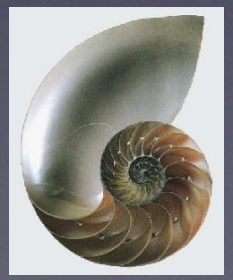
Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification

Andrea Frome
Yoram Singer
Fei Sha
Jitendra Malik

ICCV 2007



test image

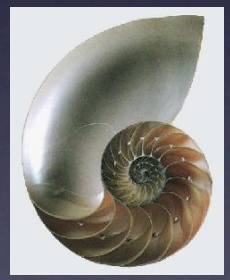
"dalmatian"

nearest neighbor

"dalmatian"

"dalmatian"

"buddha"



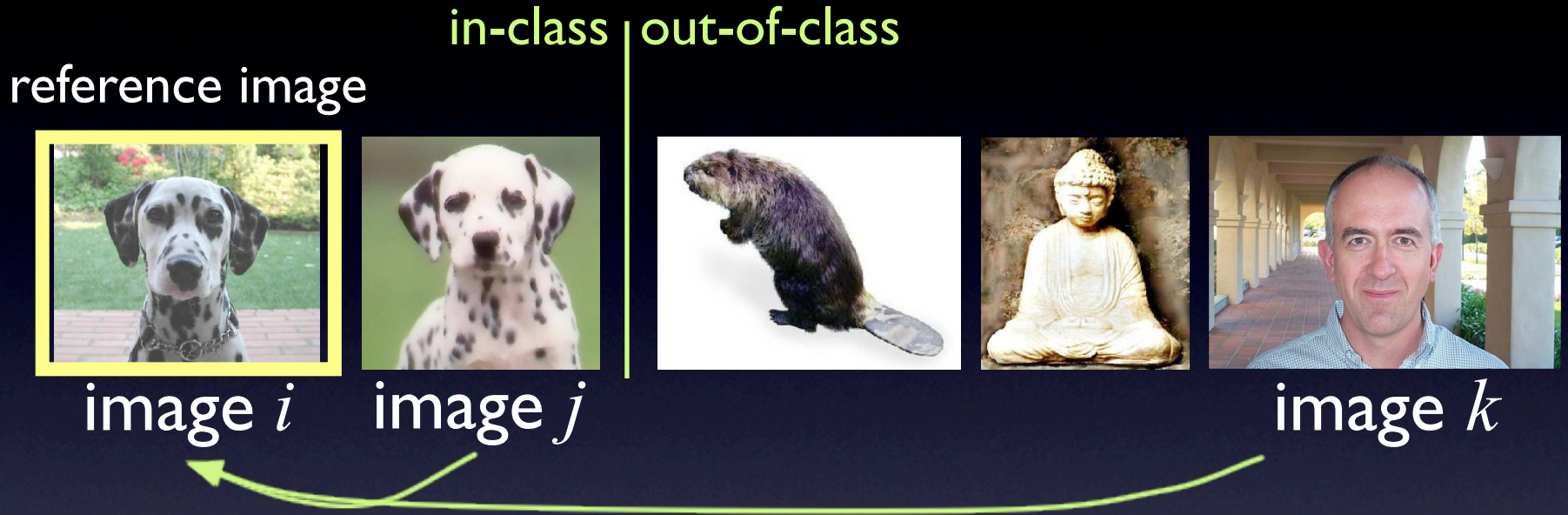
"nautilus"

"beaver"

e.g., NN, SVM

learning algorithm

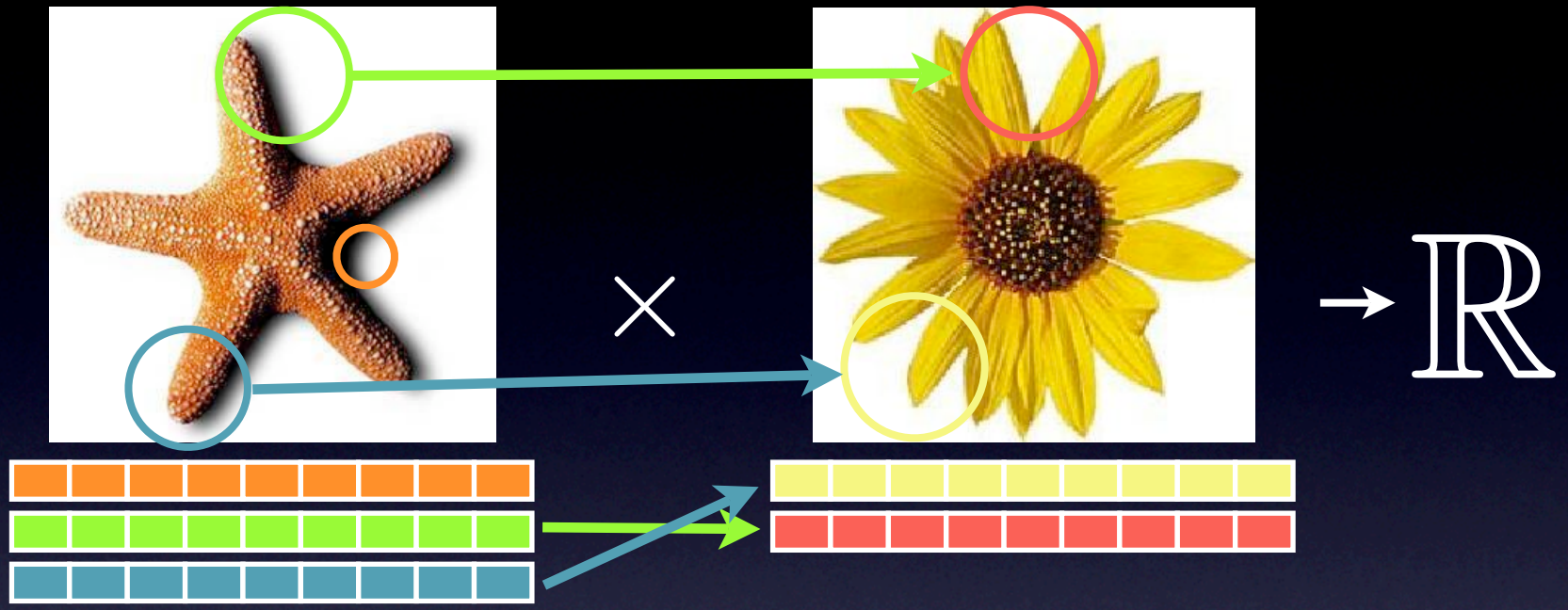
ranking: learn from triplets of training images



$$D(\text{image } k, \text{image } j) > D(\text{image } k, \text{image } i)$$

$$D_{ki} > D_{ji}$$

patch-based features



relaxations to correspondence

fast set matching (Grauman&Darrell 2006, Lazebnik,et al. 2006, Bosch,et al. 2007)

quantize feature space (bag of features) (Lazebnik, et al. 2006)

ignore spatial information (Grauman&Darrell 2006)

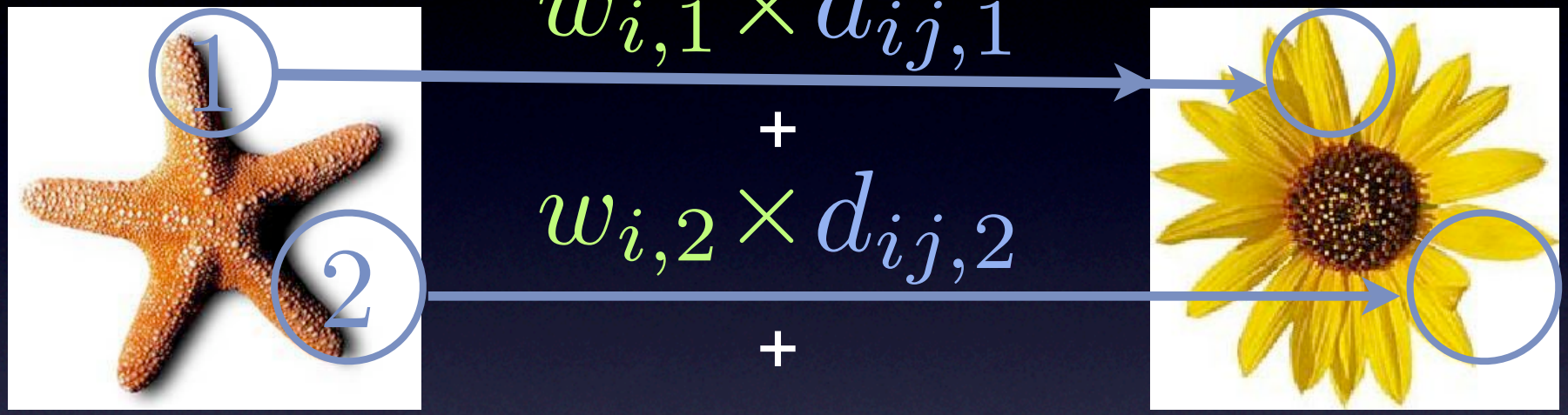
use absolute position information (Zhang,et al. 2006, Lazebnik, et al. 2006, Mutch&Lowe 2006)

D_{ij} : distance from image i to image j
(not symmetric)

image i

feature-to-image distance

image j



$$w_{i,1} \times d_{ij,1}$$

$$+$$

$$w_{i,2} \times d_{ij,2}$$

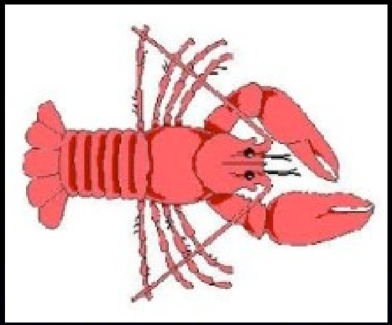
$$+$$

$$\ldots$$

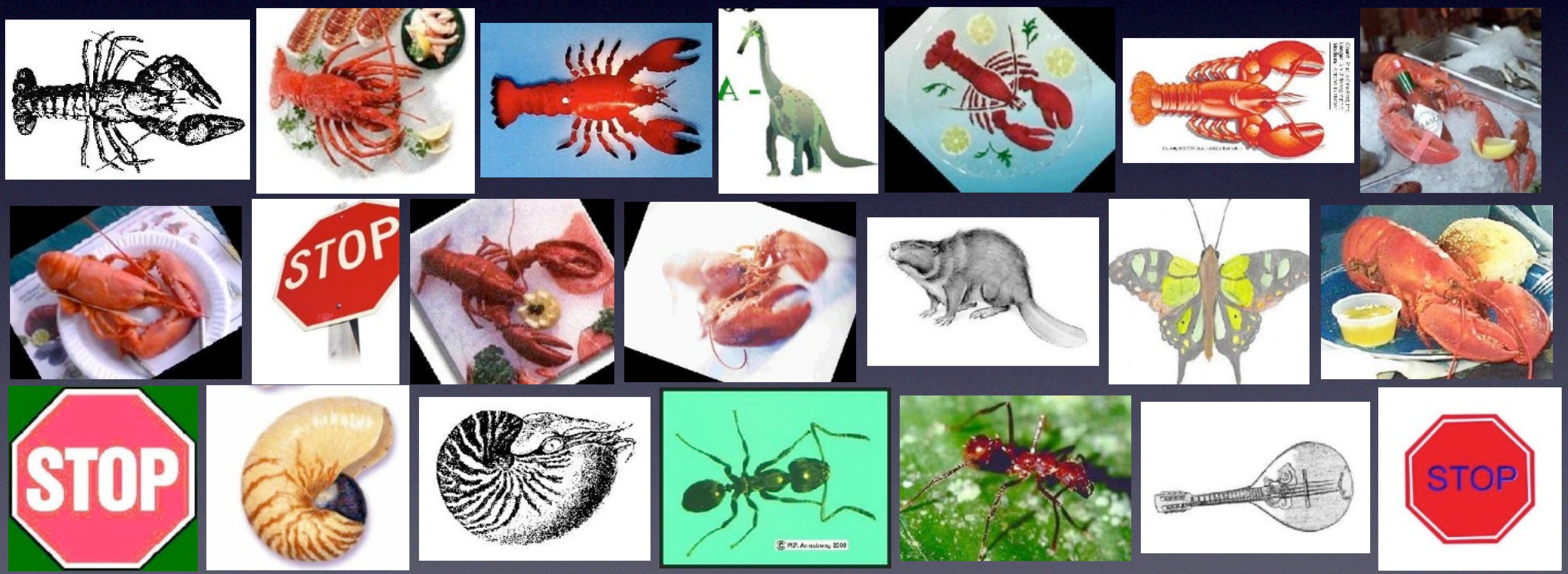
$$D_{ij} = \sum_{m=1}^{M} w_{i,m} d_{ij,m} = \mathbf{w}_i \cdot \mathbf{d}_{ij}$$

distance function can be evaluated from image i to any other image

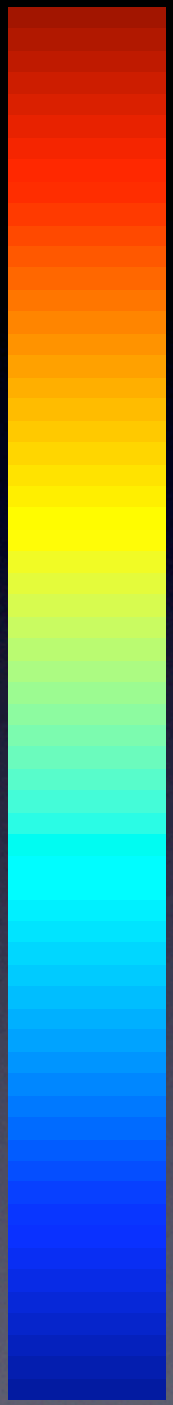
query image



retrieval results (weighted training images):



highest weight



lowest weight

why learn for every image?

clutter & occlusion

importance of a feature changes within a category

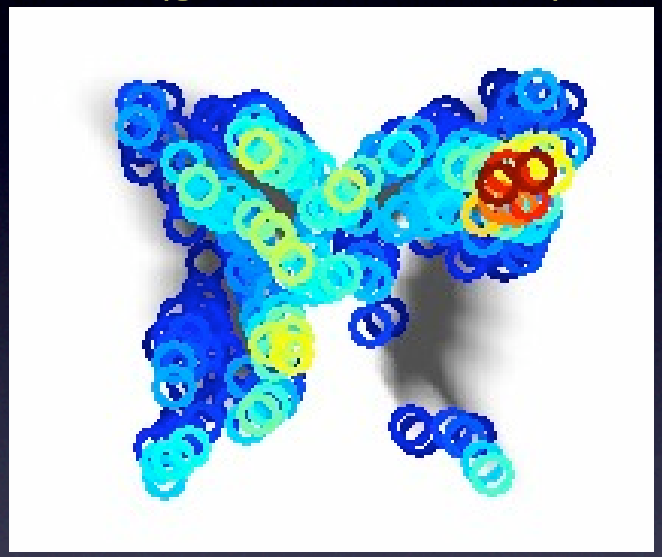
pose & articulation

large variation

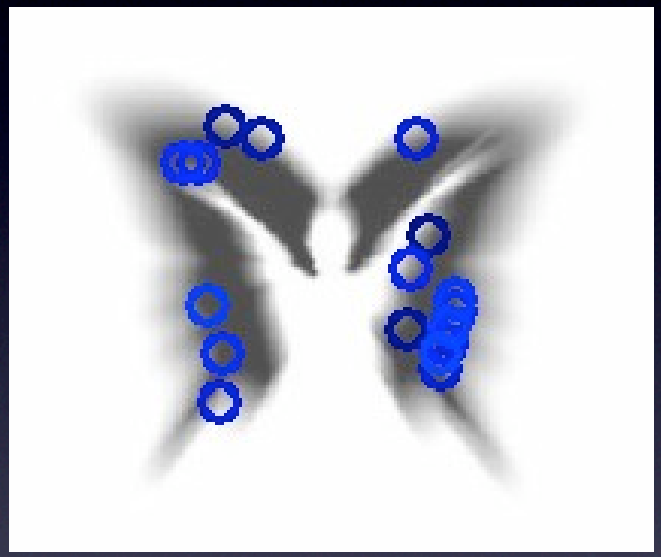
psychology: Rosch's family resemblances

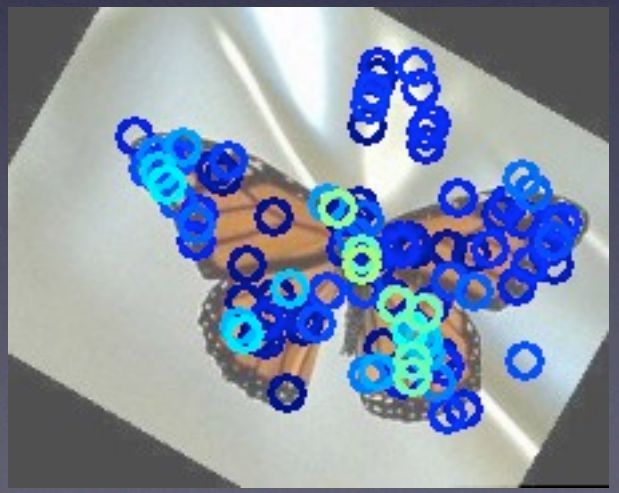
Combine features in a way appropriate to *each image*

large-extent shape feature
(geometric blur)



color



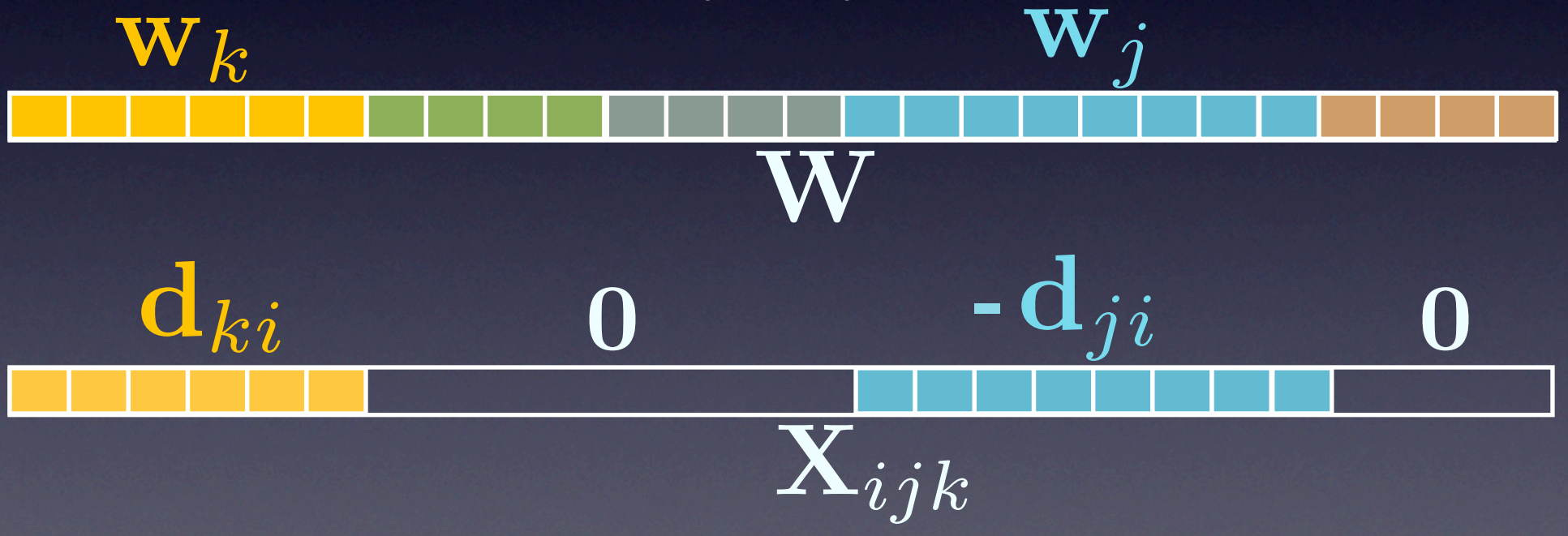


- mathematical formulation
- relationship to other distance learning approaches
- selection of triplets
- results

$$D_{ki} > D_{ji}$$

$$\mathbf{w}_k \cdot \mathbf{d}_{ki} > \mathbf{w}_j \cdot \mathbf{d}_{ji}$$

$$\mathbf{w}_k \cdot \mathbf{d}_{ki} - \mathbf{w}_j \cdot \mathbf{d}_{ji} > 0$$



$$\mathbf{W} \cdot \mathbf{X}_{ijk} > 0$$

$$\mathbf{W} \cdot \mathbf{X}_{ijk} > 0$$
$$\mathbf{W} \cdot \mathbf{X}_{ijk} \geq 1$$

empirical loss: $$\sum_{i,j,k \in \text{triplets}} [1 - \mathbf{W} \cdot \mathbf{X}_{ijk}]_+$$

$$\begin{aligned} \min_{\mathbf{W},\xi} \quad & \frac{1}{2}\|\mathbf{W}\|^2 + C \sum_{ijk} \xi_{ijk} \\ \text{s.t.} \quad & \mathbf{W} \cdot \mathbf{X}_{ijk} \geq 1 - \xi_{ijk} \\ & \xi_{ijk} \geq 0 \\ & \mathbf{W} \succeq 0 \end{aligned}$$

Schultz,Joachims NIPS 2003
Frome,Singer,Malik NIPS 2006

problem scale
(15 images/category, 101 categories)

~1,200 features/image: weight vector has 1.8M elements

using in- vs. out-of-class,

exhaustive set of triplets is 31.8 M triplets

speeding it up

pare down to 15.7 M triplets

solve the dual problem

similar to on-line algorithms

early stopping: 10 hours to 1 hour

set trade-off parameter: one run through triplets

weight vectors are surprisingly sparse.

on average, 68% of weights are zero

Selecting triplets: 15 images/category

select 15.7 M out of 31.8 M triplets
many are easy, some are too hard

easy triplet

"reference"

hard triplet

"reference"

Heuristic using independent feature-to-image distances.

Relationship to other distance learning work

Zhang,Malik (CVPR 2003)
Bosch,Zisserman,Munoz (CVPR 2007)

learn a distance function for all images (global)

per category

one per image (local)

Xing,Ng,Jordan,Russell (NIPS 2002)
Schultz,Joachims (NIPS 2003)
Shalev-Shwartz,Singer,Ng (ICML 2004)
Weinberger,Blitzer,Saul (NIPS 2005)
Globerson,Roweis (NIPS 2005)
Grangier,Monay,Bengio (ECML 2006)
Grauman,Darrell (NIPS 2006)
Varma,Ray (ICCV 2007)

Frome,Singer,Malik (NIPS 2006)
Frome,Singer,Sha,Malik (ICCV 2007)

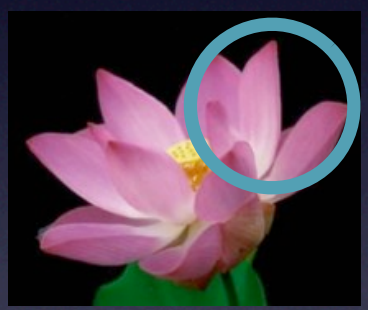
exploit collection of partial descriptors (patch-based features)

experiments

Caltech-101 (without using absolute position)

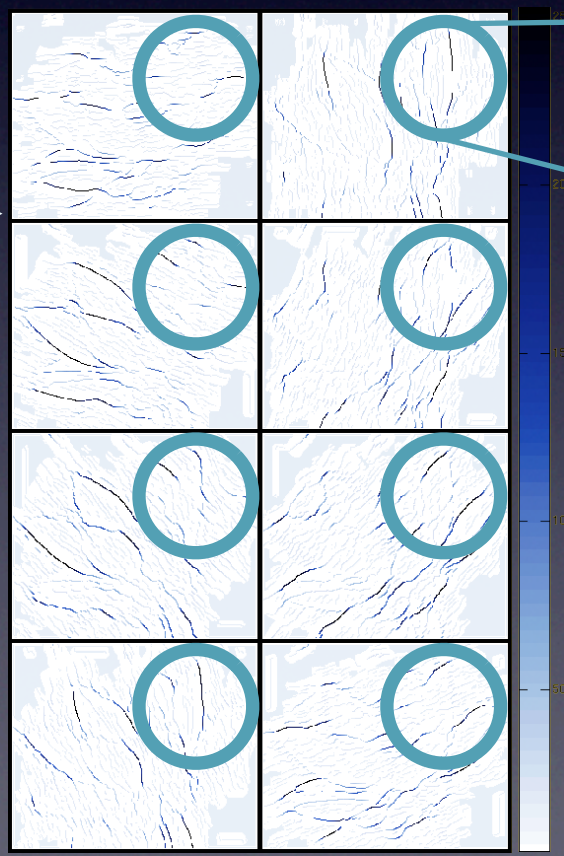
features: geometric blur (2 sizes) and color

L_2 feature-to-image distance

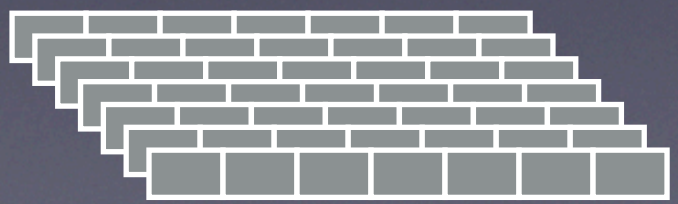


42 & 70 pixel radius, 4 channels

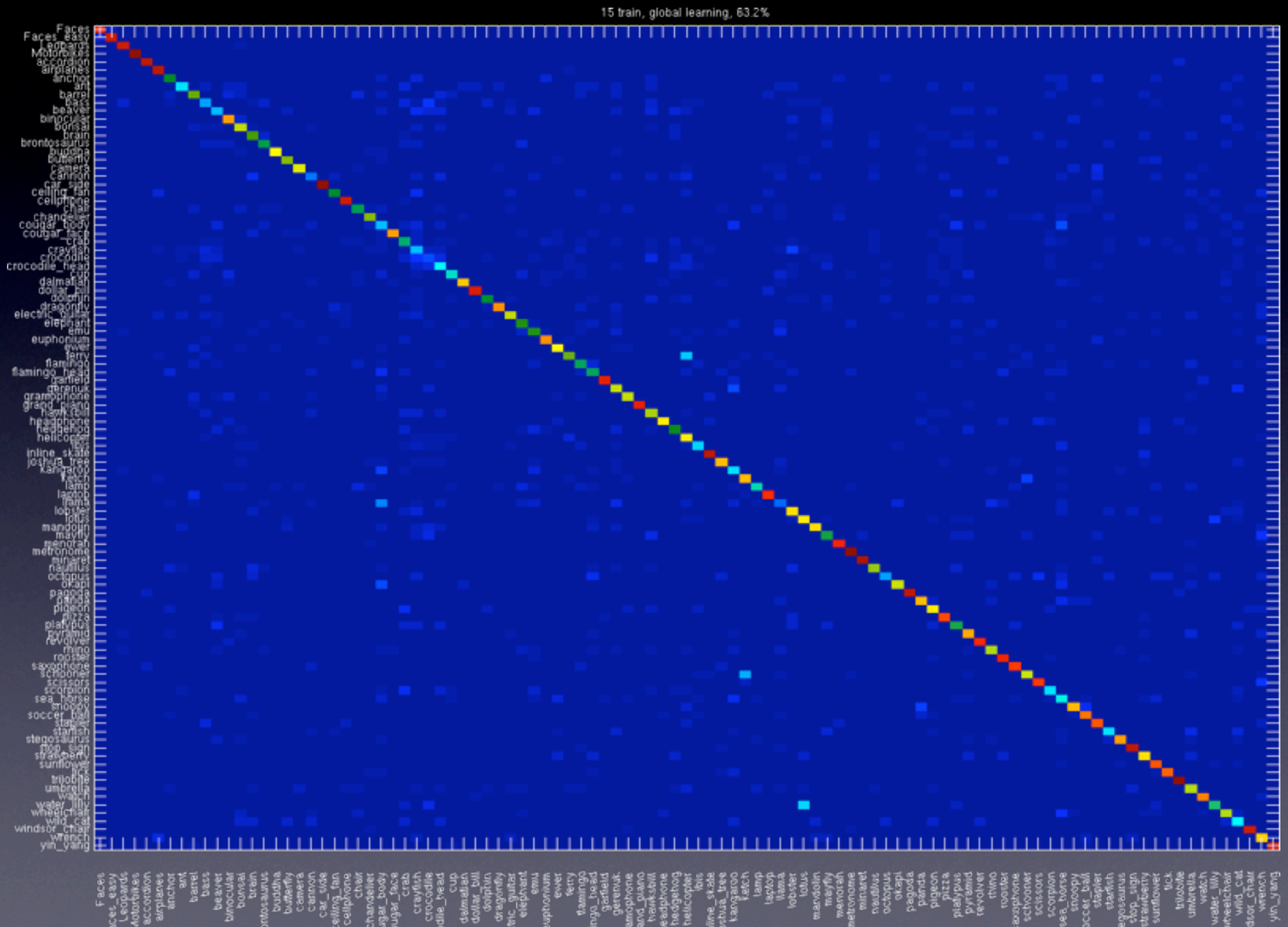
Berg,Berg,Malik CVPR 2005



geometrically blur & sample

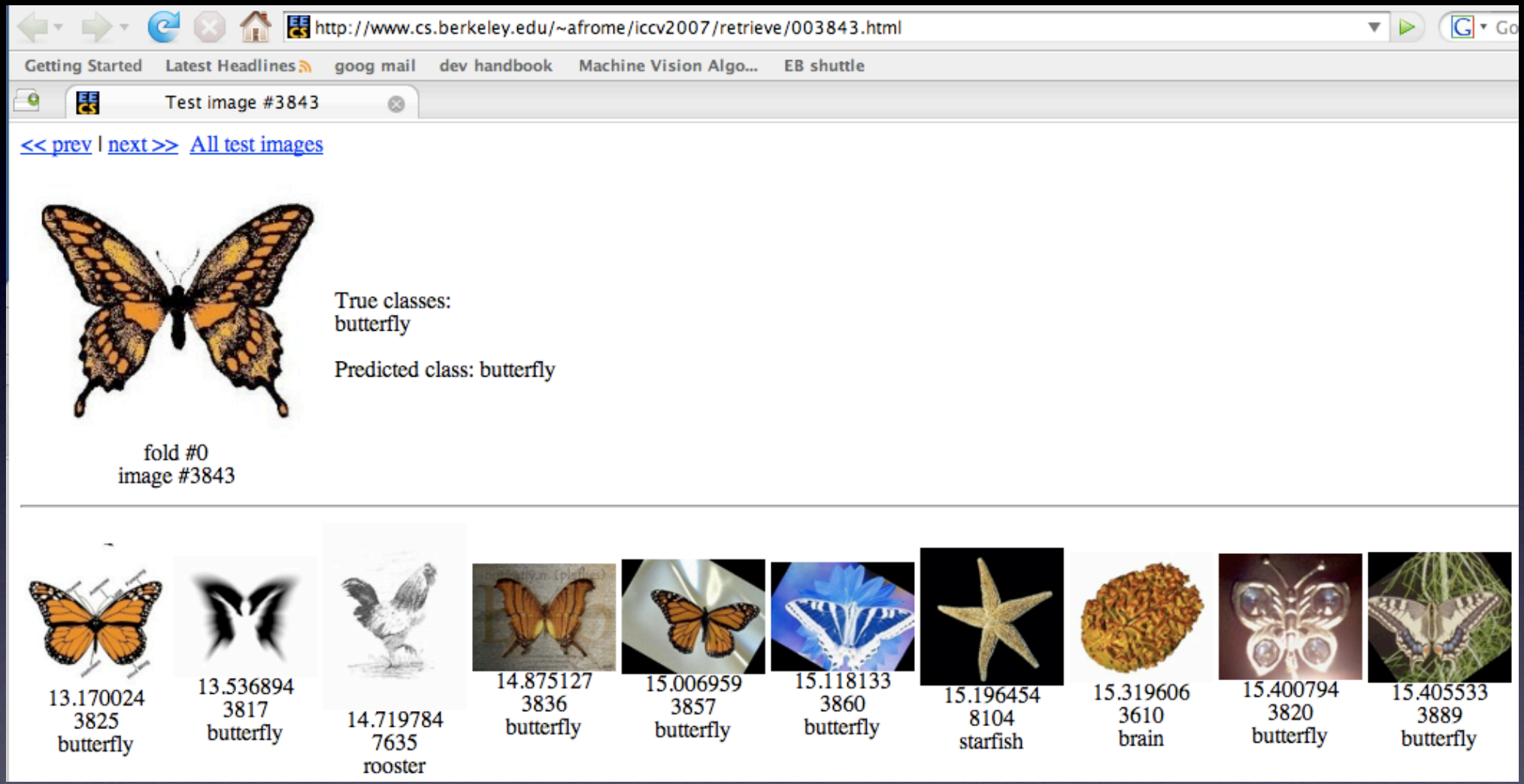


confusion, 15 images/class (63.2%)




training examples per class
mean recognition rate
0
5
10
15
20
25
30
35
70
65
60
55
50
45
40
35
30
25
20
15
10
Frome,Singer,Sha,Malik, ICCV 2007
Griffin,Holub,Perona, Tech Report, 2007
Frome,Singer,Malik, NIPS 2006
Zhang, et al., CVPR 2006
Lazebnik, et al., CVPR 2006
Berg, PhD Thesis 2005
Mutch & Lowe, CVPR 2006
Grauman & Darrell, Tech Report, Mar 2006
Berg, et al., CVPR 2005
Wang, et al., CVPR 2006
Holub, et al., ICCV 2005
Serre, et al., CVPR 2005
Fei–Fei, et al., 2004

http://www.cs.berkeley.edu/~afrome/iccv2007



thank you.