Regression, Classifiers and Correspondence

CS - 4(5)98 Signals AI

Regression

- General idea
  - predictive density of y from observation x
  - ideal: report P(y|x)
    - but usually we don't know this
    - instead, build a model of P(y|x)
- Loss
  - we need a model of the cost of mispredicting y
  - we will minimize expected loss

Common cases

- Model y as f(x)+additive gaussian noise
  - P(Y|X) is N(f(x), Cov)
- Loss is

$$L(y \rightarrow f(x)) = (y - f(x))^2$$

  - known as squared error loss
- f is a linear function
  - we could put a 1 in x if we wanted

$$f(x) = x'\beta$$

Linear Regression

- Stack y's into vector, x's into matrix
  - expected loss over example points is then

$$\frac{1}{N}(Y - X'\beta)'(Y - X'\beta)$$

  - minimize this; minimum is at

$$(XX')^{-1}(XY)$$

  - often problems with rank

K-nearest neighbours

$$f(x) = \frac{1}{K} \sum_{i \in k \text{ examples closest to } x} y_i$$

- This approximates E[Y|X=x]
  - assuming that y changes slowly compared with scatter of samples
  - notice as N, k go to infinity, if k/N goes to zero,
    - then f(x) usually goes to E[Y|X=x]
    -

Locally weighted regression

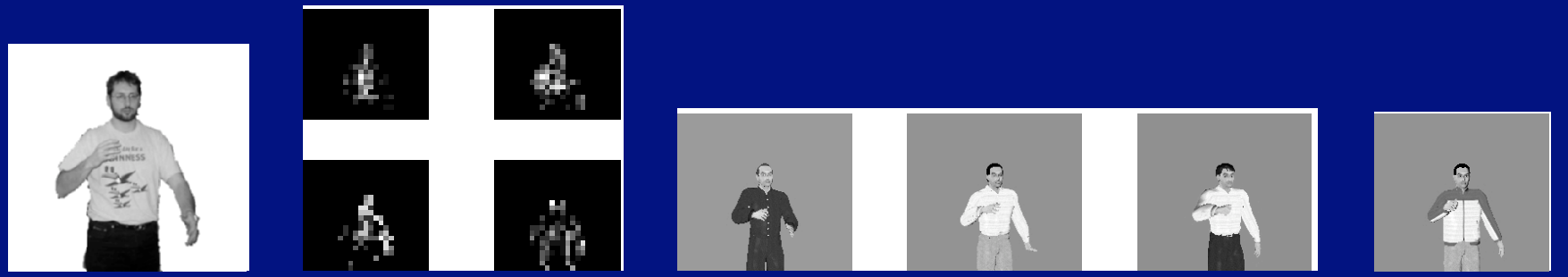
- One can weight the errors in a linear regression

$$(Y - X'\beta)'W(Y - X'\beta)$$

  - solution at

$$(XWX')^{-1}(XWY)$$

- LWR: weights chosen as a function of query x
  - weight points higher if they're closer
  - cost: new linear system every time you want to predict
  - advantage: locally smoothed predictions

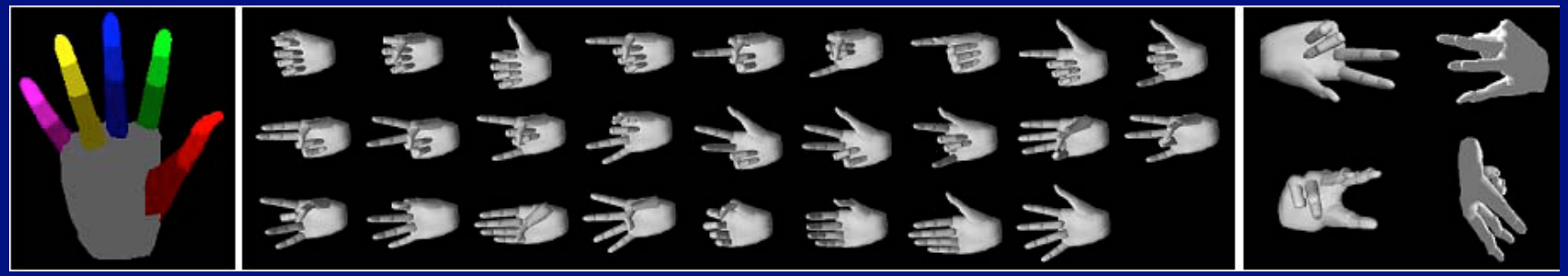


(a) Input (b) A subset of the features (c) 3 top matches in PSH, left to right (d) Robust LWR

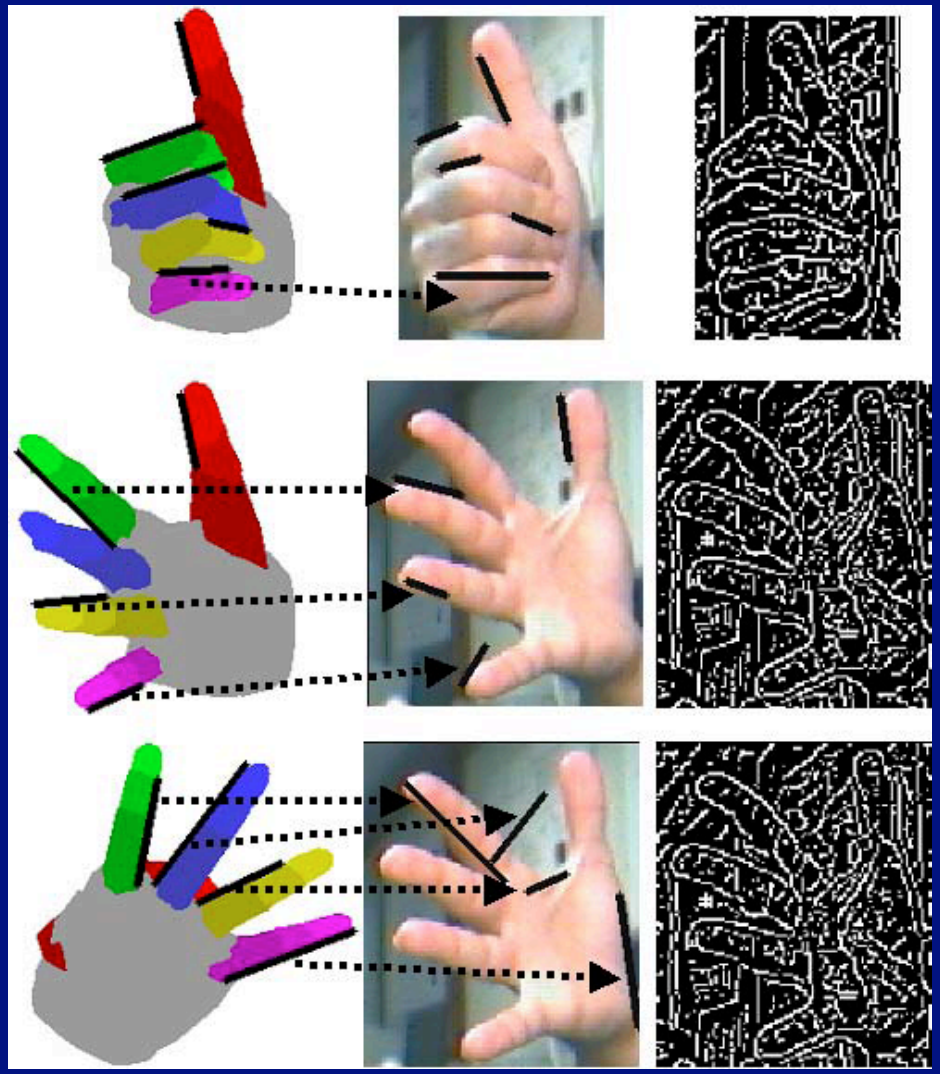
Figure 1. Pose estimation with parameter-sensitive hashing and local regression.

Shakhnarovich, Viola, Darrell, 03

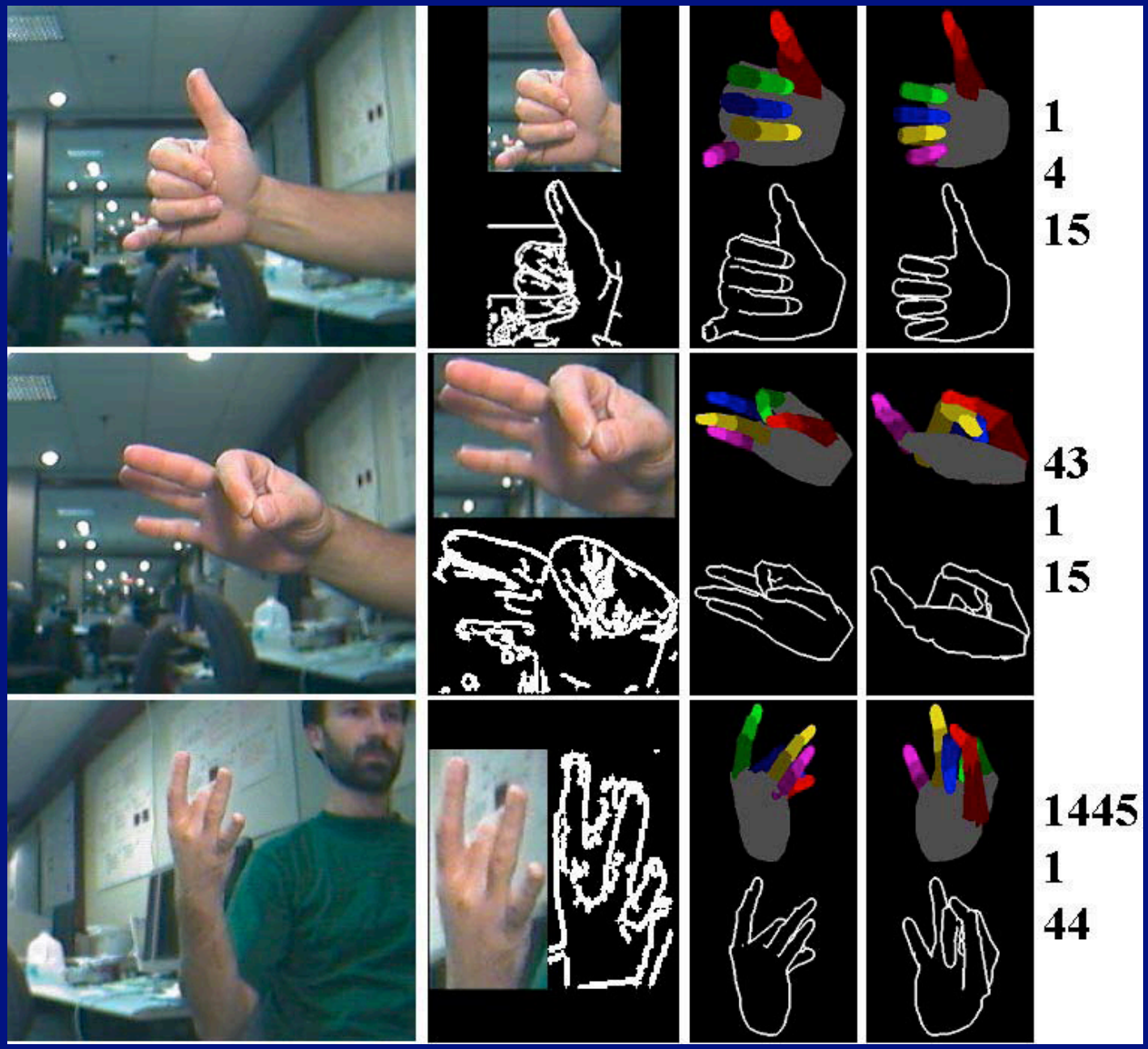
Shakhnarovich, Viola, Darrell, 03



Athitsos+Sclaroff 03



Athitsos+Sclaroff 03



Athitsos+Sclaroff 03

The curse of dimension

- In high dimensions, volume is on the “skin” of a body
  - e.g. high dimensional cube
- Example: uniform data in unit cube in dimension p
  - want fraction r of data to be in subcube
  - so subcube must have volume r
  - so edge length must be r^(1/p)

  - numbers: p=10, r=0.1 gives edge length of 0.794
  - hardly local!

Bias and variance

- Consider some deterministic function y=f(x)
- draw a bunch of (x, y) samples using a sampling density T
- estimate value with 1-NN

$$
\begin{aligned}
\mathrm{MSE}(x_o) &= E_T[(f(x_0) - \hat{y}_0)^2] \\
&= f(x_0)^2 - 2E_T[y_0]f(x_0) + E_T[y_0^2] \\
&= f(x_0)^2 - 2E_T[y_0]f(x_0) + (E_T[y_o])^2 \\
&\quad +(E_T[y_0])^2 - 2(E_T[y_0])^2 + E_T[y_0^2] \\
&= (f(x_0) - E_T[y_0])^2 + E_T[(y_0 - E_T[y_0])]^2
\end{aligned}
$$

BIAS VARIANCE

Bias and Variance

- Bias
  - what happens because your model cannot fit the data, however well the parameters
- Variance
  - the result of not being able to estimate the parameters correctly, which occurs because different sets of samples of the same underlying density give different estimates

Regularizing linear regression

- Options
  - throw out variables
    - but which ones? search
    - variance could go up, because different samples might result in different models
  - penalize large coefficients
    - rank interpretation
    - noise interpretation
    - usually a really bad idea to penalize the constant offset!
      - center data

Ridge regression

- Penalize large coefficients, so minimize

$$(Y - X'\beta)'(Y - X'\beta) + \lambda\beta'\beta$$

- This gives

$$\beta^{ridge} = (XX' + \lambda I)^{-1}XY$$

- Notice this is equivalent to min $(Y - X'\beta)'(Y - X'\beta)$
  - subject to

$$\beta'\beta \leq s$$

Ridge regression

- Notice this isn't covariant under scaling of inputs
  - be careful about relative scaling of variables
- We can show that ridge regression scales directions of SVD of X by

$$\frac{d^2}{d^2 + \lambda}$$

  - shrinkage is most pronounced in directions of x with low variance
    - because we have poor estimates of the gradient of y in these directions

The Lasso

$$\beta^{\text{lasso}} = \arg\min \sum_{i=1}^{i=N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Subject to:

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

The Lasso

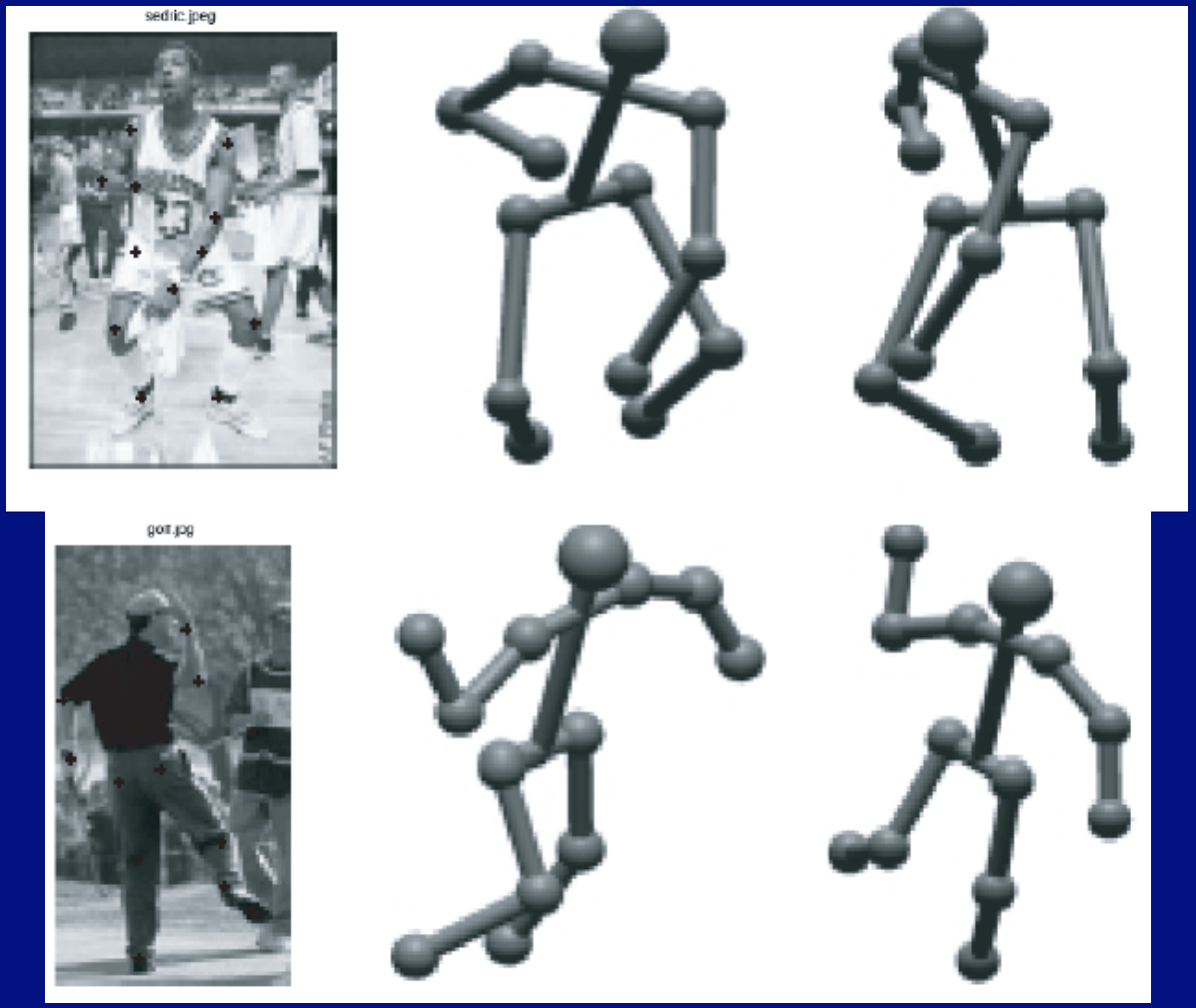
- No longer use linear algebra,
  - optimization problem is nasty
- Sufficiently small t forces components to be zero

3D from 2D for single views of humans

- Can one solve?
  - yes
- How?
  - rather naturally regression

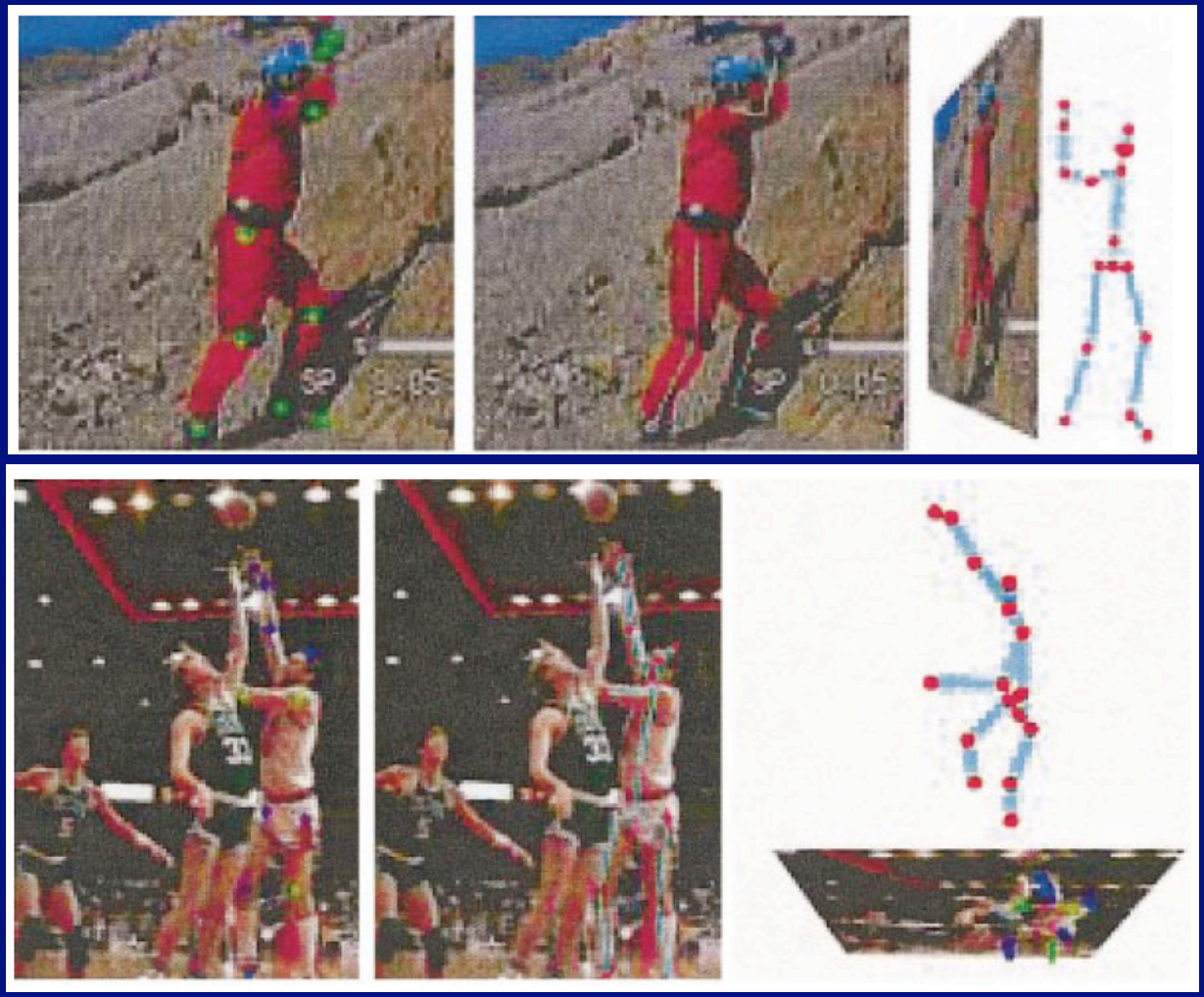
- Is this problem ambiguous?
  - a large literature says yes
  - a large literature says no

Unambiguous



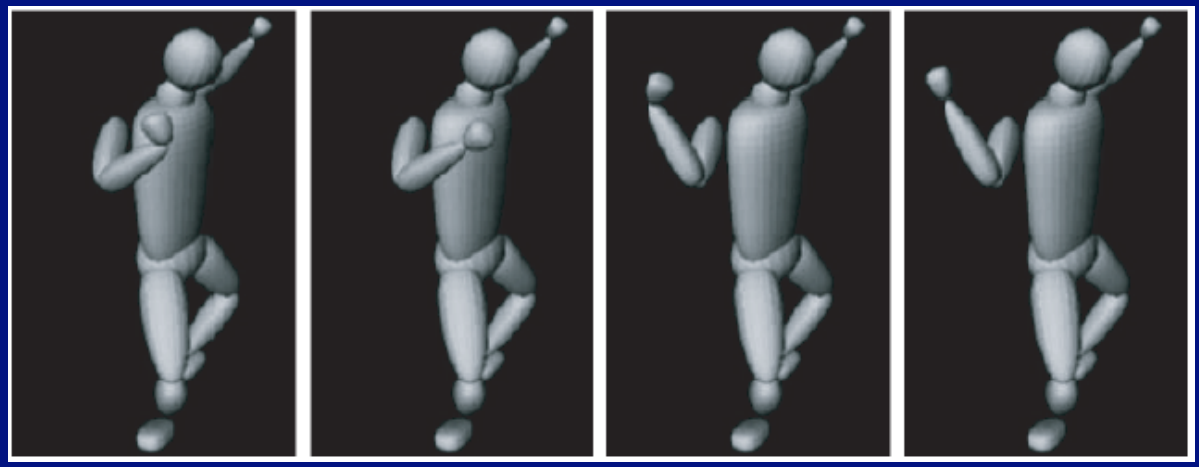
Taylor, 00

Unambiguous



Barron +Kakadiaris, 00

Ambiguous



Sminchisescu+Triggs, 03

Ambiguous

Agarwal and Triggs 05

Does ambiguity exist?

- Human tracking is multimodal
  - popular opinion
- Not even in single frames
  - Taylor 2000; Barron+Kakadiaris 00; Shakhnarovich ea 03;
- In single frames, but not over somewhat longer time scales
  - Sminchicescu+Triggs 01, 03, 03; Agarwal+Triggs 05, 06
- Not if you lift from multiple frames
  - Howe ea 00; Howe 04; Ramanan+Forsyth 03
- Some ambiguities persist over very long time scales
  - left+right leg, for example

Ambiguity

- The literature is confused
- There are not reliable or compelling experiments
- There *may* be very little ambiguity or quite a lot
- It is crucial to clean this point up
  - however, one must be careful --- frequency arguments are dangerous.

Basic ideas in classifiers

- Loss
  - one can refuse to classify
- Total risk

$$R(s) = Pr\{1 \to 2|\text{using } s\} L(1 \to 2) + Pr\{2 \to 1|\text{using } s\} L(2 \to 1)$$

- Expected loss of classifying a point gives

1 if $p(1|\boldsymbol{x})L(1 \to 2) > p(2|\boldsymbol{x})L(2 \to 1)$

2 if $p(1|\boldsymbol{x})L(1 \to 2) < p(2|\boldsymbol{x})L(2 \to 1)$

Known class-conditional densities

- Assume class-conditional densities are Gaussian

$$p(\underline{x}|k) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_k \mid^{1/2}} \exp\left(\frac{-1}{2}(\underline{x} - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{x} - \underline{\mu}_k)\right)$$

$$p(k|\underline{x}) \propto \frac{1}{(2\pi)^{d/2} \mid \Sigma_k \mid^{1/2}} \exp\left(\frac{-1}{2}(\underline{x} - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{x} - \underline{\mu}_k)\right) \pi_k$$

For some cases, you could evaluate this, perhaps by estimating each mean and covariance

Mahalonobis distance

- Pick class that minimizes

$$\left((\underline{x} - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{x} - \underline{\mu}_k)\right) - 2 \log \pi_k$$

- Notice that if the covariance is the same for each class
  - test boils down to a linear expression

Classification by regression

- Regress a class label against data
  - works well for k-NN
  - not so well for linear regression with least squares
    - problem with the loss - it overcharges for bad mistakes

Histogram based classifiers

- Represent class-conditional densities with histogram
- Advantage:
  - estimates become quite good
    - (with enough data!)
- Disadvantage:
  - Histogram becomes big with high dimension
    - but maybe we can assume feature independence?

Finding skin

- Skin has a very small range of (intensity independent) colours, and little texture
  - Compute an intensity-independent colour measure, check if colour is in this range, check if there is little texture (median filter)
  - See this as a classifier - we can set up the tests by hand, or learn them.

Histogram classifier for skin

$$\frac{P(rgb \mid skin)}{P(rgb \mid \neg skin)} \geq \Theta$$

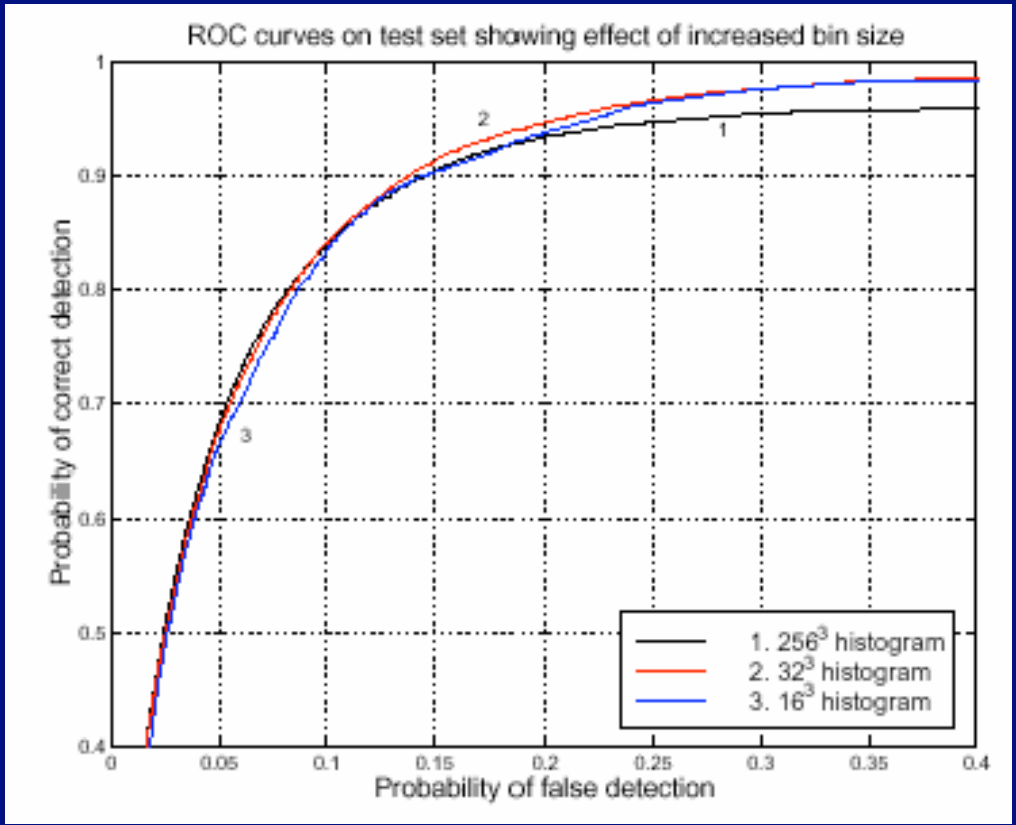


Figure from Jones+Rehg, 2002

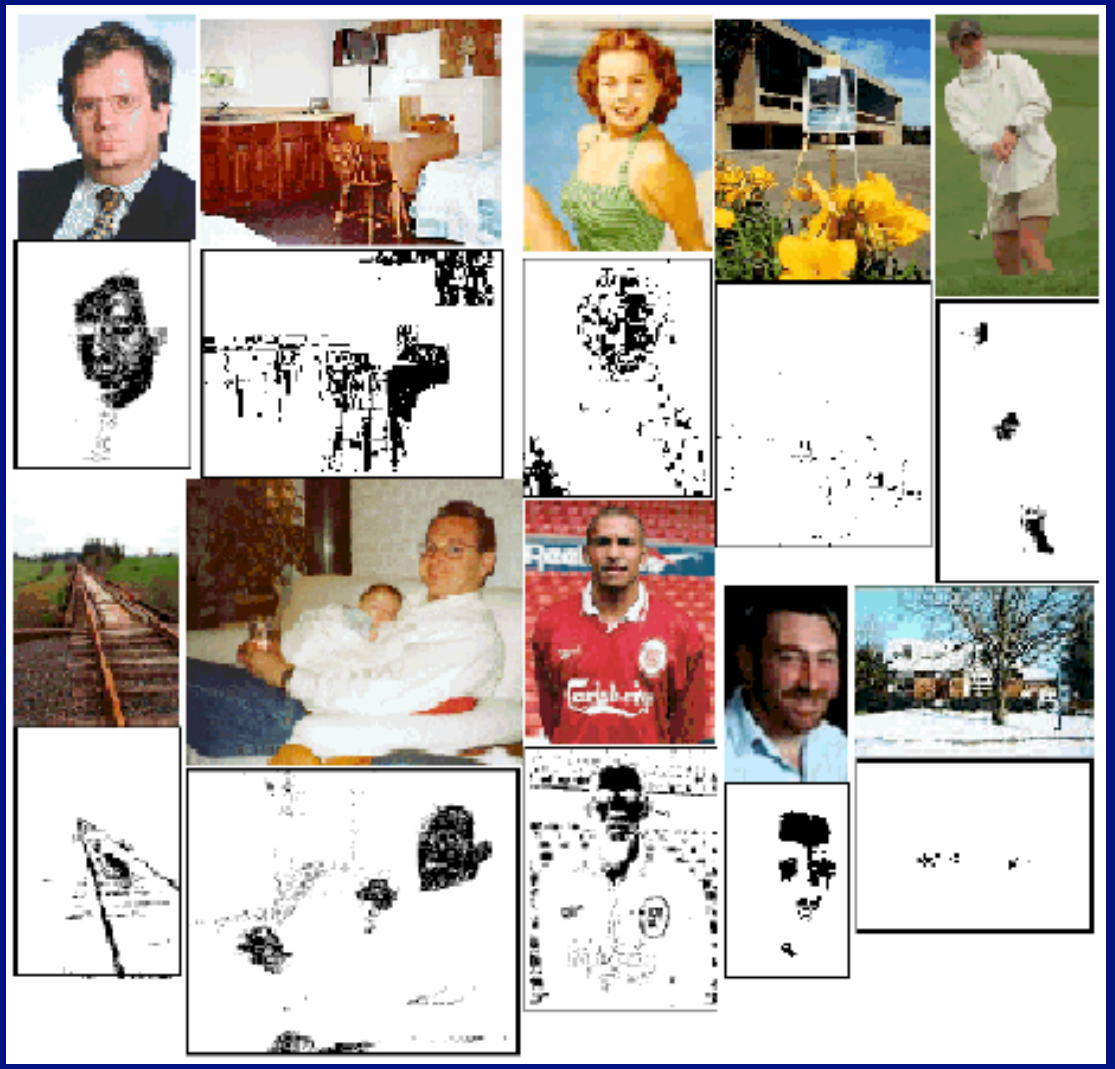


Figure from Jones+Rehg, 2002

Curse of dimension - I

- This won’t work for many features
  - try R, G, B, and some texture features
  - too many histogram buckets

Finding faces

- Faces “look like” templates (at least when they’re frontal).
- General strategy:
  - search image windows at a range of scales
  - Correct for illumination
  - Present corrected window to classifier
- Issues
  - How corrected?
  - What features?
  - What classifier?
  - what about lateral views?

The Thatcher Illusion
Figures by Henry Rowley,
http://www.cs.cmu.edu/~har/puzzle.html

The Thatcher Illusion
Figures by Henry Rowley,
http://www.cs.cmu.edu/~har/puzzle.html

Naive Bayes

- Previously, we detected with a likelihood ratio test

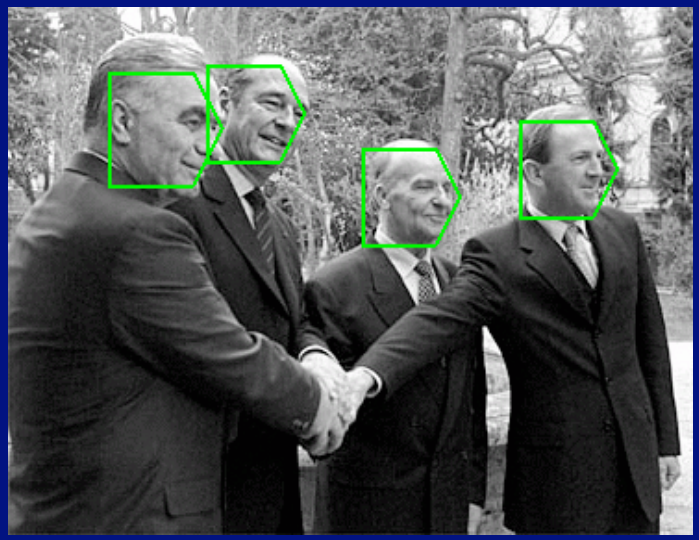
$$\frac{P(\text{features}|\text{event})}{P(\text{features}|\text{not event})} > \text{threshold}$$

- Now assume that features are conditionally independent given event

$$P(f_0, f_1, f_2, \ldots, f_n|\text{event}) = P(f_0|\text{event})P(f_1|\text{event})P(f_2|\text{event})\ldots P(f_n|\text{event})$$

Naive Bayes

- (not necessarily perjorative)
- Histogram doesn't work when there are too many features
  - the curse of dimension, first version
  - assume they're independent conditioned on the class, cross fingers
  - reduction in degrees of freedom
  - very effective for face finders
    - relations may not be all that important
  - very effective for high dimensional problems
    - bias vs. variance



Work by Schneiderman and Kanade,
http://www.cs.cmu.edu/afs/cs.cmu.edu/user/hws/www/hws.html

Work by Schneiderman and Kanade,
http://www.cs.cmu.edu/afs/cs.cmu.edu/user/hws/www/hws.html

Many more face finders on the face detection home page
http://home.t-online.de/home/Robert.Frischholz/face.htm

Face Recognition

- Whose face is this? (perhaps in a mugshot)
- Issue:
  - What differences are important and what not?
  - Reduce the dimension of the images, while maintaining the “important” differences.
- One strategy:
  - Principal components analysis, then nearest neighbours
  - Many face recognition strategies at http://www.cs.rug.nłusers/peterkr/FACE/face.html

Curse of dimension-II

- General phenomenon of high dimensions
  - volume is concentrated at the boundary
- Parameter estimation is hard for high dimensional distributions
  - even Gaussians
    - where probability is concentrated further and further from the mean
    - and covariance has too many parameters
      - dodge: assume covariance is diagonal
- Idea: reduce the dimension of the feature set
  - Principal components
  - Linear discriminants

Principal components

- Find linear features that explain most of the variance of the data

Assume we have a set of n feature vectors $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ in $\mathbb{R}^d$. Write

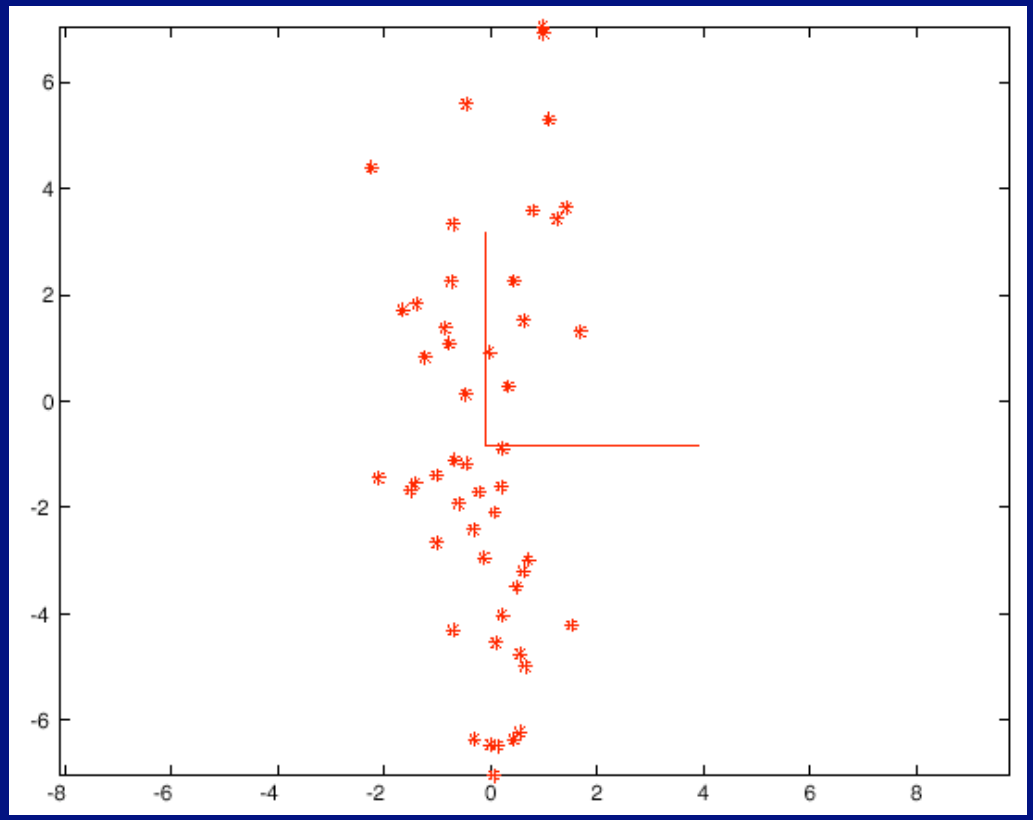
$$\boldsymbol{\mu} = \frac{1}{n} \sum_i \boldsymbol{x}_i$$

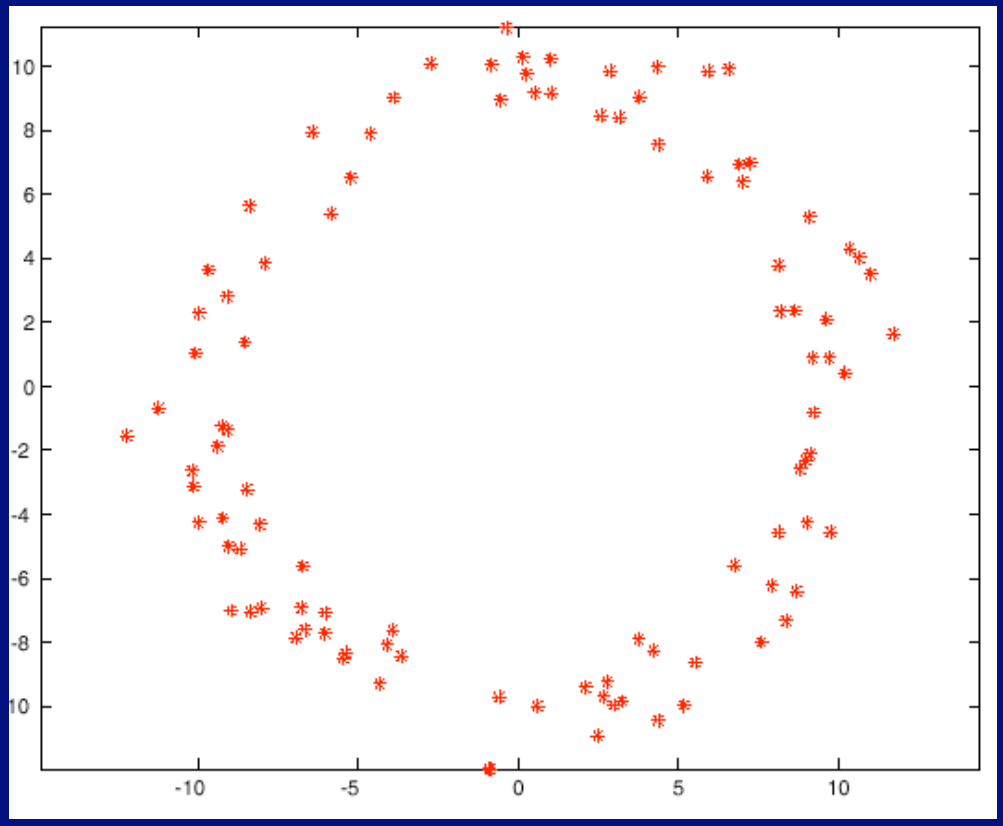
$$\Sigma = \frac{1}{n-1} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

The unit eigenvectors of Σ — which we write as $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$, where the order is given by the size of the eigenvalue and $\boldsymbol{v}_1$ has the largest eigenvalue — give a set of features with the following properties:

- They are independent.
- Projection onto the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ gives the k-dimensional set of linear features that preserves the most variance.

Algorithm 22.5: *Principal components analysis identifies a collection of linear features that are independent, and capture as much variance as possible from a dataset.*

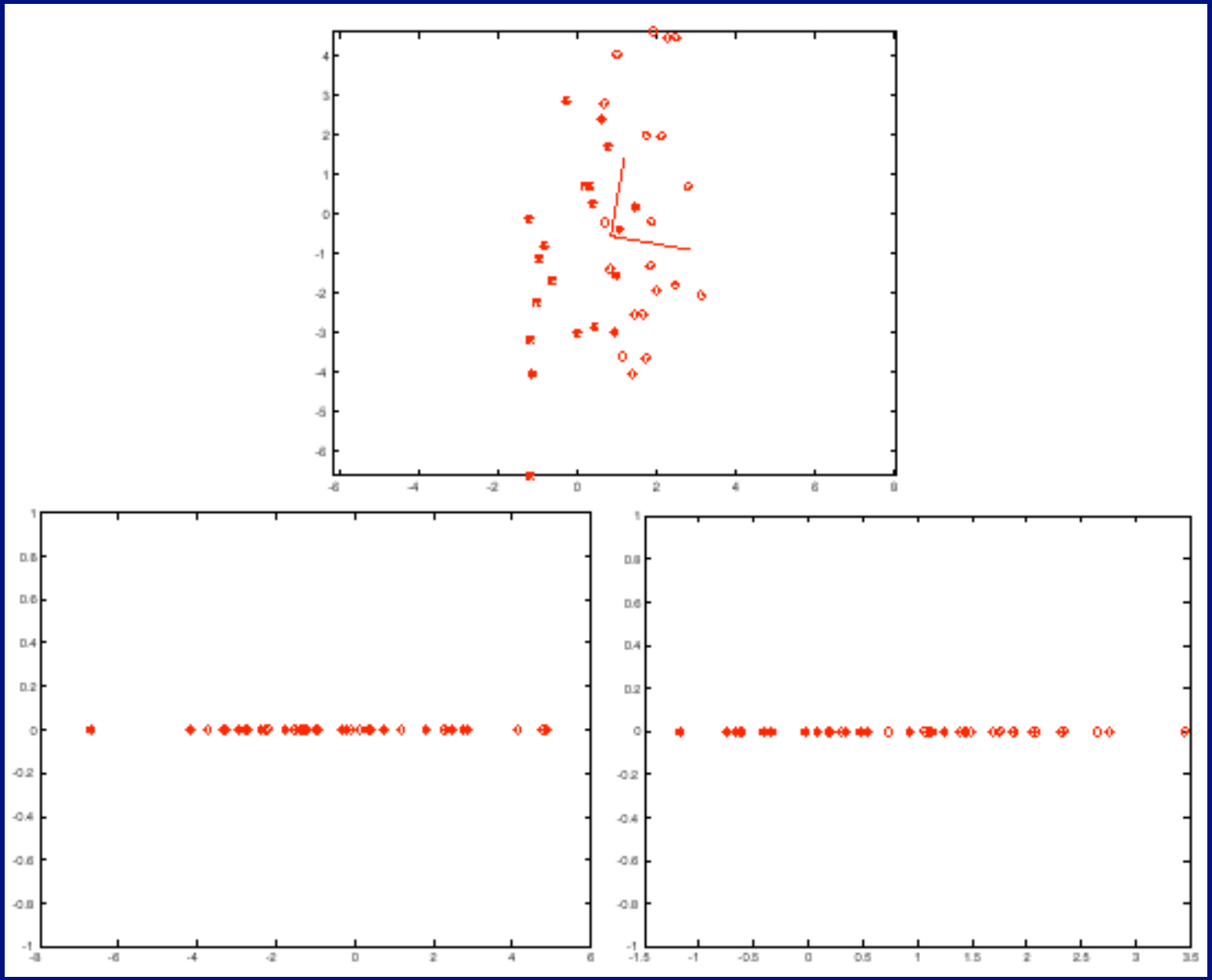

6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6
8

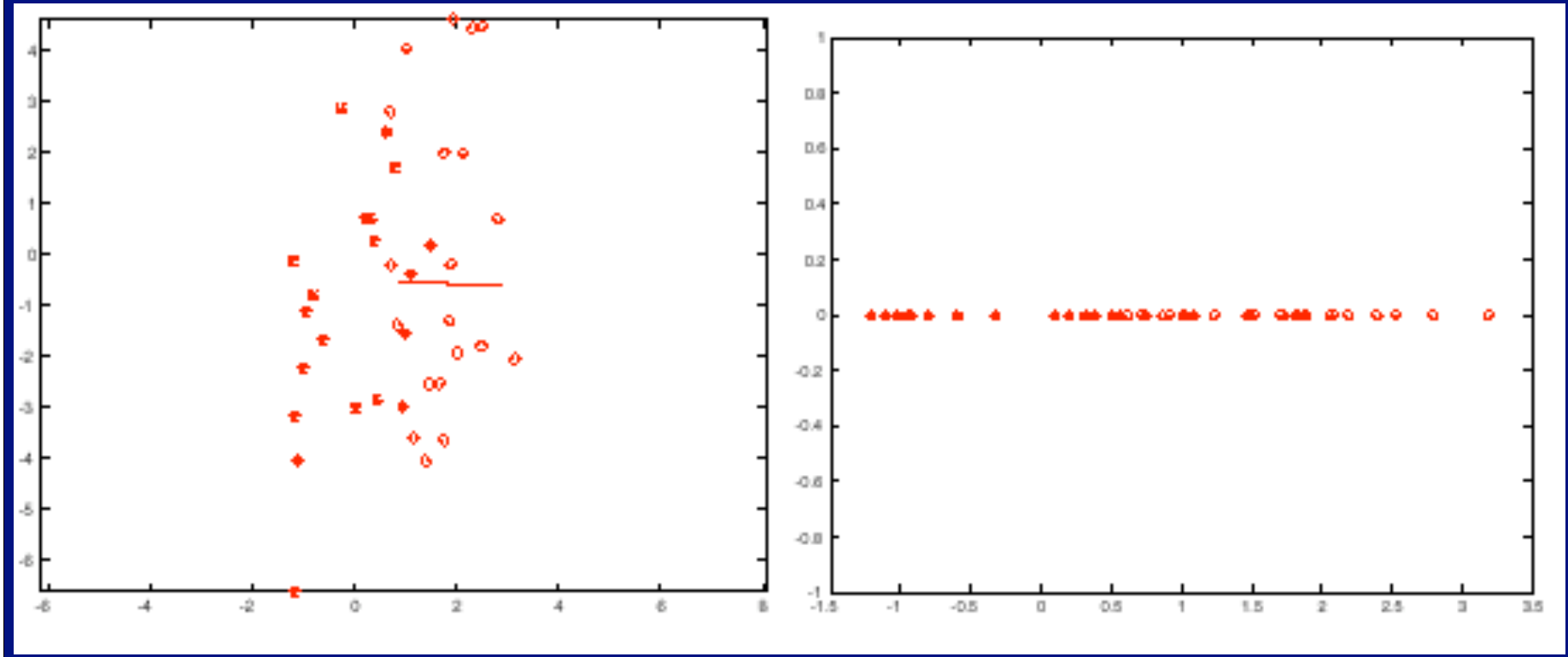

10
8
6
4
2
0
-2
-4
-6
-8
10
-10
-5
0
5
10

Principal components for face images, from
http://vismod.www.media.mit.edu/vismod/demos/facerec/basic.html

Linear discriminant analysis

- Principal components do not preserve discrimination
  - so we could have features that don't distinguish, see picture
- Assume (pretend) class conditional densities are normal, with the same covariance
  - Choose linear features so that
    - between class variation is big compared to within class variation
      - between class variation
        - covariance of class means
      - within class variation
        - class covariance




-6
-4
-2
0
2
4
6
8
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5

Assume that we have a set of data items of g different classes. There are n_k items in each class, and a data item from the k'th class is $\boldsymbol{x}_{k,i}$, for $i \in \{1, \ldots, n_k\}$. The j'th class has mean $\boldsymbol{\mu}_j$. We assume that there are p features (i.e. that the $\boldsymbol{x}_i$ are p-dimensional vectors).
Write $\overline{\boldsymbol{\mu}}$ for the mean of the class means, i.e.

$$\overline{\boldsymbol{\mu}} = \frac{1}{g} \sum_{j=1}^{g} \boldsymbol{\mu}_j$$

Write

$$\mathcal{B} = \frac{1}{g-1} \sum_{j=1}^{g} (\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}})^T$$

Assume that each class has the same covariance Σ, which is either known or estimated as

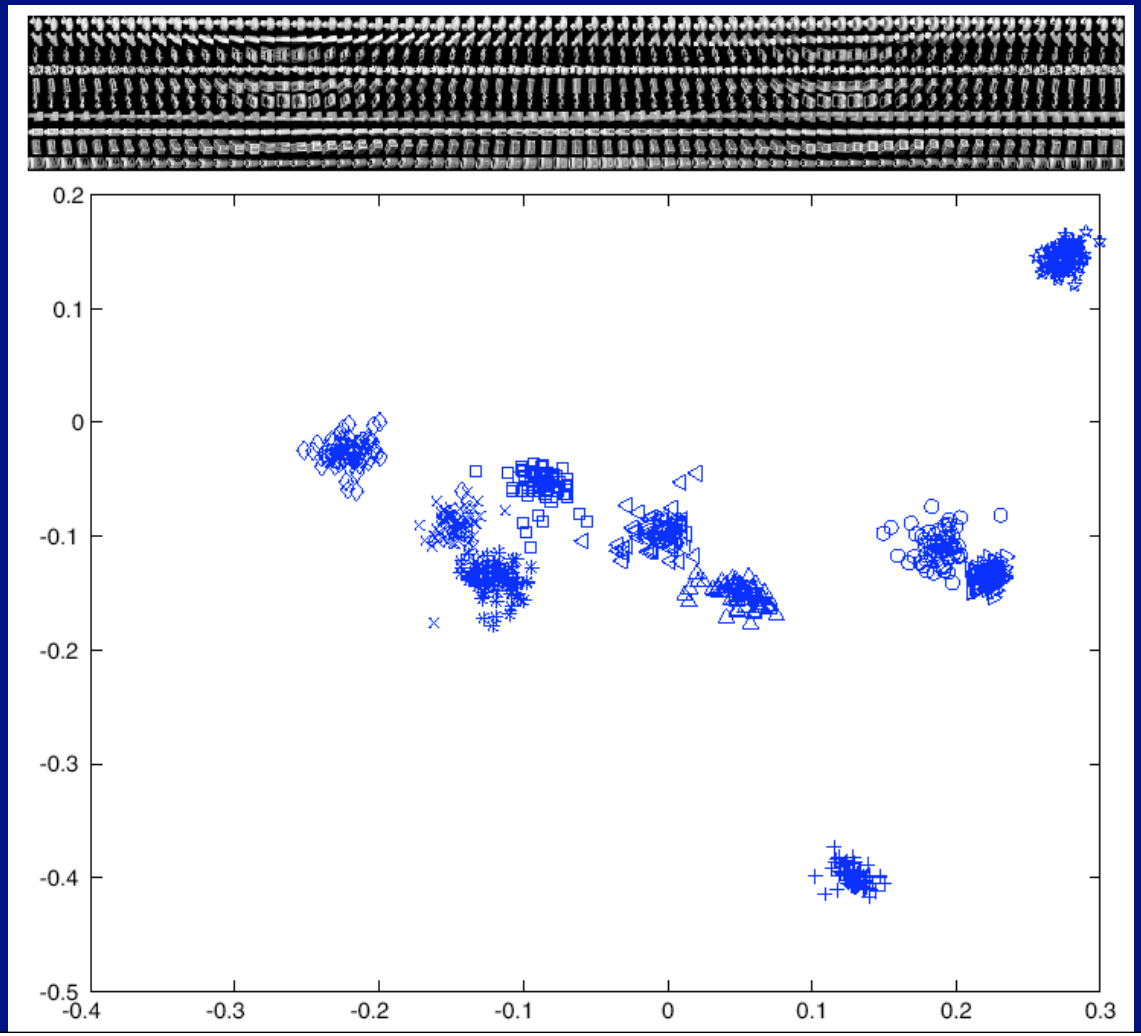
$$\Sigma = \frac{1}{N-1} \sum_{c=1}^{g} \left\{ \sum_{i=1}^{n_c} (\boldsymbol{x}_{c,i} - \boldsymbol{\mu}_c)(\boldsymbol{x}_{c,i} - \boldsymbol{\mu}_c)^T \right\}$$

The unit eigenvectors of $\Sigma^{-1}\mathcal{B}$ — which we write as $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$, where the order is given by the size of the eigenvalue and $\boldsymbol{v}_1$ has the largest eigenvalue — give a set of features with the following property:

- Projection onto the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ gives the k-dimensional set of linear features that best separates the class means.

Algorithm 22.6: *Canonical variates identifies a collection of linear features that separating the classes as well as possible.*

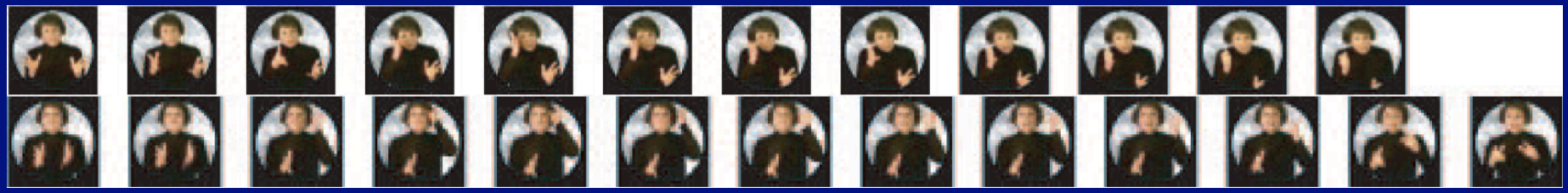
First two canonical variates for well known image collection



ASL translation

ASL

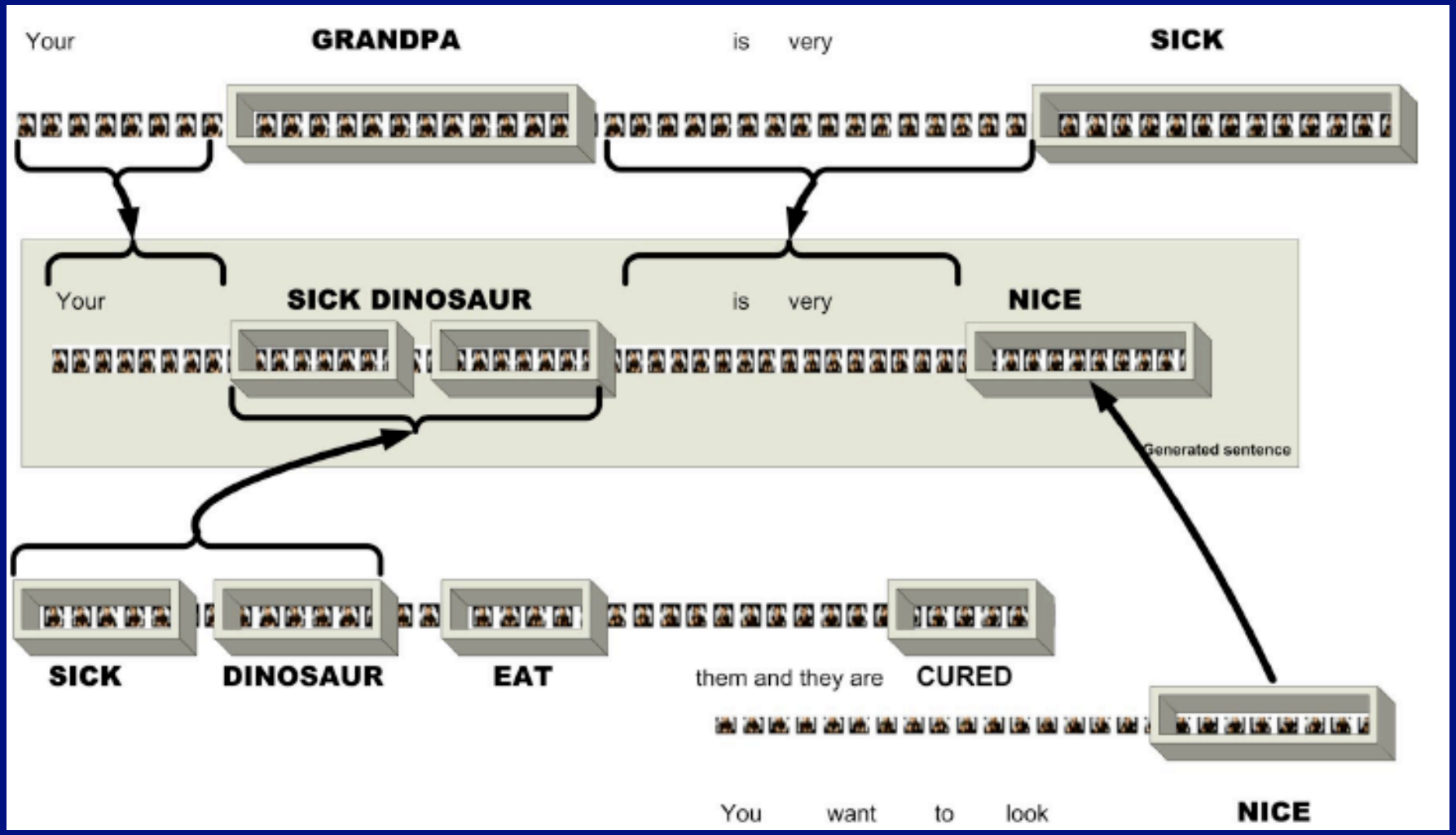
Grandma



Grandpa

- Notice:
  - intra-class variation (timing, role of hands)
  - carefully dressed narrator
  - low resolution
- Would like to spot words in text

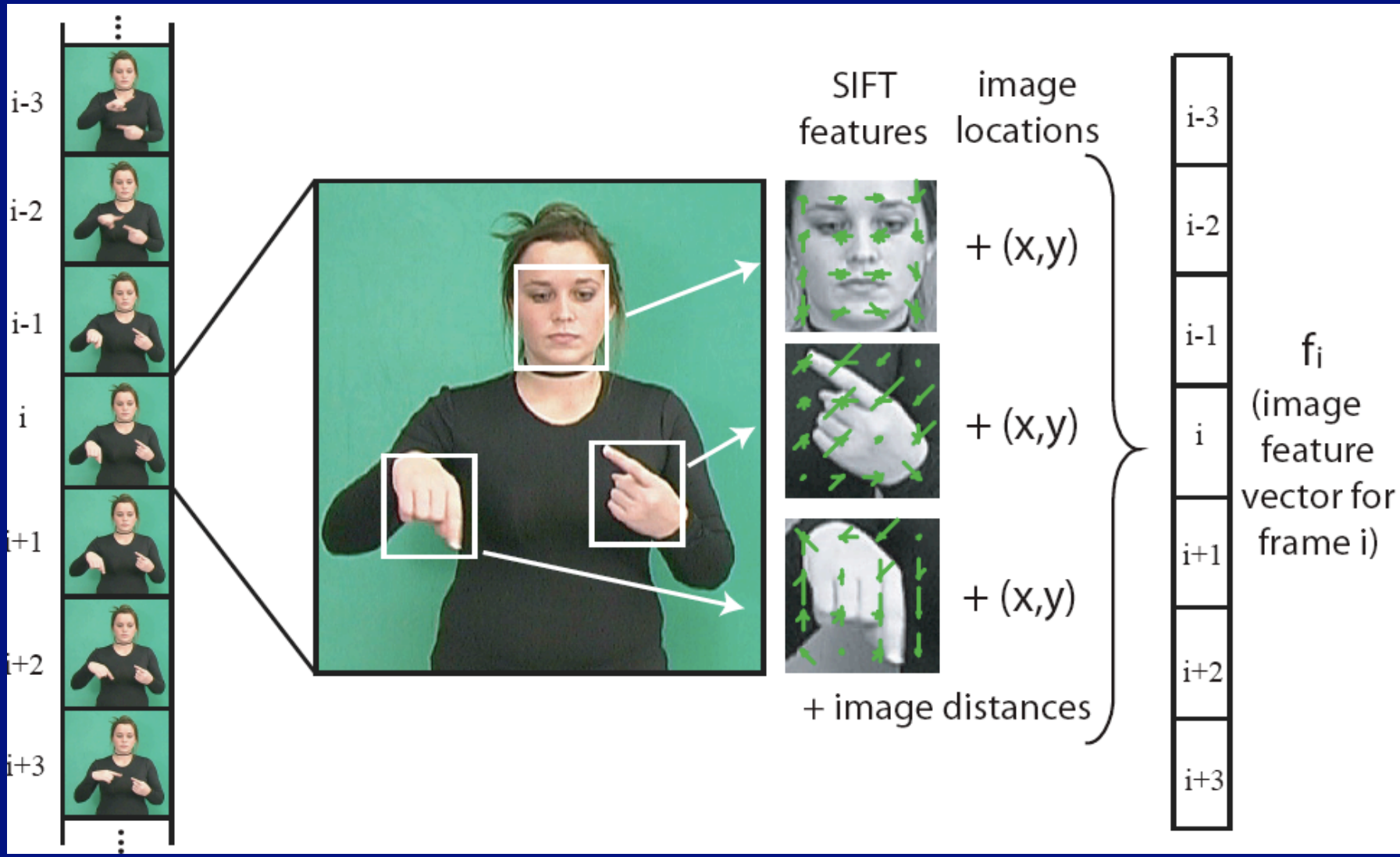
Wordspotting allows translations



Wordspotting

- Difficulties
  - building a discriminative word model for a large vocabulary is hard
    - need lots of examples of each word
  - building a generative word model is hard, too
    - no widely available pronunciation dictionary
  - very large number of features
    - next slide
- Strategy
  - build spotters for some words (base words)
  - now use the output of those spotters as features for other words
    - uses less training data
    - because features are discriminative

Features

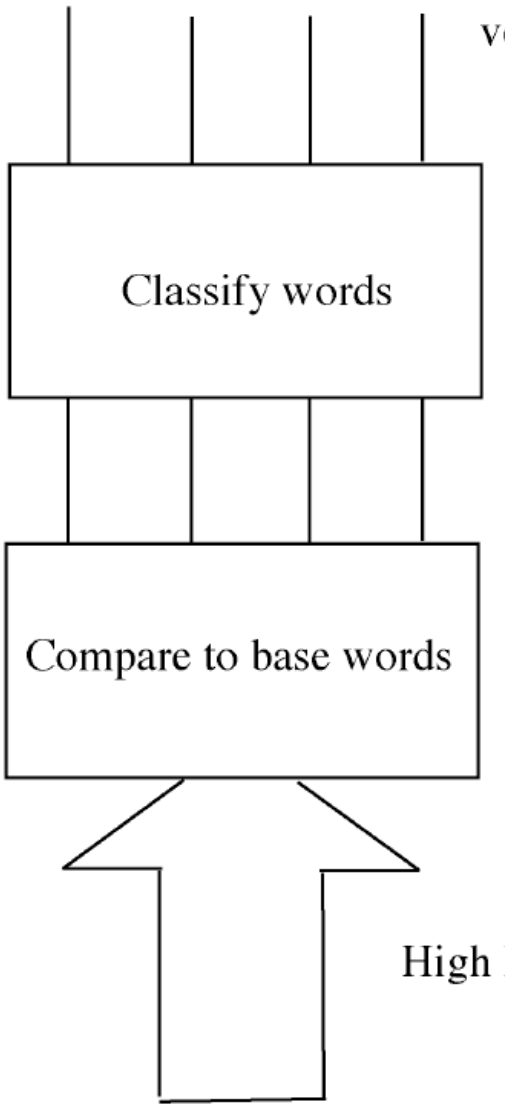

i-3
i-2
i-1
i
i+1
i+2
i+3
SIFT features
image locations
+ (x,y)
+ (x,y)
+ (x,y)
+ image distances
i-3
i-2
i-1
i
i+1
i+2
i+3
f_i
(image feature vector for frame i)

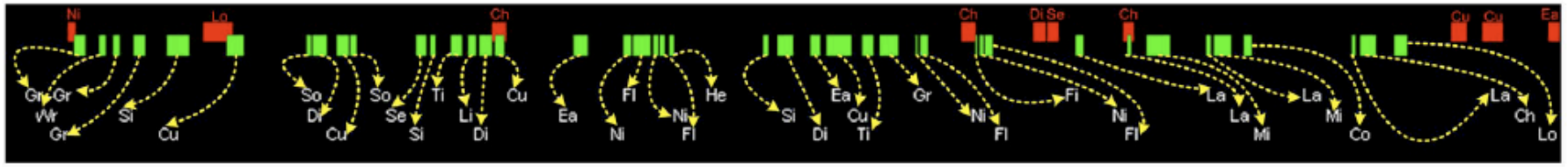
Logistic Regression

- Build a parametric model of the posterior,
  - p(class|information)
- For a 2-class problem, assume that
  - log(P(1|data))-log(P(0|data))=linear expression in data
- Training
  - maximum likelihood on examples
  - problem is convex
- Classifier boundary
  - linear

Multiclass classification

- Many strategies
  - 1-vs-all
    - for each class, construct a two class classifier comparing it to all other classes
    - take the class with best output
      - if output is greater than some value
  - Multiclass
    - log(P(i|features))-log(P(k|features))=(linear expression)
    - many more parameters
    - harder to train with maximum likelihood
    - still convex

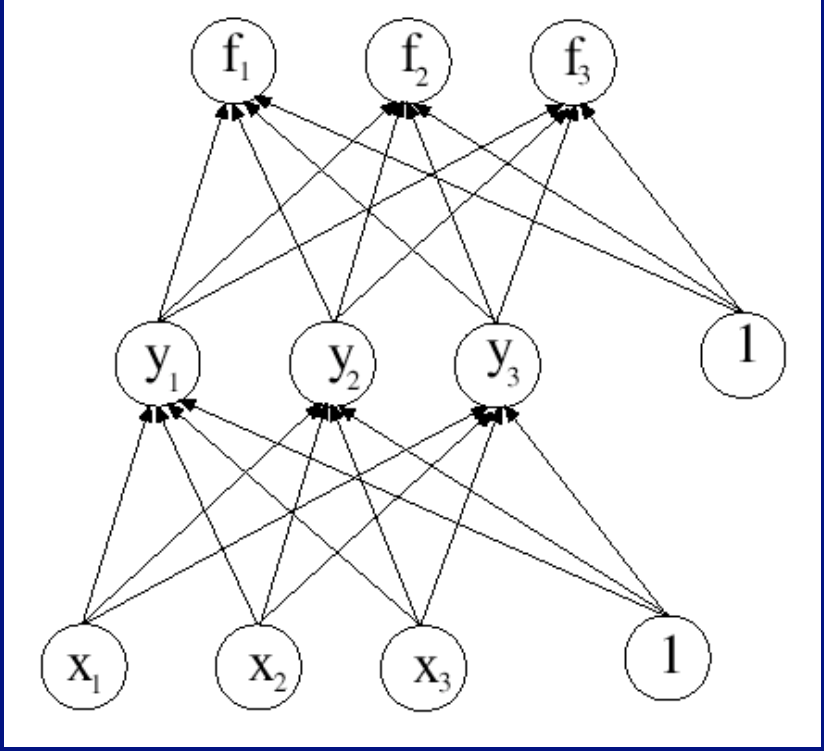

One-vs-All classifier outputs for vocabulary
Classify words
One-vs-All classifier outputs for base words (30 features, all rather discriminative)
Compare to base words
High Dimensional Input Features



Grandma? What is it? What's wrong with grandpa ?
your grandpa is very sick,littlefoot.
ill ? he'll be cured, won't he ?
I don't know, littlefoot. some dinosaurs cure, and some don't.
I've seen this sickness many times in my life. No dinosaur ever cure, unless--
Unless what ?
Unless they eat the golden petals of the night flower.
The night flower ? did you hear that ? the night flower ?
Yes, golden petals. sick dinosaurs eat them and are cured, if they eat them in time.
Grandma, we have to get the night flower for grandpa.
Old one, where can I find the night flower ?
In the land we came from, the land of mists.
The land of mists.
Cousins, who will take me to the night flower ?
Not me. I'm not going back there.
The land has changed too much. Long necks are not welcome there.

Neural networks

- Logistic regression heavily generalized



$$\boldsymbol{g}(\boldsymbol{x}) \approx \boldsymbol{f}(\boldsymbol{x}) = [\phi(\boldsymbol{w}_{21} \cdot \boldsymbol{y}), \phi(\boldsymbol{w}_{22} \cdot \boldsymbol{y}), \ldots \phi(\boldsymbol{w}_{2n} \cdot \boldsymbol{y})]$$

$$\boldsymbol{y}(\boldsymbol{z}) = [\phi(\boldsymbol{w}_{11} \cdot \boldsymbol{z}), \phi(\boldsymbol{w}_{12} \cdot \boldsymbol{z}), \ldots \phi(\boldsymbol{w}_{1m} \cdot \boldsymbol{z}), 1]$$

$$\boldsymbol{z}(\boldsymbol{x}) = [x_1, x_2, \ldots, x_p, 1]$$

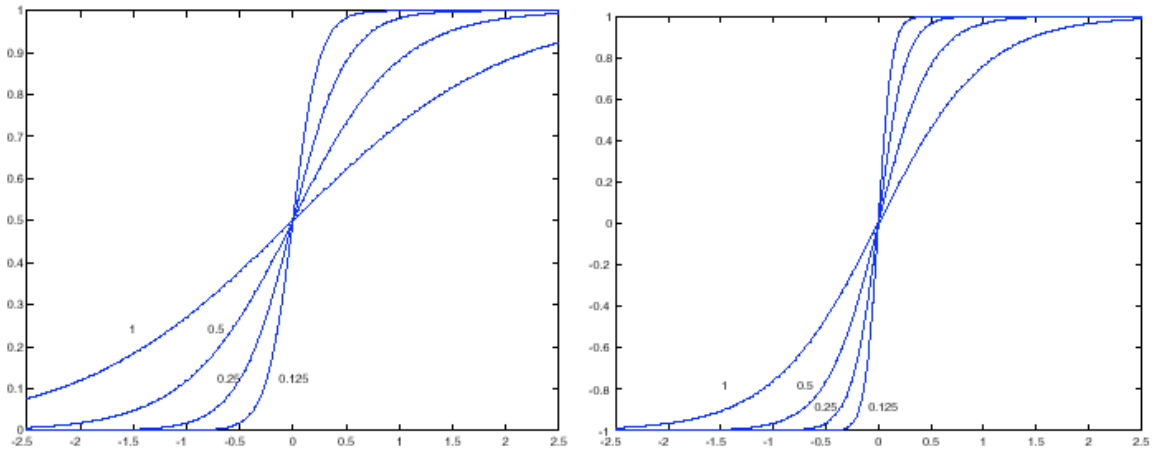


Figure 22.14. On the **left**, a series of squashing functions obtained using $\phi(x;\nu) = \frac{e^{x/\nu}}{1+e^{x/\nu}}$, for different values of ν indicated on the figure. On the **right**, a series of squashing functions obtained using $\phi(x;\nu,A) = A\tanh(x/\nu)$ for different values of ν indicated on the figure. Generally, for x close to the center of the range, the squashing function is linear; for x small or large, it is strongly non-linear.

Training

- Choose parameters to minimize error on training set

$$Error(\boldsymbol{p}) = \left(\frac{1}{2}\right) \sum_{e} | \boldsymbol{n}(\boldsymbol{x}^e; \boldsymbol{p}) - \boldsymbol{o}^e |^2$$

- Stochastic gradient descent, computing gradient using trick (backpropagation, aka the chain rule)
- Stop when error is low, and hasn't changed much

Rowley-Baluja-Kanade face finder (1)

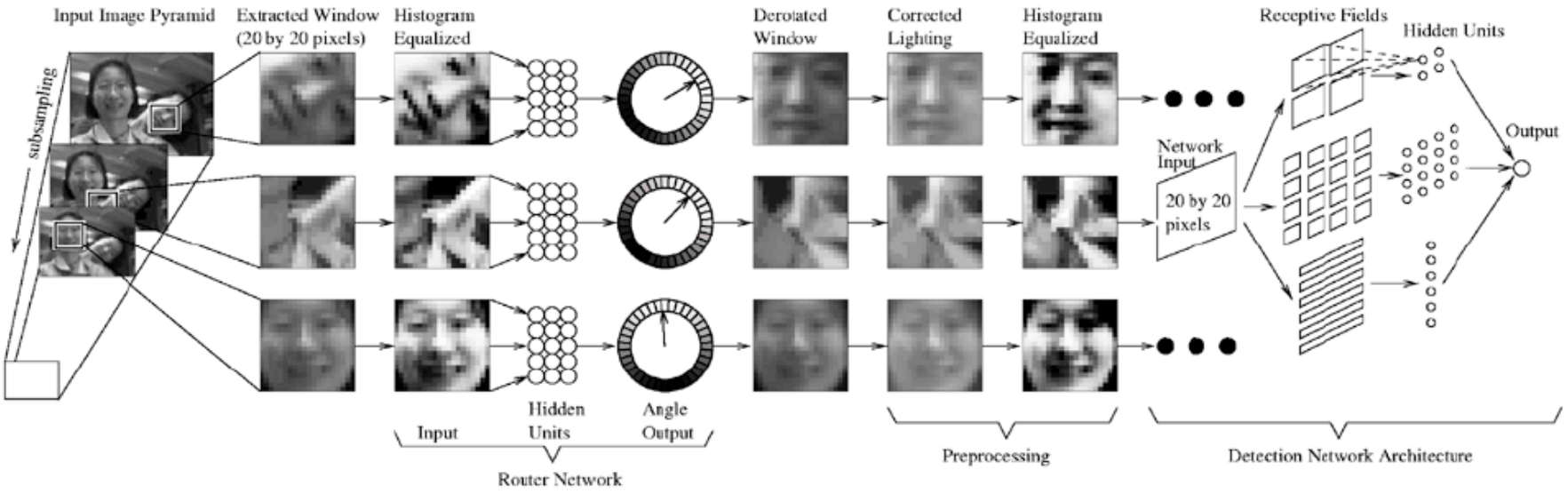


Figure from "Rotation invariant neural-network based face detection,"
H.A. Rowley, S. Baluja and T. Kanade, Proc. Computer Vision and Pattern Recognition,
1998, c 1998, IEEE as shown in Forsyth and Ponce, p589

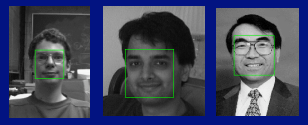
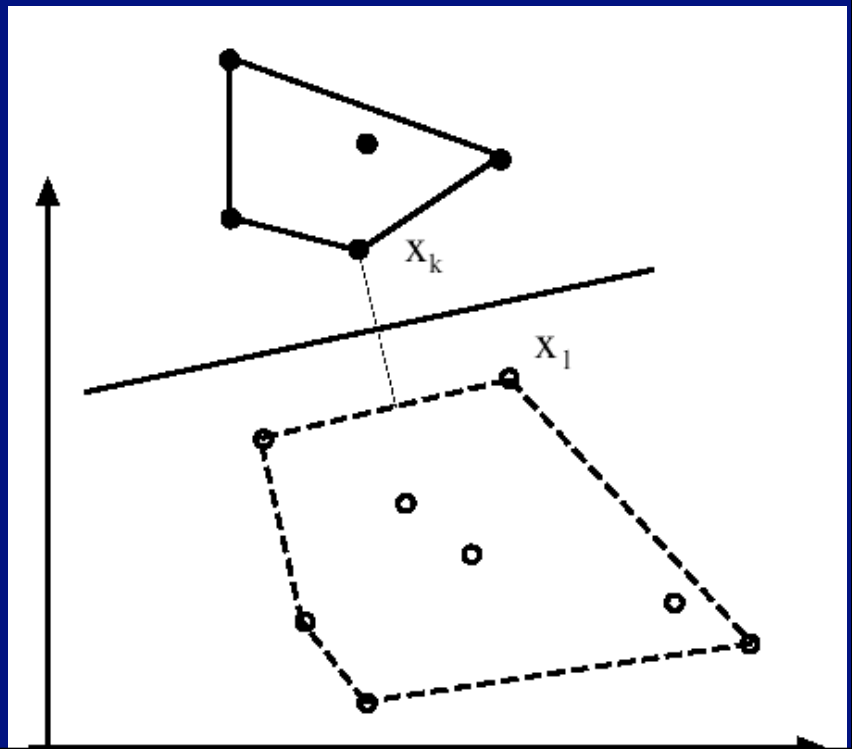


Figure from "Rotation invariant neural-network based face detection,"
H.A. Rowley, S. Baluja and T. Kanade, Proc. Computer Vision and Pattern Recognition,
1998, c 1998, IEEE as shown in Forsyth and Ponce, p589

Decision boundaries

- The boundary matters
  - but the details of the probability model may not
- Seek a boundary directly
  - when we do so, many or most examples are irrelevant
- Support vector machine



Support Vector Machines, easy case

- Classify with sign(w.x+b)
- Linearly separable data means $y_i\,(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)>0$
- Choice of hyperplane means $y_i\,(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geq 1$
- Hence distance

$$\begin{aligned} dist(\boldsymbol{x}_k,\text{hyperplane})+dist(\boldsymbol{x}_l,\text{hyperplane}) &= (\frac{\boldsymbol{w}}{|\boldsymbol{w}|}\cdot\boldsymbol{x}_k+\frac{b}{|\boldsymbol{w}|})-(\frac{\boldsymbol{w}}{|\boldsymbol{w}|}\cdot\boldsymbol{x}_1+\frac{b}{|\boldsymbol{w}|}) \\ &= \frac{\boldsymbol{w}}{|\boldsymbol{w}|}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)=\frac{2}{|\boldsymbol{w}|} \end{aligned}$$

Support Vector Machines, separable case

$$\begin{aligned} \text{minimize} \quad & (1/2)\boldsymbol{w} \cdot \boldsymbol{w} \\ \text{subject to} \quad & y_i \left(\boldsymbol{w} \cdot \boldsymbol{x}_i + b\right) \geq 1 \end{aligned}$$

By being clever about what x means, I can have much more interesting boundaries.

Dual SVM problem

Lagrangian

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0$$

Gradient with respect to w, b must vanish

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\sum_i \alpha_i y_i = 0.$$

Dual SVM problem - II

Substitute

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

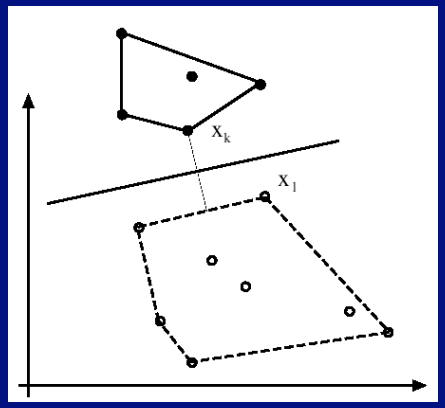
Constraints

$$\sum_i \alpha_i y_i = 0.$$

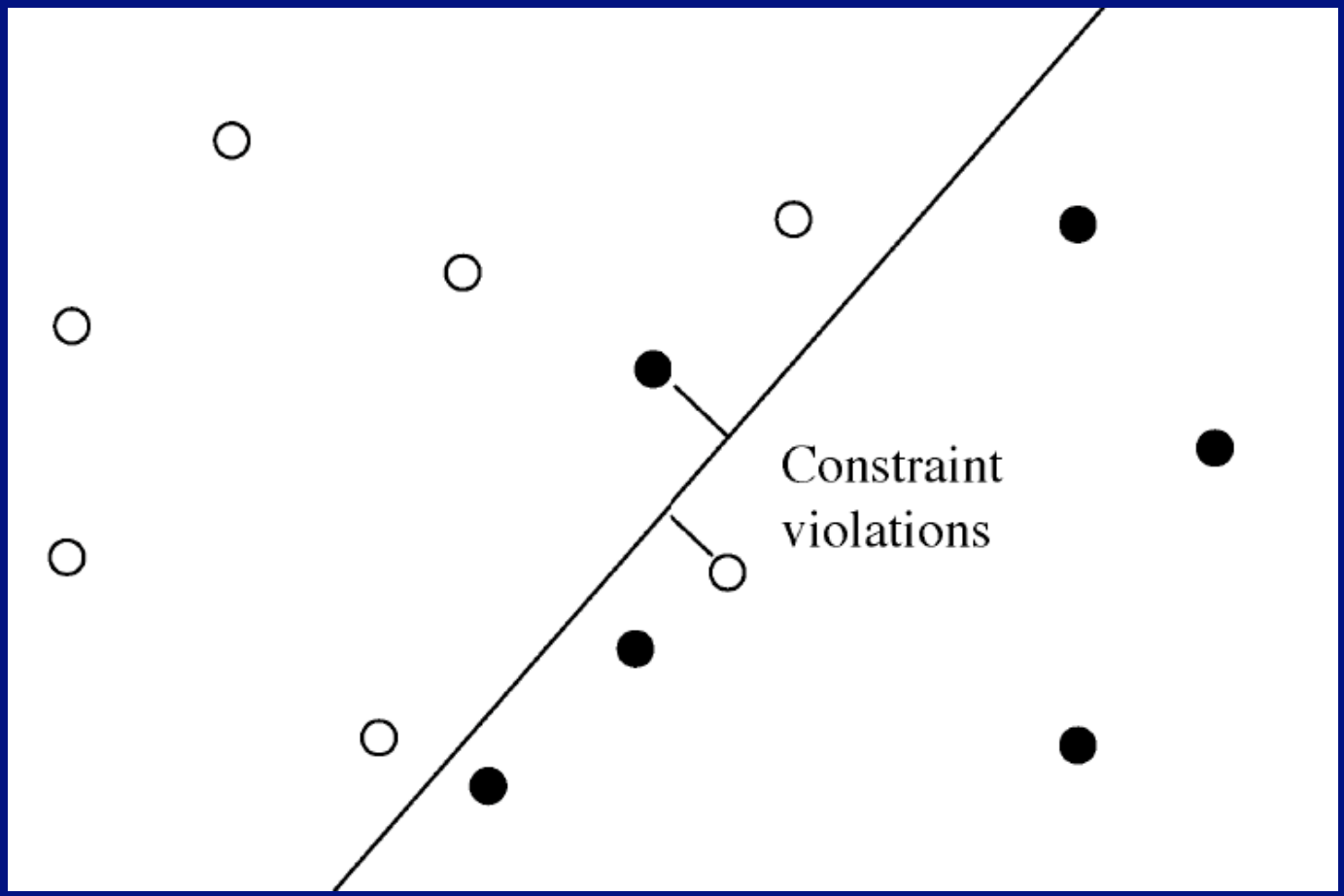
$$\alpha_i \geq 0$$

Solution

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$



Data not linearly separable



Data not linearly separable

Constraints become

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b &\geq +1 - \xi_i \quad \text{for } y_i = +1 \\ \mathbf{x}_i \cdot \mathbf{w} + b &\leq -1 + \xi_i \quad \text{for } y_i = -1 \\ \xi_i &\geq 0 \ \forall i. \end{aligned}$$

Objective function becomes

$$\|\mathbf{w}\|^2/2 + C\left(\sum_i \xi_i\right)$$

Data not linearly separable - II

Maximize:

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

subject to:

$$0 \leq \alpha_i \leq C,$$

$$\sum_i \alpha_i y_i = 0.$$

The solution is again given by

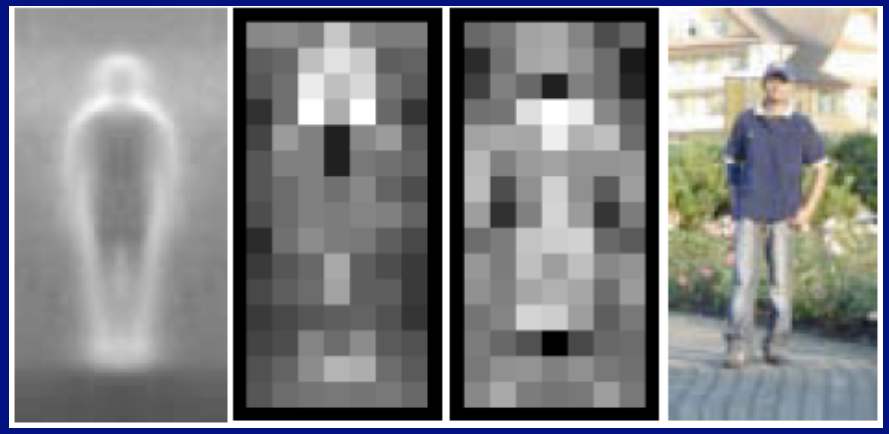
$$\mathbf{w} = \sum_{i=1}^{N_S} \alpha_i y_i \mathbf{x}_i .$$

Pedestrian detection with an SVM

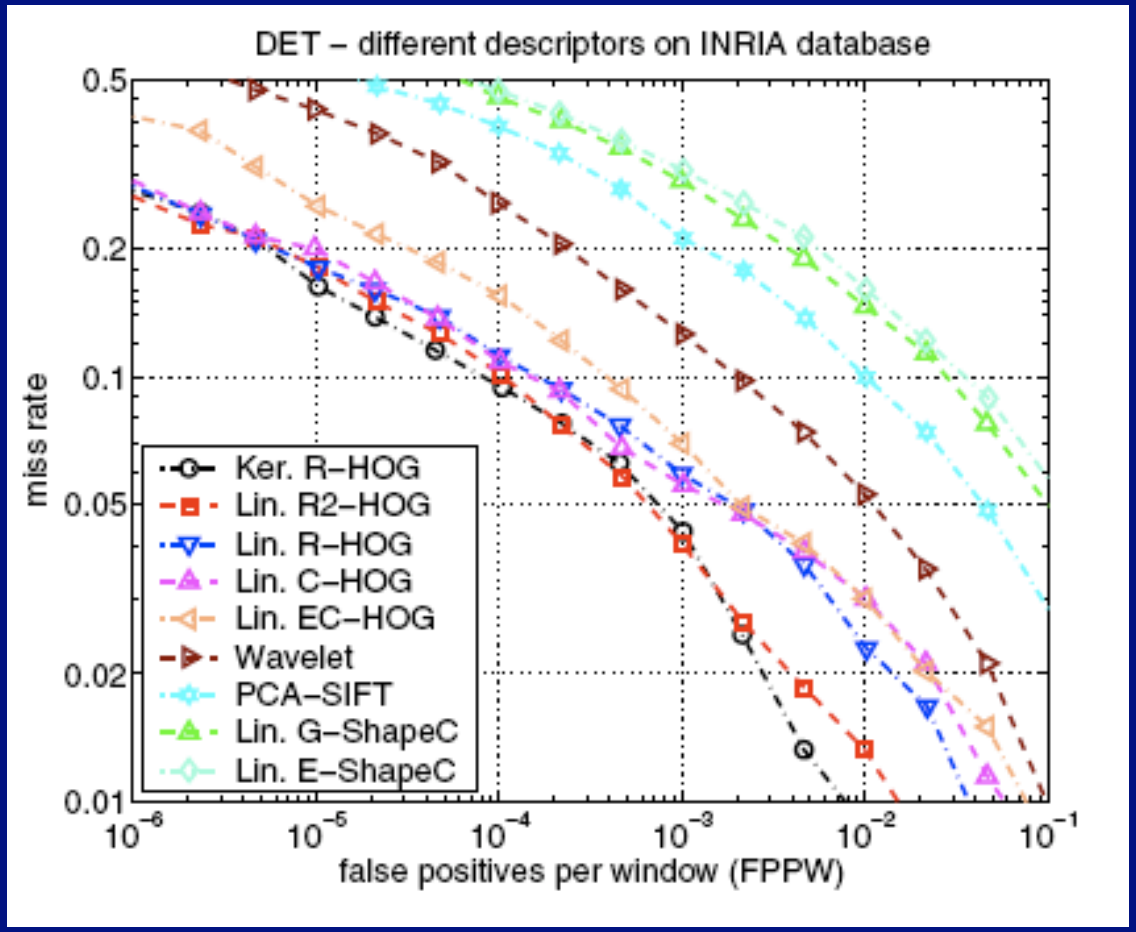
Dalal+Triggs 05

Features

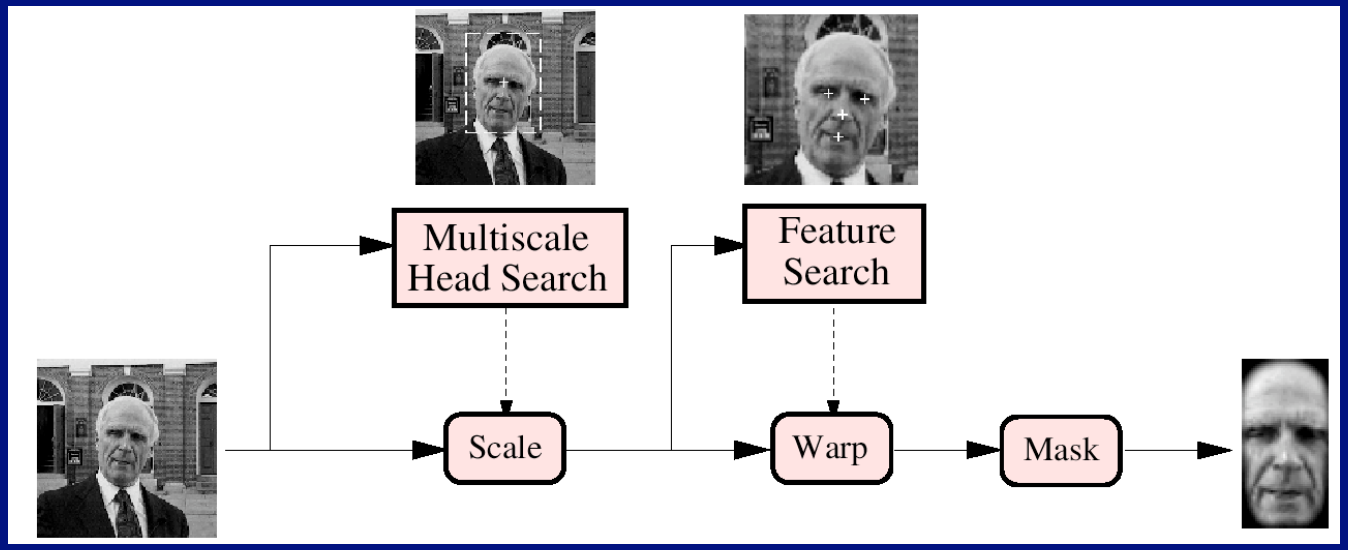
Dalal+Triggs 05



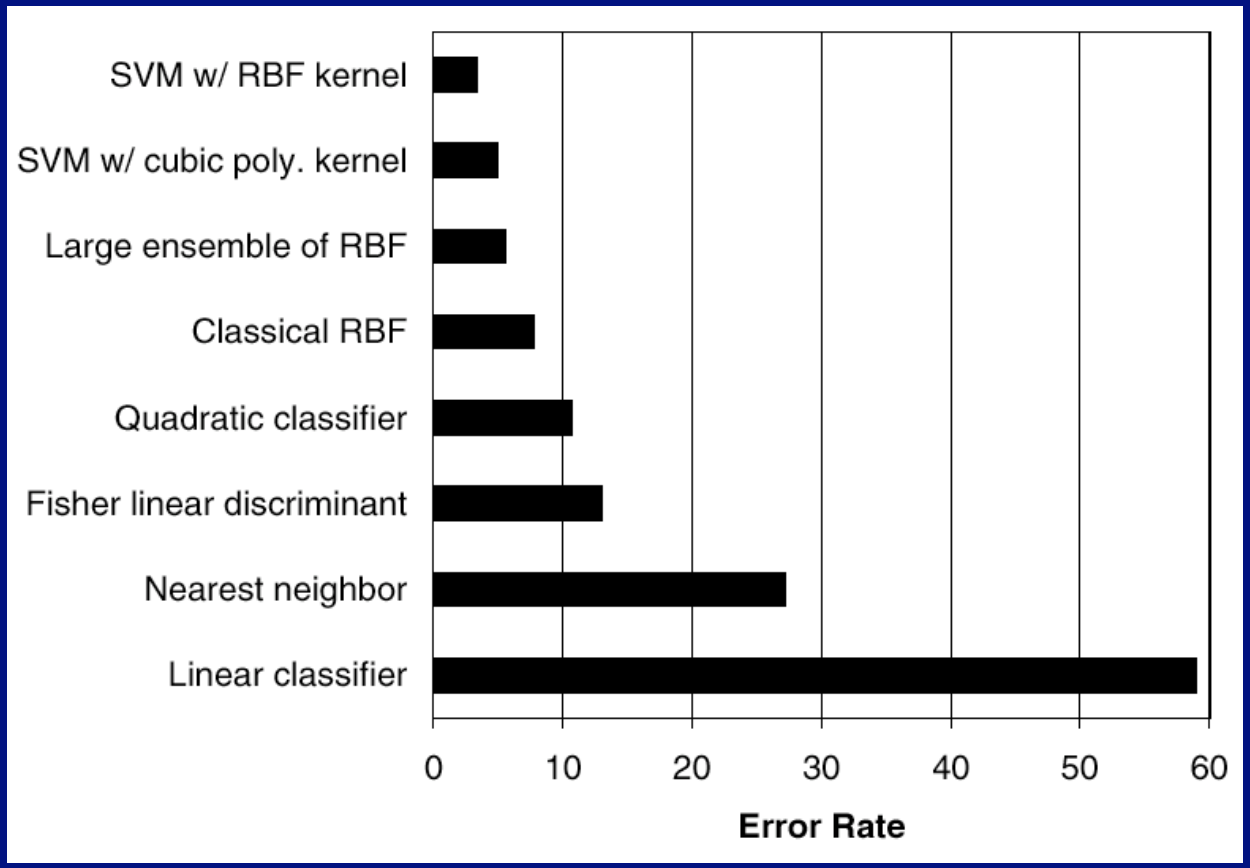
Performance

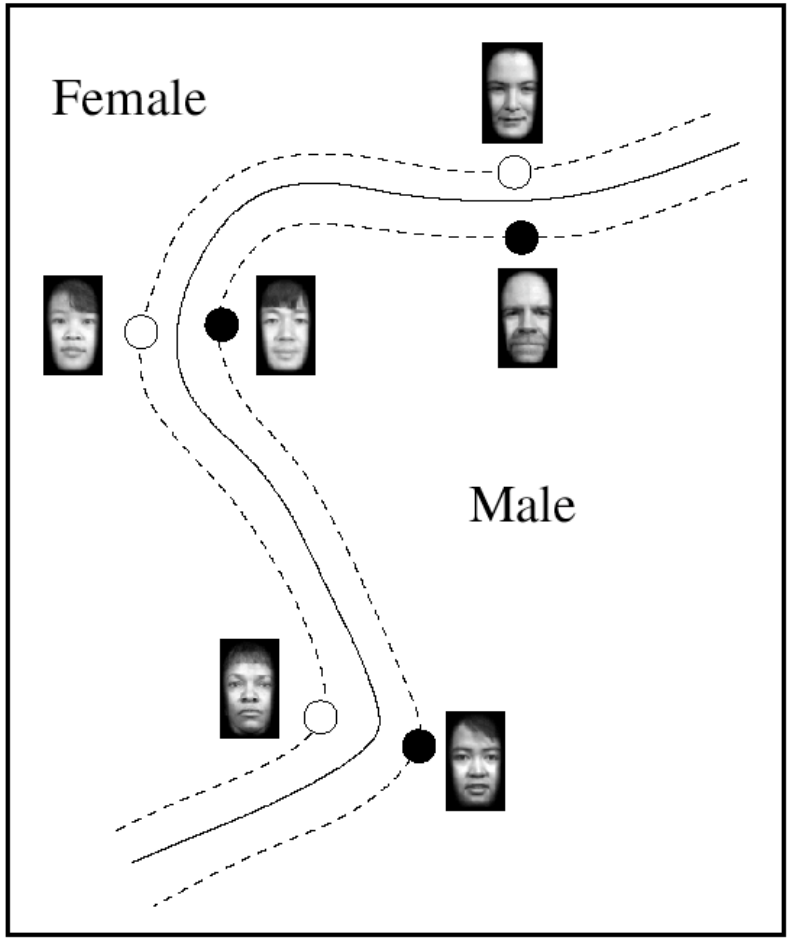


Dalal+Triggs 05



From “Gender Classication with
Support Vector Machines”
Baback Moghaddam
Ming-Hsuan Yang, MERL TR


SVM w/ RBF kernel
SVM w/ cubic poly. kernel
Large ensemble of RBF
Classical RBF
Quadratic classifier
Fisher linear discriminant
Nearest neighbor
Linear classifier
0
10
20
30
40
50
60
Error Rate


Female
Male


% Error Rates
30
25
20
15
10
5
0
Low−Res
Hi−Res
SVM
Human

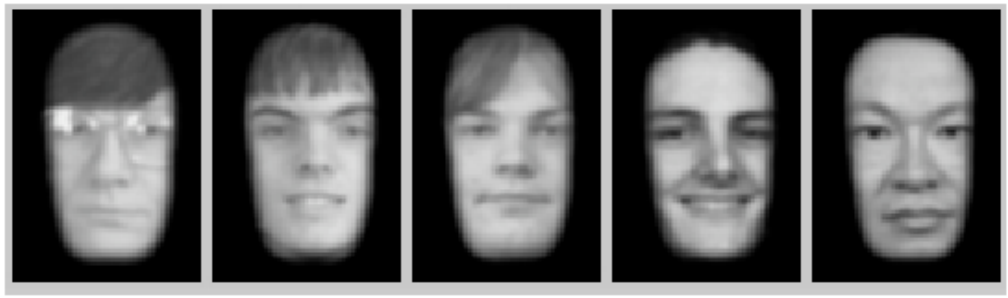


Figure 7. Top five human misclassifications

LOTS of BIG collections of images

Corel Image Data	40,000 images
Fine Arts Museum of San Francisco	83,000 images online
Cal-flora	20,000 images, species information
News photos with captions (yahoo.com)	1,500 images per day available from yahoo.com
Hulton Archive	40,000,000 images (only 230,000 online)
internet.archive.org	1,000 movies with no copyright
TV news archives (televisionarchive.org, informedia.cs.cmu.edu)	Several terabytes already available
Google Image Crawl	>330,000,000 images (with nearby text)
Satellite images (terrarserver.com, nasa.gov, usgs.gov)	(And associated demographic information)
Medial images	(And associated with clinical information)

* and the BBC is releasing its video archive, too;
and we collected 500,000 captioned news images;
and it's easy to get scanned mediaeval manuscripts;
etc., etc.,

Imposing order

- Iconic matching
  - child abuse prosecution
  - managing copyright (BayTSP)

  Current, practical applications
- Clustering
  - Browsing for:
    - web presence for museums (Barnard et al, 01)
    - home picture, video collections
    - selling pictures

  Maybe applications
- Searching
  - scanned writing (Manmatha, 02)
  - collections of insects

  Maybe applications
- Building world knowledge
  - a face gazetteer (Miller et al, 04)

Search is well studied

- Metadata indexing
  - keywords, date of photo, place, etc.
- Content based retrieval
  - query by example with
    - global features
      - (e.g. Flickner et al. 95, Carson et al. 99, Wang 00, various entire conferences)
    - local features
      - (e.g. Photobook - Pentland et al 96; Blobworld - Carson et al, 98)
    - relevance feedback
      - (e.g. Cox et al 00; Santini 00; Schettini 02; etc.)
  - query by class
    - naughty pictures
      - (eg Forsyth et al. 96, 99; Wang et al. 98; Chan et al 99)

What will users pay for?

- Work by Peter Enser and colleagues on the use of photø movie collections
  (Enser McGregor 92; Ornager 96; Armitage Enser 97; Markkula Sormunen 00; Frost et al 00; Enser 00)
- Typical queries:

What is this about?

"… smoking of kippers…"

"The depiction of vanity in painting, the depiction of the female figure looking in the mirror, etc."

"Cheetahs running on a greyhound course in Haringey in 1932"

Annotation results in complementary words and pictures

Query on

“Rose”

Example from Berkeley
Blobworld system

Annotation results in complementary words and pictures

Query on

Example from Berkeley
Blobworld system

Annotation results in complementary words and pictures

Query on

“Rose”

and

Example from Berkeley Blobworld system

Exploiting complementary information

- A probability model linking images and annotations
  - exploit co-occurence
  - better estimates of "meaning" for clustering and browsing
  - soft search, auto illustration, auto annotation
- Predicting words from image regions
  - explicitly encode and infer correspondence
    - aligned bitext
    - no alignment
  - rather like recognition
  - pinch techniques from statistical natural language processing
- Linking face images with names
  - an important special case
  - datasets of an epic scale available
  - like face recognition, but easier
  - breaking correspondence by clustering

Browsing

- Searching big, unknown collections is hard for naive user
  - skilled users don't benefit from vision-based tools
  - problem of overrated significance
- Browsing?
  - seems to be preferred by naive users (Frost et al, `00)
  - but browsing requires organization too
  - generally underrated problem

*Notable exceptions ---Sclaroff, Taycher, and La Cascia, 98; Rubner, Tomasi, and Guibas, 00; Smith Kanade, 97.

Clustering words and pictures

- Build a joint probability model linking words and pictures
- Use Hoffman's hierarchical aspect model
  - which is a form of clusterer [Hofmann 98; Hofmann & Puzicha 98]
- Lay out and browse the clusters

Input

"This is a picture of the sun setting over the sea with waves in the foreground"

Image processing*

Each blob is a large vector of features
- Region size
- Position
- Colour
- Oriented energy (12 filters)
- Simple shape features

Language processing

sun sky waves sea

* Thanks to Blobworld team [Carson, Belongie, Greenspan, Malik], N-cuts team [Shi, Tal, Malik]

FAMSF Data

Web number: 4359202410830012

rec number: 2

Description:
serving woman stands in a dressing room, in front of vanity with chair, mirror and mantle, holding a tray with tea and toast

Title: Le Matin

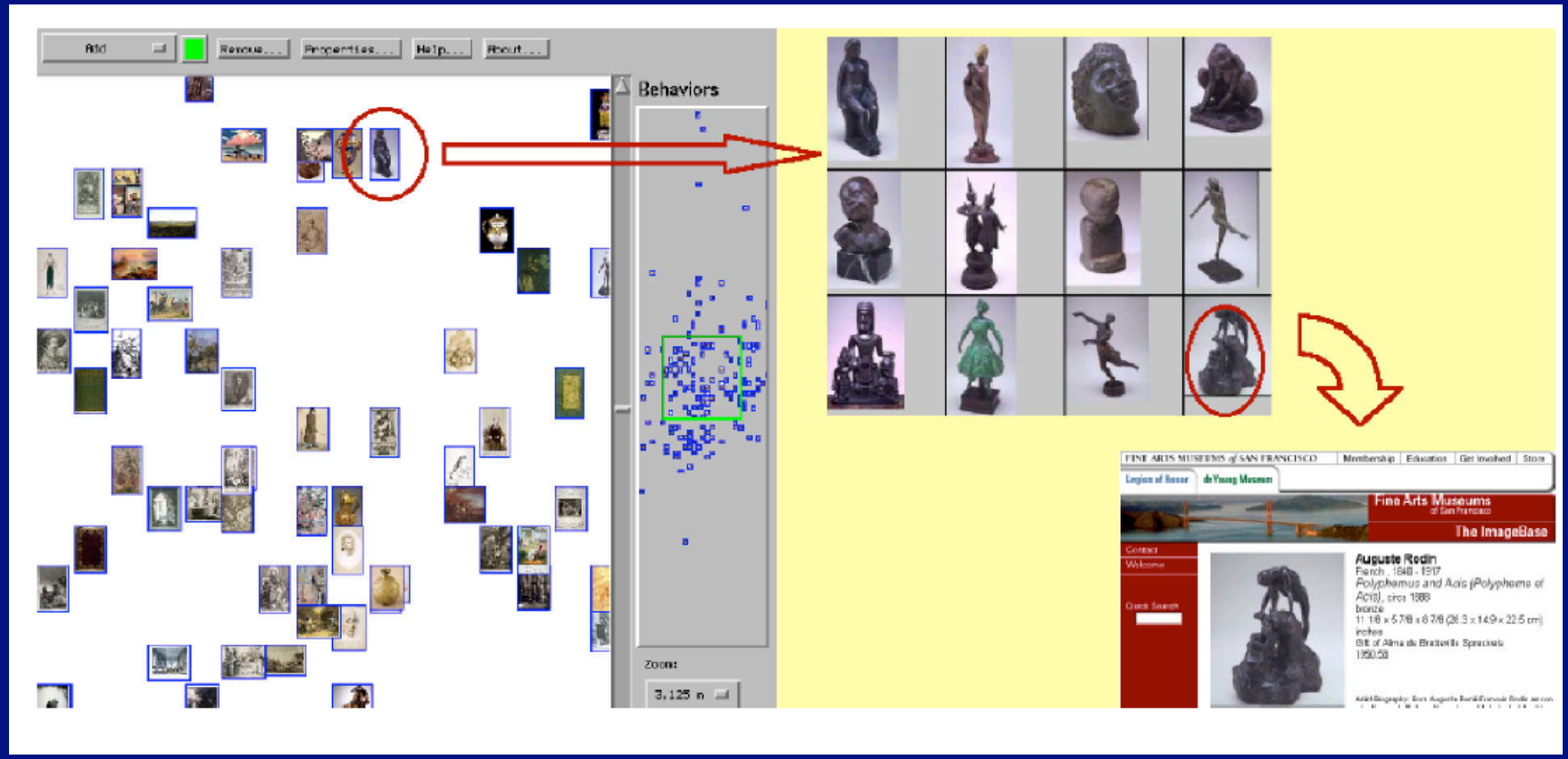
Primary class: Print

Display date: 1886

Artist: Tissot

Country: France

83,000 images online, we clustered 8000


Add
Revoke...
Properties...
Help...
About...
Behaviors
Zoom:
3.125 m
FINE ARTS MUSEUMS of SAN FRANCISCO
Membership
Education
Get involved
Store
Legion of Honor
de Young Museum
Fine Arts Museums
of San Francisco
The ImageBase
Contact
Welcome
Quick Search
Auguste Rodin
French, 1840 - 1917
Polyphemus and Acis (Polyphème et Acis), circa 1888
bronze
11 1/8 x 5 7/8 x 8 7/8 (28.3 x 14.9 x 22.5 cm)
inches
Gift of Alma de Bretteville Spreckels
1950.59

Searching

Compute P(document | query_items)

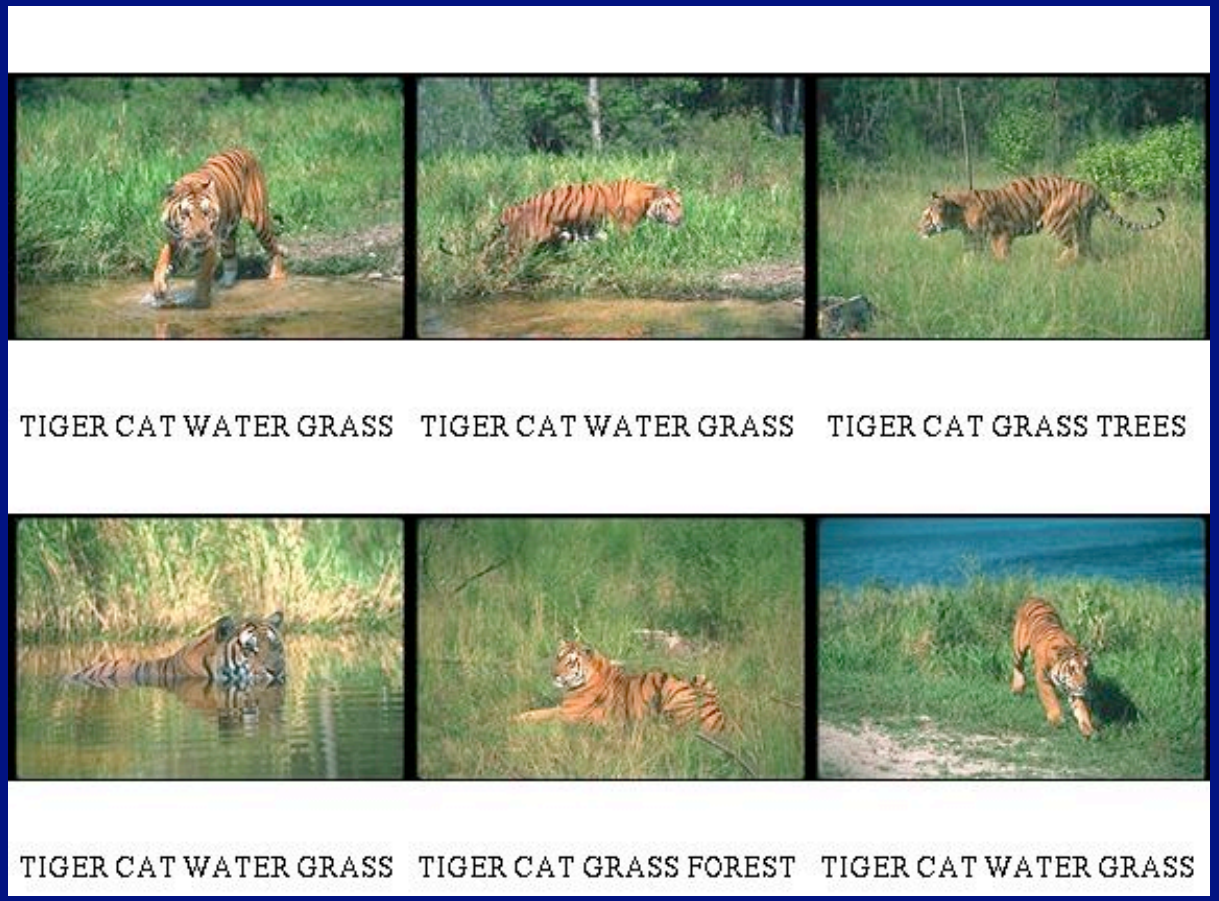
query_items can be words, features, or both

Natural way to express “soft queries”

Related retrieval work: Cascia, Sethi, and Sclaroff, 98; Berger and Lafferty, 98; Papadimitriou et al., 98

Query: “river tiger” from 5,000 Corel images
(The words never occur together.)

Retrieved items: rank order P(document | query)



Search

Compute P(document | query_items)

Related retrieval work: Cascia, Sethi, and Sclaroff, 98; Berger and Lafferty, 98; Papadimitriou et al., 98

Query: "water sky cloud  "

Retrieved items:

Pictures from Words (Auto-illustration)

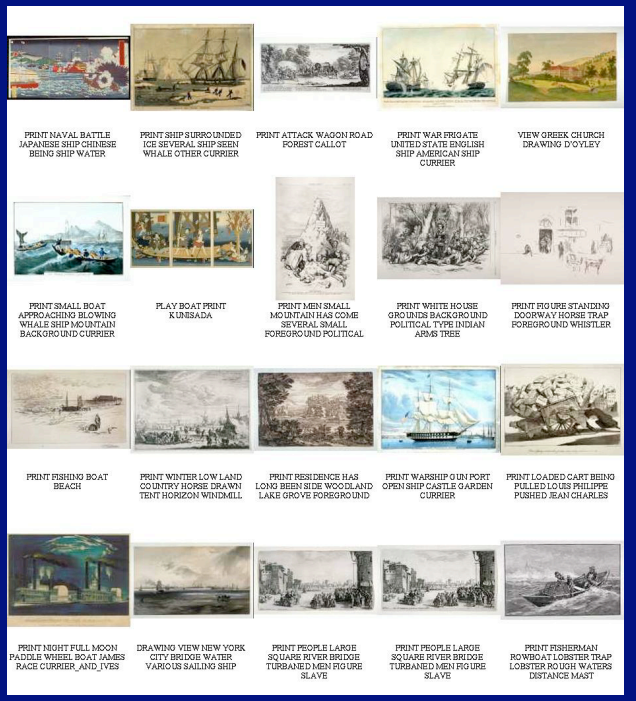
Text Passage (Moby Dick)

"The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship …"

Extracted Query

large importance attached fact old dutch century more command whale ship was person was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch ...

Retrieved Images



PRINT NAVAL BATTLE JAPANESE SHIP CHINESE BEING SHIP WATER

PRINT SHIP SURROUNDED ICE SEVERAL SHIP SEEN WHALE OTHER CURRIER

PRINT ATTACK WAGON ROAD FOREST CALLOT

PRINT WAR FRIGATE UNITED STATE ENGLISH SHIP AMERICAN SHIP CURRIER

PRINT SMALL BOAT APPROACHING BLOWING WHALE SHIP MOUNTAIN BACKGROUND CURRIER

PLAY BOAT PRINT KUNISADA

PRINT MEN SMALL MOUNTAIN HAS COME SEVERAL SMALL FOREGROUND POLITICAL

PRINT WHITE HOUSE GROUNDS BACKGROUND POLITICAL TYPE INDIAN ARMS TREE

Auto-annotation

- Predict words from pictures
  - Obstacle:
    - Hoffman's model uses document specific level probabilities
  - Dodge
    - smooth these empirically
- Attractions:
  - easy to score
  - large scale performance measures (how good is the segmenter?)
  - possibly simplify retrieval (Li+Wang, 03)

Keywords
GRASS TIGER CAT FOREST
Predicted Words (rank order)

tiger cat grass people water bengal buildings ocean forest reef

---

Keywords
HIPPO BULL mouth walk
Predicted Words (rank order)

water hippos rhino river grass reflection one-horned head plain sand

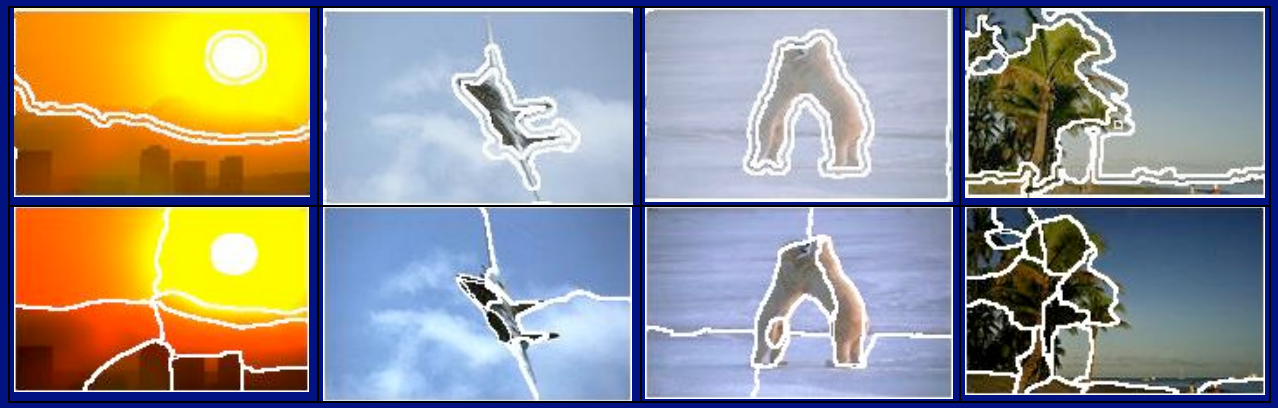
---

Keywords
FLOWER coralberry LEAVES PLANT

Predicted Words (rank order)

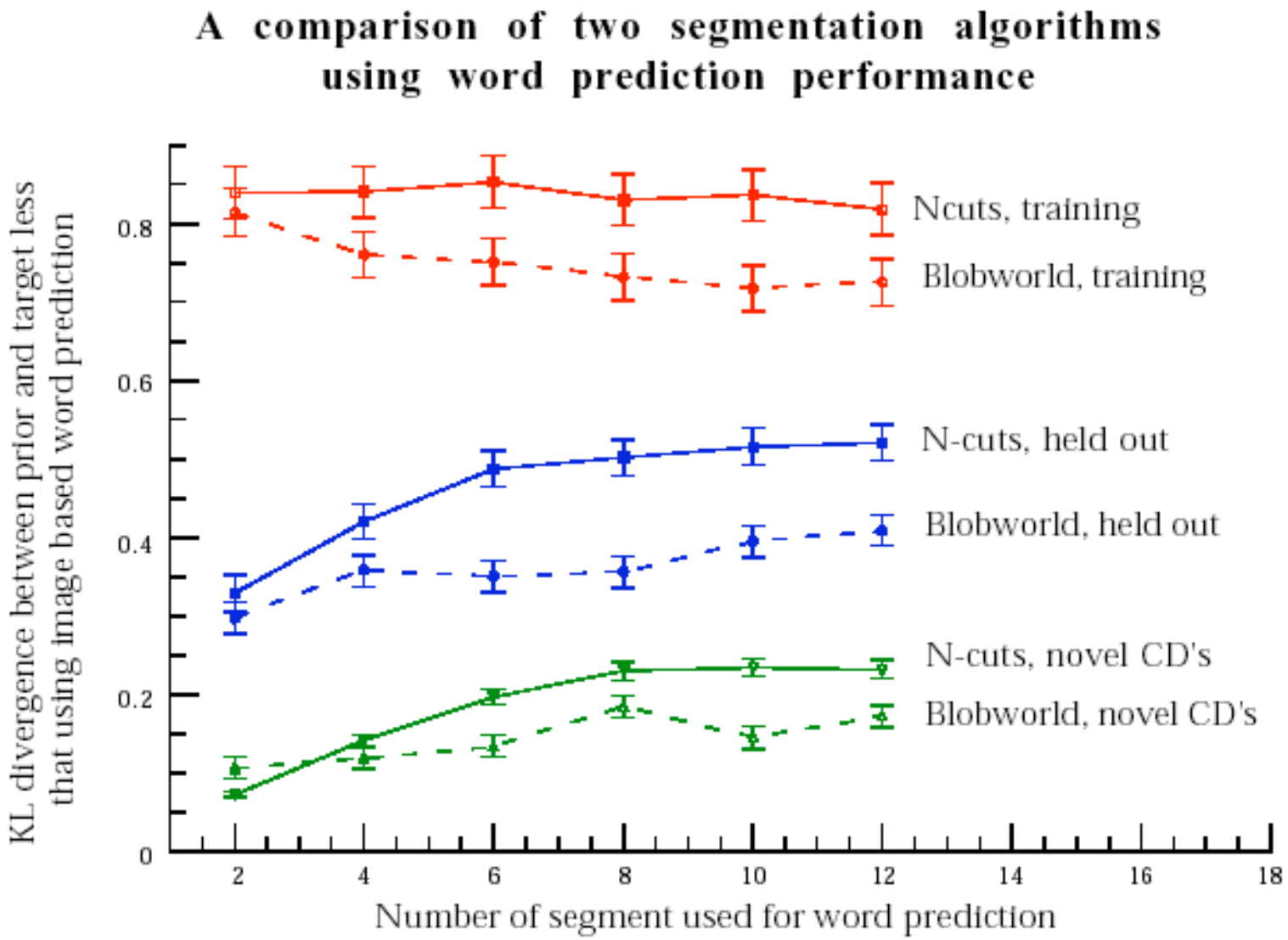
fish reef church wall people water landscape coral sand trees

Blobworld segmentations



N-cuts segmentations

KL divergence based word prediction measure (compared with prior, bigger is better)



Barnard et al. ‘03

Exploiting complementary information

- A probability model linking images and annotations
  - exploit co-occurence
  - better estimates of "meaning" for clustering and browsing
  - soft search, auto illustration, auto annotation
- Predicting words from image regions
  - explicitly encode and infer correspondence
    - aligned bitext
    - no alignment
  - rather like recognition
  - pinch techniques from statistical natural language processing
- Linking face images with names
  - an important special case
  - datasets of an epic scale available
  - like face recognition, but easier
  - breaking correspondence by clustering

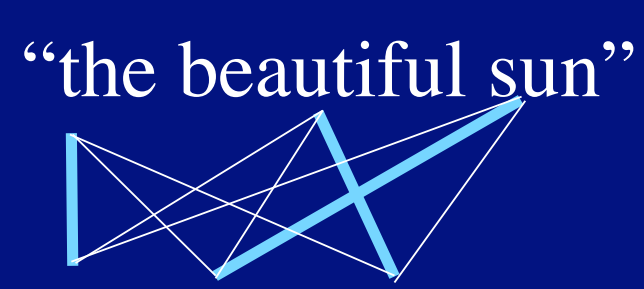
Annotation vs Recognition

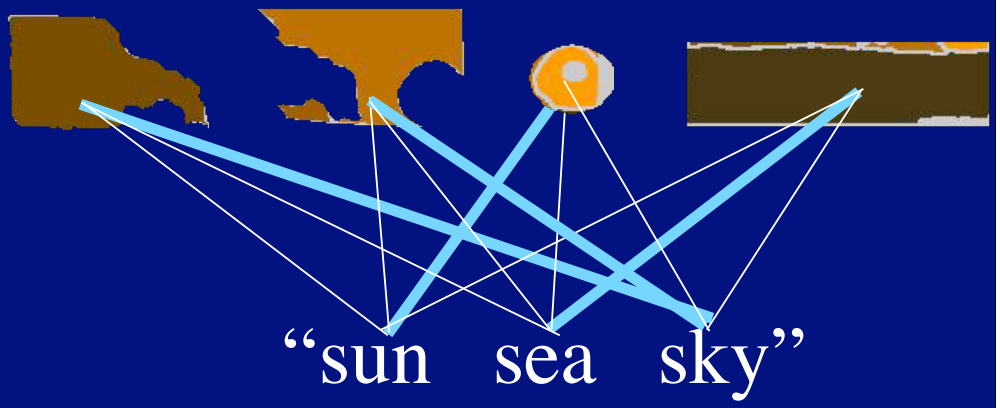
?

tiger cat grass

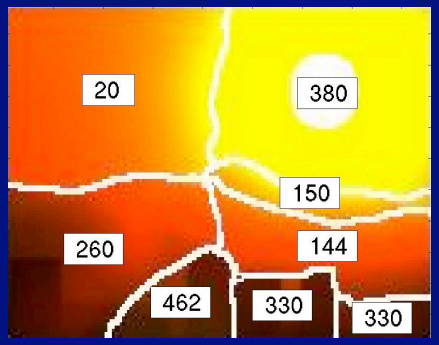
Lexicon building

- In its simplest form, missing variable problem
- Pile in with EM
  - given correspondences, conditional probability table is easy (count)
  - given cpt, expected correspondences could be easy
- Caveats
  - might take a lot of data; symmetries, biases in data create issues

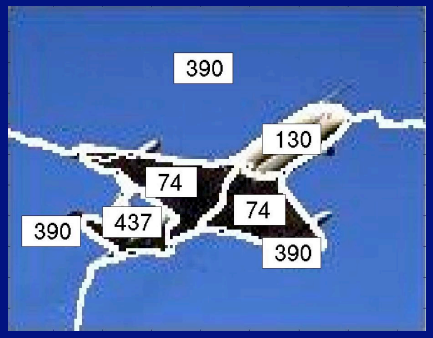




Brown, Della Pietra, Della Pietra & Mercer 93; Melamed 01



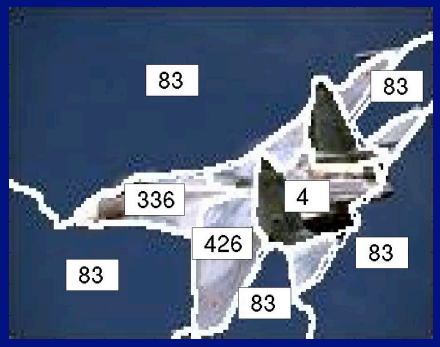
city mountain sky sun



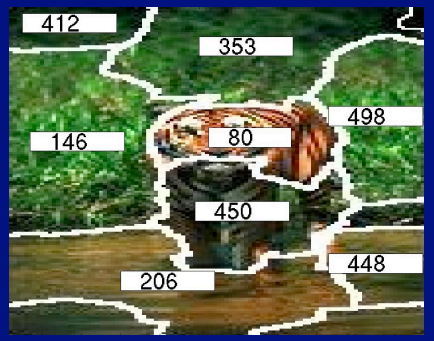
jet plane sky

cat forest grass tiger

beach people sun water



jet plane sky



cat grass tiger water

“Lexicon” of “meaning”

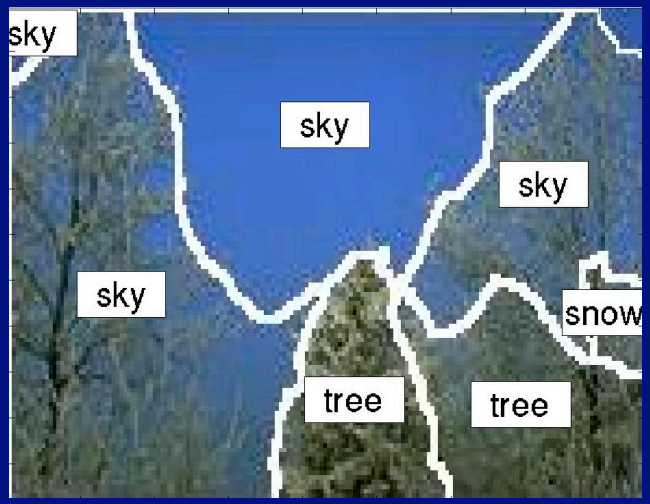
sun

sky

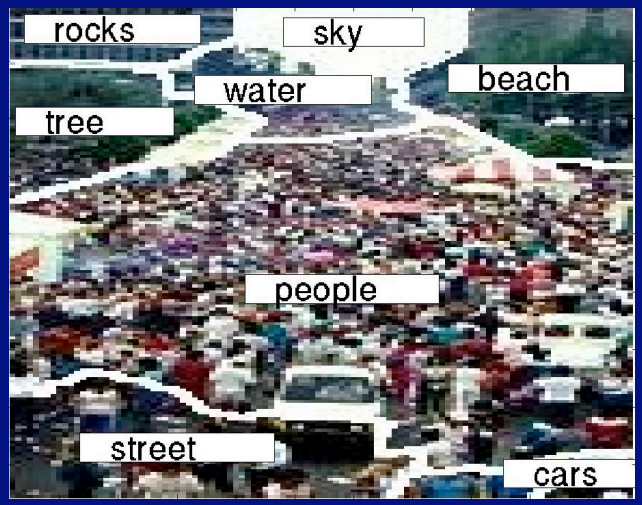
cat

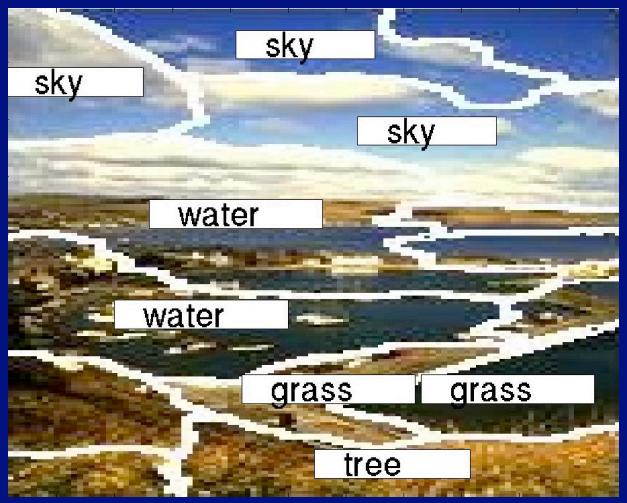
horse

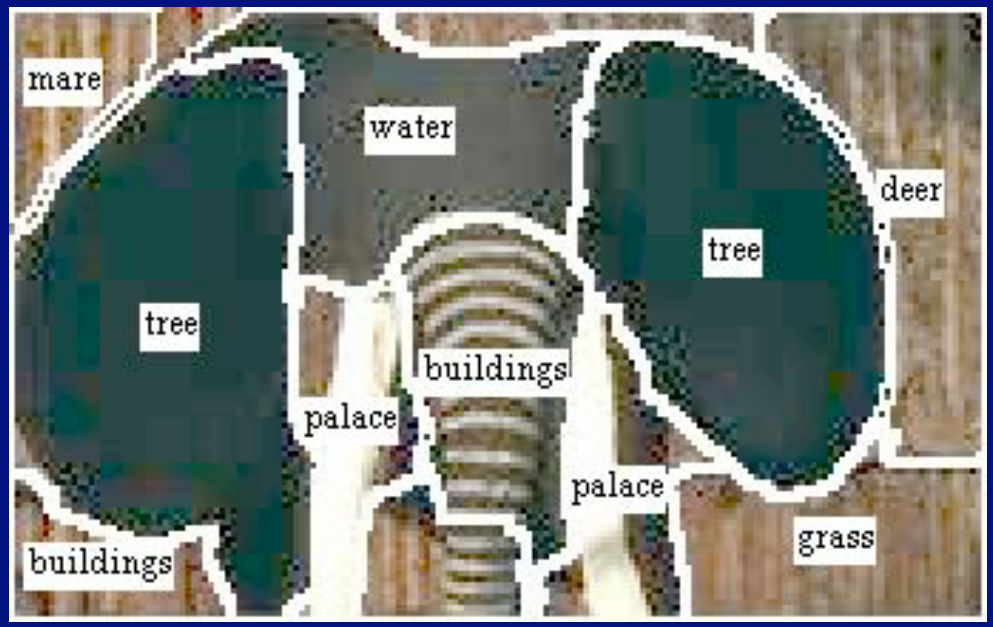
This could be either a conditional probability table or a joint probability table; each has significant attractions for different applications


sky
sky
sky
sky
snow
tree
tree


sky
sky
water
buildings
sky
tree
buildings
buildings


rocks
sky
water
beach
tree
people
street
cars


sky
sky
sky
water
water
grass
grass
tree

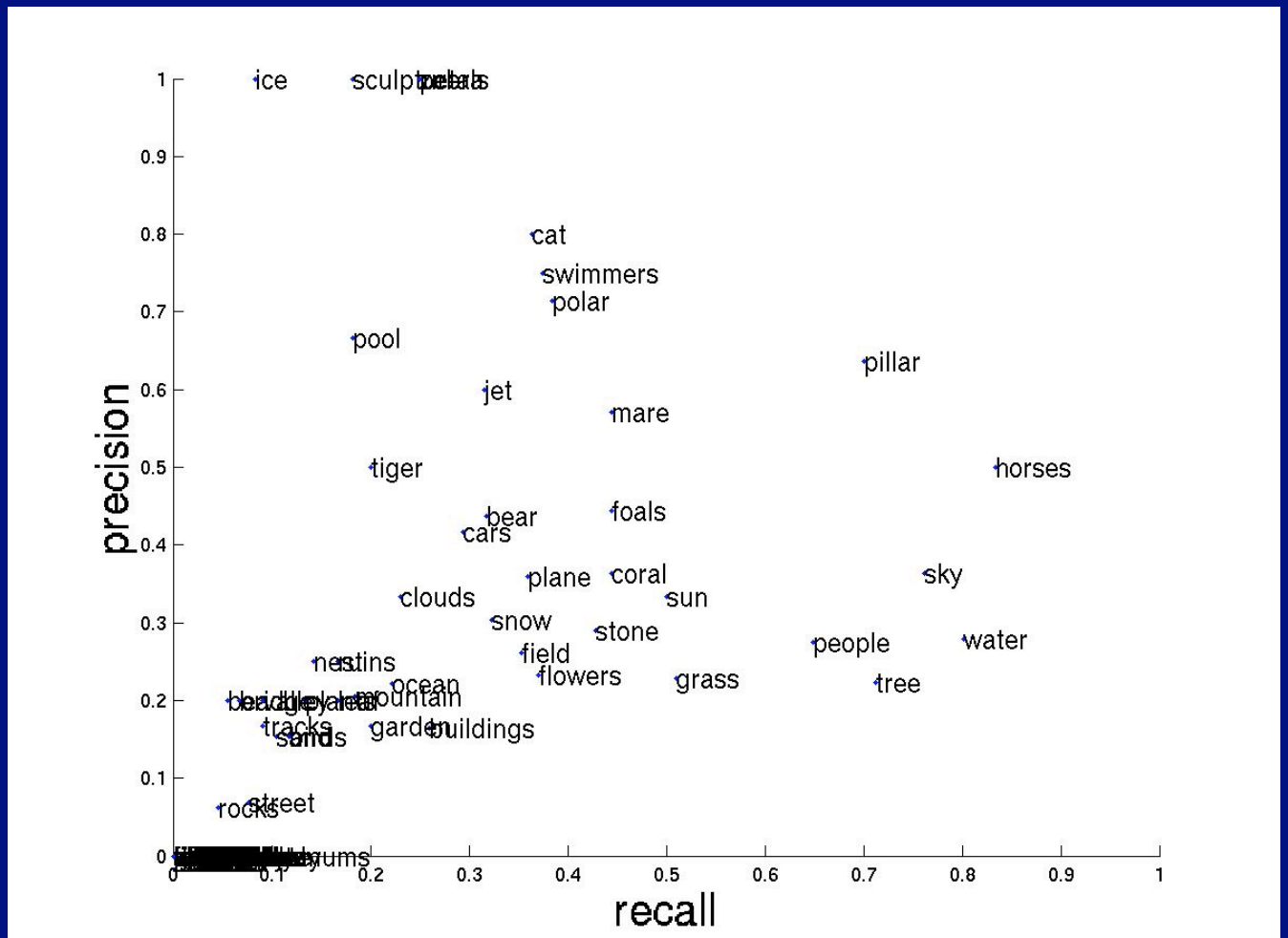

mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

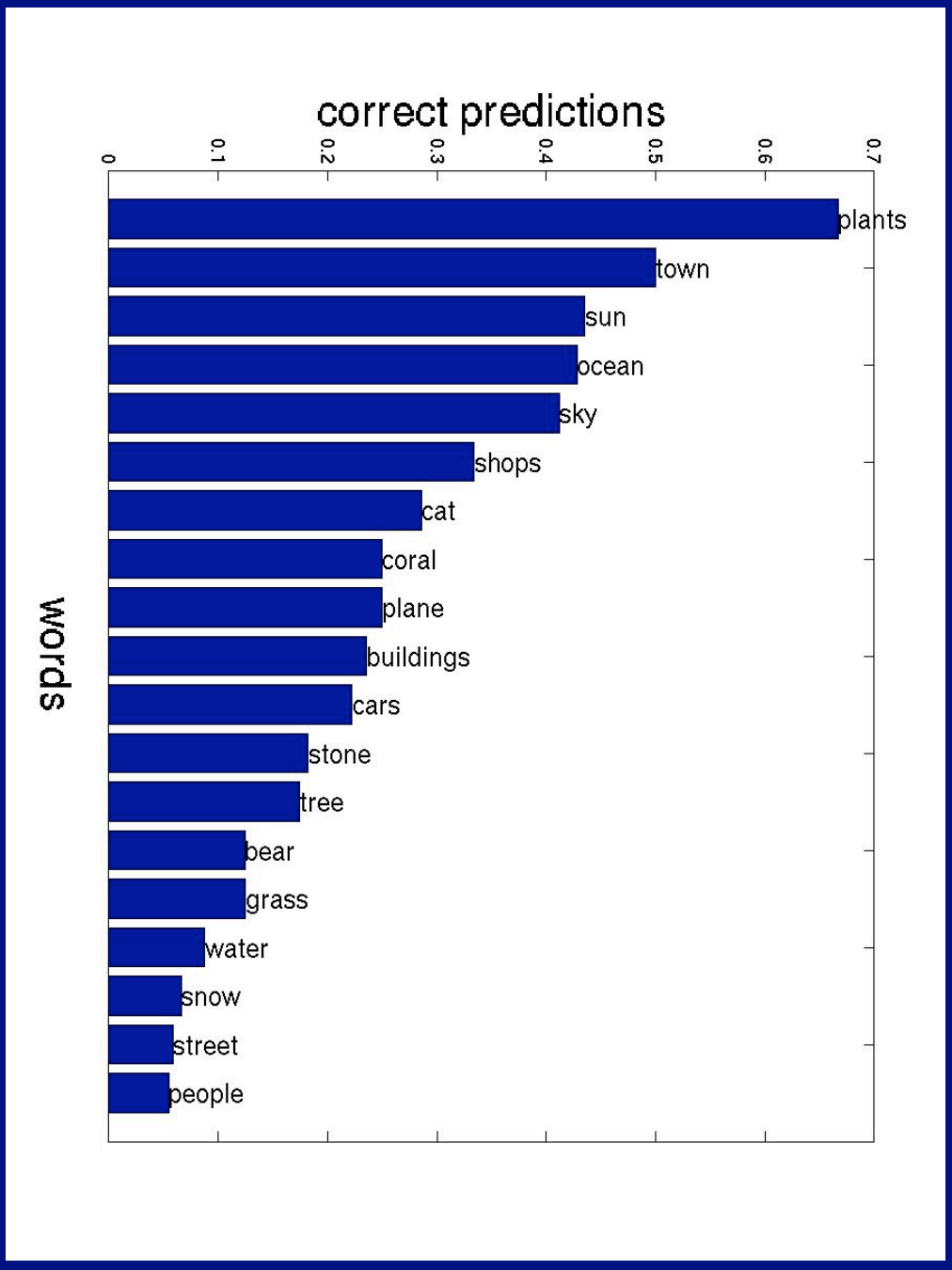
Performance measurement

By hand

By proxy

Grass Cat Buildings
Horses Tiger Mare


precision
recall
ice
cat
swimmers
polar
pool
pillar
jet
mare
tiger
horses
bear
foals
cars
plane
coral
sky
clouds
sun
snow
stone
people
water
field
flowers
grass
tree
ocean
mountain
tracks
garden
buildings
rocks
street
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


correct predictions
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
words
plants
town
sun
ocean
sky
shops
cat
coral
plane
buildings
cars
stone
tree
bear
grass
water
snow
street
people

Exploiting complementary information

- A probability model linking images and annotations
  - exploit co-occurence
  - better estimates of "meaning" for clustering and browsing
  - soft search, auto illustration, auto annotation
- Predicting words from image regions
  - explicitly encode and infer correspondence
    - aligned bitext
    - no alignment
  - rather like recognition
  - pinch techniques from statistical natural language processing
- Linking face images with names
  - an important special case
  - datasets of an epic scale available
  - like face recognition, but easier
  - breaking correspondence by clustering

News dataset

- Approx 5e5 news images, with captions
  - Easily collected by script from Yahoo over the last 18 months or so
- Mainly people
  - politicians, actors, sportsplayers
  - long, long tails distribution
- Face pictures captured “in the wild”
- Correspondence problem
  - some images have many (resp. few) faces, few (resp. many) names (cf. Srihari 95)

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

Data examples

Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

President George W. Bush waves as he leaves the White House for a day trip to North Carolina, July 25, 2002. A White House spokesman said that Bush would be compelled to veto Senate legislation creating a new department of homeland security unless changes are made. (Kevin Lamarque/Reuters)

Process

- Extract proper names
  - rather crudely, at present
- Detect faces
  - with Cordelia Schmid's face detector, (Vogelhuber Schmid 00)
- Rectify faces
  - by finding eye, nose, mouth patches, affine transformation
- Kernel PCA rectified faces
- Estimate linear discriminants
- Now have (face vector; name_1,...., name_k)

Scale

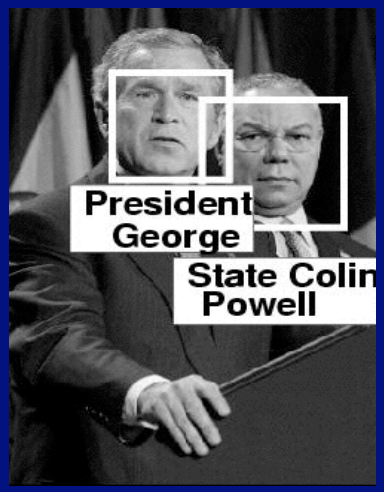
44773 big face responses

34623 properly rectified

27742 for k<=4

Building a face dictionary

- Compute linear discriminants
  - using single name, single face data items
  - we now have a set of clusters
- Now break correspondence with modified k-means
  - assign face to cluster with closest center,
    - chosen from associated names
  - recompute centers, iterate
  - using distance in LD space
- Now recompute discriminants, recluster with modified k-means



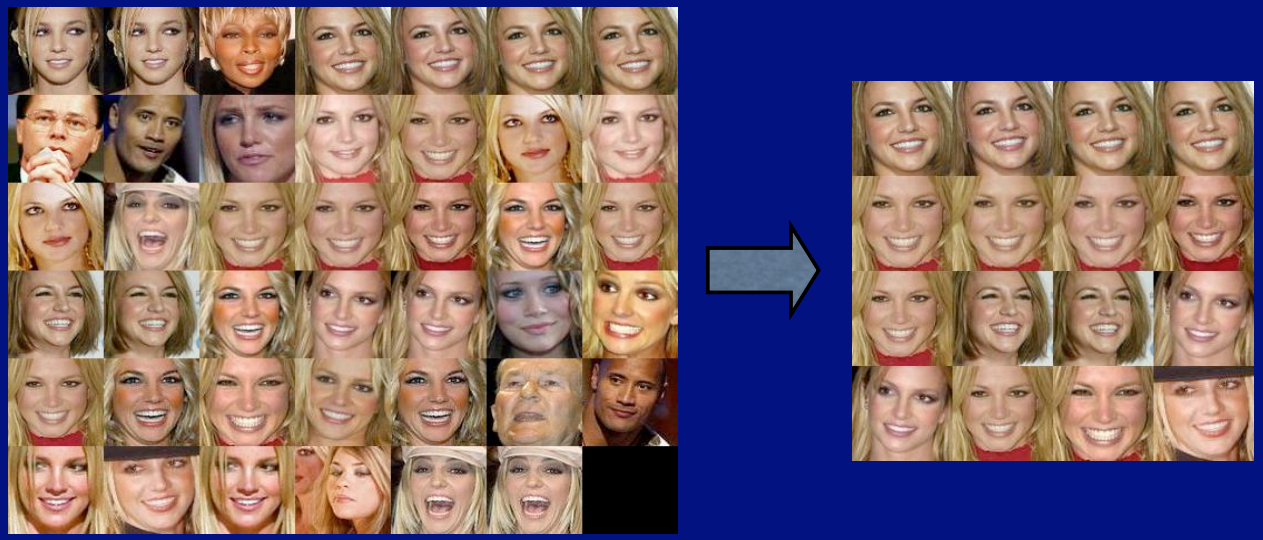
US President George W. Bush (L) makes remarks while Secretary of State Colin Powell (R) listens before signing the US Leadership Against HIV /AIDS , Tuberculosis and Malaria Act of 2003 at the Department of State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/ Luke Frazza)

German supermodel Claudia Schiffer gave birth to a baby boy by Caesarian section January 30, 2003, her spokeswoman said. The baby is the first child for both Schiffer, 32, and her husband, British film producer Matthew Vaughn, who was at her side for the birth. Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)

British director Sam Mendes and his partner actress Kate Winslet arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars Tom Hanks as a Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung

Pruning

- Using a likelihood model
- Tradeoff: size vs accuracy



Merging

Venezuelan
President Chavez

Hugo Chavez

Ryan's clean demo http://www.eecs.berkeley.edu/~ryanw/clustersFull1theta15/index.html

Tamara's demo http://www.cs.berkeley.edu/~millert/faces/faceDict/starClust/

How well does it work?

- Draw a cluster from the list, and an image from that cluster
  - frequency that that image is of someone else

#Images	#Clusters	error rate
19355	2357	26%
7901	1510	11%
4545	765	5.2%
3920	725	7.5%
2417	328	6.6%

- How many bits are required to fix result?

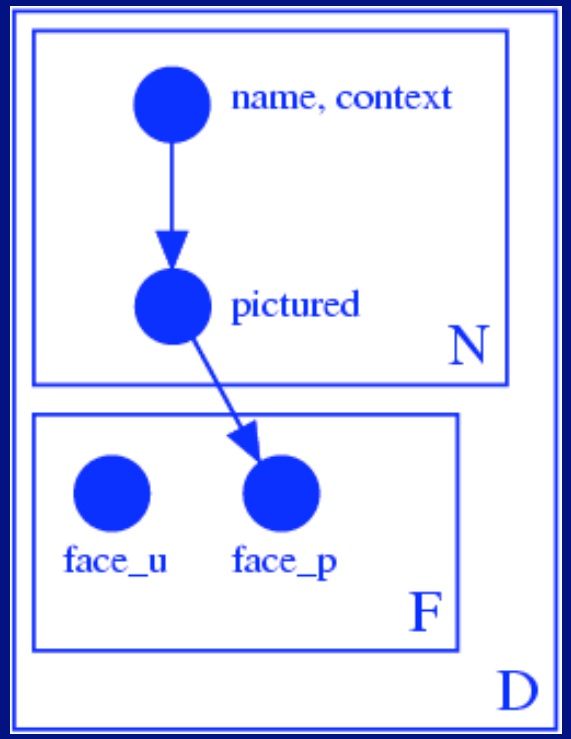
Works - but

- We are missing language cues

Sahar Aziz, left, a law student at the University of Texas, hands the business card identifying Department of the Army special agent Jason D. Treesh to one of her attorneys, Bill Allison, right, during a news conference on Friday, Feb. 13, 2004, in Austin, Texas. In the background is Jim Harrington, director of the Texas Civil Rights Project. (AP Photo Harry Cabluck)

Training a language module

- Idea:
  - a set of named faces is supervised training data for a "who's in the picture" module
  - actually, do EM (or maximize?) over missing correspondences

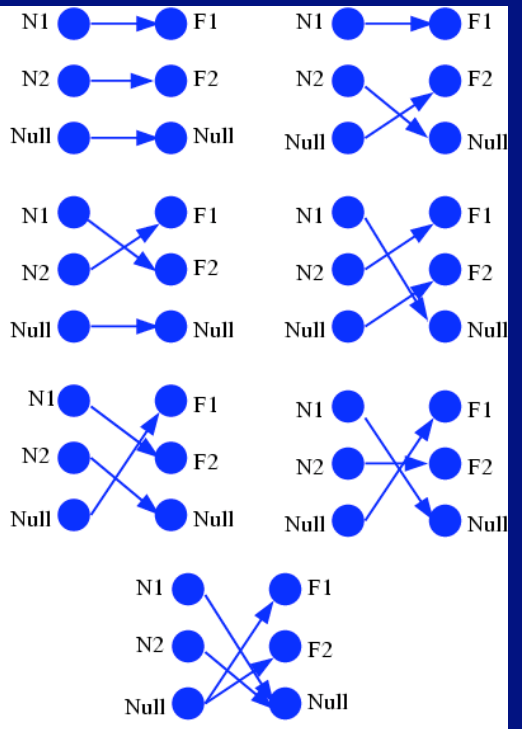


President and Chief Operating Officer of the Vancouver, British Columbia 2010 Bid Corporation John Furlong (rear) smiles while celebrating with compatriots their victory in obtaining the 2010 Winter Olympics bid on late July 2, 2003 in Prague. Vancouver won with 56 votes against 53 votes for Pyeonchang in the second round of balloting at an IOC gathering in Prague. REUTERS/Petr Josek

F1 =

F2 =

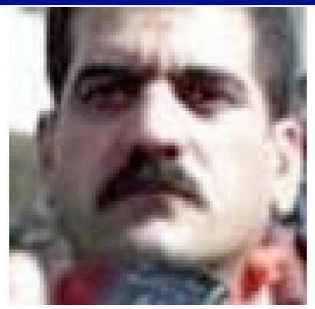
N1 = John Furlong,
N2 = Winter Olympics,



All possible name/face assignments

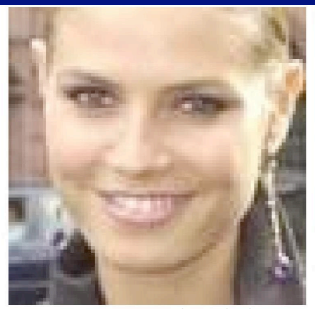
Language improves naming,

before – CEO Summit
after – Martha Stewart

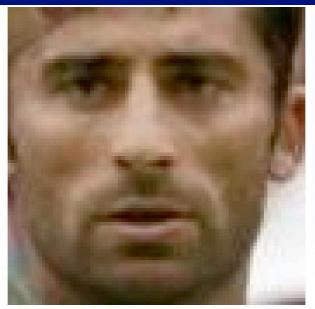


before – U.S. Joint
after – Null

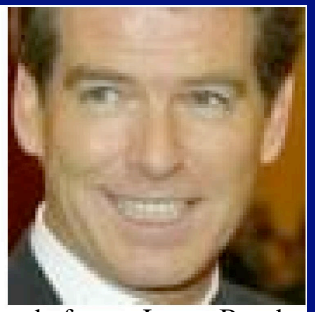
before – Angelina Jolie
after – Jon Voight



before – Ric Pipino
after – Heidi Klum

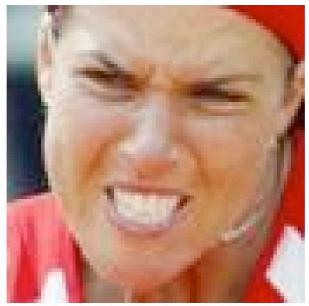


before – U.S. Open
after – David Nalbandian



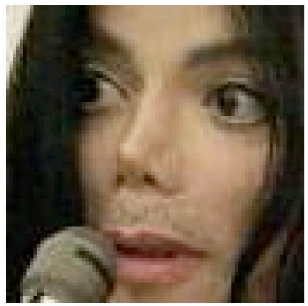
before – James Bond
after – Pierce Brosnan

before – U.S. House
after – Andrew Fastow

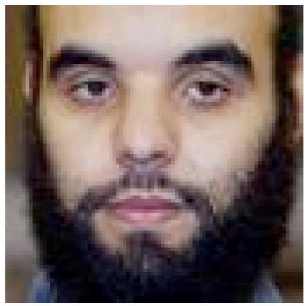


before – Julia Vakulenko
after – Jennifer Capriati

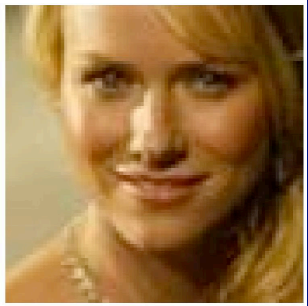
before – Vice President
Dick Cheney
after – President George W.



before – Marcel Avram
after – Michael Jackson



before – al Qaeda
after – Null

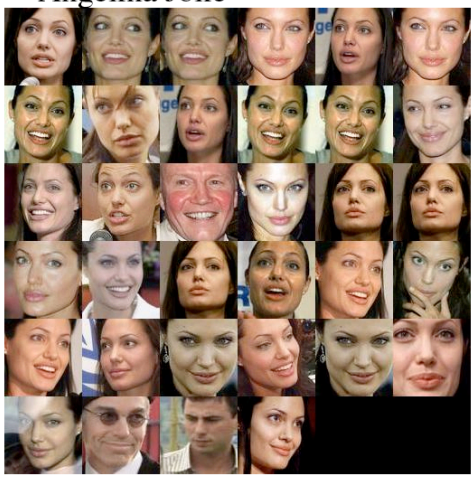


before – James Ivory
after – Naomi Watts

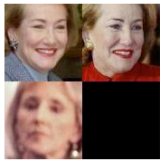
Model	EM	MM
Appearance Model, No Lang Model	56%	67%
Appearance Model + Lang Model	72%	77%

Clusters,

Angelina Jolie



Elizabeth Dole

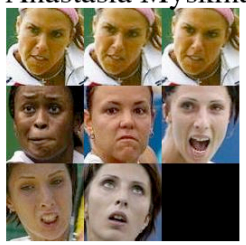


Donald Fehr

U.S. Open

Abraham Lincoln

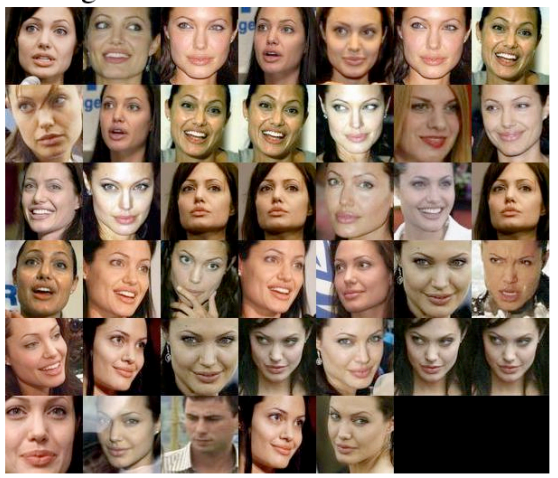
Anastasia Myskina



Without language model

Elizabeth Dole

Angelina Jolie



Donald Fehr

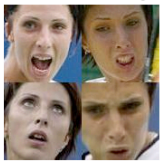
Abraham Lincoln

empty

U.S. Open

empty

Anastasia Myskina



With language model

and yields a useful little NLP module, too

IN Pete Sampras IN of the U.S. celebrates his victory over Denmark's **OUT Kristian Pless OUT** at the **OUT U.S. Open OUT** at Flushing Meadows August 30, 2002. Sampras won the match 6-3 7- 5 6-4. REUTERS/Kevin Lamarque
Germany's **IN Chancellor Gerhard Schroeder IN**, left, in discussion with France's **IN President Jacques Chirac IN** on the second day of the EU summit at the European Council headquarters in Brussels, Friday Oct. 25, 2002. EU leaders are to close a deal Friday on finalizing entry talks with 10 candidate countries after a surprise breakthrough agreement on Thursday between France and Germany regarding farm spending.(AP Photo/European Commission/HO)
'The Right Stuff' cast members **IN Pamela Reed IN**, (L) poses with fellow cast member **IN Veronica Cartwright IN** at the 20th anniversary of the film in Hollywood, June 9, 2003. The women played wives of astronauts in the film about early United States test pilots and the space program. The film directed by **OUT Philip Kaufman OUT**, is celebrating its 20th anniversary and is being released on DVD. REUTERS/Fred Prouser
Kraft Foods Inc., the largest U.S. food company, on July 1, 2003 said it would take steps, like capping portion sizes and providing more nutrition information, as it and other companies face growing concern and even lawsuits due to rising obesity rates. In May of this year, San Francisco attorney **OUT Stephen Joseph OUT**, shown above, sought to ban Oreo cookies in California – a suit that was withdrawn less than two weeks later. Photo by Tim Wimborne/Reuters REUTERS/Tim Wimborne

Classifier	labels correct	IN correct	OUT correct
Baseline	67%	100%	0%
EM Labeling with Language Model	76%	95%	56%
MM Labeling with Language Model	84%	87%	76%

Faces - To do

- Better image features
- More sophisticated probability model, EM
- Estimate P (no pic | name) using EM
- Better named entity recognition
- Co-reference resolution (across languages?) using faces
- Use non-parametric face model (animation?)
- Start looking at face recognition

Partially supervised data == Missing correspondence

- Supervised data, but with a little bit missing
  - There's not all that much unsupervised data but lots of semi-supervised
- Linking and association
  - picture is labelled, but object not segmented
    - Faces (Leung, Burl, Perona, 95); Faces and cars (Weber Perona 01); Faces,cars,motorbikes,planes,tigers (Fergus Zisserman Perona 03); Animal pix (Schmid 01); Clustering (Barnard et al, 01, 01); word prediction (Barnard et al 03; Wang et al, 02; Lia et al, 03;); album cover-music (Brochu et al; 02); objects (Duygulu et al, 02; Barnard et al 03); names and faces (Miller et al 04); speech and pictures (Fleck et al, 04 patent).
  - Words, metadata should be linked to picture
    - Face pix (Srihari, 95); Corel (Barnard et al 01; Li+Wang 03); Art (Barnard et al. 01);
- Coherence
  - Objects of interest look coherent from frame to frame in video
    - People tracking (Ramanan+Forsyth '03); Animals (Ramanan+Forsyth '03)
  - Picture posesses noisy label; which labels are right?
    - Image search results (Fergus et al 04)
- missing data tends to be correspondence

Conclusions

- There's more data out there about the visual world than immediately meets the eye
- Visual information should be linked with other forms of information
  - so one can work where it's easiest
- Doing so may yield useful artifacts and insights