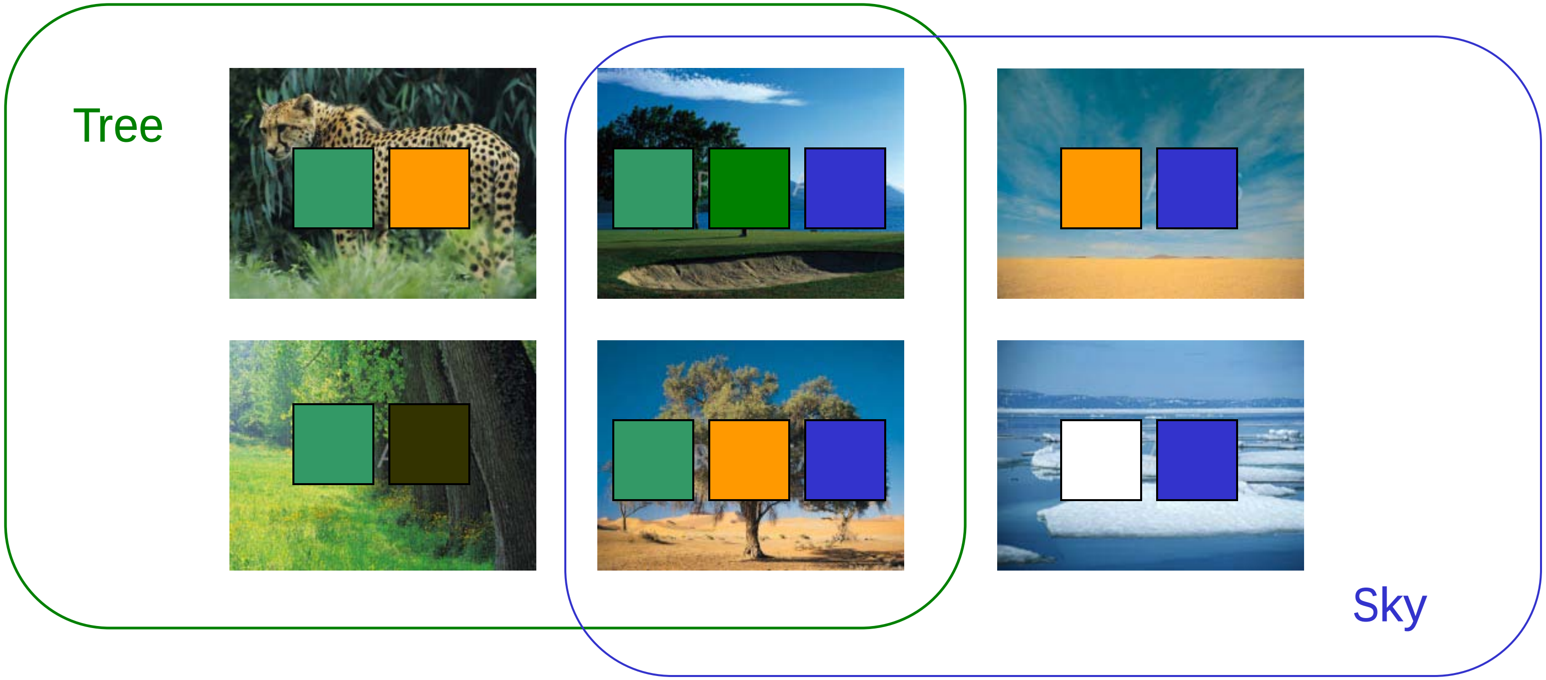
- The feature distribution of each object within a region is a Gaussian;

- Each image is a set of regions;

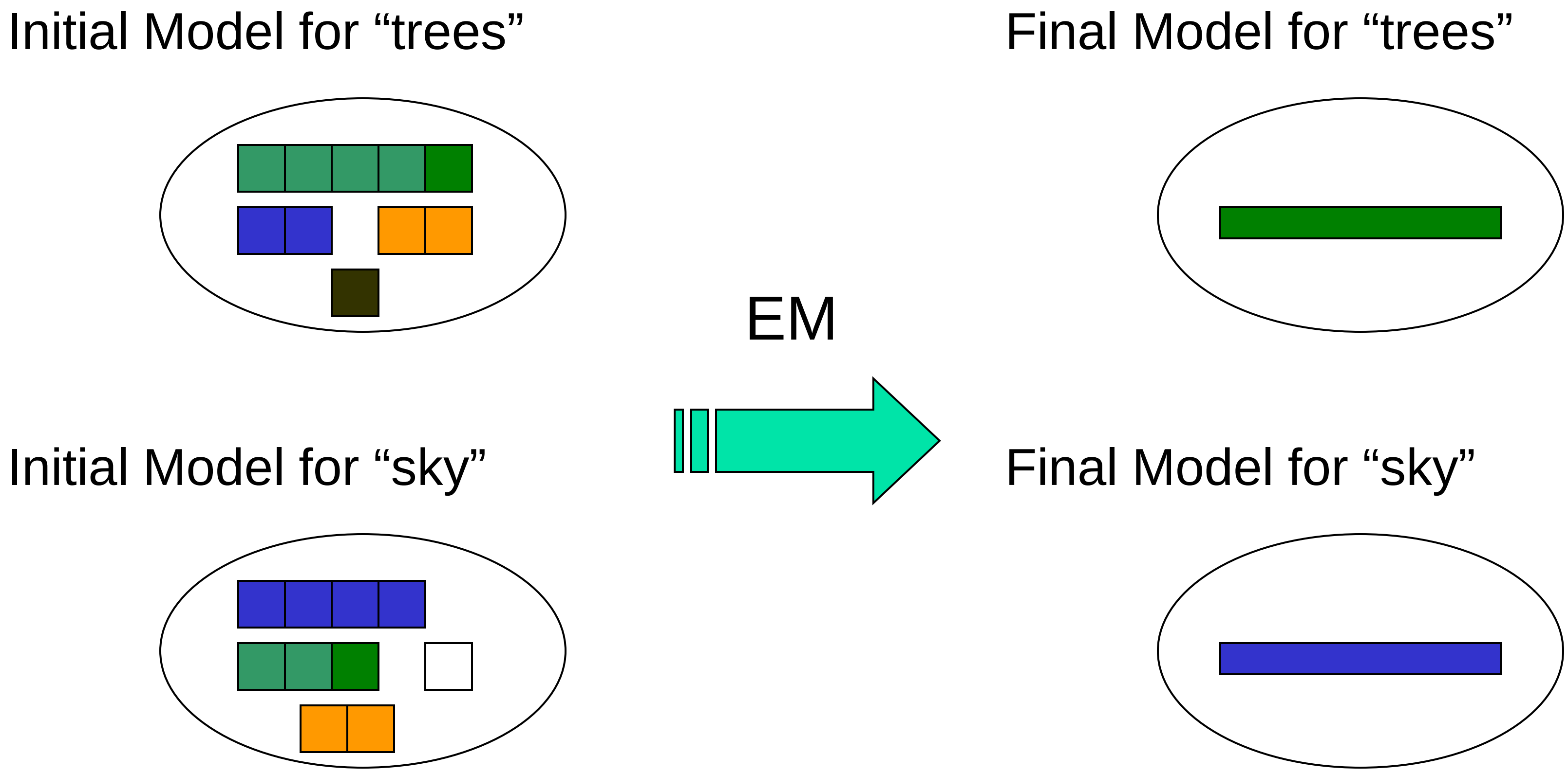
  each region can be modeled as a mixture of multivariate Gaussian distributions.

Model Initial Estimation

- Estimate the initial model of an object using all the region features from all images that contain the object



EM Variant


Initial Model for “trees”
Final Model for “trees”
EM
Initial Model for “sky”
Final Model for “sky”

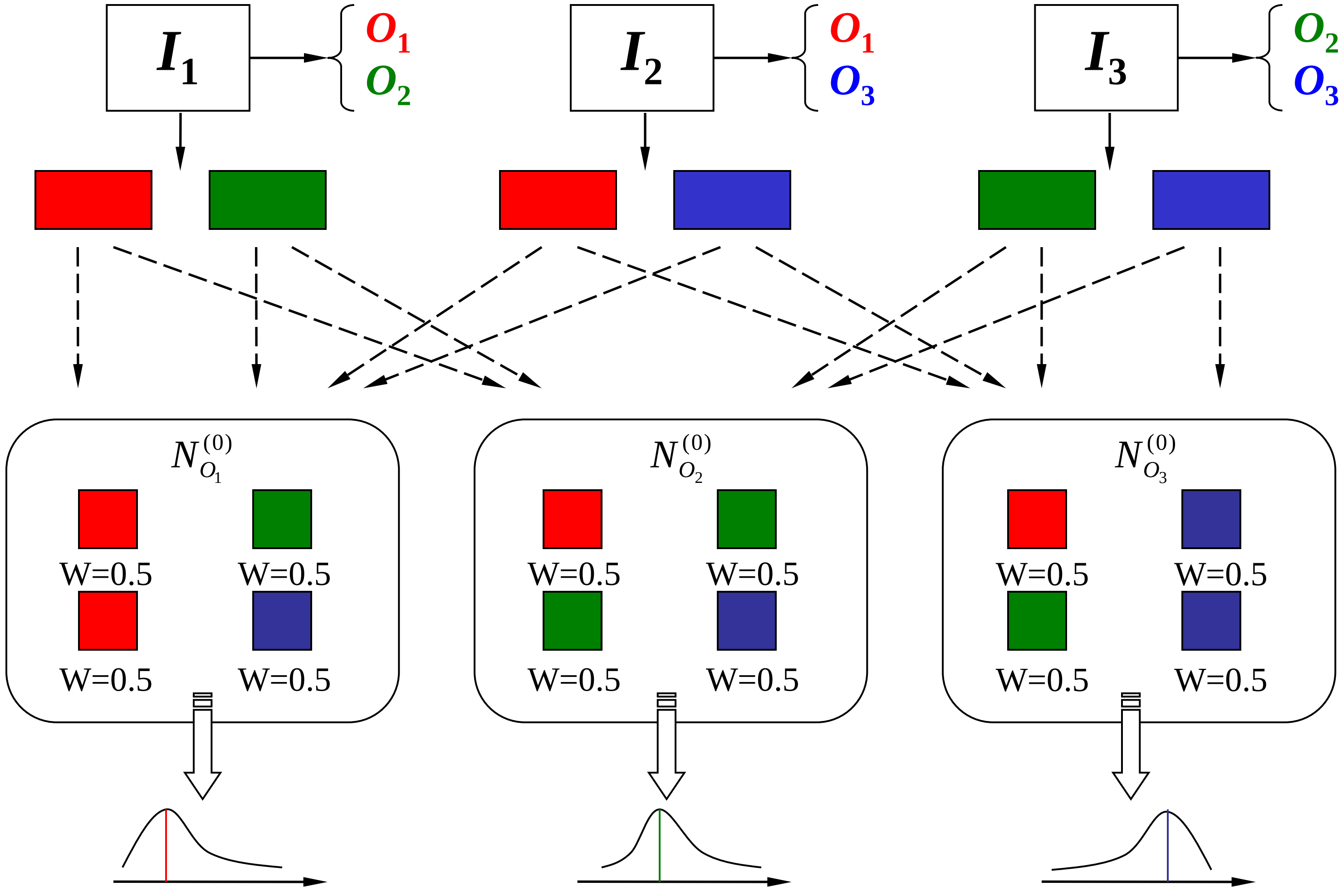
EM Variant

- Fixed Gaussian components (one Gaussian per object class) and fixed weights corresponding to the frequencies of the corresponding objects in the training data.
- Customized initialization uses only the training images that contain a particular object class to initialize its Gaussian.
- Controlled expectation step ensures that a feature vector only contributes to the Gaussian components representing objects present in its training image.
- Extra background component absorbs noise.

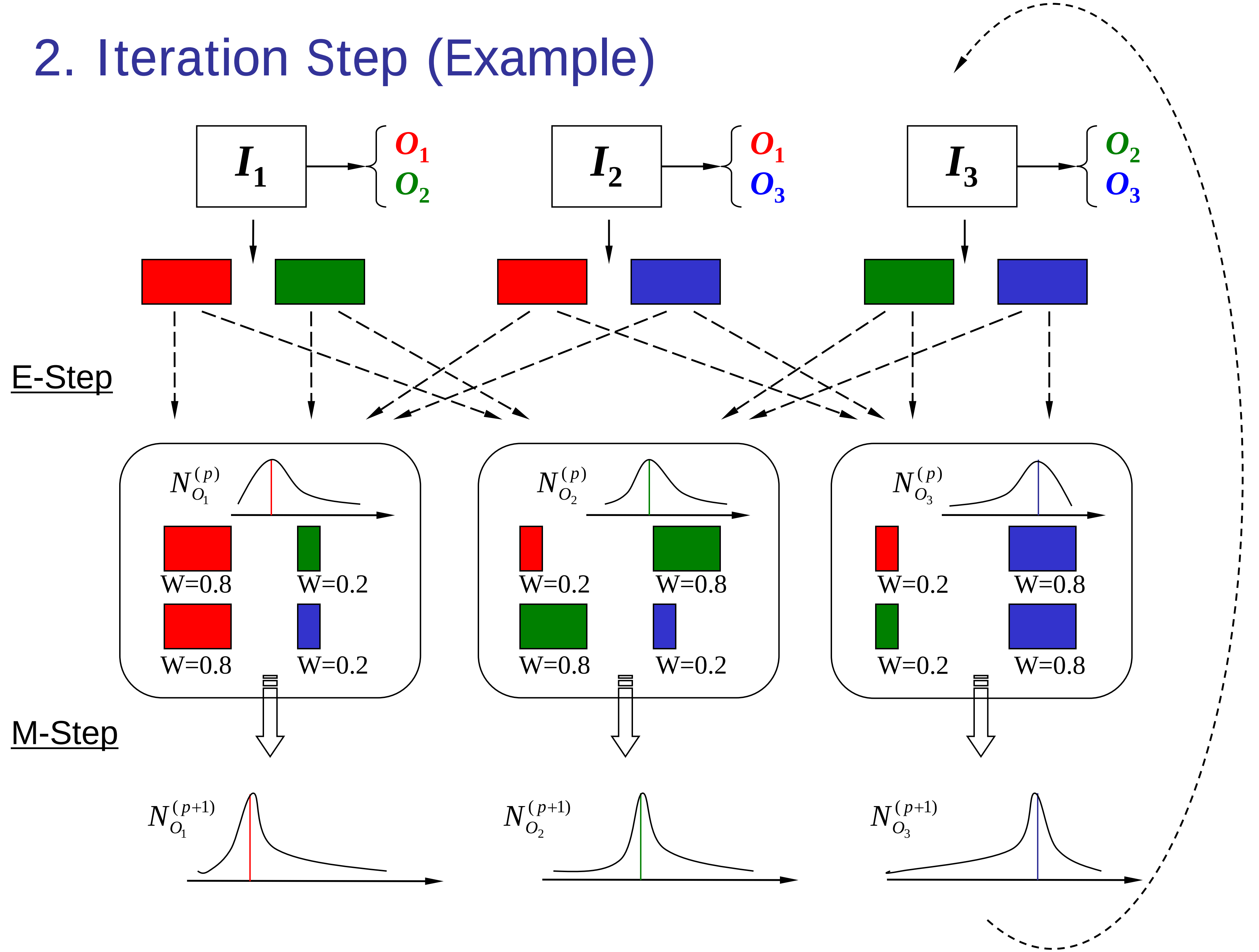
Gaussian for trees	Gaussian for buildings	Gaussian for sky	Gaussian for background

1. Initialization Step (Example)

Image & description



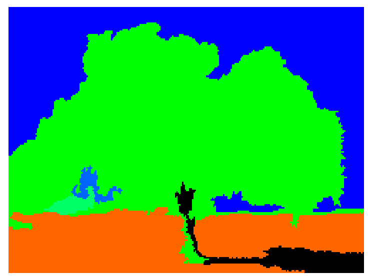
2. Iteration Step (Example)



Recognition

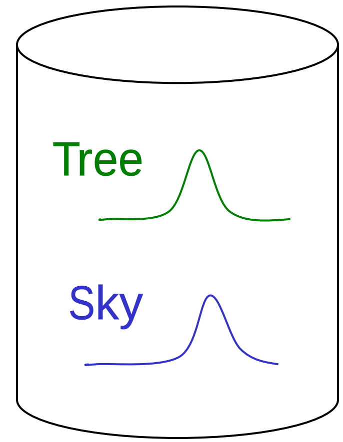
Test Image

Color Regions



compare

Object Model Database



To calculate $p(tree \mid image)$

$p(tree \mid image) = f$
$p(\, tree \mid$  $)$
$p(\, tree \mid$  $)$
$p(\, tree \mid$  $)$
$p(\, tree \mid$  $)$

$$p(o \mid F_I^a) = \underset{r^a \in F_I^a}{f} \left(p(o \mid r^a)\right)$$

f is a function that combines probabilities from all the color regions in the image.

What could it be?

Combining different abstract regions

- Treat the different types of regions independently and combine at the time of classification.

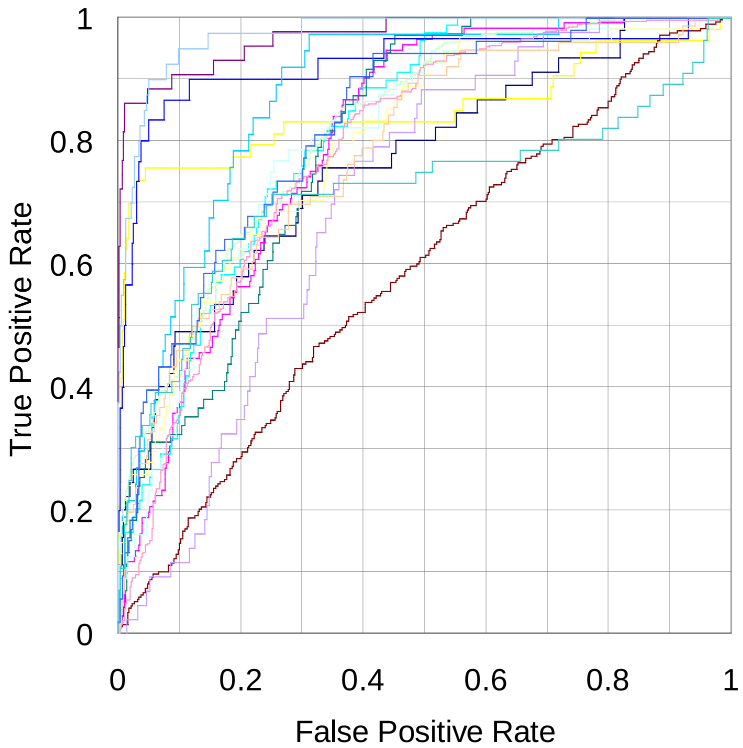
$$p(o \mid \{F_I^a\}) = \prod_a p(o \mid F_I^a)$$

- Form intersections of the different types of regions, creating smaller regions that have both color and texture properties for classification.

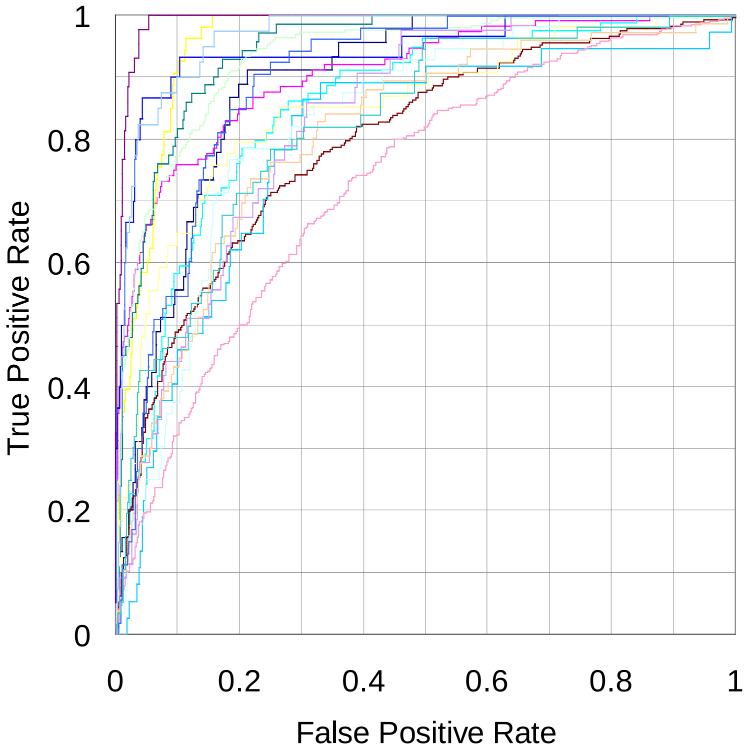
Experiments (on 860 images)

- 18 keywords: mountains (30), orangutan (37), track (40), tree trunk (43), football field (43), beach (45), prairie grass (53), cherry tree (53), snow (54), zebra (56), polar bear (56), lion (71), water (76), chimpanzee (79), cheetah (112), sky (259), grass (272), tree (361).
- A set of cross-validation experiments (80% as training set and the other 20% as test set)
- The poorest results are on object classes “tree,” “grass,” and “water,” each of which has a high variance; a single Gaussian model is insufficient.

ROC Charts



Independent Treatment of Color and Texture



Using Intersections of Color and Texture Regions

Sample Retrieval Results

cheetah

Sample Results (Cont.)

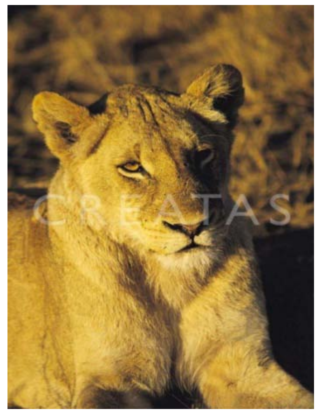
grass

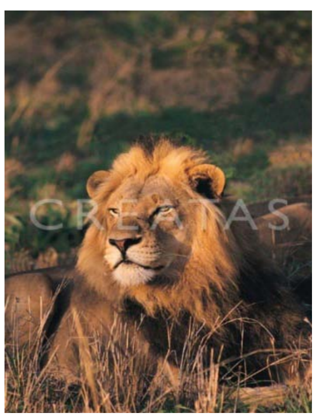
Sample Results (Cont.)

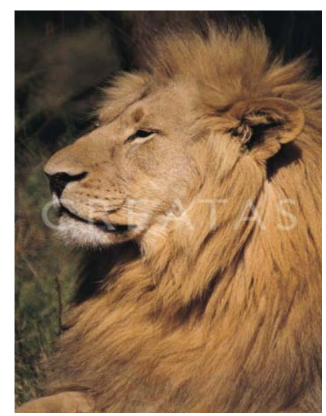
cherry tree

Sample Results (Cont.)

lion







Summary

- Designed a set of abstract region features: color, texture, structure, . . .
- Developed a new semi-supervised EM-like algorithm to recognize object classes in color photographic images of outdoor scenes; tested on 860 images.
- Compared two different methods of combining different types of abstract regions. The intersection method had a higher performance

Our New Approach to Combining Different Feature Types

Phase 1:

- Treat each type of abstract region separately
- For abstract region type a and for object class o, use the EM algorithm to construct a model that is a mixture of multivariate Gaussians over the features for type a regions.

Consider only abstract region type color (c) and object class object (o)

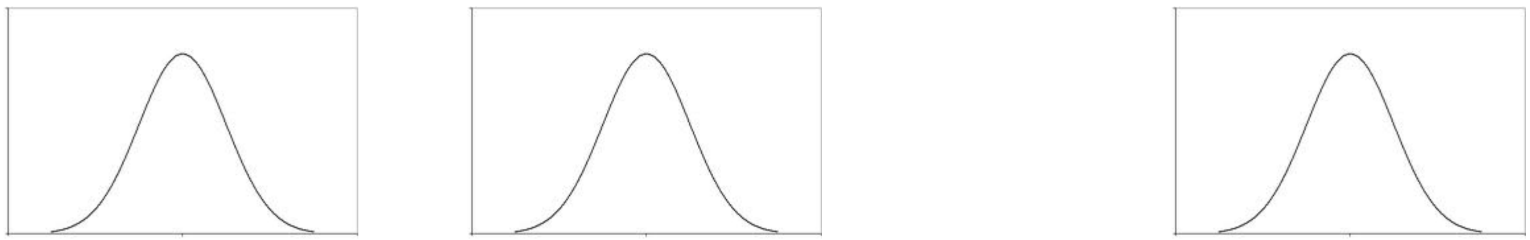
- At the end of Phase 1, we can compute the distribution of color feature vector in an image containing object o.

$$P(X^c|o) = \sum_{m=1}^{M^c} w_m^c \cdot N(X^c; \mu_m^c, \Sigma_m^c)$$

- M^c is the number of components.
- The w's are the weights of the components.
- The μ's and $\sum$'s are the parameters of the components

Color Components for Class o

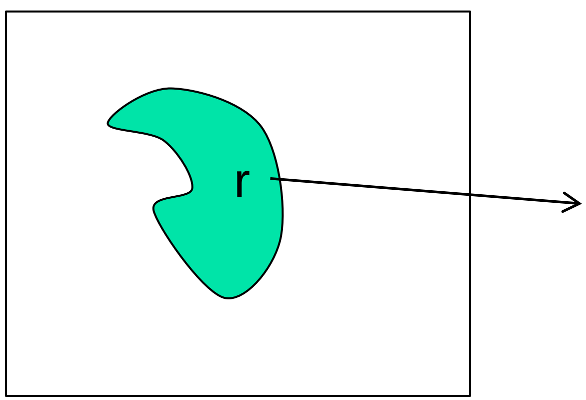
$$P(X^c|o) = \sum_{m=1}^{M^c} w_m^c \cdot N(X^c; \mu_m^c, \Sigma_m^c)$$



component 1 μ_1, Σ_1, w_1

component 2 μ_2, Σ_2, w_2

component M^c μ_M, Σ_M, w_M



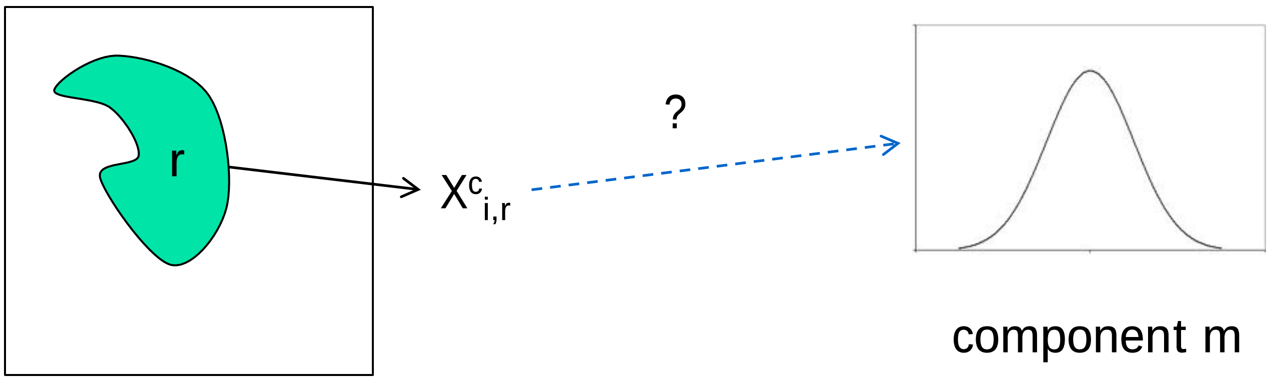
color feature vector X^c for region r

Now we can determine which components are likely to be present in an image.

- The probability that the feature vector X from color region r of image I_i comes from component m is given by

$$P(X_{i,r}^c, m^c) = w_m^c \cdot N(X_{i,r}^c, \mu_m^c, \Sigma_m^c)$$

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

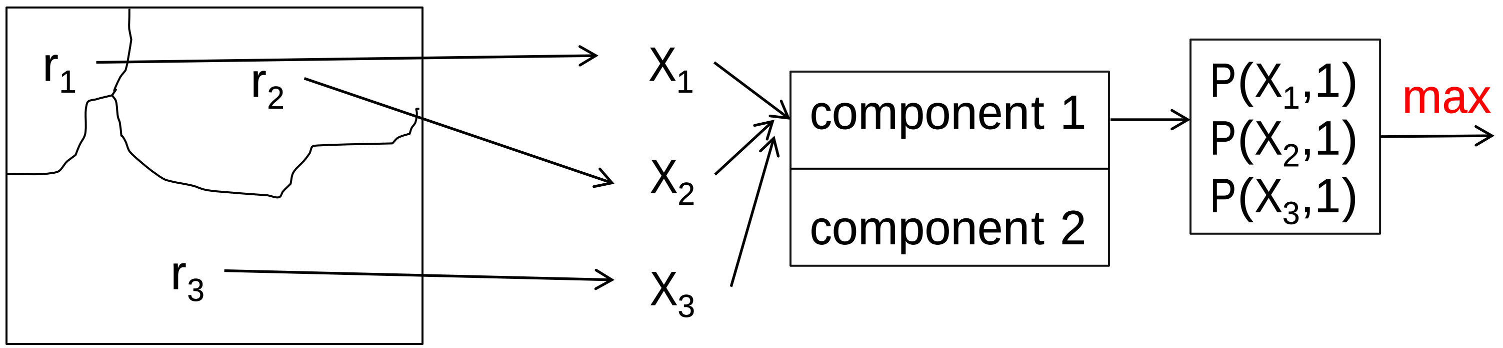


And determine the probability that the whole image is related to component m as a function of the feature vectors of all its regions.

- Then the probability that image I_i has a region that comes from component m is

$$P(I_i, m^c) = f(\{P(X^c_{i,r}, m^c) | r = 1, 2, \ldots\})$$

- where f is an aggregate function such as mean or max



Aggregate Scores for Color

Components

1	2	3	4	5	6	7	8

beach

.93	.16	.94	.24	.10	.99	.32	.00

beach

.66	.80	.00	.72	.19	.01	.22	.02

not beach

.43	.03	.00	.00	.00	.00	.15	.00

We now use positive and negative training images, calculate for each the probabilities of regions of each component, and form a training matrix.

$$
\begin{array}{c}
I_1^+ \\ I_2^+ \\ \vdots \\ I_1^- \\ I_2^- \\ \vdots
\end{array}
\begin{bmatrix}
P(I_1^+, 1^c) & P(I_1^+, 2^c) & \cdots & P(I_1^+, M^c) \\
P(I_2^+, 1^c) & P(I_2^+, 2^c) & \cdots & P(I_2^+, M^c) \\
\vdots & & & \\
P(I_1^-, 1^c) & P(I_1^-, 2^c) & \cdots & P(I_1^-, M^c) \\
P(I_2^-, 1^c) & P(I_2^-, 2^c) & \cdots & P(I_2^-, M^c) \\
\vdots & & &
\end{bmatrix}
$$

Phase 2 Learning

- Let C_i be row i of the training matrix.
- Each such row is a feature vector for the color features of regions of image I_i that relates them to the Phase 1 components.
- Now we can use a second-stage classifier to learn $P(o|I_i)$ for each object class o and image I_i .

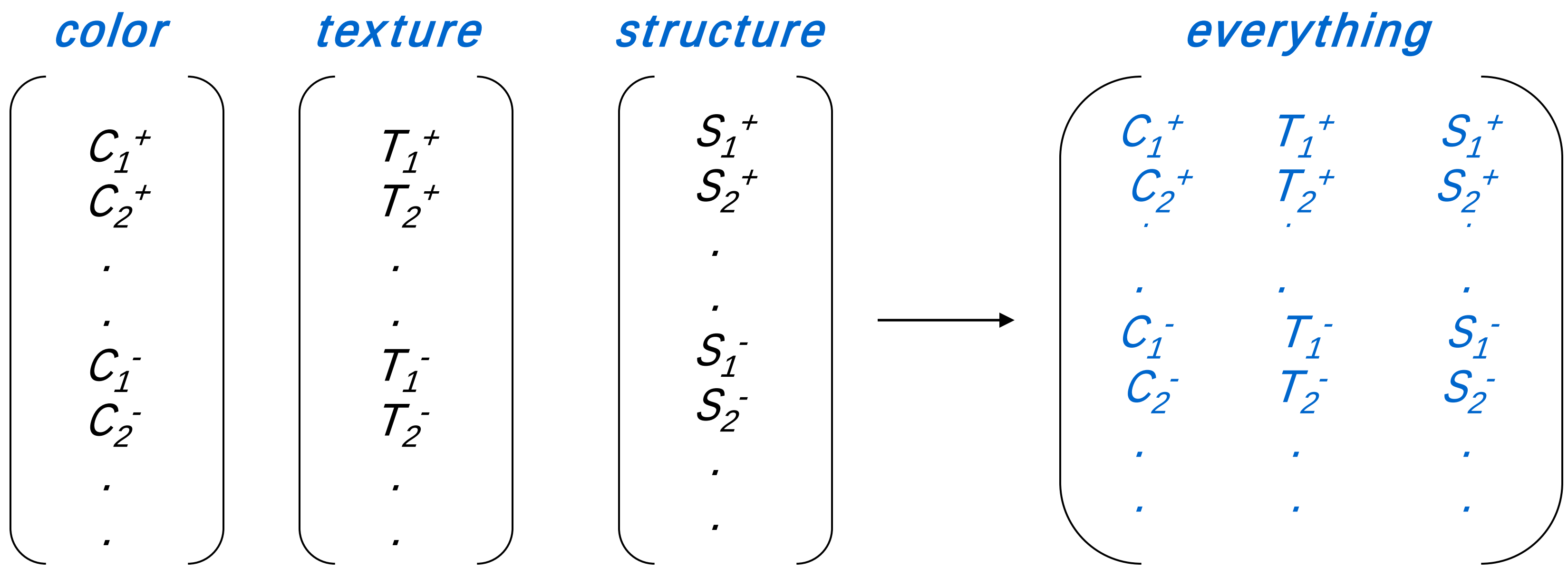
Multiple Feature Case

- We calculate separate Gaussian mixture models for each different features type:

- Color: C_i
- Texture: T_i
- Structure: S_i

- and any more features we have (motion).

Now we concatenate the matrix rows from the different region types to obtain a multi-feature-type training matrix.



ICPR04 Data Set with General Labels

	EM-variant with single Gaussian per object	EM-variant extension to mixture models	Gen/Dis with Classical EM clustering	Gen/Dis with EM-variant extension
African animal	71.8%	85.7%	89.2%	90.5%
arctic	80.0%	79.8%	90.0%	85.1%
beach	88.0%	90.8%	89.6%	91.1%
grass	76.9%	69.6%	75.4%	77.8%
mountain	94.0%	96.6%	97.5%	93.5%
primate	74.7%	86.9%	91.1%	90.9%
sky	91.9%	84.9%	93.0%	93.1%
stadium	95.2%	98.9%	99.9%	100.0%
tree	70.7%	79.0%	87.4%	88.2%
water	82.9%	82.3%	83.1%	82.4%
MEAN	**82.6%**	**85.4%**	**89.6%**	**89.3%**

Comparison to ALIP: the Benchmark Image Set

- Test database used in SIMPLIcity paper and ALIP paper.
- 10 classes (*African people*, *beach*, *buildings*, *buses*, *dinosaurs*, *elephants*, *flowers*, *food*, *horses*, *mountains*). 100 images each.

Comparison to ALIP: the Benchmark Image Set

	ALIP	cs	ts	st	ts+st	cs+st	cs+ts	cs+ts+st
African	52	69	23	26	35	79	72	74
beach	32	44	38	39	51	48	59	64
buildings	64	43	40	41	67	70	70	78
buses	46	60	72	92	86	85	84	95
dinosaurs	100	88	70	37	86	89	94	93
elephants	40	53	8	27	38	64	64	69
flowers	90	85	52	33	78	87	86	91
food	68	63	49	41	66	77	84	85
horses	60	94	41	50	64	92	93	89
mountains	84	43	33	26	43	63	55	65
MEAN	**63.6**	**64.2**	**42.6**	**41.2**	**61.4**	**75.4**	**76.1**	**80.3**

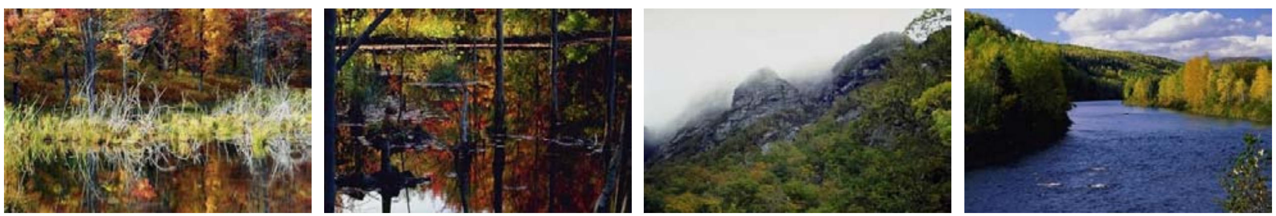
Comparison to ALIP: the 60K Image Set

- 59,895 COREL images and 599 categories;
- Each category has about 100 images;
- 8 images per category were reserved for testing.
- To train on one category, all the available 92 positive images were used find the clusters. Those positive images, along with 1,000 randomly selected negative images were then used to train the MLPs.

Comparison to ALIP: the 60K Image Set

0. Africa, people, landscape, animal

1. autumn, tree, landscape, lake



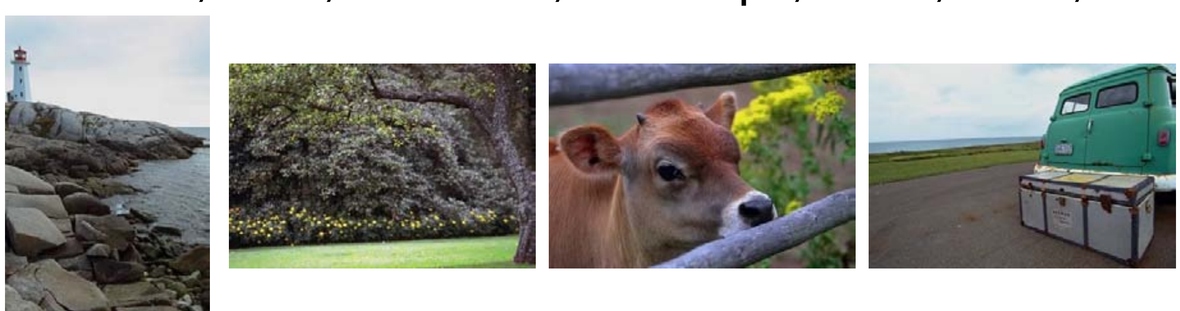
2. Bhutan, Asia, people, landscape, church

Comparison to ALIP: the 60K Image Set

3. California, sea, beach, ocean, flower

4. Canada, sea, boat, house, flower, ocean

5. Canada, west, mountain, landscape, cloud, snow, lake



Comparison to ALIP: the 60K Image Set

Number of top-ranked categories required	1	2	3	4	5
ALIP	11.88	17.06	20.76	23.24	26.05
Gen/Dis	11.56	17.65	21.99	25.06	27.75

The table shows the percentage of test images whose true categories were included in the top-ranked categories.

Groundtruth Data Set

- UW Ground truth database (1224 images)
- 31 elementary object categories: *river* (30), *beach* (31), *bridge* (33), *track* (35), *pole* (38), *football field* (41), *frozen lake* (42), *lantern* (42), *husky stadium* (44), *hill* (49), *cherry tree* (54), *car* (60), *boat* (67), *stone* (70), *ground* (81), *flower* (85), *lake* (86), *sidewalk* (88), *street* (96), *snow* (98), *cloud* (119), *rock* (122), *house* (175), *bush* (178), *mountain* (231), *water* (290), *building* (316), *grass* (322), *people* (344), *tree* (589), *sky* (659)
- 20 high-level concepts: *Asian city , Australia, Barcelona, campus, Cannon Beach, Columbia Gorge, European city, Geneva, Green Lake, Greenland, Indonesia, indoor, Iran, Italy, Japan, park, San Juans, spring flowers, Swiss mountains, and Yellowstone*.

beach, sky, tree, water

people, street, tree

building, grass, people, sidewalk, sky, tree

building, bush, sky, tree, water

flower, house, people, pole, sidewalk, sky

flower, grass, house, pole, sky, street, tree

building, flower, sky, tree, water

boat, rock, sky, tree, water

building, car, people, tree

car, people, sky

boat, house, water

building

Groundtruth Data Set: ROC Scores

street	60.4	*tree*	80.8	*stone*	87.1	*columbia gorge*	94.5
people	68.0	*bush*	81.0	*hill*	87.4	*green lake*	94.9
rock	73.5	*flower*	81.1	*mountain*	88.3	*italy*	95.1
sky	74.1	*iran*	82.2	*beach*	89.0	*swiss moutains*	95.7
ground	74.3	*bridge*	82.7	*snow*	92.0	*sanjuans*	96.5
river	74.7	*car*	82.9	*lake*	92.8	*cherry tree*	96.9
grass	74.9	*pole*	83.3	*frozen lake*	92.8	*indoor*	97.0
building	75.4	*yellowstone*	83.7	*japan*	92.9	*greenland*	98.7
cloud	75.4	*water*	83.9	*campus*	92.9	*cannon beach*	99.2
boat	76.8	*indonesia*	84.3	*barcelona*	92.9	*track*	99.6
lantern	78.1	*sidewalk*	85.7	*geneva*	93.3	*football field*	99.8
australia	79.7	*asian city*	86.7	*park*	94.0	*husky stadium*	100.0
house	80.1	*european city*	87.0	*spring flowers*	94.4		

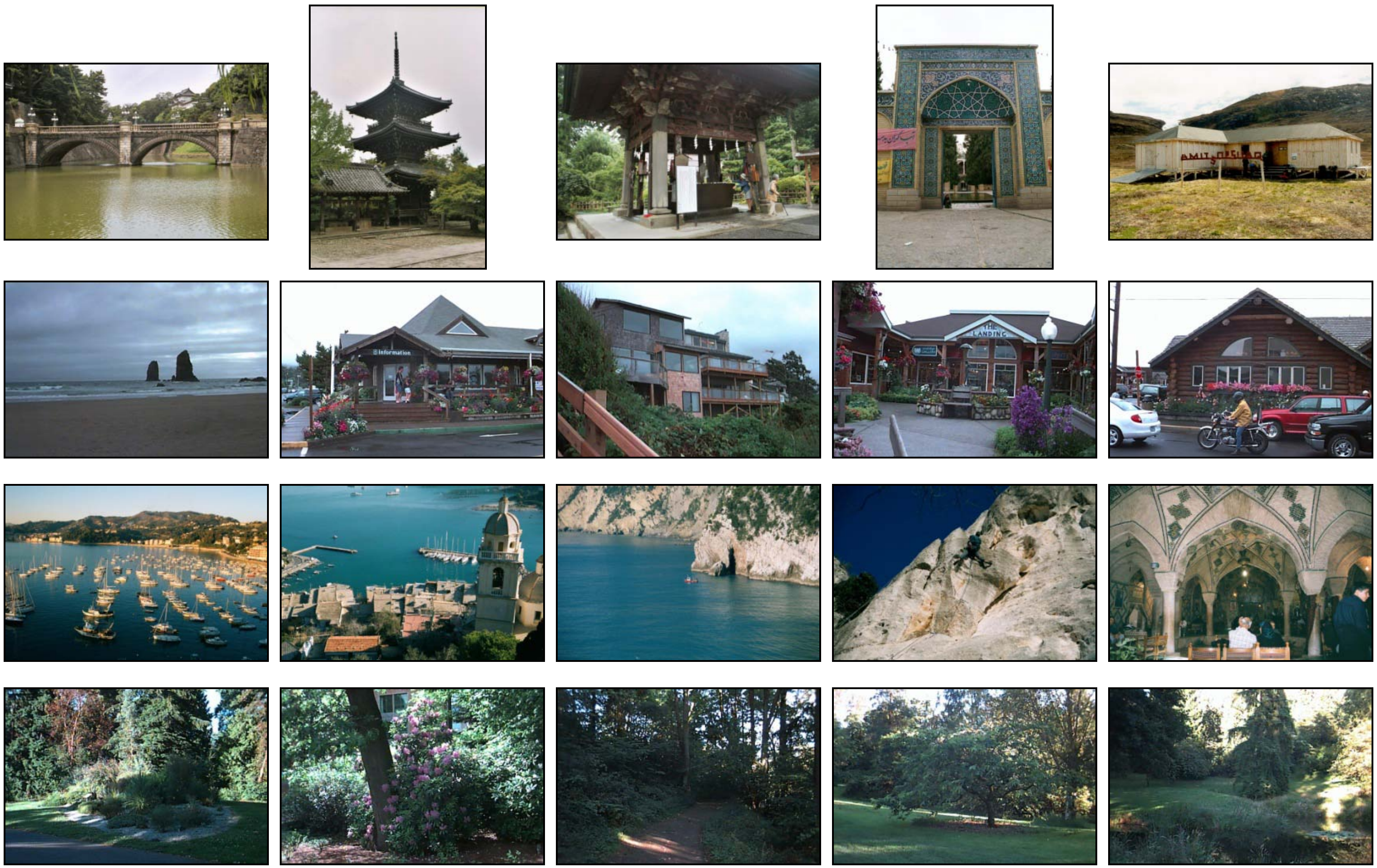
Groundtruth Data Set: Top Results

Asian city

Cannon beach

Italy

park



Groundtruth Data Set: Top Results

sky

spring flowers

tree

water

Groundtruth Data Set: Annotation Samples

tree(97.3), **bush**(91.6), **spring flowers**(90.3), **flower**(84.4), park(84.3), **sidewalk**(67.5), **grass**(52.5), **pole**(34.1)

sky(99.8), **Columbia gorge**(98.8), lantern(94.2), **street**(89.2), house(85.8), bridge(80.8), car(80.5), hill(78.3), boat(73.1), pole(72.3), **water**(64.3), mountain(63.8), **building**(9.5)

sky(95.1), **Iran**(89.3), house(88.6), **building**(80.1), boat(71.7), bridge(67.0), **water**(13.5), **tree**(7.7)

Italy(99.9), grass(98.5), **sky**(93.8), rock(88.8), **boat**(80.1), **water**(77.1), Iran(64.2), stone(63.9), bridge(59.6), **European**(56.3), sidewalk(51.1), **house**(5.3)

Comparison to Fergus and to Dorko/Schmid using their Features

Using their features and image sets, we compared our generative / discriminative approach to those of Fergus and Dorko/Schmid.

The image set contained 1074 airplane images, 826 motor bike images, 450 face images, and 900 background. Half were used to train and half to test. We added half the background images to the training set for our negative examples.

	Fergus	Dorko/Schmid	Ours
airplanes	90.2%	96.0%	96.6%
faces	96.4%	96.8%	96.5%
motorbikes	92.5%	98.0%	99.2%

Structure Feature Experiments

(from other data sets with more manmade structures)

- 1,951 total from freefoto.com
- bus (1,013) house/building (609) skyscraper (329)

Structure Feature Experiments: Area Under the ROC Curves

1. Structure (with color pairs)
   - Attributes (10)
     - Color pair
     - Number of lines
     - Orientation of lines
     - Line overlap
     - Line intersection

2. Structure (with color pairs) + Color Segmentation

3. Structure (without color pairs) + Color Segmentation

	bus	house/ building	skyscraper
Structure only	0.900	0.787	0.887
Structure + Color Seg	0.924	0.853	0.926
Structure^2 + Color Seg	0.940	0.860	0.919