Top–down learning of low-level vision tasks

Michael J. Jones, Pawan Sinha, Thomas Vetter and Tomaso Poggio

Perceptual tasks such as edge detection, image segmentation, lightness computation and estimation of three-dimensional structure are considered to be low-level or mid-level vision problems and are traditionally approached in a bottom–up, generic and hard-wired way. An alternative to this would be to take a top–down, object-class-specific and example-based approach. In this paper, we present a simple computational model implementing the latter approach. The results generated by our model when tested on edge-detection and view-prediction tasks for three-dimensional objects are consistent with human perceptual expectations. The model's performance is highly tolerant to the problems of sensor noise and incomplete input image information. Results obtained with conventional bottom–up strategies show much less immunity to these problems. We interpret the encouraging performance of our computational model as evidence in support of the hypothesis that the human visual system may learn to perform supposedly low-level perceptual tasks in a top–down fashion.

Address: E25–201, Center for Biological and Computational Learning,
Massachusetts Institute of Technology, 45 Carleton Street,
Cambridge, Massachusetts 02142, USA.

Correspondence: Tomaso Poggio
E-mail: tp@ai.mit.edu

Received: 11 August 1997
Revised: 29 September 1997
Accepted: 2 October 1997

Current Biology 1997, 7:991–994
http://biomednet.com/elecref/0960982200700991

Results and discussion

The extraction of edges from images is believed to be a key objective of early visual processing. As many years of work on edge detection have shown [1–3], the problem is difficult, in part because physical edges — meant as discontinuities in three-dimensional structure and albedo that convey information about the object's shape and identity — do not always generate intensity edges in the image. Conversely, intensity edges are often due to shading effects produced by illumination and, therefore, do not always reflect intrinsic properties of the object.

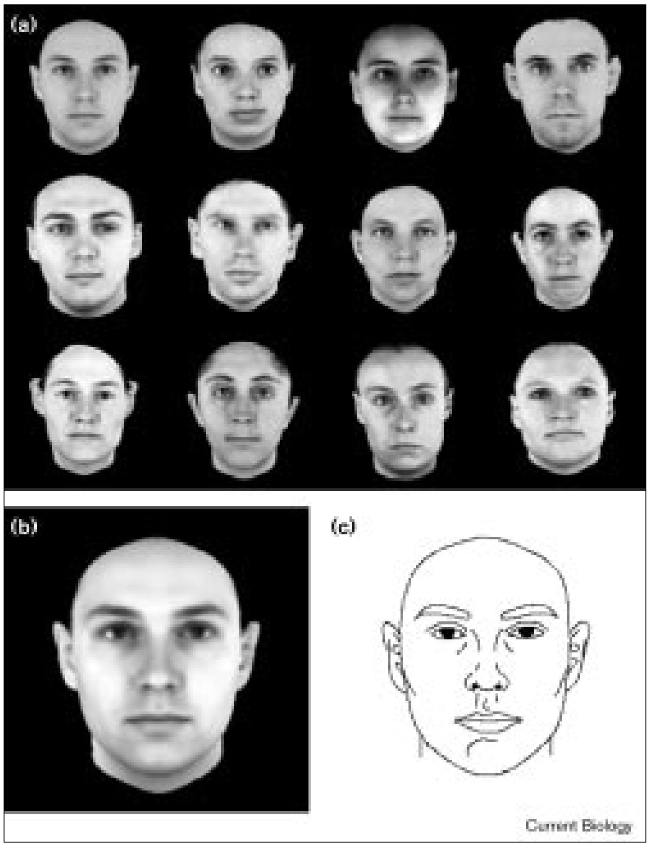
We propose a novel solution to this problem that involves casting it as a learning task. Given a set of sample prototypical, grey-level, face images and the corresponding line drawings, drawn by an artist, the task is to learn the mapping that associates an 'ideal' line drawing of a face to its grey-level image. We discuss next a specific algorithm to realize this general idea.

Our algorithm is based on the flexible model introduced by Vetter, Jones and Poggio [4–6]. The algorithm first establishes pixelwise correspondences between a reference image and the other prototype images using an optical flow scheme or recent extensions [4–7]. Once the correspondences are computed, an image is represented as a shape vector and a texture vector. The shape vector, which specifies how the two-dimensional shape of the example differs from a reference image, corresponds to the flow field between the two images. The texture vector specifies how the texture differs in correspondence from the reference texture. Here we use the term 'texture' to mean, simply, the pixel intensities — grey level or colour values — of the image. Our flexible model for an object class is then a linear combination of the example shape and texture vectors. The matching of the model to a novel image consists of optimizing the linear coefficients of the shape and texture components.

The flexible model can be used for learning a simple visual task, like 'ideal' edge detection, in the following way. Assume that a good line drawing is available for each of the prototypical grey-level images. Then, given the image of a novel face, the approach is to estimate the parameters, or linear coefficients, of the best-fitting grey-level flexible model and to plug the same parameter values into a second flexible model built from the prototypical line drawings [8]. In general, the mapping from grey-level images to line drawings can be learned, for instance, by an approximation scheme if enough additional example pairs — images and associated line drawings — are available.

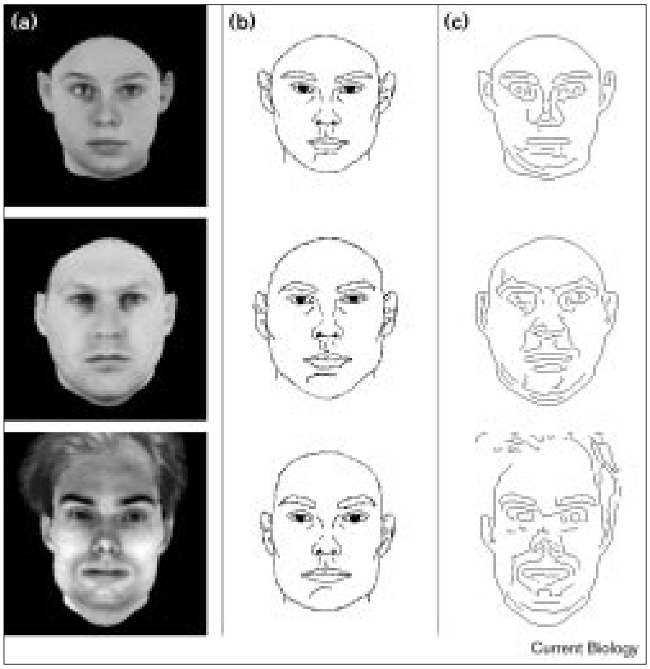
We have implemented an even simpler version of the scheme. We assume that the ideal line drawing corresponding to the average prototype is available from an artist (P.S.), such as that shown in Figure 1b and c. The matching of the flexible model obtained from the prototypes — some of which are shown in Figure 1a — to a novel grey-level image provides a shape vector that is a linear combination of the prototypes and that effectively tells how to warp the average shape of the grey-level prototype in order to match the shape of the novel grey-level image. Because the line drawings are supported on a subset of the pixels of the corresponding grey-level images, the line drawings of novel images can be obtained by warping the line drawing of the reference prototype using the estimated shape vector.

Figure 1



(a) Twelve of the 100 prototype faces that were set in pixelwise correspondence and then used for creating a flexible model of human faces. **(b)** The reference face and **(c)** its corresponding line drawing created by an artist.

Figure 2



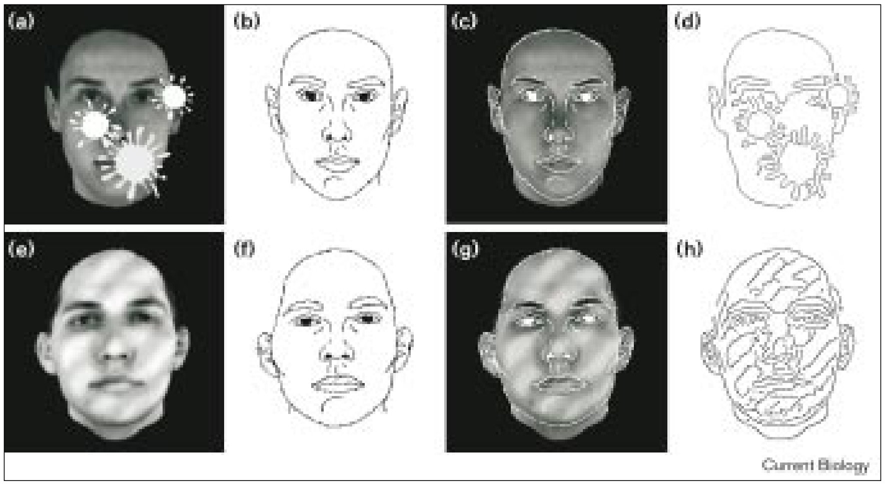
Examples of ideal edges found by the algorithm described in the paper. Column **(a)** shows the input novel images. Column **(b)** shows the line drawings estimated automatically by the algorithm that matches the flexible models to the novel images and then appropriately modifies the ideal edges of the reference image. For comparison, column **(c)** shows the edges found by a bottom–up edge detector (see [9]). Note that the ideal edges emphasize the perceptually significant features of the face much better than the Canny edges.

Figure 2 shows a few examples of novel images, not included in the set of prototypical examples, and the line drawing estimated from each of them by our ‘ideal edge detector’. To contrast this approach to a low-level gradient-based approach, Figure 2 also shows the edges found for each face image by a Canny edge detector [3]. Figure 3 shows the ideal edge-map estimated for two different face images with, potentially, many irrelevant edges and partial occlusion of the relevant ones. As is evident from the examples, our algorithm can detect and complete edges that do not correspond to any intensity gradients in the image and ignores those that are not intrinsic to faces. The power of the algorithm derives from the high-level knowledge of faces, learned from the set of prototypical images. As for how the system recognizes that the image is that of a face, we suggest that a prior step of object detection [9,10] provides information about which images or image regions are likely to contain faces.

One issue that deserves discussion here is the amount of variability in input images that a scheme such as the one proposed here can handle. All of the images we have used to illustrate our ideas are relatively similar. For our scheme to work with very dissimilar faces — such as with different hair-styles or under different lighting conditions or from different viewpoints — it would need to be trained using representative examples (see, for instance, [11]). In this proof-of-concept demonstration, limiting the variability of the face-set allowed us to test the scheme using a relatively small number of training examples.

Other supposedly early or mid-level visual tasks can be learned in a similar way. We describe briefly two of them: the generation of ‘virtual’ views and the estimation of three-dimensional structure from single images. Consider the case in which only one example image of an object is available. This may occur in object recognition tasks in which an object has to be recognized from a novel view. Vetter and Poggio [8] have considered this problem for linear object classes. An object belongs to a linear class if its three-dimensional structure can be described exactly as a linear combination of the three-dimensional structure of a small number of prototypes [4]. On the basis of this assumption, it can be proven that a new virtual view of an object belonging to a linear class

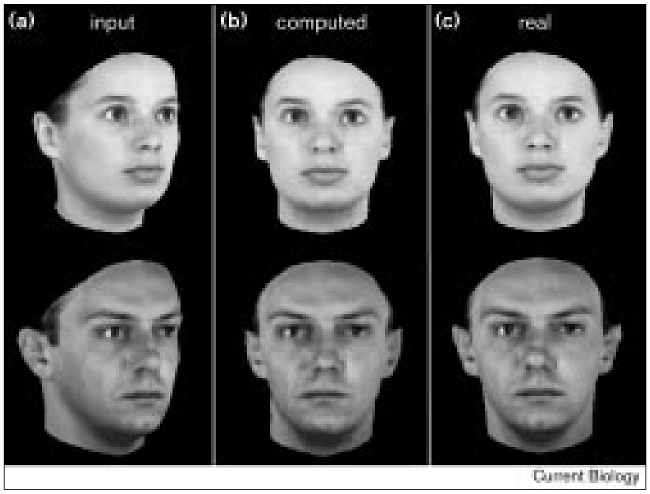
Figure 3



(a–d) Example of ideal edges (b) found for a partially occluded input face (a). The image (c) shows these edges overlaid on the unoccluded input face. (d) shows the Canny edge map for comparison. (e–h) An example of an image with non-intrinsic edges such as those due to shadows (e). Our method for finding ideal edges effectively ignores the spurious edges (f). The accuracy of these edges can be assessed from the image in (g) showing the edges overlaid on the image. (h) highlights the susceptibility of bottom–up edge-extraction approaches [9; and this study] to spurious image artefacts.

can be generated exactly from a single example view, and represented as a two-dimensional shape vector, provided appropriate prototypical views of other objects in the same class are available (under orthographic projection). In this way, and as illustrated in Figure 4, new views of a specific face with a different pose can be estimated and synthesized from a single view. The procedure is exact for linear classes; empirically, faces seem to be close to a linear class so the procedure above provides a good approximation for pose and expression. Again, this procedure can be formulated in terms of the learning metaphor in which a learning box is trained with input–output pairs of prototypical views representing each prototype in the initial and in the desired pose. Then, for a new input image the system synthesizes a virtual view in the desired pose [8].

Figure 4



An illustration of how the top–down strategy can be used to generate virtual views of three-dimensional objects – here, two human heads. (a) Input images. (b) Computed virtual frontal views. To allow the reader to assess the fidelity of the computed virtual views, we have included the real frontal views (c).

The estimation of three-dimensional structure from a single image would proceed in a very similar way, provided the image and the three-dimensional structures of a sufficient number of prototypical objects of the same class are available [4,8]. In our learning box metaphor, the system, trained with pairs of prototype images as inputs — represented as two-dimensional shape vectors — and their three-dimensional shapes as output, would effectively compute shape for novel images of the same class (compare with the somewhat different approach of [12]). A similar approach may be extended to problems of colour constancy and motion analysis, in which the desired information about colour or motion is provided in a learning-from-examples scheme based on the use of a class-specific flexible model.

The benefits that a top–down strategy potentially confers on a visual system include heightened immunity to noise and the ability to fill in missing input information on the basis of previously acquired class-specific information. An experimentally verifiable prediction, of the hypothesis that our visual system uses such a strategy, is that near-threshold contrast images of familiar classes of objects embedded in noise will be more easily detectable than similar contrast images of unfamiliar objects. The ability to fill in missing information in a top–down fashion can prove invaluable, for instance, in a situation where image quality is insufficient to allow for accurate three-dimensional shape recovery because of lack of binocular input and poor shading information. Top–down shape estimation can, in such circumstances, guide processes such as reaching and prehension.

In summary, we have provided here a class of algorithms that can be used to learn to perform visual tasks in a top–down way, specific to object classes. We conjecture — as some others have [13–15] — that perception in humans may rely on such processes to a greater extent than commonly assumed. Of course, biological vision may use bottom–up verification routines to validate the top–down 'hallucination' [14,15]. A similar verification approach (top–down and bottom–up) could also be effectively used in machine-vision implementations like the one described here.

Logically, our conjecture consists of two somewhat independent parts. The first one is that, at least in some cases, the visual system may solve low-level vision problems by exploiting prior information specific to the task and to the type of visual input. This argument may apply to several visual processes, such as motion, stereopsis, colour and computation of shape from a variety of cues (shape-from-X) [16]. (The inclusion of colour in this list may seem somewhat counter-intuitive given that colour, as a surface property, is often assumed to be independent of shape; however, illusions such as the McCollough effect [17] support our conjecture of shape-dependent and, perhaps, top–down influences on colour perception.)

The second part of the conjecture relates to how these specific algorithms may be synthesized by our visual system. The idea — the main point of this paper — is that visual systems may learn algorithms specific to a class of objects by associating in each 'prototypical' example an 'ideal' output to the input view. The 'ideal' outputs may be available through other sensory modalities, sequences of images in time or even explicit instruction. It is unlikely to be the case that a human necessarily needs explicit instructions to be able to produce line drawings corresponding to continuous-tone images. Information about which edges are the significant ones in a face might be derived by observing their variability across time or multiple instances. Explicit instruction might, however, facilitate this process just as taking an art course renders one more aware of some subtle characteristics of the visual world. Also, the notion of what constitutes an 'ideal' output corresponding to a certain class of inputs may change and evolve over time as the learning process encounters new examples. This second part of the conjecture predicts that human subjects should be able to learn to associate arbitrary outputs with input images and to generalize from these learned associations. There is a weak and a strong form of the conjecture. The strong form is that the learning follows the linear combination algorithm we have used here in our plausibility demonstration. The weak form of the conjecture, which we favour, leaves open the specific learning scheme. Preliminary psychophysical evidence favours the conjecture [18], with more experiments under way. Further work may enable us to verify whether the strong or weak form is to be preferred and, in the latter case, which learning scheme may be used by the visual system.

Acknowledgements

We are grateful to Peter Dayan, Federico Girosi, Eric Grimson, Edward Adelson and Shimon Ullman for insightful comments and suggestions. This research is sponsored by grants from ONR under contract N00014-93-1-0385 and from ARPA-ONR under contract N00014-92-J-1879; and by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program). Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Sumitomo Metal Industries, Kodak, Daimler-Benz and Siemens AG. Support for the A.I. Laboratory's artificial intelligence research is provided by ARPA-ONR contract N00014-91-J-4038. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at MIT's Whitaker College.

References

1. Marr D, Hildreth E: Theory of edge-detection. *Proc R Soc Lond [Biol]* 1980, 207:187-217.
2. Haralick RM: Edge and region analysis for digital image data. *Comp Graph Image Proc* 1980, 12:60-73.
3. Canny JF: A computational approach to edge-detection. *IEEE Trans Patt Anal Mach Vis* 1986, 8:679-698.
4. Poggio T, Vetter T: *Recognition and Structure from One 2D Model View: Observations on Prototypes, Object Classes and Symmetries.* A.I. Memo No. 1347. Cambridge, Massachusetts: Artificial Intelligence Laboratory, Massachusetts Institute of Technology; 1992.
5. Jones M, Poggio T: Model-based matching by linear combinations of prototypes. In *Proceedings of the Fifth International Conference on Computer Vision*. Los Alamitos, California: IEEE Computer Society Press; 1995:531-536.
6. Beymer D, Poggio T: Image representations for visual learning. *Science* 1996, 272:1905-1909.
7. Bergen JR, Hingorani R: *Hierarchical Motion-based Frame-rate Conversion.* Princeton, New Jersey: Technical Report, David Sarnoff Research Center; 1991.
8. Vetter T, Poggio T: Image synthesis from a single example view. In *Computer Vision – ECCV '96, Notes in Computer Science.* Cambridge, UK: Springer; 1996:1065.
9. Sung KK, Poggio T: Example-based learning for view-based human face detection. In *Proceedings of the Image Understanding Workshop*. Monterey, California: Kauffman; 1994.
10. Rowley H, Baluja S, Kanade T: *Human Face Detection in Visual Scenes.* Pittsburgh: Technical Report, Carnegie Mellon University, Computer Science; 1995: 95-158.
11. Jones M: Multidimensional morphable models: a framework for representing and matching object classes. Doctoral dissertation, submitted to the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology; 1997.
12. Redlich AN, Atick JJ, Griffin PA: Statistical approach to shape from shading: deriving 3D face surfaces from single 2D images. In *Comput Neural Syst* 1996, 7:1.
13. Cavanagh P: What's up in top-down processing? In *Representations of Vision*. Cambridge, UK: Cambridge University Press; 1991.
14. Mumford D: On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybernet* 1992, 66:241-251.
15. Ullman S: Sequence seeking and counter streams: a model for bidirectional information flow in the visual cortex. *Cereb Cortex* 1995, 5:1-11.
16. Yuille AL: *Shape From Shading, Occlusion and Texture.* A.I. Memo No. 885. Cambridge Massachusetts: Artificial Intelligence Laboratory, Massachusetts Institute of Technology; 1987.
17. McCollough C: Color adaptation of edge-detectors in the human visual system. *Science* 1965, 149:3688.
18. Sinha P, Poggio T: Role of learning in three-dimensional form perception. *Nature* 1996, 384:460-463.