Recognition: Past, present, future?

Benozzo Gozzoli, Journey of the Magi, c. 1459

Outline

- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Brief history
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation
  - From vision to action

Discrimination

Use some procedure to attach a label to

- an image; some images; video; range data; lidar data; etc, etc

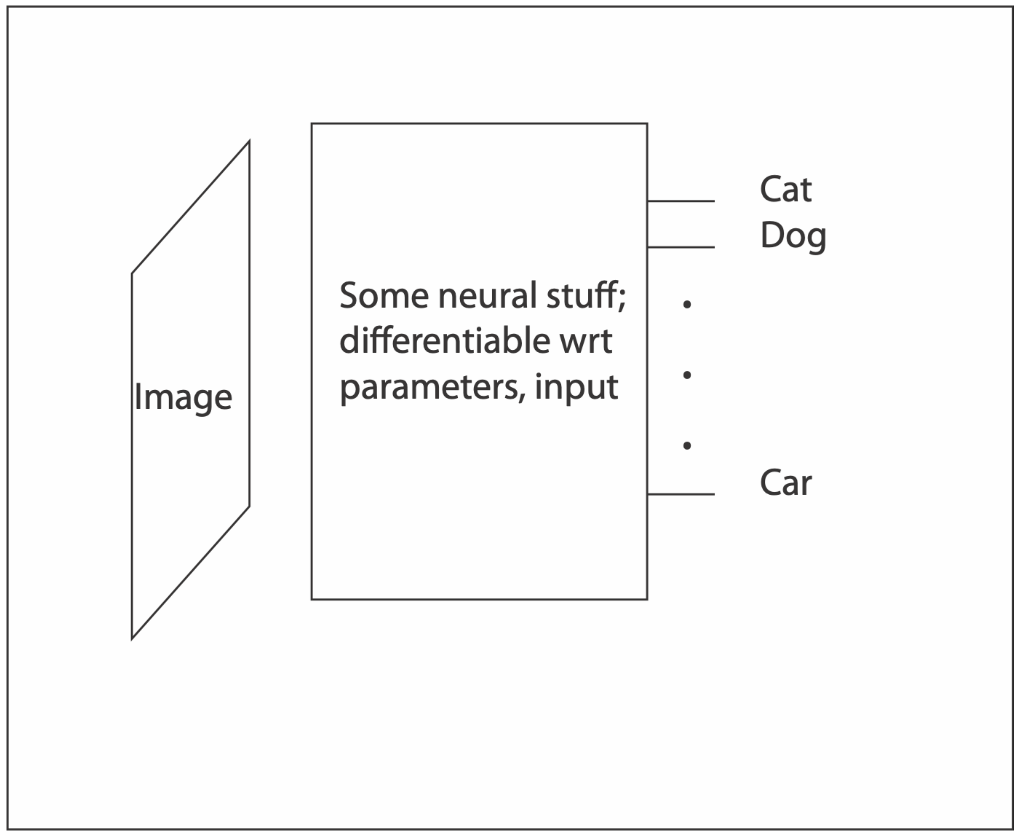
“Label” can be very loosely interpreted

- Name of the main object in the image
- Sentence describing the image
- Direction the car should turn

“Procedure” could be

- learned
- hand-tuned
- determined by physics; the problem; etc
- All three

Typical picture of image classification



Key ideas

Goal:

- Adjust classifier so that it accurately classifies *UNSEEN* data
- ie on *unseen* data, the predicted labels have low loss

Loss

- Cost of using predicted labels instead of true
- Eg error rate; quality of sentences; number of accidents

Procedure:

- Adjust so that it
  - classifies training data well
  - generalizes
- regularization term, either explicit or implicit

Evaluation:

- Use held out data to check accuracy on *UNSEEN* data

Recognition: What type of content?

Object *instance* recognition

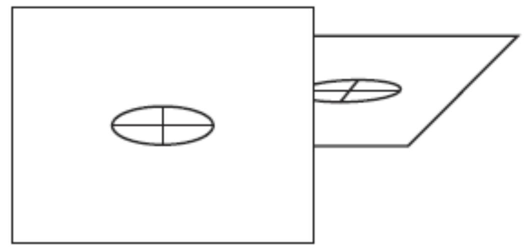
Object *category* recognition

Texture recognition

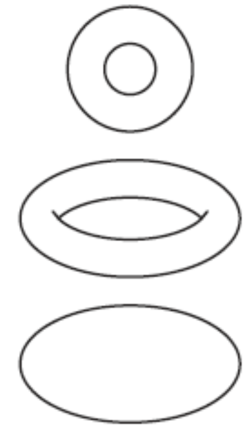
Scene recognition

- Beyond still images: video, RGBD data, point clouds, multimodal data…

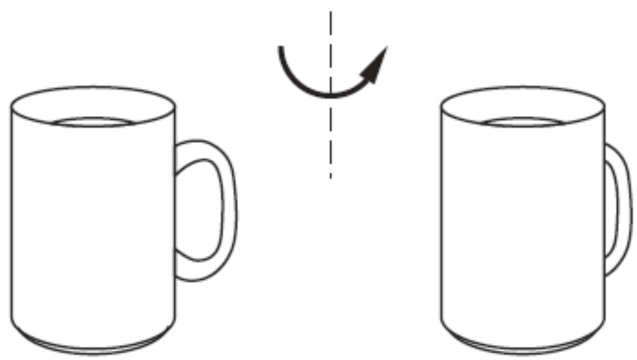
Standard difficulties



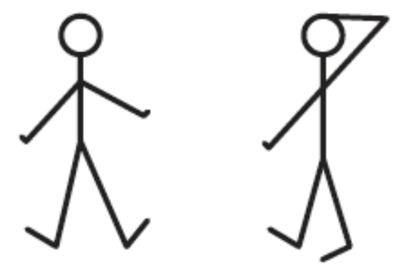
Foreshortening



Aspect



Occlusion



Deformation

Classification vs detection vs segmentation

Classification:

- there is an X in this image
  – what

Detection:

- there is an X HERE in this image
  – what AND where (variants depending on different notions of where)

Semantic segmentation:

- These pixels are sky, these road, these person, etc

Semantic instance segmentation:

- Semantic +
- These pixels are person 1, these person 2, these person 3, etc

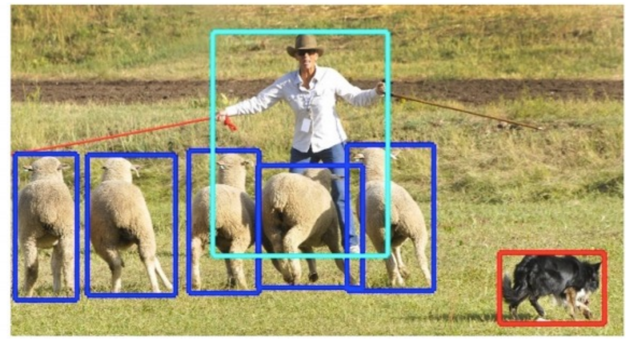
Recognition: What type of output?

- Classification: labels
- Regression: continuous values
- Dense prediction: an output at every image location
- Structured prediction: combinatorial structures
- Natural language
- Etc.

Recognition: What type of output?

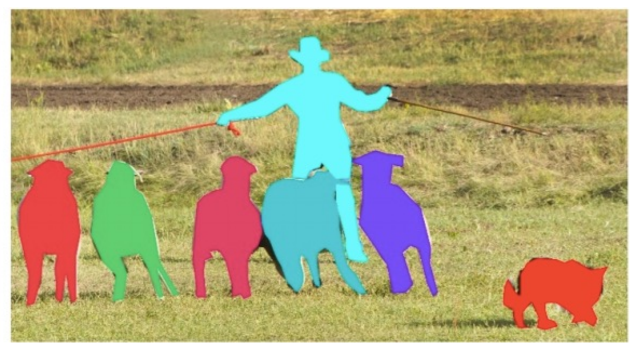
Image classification

Object detection



Semantic segmentation

Instance segmentation



- And beyond: depth/3D structure prediction, image description, etc.

Classification vs detection vs segmentation

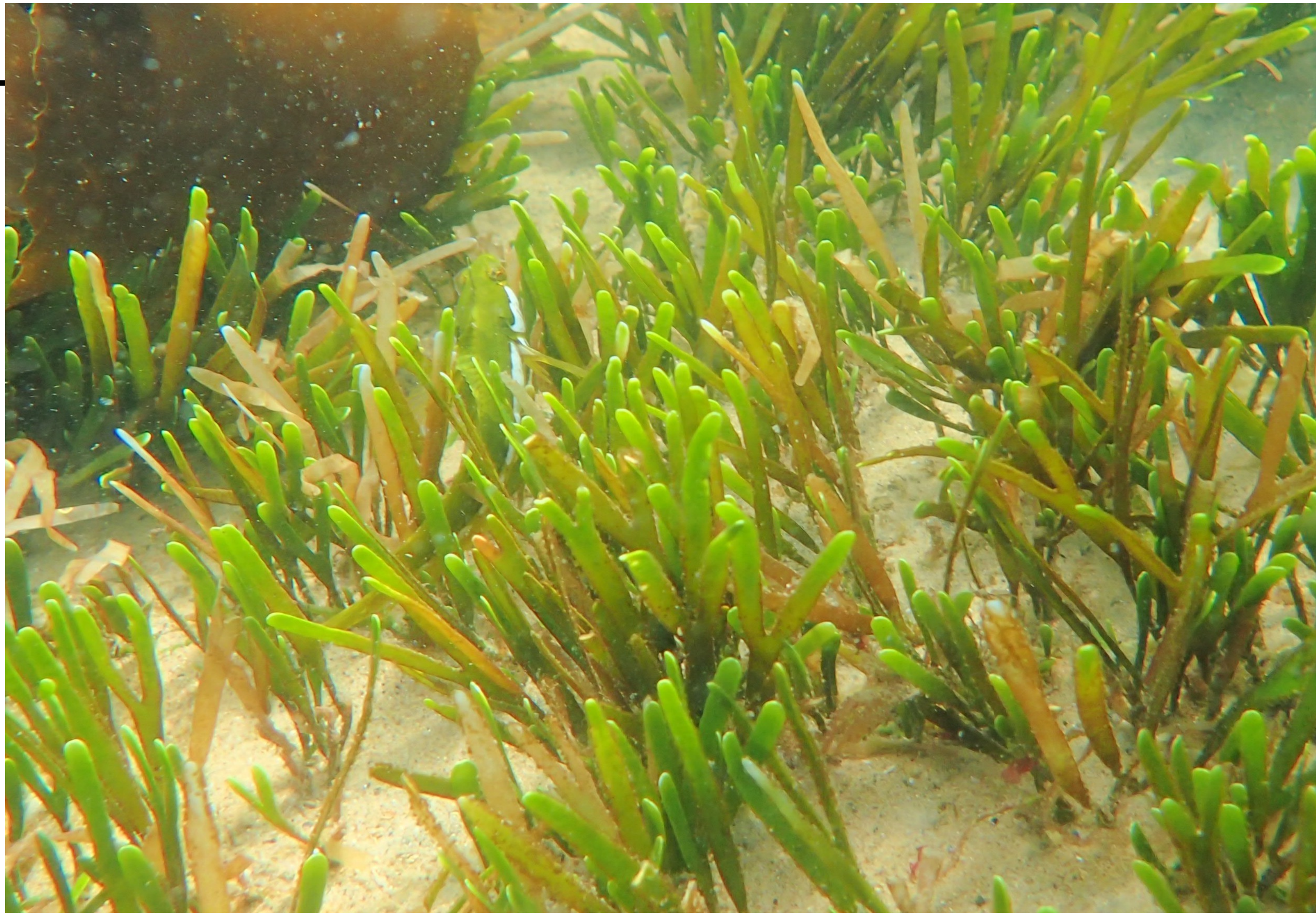
Key issues

- How to classify
- how to specify where
- relationship between what and where
  - efficiency, etc
  - Predict where first; or what first; or both in parallel?
- evaluation
  - Evaluating detection is surprisingly fiddly

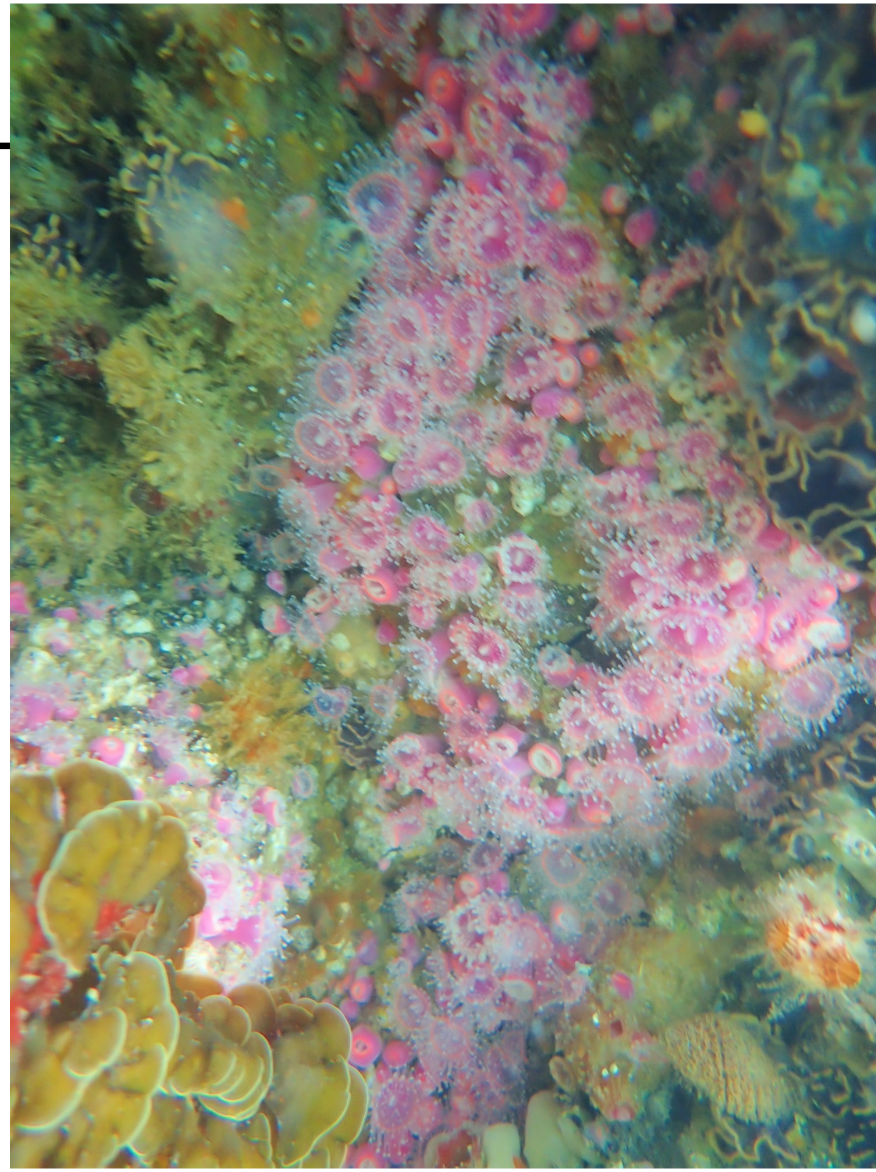
It can be hard to know what to report

Is t

It can be hard to find things



You may not know the right label



Our current state

We do wonderful things when labels are available

What we do poorly

- Ambiguous prediction
- Descriptions without labels
- Narrative and models

An extreme example

People do what they do for reasons

- these are sometimes about the physical world
- and sometimes because they have internal goals, etc

What we need to understand this

U Selection

- (the cart and people are worth talking about; the buildings are not)

U Attributes

U=under attack

- try to describe unfamiliar things in familiar terms

Geometric representations that generalize

? - eg carts can rock on axles

Situating objects in space with respect to one another

U - contact; potential; etc

Predicting who/what can do what

? - so we notice when they don't

Some form of narrative structure

? - in terms of goals, intentions, etc.

- associating potential outcomes with objects

RapidABC data

Questions you can't answer

How many RJ11 jacks in the wall near the camera?

Questions you can answer

About feelings

- How is the mother feeling?
- How is the interviewer feeling?
- How is the child feeling?

Because

- it tells you what might happen next

Questions you can answer

About feelings

- How is the mother feeling?
- How is the interviewer feeling?
- How is the child feeling?

Because

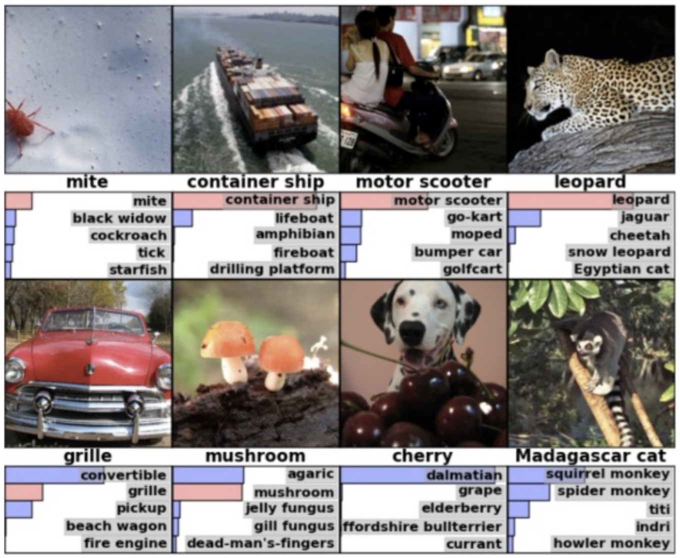
- it tells you what might happen next

Questions you can’t answer

How many RJ11 jacks in the wall near the camera?

Image Classification

ILSVRC



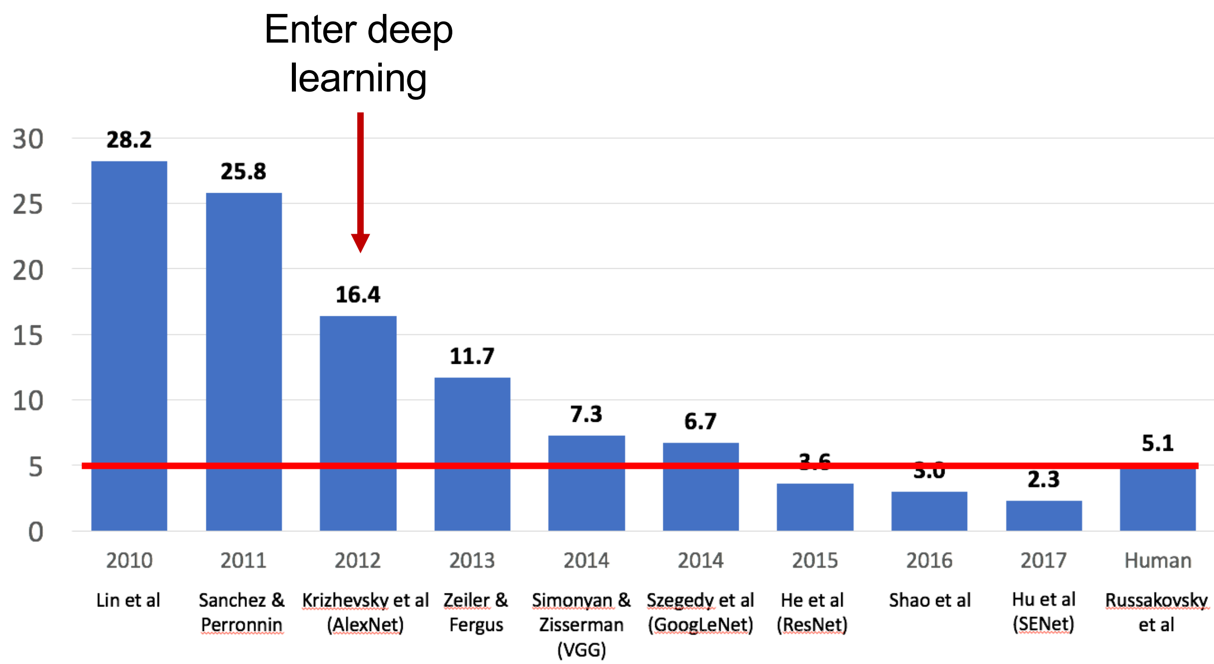


Figure source

What should recognition say?

Report names of all object categories (?!?)

- but we might not have names
- and some might not be important

Make useful reports about what's going on

- what is going to happen?
- how will it affect me?
- who's important?

Do categories exist?

- allow generalization
  - future behavior; non-visual properties of activities

A belief space about recognition

Platonism?

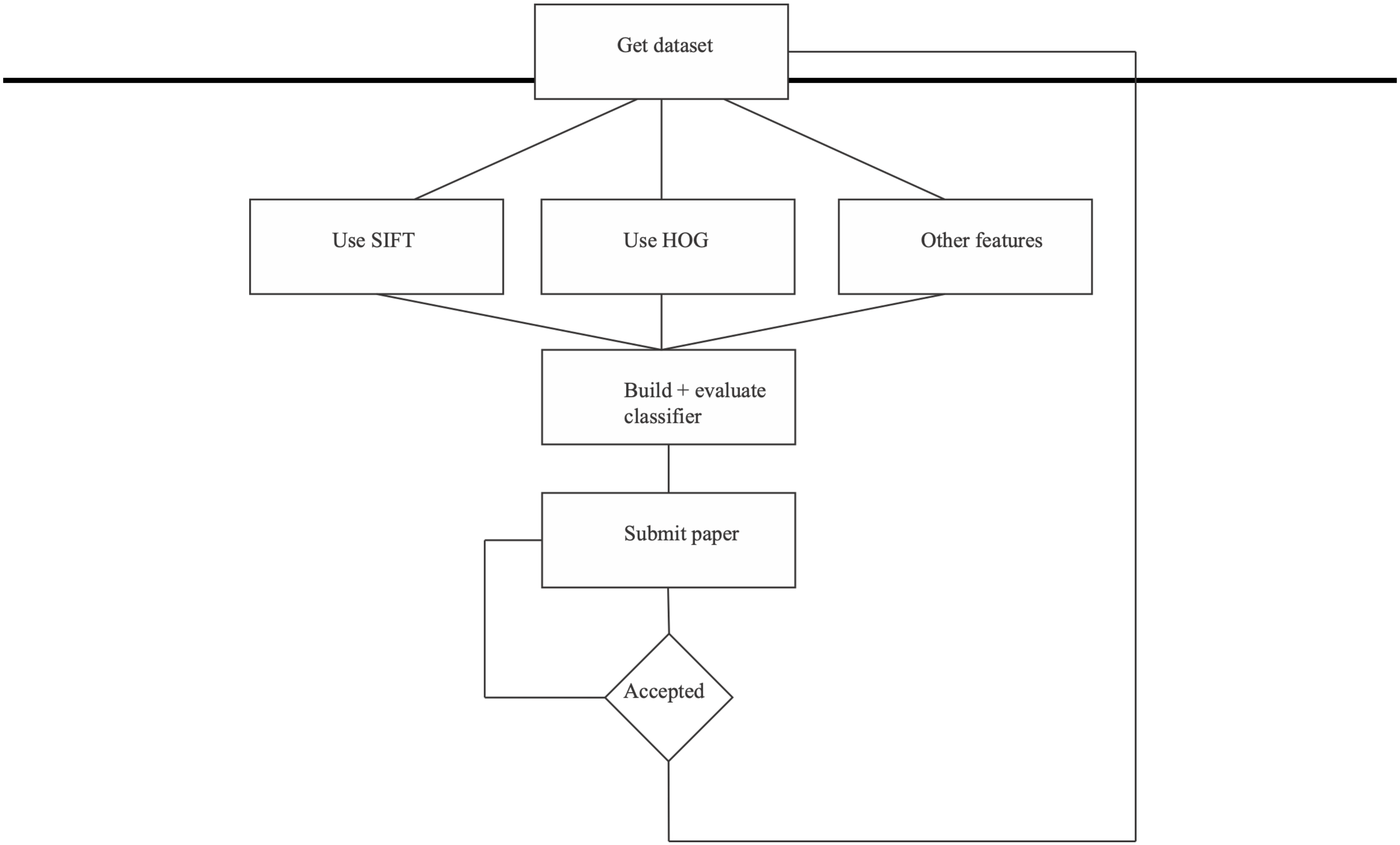
Object categories are fixed and known

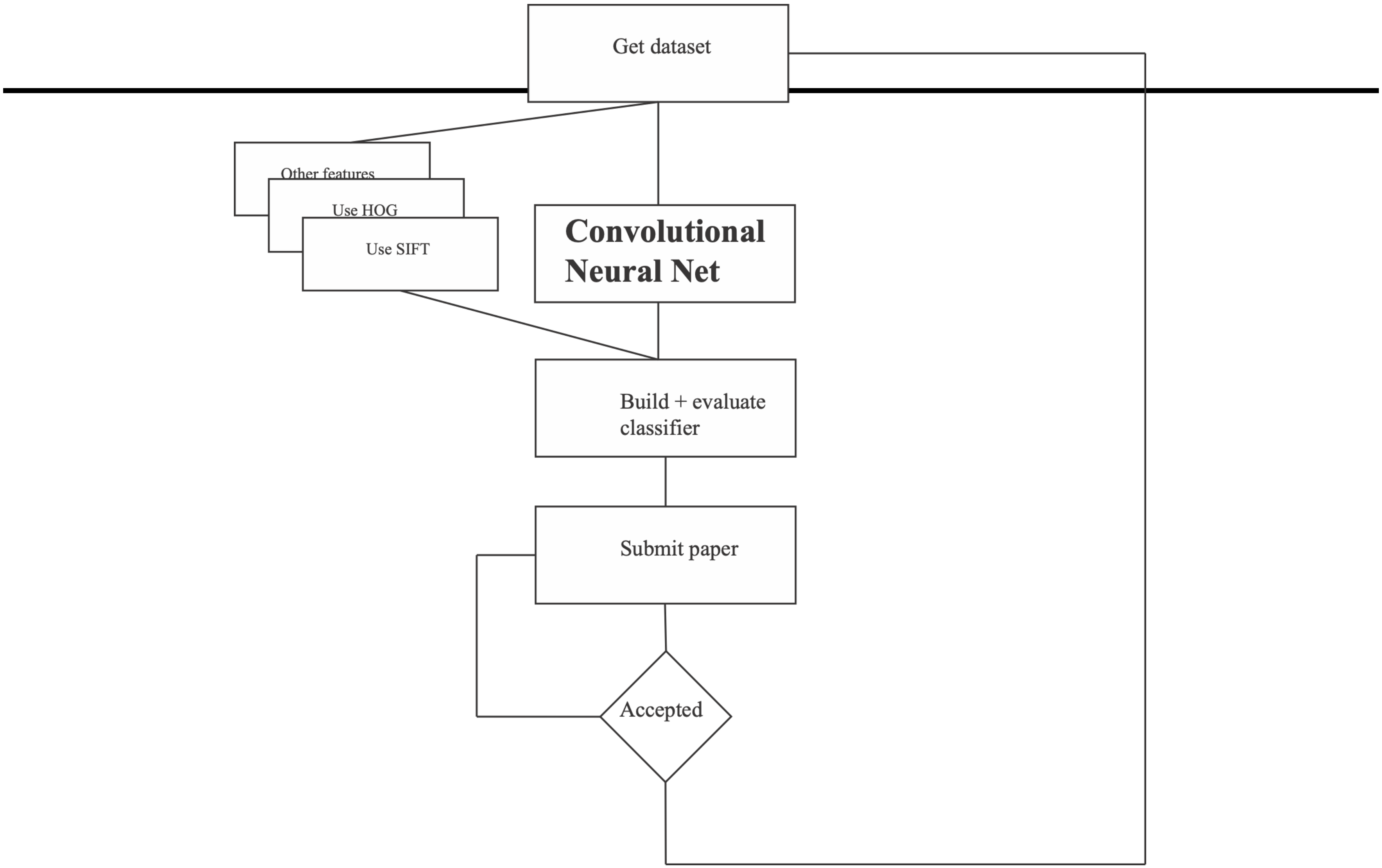
- Each instance belongs to one category of k

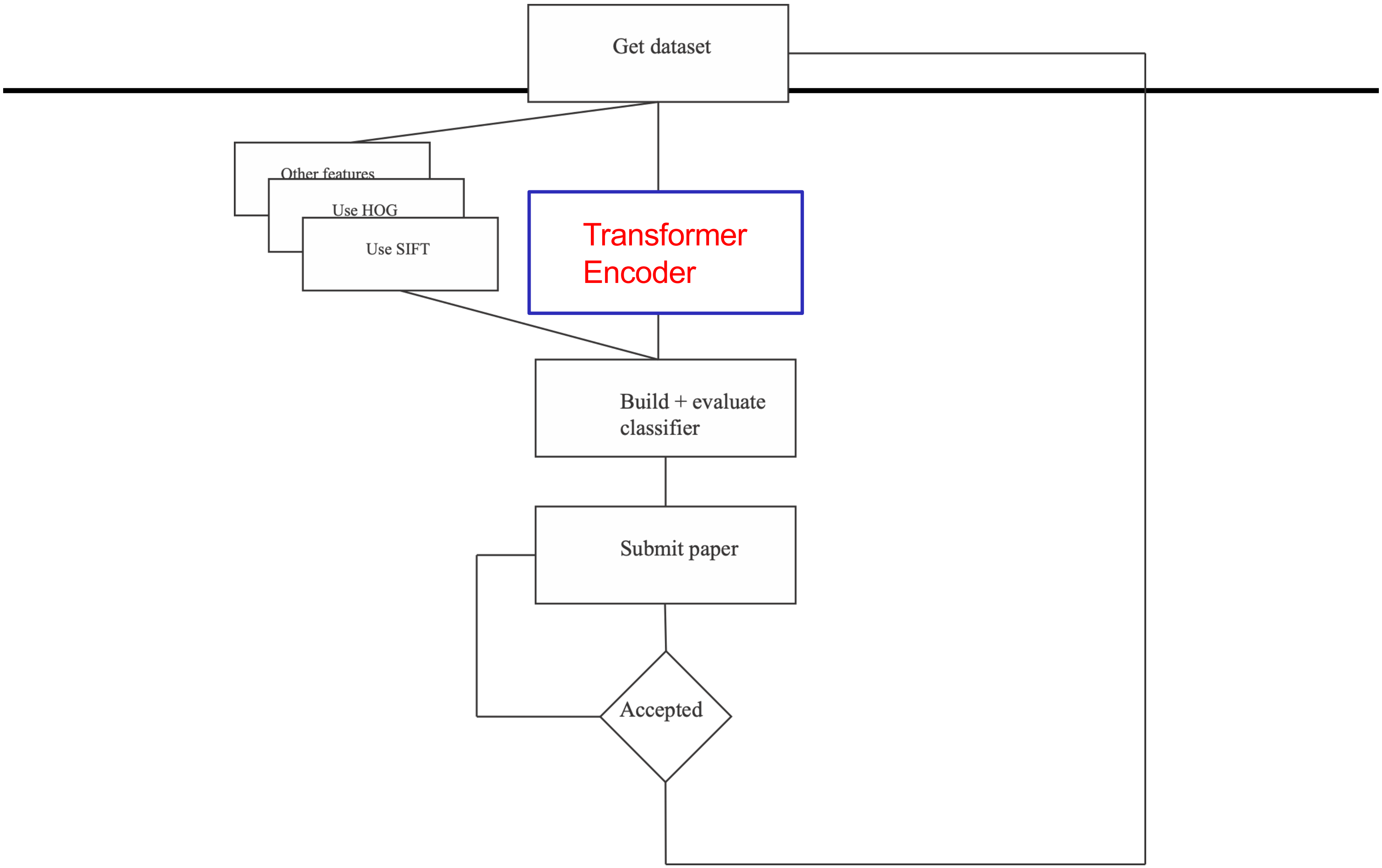
Good training data for categories is available

Object recognition=k-way classification

Detection = lots of classification


Get dataset
Use SIFT
Use HOG
Other features
Build + evaluate classifier
Submit paper
Accepted


Get dataset
Other features
Use HOG
Use SIFT
Convolutional Neural Net
Build + evaluate classifier
Submit paper
Accepted



Get dataset
Other features
Use HOG
Use SIFT
Transformer Encoder
Build + evaluate classifier
Submit paper
Accepted

A belief space about recognition

Platonism?

Object categories are fixed and known

- Each instance belongs to one category of k

Good training data for categories is available

Obvious nonsense

Obvious nonsense

Object recognition=k-way classification

Obvious nonsense

Detection = lots of classification

Are these monkeys?

What does recognition do?

Lists object names

Lists object descriptions

Evokes emotional states

- but what do we do about this?

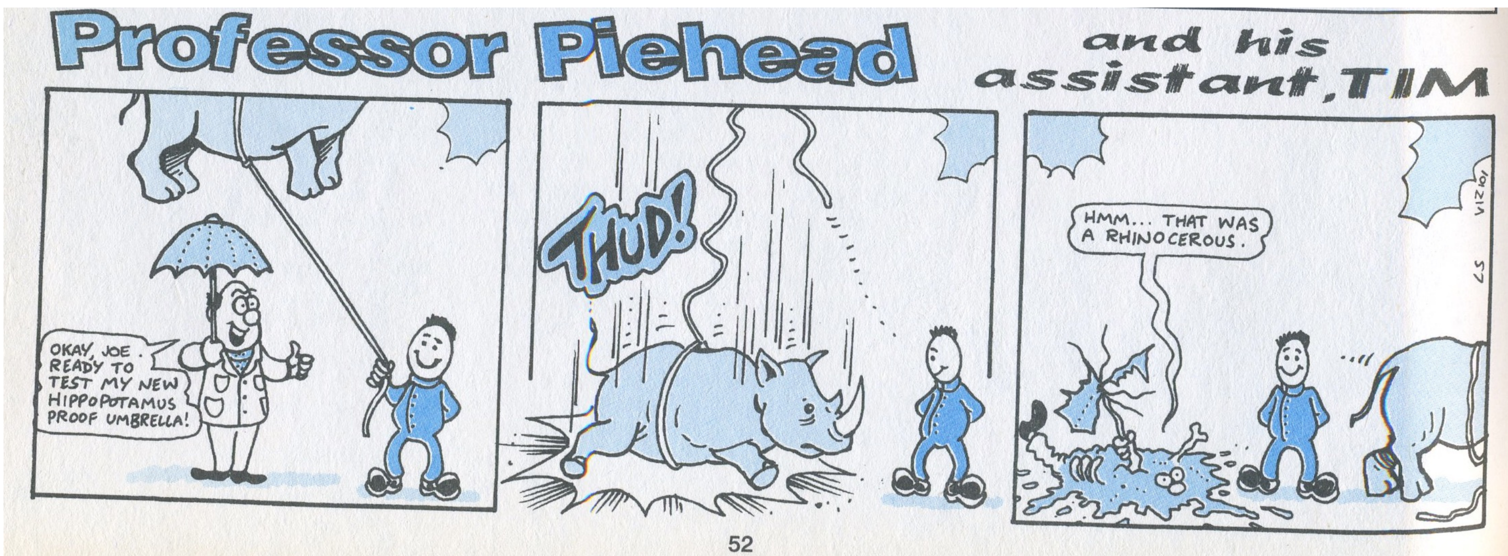
Exposes possible futures

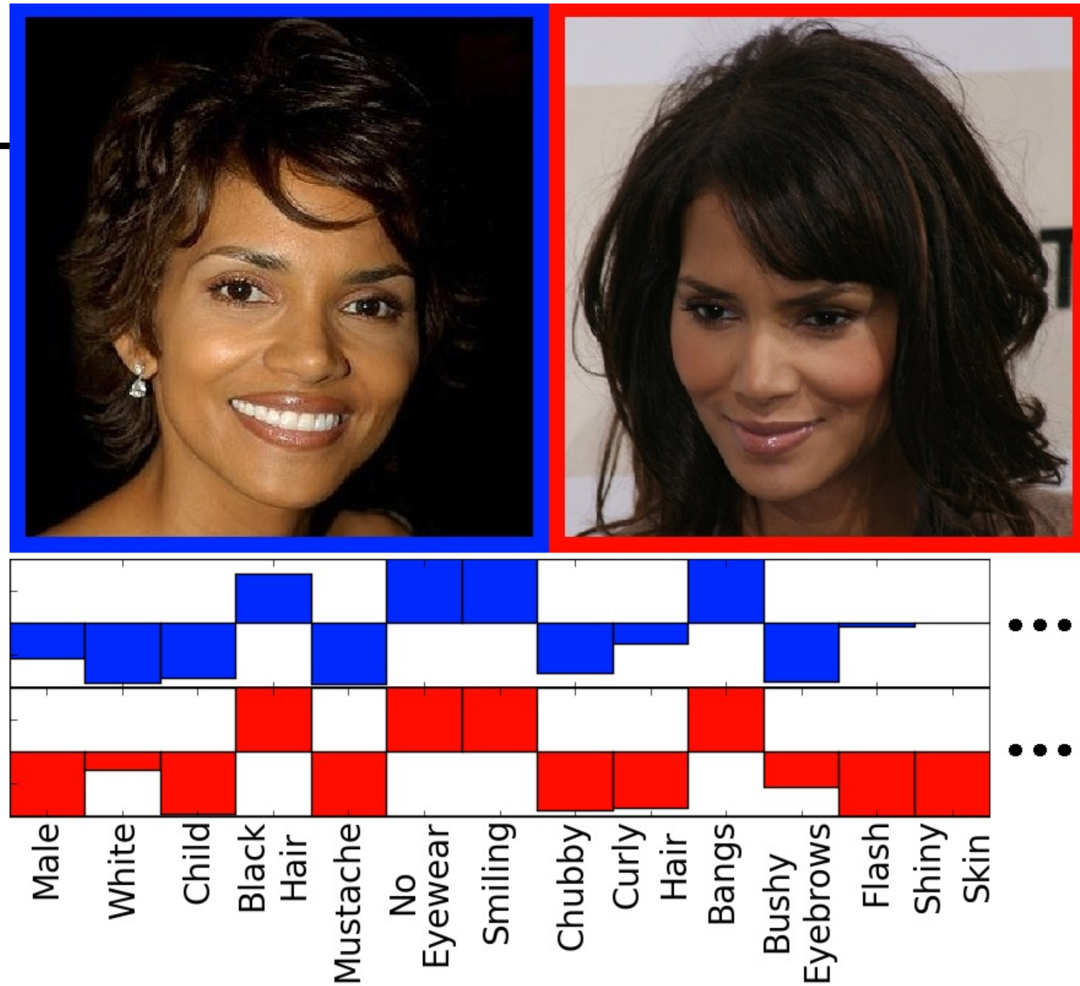
- What could happen
- Where you could go
- Who could move close to you
- What could be useful for

We should think about potential, rather than just or as well as, actual

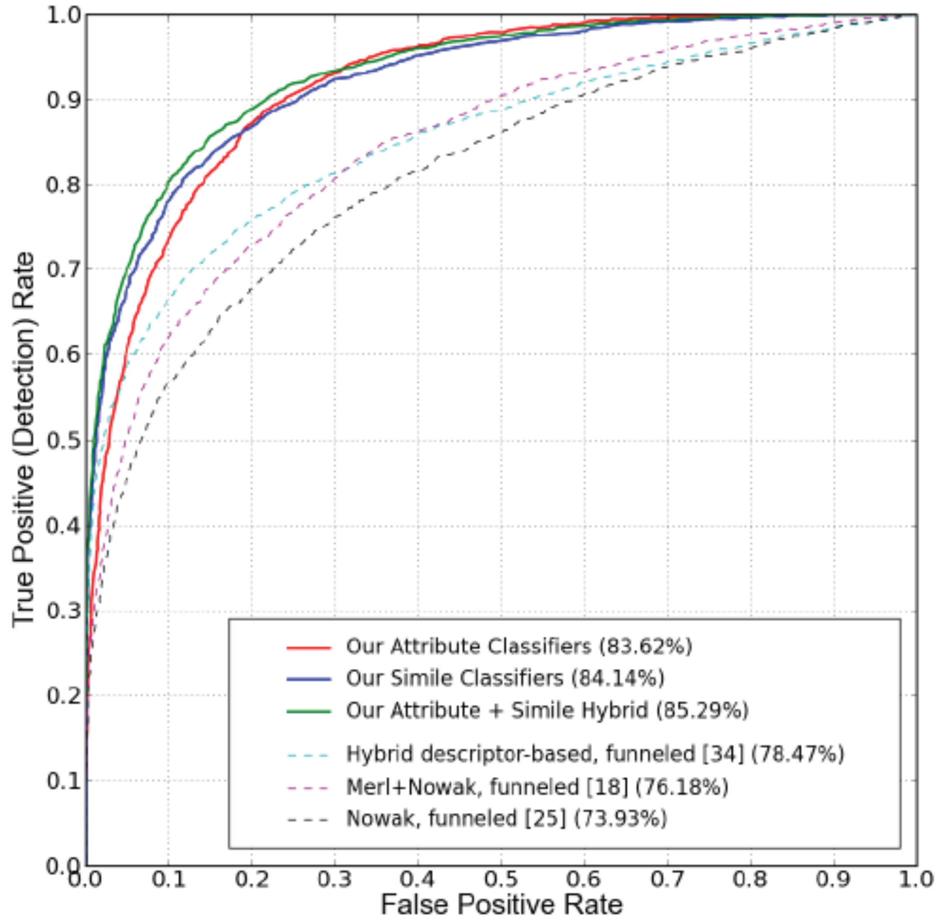
What is an object like?

Professor Piehead



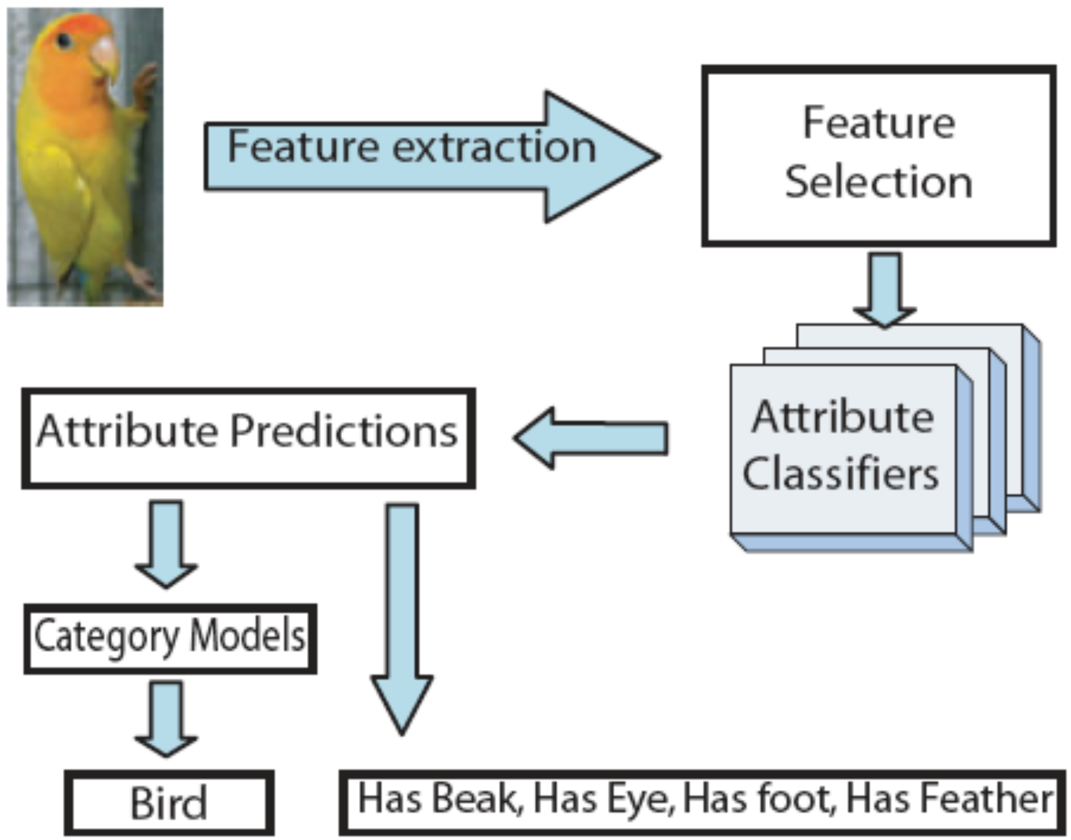


"Attribute and Simile Classifiers for Face Verification," ICCV 2009. (N. Kumar, A. Berg, P. Belhumeur, S. K. Nayar)



True Positive (Detection) Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Our Attribute Classifiers (83.62%)
Our Simile Classifiers (84.14%)
Our Attribute + Simile Hybrid (85.29%)
Hybrid descriptor-based, funneled [34] (78.47%)
Merl+Nowak, funneled [18] (76.18%)
Nowak, funneled [25] (73.93%)

General architecture



A.Farhadi, I. Endres, D. Hoiem, D.A. Forsyth, “Describing objects by their attributes”, CVPR 2009

Regression

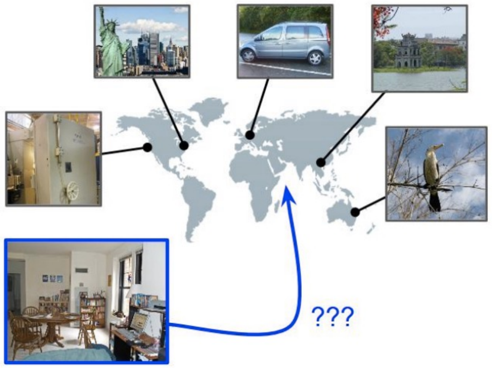
Date prediction

Vittayakorn et al. (2017)

Image colorization

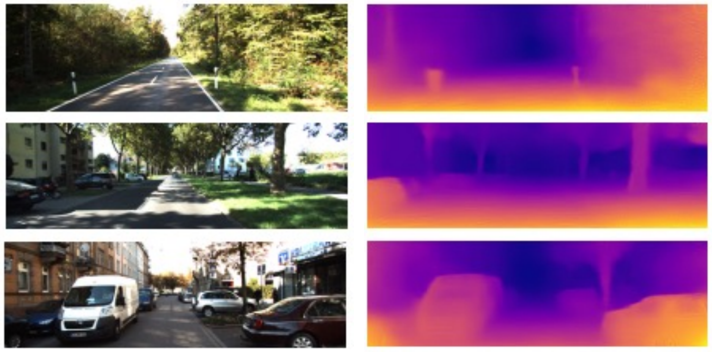
Zhang et al. (2016)

Location prediction



Vo et al. (2017)

Depth prediction



Wang et al. (2017)

Dense and structured prediction

Bounding box prediction,
dense prediction

Keypoint prediction

K. He, G. Gkioxari, P. Dollar, and R. Girshick, [Mask R-CNN](), ICCV 2017

Natural language prediction

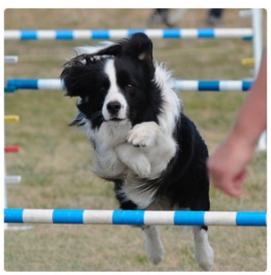
Image captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

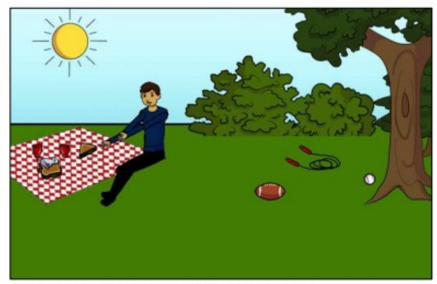
"young girl in pink shirt is swinging on swing."

A. Karpathy, L. Fei-Fei. [Deep Visual-Semantic Alignments for Generating Image Descriptions](). CVPR 2015

Visual question answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

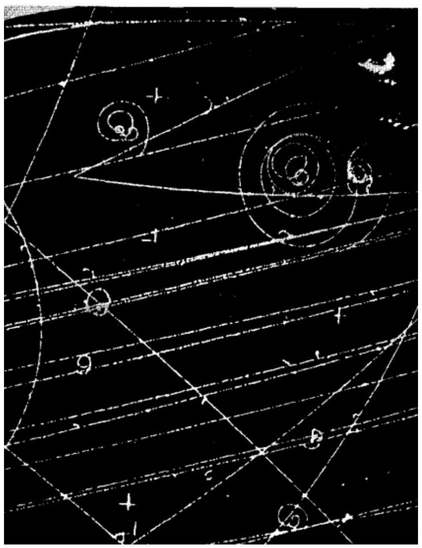
Does it appear to be rainy?
Does this person have 20/20 vision?

S. Antol et al. VQA: [Visual question answering](). ICCV 2015

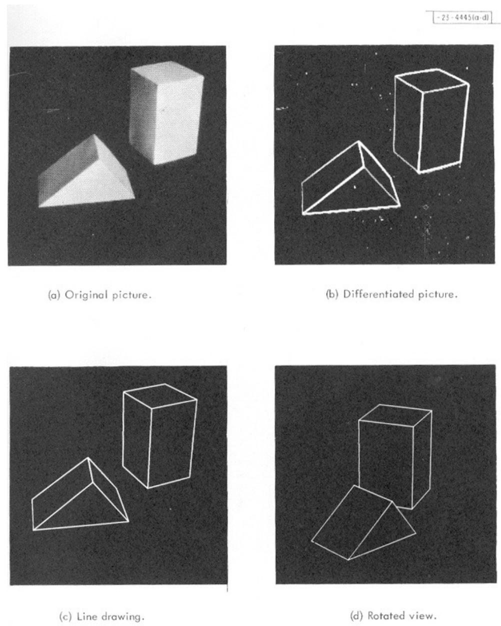
Outline

- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Brief history
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation
  - From vision to action

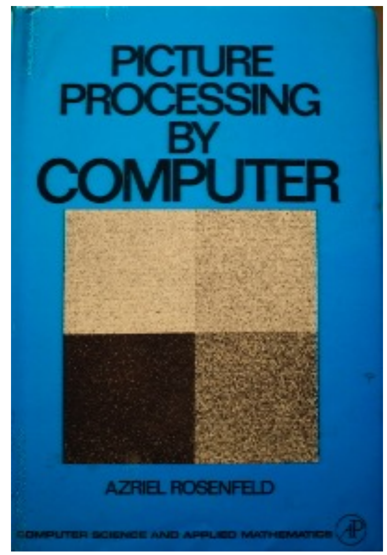
Recall: Origins of computer vision



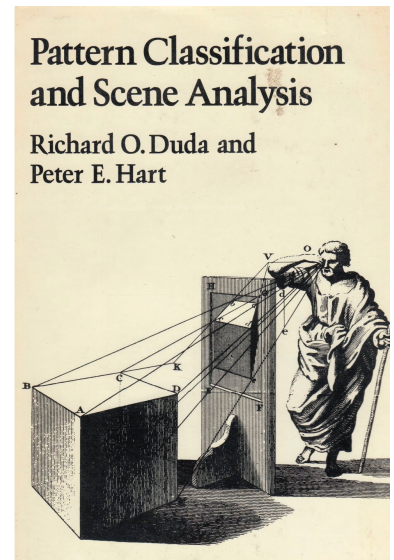
Hough, 1959



Roberts, 1963



Rosenfeld, 1969



Duda & Hart, 1972

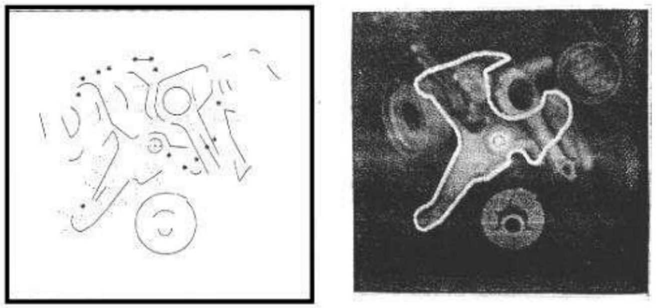
Idea: geometric alignment

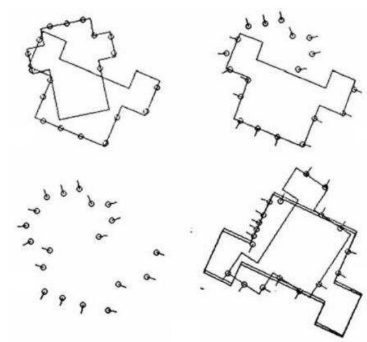
Imagine you have a set of geometric models

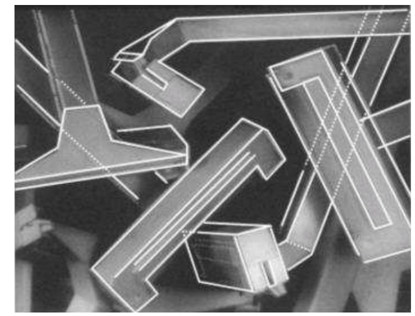
To detect objects in an image:

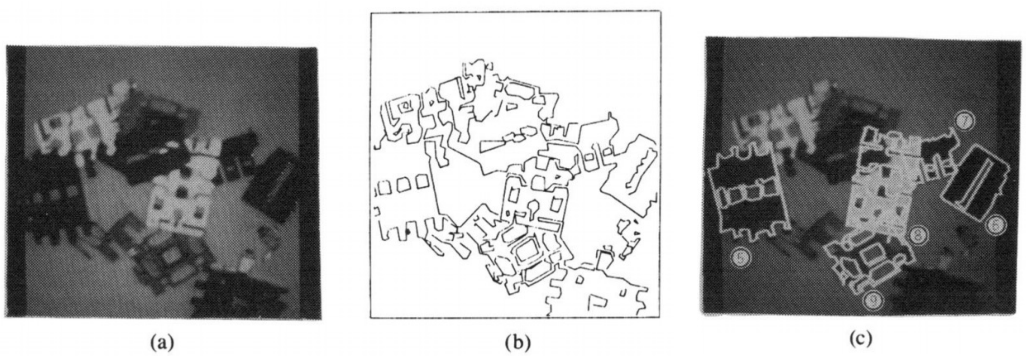
- Repeat:
  - Find image features (edges, corners, etc)
  - Hypothesize a correspondence to model features
  - Compute camera intrinsics
  - Project model into image and check for validation

History of recognition: Geometric alignment

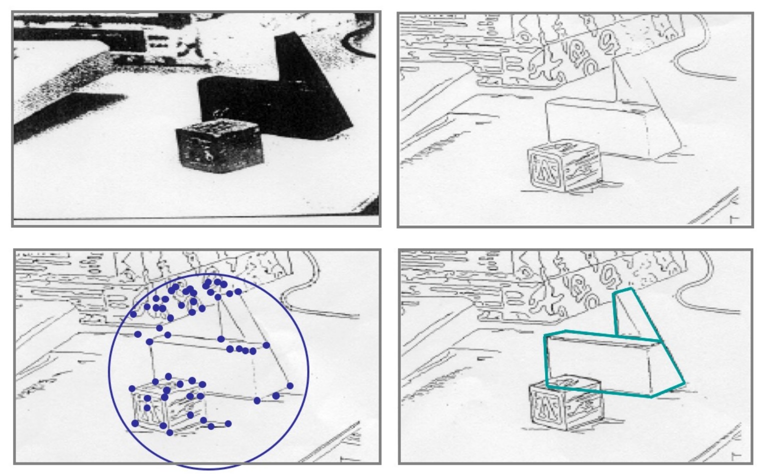

Perkins (1978)


Grimson & Lozano-Perez (1984)


Lowe (1985)



Ayache & Faugeras (1986)


Huttenlocher & Ullman (1987)

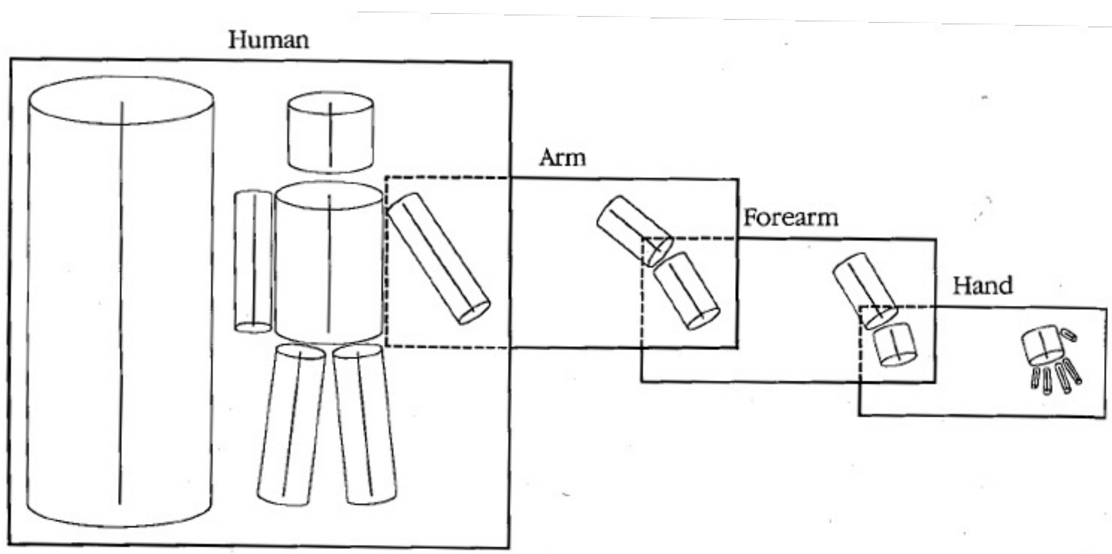
Idea: part hierarchies

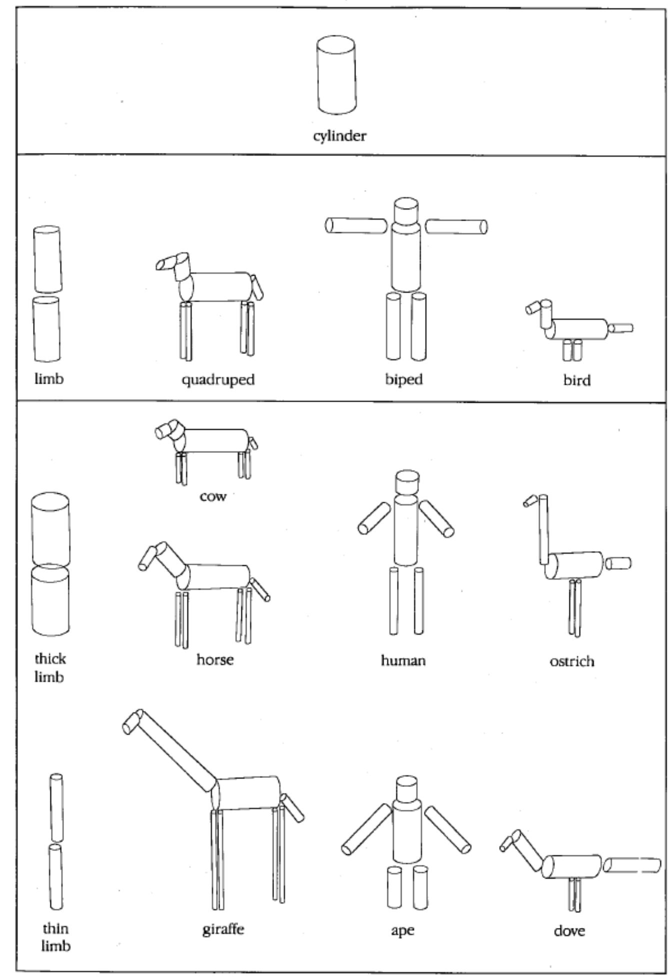
Alignment is inefficient

Assume each object is made up of a small number of parts

Use alignment to find parts, then impute object presence
by reasoning about relations

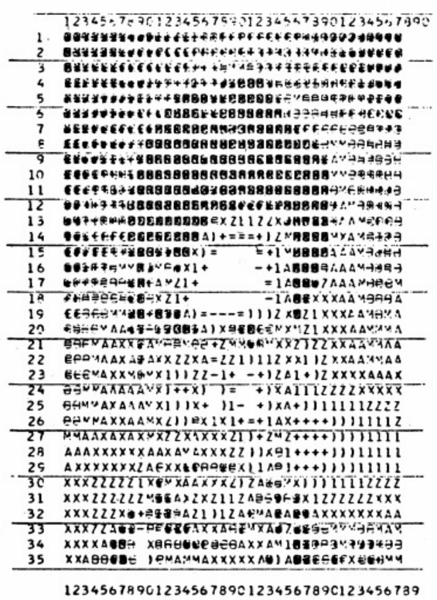
History of recognition: Hierarchies of parts



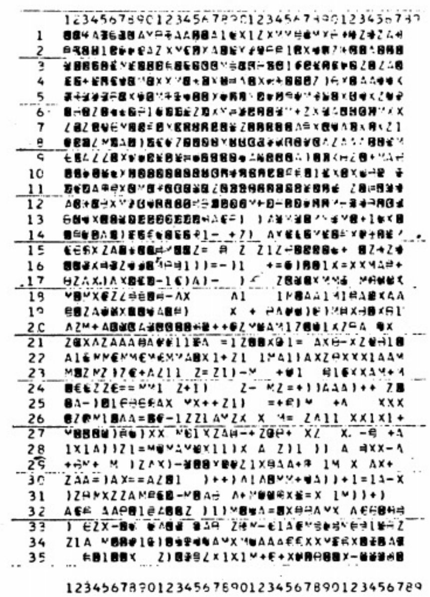


Figures from Marr's Vision (1982)

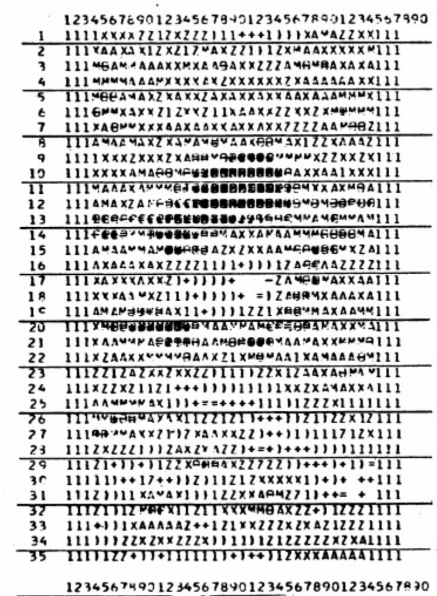
History of recognition: Deformable templates



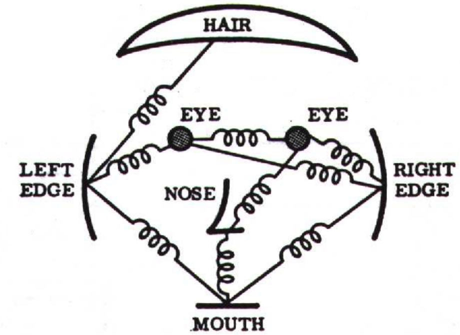
Original picture.



Noisy picture (sensed scene) as used in experiment.



L(EV)A for hair. (Density at a point is proportional to probability that hair is present at that location.)

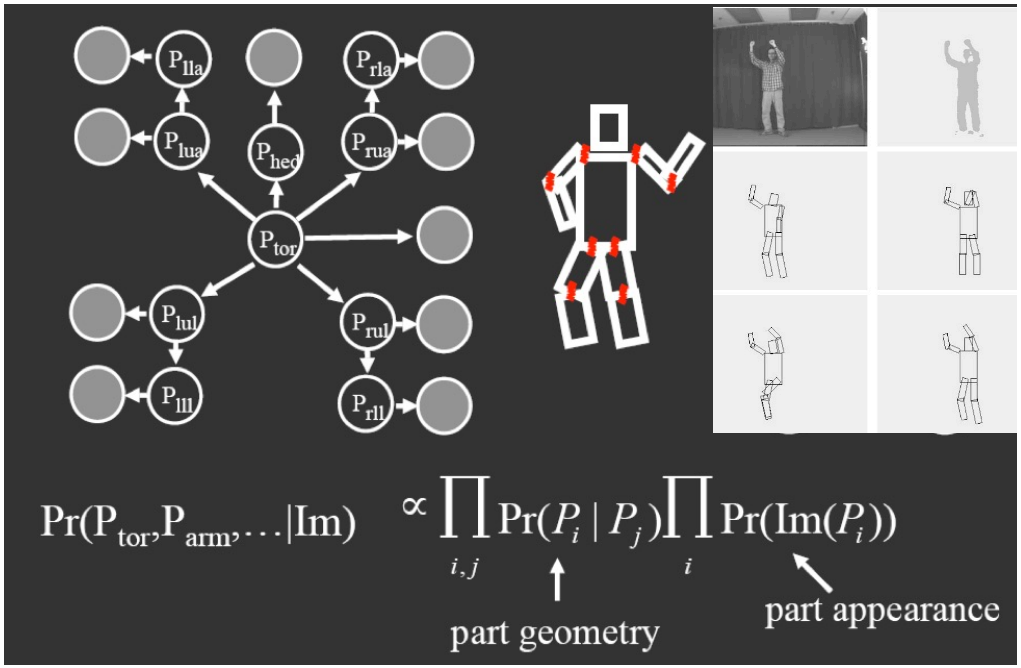


HAIR WAS LOCATED AT (11, 21)
L/EDGE WAS LOCATED AT (25, 11)
R/EDGE WAS LOCATED AT (25, 24)
L/EYE WAS LOCATED AT (21, 15)
R/EYE WAS LOCATED AT (21, 21)
NOSE WAS LOCATED AT (26, 18)
MOUTH WAS LOCATED AT (29, 17)

M. Fischler and R. Elschlager, The representation and matching of pictorial structures, IEEE Trans. on Computers, 1973

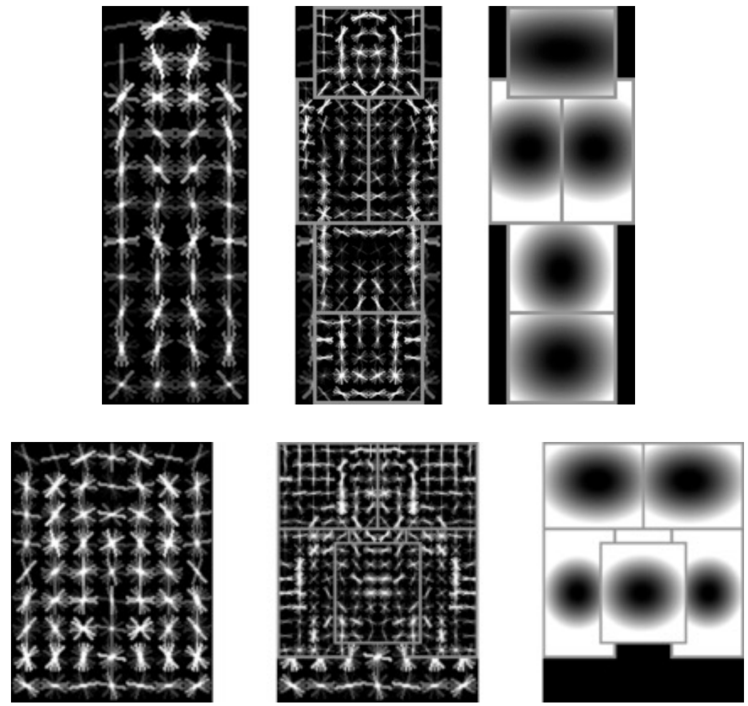
History of recognition: Deformable templates

Pictorial structures revisited



Felzenszwalb & Huttenlocher (2000)

Discriminatively trained deformable part-based models



Felzenszwalb et al. (2008)

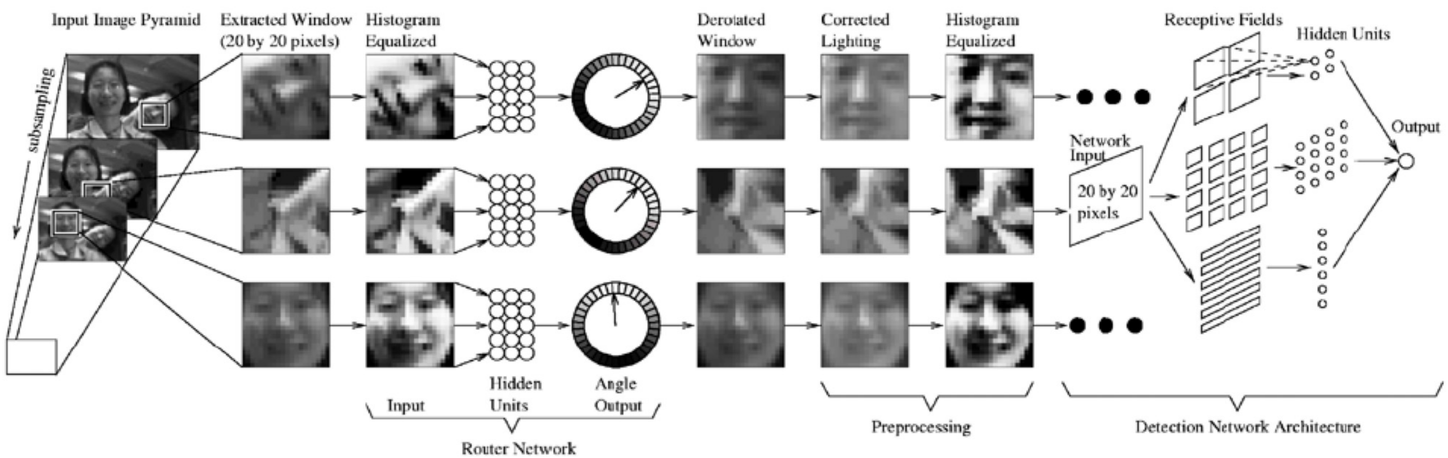
Idea: templates

Assume object has a characteristic appearance
    (from any viewpoint)

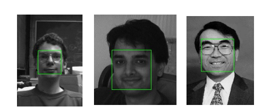
Build something that finds that appearance

Spectacular successes with face finding

Rowley-Baluja-Kanade face finder (1)

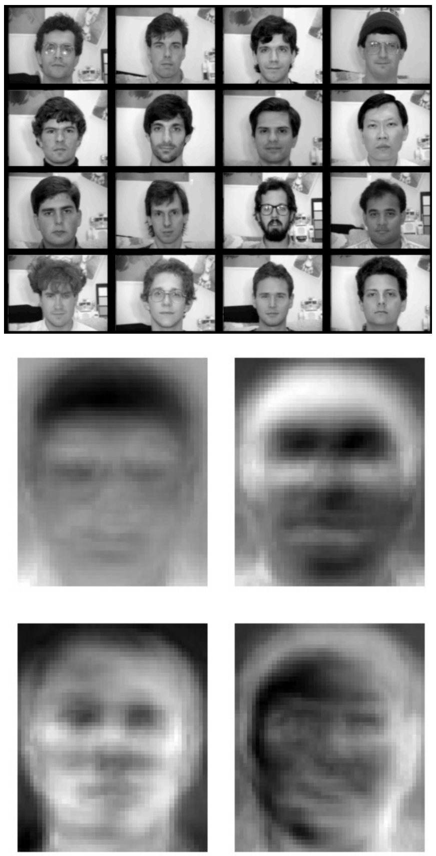


"Rotation invariant neural-network based face detection,"
H.A. Rowley, S. Baluja and T. Kanade, CVPR 1998

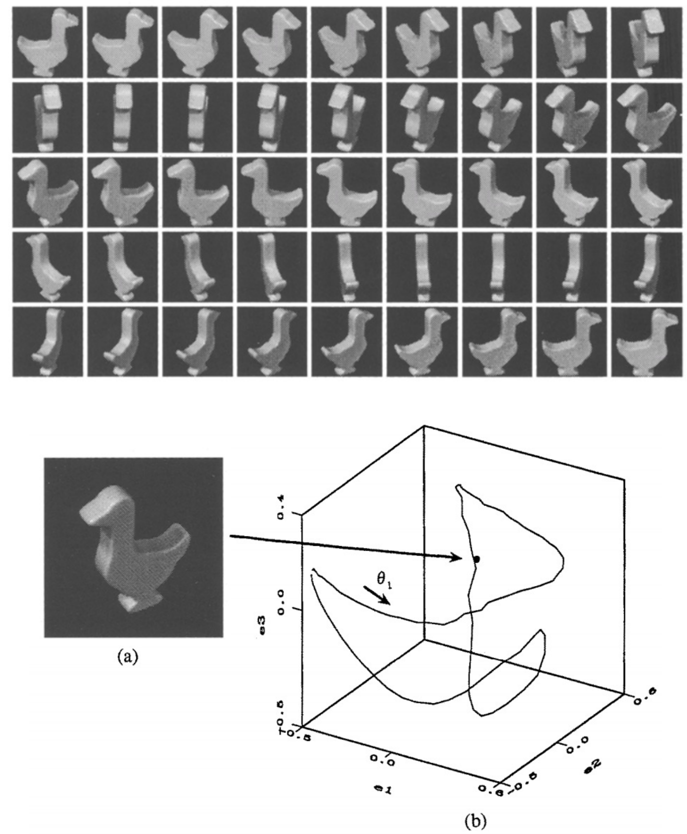


“Rotation invariant neural-network based face detection,”
H.A. Rowley, S. Baluja and T. Kanade, CVPR 1998

History of recognition: Appearance-based models



M. Turk and A. Pentland, [Face recognition using eigenfaces](), CVPR 1991



H. Murase and S. Nayar, [Visual learning and recognition of 3-d objects from appearance](), IJCV 1995

Idea: more complicated templates

Assume object has a characteristic appearance

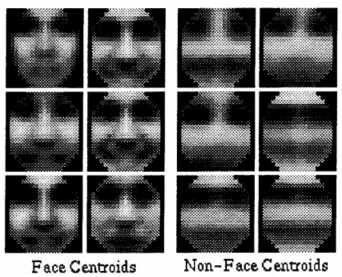
(from any viewpoint)

that might be hard to encode

Build rich encoding of that appearance

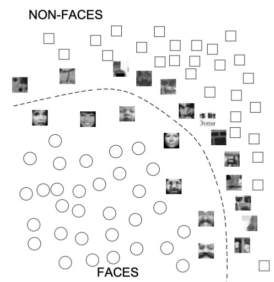
History of recognition: Features and classifiers

Appearance manifolds + neural network



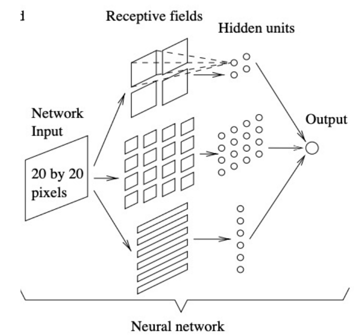
Sung & Poggio (1994)

Support vector machines



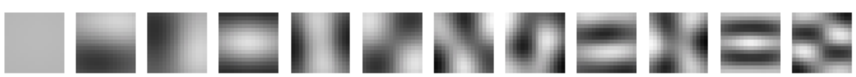
Osuna, Freund, Girosi (1997)

Neural network



Rowley, Baluja, Kanade (1998)

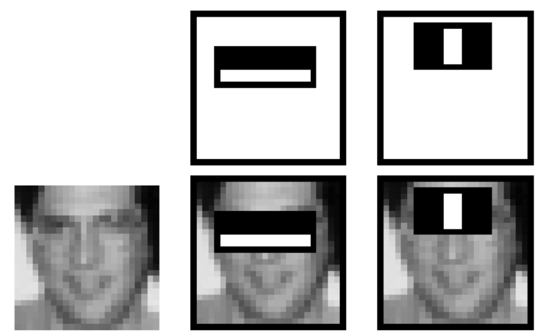
Statistics of feature responses, probabilistic classifier



$$\prod_{j=1}^{n_{magn}}\prod_{i=1}^{n_{subs}} \frac{P(q1_i^j|\text{object})P(pos_i^{j'}|q2_i,\text{object})}{\frac{P(q1_i^j|\overline{\text{object}})}{n_{subs}}} \underset{\overline{\text{object}}}{\overset{\text{object}}{\gtrless}} \lambda = \frac{P(\overline{\text{object}})}{P(\text{object})}$$

Schneiderman & Kanade (1998)

Rectangle features, boosting



Viola & Jones (2001)

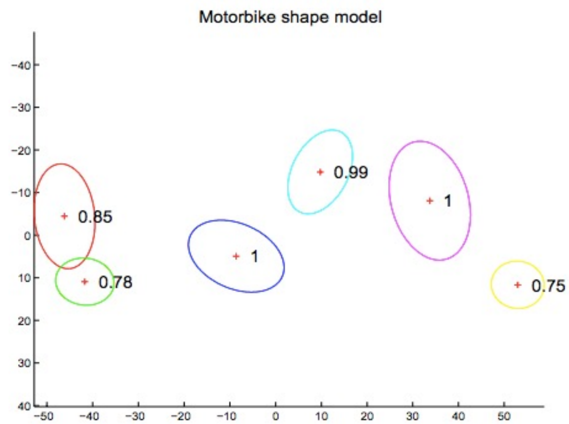
Idea: probabilistic templates

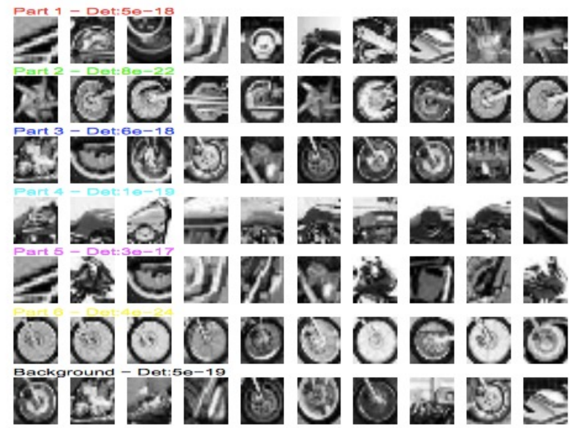
Assume "parts" that have characteristic appearance
(from any viewpoint)

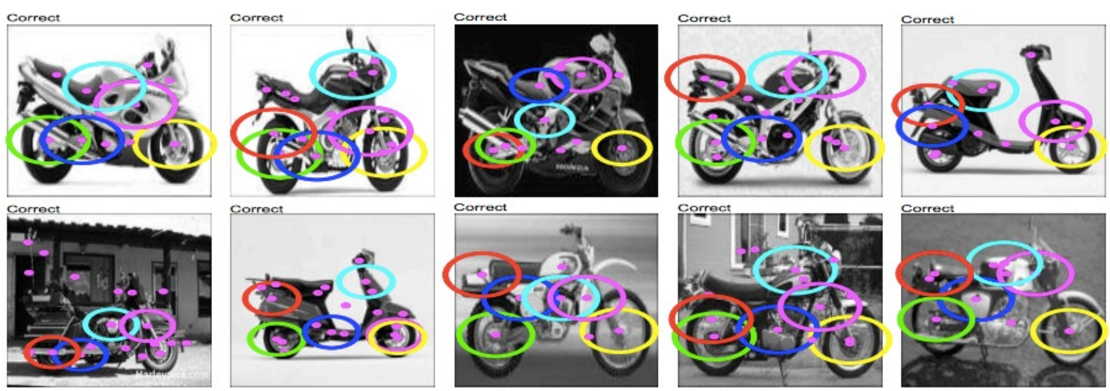
If enough parts are found in about the right relation to one another, the object is there

"About the right relation" - probabilistic

History of recognition: Constellation models





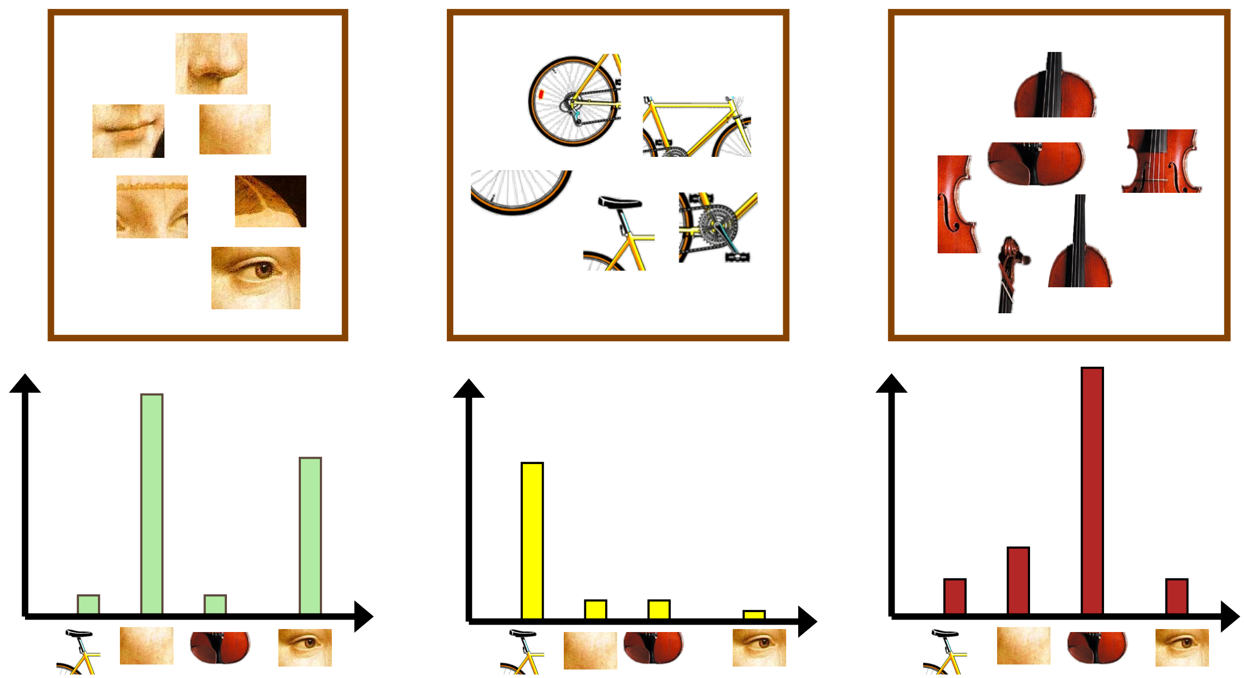


Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

Idea: bags of features

If enough distinctive features are there, the object is present
(you can ignore the relations which create complexity)
(and you might even be able to use voting to figure out
which features are reliable; where the object is; prev lecture)

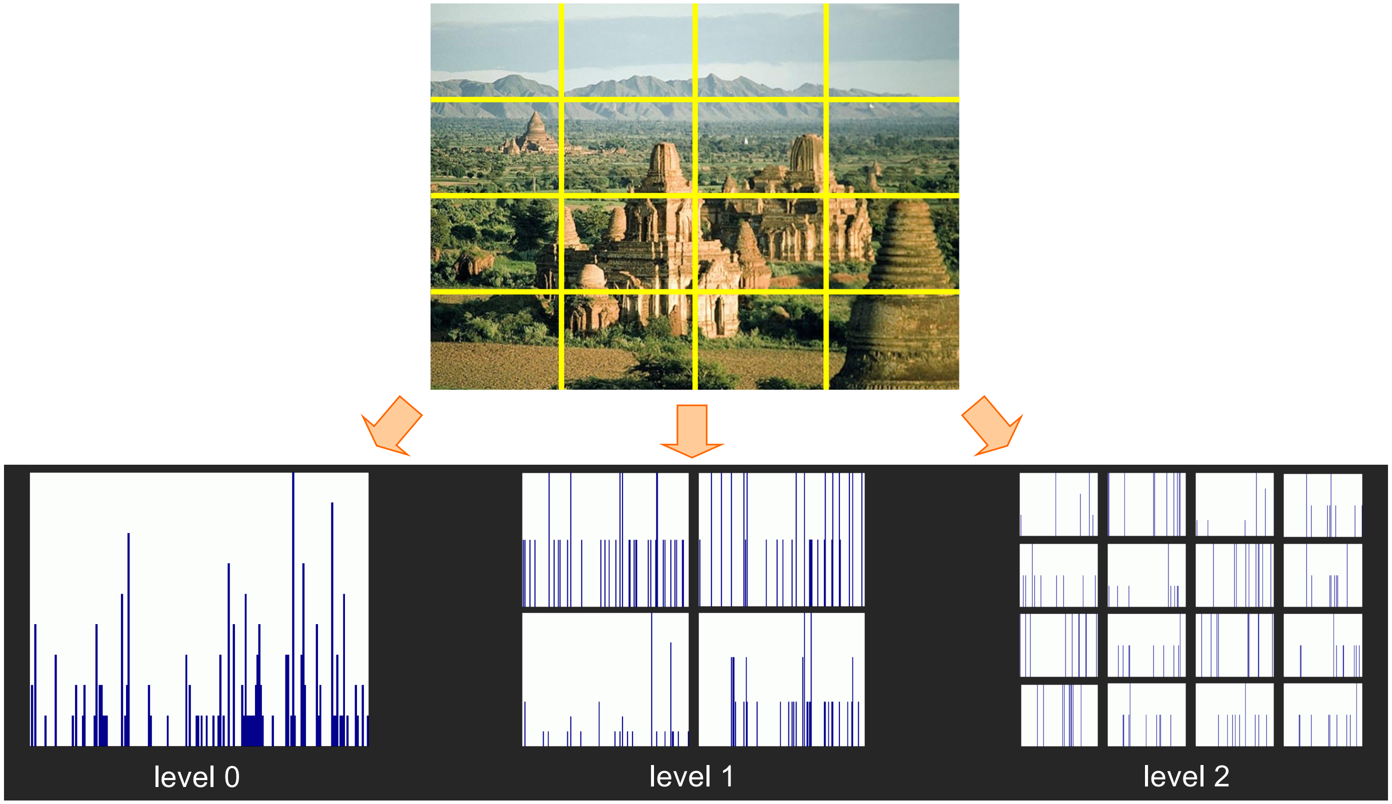
History of recognition: Bags of keypoints



Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

Spatial pyramids

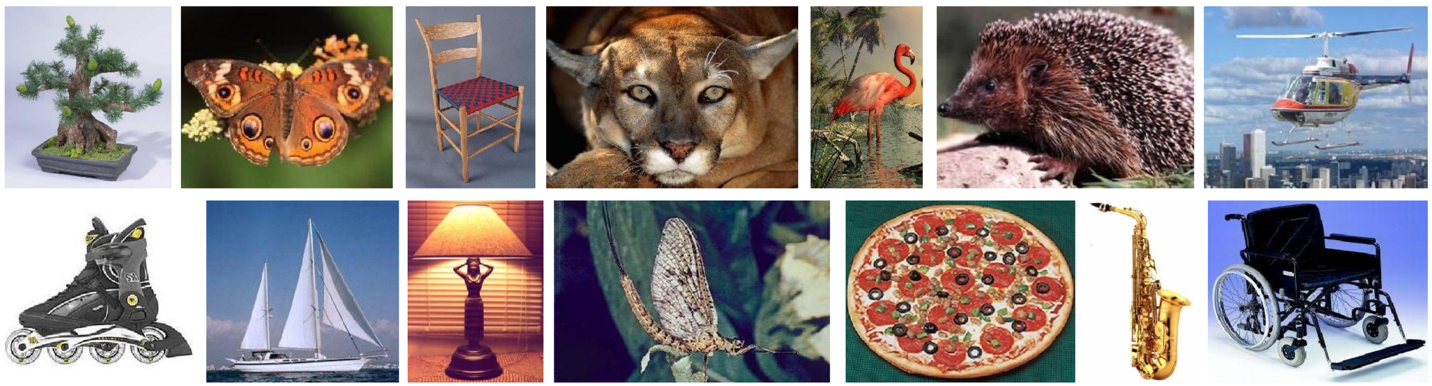
- Orderless pooling of local features over a coarse grid



Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramids

- Caltech101 classification results:



	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

Idea: objects are patterns of patterns of patterns…

And a simple pattern detector is easy to build…

Convolve with template, check against threshold
(= simple unit in neural network)

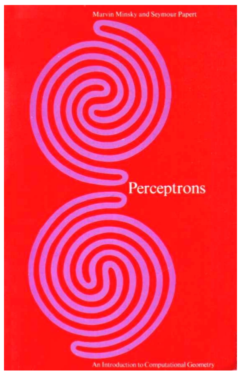
Stack these in layers, and you're there

Convolution kernels? Learn these to get the right behavior

History of recognition: Neural networks

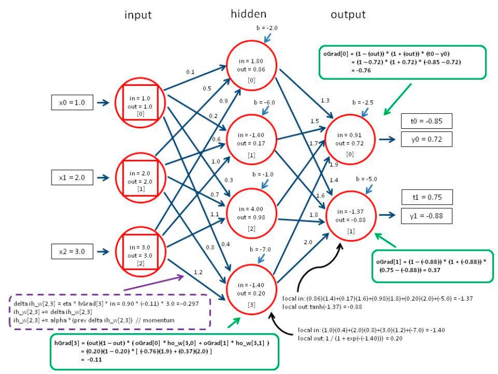
Perceptrons

Rosenblatt (1958)



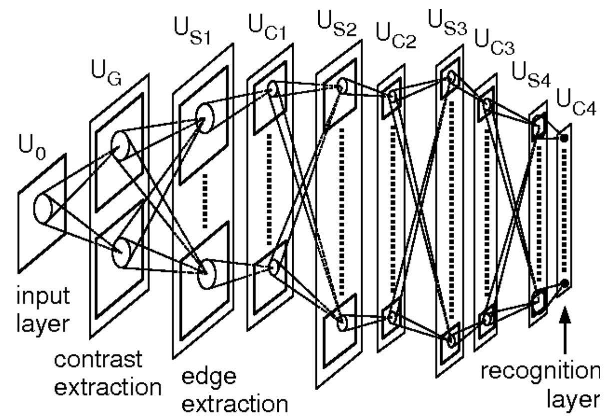
Minsky & Papert (1969)

Back-propagation



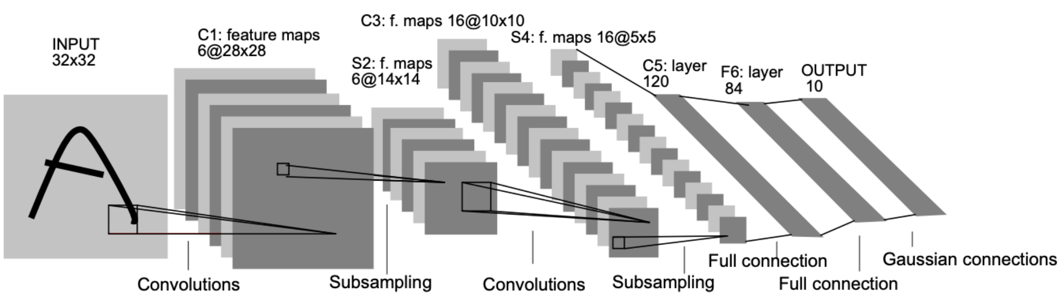
Rumelhart, Hinton & Williams (1986)

Neocognitron



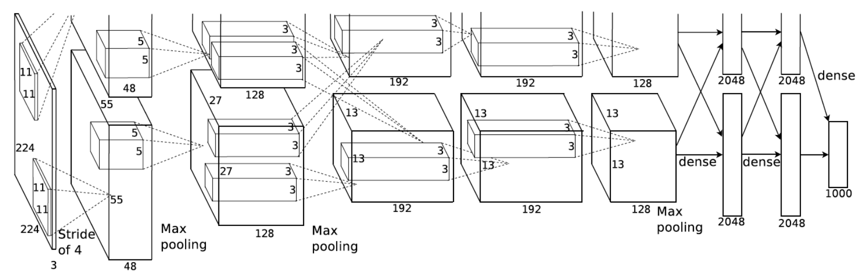
Fukushima (1980)

LeNet-5



LeCun et al. (1998)

AlexNet



Krizhevsky et al. (2012)

Announcements and reminders

- **Assignment 5** is due December 11
- **Final project reports** are due 23h59 CST
  - Thursday, December 17

- Anything missing that you hope we might be willing to give you credit for
  - Due 23h59 CST Thursday December 17