Object Recognition

Chris Williams

Neural Information Processing
School of Informatics, University of Edinburgh

January 15, 2018

Overview

- Neurobiology of Vision
- Computational Object Recognition: What's the Problem?
- Fukushima's Neocognitron
- HMAX model and more recent versions
- Some other approaches

Neurobiology of Vision

- WHAT pathway: V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ IT
- WHERE pathway: V1 $\rightarrow$ V2 $\rightarrow$ V3 $\rightarrow$ MT/V5 $\rightarrow$ parietal lobe
- IT (Inferotemporal cortex) has been shown to have cells that are relatively invariant to size and position of objects (e.g. face cells), but many are variable wrt view
- In the end what and where information must be combined, but it is not yet known where this happens

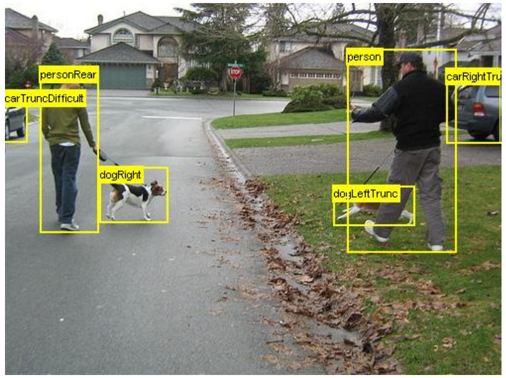
Invariances in higher visual cortex

mvrfigs/sheinberg1.png

Computational Object Recognition

- The big problem is creating *invariance* to scaling, translation, rotation (both in-plane and out-of-plane), and dealing with partial occlusion, while at the same time being selective
- However, note that humans/animals are not perfectly invariant, especially wrt 3D rotations
- Objects are not generally presented against a neutral background, but are embedded in *clutter*
- Object *class* recognition vs specific object recognition
- Tasks: classification, localization, segmentation and more ...

- Classification
  - Is there a dog in this image?
- Detection
  - Localize all the people (if any) in this image



- Segmentation
  - Label each pixel as class x or background

Some Computational Models

Two extremes:

- Extract 3D description of the world, and match it to stored 3D structural models (e.g. human as generalized cylinders)
- Collection of 2D views

Some other methods

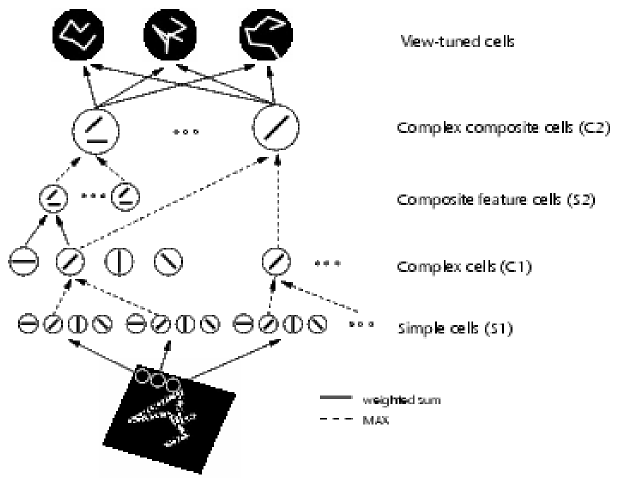
- 2D structural description (parts and spatial relationships)
- Match image features to model features, or do pose-space clustering (Hough transforms)
  - What are good types of features?
- Feedforward neural network (large input dimension, needs huge training set; no invariances apriori)
- Bag-of-features (no spatial structure; but what about the “binding problem”?)
- Scanning window methods to deal with translation/scale

Fukushima's Neocognitron

Fukushima (1980), Fukushima (1988)

- We wish to deal with imprecise scaling and location information
- Strategy is to "clone" (or replicate) a detector over a region of space, and then pool the responses of the cloned units; this trades off selectivity and invariance
- This strategy can then be repeated at higher levels, giving rise to greater invariance
- S-cells (simple cells) do convolution with local filters
- C-cells (complex cells) do pooling (sum or maximum) and down-sampling
- Object detection is based on the output of C2 complex cells
- Note that penultimate layer is like a "bag of features"
- See also Le Cun et al (1990), convolutional neural networks

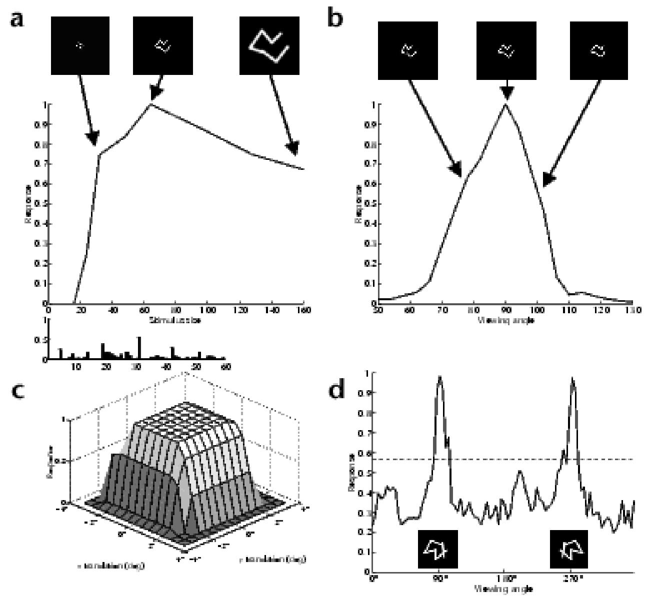
HMAX model



Reisenhuber and Poggio (1999)

HMAX model II

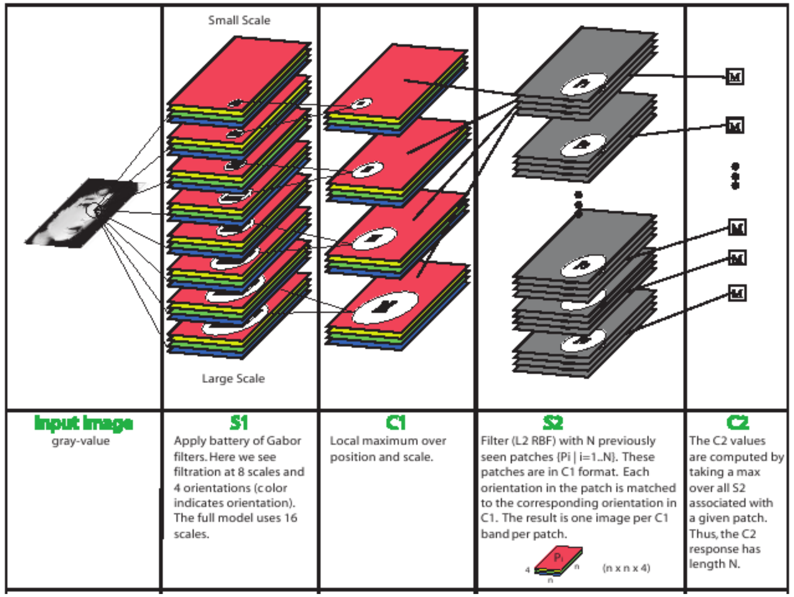
- S1 detectors based on Gabor filters at various scales, rotations and positions
- Riesenhuber and Poggio hand-coded S2 cells based on conjunctions of C1 cells (simple unsupervised learning)
- They used "paper clip" style stimuli
- Were able to show broad tuning curves wrt size, translation
- Scrambling of the input image does not give rise to object detections: not all conjunctions are preserved



Reisenhuber and Poggio (1999)

Serre et al (2007)

- Used real images as inputs
- As before, use Gabor filters at various orientations and scales as S1 features
- C1 takes max of S1 features over a range of scales and positions
- S2 layer of RBF units trained by using patterns of activation of the C1 layer patches as templates
- S2 units respond to patterns of edge/bar conjunctions
- Obtain K S2-layer maps, one for each C1 patch ($K \leq 1000$)
- C2 computes max over all positions and scales of each S2 map
- Use a SVM classifier on C2 outputs



Serre, Wolf, Bileschi, Riesenhuber, Poggio (2007)

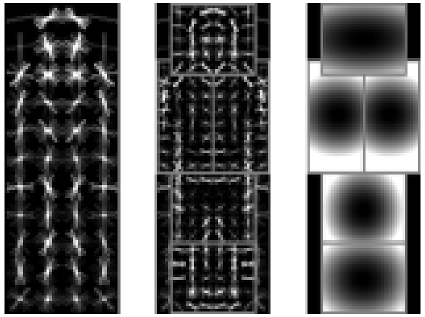
mvrfigs/serre_07-f1.png

Results

- Claimed as a model on a “rapid categorization task”, where back-projections are inactive
- Results on a animal vs non-animal rapid categorization task closely match human performance
- Classification results (Caltech 101) are state-of-the-art
- Localization can be achieved by using a sliding-window method
- The model doesn’t do segmentation (as opposed to bounding boxes)
- Similar performance can be obtained by bag-of-features models which don’t use the same S1/C1 representations

Felzenszwalb et al (2010)

- Current leading method for object localization in PASCAL VOC competitions (20 classes)
- The model is defined by a coarse root filter (a), several higher resolution part filters (b) and a spatial model for the location of each part relative to the root (c)
- The filters specify weights for histogram of oriented gradients features. Their visualization show the positive weights at different orientations.



(a) (b) (c)

- Histogram of oriented gradients (HOG) features are local histogram of oriented gradient responses (cf C1 units in Serre et al, and Lowe's SIFT descriptors (2004))
- The visualization of the spatial models reflects the "cost" of placing the center of a part at different locations relative to the root.
- Scanning window approach to object localization

Summary and Discussion

- Hierarchical feedforward pooling architectures are a common model for object recognition
- There are other possibilities: generative as opposed to discriminative models e.g. Sudderth et al (2005). Allows unsupervised training.
- Not much rôle for top-down influences in these models (e.g. for figure/ground separation)
- Many object recognition models are rather weak models of shape, and tend to focus on local texture descriptions
- Evaluation on standard datasets, e.g. PASCAL VOC competitions
- There is still much to be done to obtain human level performance!

References I

- P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32(9) (2010)
- Fukushima, K. Neocognitron: A self-organising multi-layered neural network. Biol. Cybern., 20:121-136 (1980)
- Fukushima, K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Networks, 1:119-130 (1988)
- Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Handwritten digit recognition with a back-propagation network. In Advances in Neural Information Processing Systems 2 (1990)

References II

- Logothetis, N. K. and Sheinberg, D. L. Visual Object Recognition. Ann Rev Neurosci 19 577-621 (1996)
- Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. IJCV 60(2) 91-110 (2004)
- Riesenhuber, M. and Poggio, T. Hierarchical models of object recognition in cortex. Nat. Neuro., 2:1019-1025 (1999)
- T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio. Object recognition with cortex-like mechanisms. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (3), pp. 411-426 , (2007)
- Serre, T., Oliva, A., and Poggio, T. A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci USA, 104(15):6424-6429 (2007)

References III

- Sudderth, E., Torralba, A., Freeman, W. T., and Willsky, A. S. Learning Hierarchical Models of Scenes, Objects and Parts. In ICCV 2005.