

# Recognizing objects and actions in images and video

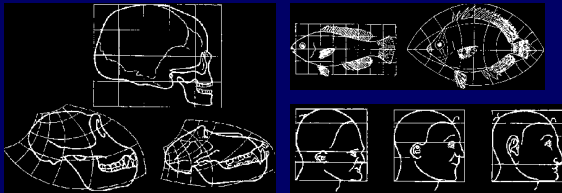
Jitendra Malik

U.C. Berkeley

## Outline

- Finding boundaries
- Recognizing objects
- Recognizing actions

## Biological Shape

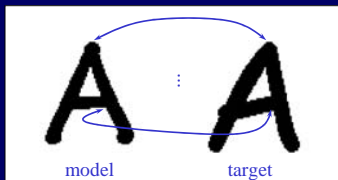


- D'Arcy Thompson: *On Growth and Form*, 1917  
– studied transformations between shapes of organisms

## Deformable Templates: Related Work

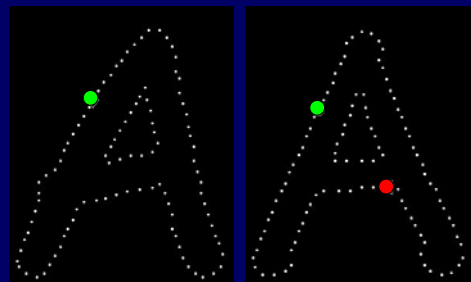
- Fischler & Elschlager (1973)
- Grenander et al. (1991)
- von der Malsburg (1993)

## 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

University of California Berkeley Computer Vision Group

### Shape Context

University of California Berkeley Computer Vision Group

### 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

University of California Berkeley Computer Vision Group

### Comparing Shape Contexts

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]

University of California Berkeley Computer Vision Group

### Matching Framework

model target

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

University of California Berkeley Computer Vision Group

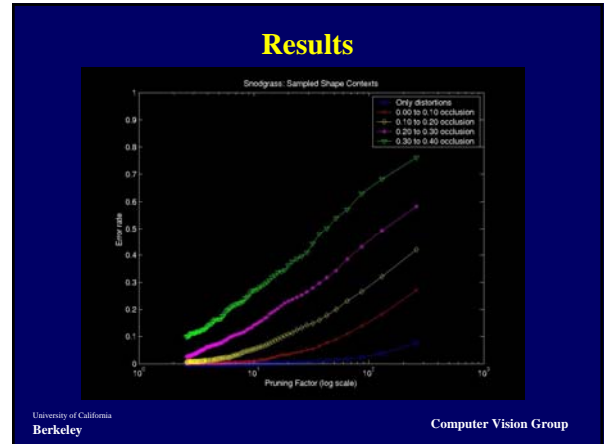
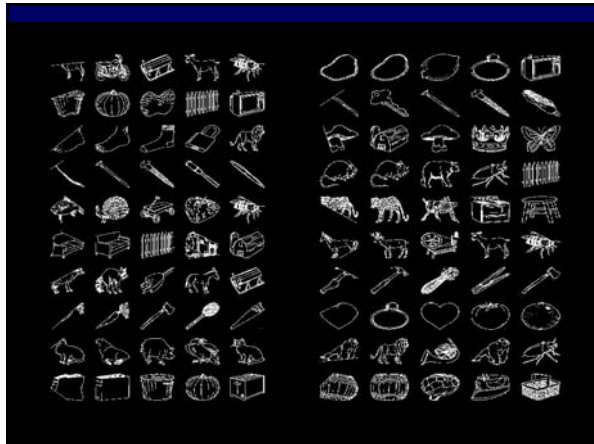
### Fast pruning

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

$$dist(S_{query}, S_i) = \sum_{j=1}^r \chi^2(SC_{query}^j, SC_i^j)$$

$$SC_i^* = \arg \min_u \chi^2(SC_{query}^j, SC_i^u)$$

University of California Berkeley Computer Vision Group



### Matching Framework

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

University of California Berkeley Computer Vision Group

### Thin Plate Spline Model

- 2D counterpart to cubic spline:

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

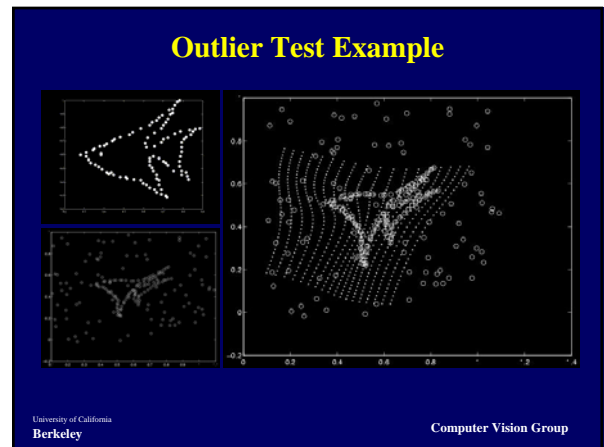
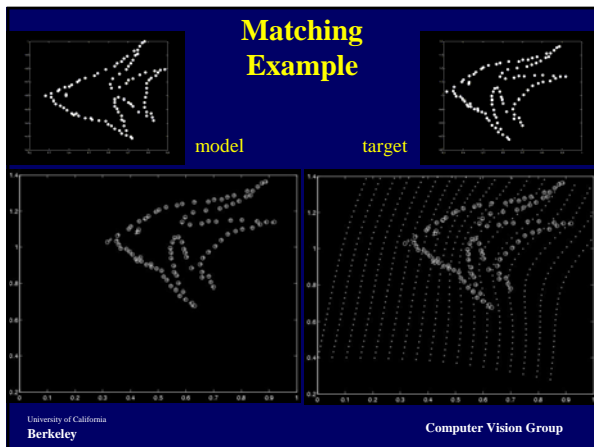
- Minimizes *bending energy*:

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

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

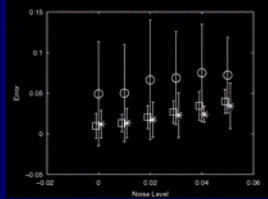
Duchon (1977), Meinguet (1979), Wahba (1991)

University of California Berkeley Computer Vision Group

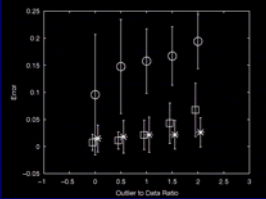


## Synthetic Test Results

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

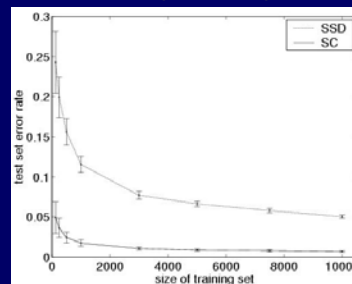
- Handwritten digits
- COIL 3D objects (Nayar-Murase)
- Human body configurations
- Trademarks

## 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%



## Results: Digit Recognition



1-NN classifier using:

Shape context + 0.3 \* bending + 1.6 \* image appearance

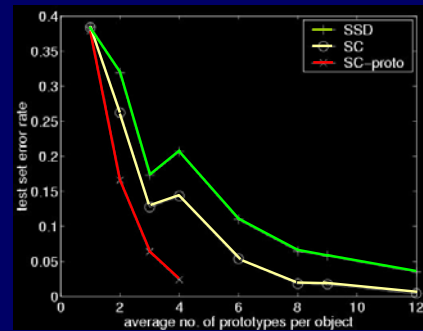
## COIL Object Database



University of California  
Berkeley

Computer Vision Group

## Error vs. Number of Views



University of California  
Berkeley

Computer Vision Group

## Prototypes Selected for 2 Categories



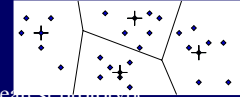
Details in Belongie, Malik & Puzicha (NIPS2000)

University of California  
Berkeley

Computer Vision Group

## Editing: K-medoids

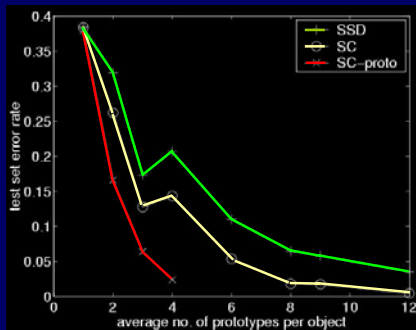
- **Input:** similarity matrix
- **Select:** K prototypes
- **Minimize:** mean distance to nearest prototype
- **Algorithm:**
  - iterative
  - split cluster with most errors
- **Result:** Adaptive distribution of resources (cfr. aspect graphs)



University of California  
Berkeley

Computer Vision Group

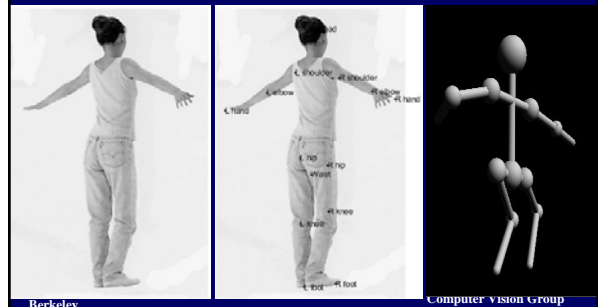
## Error vs. Number of Views



University of California  
Berkeley

Computer Vision Group

## Human body configurations

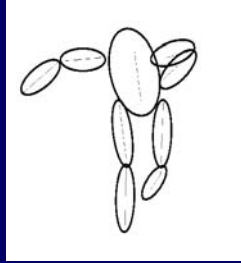


Berkeley

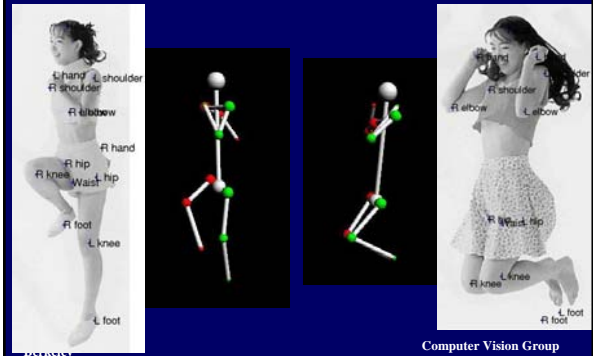
Computer vision Group

## Deformable Matching

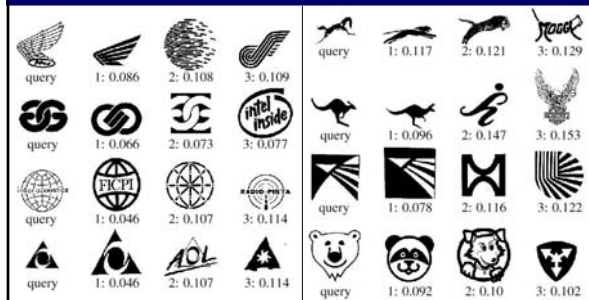
- Kinematic chain-based deformation model
- Use iterations of correspondence and deformation
- Keypoints on exemplars are deformed to locations on query image



## Results



## 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)<sup>2</sup> \* 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