Recognizing objects and actions in images and video

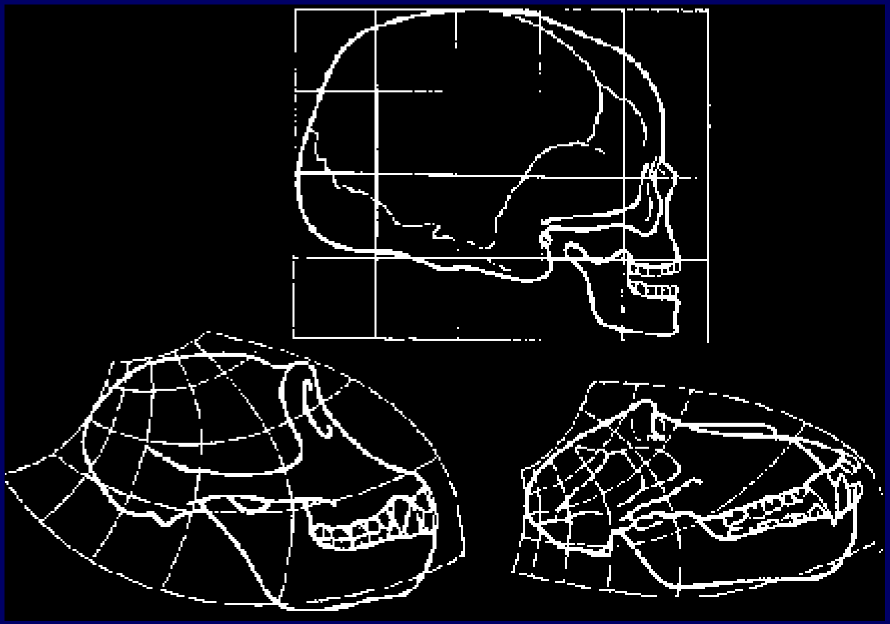
Jitendra Malik

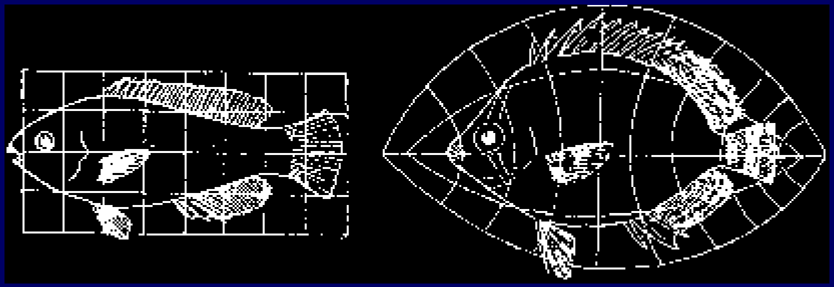
U.C. Berkeley

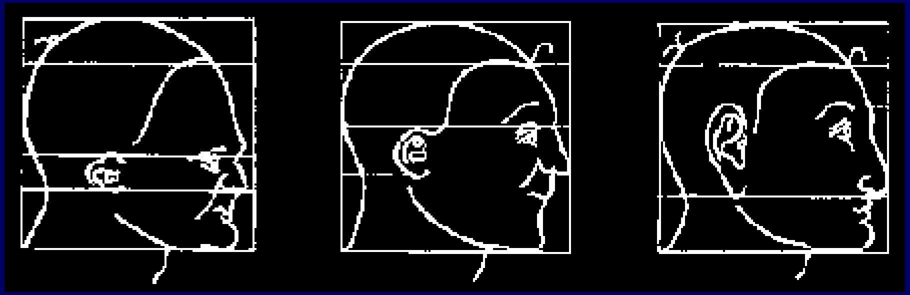
Outline

- Finding boundaries
- Recognizing objects
- Recognizing actions

Biological Shape





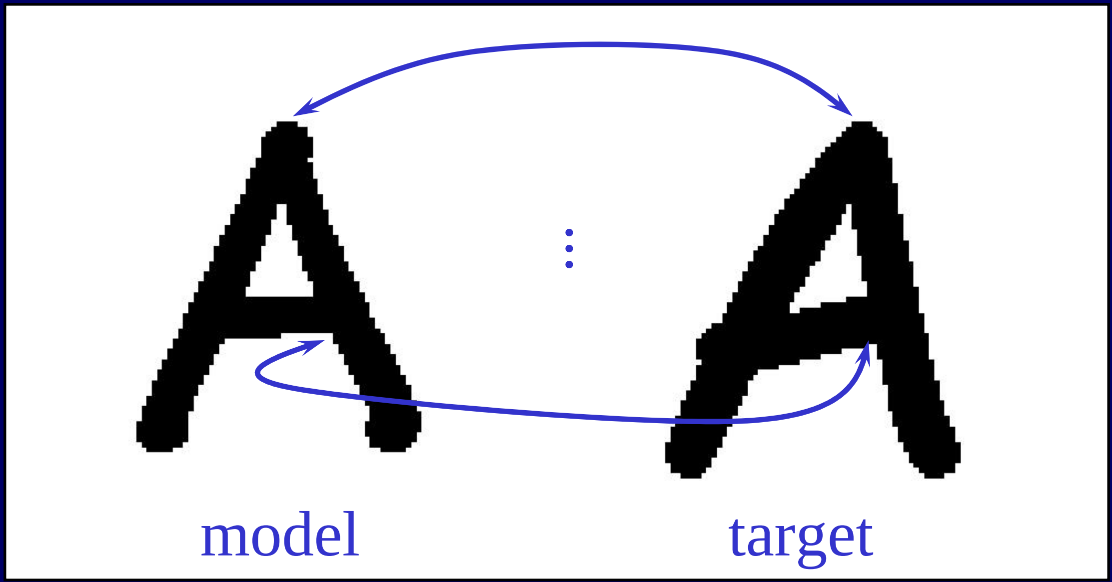


- D'Arcy Thompson: *On Growth and Form*, 1917
  - studied transformations between shapes of organisms

Deformable Templates: Related Work

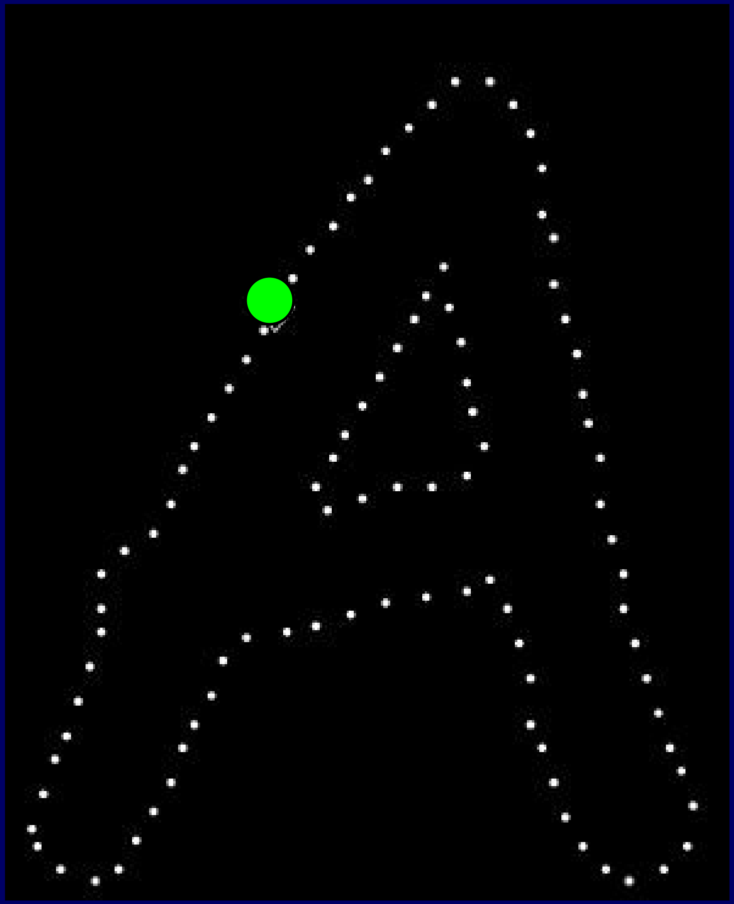
- Fischler & Elschlager (1973)
- Grenander et al. (1991)
- von der Malsburg (1993)

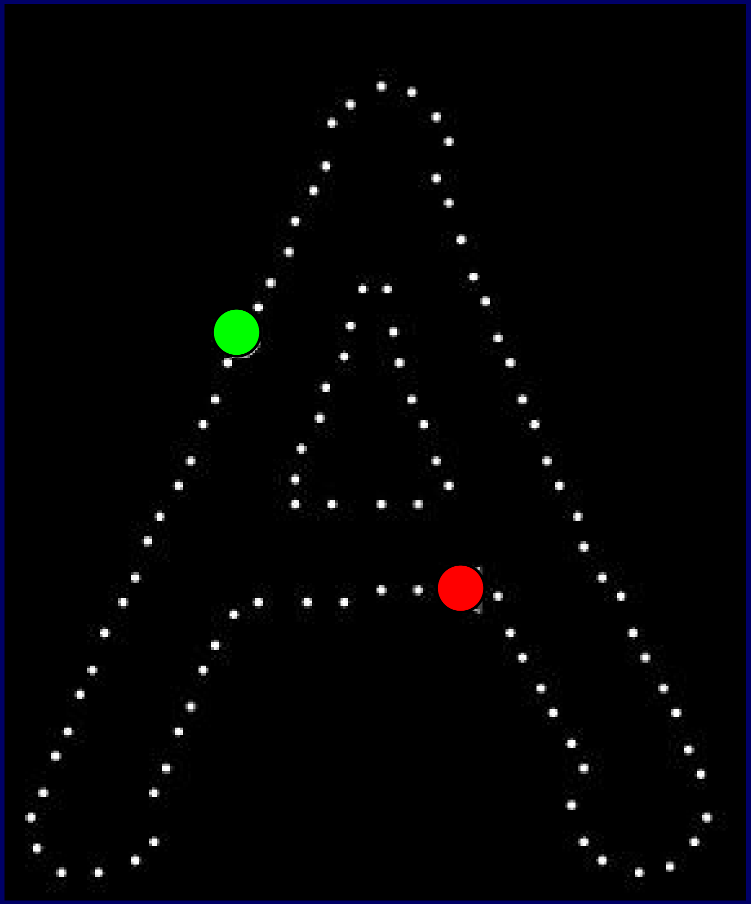
Matching Framework



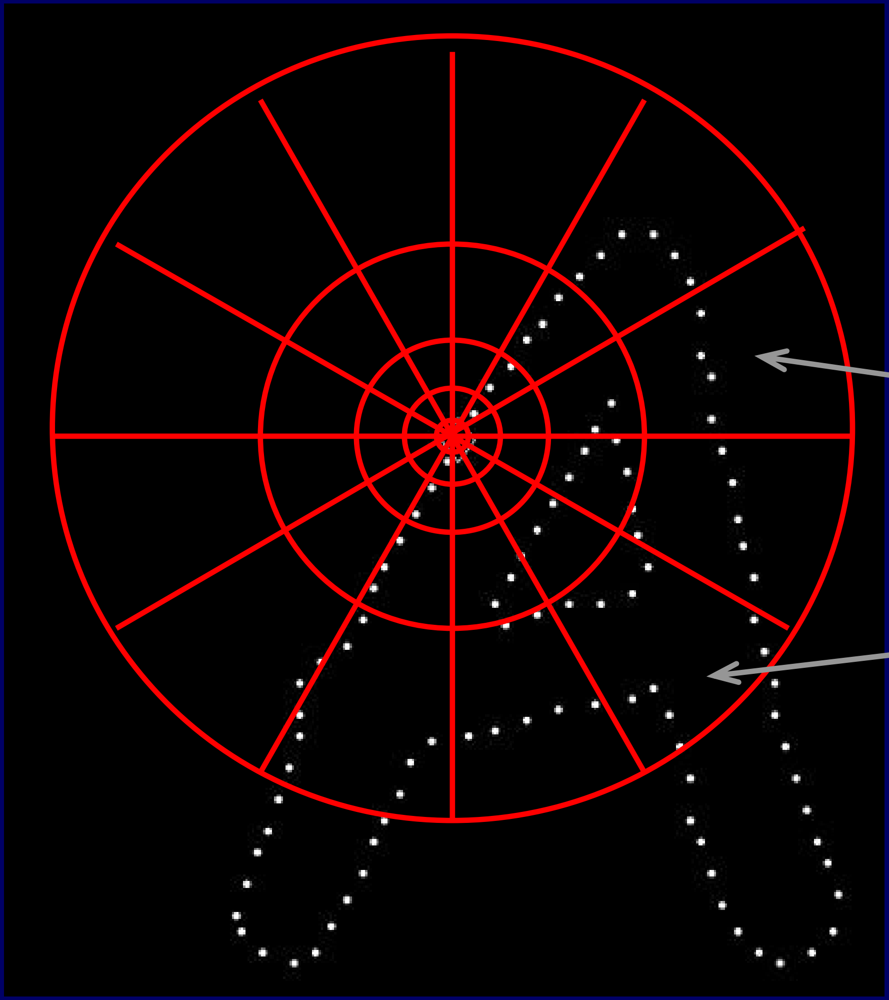
- Find correspondences between points on shape
- Fast pruning
- Estimate transformation & measure similarity

Comparing Pointsets





Shape Context



Count the number of points inside each bin, e.g.:

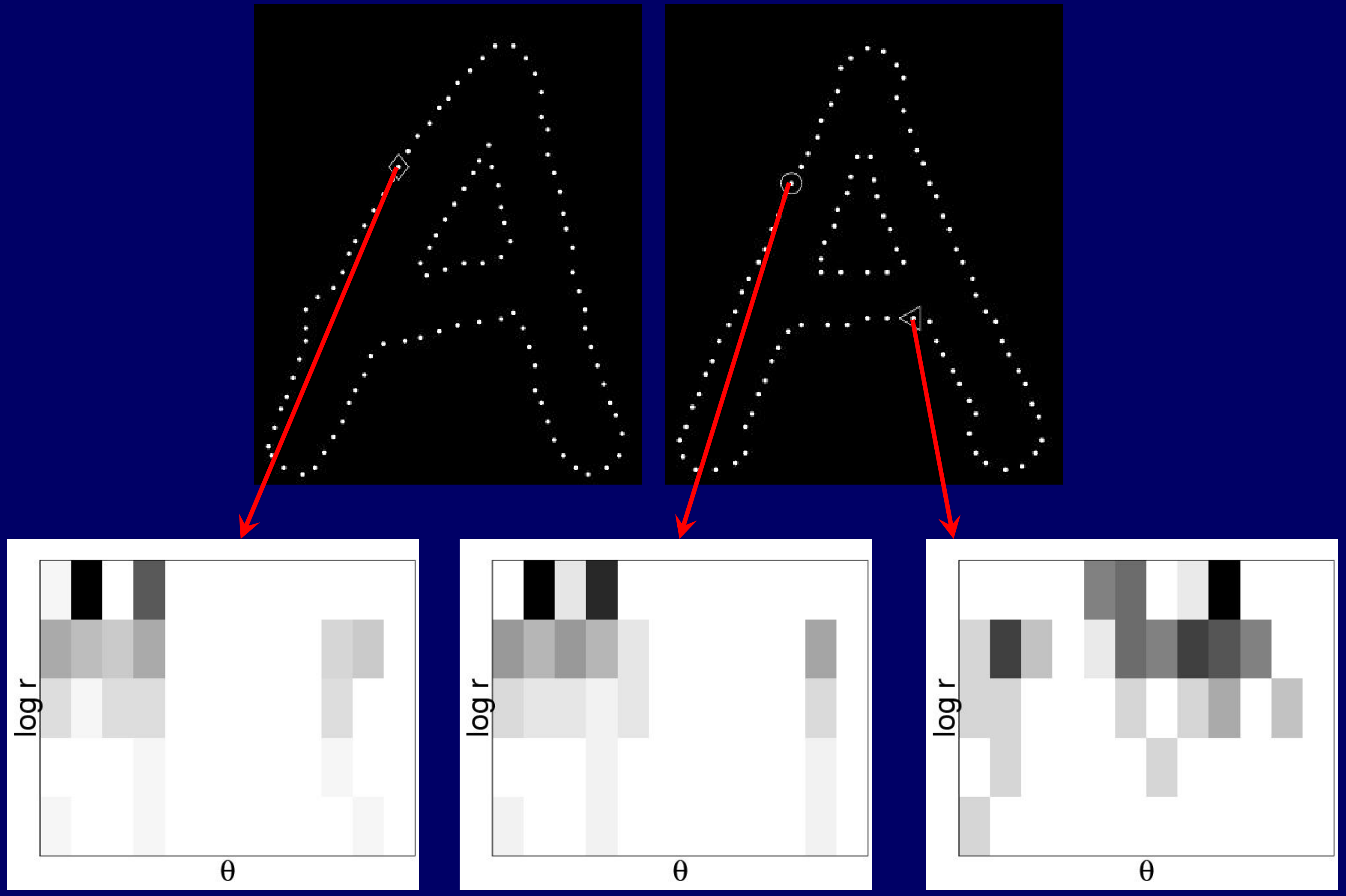
Count = 4

⋮

Count = 10

☞ Compact representation of distribution of points relative to each point

Shape Context



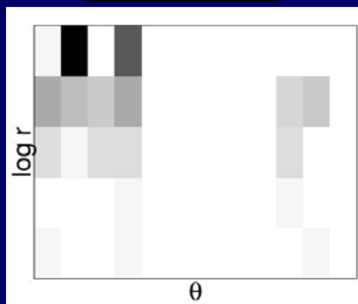
Shape Contexts

- Invariant under translation and scale
- Can be made invariant to rotation by using local tangent orientation frame
- Tolerant to small affine distortion
  - Log-polar bins make spatial blur proportional to *r*

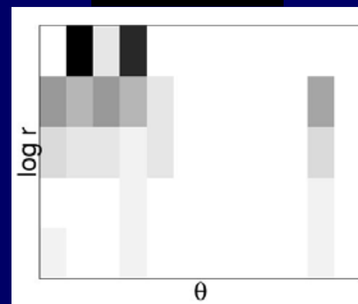
Cf. Spin Images (Johnson & Hebert) - range image registration

Comparing Shape Contexts

$h_i(k)$

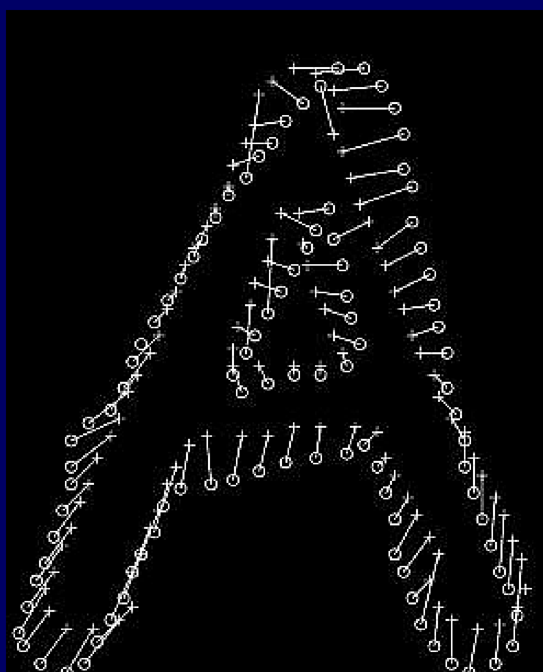


$h_j(k)$



Compute matching costs using Chi Squared distance:

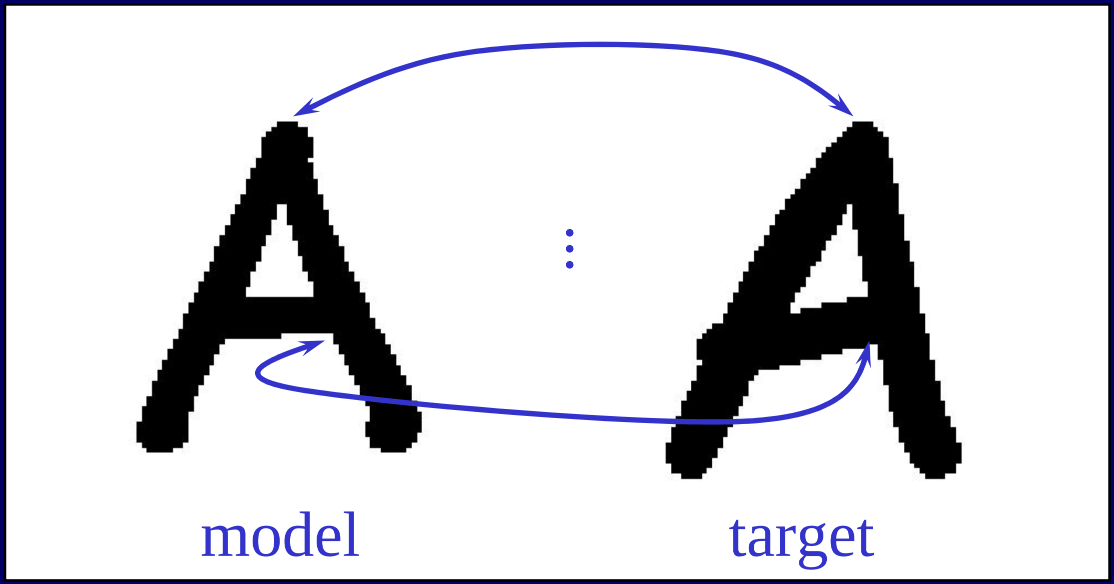
$$C_{ij} = \frac{1}{2} \sum_{k=1}^{K} \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)}$$



Recover correspondences by solving linear assignment problem with costs C_{ij}

[Jonker & Volgenant 1987]

Matching Framework



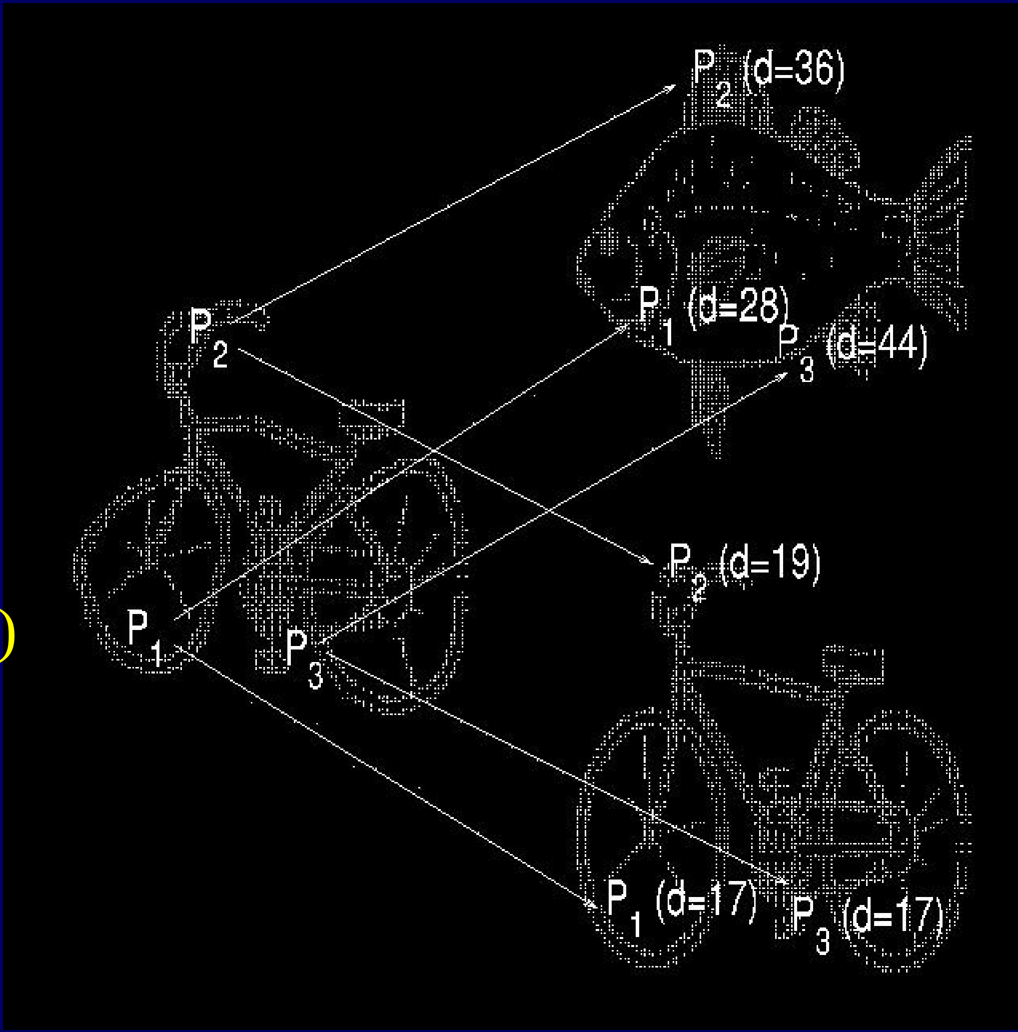
- Find correspondences between points on shape
- Fast pruning
- Estimate transformation & measure similarity

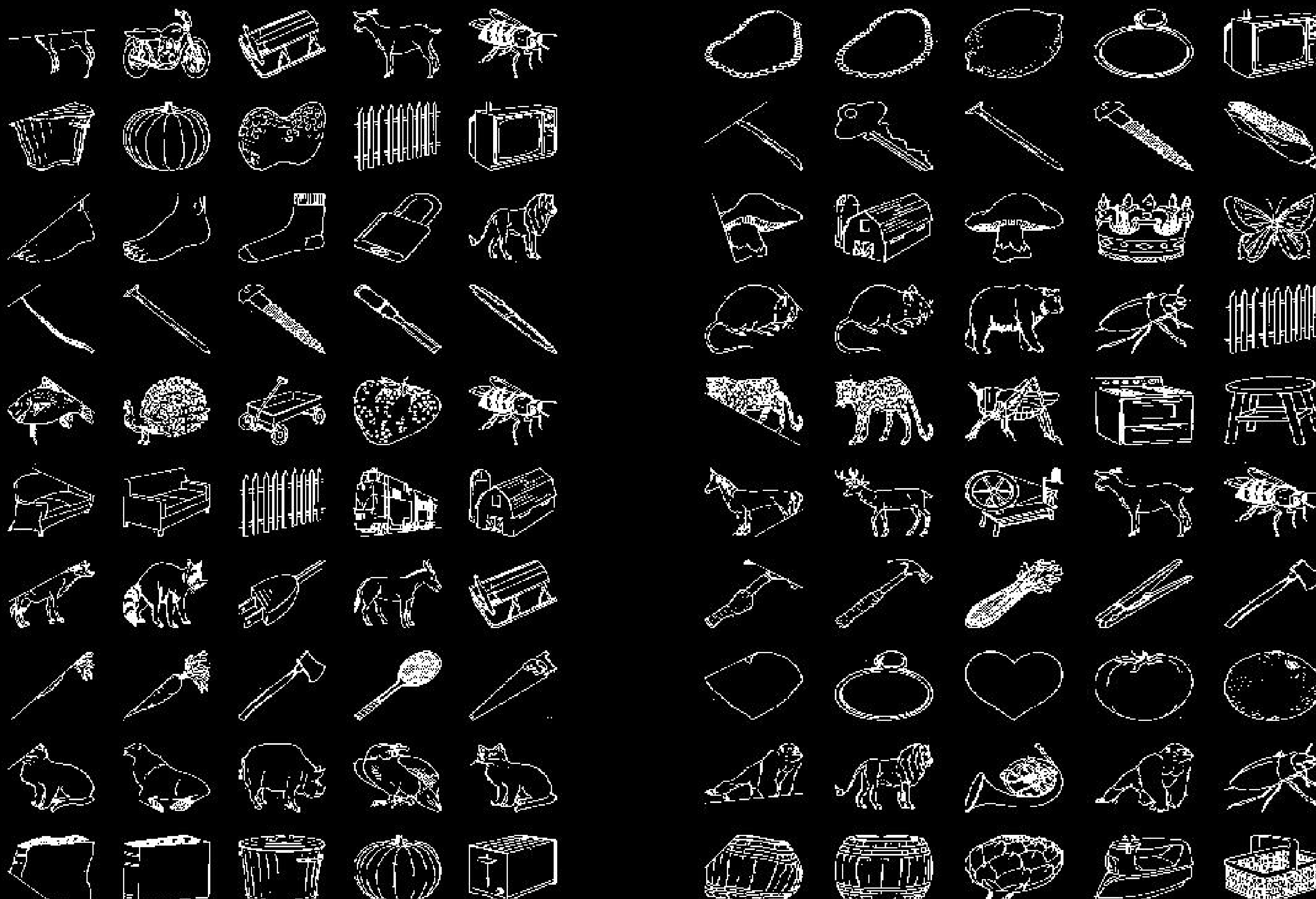
Fast pruning

- Find best match for the shape context at only a few random points and add up cost

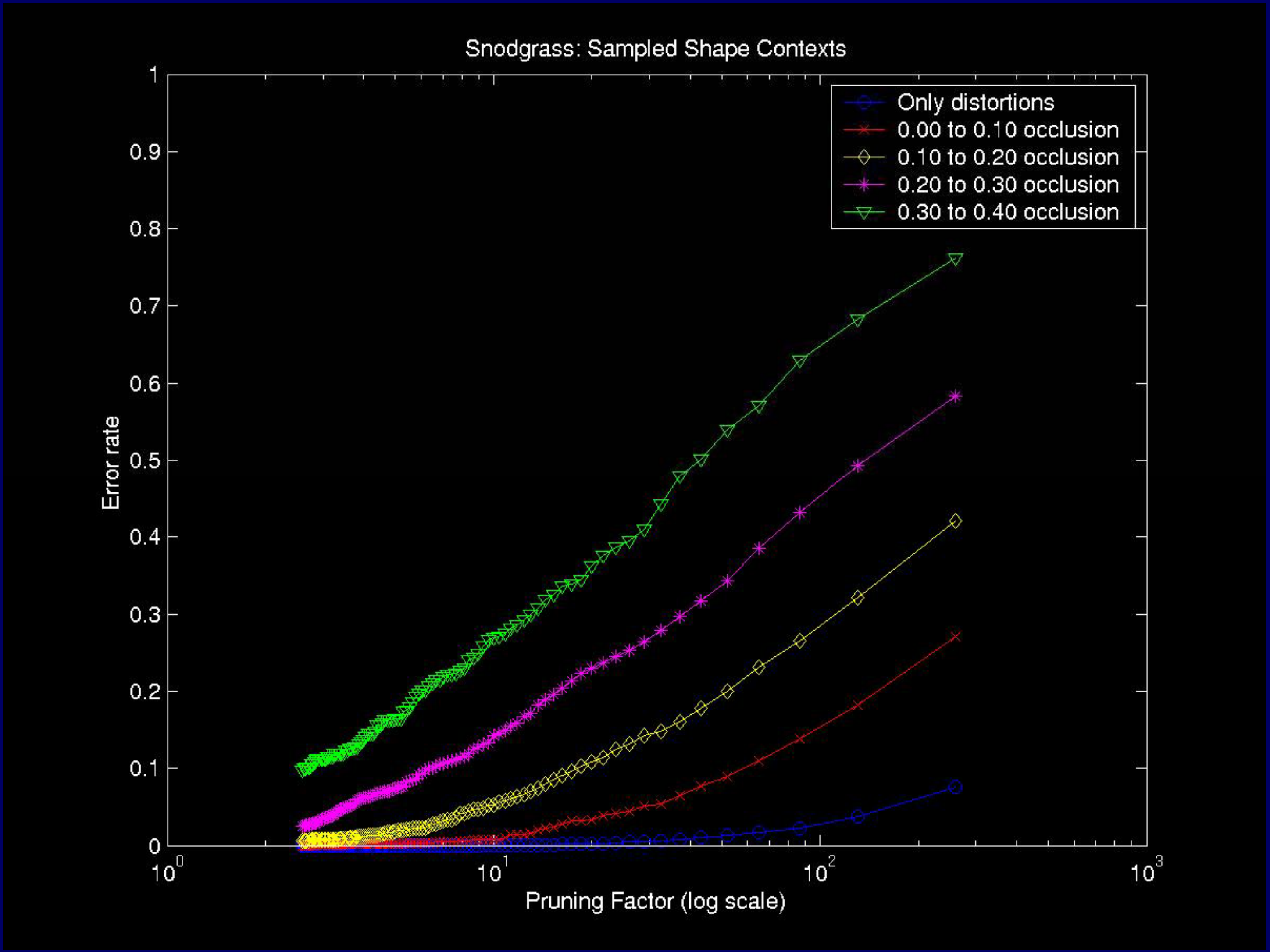
$$dist(S_{query}, S_i) = \sum_{j=1}^{r} \chi^2(SC^j_{query}, SC^*_i)$$

$$SC^*_i = \arg\min_u \chi^2(SC^j_{query}, SC^u_i)$$



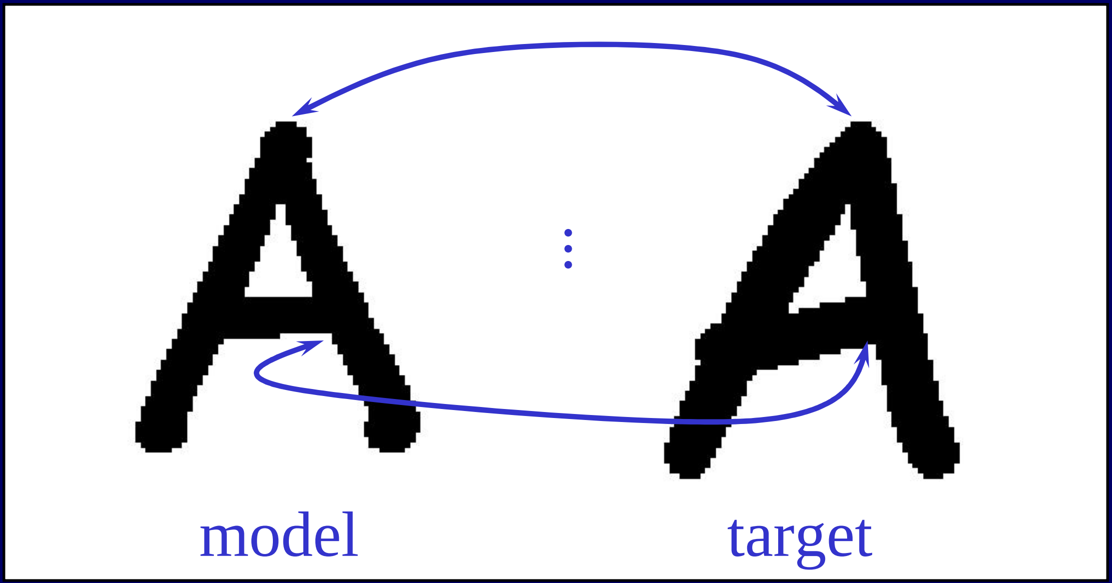


Results



Snodgrass: Sampled Shape Contexts
Only distortions
0.00 to 0.10 occlusion
0.10 to 0.20 occlusion
0.20 to 0.30 occlusion
0.30 to 0.40 occlusion
Error rate
Pruning Factor (log scale)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10^0
10^1
10^2
10^3

Matching Framework



- Find correspondences between points on shape
- Fast pruning
- Estimate transformation & measure similarity

Thin Plate Spline Model

- 2D counterpart to cubic spline:

$$U(r) \;=\; r^2 \log r, \quad r > 0$$

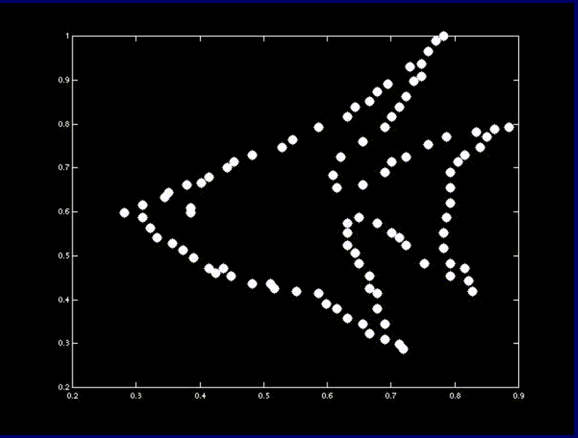
- Minimizes *bending energy*:

$$I_f \;=\; \int\!\!\int_{\mathbb{R}^2} \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 dx\,dy$$

- Solve by inverting linear system
- Can be regularized when data is inexact

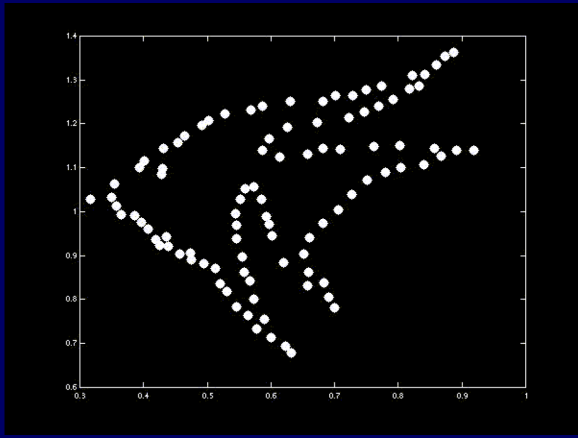
Duchon (1977), Meinguet (1979), Wahba (1991)

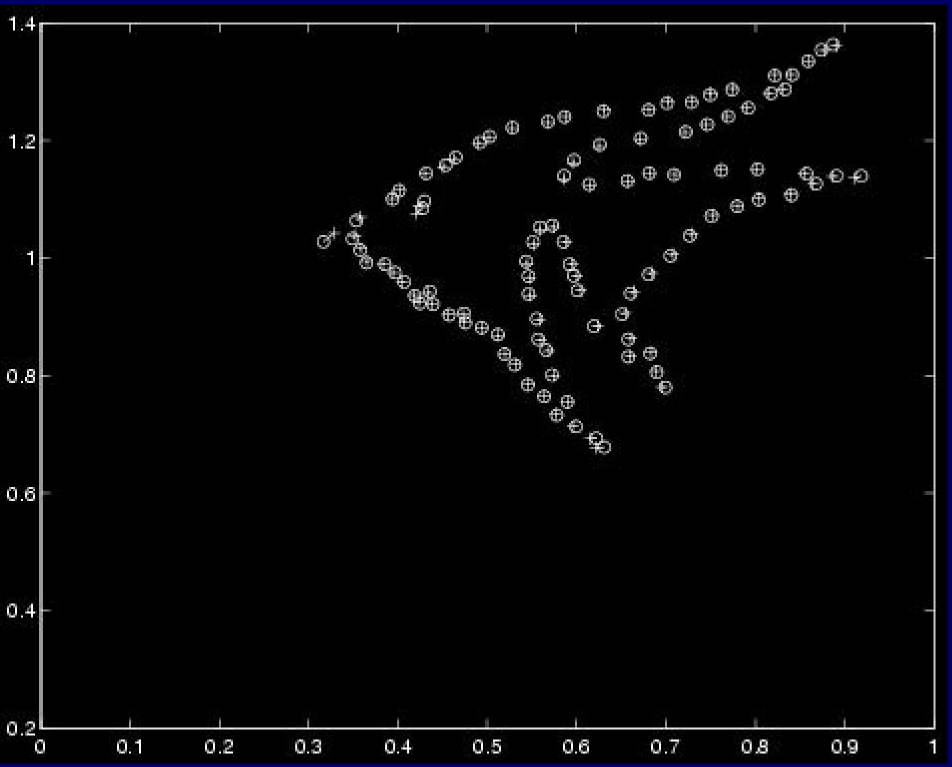
Matching Example

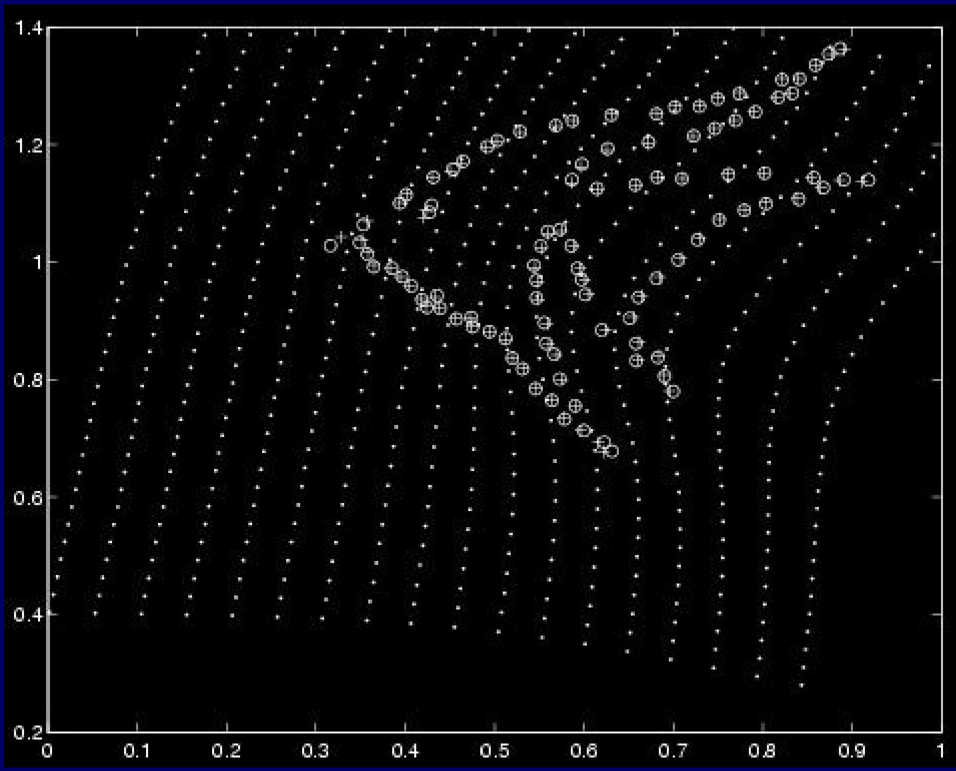


model

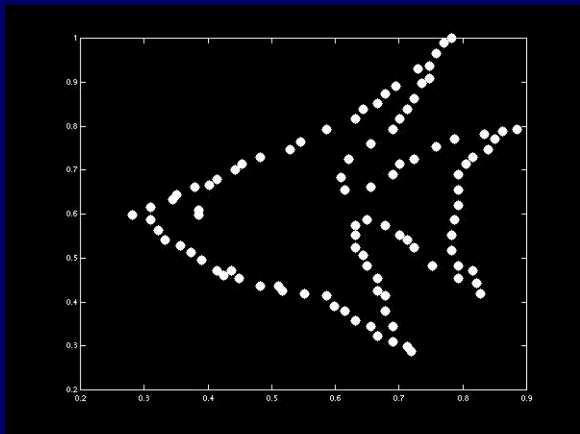
target

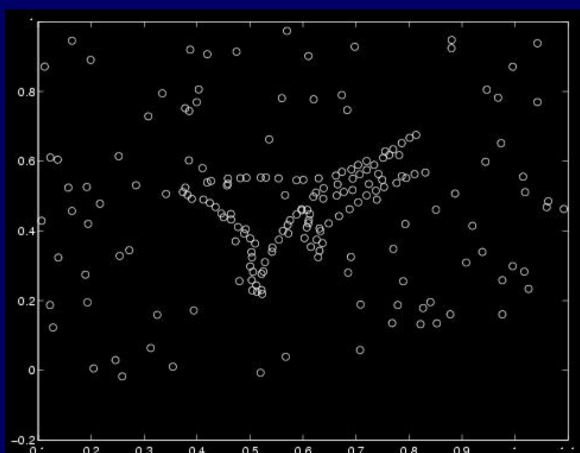


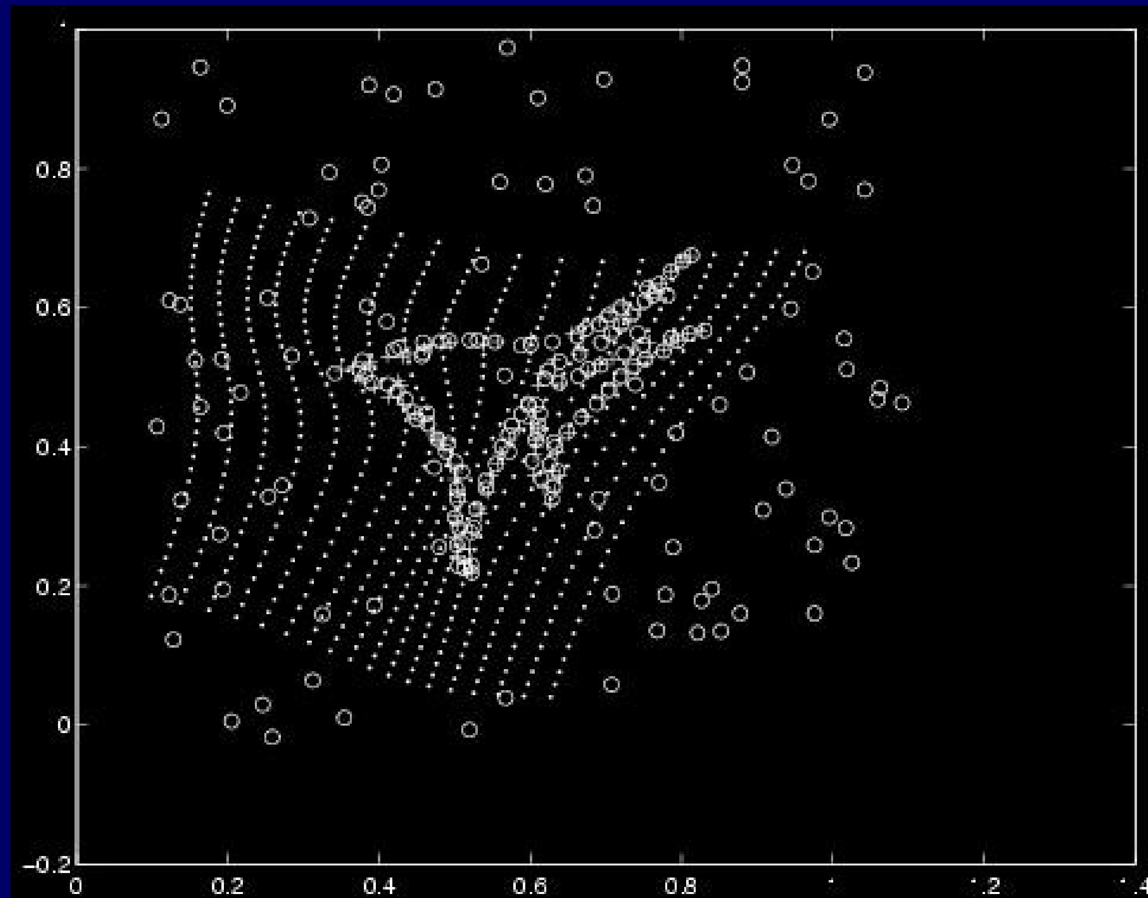




Outlier Test Example

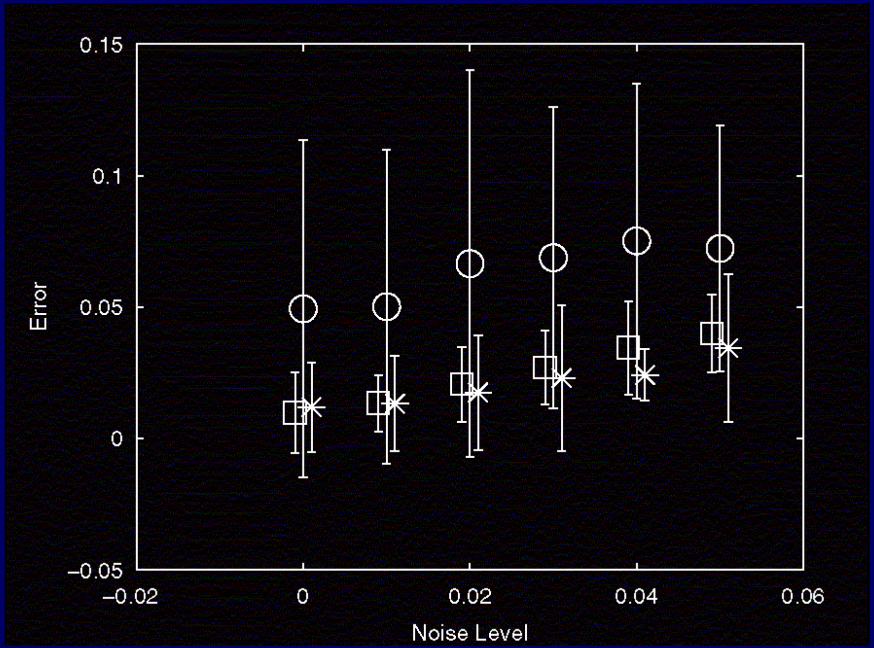




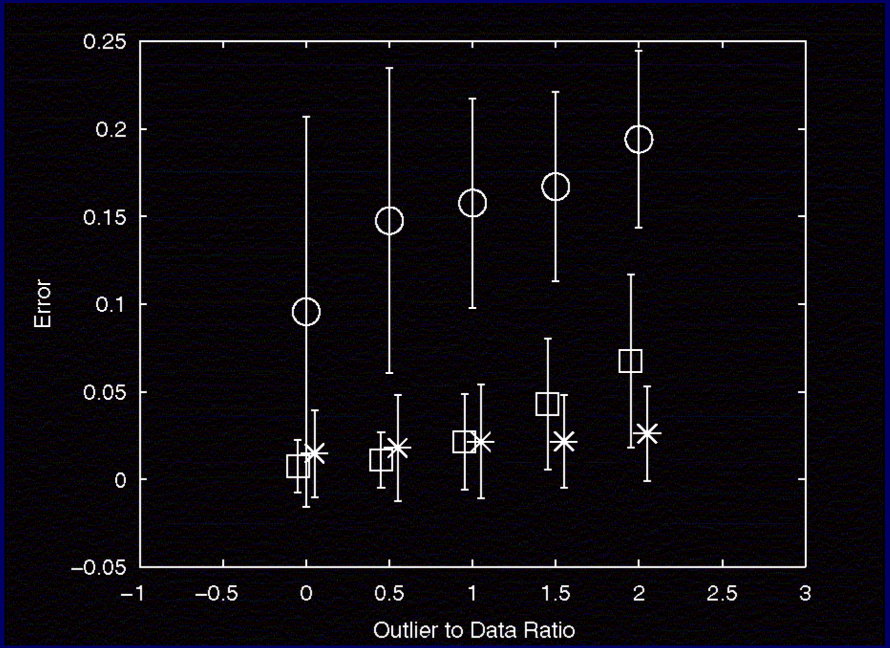


Synthetic Test Results

Fish - deformation + noise



Fish - deformation + outliers



○ ICP □ Shape Context ✱ Chui & Rangarajan

Terms in Similarity Score

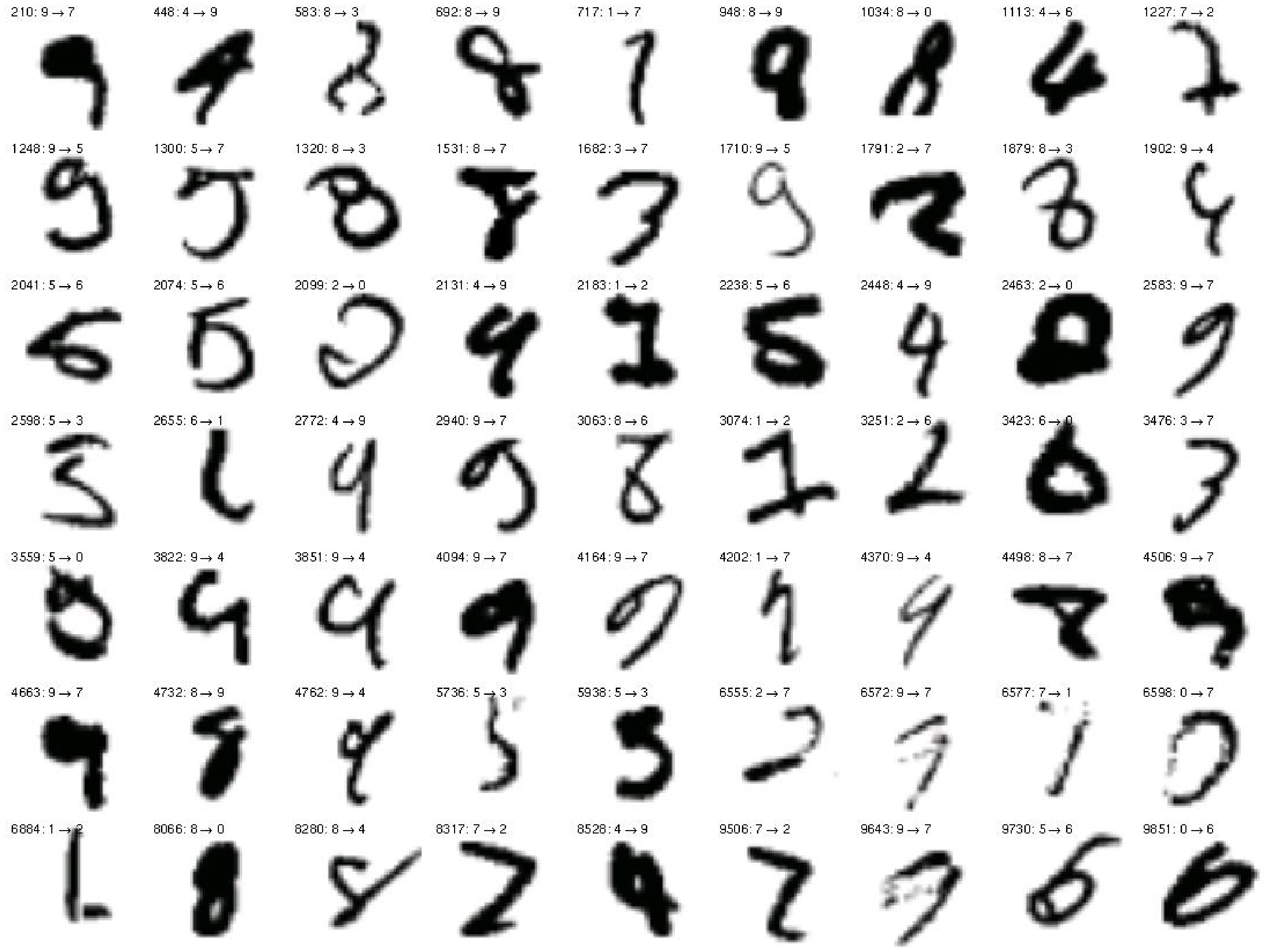
- Shape Context difference
- Local Image appearance difference
  - orientation
  - gray-level correlation in Gaussian window
  - … (many more possible)
- Bending energy

Object Recognition Experiments

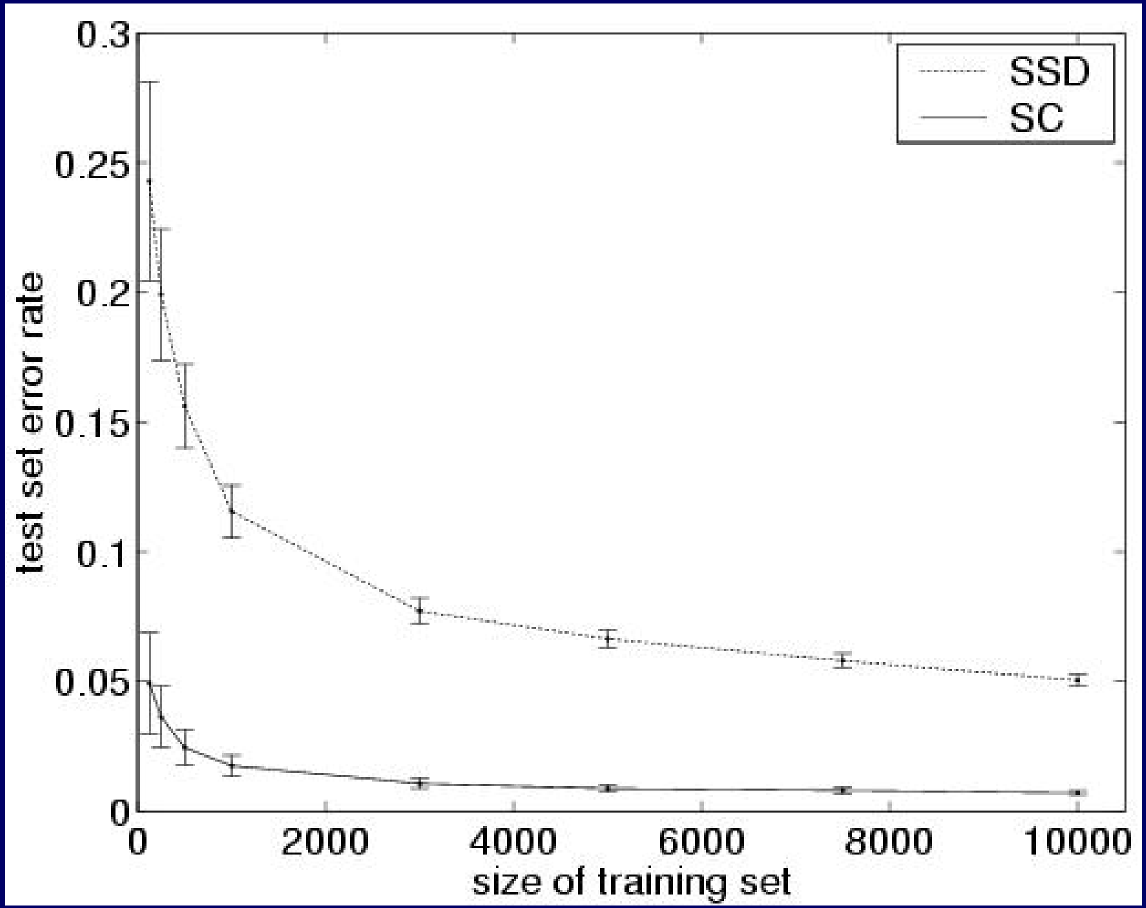
- Handwritten digits
- COIL 3D objects (Nayar-Murase)
- Human body configurations
- Trademarks

Handwritten Digit Recognition

- MNIST 60 000:
  - linear: 12.0%
  - 40 PCA+ quad: 3.3%
  - 1000 RBF +linear: 3.6%
  - K-NN: 5%
  - K-NN (deskewed): 2.4%
  - K-NN (tangent dist.): 1.1%
  - SVM: 1.1%
  - LeNet 5: 0.95%
- MNIST 600 000 (distortions):
  - LeNet 5: 0.8%
  - SVM: 0.8%
  - Boosted LeNet 4: 0.7%
- MNIST 20 000:
  - K-NN, Shape Context matching: 0.63%


210: 9 → 7
448: 4 → 9
583: 8 → 3
692: 8 → 9
717: 1 → 7
948: 8 → 9
1034: 8 → 0
1113: 4 → 6
1227: 7 → 2
1248: 9 → 5
1300: 5 → 7
1320: 8 → 3
1531: 8 → 7
1682: 3 → 7
1710: 9 → 5
1791: 2 → 7
1879: 8 → 3
1902: 9 → 4
2041: 5 → 6
2074: 5 → 6
2099: 2 → 0
2131: 4 → 9
2183: 1 → 2
2238: 5 → 6
2448: 4 → 9
2463: 2 → 0
2583: 9 → 7
2598: 5 → 3
2655: 6 → 1
2772: 4 → 9
2940: 9 → 7
3063: 8 → 6
3074: 1 → 2
3251: 2 → 6
3423: 6 → 0
3476: 3 → 7
3559: 5 → 0
3822: 9 → 4
3851: 9 → 4
4094: 9 → 7
4164: 9 → 7
4202: 1 → 7
4370: 9 → 4
4498: 8 → 7
4506: 9 → 7
4663: 9 → 7
4732: 8 → 9
4762: 9 → 4
5736: 5 → 3
5938: 5 → 3
6555: 2 → 7
6572: 9 → 7
6577: 7 → 1
6598: 0 → 7
6884: 1 → 2
8066: 8 → 0
8280: 8 → 4
8317: 7 → 2
8528: 4 → 9
9506: 7 → 2
9643: 9 → 7
9730: 5 → 6
9851: 0 → 6

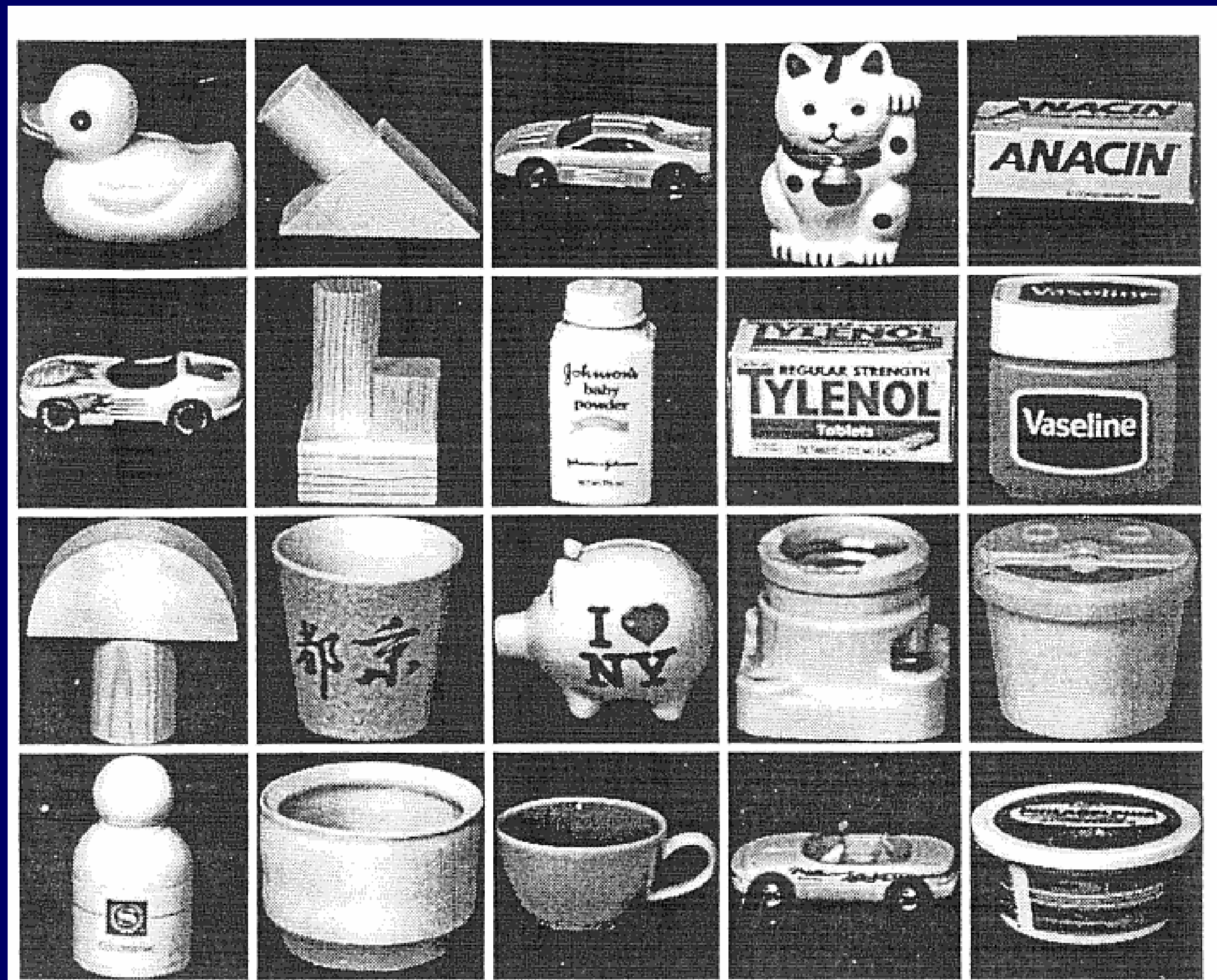
Results: Digit Recognition



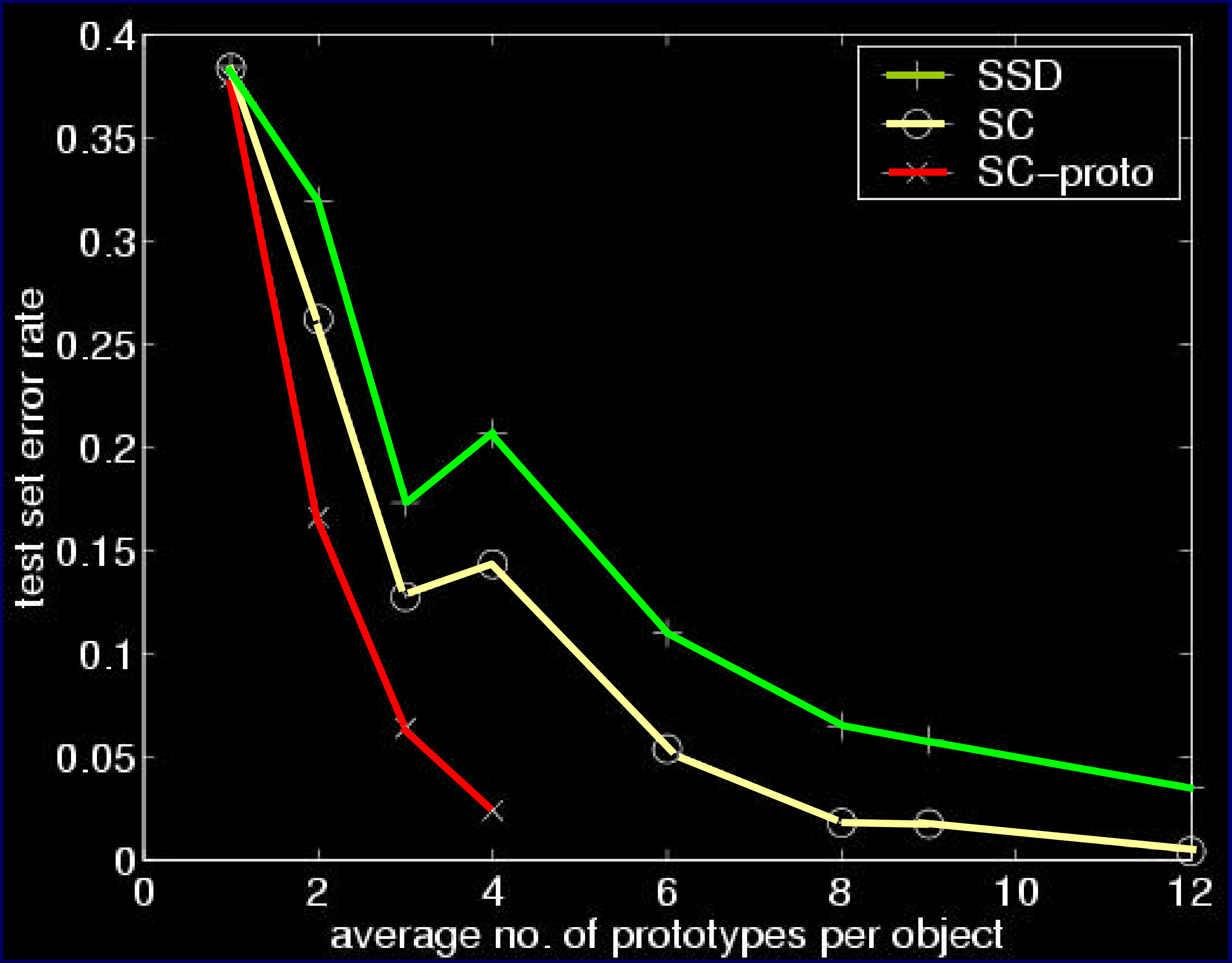
1-NN classifier using:

Shape context + 0.3 * bending + 1.6 * image appearance

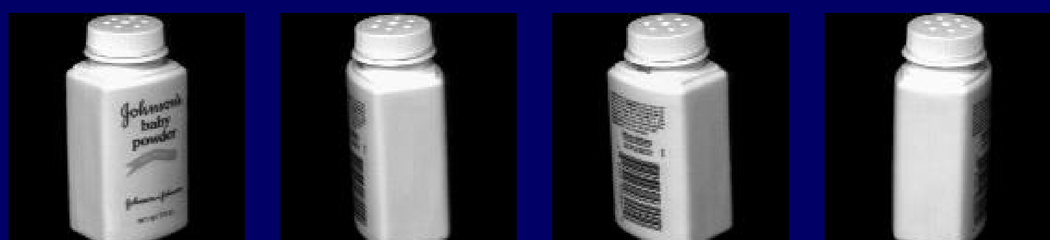
COIL Object Database



Error vs. Number of Views

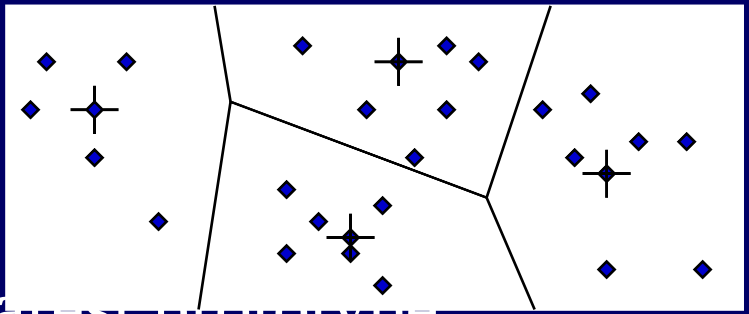


Prototypes Selected for 2 Categories



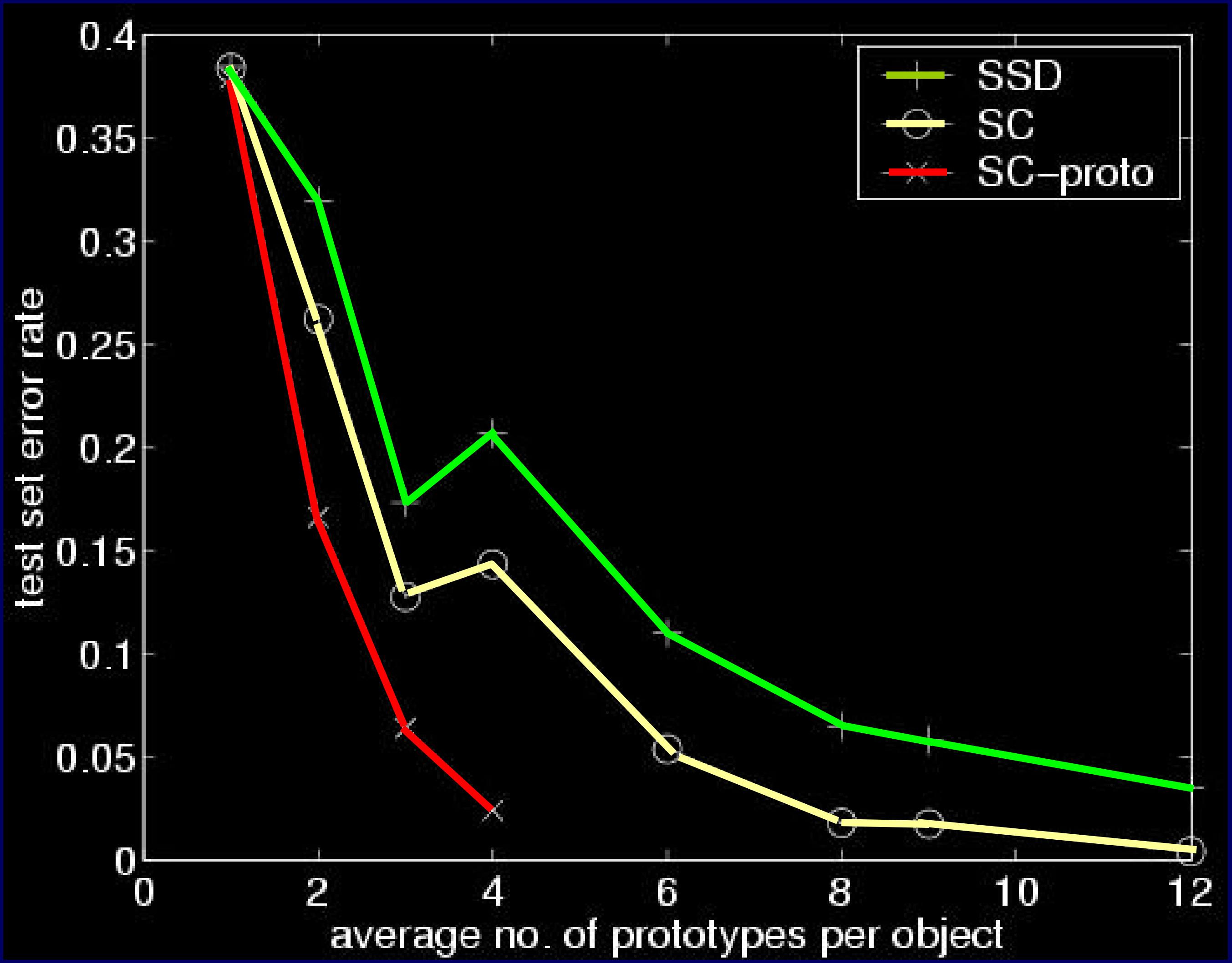
Details in Belongie, Malik & Puzicha (NIPS2000)

Editing: K-medoids

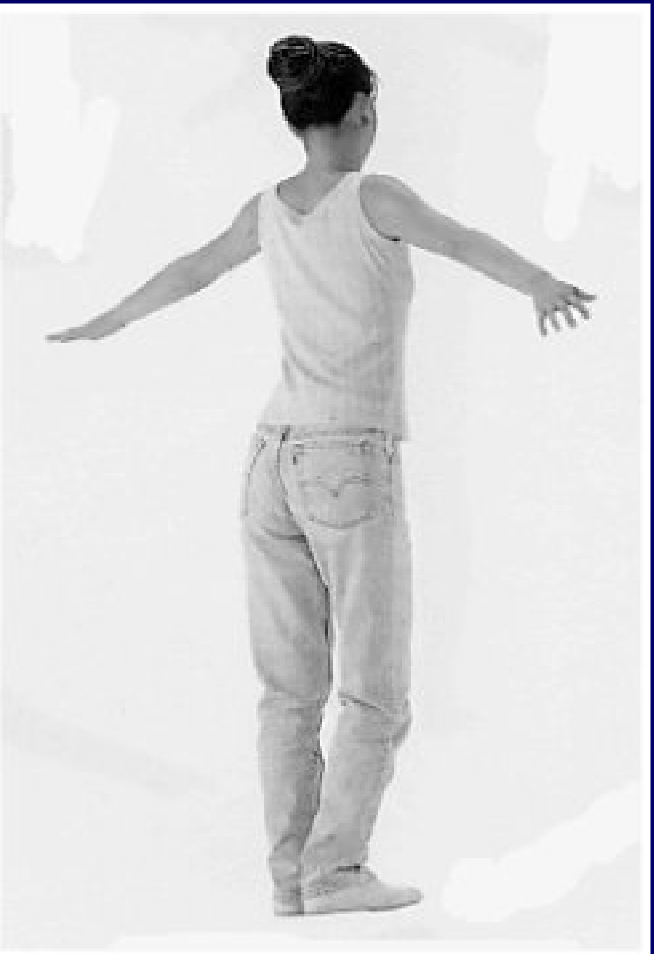


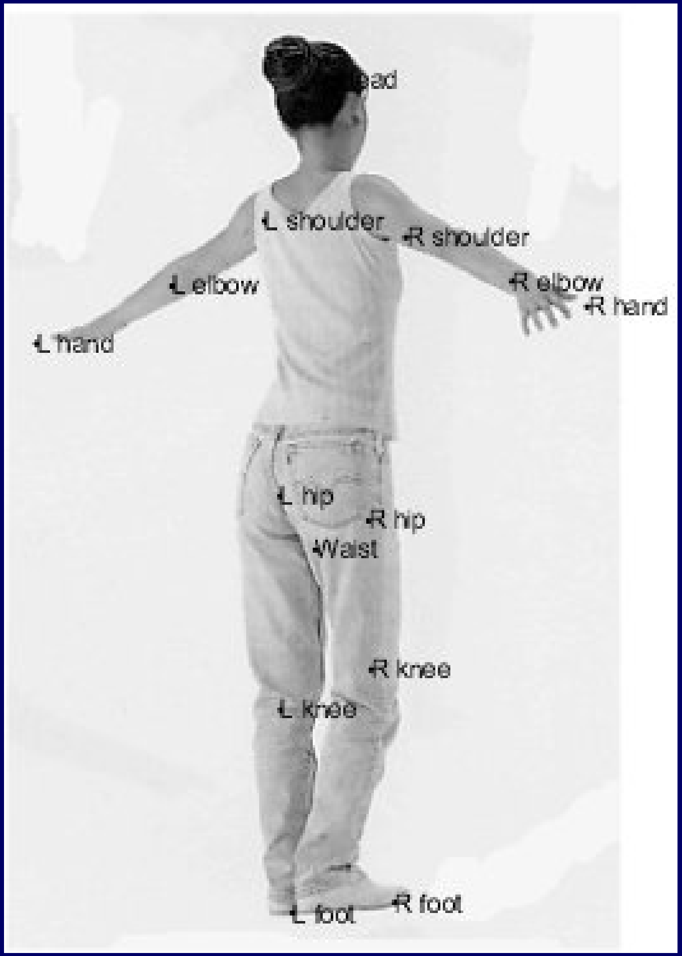
- Input: similarity matrix
- Select: K prototypes
- Minimize: mean distance to nearest prototype
- Algorithm:
  - iterative
  - split cluster with most errors
- Result: Adaptive distribution of resources (cfr. aspect graphs)

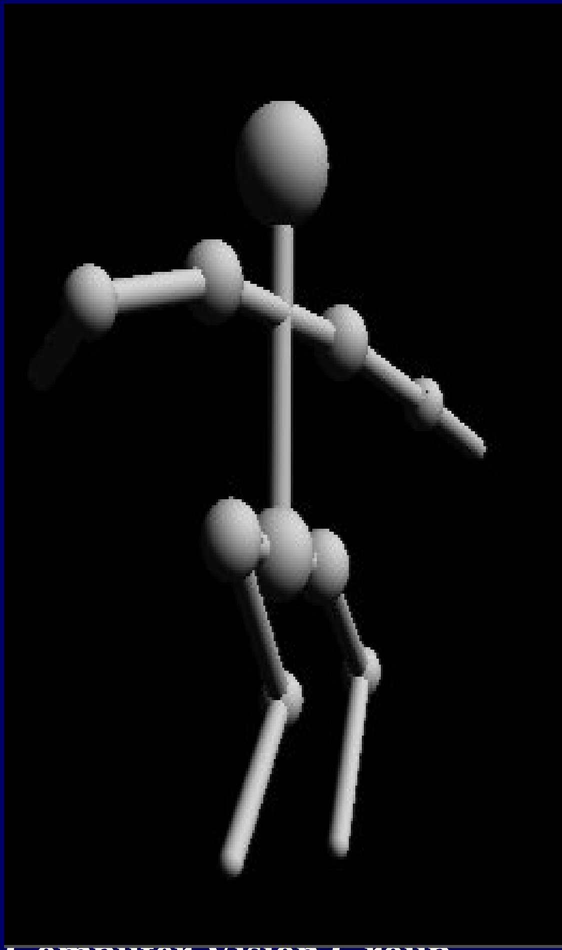
Error vs. Number of Views



Human body configurations

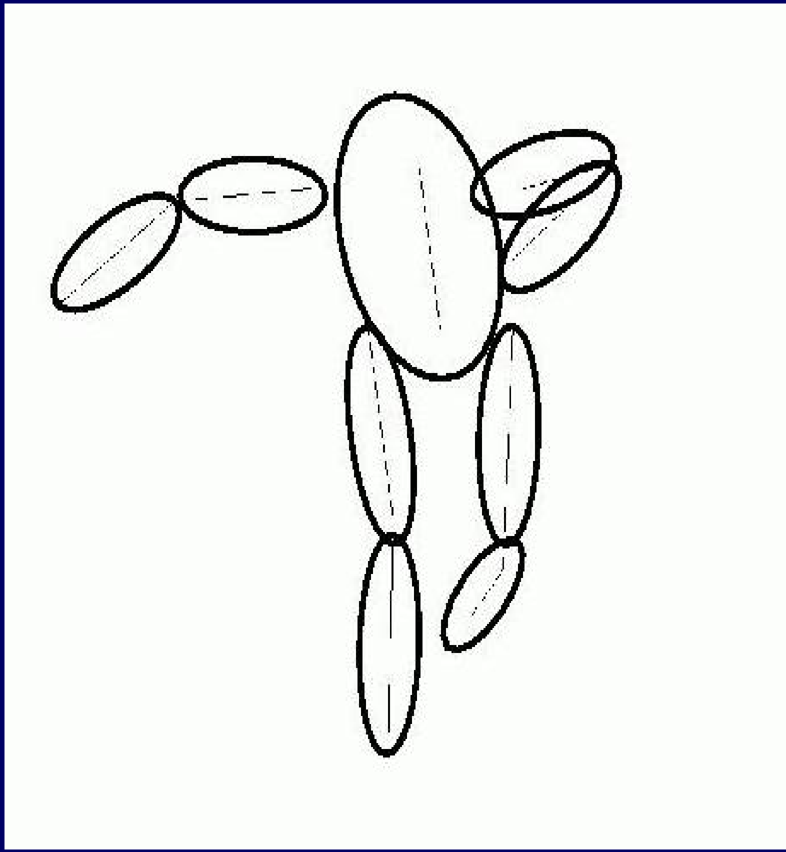




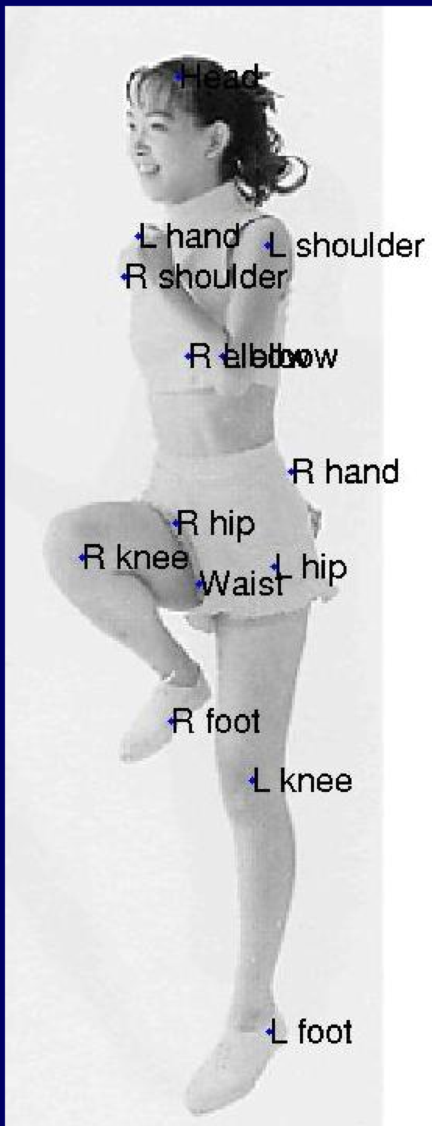


Deformable Matching

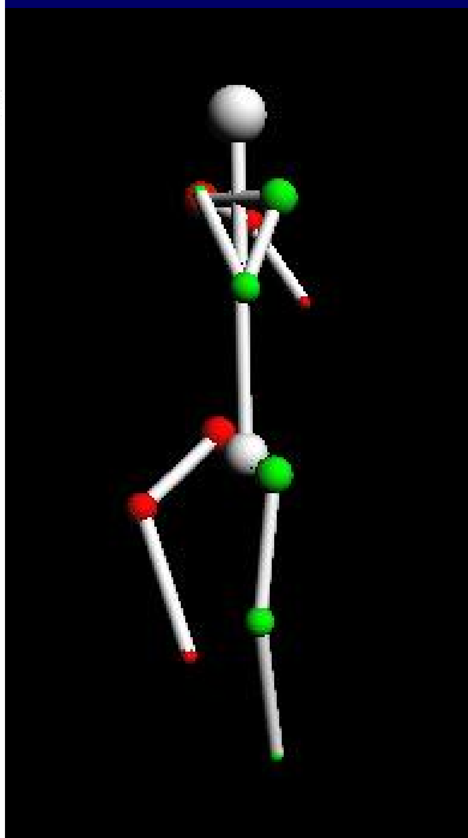
- Kinematic chain-based deformation model
- Use iterations of correspondence and deformation
- Keypoints on exemplars are deformed to locations on query image

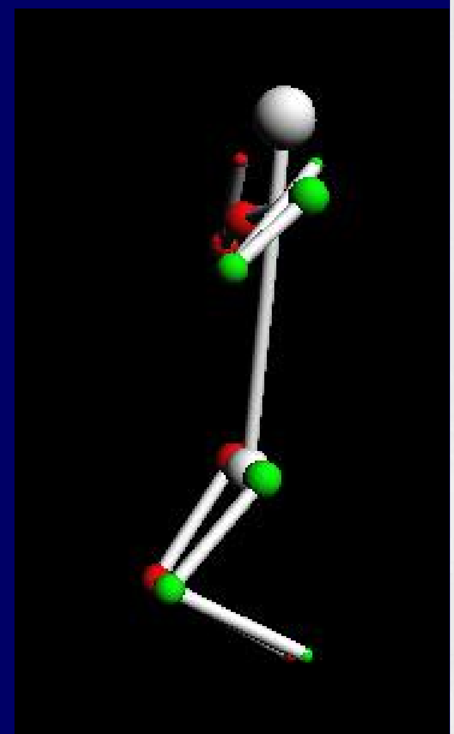


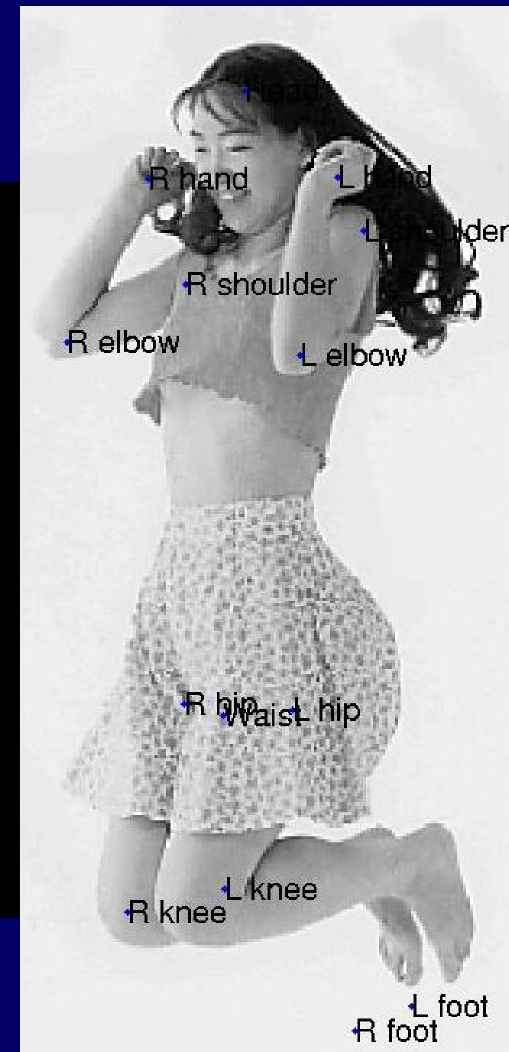
Results



Head
L hand
L shoulder
R shoulder
R elbow
L elbow
R hand
R hip
R knee
Waist
L hip
R foot
L knee
L foot

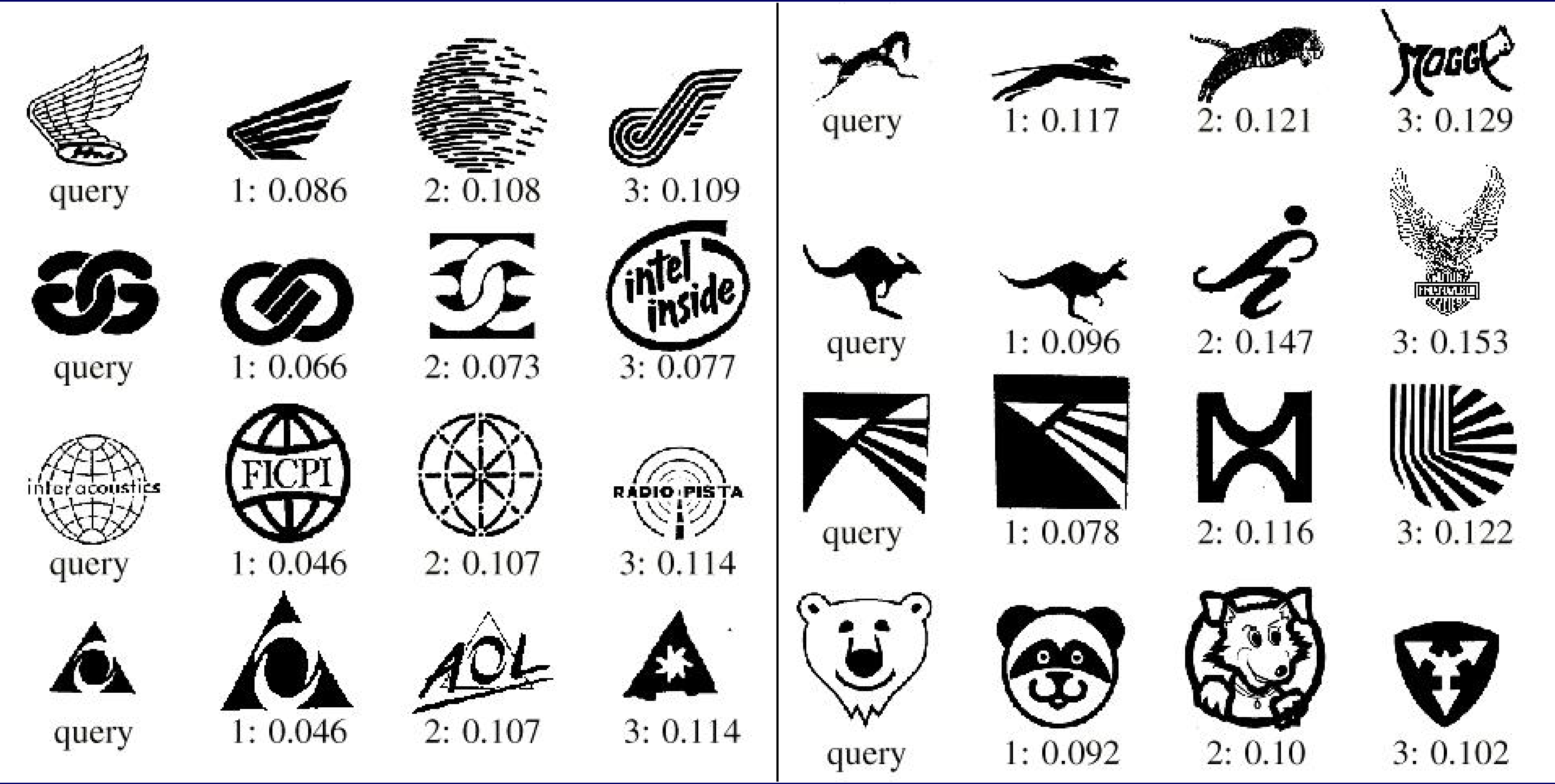




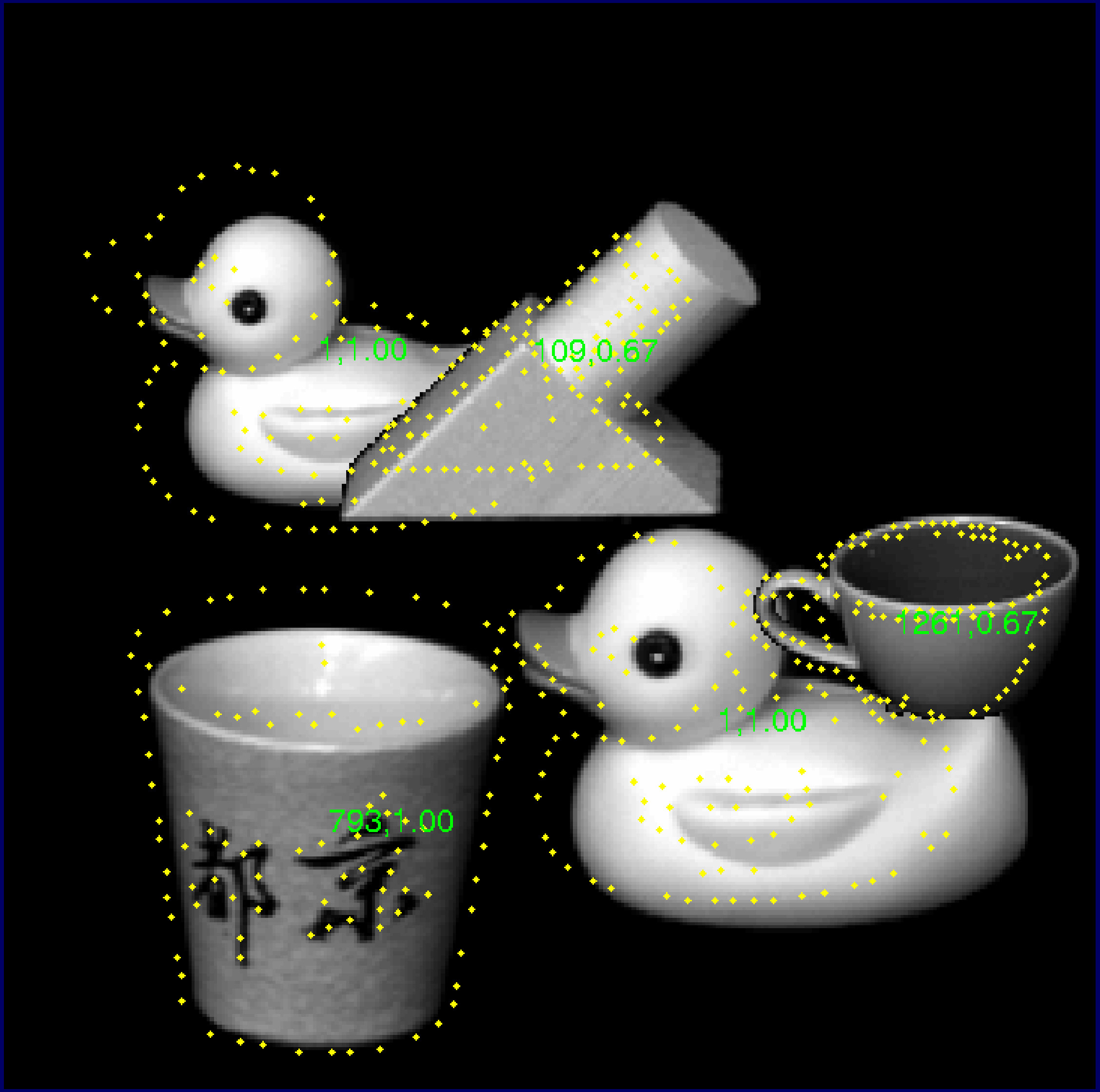


Head
R hand
L hand
L shoulder
R shoulder
R elbow
L elbow
R hip
Waist
L hip
L knee
R knee
L foot
R foot

Trademark Similarity



Recognizing objects in scenes



Shape matching using multi-scale scanning

- Shape context computation (10 Mops)
  - Scales * key-points * contour-points (10*100*10000)
- Multi scale coarse matching (100 Gops)
  - Scales * objects * views * samples * key-points* dim-sc (10*1000*10*100*100*100)
- Deform into alignment (1 Gops)
  - Image-objects * shortlist * (samples)^2 *dim-sc (10*100*10000*100)

Shape matching using grouping

- Complexity determining step: find approx. nearest neighbors of 10^2 query points in a set of 10^6 stored points in the 100 dimensional space of shape contexts.
- Naïve bound of 10^9 can be much improved using ideas from theoretical CS (Johnson-Lindenstrauss, Indyk-Motwani etc)

Putting grouping/segmentation on a sound foundation

- Construct a dataset of human segmented images
- Measure the conditional probability distribution of various Gestalt grouping factors
- Incorporate these in an inference algorithm