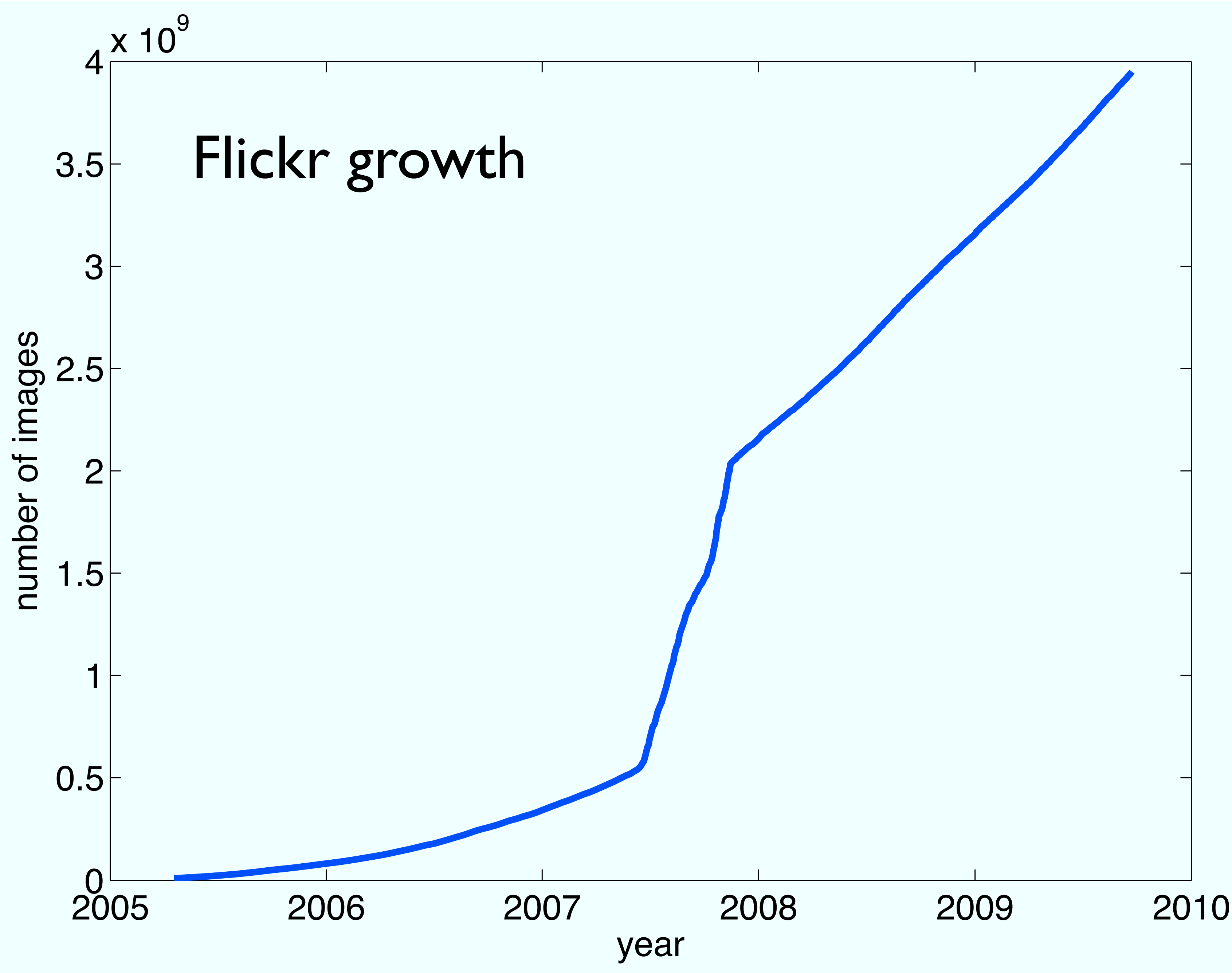
Scaling Object Recognition: Benchmark of Current State of the Art Techniques

Mohamed Aly[1], Peter Welinder[1],
Mario Munich[2], Pietro Perona[1]

(1)California Institute of Technology
(2)Evolution Robotics

Motivation



Flickr growth
x 10^9
0
0.5
1
1.5
2
2.5
3
3.5
4
number of images
2005
2006
2007
2008
2009
2010
year

Motivation

Large image collections:

- 20 billions on ImageShack
- 15 billions on Facebook
- 7 billions on Photobucket
- 4 billions on Flickr

[http://www.techcrunch.com April 2009]

Motivation

Individual object recognition:
How do current methods scale?

- CPU cycles
- RAM
- Precision / recall

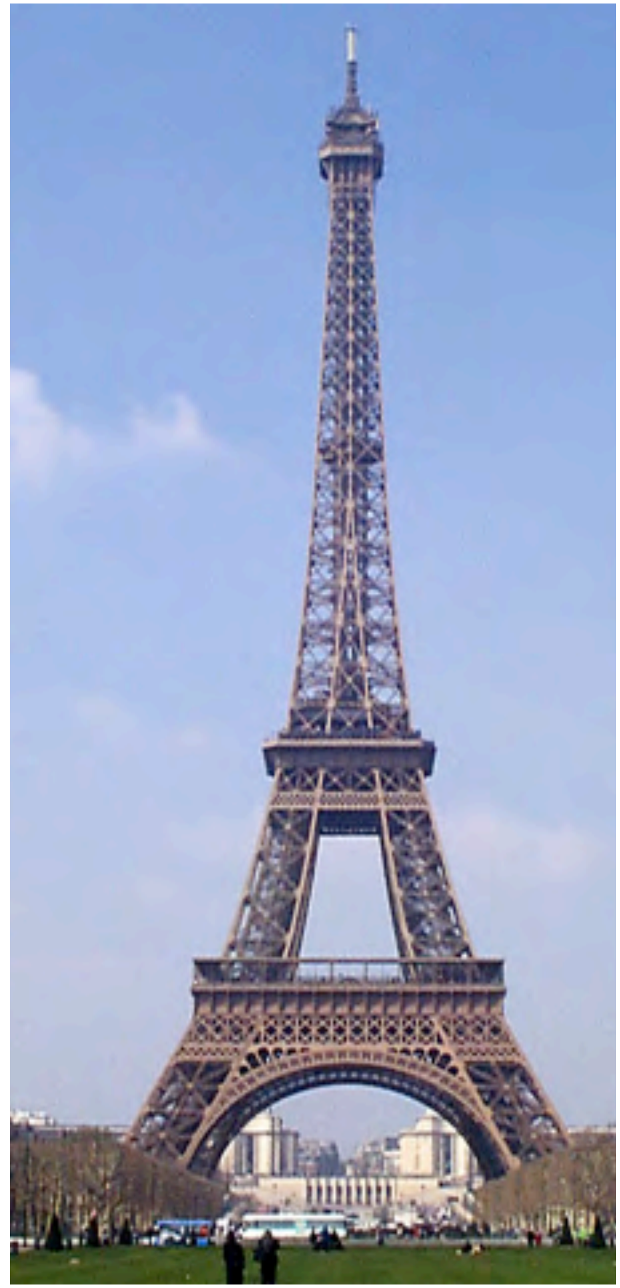
Outline

- Datasets
- Recognition Methods
- Experimental Setup
- Results
- Conclusions

Three flavors:

Item to item

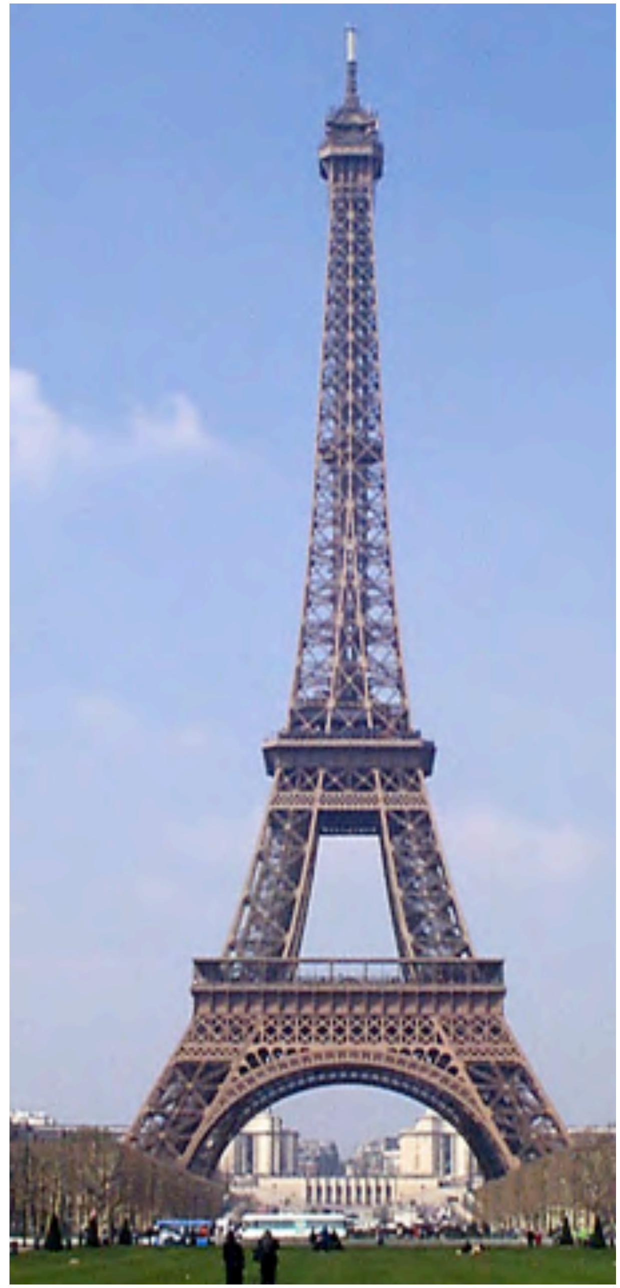
Probe



Model

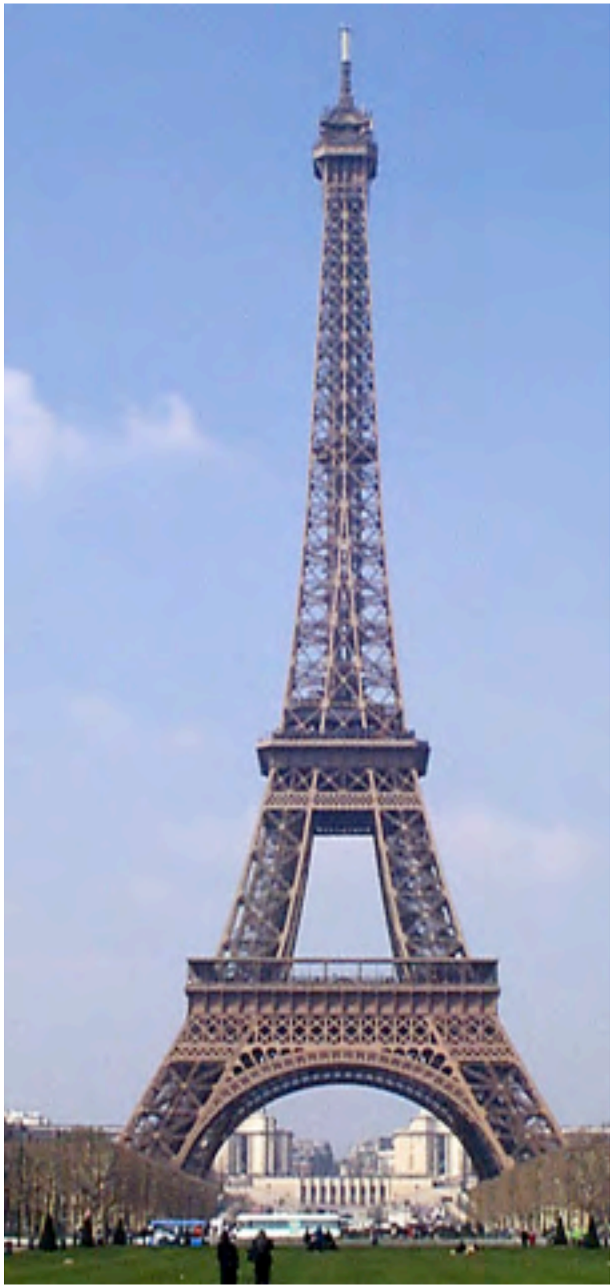
Scene to item

Probe



Model

Item to scene



Probe

Model

Three flavors:

- Item to item
- Item to scene
- Scene to item

Dataset 1: CD Covers

Model Set: ~ 132,000 unique images

downloaded from freecovers.net (available on vision.caltech.edu)

Dataset 1: CD Covers

Probe sets

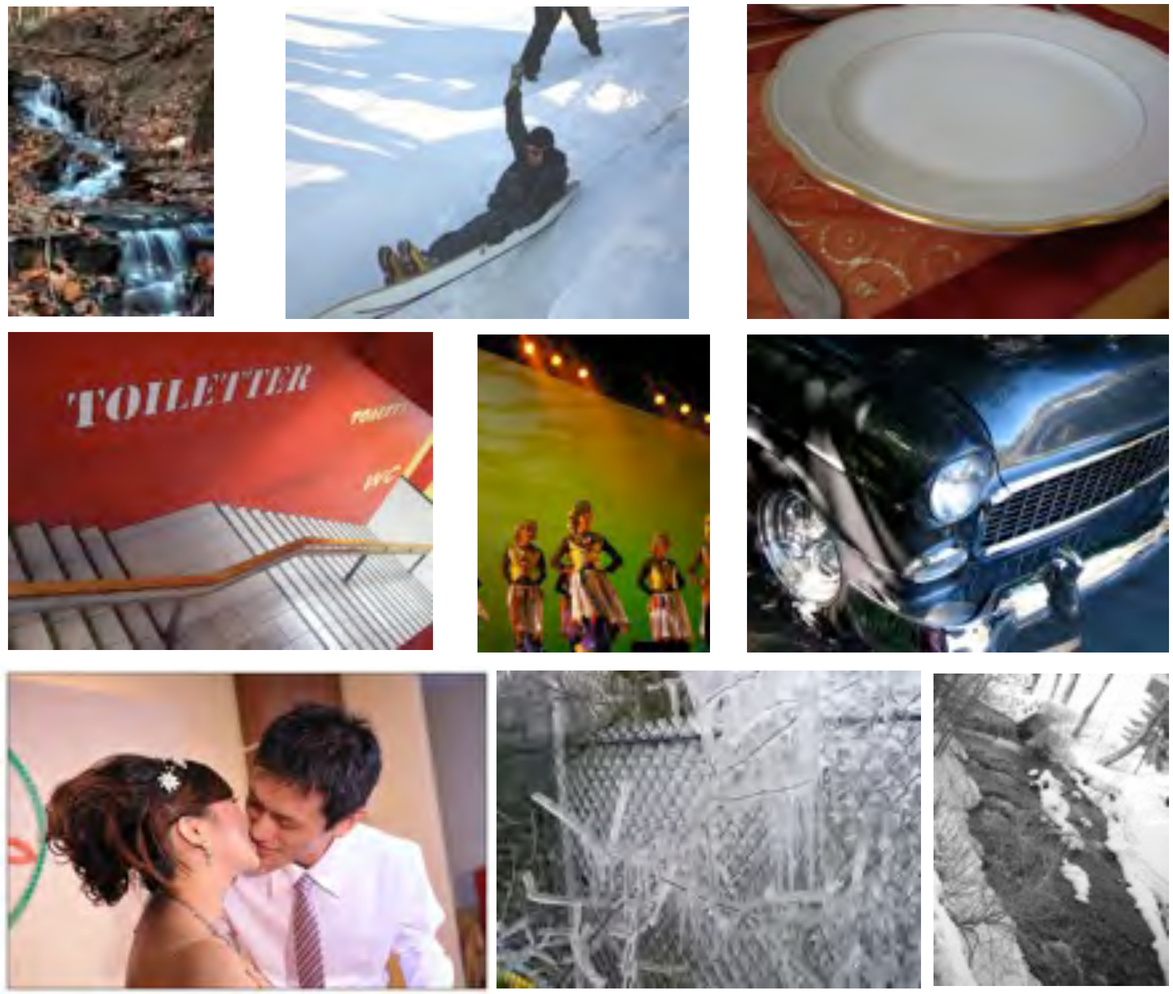
500 Synthetic transformations

388 Photographs [Nister 06]

Dataset 2: Pasadena Houses

Model set: ~10^5 photographs

125 pictures of LA houses



10^5 flickr photographs

Dataset 2: Pasadena Houses

Probe set: 625 pictures of Pasadena houses

Query 1

Model

Query 5

Different:

- viewpoint
- time of day
- camera

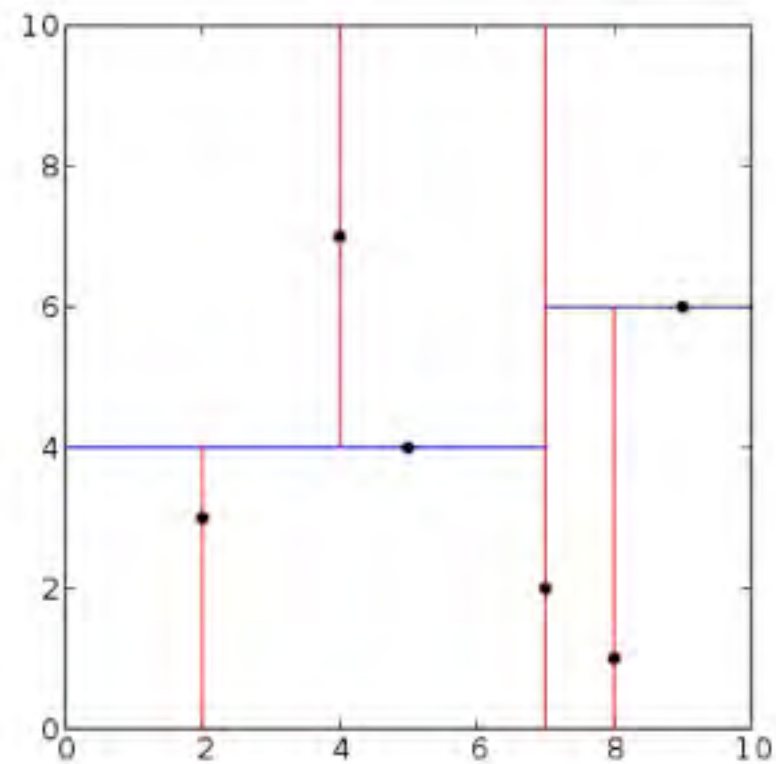
Dataset 2: Pasadena Houses

Recognition Approaches

- Sift/NN/Hough/RANSAC [Lowe '04]
- Sift/Quantize/Rank [VideoGoogle '03]

Nearest-Neighbor 1: Kd-tree

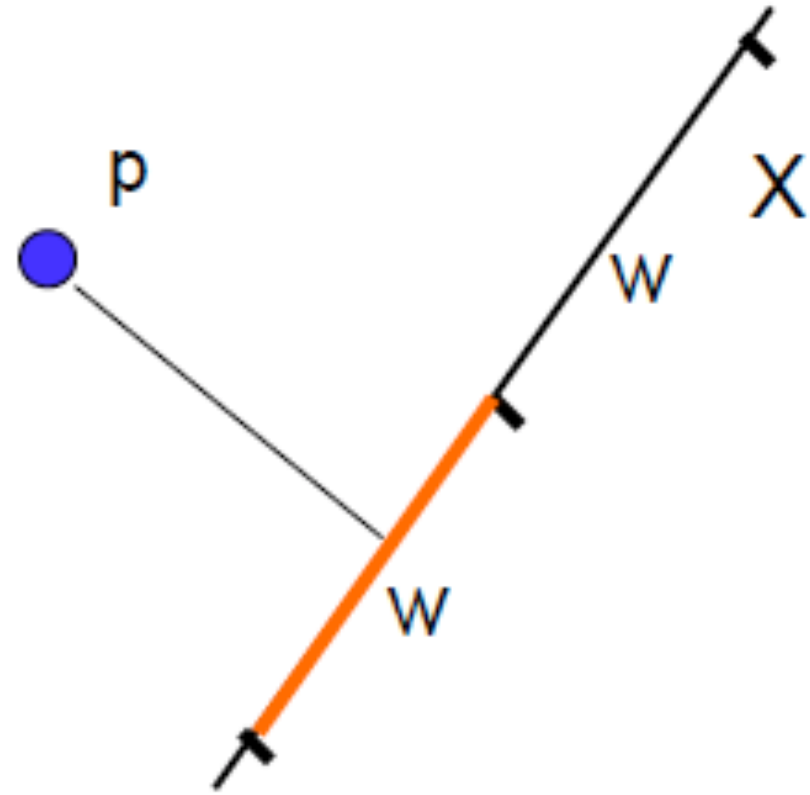
- Approximate
- Build *O(d N)*
- Search *O(log(N))*
- Multiple trees: Kd-forest



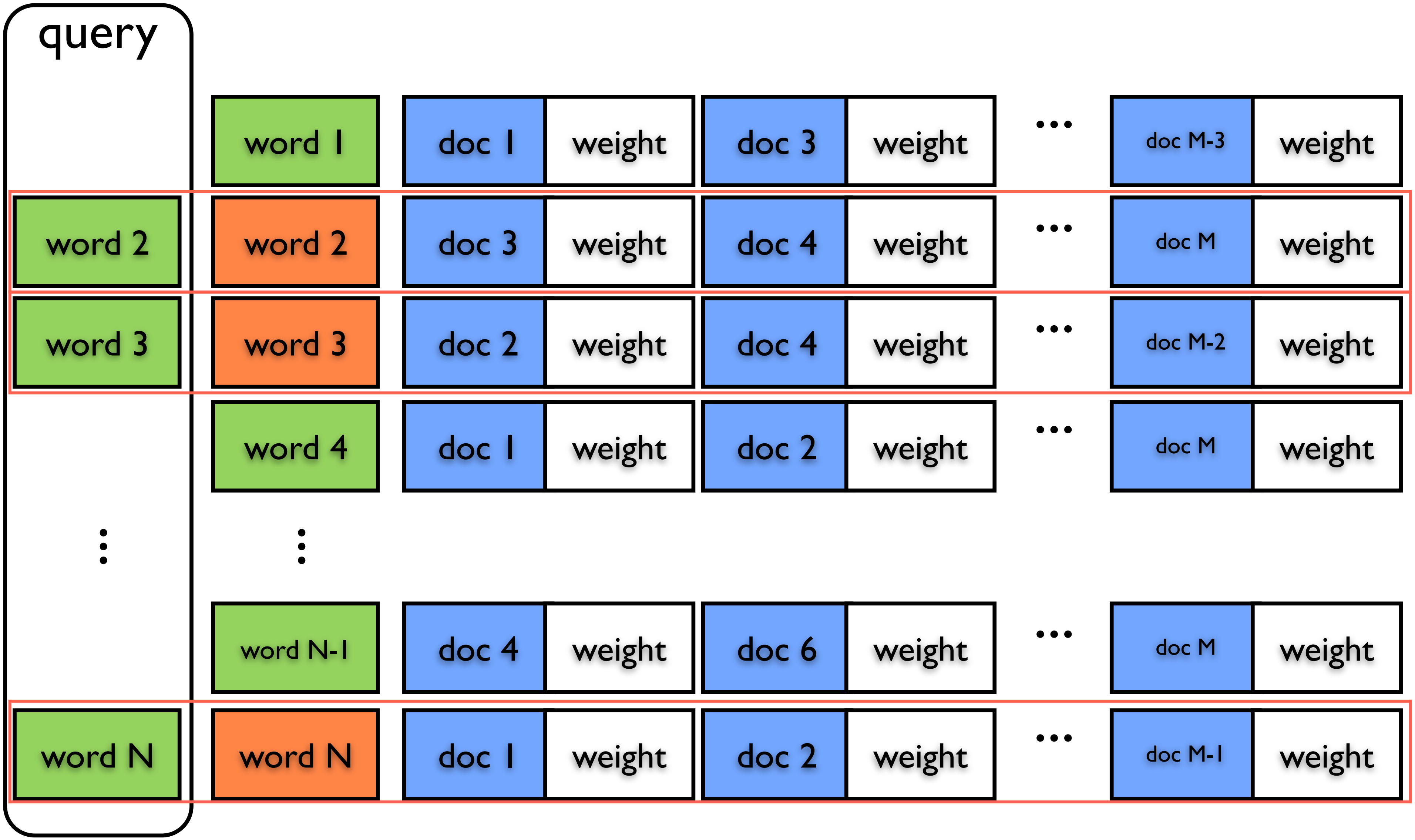
N ~ number of prototypes

Nearest-Neighbor 2: LSH

- E^2LSH package [Andoni 2004]
- Build *O(N)*
- Search *O(b) ~ O(N)*



Method 3: Bag-of-Words



[Sivic et al., Video Goggle '03]

Method 3: Bag-of-Words

- Extract Sift Features
- Quantize using Approximate K-means with Kd-forest *[Philbin et al. '07]*
- Compute word histograms *[Dorko-Schmid '03]*
- Search *O(N)*
- Fast search using Inverted File *[Sivic et al. '03]*

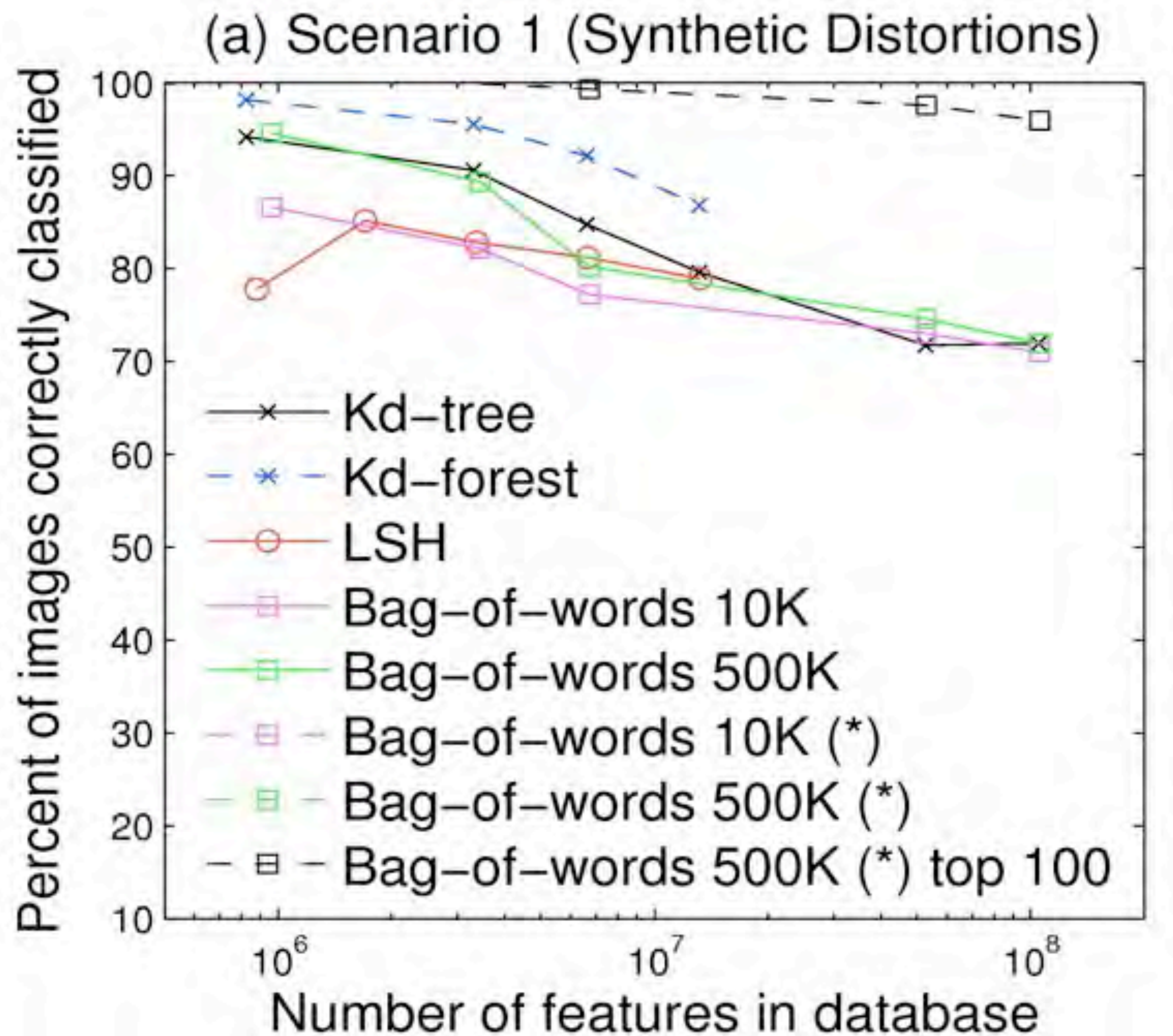
Experimental Setup

- Datasets:

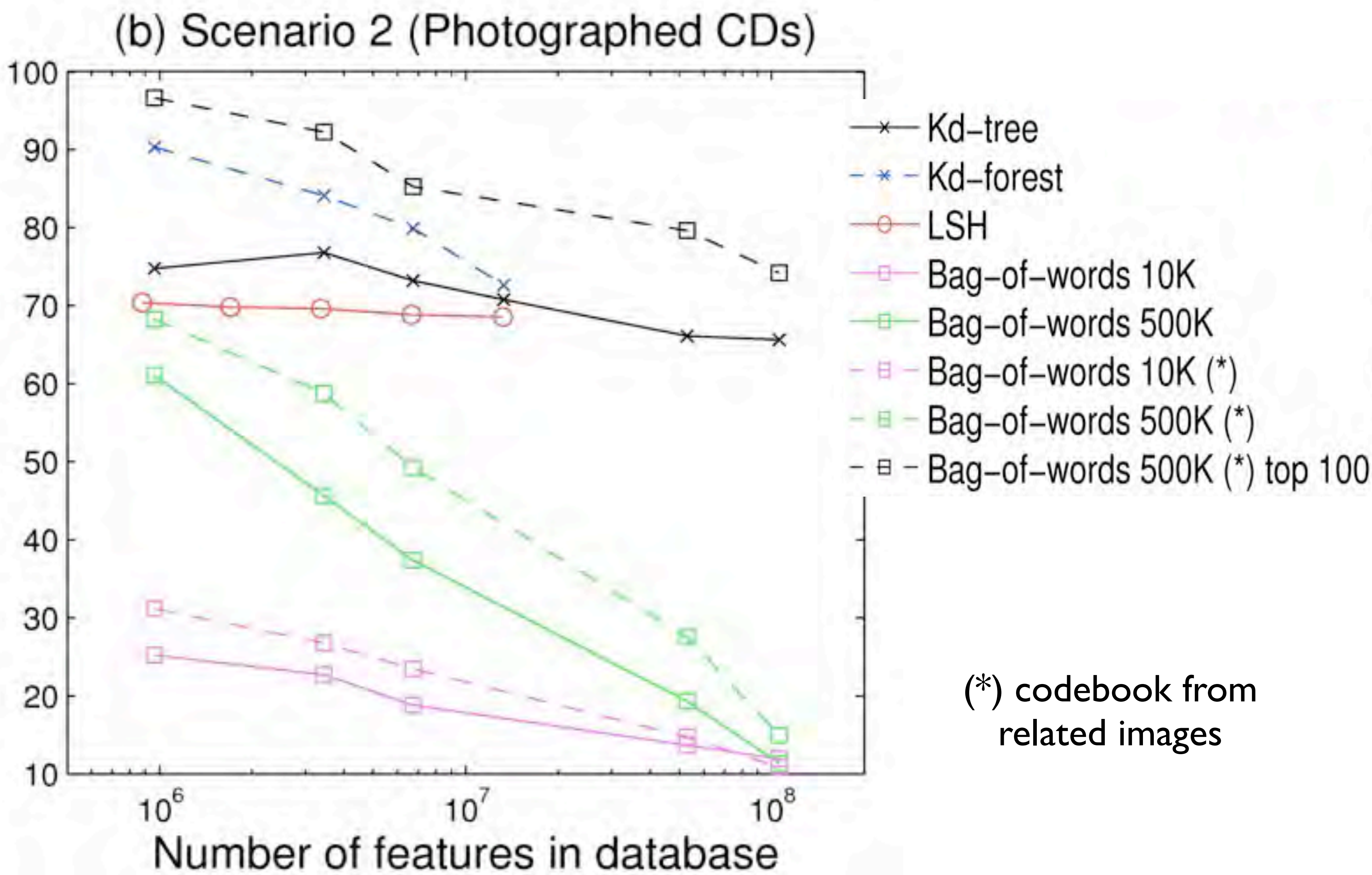
	Model Set	Probe Set	# images
Scenario 1	Covers	Synthetic	500
Scenario 2	Covers	Photographed	388
Scenario 3	Flickr	Houses	625

- One image/object in Model Set
- Rest in Probe Set
- Increase model set size: 1k, 4k, 8k, 16k, 32k, 64k, 128k images

Results: Recognition

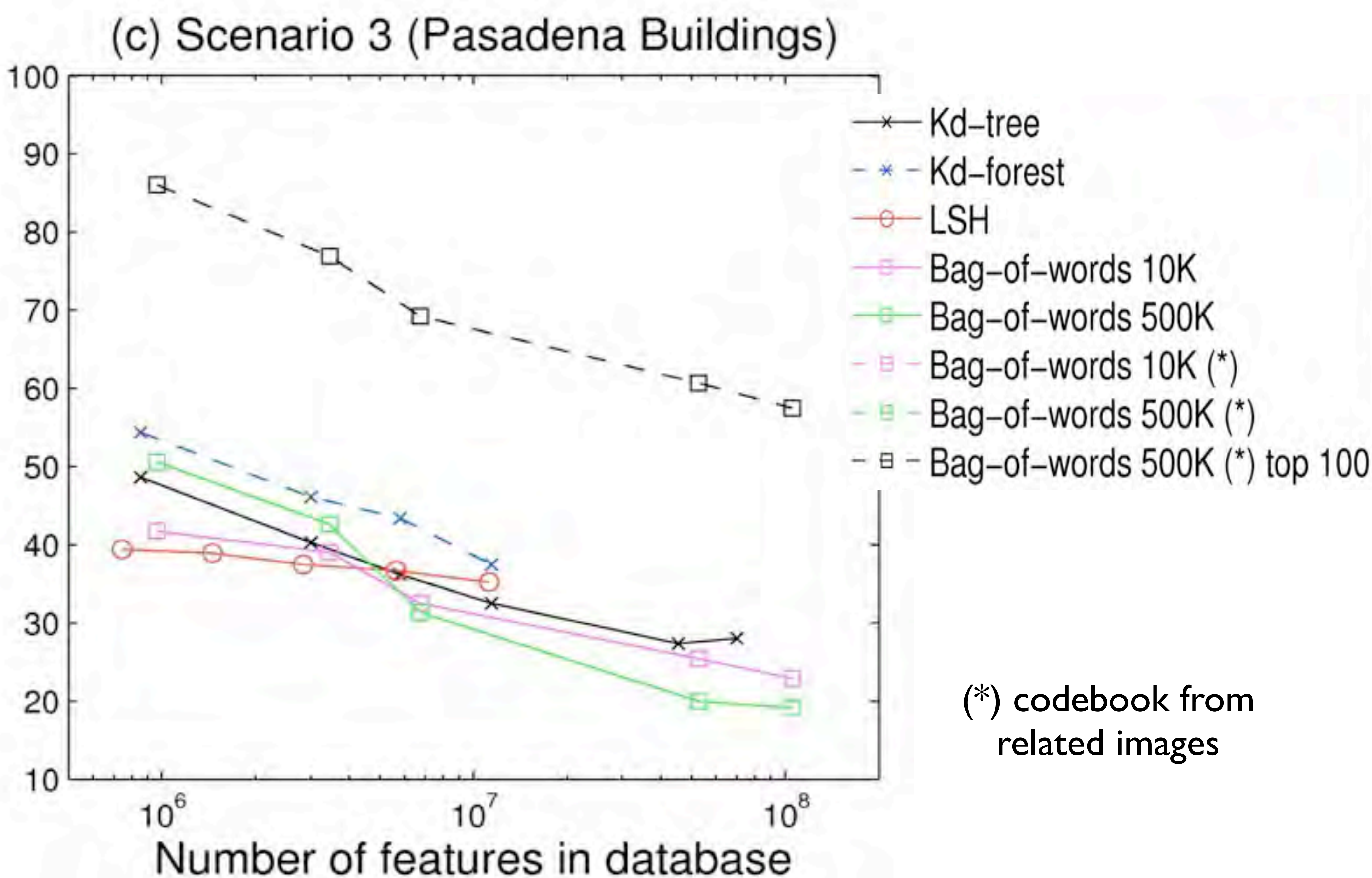


Results: Recognition



(*) codebook from related images

Results: Recognition

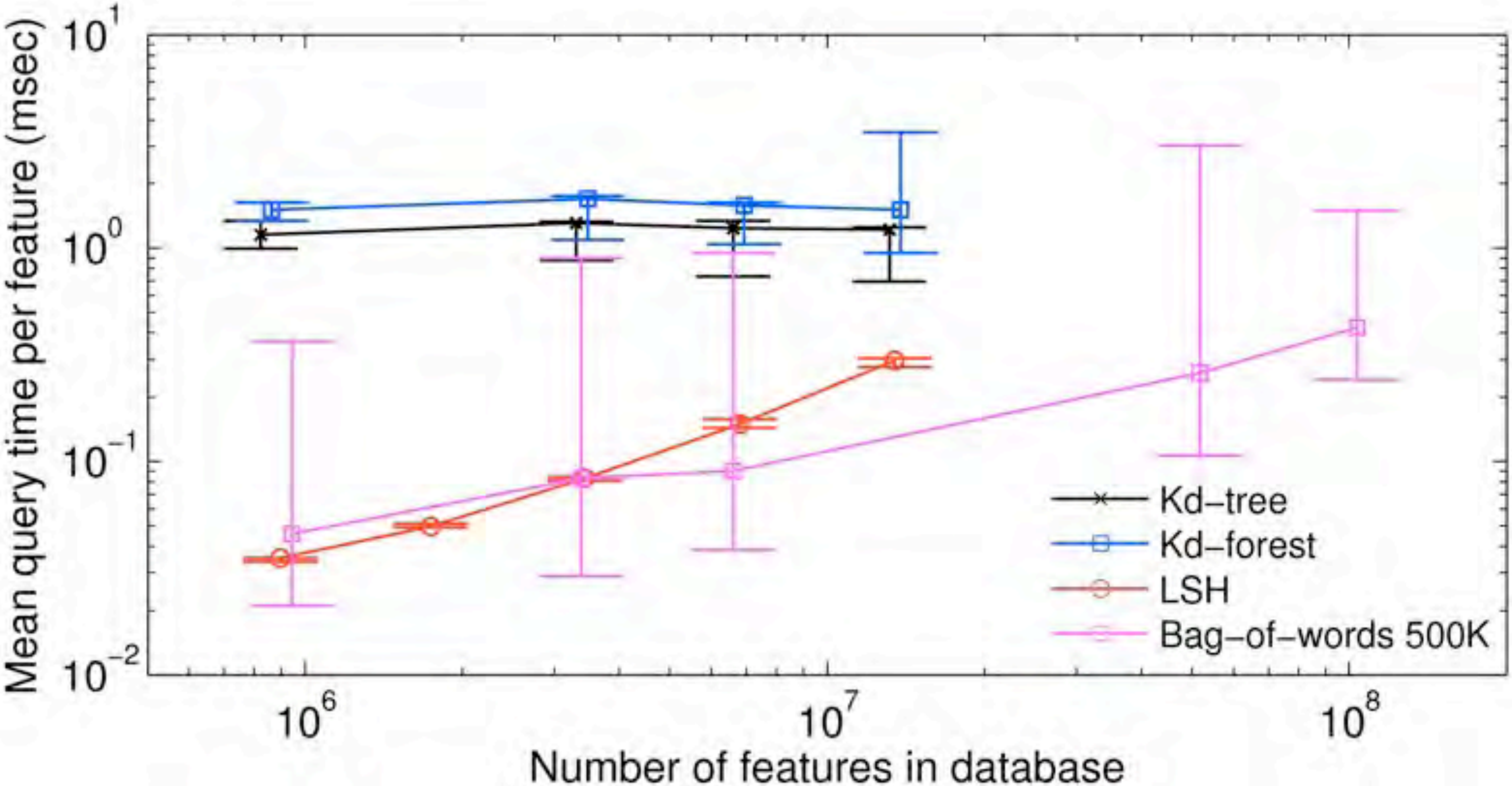


(*) codebook from related images

Conclusions I

- Synthetic distortions are useless
- Performance drops w.r. to n. of features
- LSH scales best
- KD forest scales OK
- Bag-of-words scales poorly

Results: Query Time



(Database: synthetic CD covers)

Conclusions 2

- LHS scales linearly (ouch)
- Bag-of-words scales like sqrt(N)
- KD forests have constant cost O(1)

Conclusions

- Importance of diverse datasets, natural probes
- Recall overall disappointing
- Nowhere close to 10^{10} images
- Bag-of-words recall does not scale well
- Kd-trees cost constant w.r.t. to N, unlike LSH
- Only bag-of-words fits in RAM beyond 10^5 images

- Much work still ahead of us!