Multimodal Fusion using Learned Text Concepts for Image Categorization

Qiang Zhu*, **Mei-Chen Yeh**, **Kwang-Ting Cheng**

Learning-based Multimedia Lab

Department of Electrical & Computer Engineering

University of California at Santa Barbara


EXIT 3
1 MILE
homas
Rd


EXIT 20
51st Ave
3/4 MILE


Glendale
Ave


FREEWAY ENDS
Price Rd
1/2 MILE
101


HEALTH TECHNOLOGIES


Noble Hall


SANTA BARBARA
BANK & TRUST

Image categorization based on low-level cues

Image categorization based on detected text lines

Fusion

Image Category

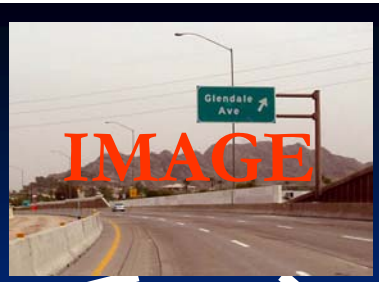


Image categorization based on low-level cues

Text Detector

Text Lines

Multiple Instance Learning

Text Concept

Minimum Distance

Bullhead City
Needles

Classification

Fusion

Image Category

False Alarm

- **Learning text concept is not a trivial task !**
  1. Multiple text lines contained in one image
  2. Not every text line is representative in the target category
  3. High false alarm for text detection

- ***Multiple Instance Learning*** **(MIL): To learn a text concept in a target category**

Multiple Instance Learning

Originally for learning concepts from ambiguity data

1. A bag is a collection of instances
2. Positive bag has at least one positive instance
3. Negative bag has only negative instances

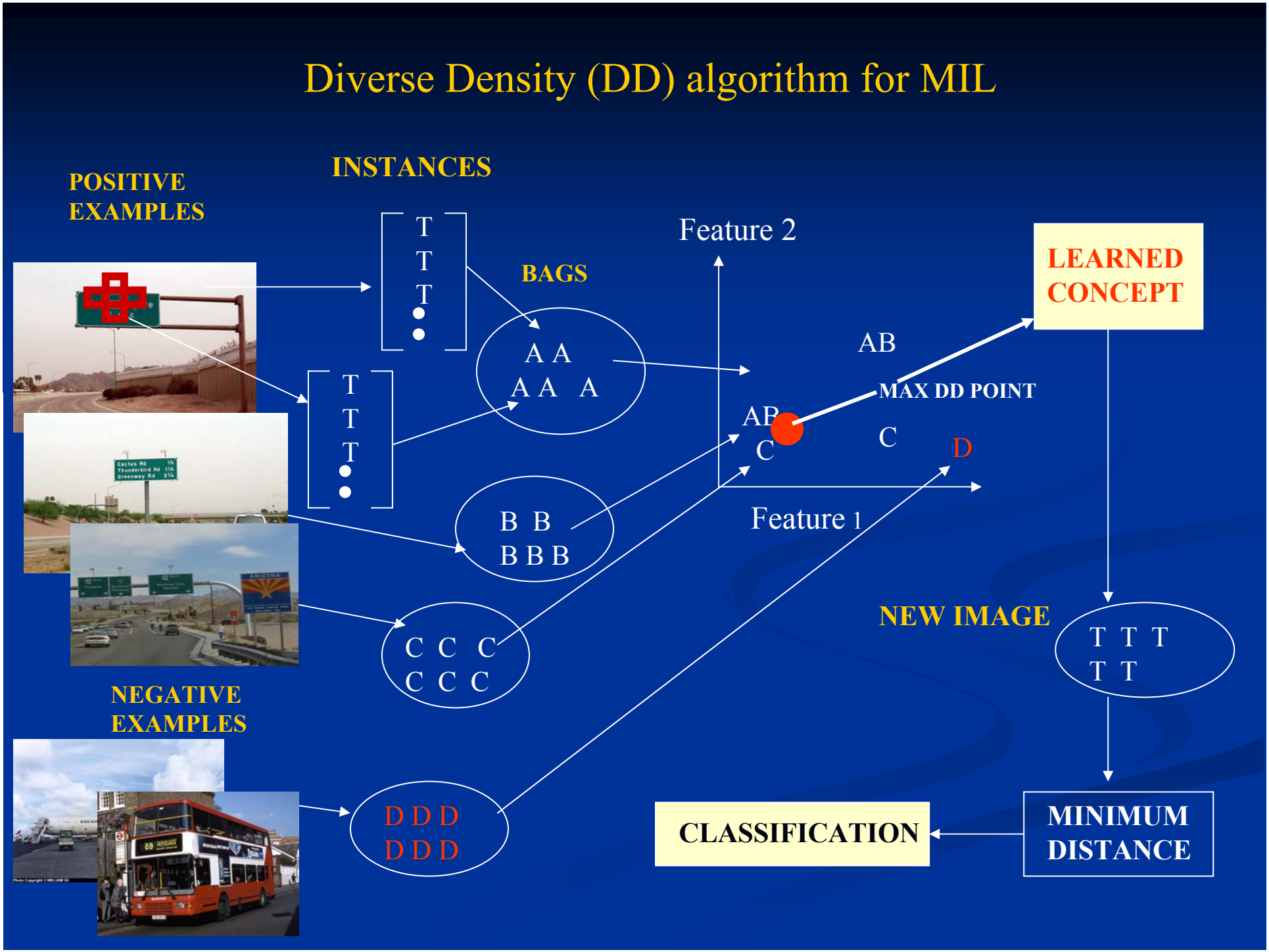
Analogy:

Image = Bag & Text line = Instance

Algorithms developed for Multiple Instance Learning

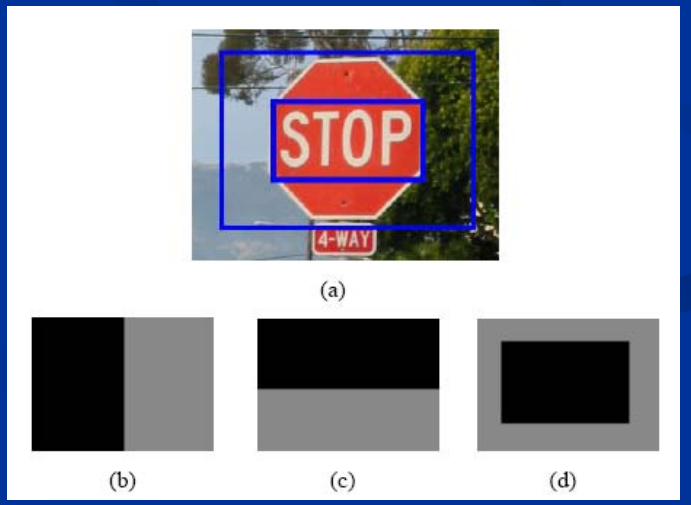
- Iterated-discrim APR [Dietterich *et al.*, 1997]
- **Diverse Density (DD) [Maron and Lozano-Perez, 1998]**
- EM-DD [Zhang and Goldman, 2001]
- Two SVM variants for MIL [Andrews *et al.*, 2002]
- Citation-kNN for MIL [Wang and Zucker, 2000]

Diverse Density (DD) algorithm for MIL



For each text line, we use a 16-d feature vector to describe this region

- Background color and foreground color by analyzing the color histogram
- Text size and location in the target image
- Global textures: edge, brightness, contrast, moments
- Local structures



Text Concept of an Image Category

Represented as two 16-d vectors:

- A 16-d feature vector with representative values across all the text lines in this category,
- A 16-d weight vector indicating impacts of each feature

Multiple Instance Learning (MIL) learns the text concept from a given set of training images and its detected text lines embedded in each image

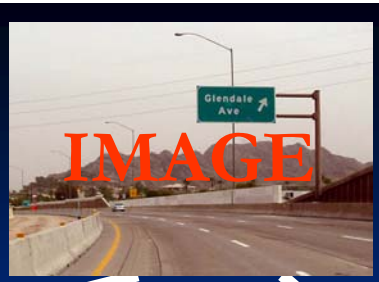


Image categorization based on low-level cues

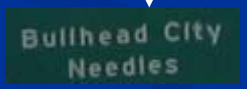
Text Detector

Text Lines

Multiple Instance Learning

Text Concept

Minimum Distance



Accuracy?

Fusion

Image Category

imET (Images with Embedded Texts) Database

(http://lbmedia.ece.ucsb.edu/resources/dataset/imET.rar)

Category name	Number of images	Average number of text lines per image
Airplane	260	1.7
Building	207	1.2
Bus	294	5.2
Camera	260	2.1
Car	220	3.3
Door	200	1.6
Highway	205	3.4
Street	250	1.6


STOP


Kalitta Charters II


Camaro

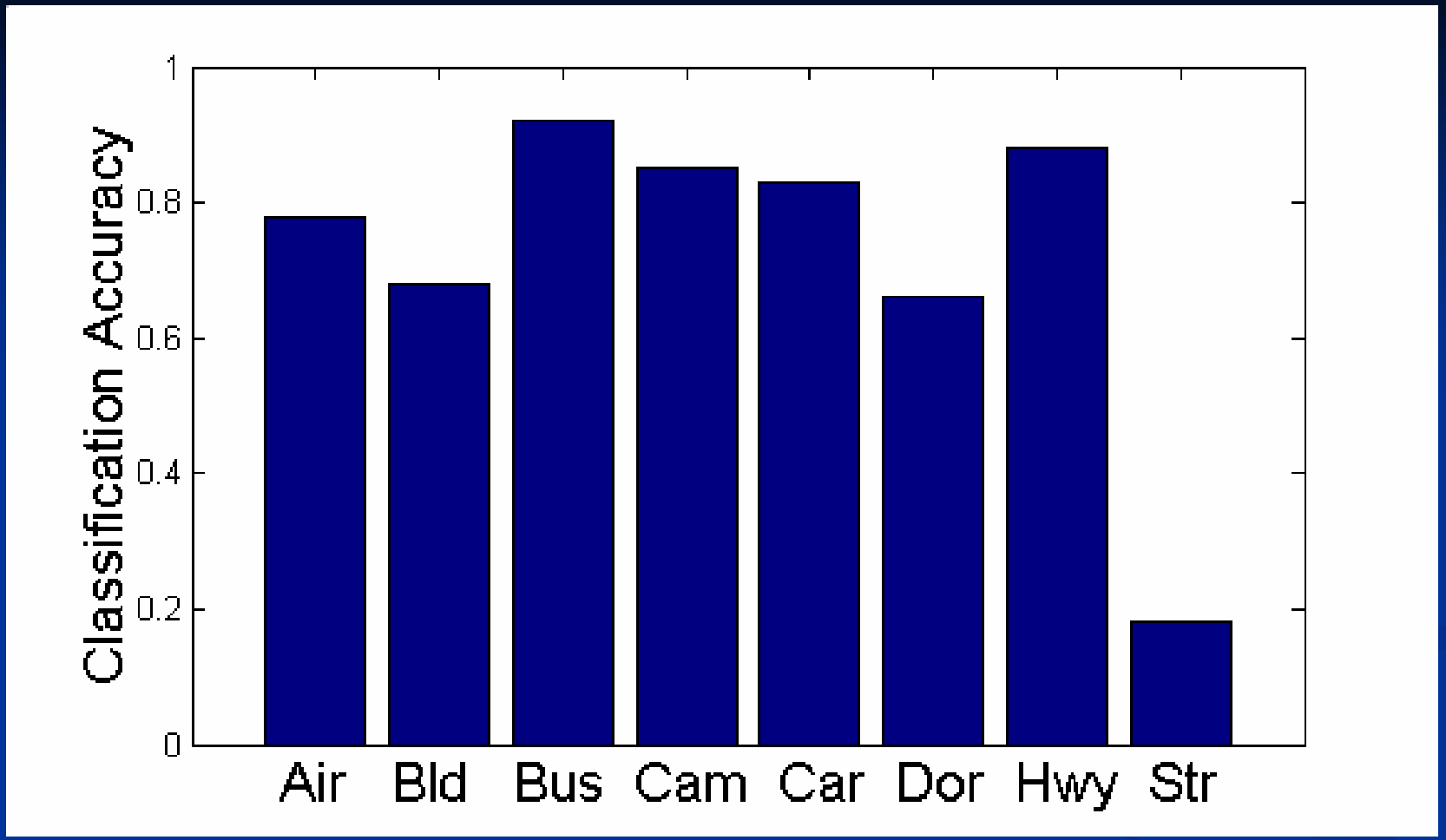

STAIRS


STAPLES


FUJIFILM
SMART SHOT
PLUS


Glendale
Ave


AA
AA
A
A



Directly apply this metric to image categorization

1. Average accuracy is 72.3%, or 80.1% if exclude "Street"
2. Low performance for "Street" category

Category	Top-3 features with maximum impacts
Airplane	{location, contrast, edge}
Building	{edge, location, local structure}
Bus	{edge, location, aspect ratio}
Camera	{size, edge, local structure}
Car	{aspect ratio, location, edge}
Door	{aspect ratio, edge, variance}
Highway	{color, mean, local structure}
Street	{local structure, location, edge}

Text concepts in different categories rely on different features

Image categorization based on low-level cues

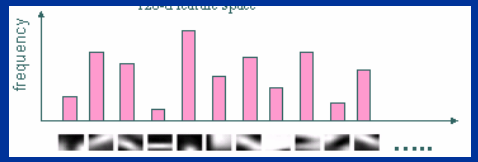
Image categorization based on detected text lines

Fusion

Image Category

Low-level visual cues

Bag-of-words model



Naïve Bayes classifier

Probability

Image categorization based on detected text lines

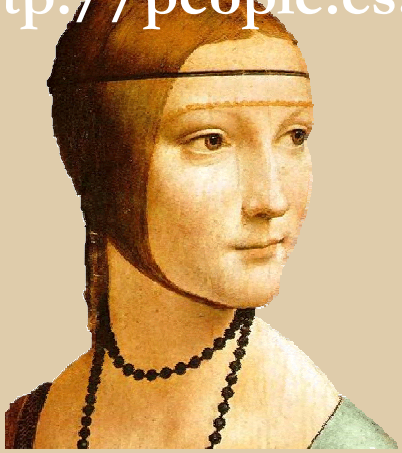
Fusion

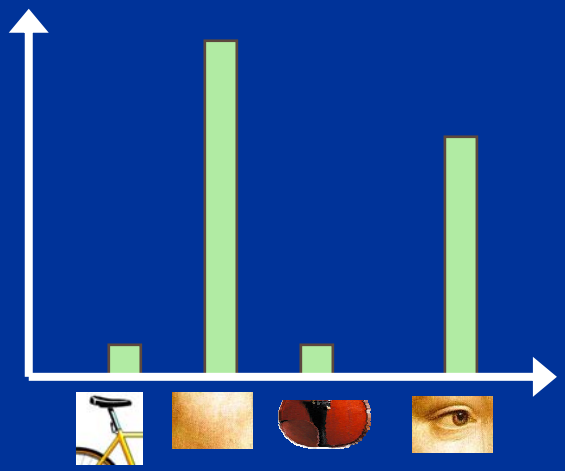
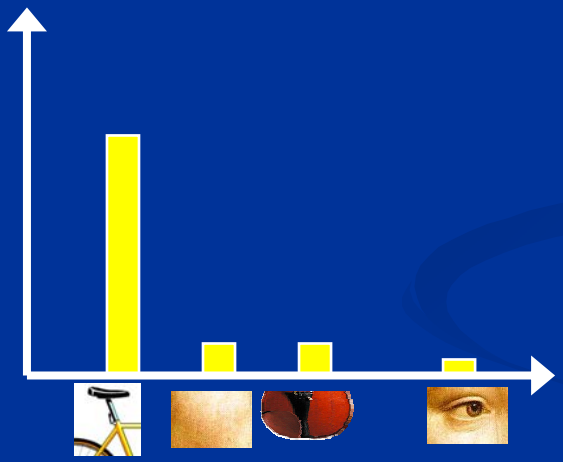
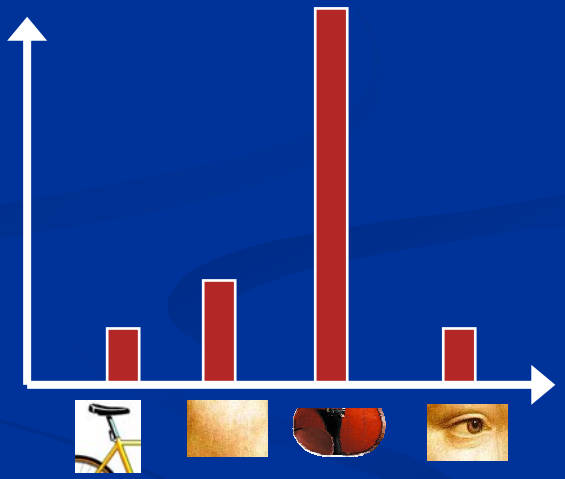
Image Category

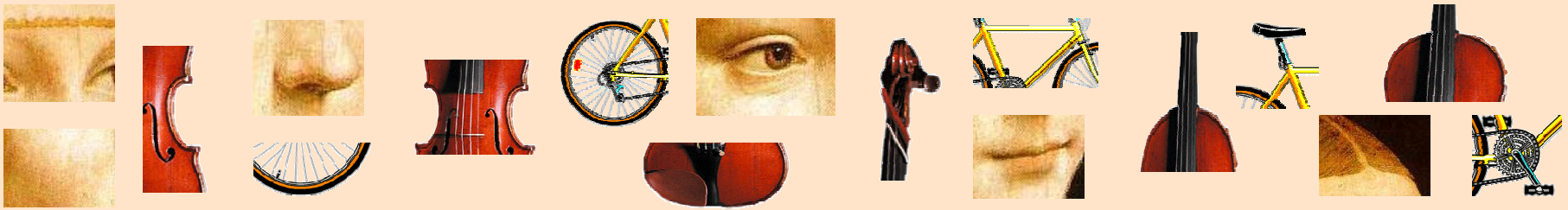
http://people.csail.mit.edu/torralba/iccv2005/

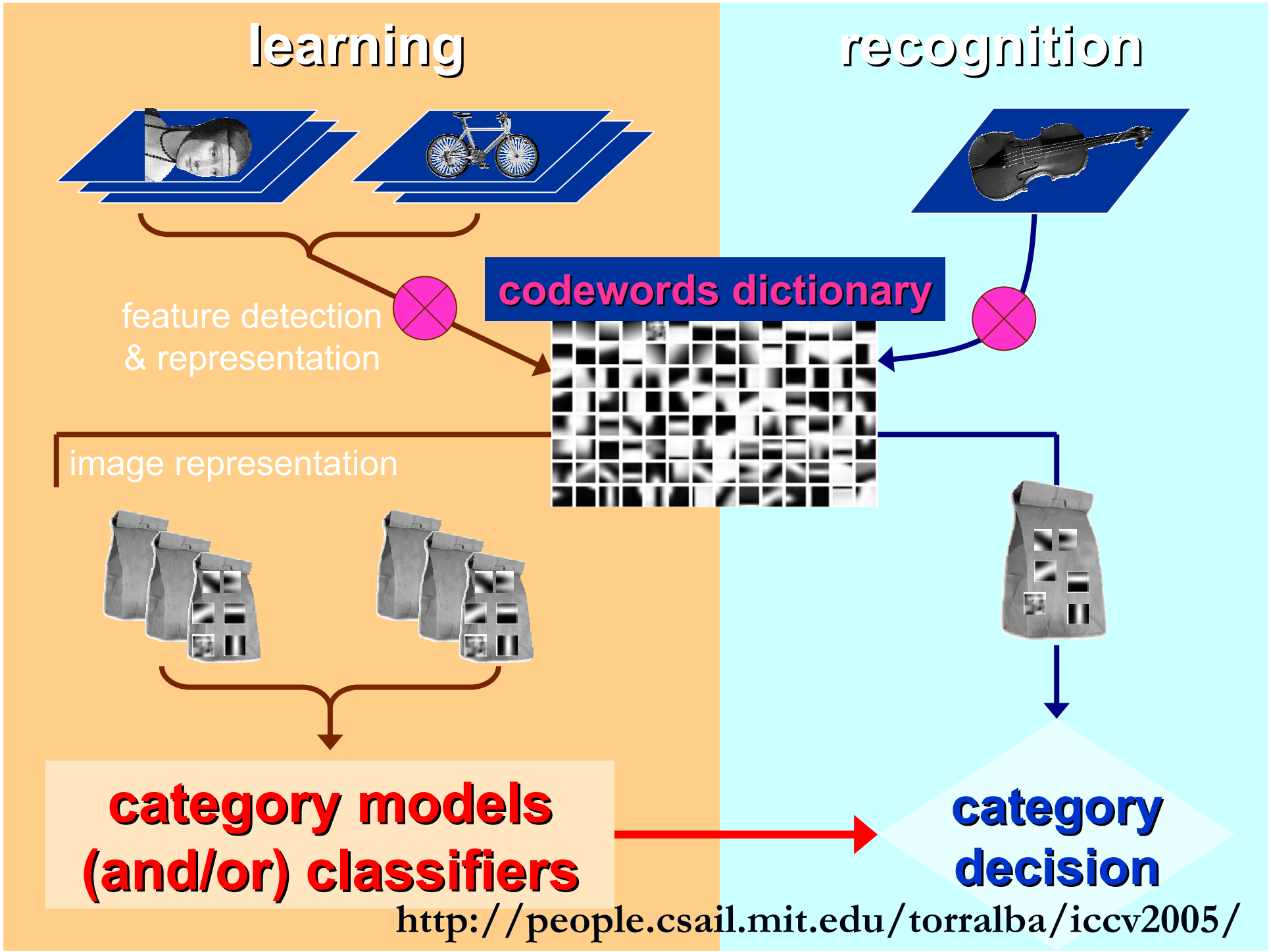
Object → Bag of 'words'

http://people.csail.mit.edu/torralba/iccv2005/



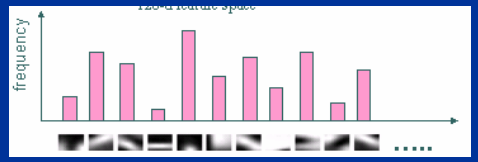







learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision
http://people.csail.mit.edu/torralba/iccv2005/

Low-level visual cues

Bag-of-words model



Naïve Bayes classifier

Accuracy?

Image categorization based on detected text lines

Fusion

Image Category

Classification Accuracy --- the Baseline

Average 81.3%	air	bldg	bus	cam	car	door	hwy	street
Airplane	**84.6**	0.00	5.13	0.00	3.85	0.00	5.13	1.28
Building	1.61	**46.8**	1.61	1.61	9.68	12.9	0.00	25.8
Bus	0.00	0.00	**95.5**	0.00	3.41	0.00	0.00	1.14
Camera	0.00	3.03	0.00	**92.4**	4.55	0.00	0.00	0.00
Car	1.28	2.56	2.56	1.28	**84.6**	3.85	0.00	3.85
Door	1.67	6.67	1.67	1.67	6.67	**78.3**	1.67	1.67
Highway	1.64	1.64	3.28	0.00	3.28	4.92	**80.3**	4.92
Street	0.00	2.67	0.00	0.00	5.33	0.00	4.00	**88.0**

Image categorization based on low-level cues

Image categorization based on detected text lines

Fusion

Image Category

Image categorization based on low-level cues

Image categorization based on detected text lines

A linear SVM classifier

1. Probabilities from a Bag-of-words model
2. Distances to the learned text concepts
3. Number of text lines

Image Category

Accuracy?

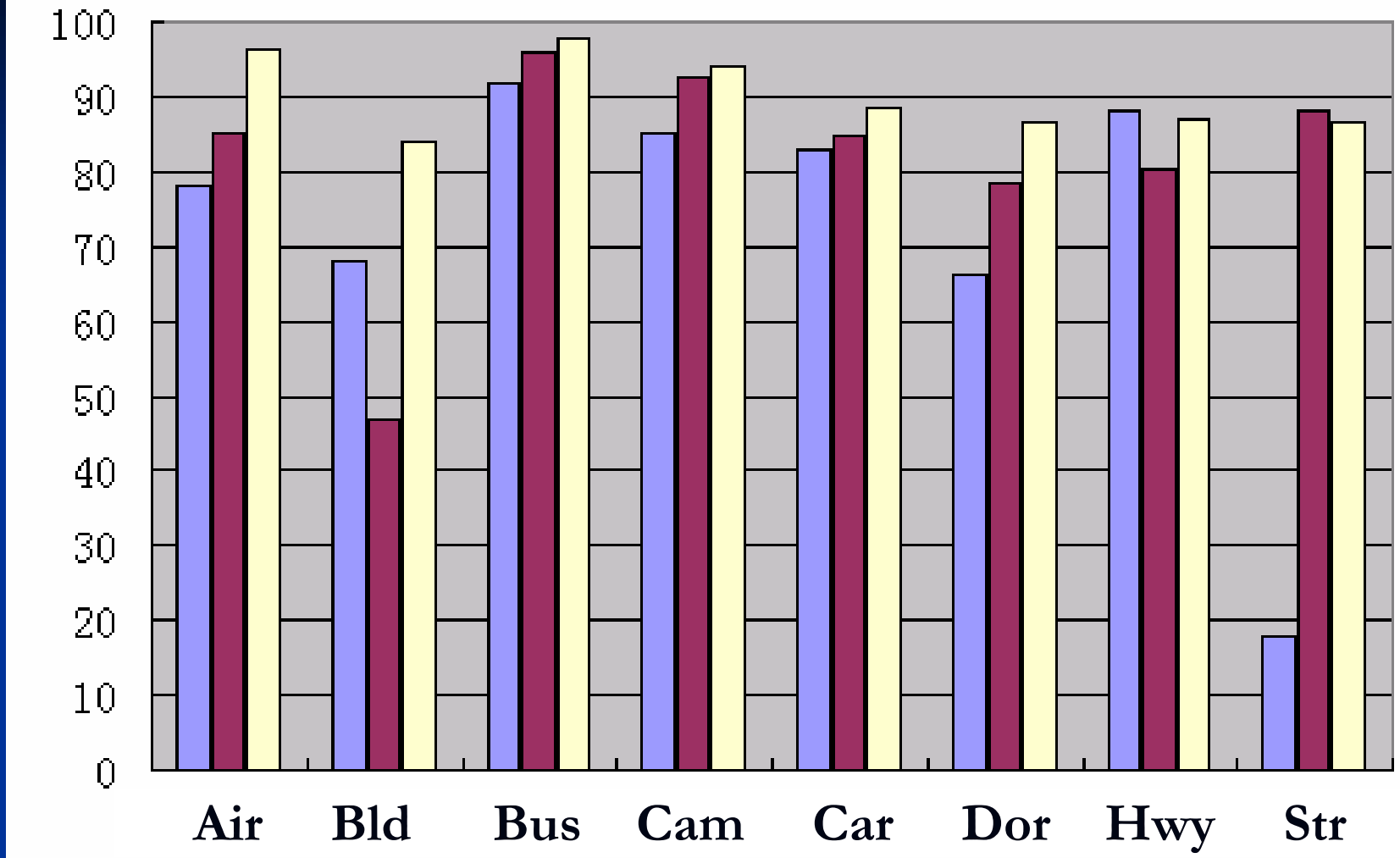


Image Categorization using Text Concept (Average accuracy =72.3)

Image Categorization using Bag-of-words (Average accuracy =81.3)

Image Categorization using fusion approach (Average accuracy =90.1)

Conclusion

- Embedded texts is valuable for image categorization
- Multiple Instance Learning efficiently finds the useful patterns in the text lines of a target category
- Robust for noise such as multiple texts, false texts
- Bridge text detection and image categorization

Thank you!

check our website http://lbmedia.ece.ucsb.edu
for project details, dataset and slides