

Statistical Models for Scene Analysis

Collaborators: Not responsible for what I say.

Donald Geman

Stuart Geman

Alain Trouvé

Elliot Bernstein

Advantages of Statistical Models

- Statistical models - a **principled method** of assigning weights and comparing different hypotheses.
- Simple data models **conditional independence (CI)** - push complexity into latent variables.
 - Deal effectively with **large numbers of classes**.
 - Learn each class **separately**.
 - No need for massive training sets.
- **Invariance** modeled explicitly through **latent** instantiation or pose variables:
deformable templates.
- CI data models allow for explicit modeling of **occlusion**.
- Easy formulation of various **alternative hypotheses**.
- Object models can be composed into **scene models**, and decomposed **into sub-object models**.

A range of applications

Using the same modeling framework we can 'solve' (address) the following:

Classify handwritten characters.

Training data per class	Error rate	SVM
10	6.05	12.61
30	3	6.17
100	1.9	3.02
1000	.8	1.15
3000	.66	?

Read zipcodes - only individual class models

84% correct (out of 1000 for CEDAR database).

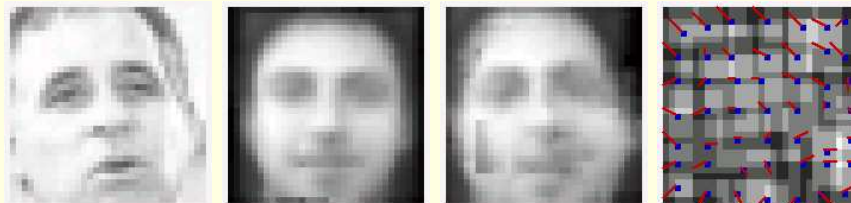


Read a license plate. 95% correct.



Detect faces with instantiation information.

(Training set 300: D=.85 FP=97)



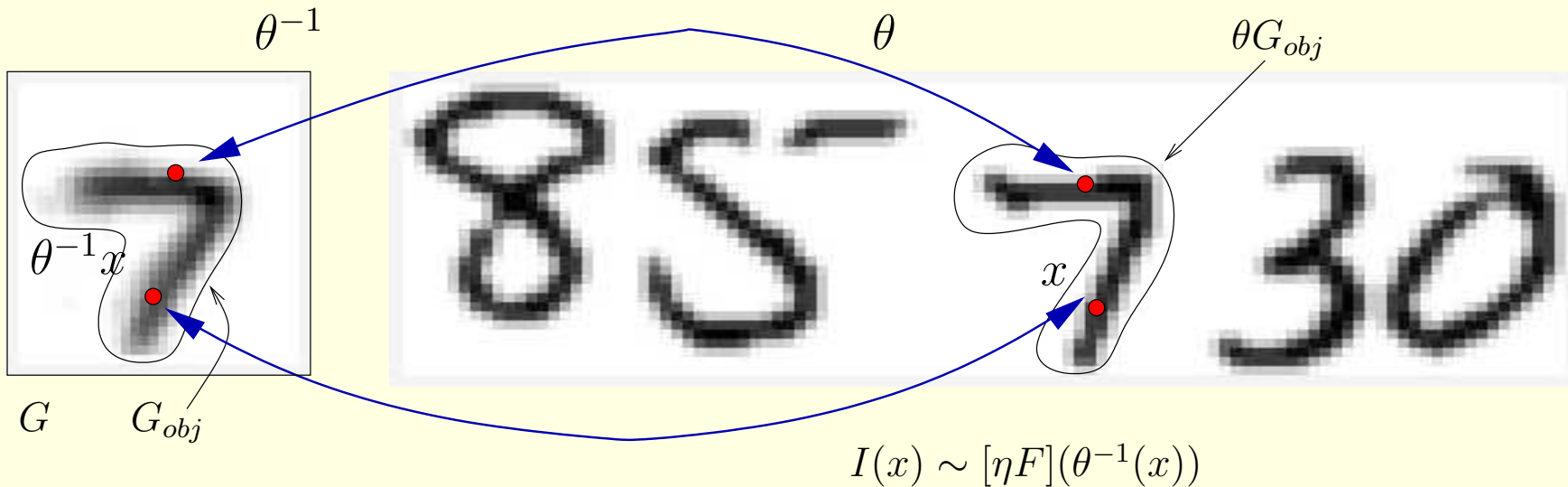
Detect cars with pose. (Training set 100).



Detect horses with instantiation information. (Training set 100).



Deformable Templates



'7' instantiated at *geometric map*: $\theta : G \rightarrow L$.

and *photometric 'map'*: η (contrast + baseline).

Define *model support*: $G_{obj} \subset G$ -

Set of $x \in G$ such that:

conditional on $\theta = \textit{identity}$,

distribution of $I_W(x)$ (nbhd. of x) - different from bgd. dist.

Data model: Conditional Independence

$$P_{obj}(I|\theta, \eta) = \prod_{x \in \theta G_{obj}} f_{obj}(I(x)|\theta, \eta) \prod_{x \notin \theta G_{obj}} f_{bgd}(I(x)|\theta, \eta).$$

$$f_{obj}(I(x)|\theta, \eta) = g(I(x); [\eta F](\theta^{-1}x)).$$

Divide by likelihood of **no** object, and take logs:

$$\log \frac{P_{obj}(I|\theta, \eta)}{P_{bgd}(I|\eta)} = \sum_{x \in \theta G_{obj}} \log \frac{f_{obj}(I(x)|\theta, \eta)}{f_{bgd}(I(x)|\eta)}$$

Sum is restricted to support θG_{obj} .

An object class c will be represented as mixture of product models

$$P_c(I|\theta, \eta) = \sum_m \pi_m P_{c,m}(I|\theta, \eta).$$

Multiple objects - Scene models

Interpretation - $\mathbf{D} = \{c_i, m_i, \theta_i, S_i, i = 1, \dots, n\}$,

ordered according to occlusion.

$S_i = \theta_i G_{c_i, m_i}$ support of detection i . Define: $T_i = \cup_{j=1}^i S_j$.

Likelihood ratio

$$\frac{P(I|\mathbf{D}, \eta)}{P_{bgd}(I|\eta)} = \prod_i \prod_{S_i \setminus T_{i-1,e}} \frac{f(I(x)|m_i, \theta_i, c_i, \eta)}{f_{bgd}(I(x)|\eta)}$$

Choose: $argmax_{\bar{c}} \max_{\bar{\theta}, \bar{m}, \eta} P(\hat{I}|\mathbf{D}, \eta) P(\mathbf{D}) P(\eta)$.

All decisions based on likelihood ratios.

Training

General framework.

- Choose parameterization of θ, η .
- Choose number M of components in mixture model.
- Choose parameterization β of template $F(x; \beta)$ + Prior.
- Choose parameterization α of $P(\theta; \alpha)$ + Prior.
- Use clean data samples from object class $I^{(t)}$ and EM framework where $\theta^{(t)}, \eta^{(t)}, m^{(t)}$ are unobserved.

Making it work - Successive approximations

- Get rid of η . **Transform** data to binary oriented edges. (David Jacobs approved it yesterday)

Bernoulli models: Defined through probability maps.

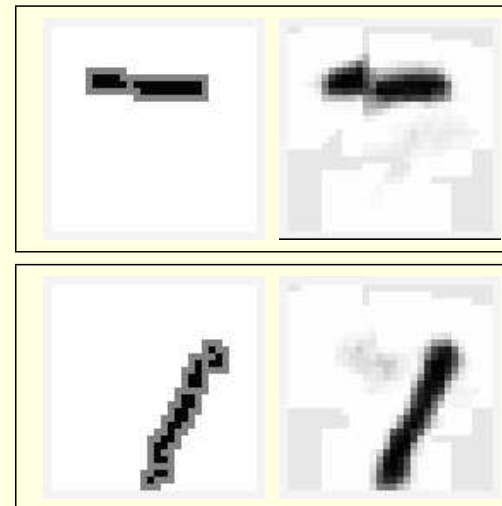
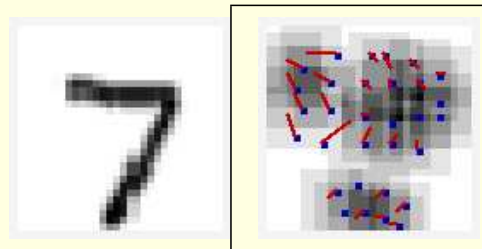
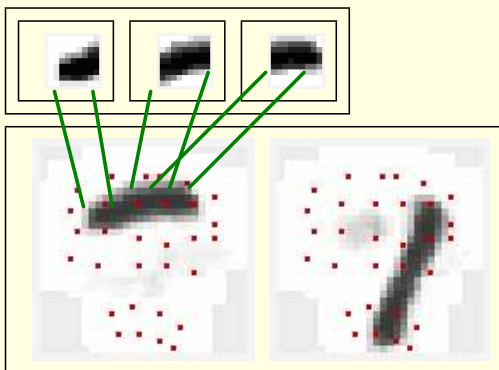
$$p_e(z), z \in G, e = 1, \dots, 8.$$

$$\text{Support: } G_{e,obj} = \{z : p_e(z) > p_{bgd}\}.$$

- Approximate θ as rigid shifts of local submodels

$$Q_i = \{p_e(z), z \in W + z_i\} -$$

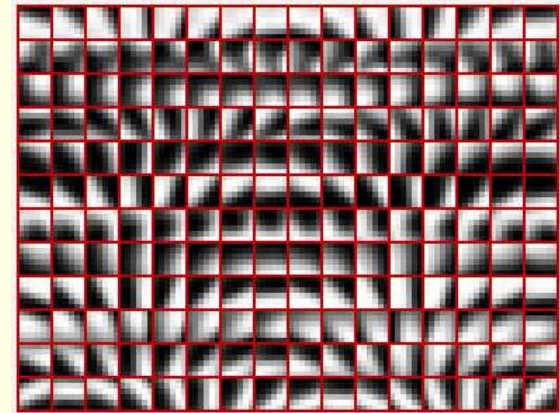
Patched together by averaging at each pixel. (POP) models.



Making it work - Further Approximations - Feed Forward

Identify **recurring submodels** - generic parts - Q_f .

Unsupervised learning of mixture models (with symmetries) for sub-windows... with structure.



- **Transform** binary edge data to binary **part** data.
- **Spread and downsample** to coarse grid.
- **Bernoulli** model on new features.

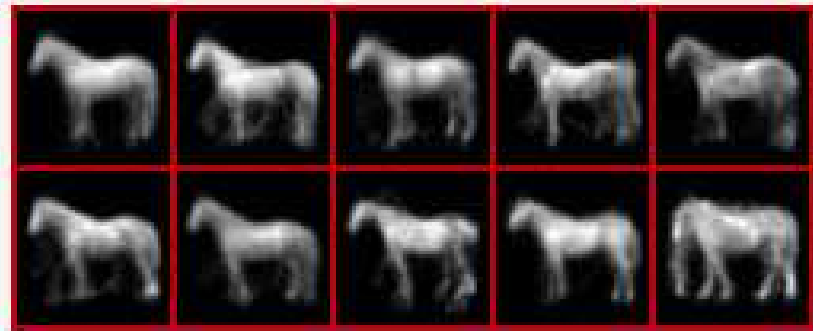


Efficient indexing +

Efficient estimation of mixture components of objects. θ eliminated.

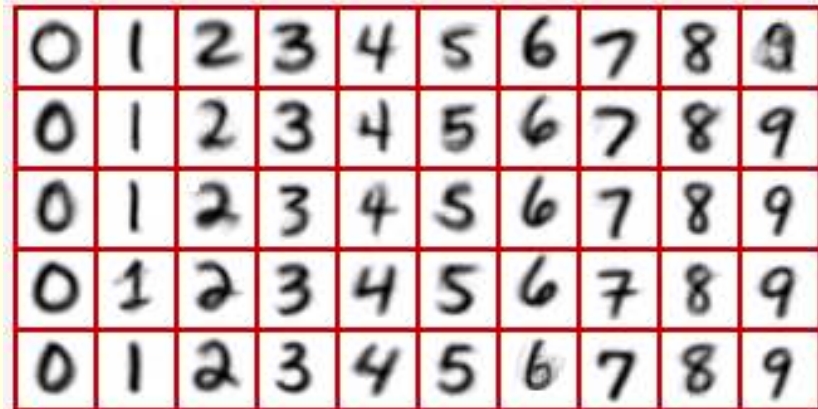
Network Architecture: *Image* \rightarrow
edges \rightarrow *spread* \rightarrow *parts* \rightarrow
spread + subsample \rightarrow *Object*
mixture models \rightarrow *Argmax*.

Refining - Feedback



- Learn fine POP models for components trained on coarse models.

Global mean image: Patch local means modulo shifts. →



- Use indexing from feed-forward to initialize refined matching of POP models.
Obtain object supports and instantiation information.
- Maximize over interpretations - HARD.
Plates, zipcodes - dynamic programming, simple prior on configurations.

Other approximations

Quantize Q_i 's to fixed library Q_f .

Or define by hand.

$$Y_f(x) = \hat{I} * W_f(x).$$

Local MAX on Y_f + subsampling.

(Reisenhuber and Poggio, 1999).

Histograms on Y_f (Lazebnik et. al., 2006)

Binarize -

$$\hat{Y}_f = \mathbf{1}\{Y_f(x) > \tau_f\}.$$

One feature per object region. *Star graph*+SPREAD - (Amit, 2000.)

Use *all* with Cond. ind. for \hat{Y}_f after spreading and subsampling.

(Bernstein and Amit, 2005).

Small number of *non-overlapping* Q_i 's +

Geom. model (θ) - simple graph. (Burl et.

al., 1998, Felzenschwalb and Huttenlocher, 2005).

Only rare \hat{Y}_f 's. Sparse labeled point

proc. Cond. ind. given θ + bgd. model.

(Burl et. al., 1996, Amit 1996).

Model image data at points ($I \cdot \hat{Y}_f$).

(Li et. al., 2006).

Determine Q_i 's based on discriminative training. (Viola and Jones, 2001, Torralba et. al. 2003).

Model $\{Y_i\} = X * W_i$ in window as cond. ind. given object/bgd. Kanade and

Schneiderman (1998), Ullman et. al. (2002).

Other connections: Nearest neighbor methods

- **Tangent distance:** (Hastie and Simard, 1998).
 - Parameterization of deformations.
 - Data model - conditional on deformation of template i.i.d Gaussian noise.
 - Every example in every class is a 'template'
 - Deformation process approximated by linearization.
- **Shape context:** (Belongie et. al., 2002)
 - Pick 100 points on reference grid - put down wedges.
 - Add probabilities from POP model of each edge orientation in each wedge.
 - Binomial model for counts in each wedge in data.

Conditional independence models for objects with latent instantiation variables.

- Easily defined *approximate* models.
- Approximations yield tractable *estimation* procedures.
- Approximate models yield efficient *indexing*. Naturally incorporated in **CTF** framework.
- Easily composed to *scene* models.
- Naturally *decomposed* to part models - *clutter* models or *alternative hypotheses*,
(see next slide.)
- Possibility of *online reestimation* of background parameters (e.g. edge density).
- Various levels of approximation are analogous to many existing algorithms. *Useful framework for discussing and comparing algorithms.*

Some References

- Amit Y. and Geman, D. and Fan X. (2004), A Coarse-to-Fine Strategy for Multi-Class Shape Detection. IEEE PAMI.
- Amit Y. and Trouve A. (2005), POP: Patchwork of Parts Models for Object Recognition, Languishing in IJCV, 14 months and counting.
- Bernstein E. and Amit Y. (2005), Part-Based Statistical Models for Object Classification and Detection, CVPR (2005).
- Amit Y. and Trouve A. (2006), Generative Models for Labeling Multi-Object Configurations in Images. SICILY Object Recognition workshop.
- **Stuart Geman (1988-2006), Personal Communications.**



Original image.



Global threshold on LRT to generic background with locally estimated edge density.



Threshold on LRT around head + back area: test against 'generic four legged alternative'.

Colored areas are supports of instantiated objects using edge based models.



Comparing two alternative occlusion ordering alternatives. White horse in front of black horse wins.