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

Andrea Frome, Yoram Singer, Fei Sha,
Jitendra Malik

ICCV 2007

Presented by Roie Kliper

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

Andrea Frome
EECS, UC Berkeley
andrea.frome@gmail.com

Yoram Singer
Google, Inc.
singer@google.com

Fei Sha
EECS, UC Berkeley
feisha@cs.berkeley.edu

Jitendra Malik
EECS, UC Berkeley
malik@cs.berkeley.edu

Abstract

We address the problem of visual category recognition by learning an image-to-image distance function that attempts to satisfy the following property: the distance between images from the same category should be less than the distance between images from different categories. We use patch-based feature vectors common in object recognition work as a basis for our image-to-image distance functions. Our large-margin formulation for learning the distance functions is similar to formulations used in the machine learning literature on distance metric learning, however we differ in that we learn local distance functions—a different parameterized function for every image of our training set—whereas typically a single global distance function is learned. This was a novel approach first introduced in Frome, Singer, & Malik, NIPS 2006. In that work we learned the local distance functions independently, and the outputs of these functions could not be compared at test time without the use of additional heuristics or training. Here we introduce a different approach that has the advantage that it learns distance functions that are globally consistent in that they can be directly compared for purposes of retrieval and classification. The output of the learning algorithm are weights assigned to the image features, which is intuitively appealing in the computer vision setting: some features are more salient than others, and which are more salient depends on the category, or image, being considered. We train and test using the Caltech 101 object recognition benchmark. Using fifteen training images per category, we achieved a mean recognition rate of 63.2% and using twenty images per category, a rate of 66.6%.

image *j* image *i*

D_{ji}

1. Introduction

Outline

- What are Distance Functions and what are the good for?
- Global vs. Local Distance Functions.
- Algorithm Outline & Details...
- Results
- Some conclusions

- (Images in this presentation are Taken from Frome's Thesis 2007)

Distance Functions

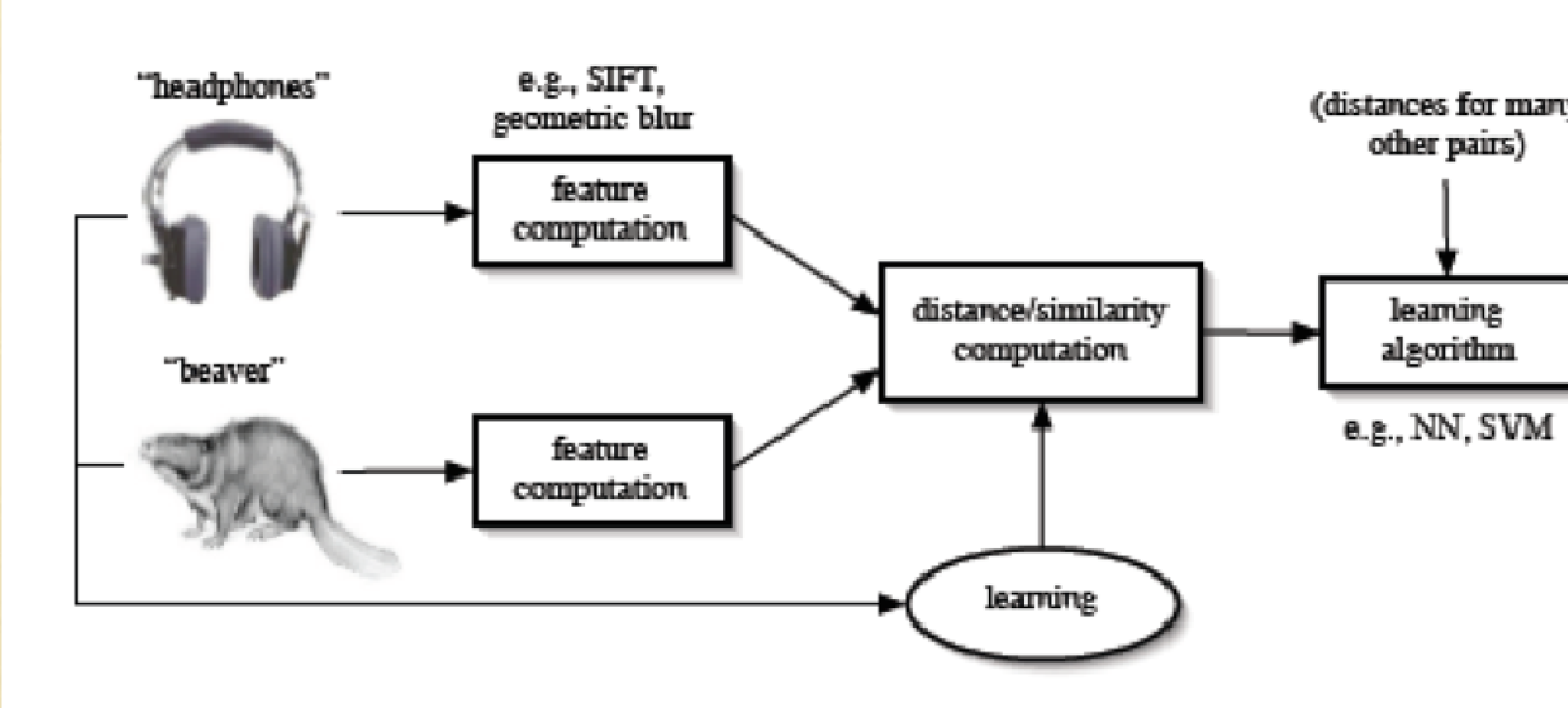
- A distance function is a function defined over pairs of data points. $D : X \times X \rightarrow \Re$
- A distance metric is a mapping to the non-negative real number $D : X \times X \rightarrow \Re^{+}$,that also obeys the following properties:
  - Isolation.
  - Symmetry.
  - Triangular Inequality.

- Similarity function is inversely related to a distance functio
- There are several intuitive ways of transforming a similarity function into a distance function and vice-versa.
- One important and widely used type of similarity functions are kernel functions.
  - A kernel function K can be used to define a Gram matrix (also known as a kernel matrix).
  - The Gram matrix is in essence a similarity matrix of the set of vectors S.
  - It also has several other appealing properties: it is symmetric and positive semi-definite. (In fact it can be shown that any positive semi-definite symmetric matrix corresponds to some kernel function k.)

important?

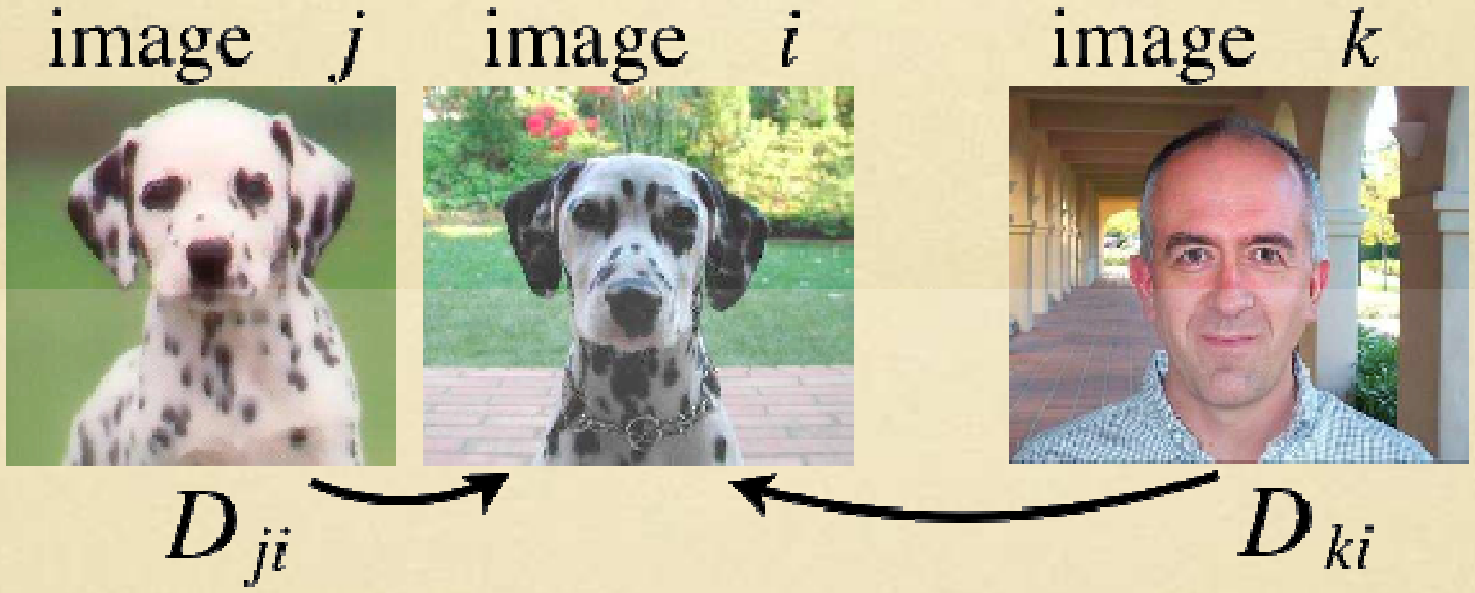
- A distance functions defines a topology over the space.
- Relieves interesting manifolds (Manifold learning)
- Many Distance-based algorithms:
  - NN classification
  - Radial basis function network
  - K-means clustering,
  - Various graph based methods (linkage clustering)
  - kernel methods.
- Many others....

Categorization scenario



General Idea:

- Distance between images from the same category should be less than distance between images from different categories.



Global vs Local

- Global: (Ambitious) In general try to capture the global dimensions that are responsible to the data categorization: $D : X \times X \rightarrow \Re$
  - e.g: Mahalanobis
- Local: (modest) Learn a distance that is specific for a given focal point (image). $D_a : X \rightarrow \Re$
- Make it globally consistent....

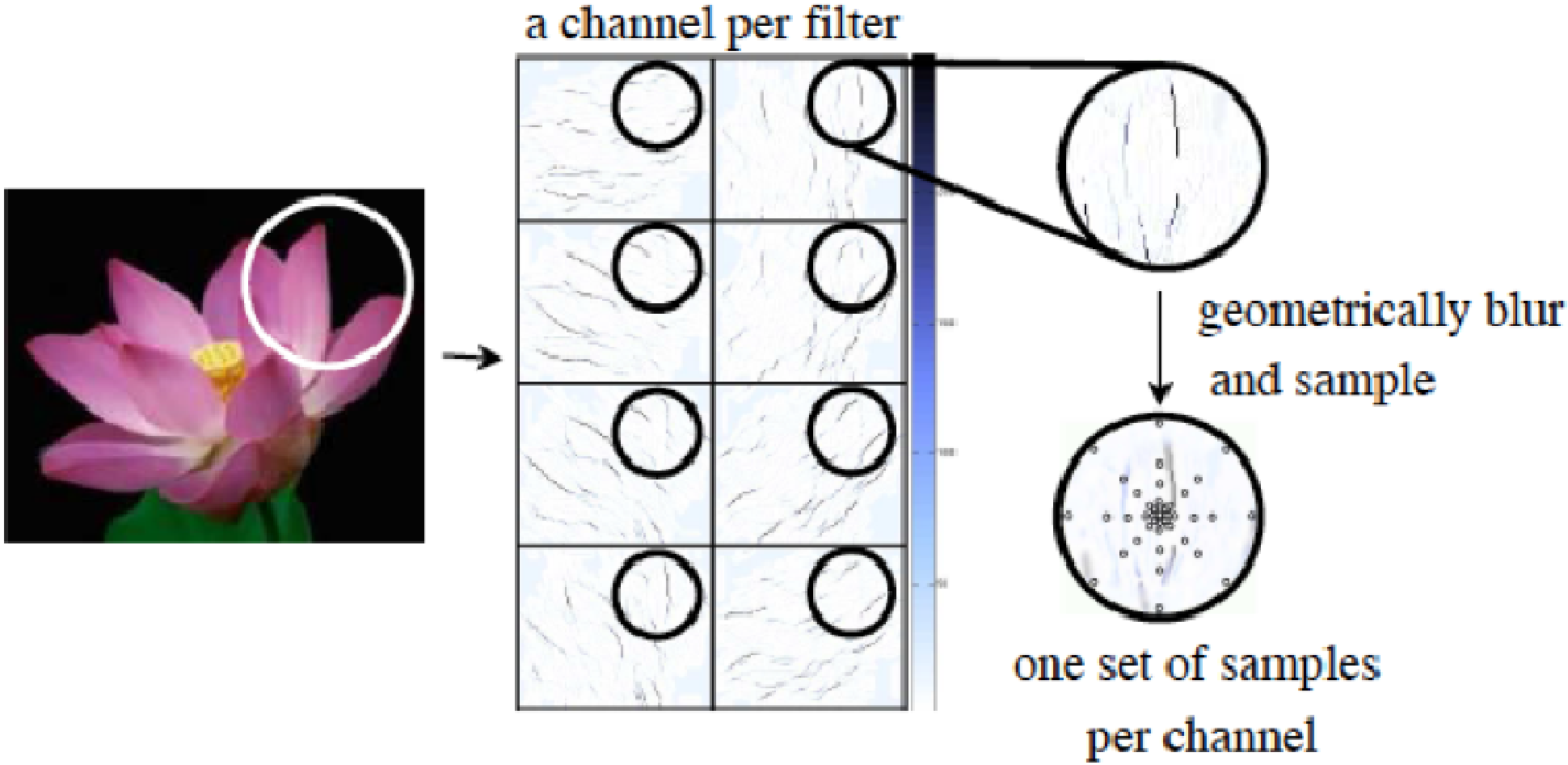
Algorithm Outline

- Input: patch-based feature vectors such as SIFT or geometric blur. (Using geometric blur and simple color features)
- Parametrize the image-to-image distance functions using a weighted linear combination of distances between patch based shape feature descriptors.
- Generate Triplet based constraints.
- Large margin optimization formulation
- Solve (of the shelf solver)
- Output:
  - weights assigned to the image features
  - Local distance function which are globally consistent (thus can be compared)

descriptors extraction and use:

- SIFT, Geometric Blur
  - Robust under affine transformations.
  - These features characterize image patches by fixed length vectors, which can be compared using L1 or L2 Norms.
- Different features can be combined into a sing bag of feature.

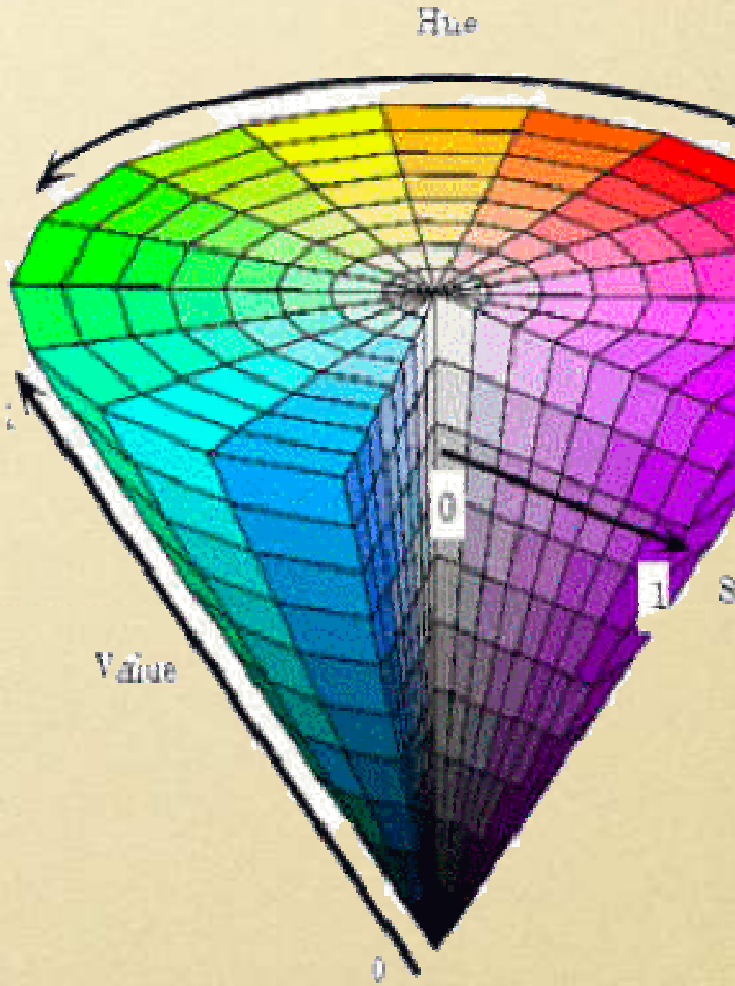
Geometric blur



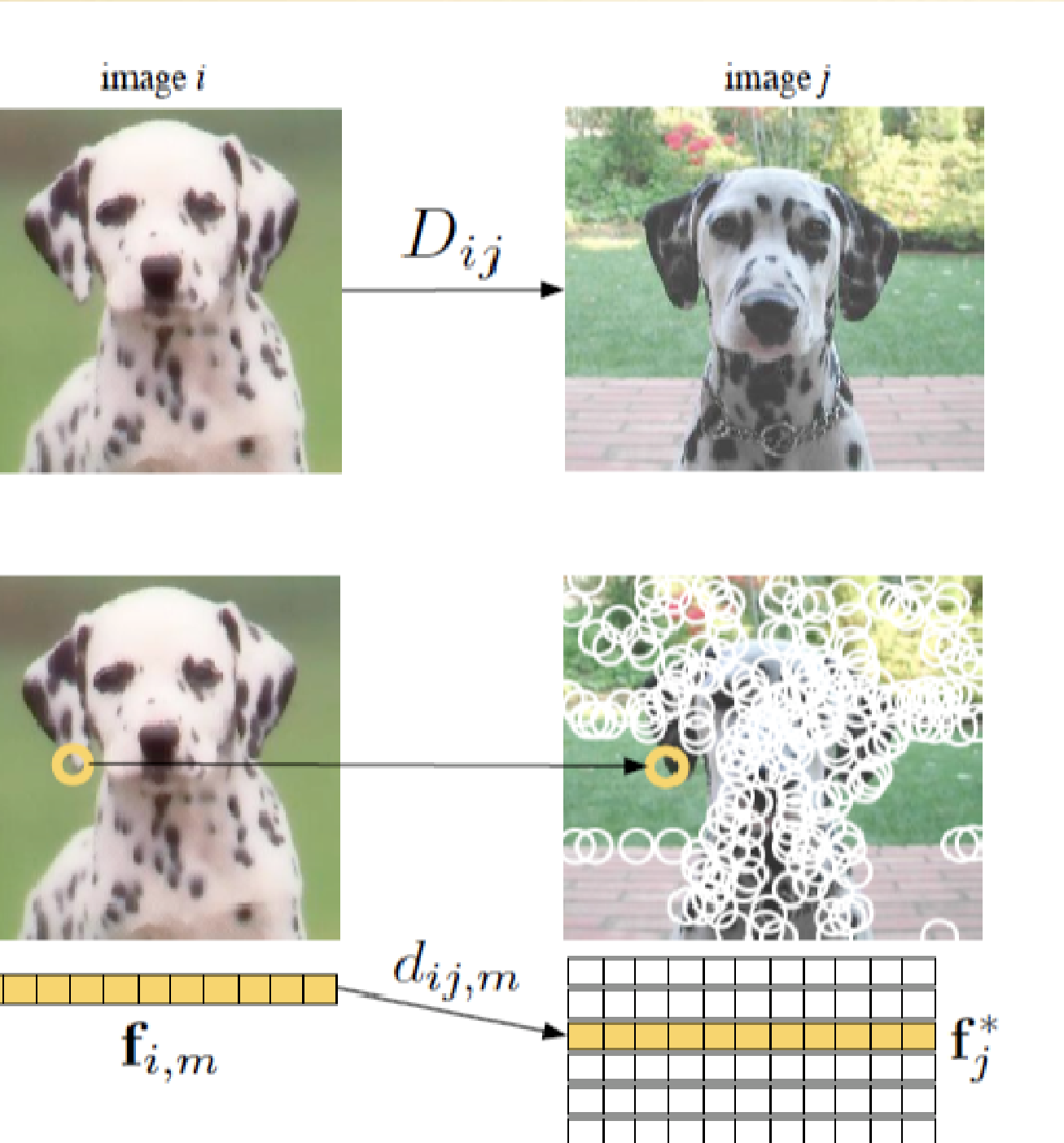
- Berg & Malik, CVPR 01

Color Features

- Color features are histograms of eight-pixel radius patches also centered at edge pixels in the image.
- Any “pixels” in a patch off the edge of the image are counted in a “undefined” bin, and we convert the HSV coordinates of the remaining points to a Cartesian space where the z direction is value and (x, y) is the Cartesian projection of the hue/saturation dimensions.
- Divide the (x, y) space into an 11 °— 11 grid, and make three divisions in the z direction.

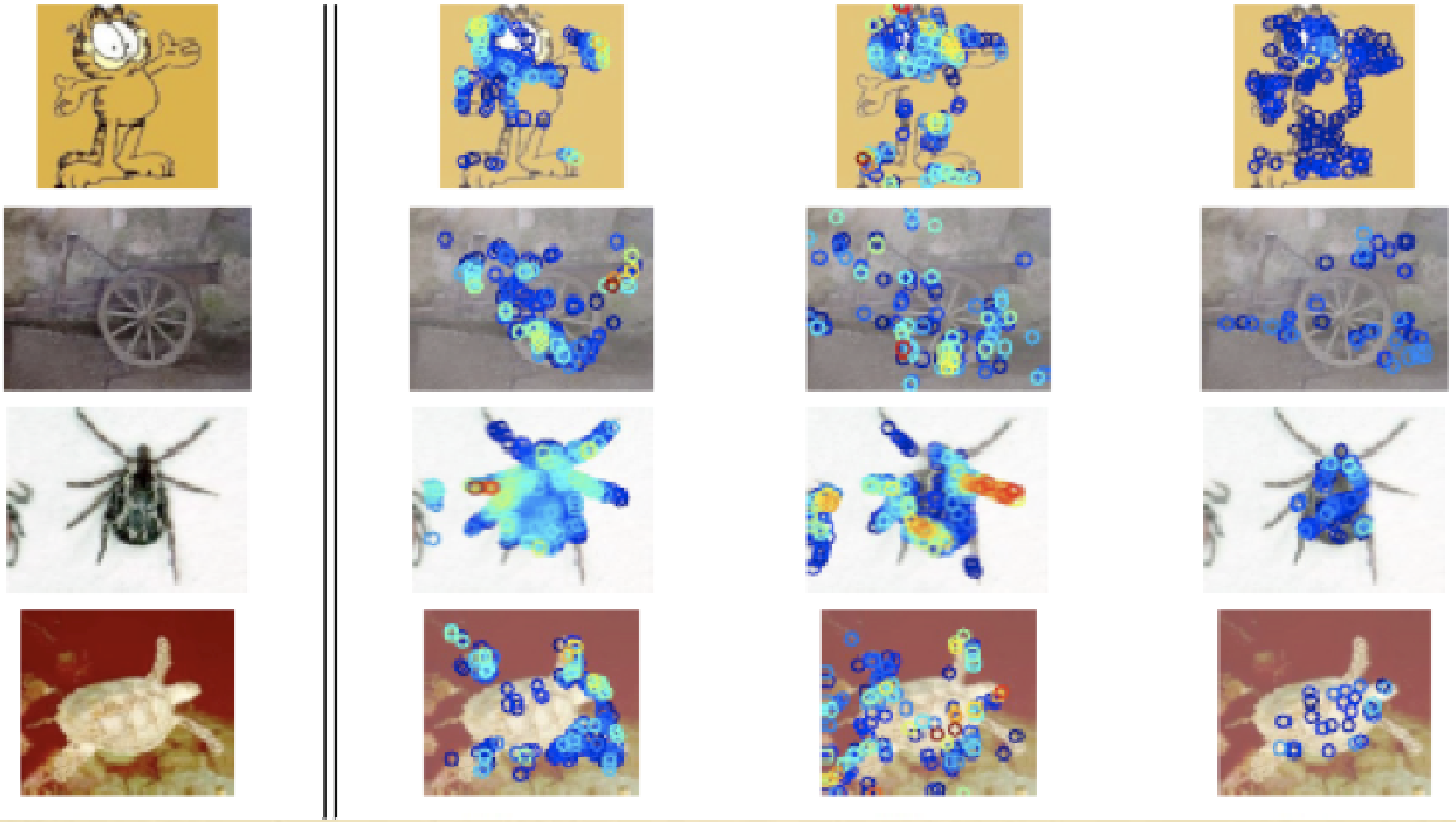


Feature to image distances



$$d_{ij,m} = \min_{k} \| \mathbf{f}_{i,m} - \mathbf{f}_{j,k} \|$$

Combining feature types



A Distance Function

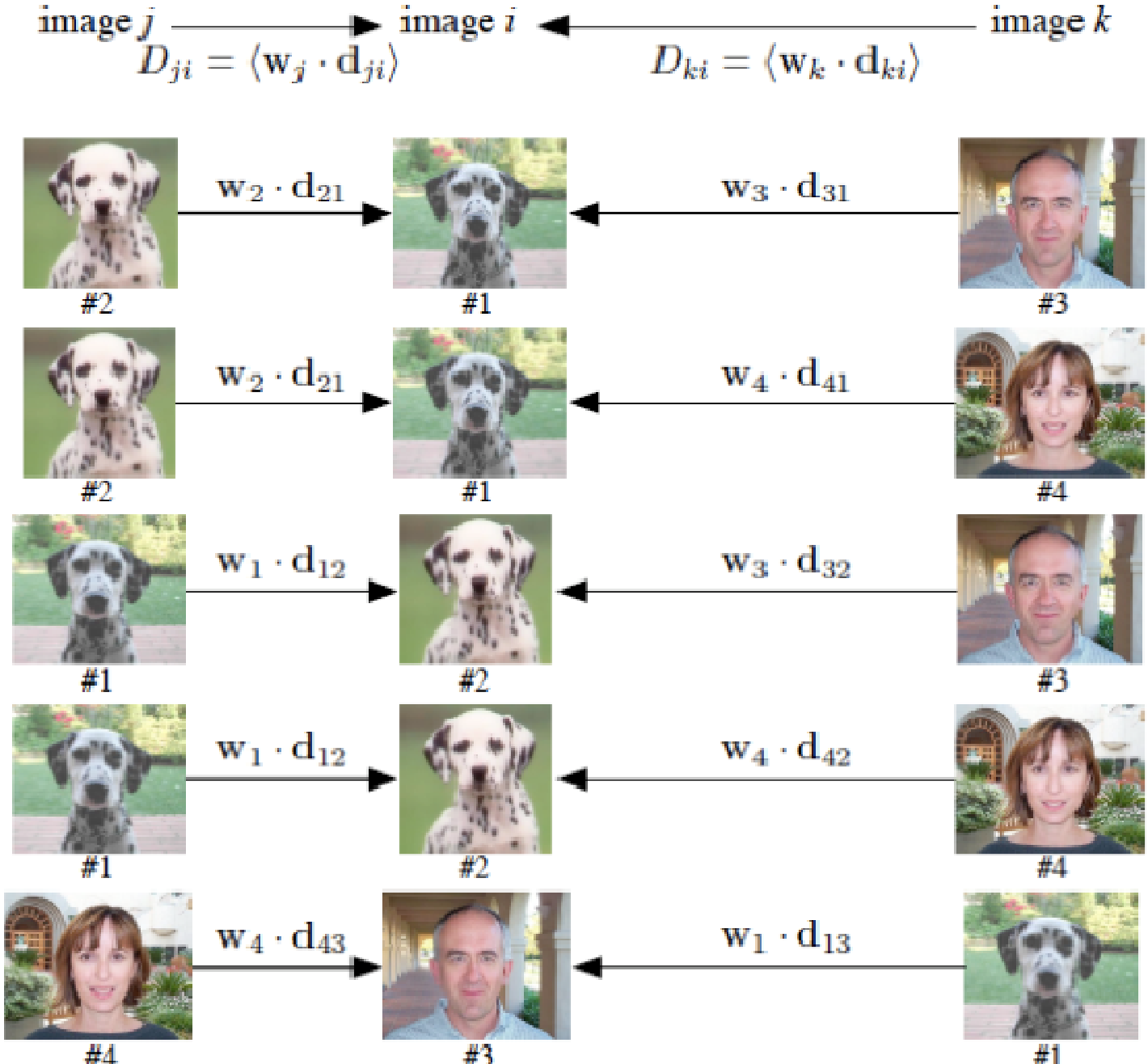
- The image to image distance function is defined as follows:

$$D_{j\to i} = \sum_{m=1}^{M} w_{j\to i,m} \cdot d_{j\to i,m} = \left\langle \vec{W}_j, \vec{d}_{j\to i} \right\rangle$$

- A Triplet constraint will look like this:

$$\left\langle \vec{W}_j, \vec{d}_{j\to i} \right\rangle > \left\langle \vec{W}_k, \vec{d}_{k\to i} \right\rangle$$

A Partial Set of Triplets



Choosing Triplets

- Choosing an exhaustive set of triplets in impractical.
  - Data is too large to fit in the memory.
  - Iteration run time increases linearly with the number of triplets.
  - Only few triplets are informative.
- Solution: pruning:
  - For each of the M elementary distances in focal image F, we find the top K closest images
  - If K contains both in-class and out-of-class images, create all triplets from (F,in-class,out-class) combinations
  - If K are all in-class images, get closest out-class image then make K triplets (reverse all K are out-class)

The optimization

- In an idealistic setting we can hope for the following to hold: $\forall_{i,j,k \in T} w_k \cdot d_{k \to j} > w_i \cdot d_{i \to j}$
- Rescaling w we can get $\forall_{i,j,k \in T} w_k \cdot d_{k \to j} > w_i \cdot d_{i \to j} +$
- In reality a relaxation of the above is needed: Using large margin terminology.(hinge loss) we get a linear penalty for deviation:

$$\sum_{i,j,k \in T} [1 - W \cdot X_{i,j,k}]_+$$

- Finally we get an optimization function:

The optimization - cont...

- Adding a Regularization term: We get the following convex problem:

$$\frac{1}{2}\|W\|^2 + C \sum_{i,j,k \in T} [1 - W \cdot X_{i,j,k}]_+$$

Optimization Solu

- Linear objective function with multip
- The Dual problem:

Slack variable as in standard soft margi

positive W forces p the distance functio

constraints:

$$\min_{W,\xi,\mu} \frac{1}{2}\left\|W - \frac{1}{2}\left\|\sum_{(ijk)\in T} C\alpha_{ijk} - \sum_{(i,j,k)\in T} X_{ijk}\xi_{ijk} + \mu\right\|^2 + \sum_{(ijk)\in T} \alpha$$

s.t.

$$\forall_{i,j,k} : \xi_{ijk} \geq 0$$

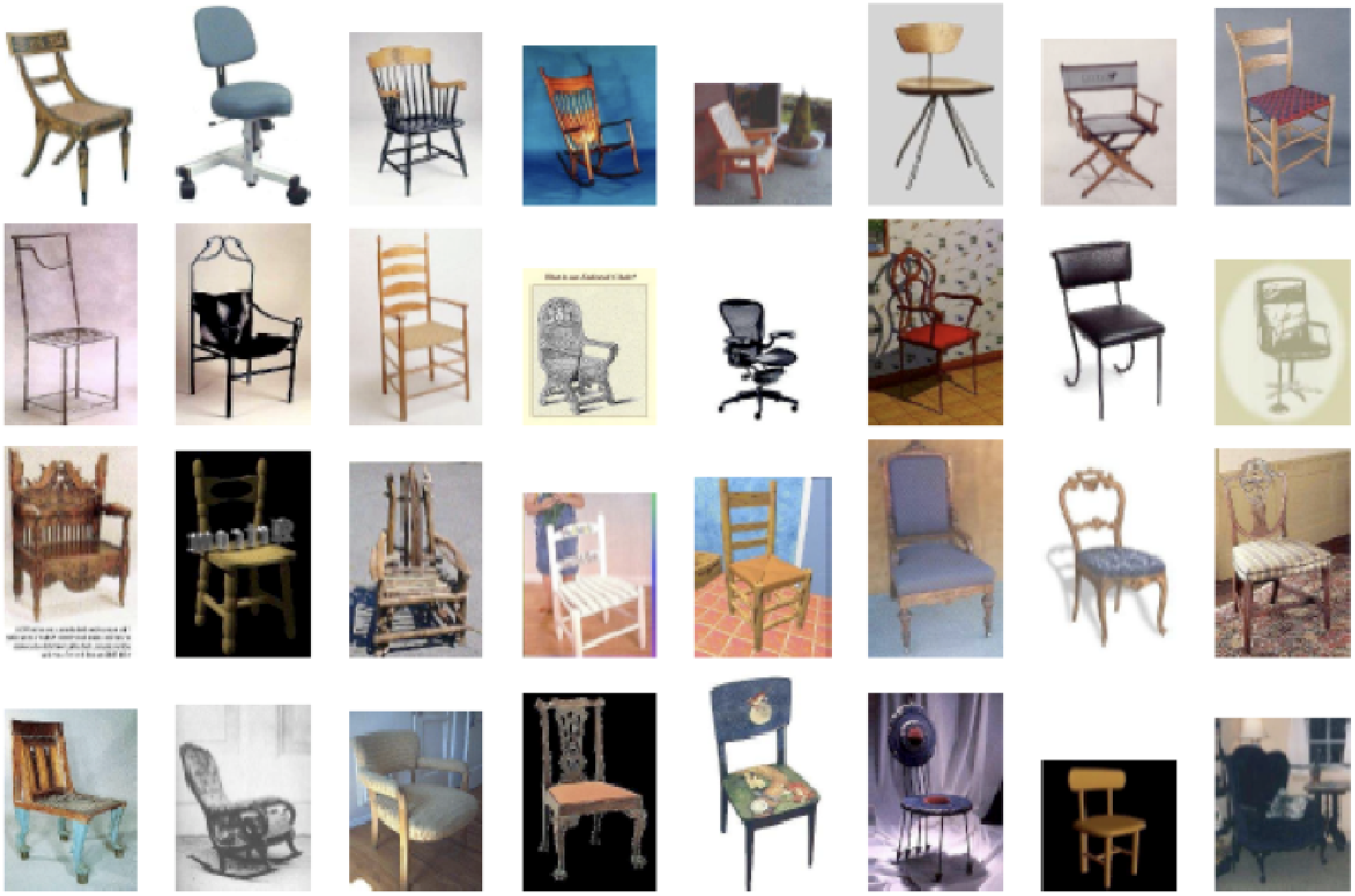
$$\forall_{i,j,k} : 0 \leq \alpha_{ijk} \leq C$$

$$\forall_{i,j,k} : W \cdot X_{ijk} \geq 1 - \xi_{ijk}$$

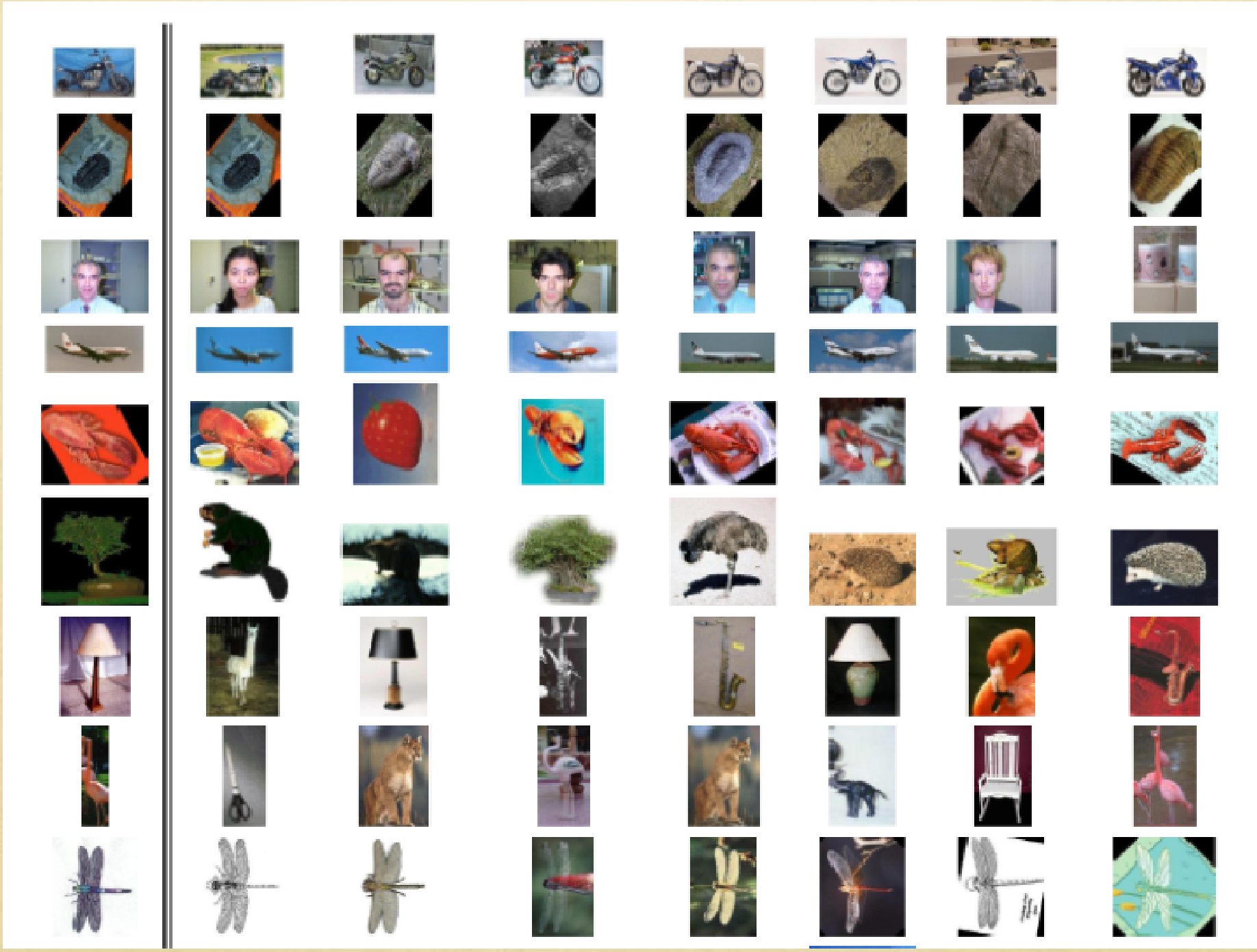
$$\forall_m : \mu_m \geq 0$$

$$\forall_m : W_m \geq 0$$

Caltech 101

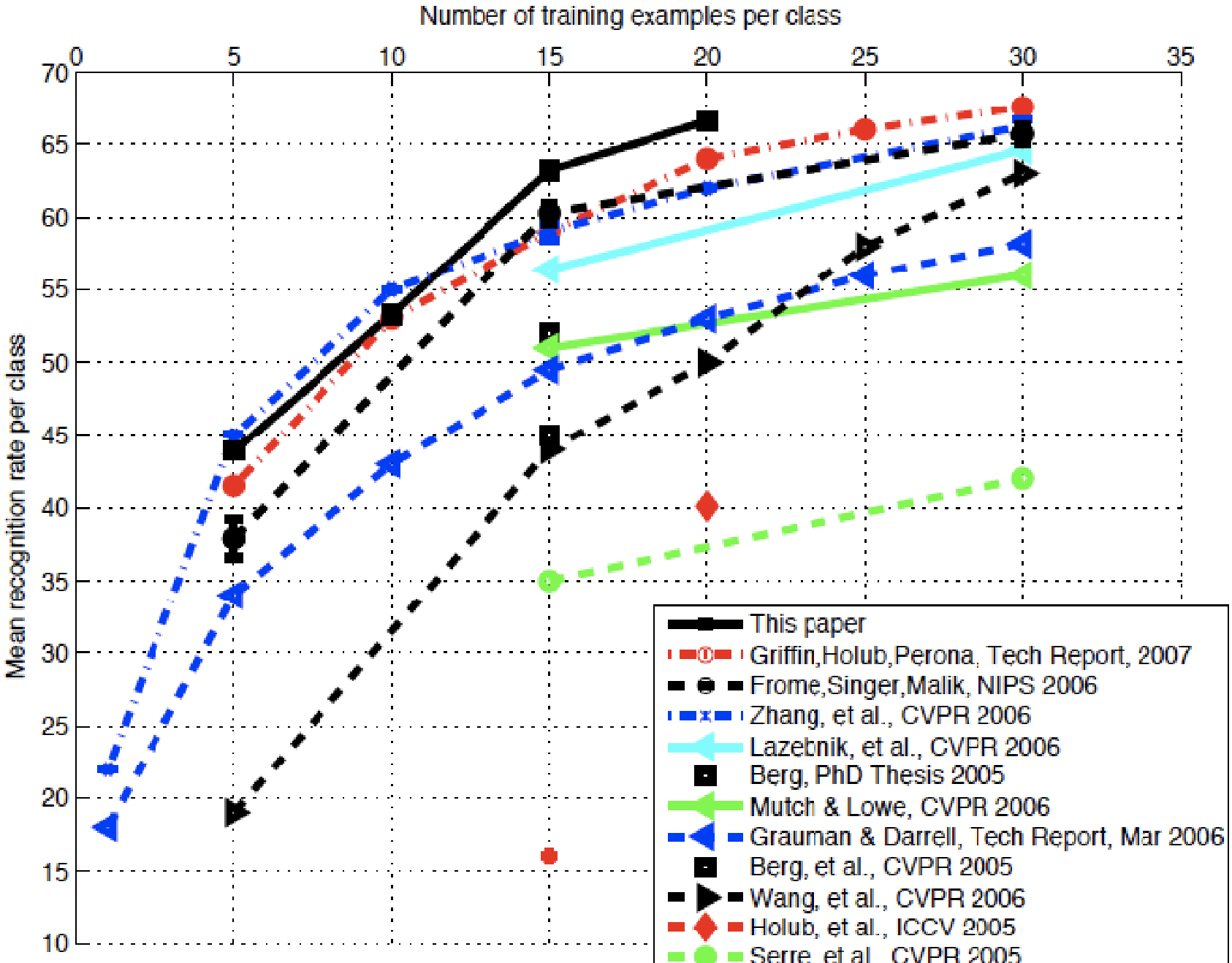


Results:

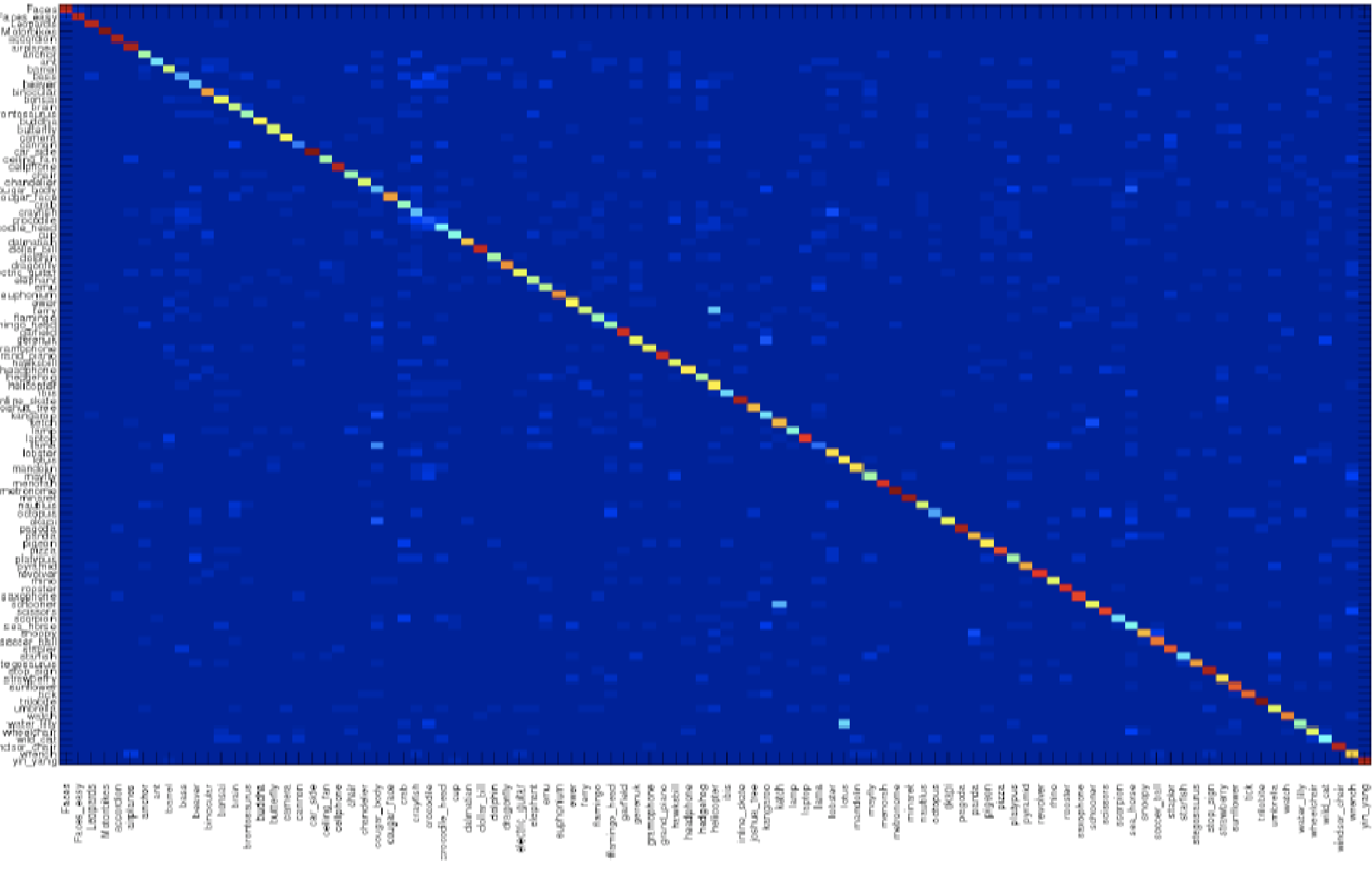


- (Taken from Frome's Thesis,2007)

Results:


Number of training examples per class
Mean recognition rate per class
0
5
10
15
20
25
30
35
70
65
60
55
50
45
40
35
30
25
20
15
10
This paper
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

Results



Take Home message

- Seems like for different visual object categories, different notions of similarity are needed: color informative for some classes, local shape for others, geometry for others.
- Though Global learning is an attractive idea, Local learning might be more adequate and meaningful.
- The visual features of an image are not equally important in determining its similarity to some prototype.
- Which features are more salient depend on the category, or even the image, being considered. (hence locality)
- The described method allows for a combination of different features in a single bag of features