Notes on 'Modeling Human Visual Recognition'

When we see…: when we see a novel object, we typically assign it to a basic-level category, of which there exist about 5000 (estimated by Biederman, 87). Other category levels exist as well, e.g. sub-ordinate and super-ordinate. This categorization can take place within 150ms (±30ms) and occurs for canonical (typical/regular) views only. Non-canonical views (e.g. chairs from below) require more time to be correctly categorized.
Scientific debates:

- how much have we understood in those 150ms? (apart from just the category label).
- is this process working only bottom-up (feedforward) or does it include also top-down processes (feedback)?

Scene Memory: once an object or scene is categorized, a huge memory for that category is retrieved (loaded). The memory allows us to rapidly analyze and understand the structure using eye movements (=saccades). The memory contains for instance information about expected object locations.

http://geon.usc.edu/%7Ebiederman/publications/Biederman_1981.pdf

Structural Variability: structural variability is the variation in exact structure between category instances (objects of the same category). This is one of the major problems why it is difficult to model the categorization process. How is it possible to capture this variability and discriminate between 5000 categories?

Fragmented Contour Image: another major problem which makes modeling categorization difficult: if one tries to extract contours, then the resulting contour image almost always appears fragmented and can vary substantially for even the same object seen under just slightly different conditions (small viewpoint changes, small illumination changes, …). It is often tempting to think that a higher resolution and using multiple scales (see also scale space later) may escape these variations. In contrast, humans can also easily understand low-resolution images (e.g. 100x200 pixels – just check the resolution of the small images on the internet). Extremely put, one can regard a gray-scale image as a noise source with a signal placed into it.

Evolvement Concepts: summarizes the concepts that have been proposed so far. It may well be possible that the (biological) visual system uses all of them in some way. The question is how exactly? What are the elementary descriptors of structure? How are those descriptors integrated? The short answer is no one knows; the long answer follows now.

Marr: work by an influential vision scientist. The approach was based on the belief that the 3D layout of our environment needs to be reconstructed in order to categorize the image. It can be understood as a local-to-global approach (or part-whole; see also previous slide):

http://www.personal.rdg.ac.uk/~sxs05ag/pub/g2007/myg2007.pdf
http://homepages.inf.ed.ac.uk/rbf/BOOKS/MARR/marr.htm

Biederman: most influential psychologist. The object recognition theory (recognition-by-components) he proposed is in some sense a refinement of Marr's ideas of recognition- by-cylinders: parts show a variety of basic geometries, which are all made of vertices (intersection of lines).

http://geon.usc.edu/~biederman/

Neural Network: Biederman's neural network (NN) implementation of his object recognition theory. The first layer extracts local orientations (corresponding to V1 in primate visual cortex), which then are integrated to more complex features (e.g. vertices) in higher layers (corresponding to V2 and higher). Other NNs can be regarded as a variant of his approach.

Image Segmentation: is the process of separating objects from their background (see axis 'foreground – background' in slide 'evolvement concepts'). Works well with high-resolution images with distinct textures/contrasts, but on low-resolution images it's limitedly applicable for the purpose of categorization.

http://en.wikipedia.org/wiki/Segmentation_(image_processing)
http://people.csail.mit.edu/seth/pubs/taskforce/paragraph3_5_0_0_3.html
http://civs.ucla.edu/Segmentation/Segment.htm

Scale Space: is the blurring of the image to arrive at coarser descriptions of the structure, which for instance can be useful for contour extraction: at coarser scales contours can be more continuous, e.g. the silhouette of a tree, but they can also be accidental due to the blurring (smearing). In my 'concept summary' this is mentioned as the fine-coarse axis.

http://en.wikipedia.org/wiki/Scale-space

Saliency Map: is the idea that a simple preprocessing mechanism can select potentially interesting (salient) points in an image. 3 basic streams (orientation [as in NN], color, intensity). Loosely motivated by psychophysical studies on human visual search (e.g. by Treisman et al.; http://en.wikipedia.org/wiki/Visual_search) but tightly connected to neurophysiological results. Based on the local-to-global principle and exploiting the coarse/fine axis. Downside from an engineering perspective: slow due to repeated DOG filtering.

http://www.scholarpedia.org/article/Saliency_map
http://ilab.usc.edu/bu/

Object Detection: some computer vision scientist develop search algorithms to localize objects in images, which has little to do with categorization (remember: categorization is the assignment of the entire image to a category and less so the assignment of individual parts/components to their respective categories). Mentioned for completion only.

http://en.wikipedia.org/wiki/Feature_detection_(computer_vision)

Image Retrieval: is the search for similar images and often corresponds to basic-level categorization. Mentioned for completion only.

http://en.wikipedia.org/wiki/Image_retrieval

Caltech 101/256: the most successful object-based categorization system so far (sub-ordinate level), developed by Perona's group. Purely computationally motivated - not based on any psychophysical data; using a variety of traditional computational methods (principal component analysis [PCA], template matching,…). In some way the system performs exceptionally well. However, that is for 101 categories only. Can they crack the problem for larger data sets using this traditional methodology?

http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Spatial Envelope: the most successful scene categorization system (super-ordinate level), developed by Oliva and Torralba. Their belief is that scene categorization can be performed without image segmentation or any grouping. One can call it a global-to-local approach. By carrying out psychophysical experiments, they determined that subjects prefer to rate 'scenes' along 5 dimensions.

Images are preprocessed by a variant of the Fourier transform, and the resulting 'Energy Spectra' are reduced using the PCA. Their system works well at the super-ordinate level.

http://people.csail.mit.edu/torralba/code/spatialenvelope/

Criticism: although the above approaches do work astonishingly well (e.g. Caltech 101, Spatial Envelope), they have several major short-comings:

1) They were designed for a manually selected set of categories.
2) Once an image is categorized, the preprocessed image can *not* be used to interpret components/parts of the object or scene, that is, it can not be used to emulate a

human visual search process using saccades. In order to do that one had to determine the geometry of individual contours for instance.

3) Can not deal with images of overlapping content.
4) How about non-canonical views? Or any pattern or texture?

In summary, the challenge of categorization is still unsolved. In particular because of item no. 2, we pursue a categorization system which is based on contours and regions (relations between contours) in some sense what NNs try to do. However, in NNs, no actual transformation of the input takes place – they essentially perform template matching.

Contour Approach: illustrates the challenge of contour boundary segregation. Where and how exactly would one segregate a complex shape (or even a complex contour) into elementary segments suitable for description? High curvature points are obvious candidates, but how would one locate them? Solution: local/global space, see next two slides.

Signature: a contour is iterated with a fixed-size window: for the selected contour segment it is determined whether it constitutes an arc and if so, its amplitude is taken and assigned to a signature.

Local/global space: the iteration is done for different window sizes. The resulting range of signatures is called a local/global space. From this space two things can be derived: 1) points of highest curvature; 2) the contour geometry (local parameters: curvature, smoothness; global parameters: arc, inflexion, alternating, straight).

Symmetric-Axis Transform: is a transform, whose output expresses the region of a shape (or relation between contours), invented by Blum. It is based on a wave-propagation process (principle ‘Fliessblatt’). The resulting sym-axes can be used to express contour relations and outline the corresponding region. The transform is part of the Gestalt thinking that a structure can be described by self-collapsing it. Local/global irrelevant.

Image Rotation Invariance: another astounding characteristic of the basic-level categorization process is that it is still performed very quickly even though an image is rotated in the image plane (in 2D; not to be confused with the rotation of an object in 3D, which results in a non-canonical view). Up-side down images however are exceptions; it requires much more time to categorize them.

Translation Invariance: and another breath-taking characteristic: images can be recognized independent of their spatial location to quite some extent. Thorpe asked subjects to perform a saccade toward the image that contained an animal (animal/no-animal discrimination task): saccades were triggered after only 120ms! No useful model exists yet, which can perform this task. Which biological NN could perform such a feat?
Scientific debates:

- is a complete categorization necessary to perform this discrimination task?
- are attentional (covert) shifts involved?

Matlab commands: type ‘help images’, which lists the commands of the image processing toolbox (if installed).

```
I = imread('cameraman.tif');   % loads image
figure(1); imshow(I);          % displays image
C = edge(I);                   % edge extraction
figure(2); imshow(C);
gf = fspecial('gaussian', 50);  % gaussian filter, 20x20 pixels
L = imfilter(I, gf);            % filtering image
figure(3); imshow(L);
C2 = edge(L);              % edge extraction on filtered image
figure(4); imshow(C2);
```

Reading

- The following two books are recommended for computer vision scientists attempting to move toward human models of perception and recognition:

Active Vision: The Psychology of Looking and Seeing

John M Findlay and Iain D Gilchrist

1) Compact overview of essential eye-movement studies.

2) A convincing summary of arguments against the excessive 'use' of attentional shifts.

Vision Science: Photons to Phenomenology

Stephen E Palmer

An opus, settled at the interface of psychology, neuroscience and computer science. Explains the topic as a whole.

http://socrates.berkeley.edu/~plab/book.htm

- The following two are recommended when plunging into psychophysical aspects of vision:

Visual Perception: Physiology, Psychology, Ecology

Vicki Bruce, Mark A. Georgeson and Patrick R. Green

Covers rather early stages of vision.

http://www.psypress.com/visualperception/

Foundations of Vision

Brian A. Wandell

Covers also early stages of vision.

- Books for natural scientists moving toward computer science/computer vision:

Digital Signal Processing: A Practical Guide for Engineers and Scientists

Steven Smith

Excellent introduction on signal processing. Contains only the essential equations for purposes of clarity.

http://www.dspguide.com/

Machine Vision

Ramesh Jain, Rangachar Kasturi and Brian G. Schunck

The basics simply explained.

Primers/Intros

Filtering: my own attempt to make the idea of filtering accessible to psych students:

http://www.allpsych.uni-giessen.de/rasche/paps/tutorial_DSP.pdf

Principal component analysis (PCA):

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Support vector machines (SVM):

http://www.nature.com/nbt/journal/v24/n12/pdf/nbt1206-1565.pdf