History of Speech Recognition

- **1965-90: looking for features (spectrum, LPC, cepstrum, cochlear feat.)**
- **1965-75: isolated-word global template matching (nearest neighbors)**
- **1975-85: deformable template matching (nearest neighbors)**
- **1980-90: structural methods / expert systems (no learning, failure)**
- **1985-90: HMMs (lots of learning, generative models, non-convex!)**
- **1990-95: global generative learning (sentence-level HMMs)**
- **1990-95: word-level discriminative learning (HMMs, non-convex!)**
  - mixtures of Gaussians, neural nets
- **1995-00: sentence-level discriminative learning (HMMs, non-convex)**
- **what made it work:**
  - lots of data+huge models, training the segmenter, generative +discriminative training, non-convex/non-linear learning

Panel on Shape Representation

- Yann LeCun: recognition architectures and representation learning.
- Martial Hebert: Shape Representation, the historical perspective
- Jean Ponce: Feature Representations, an overview

History of Handwriting Recognition

- **1965-90: looking for features**
  - edges, projections, chain code, Zernicke moments, Fourier, Haar, Hadamard, Hough,......
- **1965-75: classifiers for isolated characters**
  - nearest neighbors, linear classifiers
- **1975-85: structural methods (no learning, failure)**
- **1985-95: learning the features (lots of learning, non convex!)**
  - neural nets, convolutional nets
- **1990-00: global learning (lots of learning, context, non convex!)**
  - word-level discriminative learning (d-HMM, graph transformer nets)
- **since then, people keep re-inventing the same thing**
- **what made it work:**
  - lots of data, training the segmenter, integrated discriminative training, learning the features (deep learning), non-convex/non-lin.

History of Image Recognition

- **1965-2008: looking for features**
  - edges, countours, Hog, Sift, Shape Context,......
- **1965-08: linear classifiers (Perceptrons!), nearest neighbor classifiers**
- **1975-95: structural methods (no learning, failure)**
- **1993-01: learning the features for face detection (learning, non convex!)**
  - neural nets, convolutional nets, boosted cascades.
- **1990-00: structured output models (lots of learning, context, non convex!)**
  - word-level discriminative learning (d-HMM, graph transformer nets)
- **what made it work (so far):**
  - learning, discriminative learning, designing the right features
- **what's missing:**
  - learning the features, integrated segmenter, unsupervised/supervised learning

The Future of Image Recognition

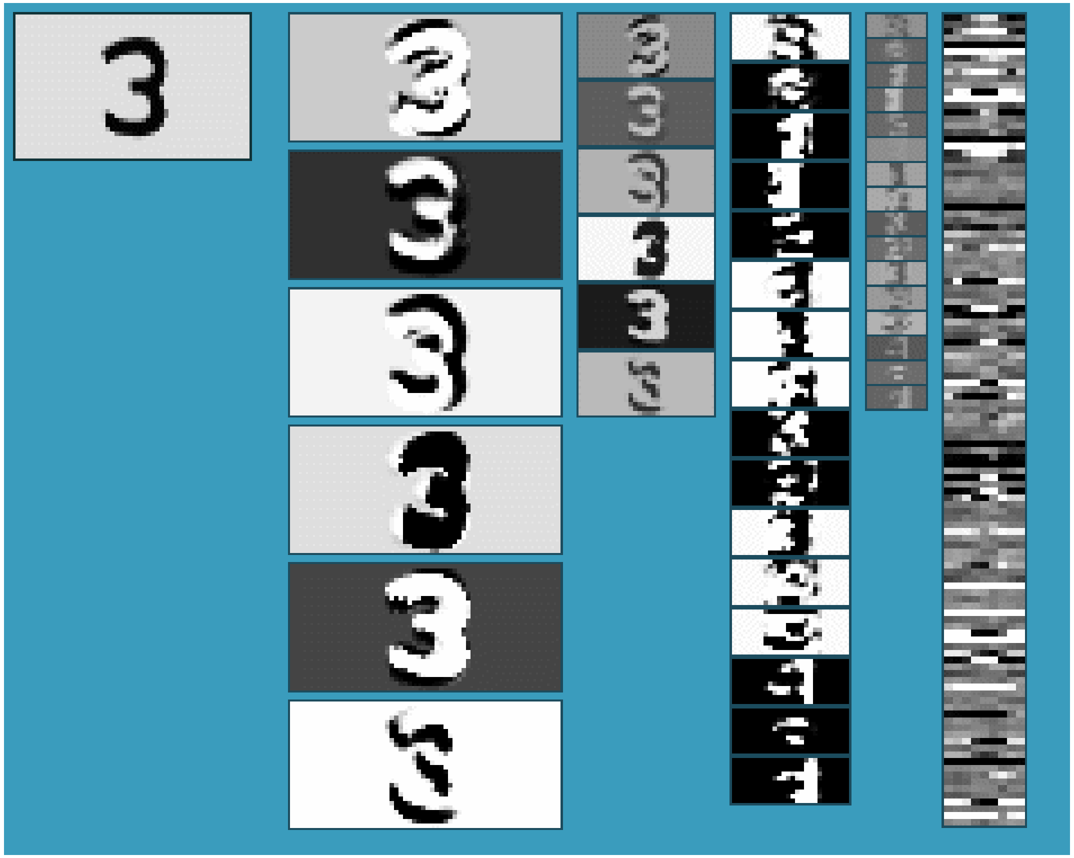
- **We are still looking for the right features**
  - we should try to learn them
  - ...but so far, feature learning for object recognition has not worked as well as for handwriting recognition
  - do we have the right learning algorithms (deep learning!)
- **We are still stuck with "linear" learning and/or nearest neighbors**
  - let's move beyond SVMs and K-NN
  - non-linear/non-convex learning was essential for speech and handwriting: mixtures of Gaussians, convolutional nets.....
- **We are just getting started with integrated (global) training**
  - training the segmenter was crucial to making speech and handwriting recognition systems work.
  - segmentation/pose are treated as latent variables.
  - This kind of approaches will be crucial for dealing with invariance
  - They will be essential for compound objects with movable parts
    - (see Ramanan/Felzenswalb/McAllester)

Do we have the right architecture?

- **Speech and Handwriting have settled on an architecture**
- **Image recognitions systems are just about to settle on an architecture**
  - 04: interest points -> global spatial pooling -> classification
  - 05: interest points -> local spatial pooling -> elastic template matching
  - 06: local feature detectors -> local spatial pooling -> classification
- **But these models are "shallow"**
  - The mammalian visual cortex is deeper
  - multiple stages of:
  - local feature detectors (simple cells) -> local pooling (complex cells)
  - Convolutional nets, HMAX.......
- **We will be converging towards the "Multistage Hubel-Wiesel Architecture"**
  - Hierarchy of increasingly invariant features
  - We will have to learn the features
  - We can design the first layer, but not the next layers!

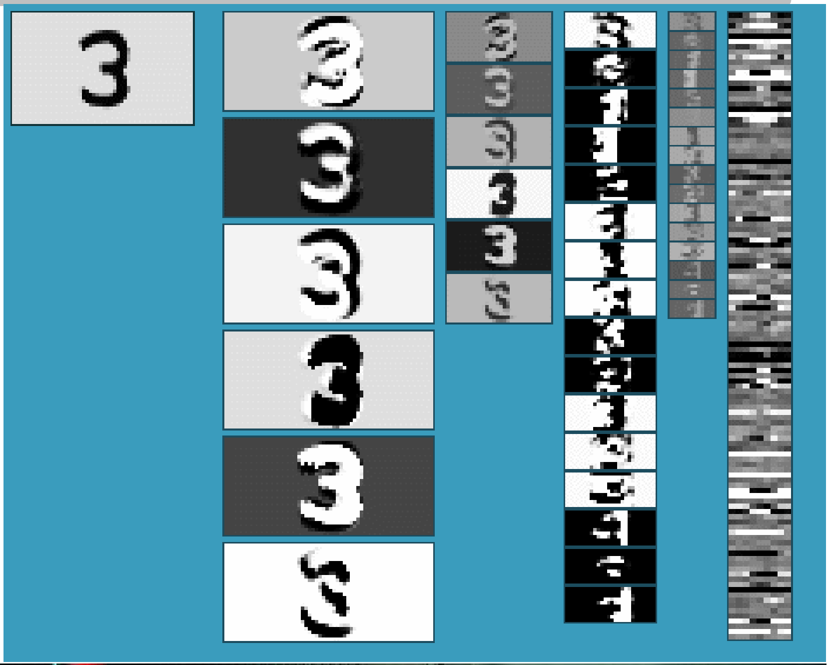
Deep Architectures for Vision: Convolutional Network

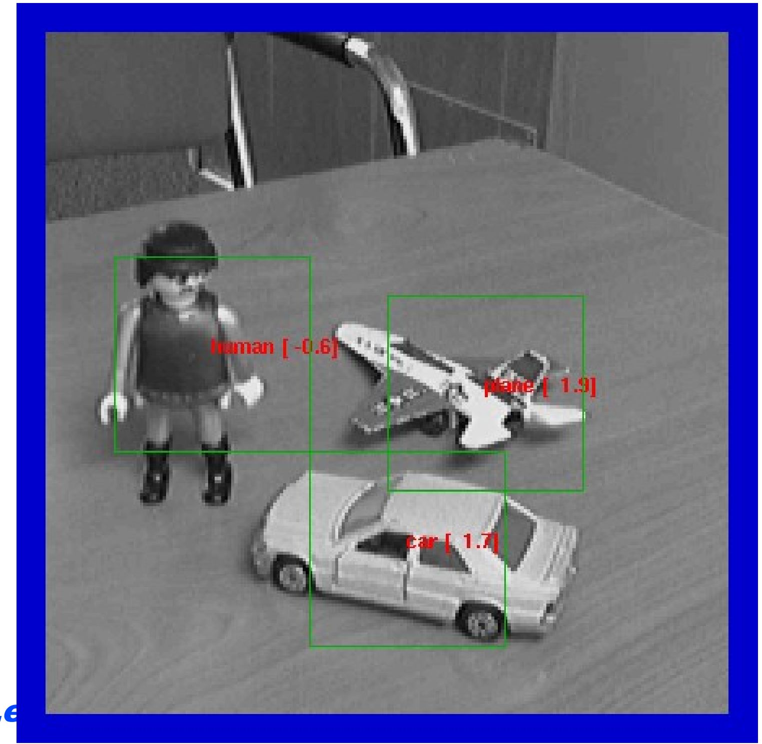
- **Building a complete artificial vision system:**
  - Stack multiple stages of simple cells / complex cells layers
  - Higher stages compute more global, more invariant features
  - Stick a classification layer on top
  - [Fukushima 1971-1982]
    - neocognitron
  - [LeCun 1988-2007]
    - convolutional net
  - [Poggio 2002-2006]
    - HMAX
  - [Ullman 2002-2006]
    - fragment hierarchy
  - [Lowe 2006]
    - HMAX

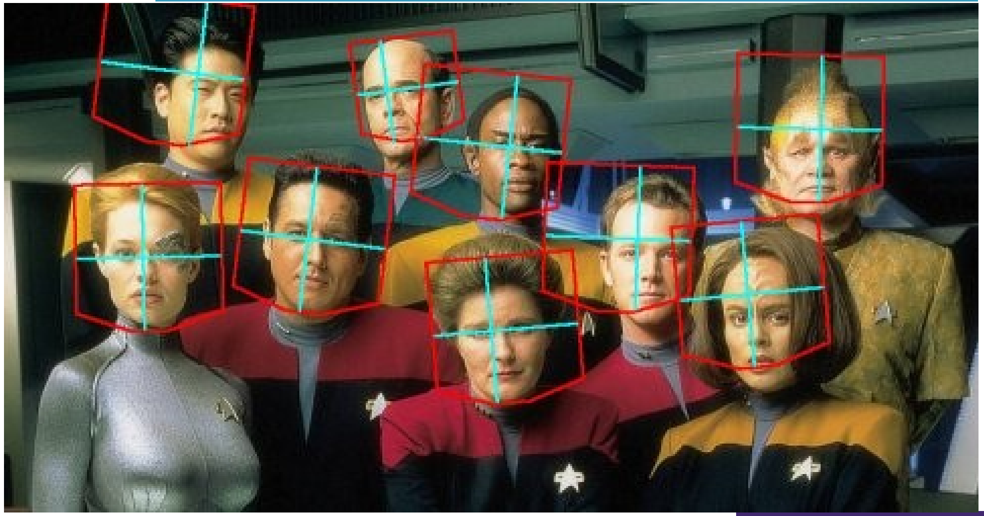


Supervised Convolutional Nets learn well with lots of data

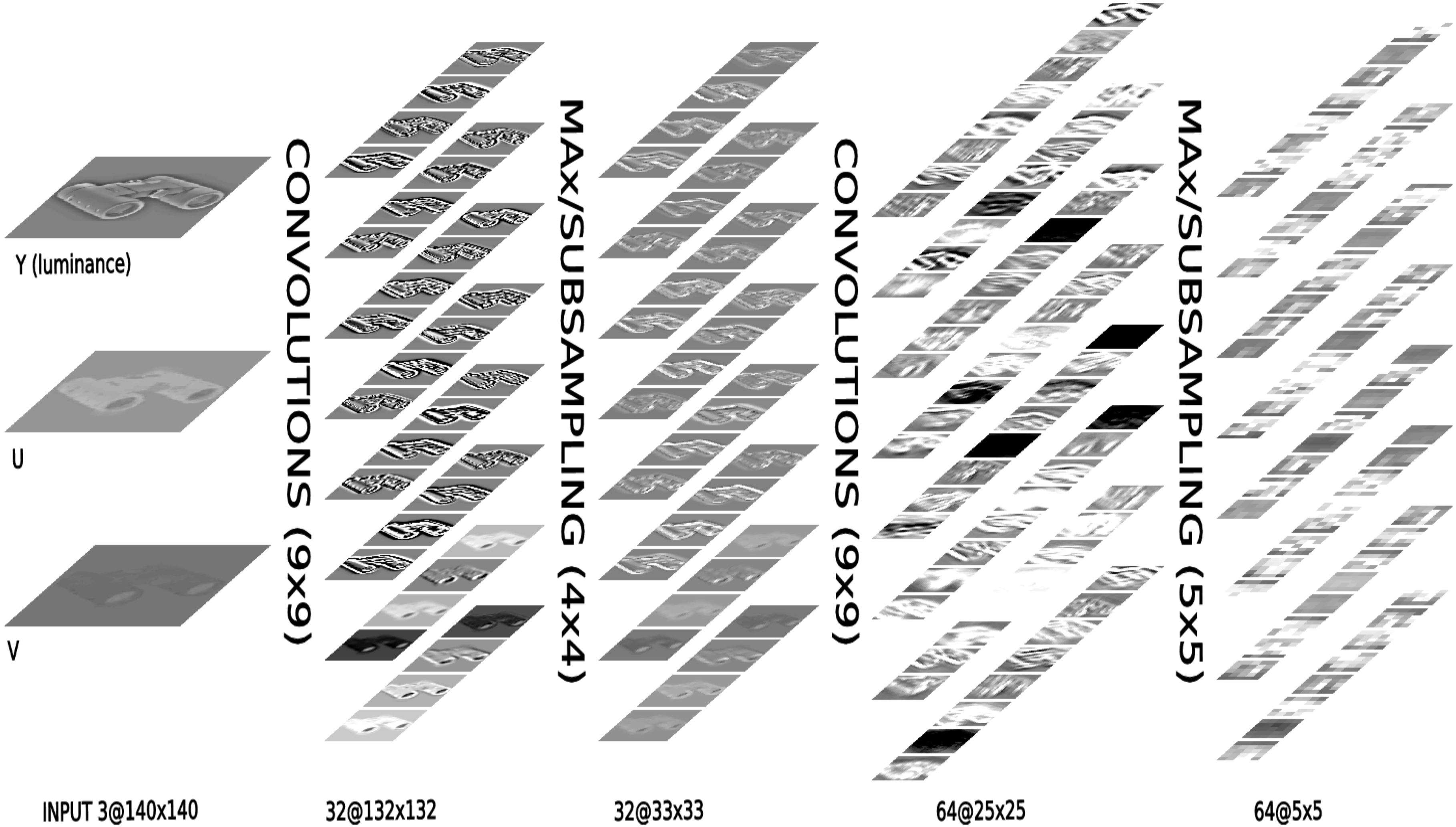
- **Supervised Convolutional nets work very well for:**
  - handwriting recognition(winner on MNIST)
  - face detection
  - object recognition with few classes and lots of training samples





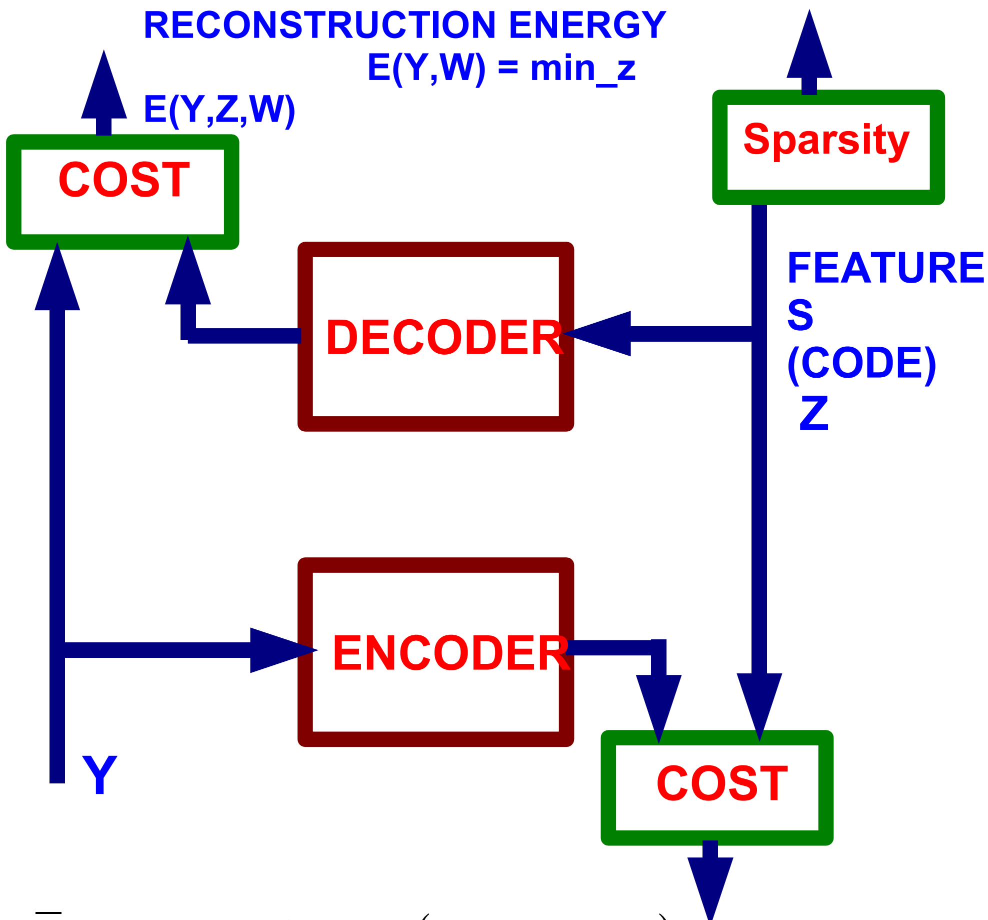


Learning a Feature Hierarchy for Object Recognition



Learning the Features?

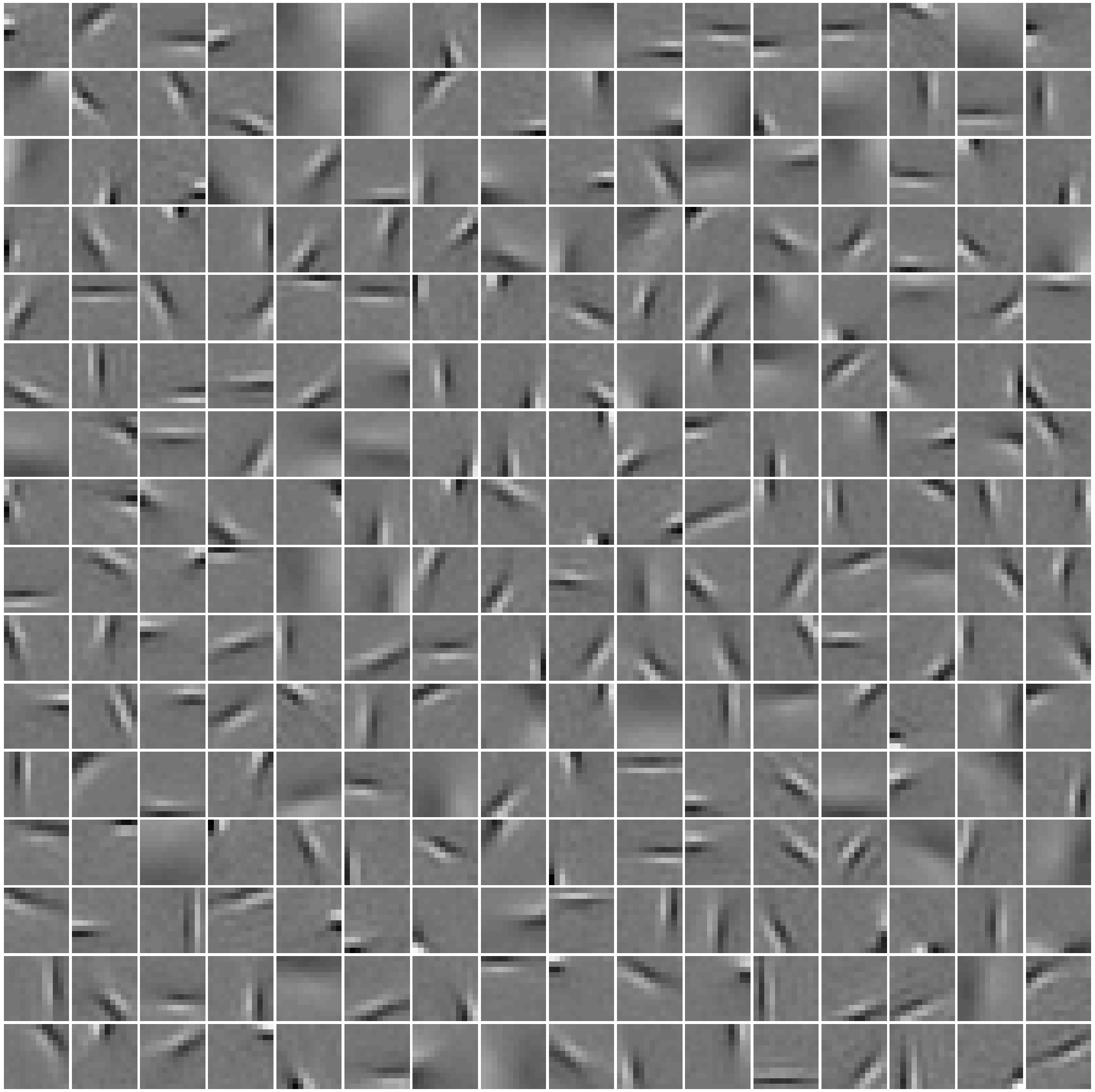
- **Decoder:**
  - Linear
- **Optional encoders of different types:**
  - None
  - Linear
  - Linear-Sigmoid-Scaling
  - Linear-Sigmoid-Linear
- **Optional sparsity penalty**
  - None, L1, Log Student-T
- **Feature Vector Z**
  - continuous



$$\bar{Z}_Y = argmin_Z E(Y, Z, W)$$

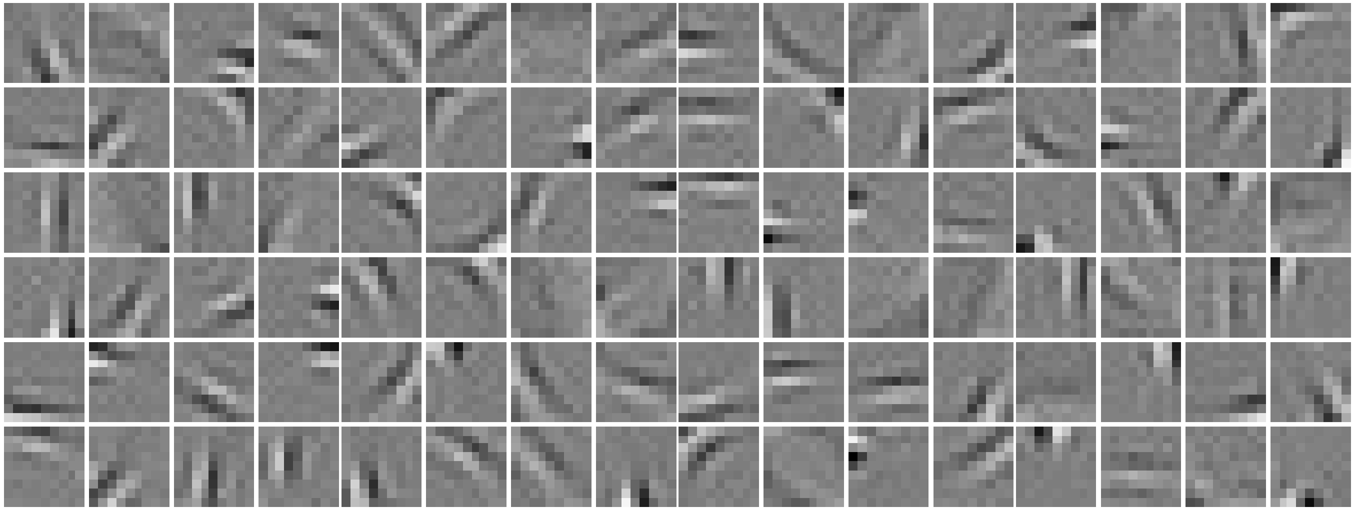
$$E(Y, W) = min_Z E(Y, Z, W)$$

Learning the Right Features?



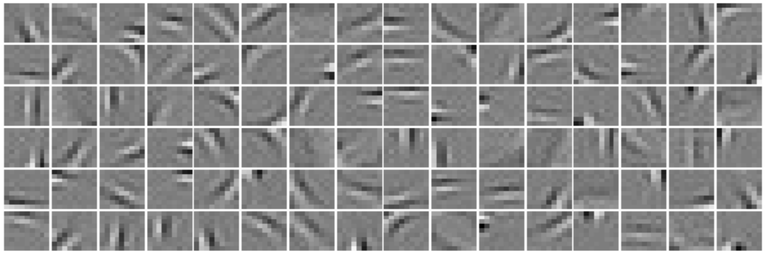
Learning the Features

- **96 filters on 9x9 patches trained with PBP**
  - with Linear-Sigmoid-Gain Encoder
- **Recognition:**
  - Normalized_Image -> Learned_Filters -> Rectification -> Local_Normalization -> Spatial_Pooling -> PCA -> Linear_Classifier
  - What is the effect of rectification and normalization?



weights :-0.9275 - 0.8688

Caltech-101 Recognition Rate



- **[96_Filters->Rectification]->Pooling->PCA->Linear_Classifier**
  - [Filters->Sigmoid] 16%
  - [Filters->Absolute_Value] 51%
  - [Local_Norm->Filters->Absolute_Value] 56%
  - [Local_Norm->Filters->Absolute_Value->Local_Norm] 58%
- **Multi-Scale Filters->Rectification->Pooling->PCA->Linear_Classifier**
  - LN->Gabor_Filters->Rectif->LN (Pinto&diCarlo 08) 59%
- **Unsupervised Convolutional Net**
  - Filt->Sigm->Pooling->Filt->Sigm->Pooling->Classifier 54%
- **Supervised Convolutional Net**
  - Filt->Sigm->Pooling->Filt->Sigm->Pooling->Classifier 20%

Martial Hebert

- Context and scene interpretation
- Background knowledge
- Parts
- Geometry
- Shape and relations

Fig. 3.3a

Fig. 3.4a

Fig. 3.5a

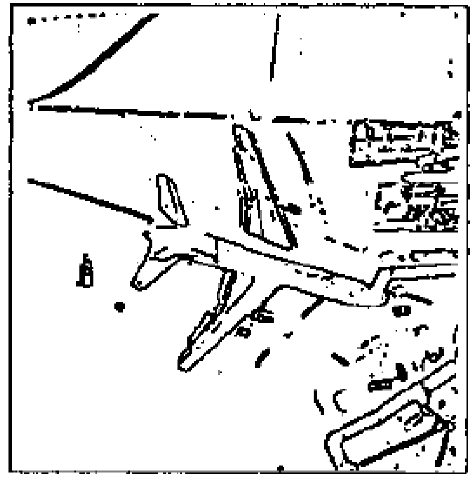
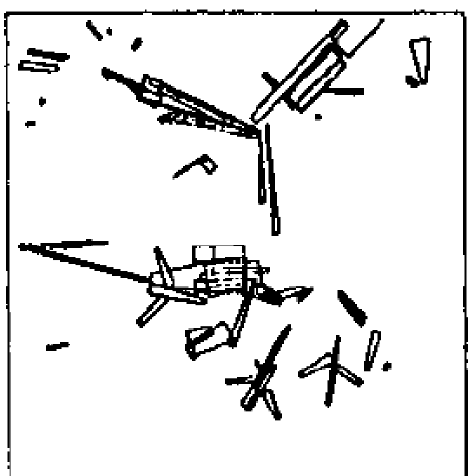
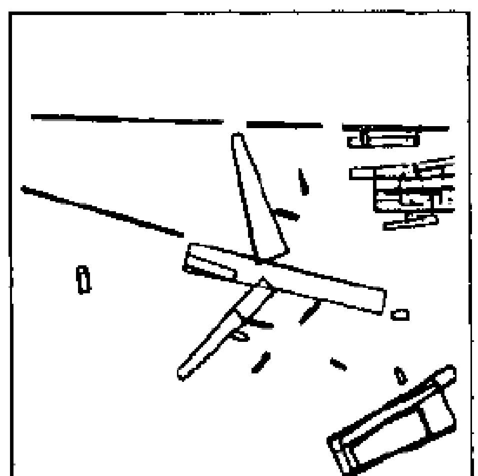


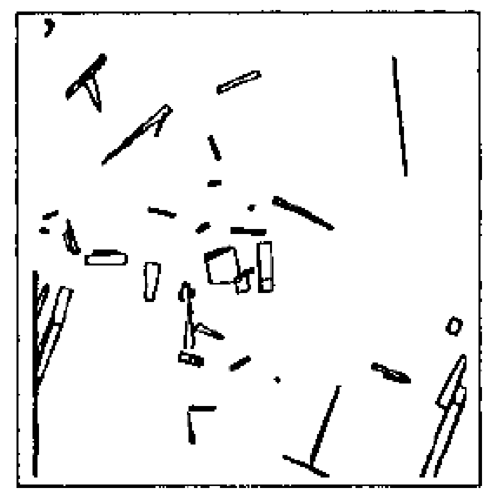
Fig. 3.3b

Fig. 3.4b

Fig. 3.5b







surface of the cylinder. It predicts that the length of the ribbon in the image will in fact be:

$$\frac{-2.42 \times CYL\_LENGTH \times \cos(-TILT)}{CYLINDER.CAMZ}$$

where 2.42 is the focal ratio of the camera and CYLINDER. CAMZ is an internal quantifier generated by the prediction module.

Both of the above approaches are used to generate back constraints to ensure coverage of all the relevant quantifiers. They are:

$$m_h \geq -2.096 \times CYL\_LENGTH \times (1/CYLINDER.CAMZ)$$
$$m_l \leq -2.338 \times CYL\_LENGTH \times (1/CYLINDER.CAMZ)$$
$$-TILT \leq -\arccos(\sup(-0.413 \times m_h \times CYLINDER.CAMZ \times (1/CYL\_LENGTH)))$$
$$-TILT \geq -\arccos(\inf(-0.413 \times m_l \times CYLINDER.CAMZ \times (1/CYL\_LENGTH)))$$

Brooks&Binford'81 (Nevatia&Binford'73)

The famous 1971 Binford Paper

Thomas O. Binford
Stanford University Artificial Intelligence Project

We describe a formal representation for a class of primitive three-dimensional shapes. Those primitive representations are combined into compound articulated representations of familiar objects. With regard to primitive representations, we discuss only the formalism, not the inference of such descriptions from visual data for complex scenes. The primary design criteria for a representation are the ease with which we can recognize an object as essentially similar to another we have seen before, or the ease with which we can identify that objects with distinct differences have important similarities (a child and an adult, or a man and a woman). This is one basis for generalization. A representation is intended to express low-level knowledge about shape, i.e. class knowledge about familiar shapes, and to serve as a basis for approximation of shape, and conjecture about missing information, for example, the hidden half of objects. The primary criterion is not the simplicity

Binford'71

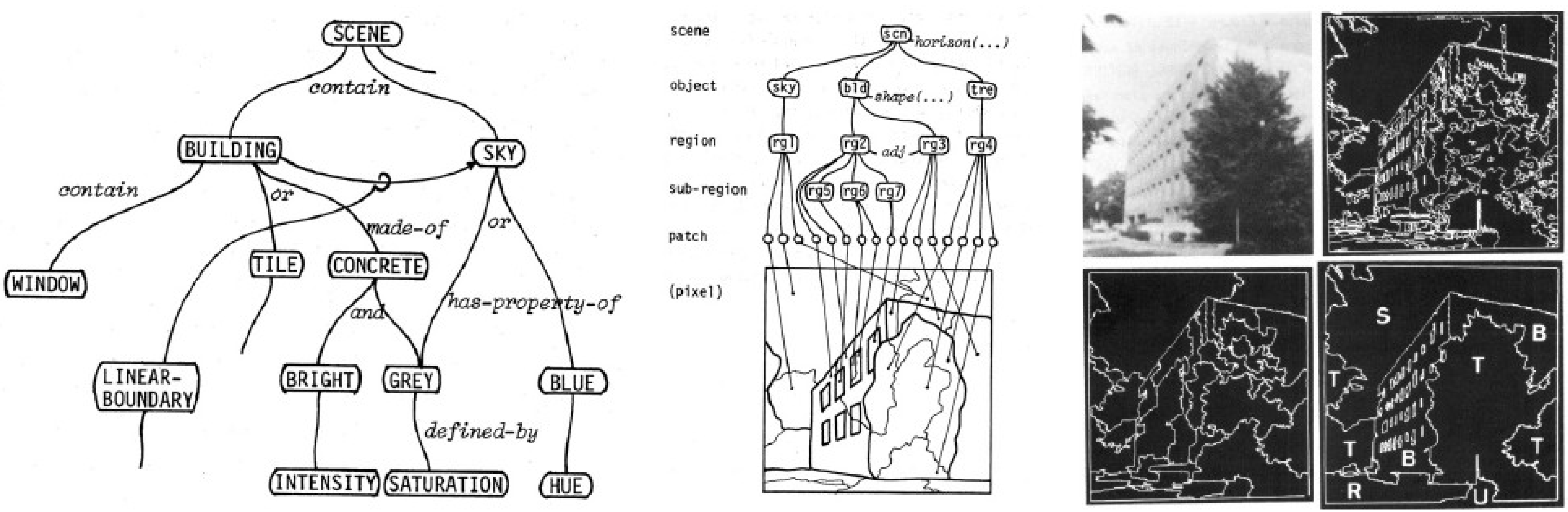
Table 2—*continued*

graphic explanation (not a part of the model)	MODEL (stored in memory)	Comments (not part of the model)
	R13 = R1 or R132	or
ellipse	R29 =	terminal
	R9 = R91 or R92	or
	R10 = R101 or R102 or R103	or or
	CUP = R2, R19, R62, R31	
	R12 = R121 or R122	or
	SUPERIOR EXTREMITY = R15–R16 or R15–R17–R16	or
	R15 = R151 or R152 or R153 or R154	or or or
	R16 = R161 or R162 or R163 or R164	or or or
	R17 = R151 or R172	or
	R18 =	terminal model
	INFERIOR EXTREMITY = R20 R151–[optional] R22 sequence	acts same as R20 – R22 or R20 – R151 – R22
	R20 = R152 or R151 or R203 or R204 or R205	or or or or

Table 2—*continued*

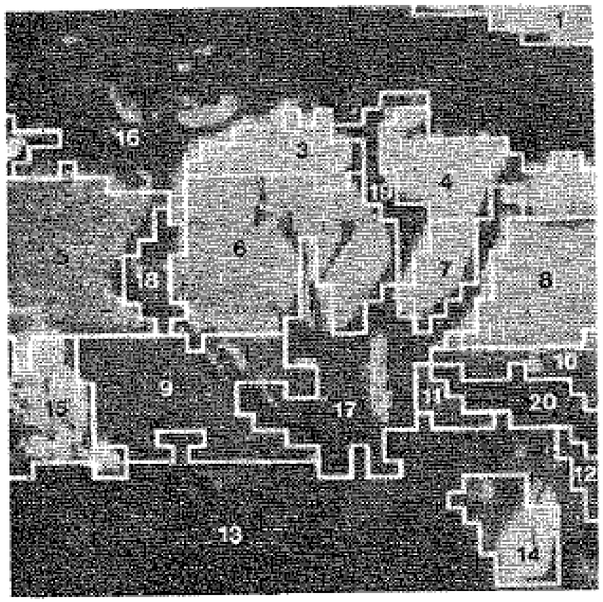
graphic explanation (not a part of the model)	MODEL (stored in memory)	Comments (not part of the model)
	R151 =	terminal
	R22 = R222 or R223 or R224 or R262	or or or
definition of boy 	BOY = HAT [optional]; R25 — HEAD — R25 [optional]; R23 [opt]; R26 or R24 [optional]; SUPERIOR EXTREMITY; SUPERIOR EXTREMITY [optional]; R18; R27; INFERIOR EXTREMITY; INFERIOR EXTREMITY [optional]	[optional] male person. young
	HEAD = F-HEAD or S-HEAD	or
	R23 =	terminal
	R24 = R261 — R261, R261, R102	
	CHINESE = R1, R2, R102, R102, R2, R103, R31, R18	R1, R2, R102, R103, R31, R18
	R26 = R261 — R261 or R262 — R262	or
	HAT = R31, R30 [optional], R290 sequence	sloppy 0.2 
	R25 =	terminal

Guzman'71



(a) Bottom-up process (b) Top-down process (c) Result

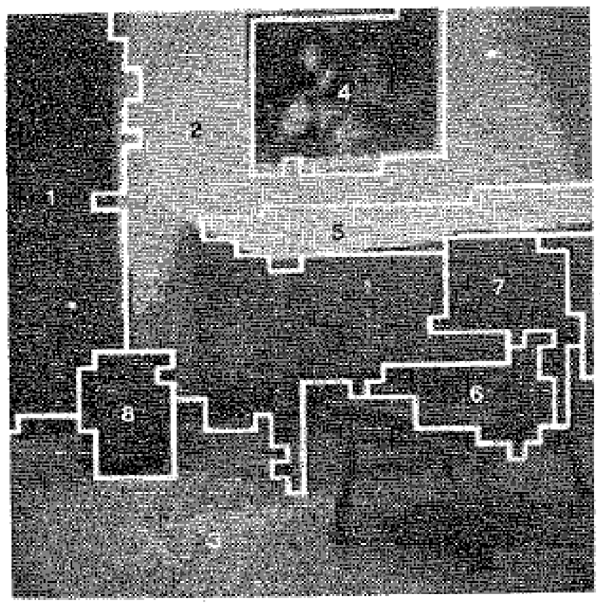
Ohta&Kanade



Final Region Interpretations

Interpretations	Regions
Sky	1,2,3,4
Mountain	5,6,7,8
Sea	9,10,11,12
Ground	13
Rock	14,15
Tree (Crown)	16
Tree (Bark)	17,18,19,20

FIG 4 Final semantic partitioning of landscape scene

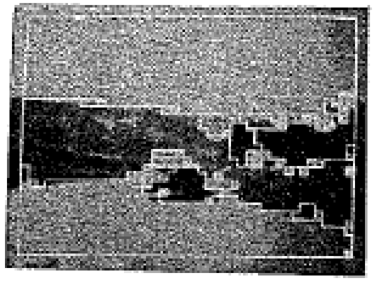


Final Region Interpretations

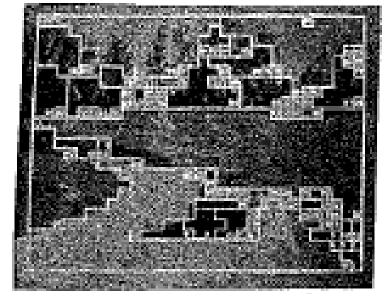
Interpretations	Regions
Door	1
Wall	2
Floor	3
Picture	4
Tabletop	5
Chairseat	6
Chairback	7
Waste Basket	8

FIG. 5. Final semantic partitioning of SRI office scene

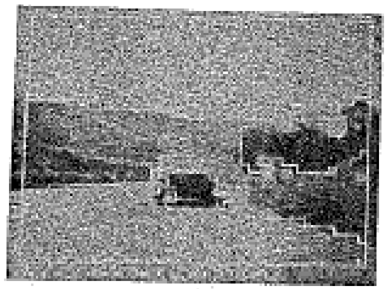
Yakimovsky&Feldman



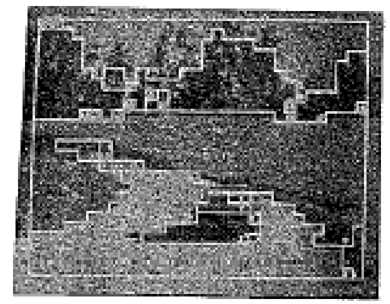
(A-6) Output of region grower based on semantics. (Melting weakest boundary first where boundary strength is computed using the semantic world model).



(B-2) Output of the non-semantic weakest boundary melted first region grower.

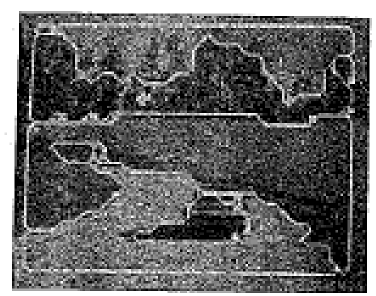


(A-7) Final grouping of regions based on the interpretation assigned to them by the world model. Regions whose meaning was assigned with confidence less than 10 are not mergable. They occur usually on the real boundary between two regions.

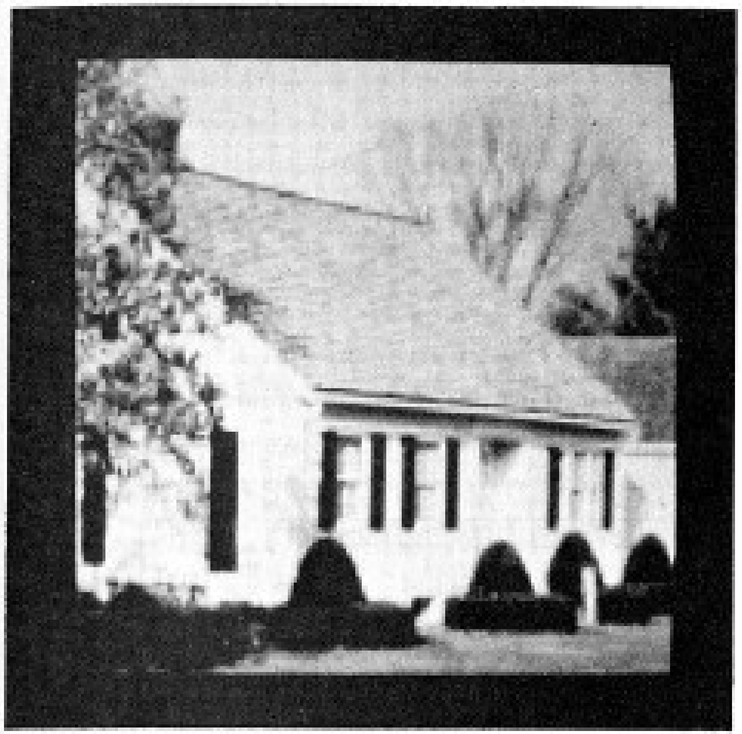


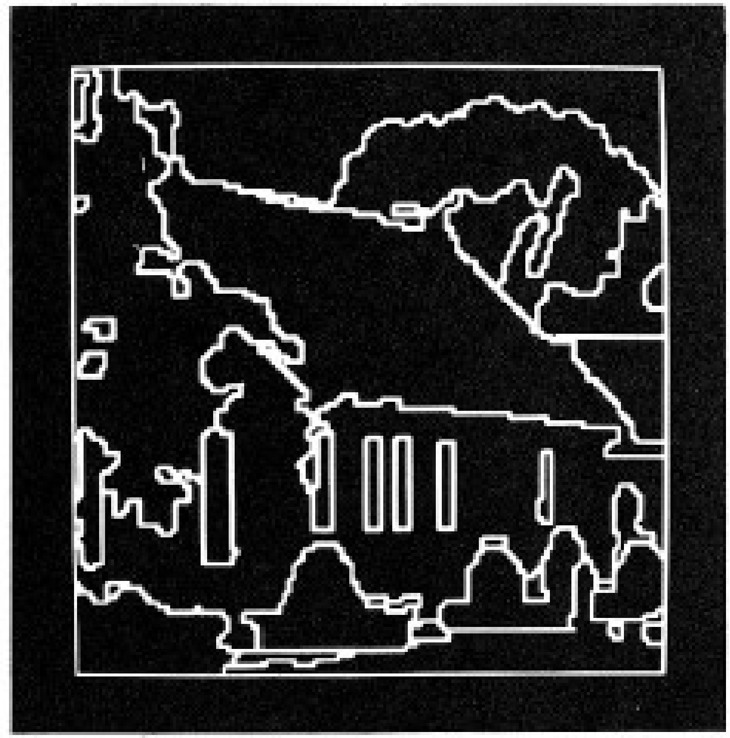
(B-3) Output of the semantic based region grower

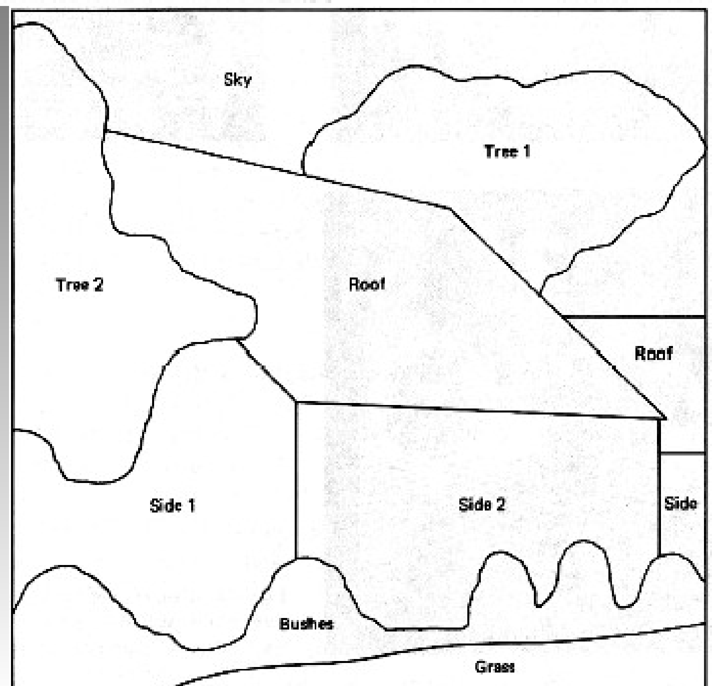
(B-1) Original picture



(B-4) Result of grouping regions by their assigned meaning. Taking only regions which were assigned meaning with confidence over 10 to be mergable



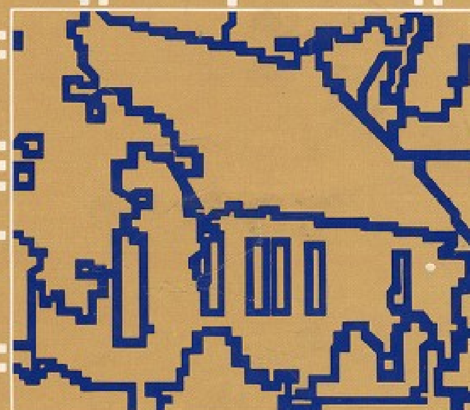




Hanson & Riseman

COMPUTER VISION

DANA H. BALLARD • CHRISTOPHER M. BROWN



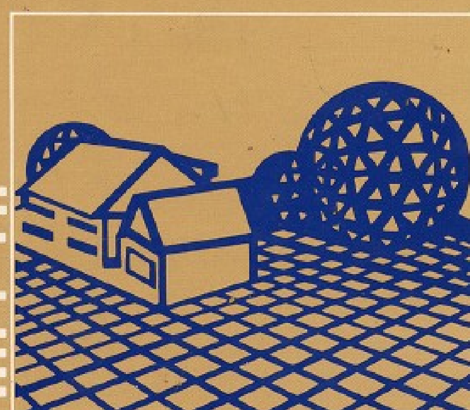
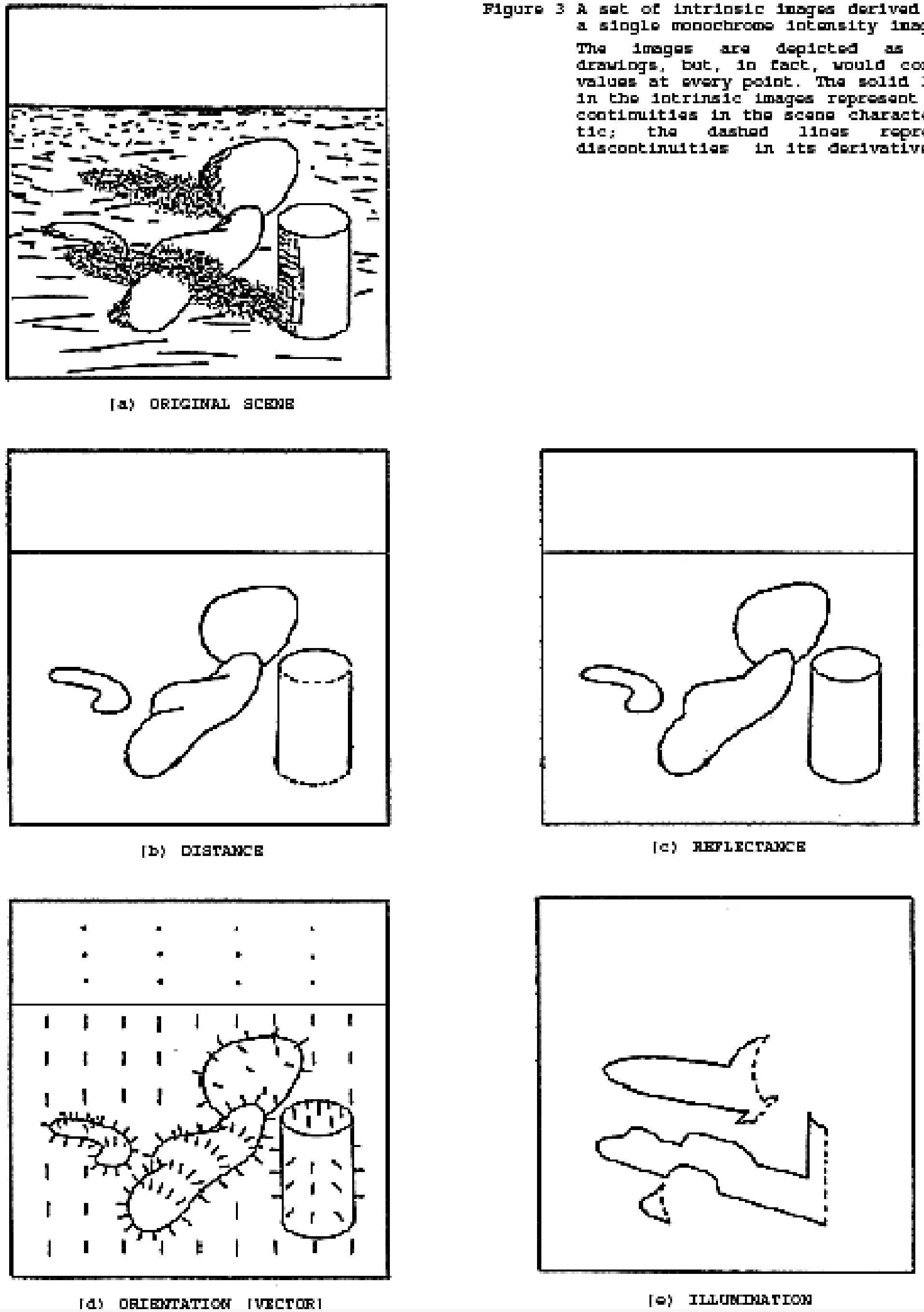


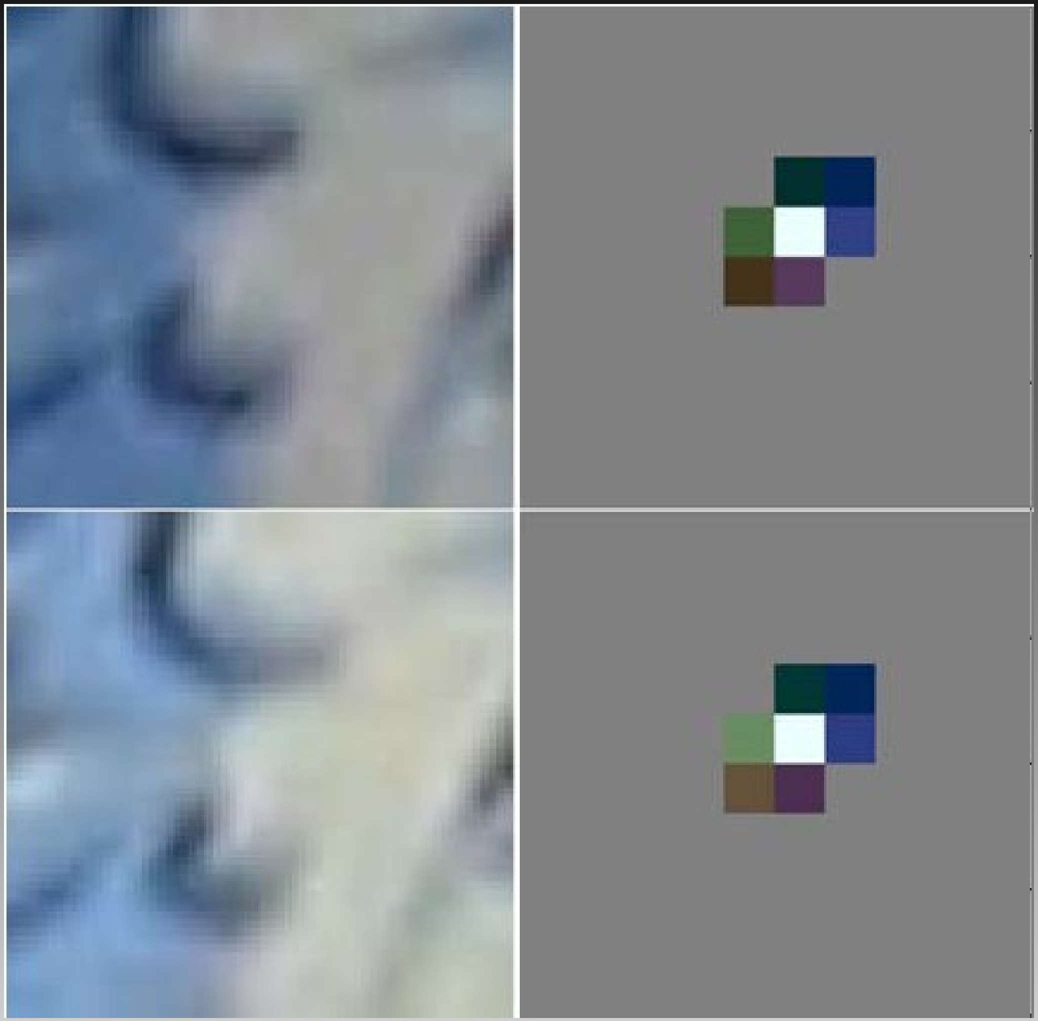
Figure 3 A set of intrinsic images derived from a single monochrome intensity image

The images are depicted as line drawings, but, in fact, would contain values at every point. The solid lines in the intrinsic images represent discontinuities in the scene characteristic; the dashed lines represent discontinuities in its derivative.

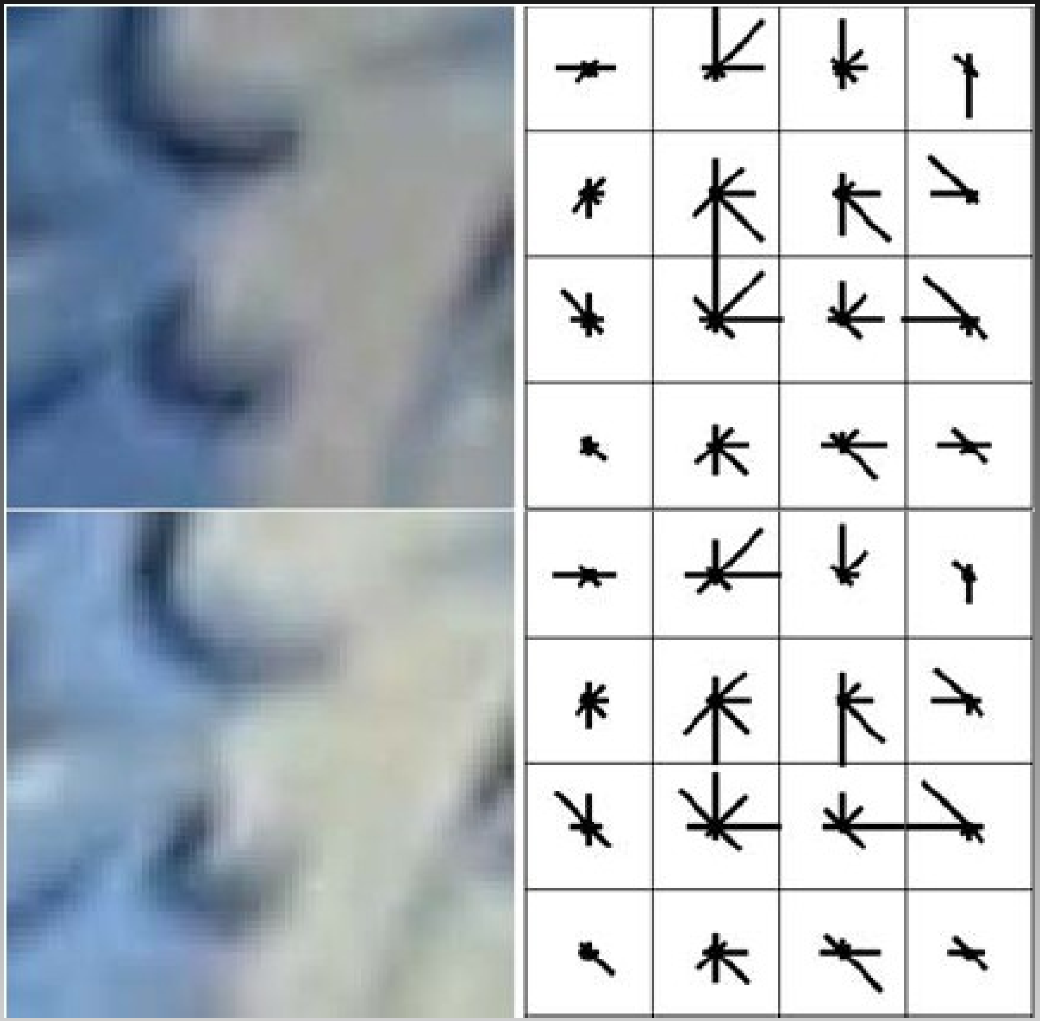


Feature Representations

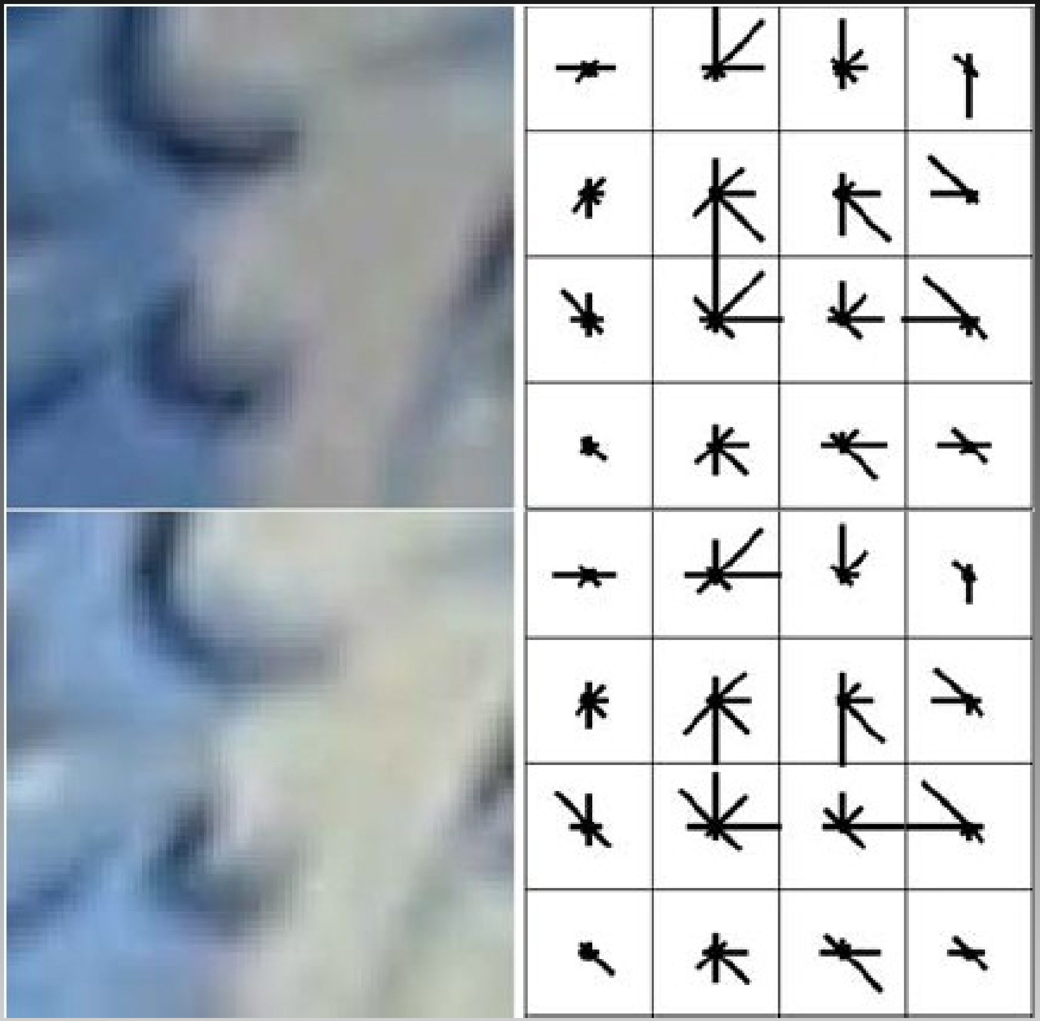
Jean Ponce



Color histograms
(Swain & Ballard'91)

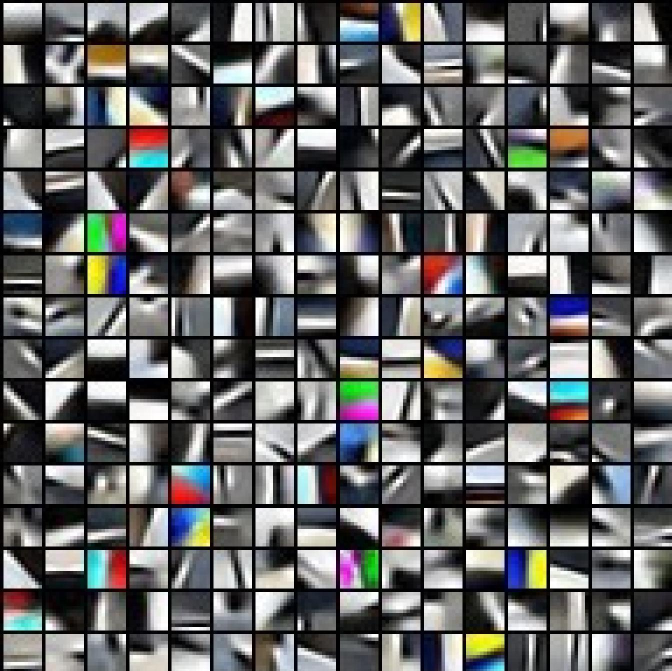


Local jets (Florack'93)
Spin images (J&H'99)
Sift (Lowe'99)
Shape contexts (B&M'95)



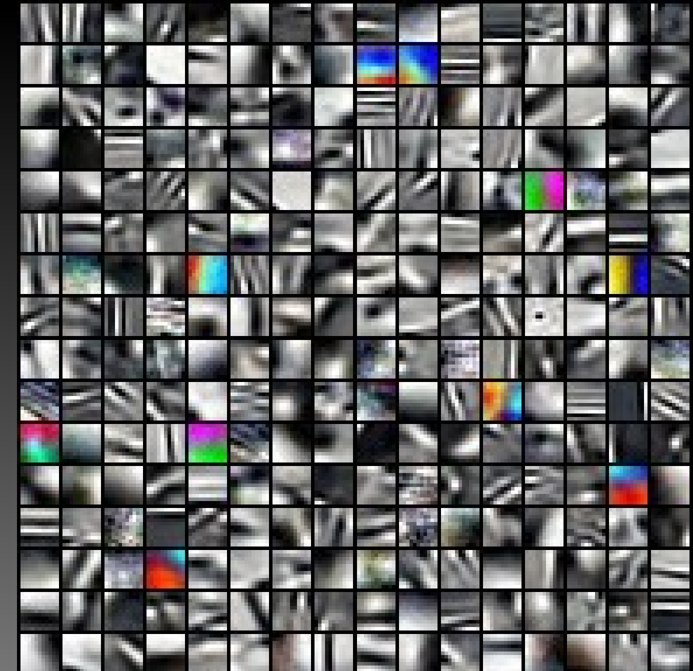
Local jets (Florack'93)
Spin images (J&H'99)
Sift (Lowe'99)
Shape contexts (B&M'95)

Texton histograms (?)
Gist (O&T'05)
Spatial pyramids (LSP'06)
Hog (D&T'06)
Phog (B&Z'07)
Convolutional nets (LC'70)

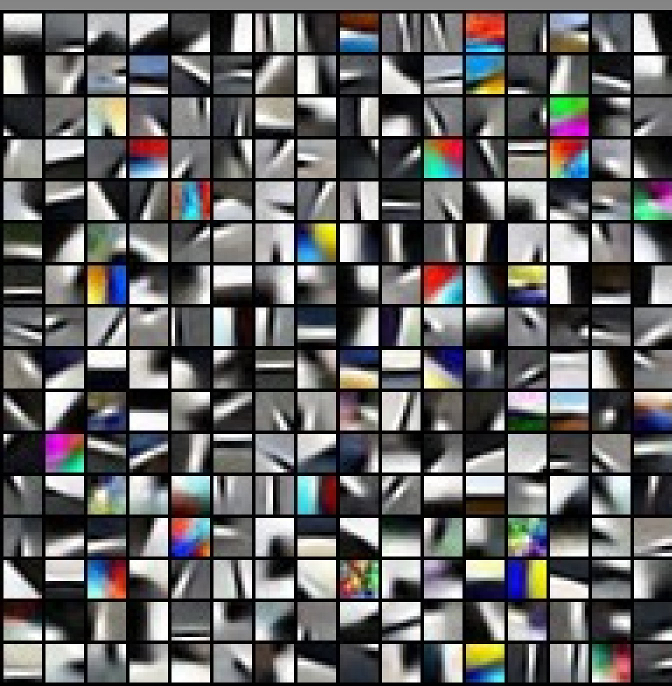


Reconstructive edge dictionary

Reconstructive background dictionary



Mairal et al.'08

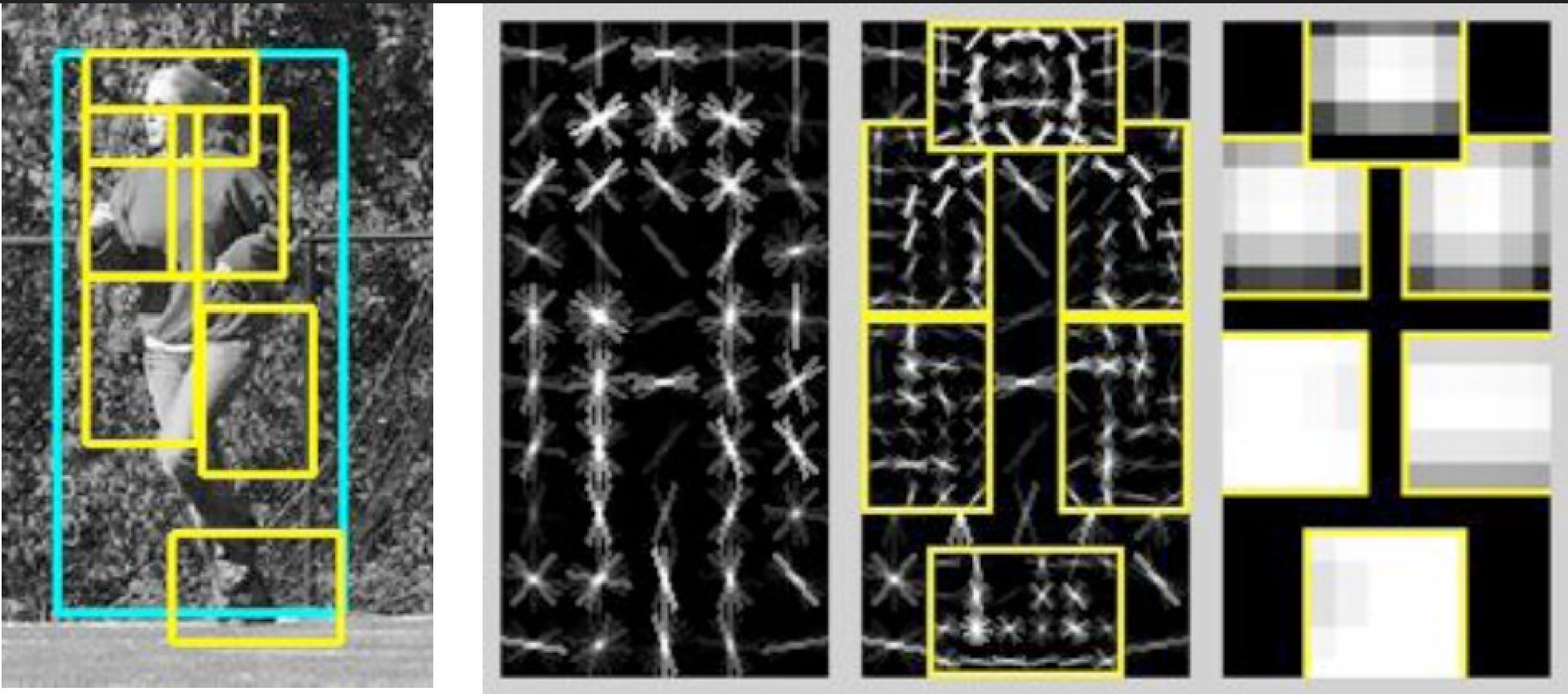


Discriminative edge dictionary

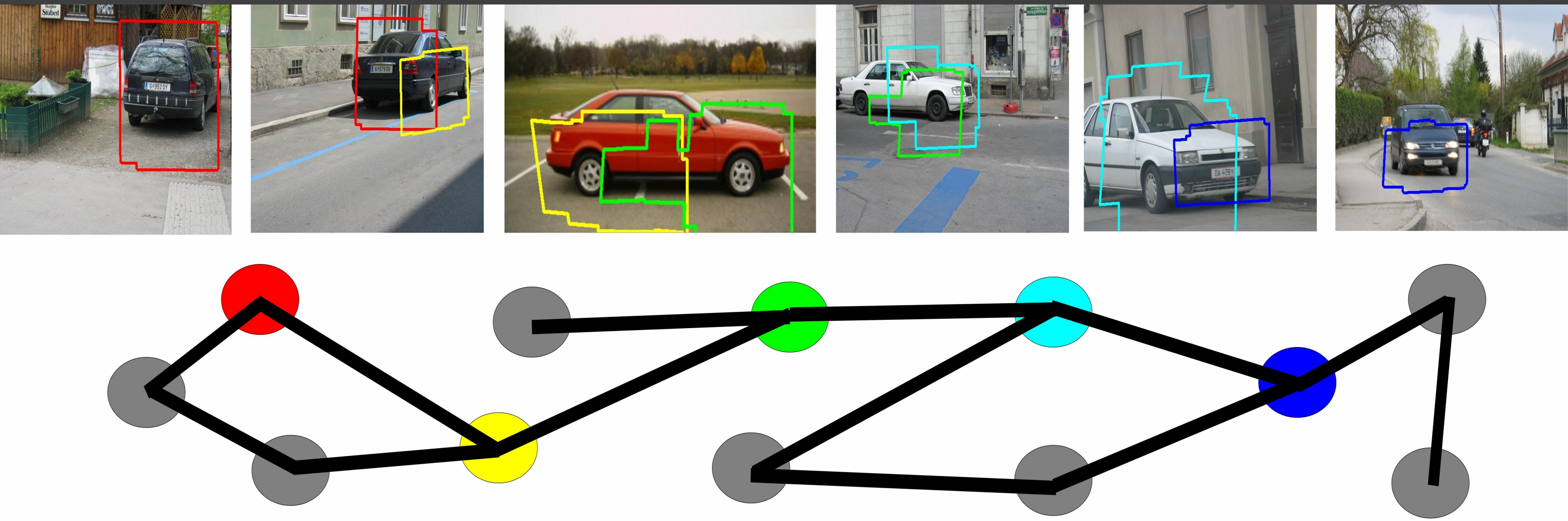
Discriminative background dictionary

Locally orderless structure of images (K&vD'99)

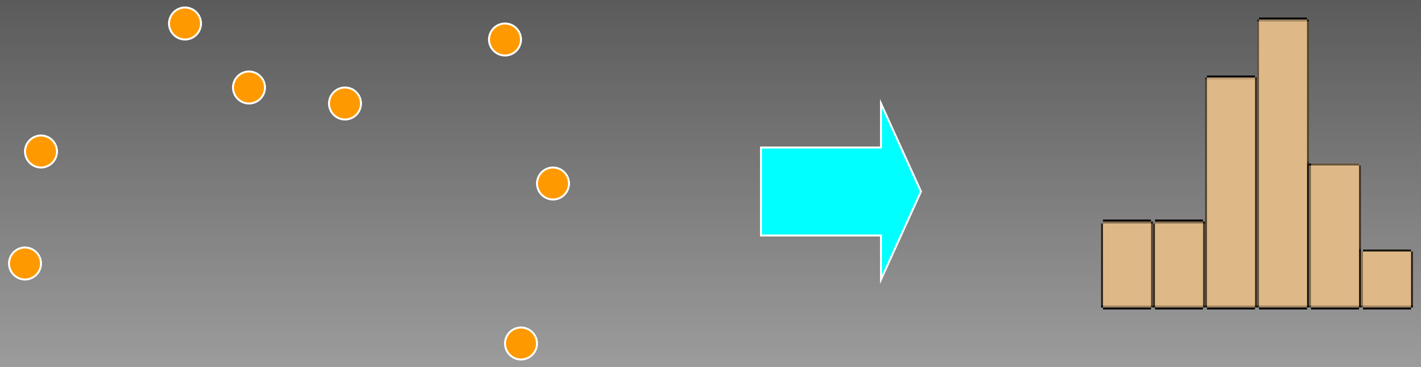
Courtesy of A. Efros



Felzwenszalb, McAllester, Ramanan, 2007

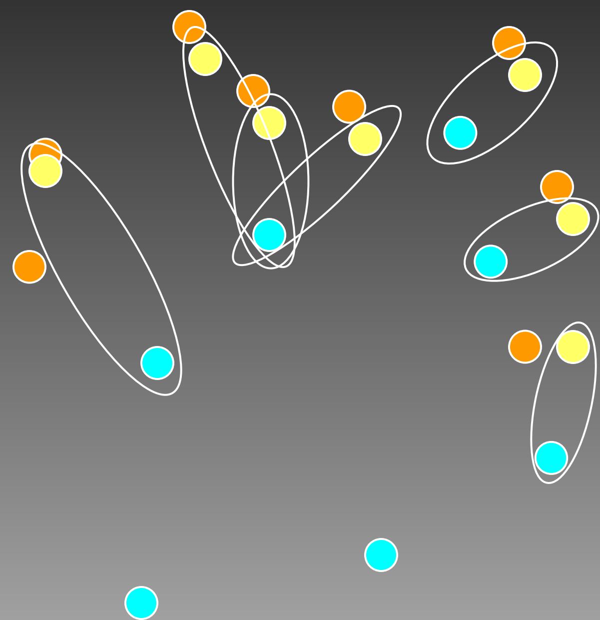


Kushal, Schmid, Ponce, CVPR'07



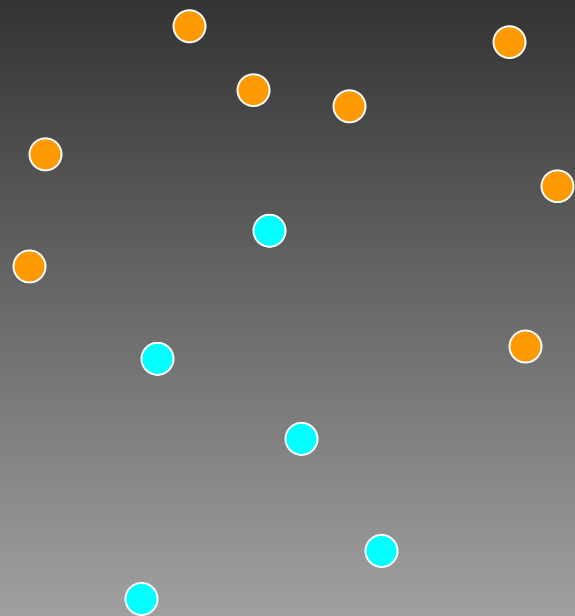
Boiman, Schechtman, Irani, CVPR'08

(73% classification rate on Caltech 101)



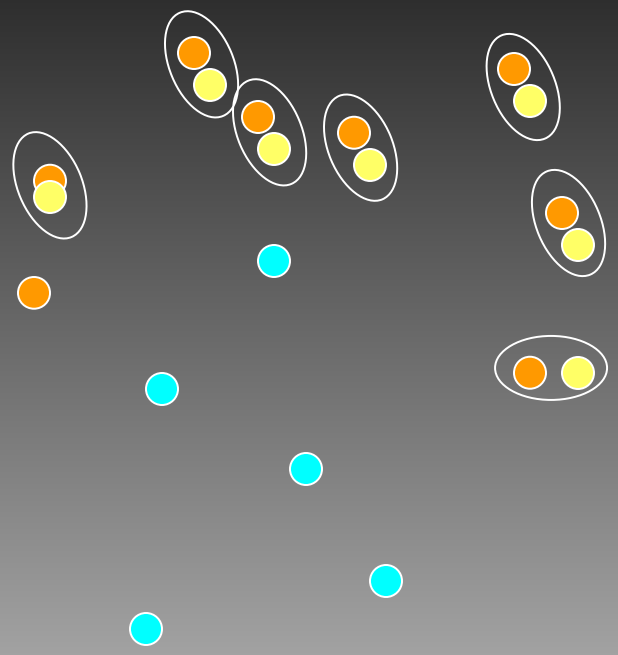
Boiman, Schechtman, Irani, CVPR'08

(73% classification rate on Caltech 101)

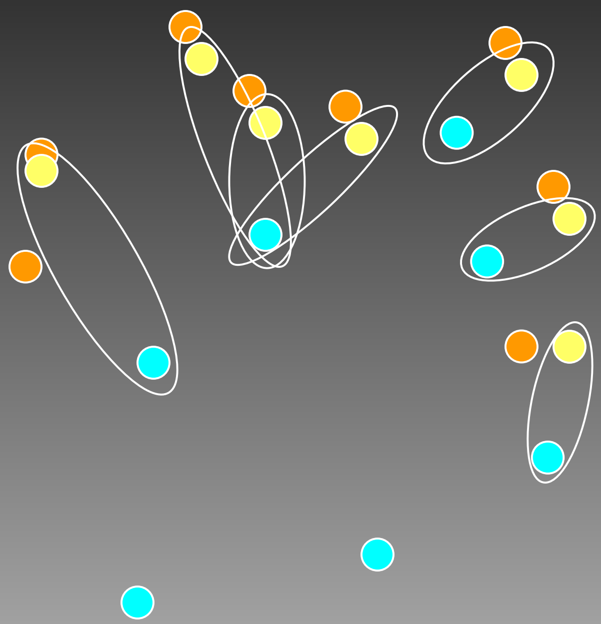


Boiman, Schechtman, Irani, CVPR'08

(73% classification rate on Caltech 101)



Boiman, Schechtman, Irani, CVPR'08
(73% classification rate on Caltech 101)



Essentially the modified Hausdorff distance for object matching of Dubuisson & Jain'95 (see also Farach-Colton & Indyk'99 for ANNs).