Image interpretation by a single bottom-up top-down cycle

Boris Epshtein*, Ita Lifshitz, and Shimon Ullman†

Department of Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Charles D. Gilbert, The Rockefeller University, New York, NY, and approved July 22, 2008 (received for review February 7, 2008)

The human visual system recognizes objects and their constituent parts rapidly and with high accuracy. Standard models of recognition by the visual cortex use feed-forward processing, in which an object's parts are detected before the complete object. However, parts are often ambiguous on their own and require the prior detection and localization of the entire object. We show how a cortical-like hierarchy obtains recognition and localization of objects and parts at multiple levels nearly simultaneously by a single feed-forward sweep from low to high levels of the hierarchy, followed by a feedback sweep from high- to low-level areas.

computer vision | object recognition | parts interpretation | cortical hierarchy | feedback processing

In the course of visual object recognition, we quickly recognize not only complete objects but also parts and subparts at different levels of detail. Hierarchical models of the visual cortex (1–3) typically perform recognition in a feed-forward manner in which recognition proceeds from the detection of simple features to more complex parts to the full object. However, the recognition of local parts is often ambiguous and depends on the object's context (Fig. 1), which is not available during feed-forward processing.

Psychological studies have also shown that the identification of a global shape and its local components proceed at similar speeds. Depending on the configuration, the global shape can either precede or follow the recognition of its local parts, and both contribute to final recognition (4, 5). Event-related potential (ERP) (6) and magnetoencephalography (MEG) (7) recordings have shown fast responses to both objects and parts, and physiological studies found that shape selectivity at different cortical levels emerges quickly and can sometimes further increase over a short time interval (8–11).

We show below how objects and their multilevel components can be detected by the cortical hierarchy efficiently and almost simultaneously, even when the local parts on their own are highly ambiguous. Unlike feed-forward models, the basic computation is a particular bottom-up (BU) top-down (TD) cycle. Feed-forward recognition was shown in past modeling to produce fast effective top-level recognition. However, we show that even when correct recognition is obtained by the BU pass, frequent errors occur at the parts level. A single TD pass is sufficient to correct almost all errors made during the BU pass, and the full cycle obtains not only object recognition but a detailed interpretation of the entire figure at multiple levels of details. We first describe below the computational model used for object and part recognition and then report testing results on natural images.

Bidirectional Hierarchical Model. In this section, we consider the problem of detecting an object C together with a set P of parts P_i of different sizes and locations, such as a face together with eyes, nose, mouth, eyebrow, nostril, upper lip, etc. The computation is performed in a hierarchical network similar to previous cortical models (2, 3). The input to the visual process consists of the detection of a set F of low-level features F_i (such as simple and complex receptive fields); all other processing stages are derived from these measurements. If a part P_i can be present in one of n different image locations, we consider it below as a variable with $n + 1$ possible values: $P_i = 0$ means that the part is not present in the image, $P_i = j$ means that P_i is present at location j. The full-recognition problem can be described as inferring the most likely values of C and all of the parts P_i from F. This can be expressed as finding values for C (the class) and P (the set of all parts) to maximize the probability $p(C, P|F)$:

$$C, P = \text{argmax}\; p(C, P|F). \qquad [1]$$

For the simulations and testing, features hierarchies for several object categories were extracted automatically from image examples. The part hierarchy used for the current bidirectional interpretation is extracted by the same process described in the past for feed-forward classification (refs. 12 and 13; similar parts used in ref. 14). In this process, object parts were first extracted from image examples by identifying common subregions with high information content (12). The same process was repeated to produce a hierarchy of informative parts and subparts (13). The hierarchy construction is described in *Methods* [and in *Hierarchies* in supporting information (SI) *Appendix*]. Examples of hierarchies obtained in this manner are illustrated in Fig. 2.

For a given feature hierarchy, we next approximate the probability $p(C, P, F)$ above using a factorization into local terms (*Methods* and *Part Model* in *SI Appendix*):

$$p(C, P, F) = \Pi p(C)p(P_i|\tilde{P}_i)p(F_k|P_k). \qquad [2]$$

Equivalently, for $p' = \log p$, as used in neural modeling (15):

$$p'(C, P, F) = p'(C) + \sum p'(P_i|\tilde{P}_i) + \sum p'(F_k|P_k). \qquad [3]$$

The P_i are parts in the hierarchy, $\tilde{P}_i$ is the parent of the part P_i (i.e., P_i is a subpart of $\tilde{P}_i$), and F_k are the input features (Fig. 2). Formally, this decomposition assumes that the nondescendents of a part P_i (all nodes other than its parent $\tilde{P}_i$ or descendents) are conditionally independent of P_i given $\tilde{P}_i$. Intuitively, it makes a "local context" assumption, namely, that the information about a part is captured by the subtree under the part together with the context supplied by the parent node.

The parameters needed for the computation: $p(C)$, $p(P_i|\tilde{P}_i)$, $p(F_k|P_k)$, are only single and pairwise probabilities that can be readily learned from observed frequencies in the training data and stored as synaptic strength in the network (15). Given a set of observed features F_k, the hierarchy is then used to determine the most likely assignment of the object and its parts at all levels.

Author contributions: S.U. designed research; B.E., I.L., and S.U. performed research; and S.U. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*Present address: Microsoft Research, Redmond, WA 98052.

†To whom correspondence should be addressed. E-mail: shimon.ullman@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/cgi/content/full/0800968105/DCSupplemental.

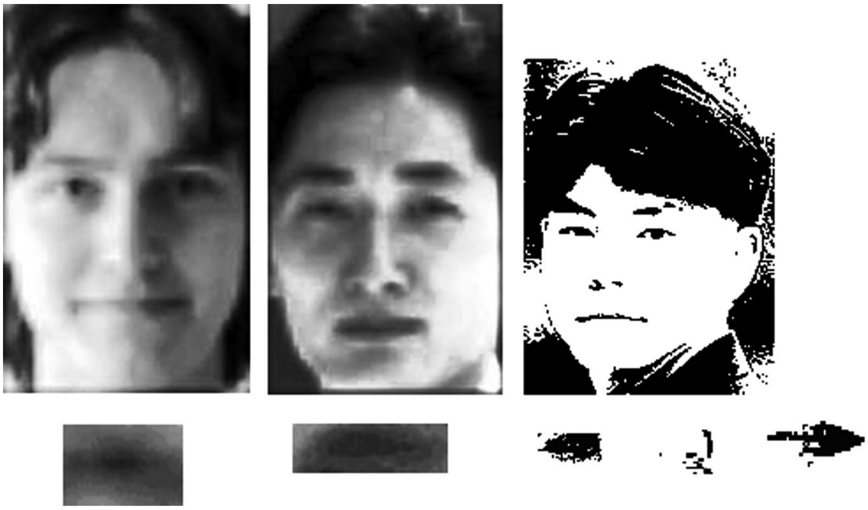

Fig. 1. Objects parts can be easily identified in the context of complete objects but become ambiguous on their own. Example images (*Upper*) and parts taken from these and similar examples (*Lower*). Such images are interpreted correctly by the BU TD cycle.

Recognizing Objects and Parts. Similar to a previous cortical model (3), the computation can be viewed as alternating stages of computing a maximal value and a summation. Unlike previous models, which compute for each part P_i a single maximal response, the current feed-forward stage finds a set of optimal values for each part, one for every possible value of the higher-level part (*Methods* and *Classification and Part Detection* in *SI Appendix*). The final selection of optimal values for all of the parts (P_i) is obtained only during the TD pass, by combining the BU and TD signals. A part that was only weakly activated during the BU pass can become activated by its selected parent, and conversely, a strongly activated part can become inhibited. In this manner, the TD pass explicitly identifies the components at all levels that constitute the object recognized at the top level. The full algorithm is a simplified version of the so-called Factor Graph or GDL computation (*Methods* and *Classification and Part Detection* in *SI Appendix*): This is a distributed process composed of local, parallel computations that was shown to be

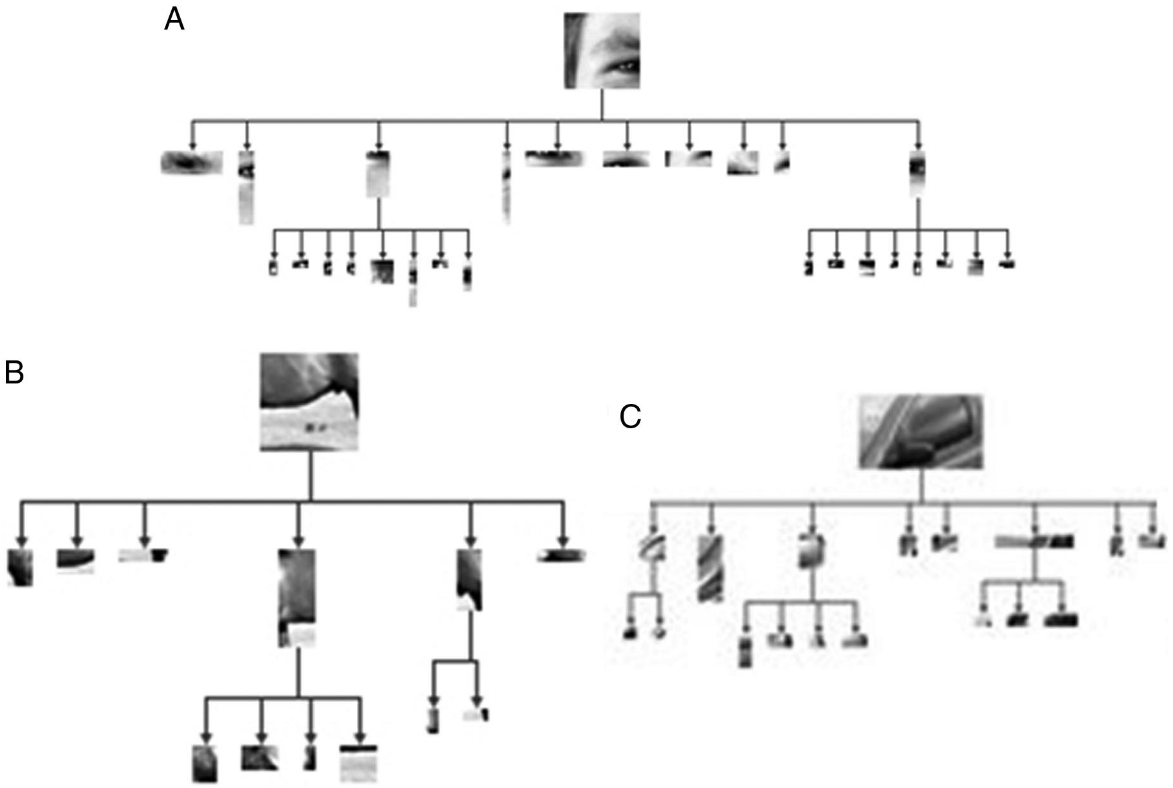

Fig. 2. An object is represented by a feature hierarchy; pieces of the feature hierarchies, extracted automatically from example images, are shown for faces (*A*), horses (*B*), and cars (*C*). The bottom layer contains the input features, which are detected in the image during recognition. Additional examples are in Fig. 3 in *SI Appendix*.

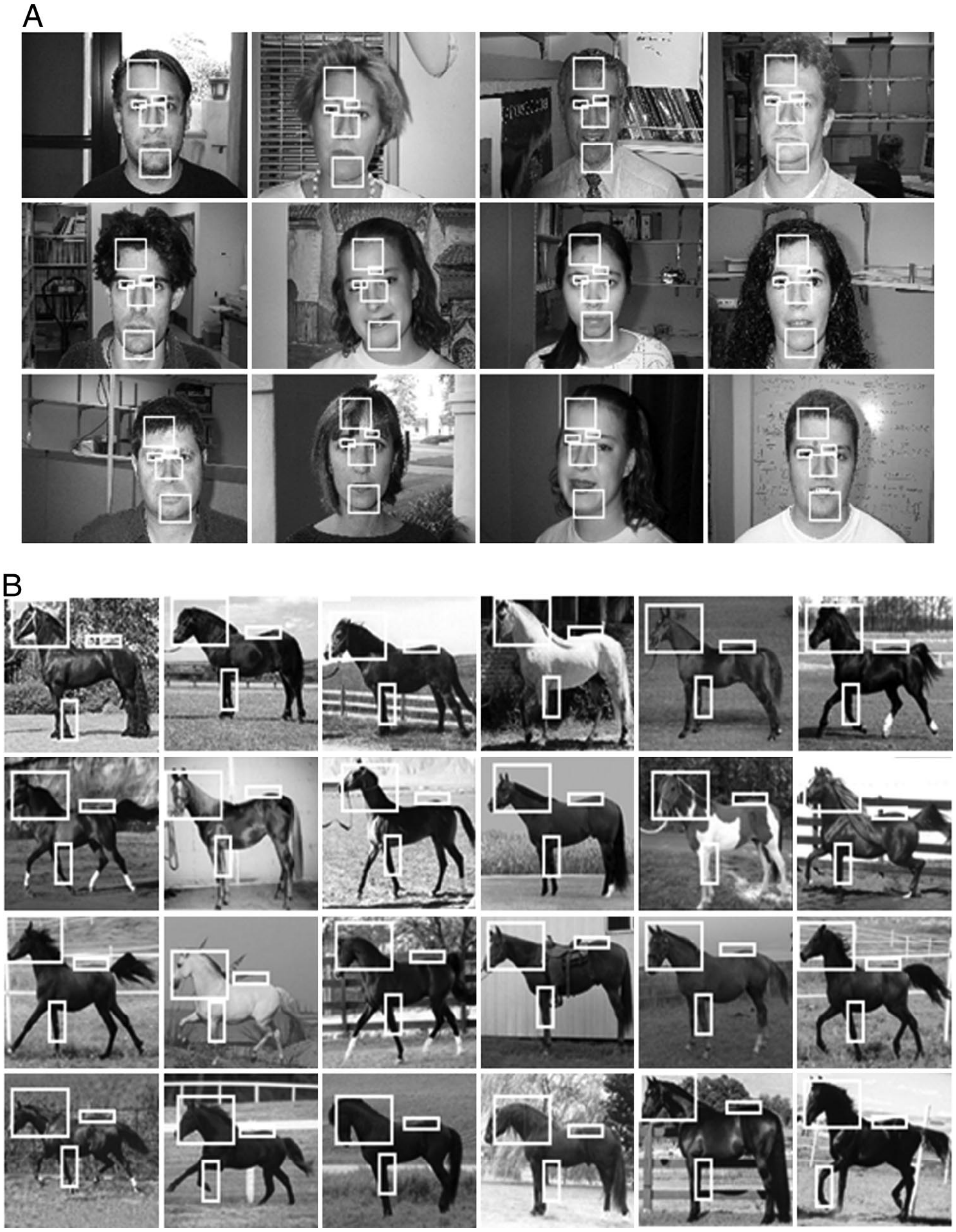

Fig. 3. (*Figure continues on the opposite page.*)

highly efficient for inference and optimization in hierarchical networks (16, 17). Possible implementations of such computations in neural models have been recently proposed (18–21).

Testing Results. Testings of the model compared the results of part interpretation by standard feed-forward models of cortical processing with the full BU–TD cycle. The goals were, first, to test the capacity of the feed-forward model to detect and localize parts: Feed-forward models were shown to produce good recognition at the object level (e.g., refs. 14 and 22), but their capacity to detect and localize parts and subparts remained unclear; second, to evaluate the improvements in part interpretation obtained by a single additional TD pass. The testing included the extraction of part hierarchies from natural images, learning the network parameters, and applying the bidirectional recognition process to test images. The results show that objects together with all their parts at multiple levels were identified and accurately localized by a single BU–TD cycle (Fig. 3 *A* and *B*), although the parts were locally ambiguous and with high variability in their appearance (Fig. 3*C*). The bidirectional scheme not only classifies the object but also identifies and localizes multilevel parts. The set of all of the detected parts covers most of the object area (Fig. 3*D*); taken together, they provide a detailed interpretation of the entire figure at multiple levels of detail. This illustrates the difference between feed-forward classification and full-image interpretation obtained by the BU–TD cycle.

The results show that the BU pass makes frequent errors at the part levels that are corrected by the TD pass. To evaluate the disagreement between the BU and TD phases, we compared the detection of each feature *E* by the BU and TD passes. For the BU detection, the part was detected by the subtree under *E*, with the detection threshold set to minimize the overall detection error. The two passes disagree if the part is detected by only one of them or if the detected locations differed by >0.1 object size. The average disagreement rate in class images was 20.8%,

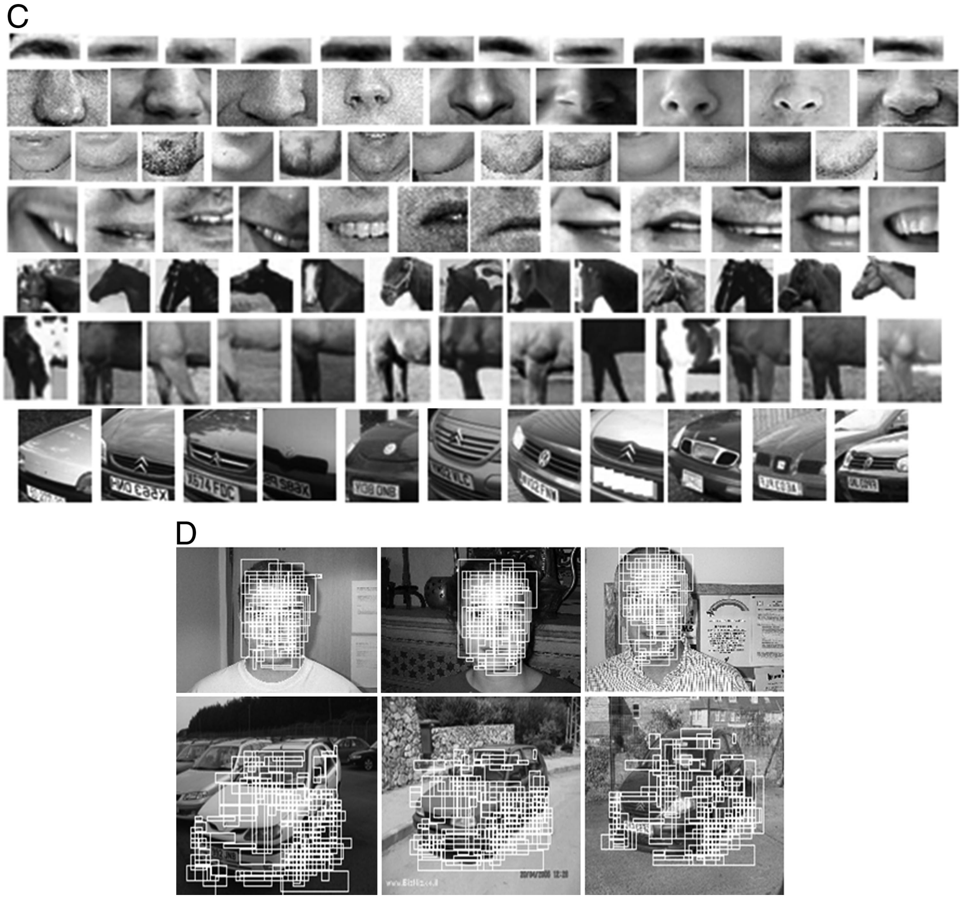

Fig. 3. Examples of parts detected by the BU TD cycle. (*A*) Detected locations of five example face-parts (hairline, right eye, left eyebrow, nose, chin); many more parts were detected by the hierarchy. (*B*) Examples of three detected horse parts (head, back, leg). Each detected part is marked by a rectangle at the detected location. (*C*) Collections of the same object parts (for face, horse, cars) detected in different images, showing the large variability in appearance. (*D*) Full interpretation: Outline rectangles mark detected parts; the collection of detected parts densely covers the entire object image. Additional examples in Figs. 4, 5, and 6 in *SI Appendix*.

42.1%, and 47.7%, at levels 1, 2, and 3–4, respectively. In the complementary set of nonclass images, object parts are often erroneously detected by the BU pass, with average detection probability 12.7%, 27.1%, and 30.8%, for parts at levels 1, 2, and 3–4, respectively. This was much higher than the TD pass (0.99%, 1.1%, and 1.8%, respectively).

We next examined the correctness of part detection to evaluate the error frequencies made by the BU and TD passes. Correctness of the BU–TD detection was verified for 12 parts in the hierarchy (three from each of four levels) in 10 groups of 38 images each, by human judgment. On this subset, overall disagreement was 23.2%. With rare exceptions ($<$0.1%), when the TD and BU agreed, their assignment was correct; most disagreements stemmed from BU errors that were corrected by the TD pass. For 18.3% (SD 1.5) of the detected parts, the TD detection was correct and BU incorrect, for 4.1% (SD 1.5), both were incorrect, and for 0.8% (SD 0.9), the TD was incorrect and the BU correct. Some parts had much higher BU error rates ($>$50%) but were still corrected by the TD pass. BU error rates in parts recognition also increase with increased noise, shadows, illumination changes, variability in the dataset, and more complex background, but TD interpretation remains robust, and BU errors are corrected effectively by the TD process (Fig. 4*b* in *SI Appendix*).

Disagreement and error rates decrease on average at higher levels because the false detection probability of the BU process decreases with part complexity, reaching, at the top level, an average error rate of ≈3% (measured at the equal error rate point) in the best reported results (22) and in the current model. The BU pass is therefore often sufficient for top-level recognition (14, 22–25), but the detection of object parts is highly unreliable without the use of disambiguating context. In contrast, correct and detailed interpretation covering the entire object is obtained by the BU–TD cycle.

Discussion

The recognition and localization of parts is important for perception and action, for example, grasping and manipulating objects, reading an instrument, identifying facial expressions, perceiving the pose of an animal, and the like. The current model shows how the detection and localization of parts at multiple levels are obtained with high accuracy and nearly simultaneously by a BU followed by TD sweep in a cortical-like hierarchy. Consistent with empirical observations, object recognition can either precede or follow parts recognition, and both are recognized within a BU–TD cycle. The two-pass model can account for human's ability to reliably recognize objects and parts, and it is consistent with timing considerations, requiring that both objects and parts be recognized quickly and at comparable speeds. A similar process could also exploit scene context to disambiguate objects in complex images.

We found in additional testing that adding lateral connections between features in the hierarchy together with applying additional iterations can play a useful role (26), e.g., in discriminating between closely similar objects and allowing improved recognition and finer discrimination at the cost of additional processing time. For example, to perform fine discrimination, it is often useful to obtain basic-level categorization by the first BU–TD cycle, and obtain a finer discrimination, based on additional features by using a second cycle (27). However, most of the part disambiguation is obtained already by the first cycle, and this can explain how objects and parts are typically recognized at similar speeds. It is likely that the use of highly informative features contributes to this rapid disambiguation.

The reported results provide a lower bound on the disambiguation that can be obtained by the TD computation, under timing constraints restricted to a single cycle. The fast disambiguation is consistent with physiology and ERP data, and is probably crucial for natural vision. The single cycle is also often sufficient to obtain recognition of objects and parts across changes in viewing direction and illumination (28). More complex aspects of scene interpretation that deal with multiple objects and their configurations may require, however, more than a single cycle.

In the testing above, computation was initialized by the BU phase. The disambiguation can probably be accelerated and refined by TD expectation, e.g., by activating the TD process before the initiation of the BU phase, and by modifying some of the lower-level representations in a task-dependent manner (29). TD processing can also allocate visual attention to specific parts in the image for further processing (30) after the initial cycle.

The feed-forward sweep in the current model is different from standard cortical modeling (1–3) because it derives for each part multiple competing alternatives with their relative likelihoods rather than successively selecting the most likely choice at each level. Each alternative is maximal over all possible assignments of the subtree under the part, unlike the BU computation in standard feed-forward models. Subsequently, the selection of optimal values for all of the parts is obtained during the TD pass. The maximal detection during the BU pass can be overruled by the TD pass, and the results show that such switches happen frequently. To allow correct detection by the TD pass, the BU detection needs to represent competing alternatives not just the optimal detection by BU criterion only.

Biologically, the structure and construction of the hierarchy could be an extension of the comprehensive standard model or similar models (2, 3, 14), with appropriate additions for bidirectional processing.

Compared with past computational modeling, previous models have used part-based object recognition (31–33) and combined BU with TD processing (18, 19, 34–36). However, past models did not study or report results on part recognition, did not examine the limitations of feed-forward models for part recognition, and did not demonstrate the contribution of a fast TD process to part detection and localization.

The results demonstrate two related aspects regarding cortical models and full object and parts recognition. First, they show the limitations of the standard feed-forward processing for part detection and localization. Previous results have shown the capacity of feed-forward models to deal with top-level object recognition (2, 3, 24, 25) but have not examined their capacity and limitations for part recognition. The current results show that even when a feed-forward process is sufficient for top-level object recognition, it is still often inadequate for parts recognition. Second, they show how a bidirectional process in a network that stores part/subpart probabilities between informative features is sufficient to obtain a combined recognition of objects and parts at multiple levels within the time constraints established by empirical studies.

The current scheme predicts that unambiguous parts that do not require additional context are identified already in the BU pass, but more ambiguous parts are resolved only during the TD pass. In the primate ventral visual stream, neurons show highly selective shape tuning already in the earliest part of their response (25, 37, 38), but the selectivity often increases within a short latency, consistent with rapid disambiguation by TD processing (8–11). We predict that the modified late response will be observed more frequently with ambiguous stimuli (e.g., blurred, noisy, partially occluded), because a larger proportion of the BU responses will be corrected by the TD signal. Because lower-level parts in the hierarchy are more ambiguous locally, we also predict that the presence of a late response with different response properties than the early response will be more frequent at the lower levels of the visual hierarchy compared with higher levels, and the model could be used to predict the ambiguity of different parts and their dependence on TD disambiguation. Because of the successive disambiguation from top to bottom, we predict that the time difference between the early and late responses will be, on average, larger at lower levels of the hierarchy.

Also, features required for fine discrimination between similar objects, e.g., facial expression, or similar individuals are often locally ambiguous. In such cases, we expect from the model that recognition at the fine level may be delayed, until the local features reach final disambiguation by the full BU–TD cycle (39). Predictions could be tested physiologically, and possibly also psychophysically, e.g., by adapting recent methods for tracking the dynamics of feature analysis during recognition (40).

The full role of the TD processing in vision is still unclear, and it is likely to participate in multiple processes (29). It has been hypothesized to play a role in several tasks, in particular figure–background segregation and grouping (41, 42), learning (43), controlling attention (30, 44), and explicit perception (45). The current study, together with empirical data as well as computational modeling (18, 19, 46), suggests that descending visual processing is used to obtain full-image interpretation by the nearly simultaneous detection and localization of object parts at multiple levels.

Methods

Image Sets. Training images for parts extraction contained 208 faces, 161 horses, 175 car gray-level images from Caltech dataset (www.vision.caltech.edu/html-files/archive.html), 120–210 pixels in each dimension. Examples of images and extracted features can be found in (Fig. 1 in *SI Appendix*); this material contains additional details on the feature extraction procedure, model, and results.

Parts Hierarchy. First level object fragments were extracted by using the procedure in ref. 12. Briefly, the process identifies fragments that deliver the maximal amount of information about the class. Candidate fragments are extracted from the training images at multiple locations and sizes. These fragments are searched for in all of the database images by using normalized correlation, and the mutual information $I(C;F)$ is computed. Detection thresholds are determined automatically for each fragment at a level that maximizes the delivered mutual information. From the initial set, a subset of the most informative fragments is selected successively (12). Features are extracted for different recognition tasks and levels of specificity by the same feature-extraction process but evaluating the information they deliver for the particular recognition task (27, 28). Hierarchical features were extracted by using the procedure in ref. 13. This process represents each object part by its own informative subparts. The entire hierarchy is extracted from the image examples by a repeated application of the same information-maximization process used for the initial extraction of informative object components. If the decomposition of fragment F into simpler features does not increase the delivered information, the decomposition terminates, and F is considered a low-level feature (Fig. 2). Additional details of the algorithms are in *Part Hierarchy* in *SI Appendix*.

Training the Model. Each part P_i in the hierarchy is considered as a variable with $n + 1$ possible values; $P_i = 0$ means that the part is not present in the image, $P_i = j$ means that P_i is present at location j (n up to 1,250). Training consists of estimating the probabilities $p(P_i|\tilde{P}_i)$ (where P_i is a part and $\tilde{P}_i$ its parent-part in the hierarchy), including $p(F_k|P_k)$, where F_k is one of the observed low-level features. These values are then used for the interpretation of novel images by using Eqs. **1** and **2** above.

To determine $p(F_k|P_k)$, the low-level feature F_k is correlated with the image, to obtain correlation values $\mathrm{Corr}_1, \ldots, \mathrm{Corr}_N$ at positions 1. . . N in the image. These are the observed values for the computation, and all pairwise probabilities in the model are estimated from them by using the standard EM algorithm, commonly used to estimate model parameters in network models. Briefly, the EM process iteratively adjusts the model parameters so as to maximize the likelihood of the observed data. A full description of the model parameter learning in found in *Training the Model* in *SI Appendix*).

Classification and Part Detection. The optimal decision at all levels is obtained efficiently by a BU flow of activation from the simple features to the class node, followed by a TD flow from the high to low levels in the hierarchy. The computation is a simplified application of the standard Factor Graphs (16) or GDL (17) method, which is related to belief propagation. This is a highly efficient optimization process that can be implemented in a network of simple interacting elements (15, 18–21) (*Classification and Part Detection* in *SI Appendix*).

Briefly, the computation proceeds by producing a set of tentative optimal values on the way up starting from the lowest-level features, and selecting the final globally optimal decision on the way down. For a given part P_i in the hierarchy, the feed-forward flow determines k optimal values, one for each possible value of $\tilde{P}_i$ (P_i's parent node). The optimal value of the parent is still unknown, but once it becomes known in the TD pass, the optimal value of P_i itself is immediately determined. At the top level, which has no parent node, the best value can be determined from the preceding level. The TD sweep proceeds to make the final selection at each level based on the level above it, propagating down to the bottom of the hierarchy.

Testing. Classification and part-detection experiments were performed on a new set of 218 face, 161 horse, 175 car, and 1,617 nonclass images. Testing was applied to novel class and nonclass images. In the class images, we estimated how often part assignment by the BU pass is overruled by the TD pass. For each part E detected in the image ($E \neq 0$ by TD pass) we compared the TD detected location E_D with the location E_U of maximal BU probability. If $|E_D - E_U| > 0.1$ object size, we consider the BU and TD to disagree. The BU process often makes multiple erroneous detections; therefore the measure gives a conservative estimate of the BU–TD disagreement. The accuracy of the final TD assignment was verified by human judgments. Additional details are in *Testing and Experimental Results* in *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the Information Society Technology Grant FP6-2005-015803 (Daisy) and Israel Science Foundation Grant 7-0369.

1. Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cyber* 36:193–202.
2. Rolls ET, Milward T (2000) A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput* 12:2547–2572.
3. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025.
4. Kimchi R (1992) Primacy of wholistic processing and global/local paradigm: A critical review. *Psychol Bull* 112:24–38.
5. Tarr M-J, Bulthoff HHB (1998) Image-based object recognition in man, monkey and machine. *Cognition* 67:1–20.
6. Bentin S, Golland Y, Flevaris A, Robertson LC, Moscovitch M (2006) Processing trees and the forest during initial stages of face perception: Electrophysiological evidence. *J Cognit Neurosci* 18:1406–1421.
7. Liu J, Harris A, Kanwisher N (2002) Stages of processing in face perception: An MEG study. *Nat Neurosci* 5:910–916.
8. Sugase Y, Yamane S, Ueno S, Kawano K (1999) Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400:869–873.
9. Tsao DY, Freiwald WA, Tootell RBH, Livingstone MS (2006) A cortical region consisting entirely of face-selective cells. *Science* 311:670–674.
10. Hegdé J, Van Essen DC (2004) Temporal dynamics of shape analysis in macaque visual area V2. *J Neurophysiol* 92:3030–3042.
11. Rainer G, Lee H, Logothetis N (2004) The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biol* 2:275–283.
12. Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. *Nat Neurosci* 5:682–687.
13. Epshtein B, Ullman S (2005) Hierarchical features for object classification. *IEEE Proc ICCV* 1:220–227.
14. Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci USA* 104:6424–6429.
15. Rao RPN (2004) Bayesian computation in recurrent cortical circuits. *Neural Comput* 16:1–38.
16. Kschischang FR, Frey BJ, Loeliger HA (2001) Factor graphs and the sum-product algorithm. *IEEE Proc Info Theor* 47:498–519.
17. Aji SM, McEliece RJ (2000) The generalized distributive law. *IEEE Trans Info Theor* 46:325–342.
18. Friston K (2002) Functional integration and inference in the brain. *Prog Neurobiol* 68:113–143.
19. Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A* 20:1434–1448.
20. Deneve S (2005) Bayesian inference in spiking neurons. *Adv Neural Info Process Syst* 17:353–360.
21. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9:1432–1438.
22. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intel PAMI* 29:411–426.
23. Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. *Nature* 381:520–522.
24. VanRullen R, Koch C (2003) Visual selective behavior can be triggered by a feed-forward process. *J Cognit Neurosci* 15:209–217.
25. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–866.
26. Tsodyks M, Gilbert CD (2004) Neural networks and perceptual learning. *Nature* 43:775–781.
27. Epshtein B, Ullman S (2006) Satellite features for the classification of visually similar classes. *Proc IEEE CVPR* 2:2079–2086.
28. Ullman S (2006). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cognit Sci* 11:58–64.
29. Gilbert CD, Sigman M (2007) Brain states: Top-down influences in sensory processing. *Neuron* 54:677–696.
30. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18:193–222.
31. Pope AR, Lowe DG (2000) Probabilistic models of appearance for 3-D object recognition. *Int J Comput Vis* 40:149–167.
32. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-Invariant learning. *Proc IEEE CVPR* 2:264–271.
33. Crandall D, Felzenszwalb P, Huttenlocher DP (2005) Spatial priors for part-based recognition using statistical models. *Proc IEEE CVPR* 1:10–17.
34. Geman S (2006) Invariance and selectivity in the ventral visual pathway. *J Physiol (Paris)* 100:212–224.
35. Jin Y, Geman S (2006) Context and hierarchy in a probabilistic image model. Object class recognition by unsupervised scale-invariant learning. *Proc IEEE CVPR* 2:2145–2152.
36. Tu ZW, Chen X, Yuille AL, Zhu SC (2005) Image parsing: Unifying segmentation, detection, and recognition. *Int J Comput Vis* 63:113–140.
37. Oram MW, Perrett DI (1992) Time course of neural responses discriminating different views of the face and head. *J Neurophysiol* 68:70–84.
38. Tovee MJ, Rolls ET, Treves A, Bellis RP (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. *J Neurophysiol* 70:640–654.
39. Grill-Spector K, Kanwisher N (2005) Visual recognition: As soon as you know it is there, you know what it is. *Psychol Sci* 16:152–160.
40. Schyns PG, Petro LS, Smith ML (2007) Dynamics of visual information integration in the brain for categorizing facial expressions. *Curr Biol* 17:1580–1585.
41. Hupé JM, *et al.* (1998) Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* 394:784–787.
42. Lamme VAF, Roelfsma PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci* 23:571–579.
43. Spratling MW, Johnson MHA (2006) Feedback model of perceptual learning and categorisation. *Vis Cognit* 13:129–165.
44. Hahnloser RHR, Douglas RJ, Hepp K (2002) Attentional recruitment of inter-areal recurrent networks for selective gain control. *Neural Comput* 14:1669–1689.
45. Hochstein S, Ahissar M (2002) View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36:791–804.
46. Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. *Annu Rev Psychol* 55:271–304.