Opinion TRENDS in Cognitive Sciences Vol.11 No.2

Object recognition and segmentation by a fragment-based hierarchy

Shimon Ullman

Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel

How do we learn to recognize visual categories, such as dogs and cats? Somehow, the brain uses limited variable examples to extract the essential characteristics of new visual categories. Here, I describe an approach to category learning and recognition that is based on recent computational advances. In this approach, objects are represented by a hierarchy of fragments that are extracted during learning from observed examples. The fragments are class-specific features and are selected to deliver a high amount of information for categorization. The same fragments hierarchy is then used for general categorization, individual object recognition and object-parts identification. Recognition is also combined with object segmentation, using stored fragments, to provide a top-down process that delineates object boundaries in complex cluttered scenes. The approach is computationally effective and provides a possible framework for categorization, recognition and segmentation in human vision.

Features for visual recognition

Categorization and recognition are basic aspects of human vision and cognition. Visual categorization (or visual classification) refers to the perception of an object as belonging to a general class, such as a dog or a building. Individual recognition is the identification of different images as depicting the same object, such as a specific face or a car, despite changes in the viewing conditions. Categorization and recognition are performed by the human brain in a natural, effortless manner and with an efficiency that is difficult to reproduce in computational models and artificial systems. General categorization has proved to be particularly elusive: how do humans, even young children, learn to distinguish between categories, such as dogs and cats, from a limited set of highly variable instances?

Typically, a recognition scheme extracts a set of features from an image and uses them during a learning phase to construct new object representations. Objects are then classified and recognized based on their feature representation [1]. Feature selection and object representation are crucial aspects of recognition: they facilitate the identification of aspects that are shared by objects in the same class, despite variability in appearance, and they support discrimination between objects and between classes that can be highly similar.

Corresponding author: Ullman, S. (shimon.ullman@weizmann.ac.il).
Available online 22 December 2006.

Different types of visual features have been used in computational models in the past, ranging from simple local-image patterns such as wavelets, Gabor filters, edges, blobs or local-edge combinations [2–4], to abstract three-dimensional shape primitives, such as so-called geons (basic geometric shapes – for example, spheres, cubes and cylinders) [5]. A common aspect of most previous features is that they are generic – a small fixed set of feature types are used to represent all objects and classes. By contrast, in the approach proposed in this Opinion article, categorization is based on representing shapes within a class by a combination of shared sub-structures called fragments [6–8].

The fragments are learned from image examples and are used as building blocks to represent object views within a given class of shapes. There are two main aspects to using these features for classification. First, unlike generic parts, these are class-specific features: for each class of objects, the appropriate visual elements are extracted and are used to distinguish objects within the class from objects in different classes. Second, the fragments are pictorial features that represent the image appearance of object components, unlike, for example, view-independent three-dimensional primitives [5,9]. The use of local pictorial features reflects the assumption that images of different objects within a visual class can be represented by similar arrangements of common sub-structures. Pictorial, informative class features have been used in recent models, and computational experimentations have demonstrated their effectiveness in making reliable categorizations of natural object classes. Next, we turn to the problem of selecting common sub-structures from examples in a manner that enables the handling of intra-class variability and the generalization to new class exemplars.

Informative class fragments

To distinguish class from non-class objects, useful features should have two main properties: distinction and frequency. For example, for face images, a fragment (F) is an effective class feature if it is likely to be found in face images but not in non-face images. These two requirements can be combined by measuring the amount of information that is supplied by the fragment about the class in question. A feature is informative if it reduces the uncertainty about the class – that is, its presence in the image increases the likelihood of the class and the likelihood decreases if the feature is absent. The difference in uncertainty with and without the use of F is defined as the information (I) that is supplied by F about the class (C; Box 1). Mathematically, an increase in the

 1364-6613/$ – see front matter doi:10.1016/j.tics.2006.11.009

Box 1. Measuring feature information

The entropy $H(X)$ measures the uncertainty of a variable X. For a discrete variable with values $x_1 \ldots x_n$, with probabilities $p(x_1) \ldots p(x_n)$, the entropy is given by the sum $-\Sigma p(x_i) \log_2 [p(x_i)]$, which was found to be a useful measure of uncertainty in information theory. For example, if X has a uniform distribution, then the uncertainty is high and so is the entropy, but if the value of X is known with high likelihood, the entropy is low. To measure the information that is delivered by a feature (F) to the classification of a class (C), the entropy $H(C)$ is computed twice, before and after the detection of F. The expected difference between the two computations is defined as the information (I) about C that is supplied by F. This is expressed in mathematical form as $I(C;F) = H(C) - H(C|F)$. In this equation, $I(C;F)$ is the information that is supplied by F about C, $H(C)$ is the initial uncertainty and $H(C|F)$ is the uncertainty when the value of F is already known. If F is a useful feature, the uncertainty about C will decrease significantly when it is known whether or not F is present in the image, reflecting the high information contribution of F.

For example, during training that consists of 100 face and 100 non-face images, F is detected 44 times in the face examples and six times in the non-face examples. Initially, $H(C) = 1.0$. In this example, the second entropy measure, $H(C|F)$, is 0.847. The average reduction in H is 0.153, and this is the information that is supplied by F.

delivered information is required to reduce the classification error. The amount of information that is supplied by F can be estimated easily from image examples: it is determined by how frequently F is detected within and outside C. Next, we describe how informative class fragments can be extracted.

Selecting informative class fragments

The principle of maximizing information for classification can be used to extract automatically a set of highly informative features from image examples, as illustrated by the process described in this section. The feature-extraction process uses both class and non-class examples; for instance, images that contain examples of the class 'horse' are used with images that do not contain horses. It is not necessary to indicate where the object is located in a given image and each image can also contain other objects. Initially, the process extracts a large number of candidate fragments at multiple positions and of multiple sizes and scale. The amount of information that each candidate fragment delivers about the class is estimated by detecting its frequency within and outside the class examples. Thus, the most informative class fragments are identified. The information that is carried by individual features is not a sufficient criterion for selection because of possible redundancy: two features can be highly informative on their own but they can also be almost identical and, therefore, redundant. Excessive redundancy is avoided by a second selection stage [8]. Fragments can be selected successively and, at each stage, the fragment that contributes the most additional information is added to the set of selected fragments. For more on the selection process, see Box 2. In a biological system, a simplified approximation to the information measure can be used, with only a minor decrease in performance (D. Levi, MSc thesis, The Weizmann Institute of Science, 2004; see http://www.wisdom.weizmann.ac.il/~danml/).

Examples of informative fragments that have been extracted for several visual classes are shown in Figure 1. The most informative features that are found for different categorization tasks are typically of intermediate complexity, including intermediate size at high resolution and larger size at intermediate resolution [8,10]. This is in contrast to visual features that are used in many previous approaches that have focused on the ends of the spectrum, either small, local features [2,3] or global features [11]. To distinguish between closely similar classes, and even individual objects, a similar process is applied, but features are evaluated by the information that they deliver for fine discrimination. This stage can extract small, local features that support the required discrimination [12] (A. Akselrod-Ballin, MSc thesis, The Weizmann Institute of Science, 2002; see http://www.wisdom.weizmann.ac.il/~shimon/students.html).

Box 2. Extracting informative fragments

The following procedure automatically selects highly informative class features from a set of class and non-class image examples (Figure 1). First, candidate fragments that have been extracted from the class images are considered, at different positions, sizes and resolutions. The information that is supplied by each candidate feature is estimated by detecting it in the training images. To detect a given fragment (F) in an image, F is searched by correlating it with the image. (Alternative similarity measures that incorporate color, texture and 3D cues can also be used.) If the similarity at any location exceeds a certain threshold (θ), then F has been detected in the image and $F = 1$; if θ is not exceeded, then $F = 0$. A binary variable $C(I)$ is used to represent the class – namely, $C(I) = 1$ if the image (I) contains a class example; if I does not contain a class example, then $C(I) = 0$. Second, the amount of information that each candidate fragment delivers about the class is estimated based on its detection frequency within and outside the class examples. The delivered information is a function of the detection threshold; therefore, the threshold for each fragment is adjusted individually to maximize the information $I(F;C)$. Third, the features are considered in the order of the information that they supply. To avoid redundancy among similar features, fragments are selected successively and, at each stage, the fragment that contributes the most additional information is added to the set of selected fragments [8]. For additional details, see Ref. [8]; for the hierarchical construction, see Ref. [13].

Empirical comparisons have shown that simple, biologically plausible processes can be used to approximate the selection of informative fragments [18] (D. Levi, MSc thesis, The Weizmann Institute of Science, 2004; see http://www.wisdom.weizmann.ac.il/~danml). In addition, although labeling training images as class or non-class facilitates learning, the selection process can be applied with some modifications without this labeling – that is, in an unsupervised manner.

Feature hierarchies

The representation of a visual class by informative components is useful for dealing with the variability in appearance of objects in a class. However, these components, like the objects themselves, can vary considerably in their appearance. Therefore, using the same principles as those discussed above, it is natural to decompose the object components into informative sub-parts. A repeated application of the feature-extraction process results in a hierarchical object representation of informative parts and sub-parts at multiple levels. Such hierarchical decomposition makes a substantial contribution to the performance and generalization capacity of the categorization scheme [13].

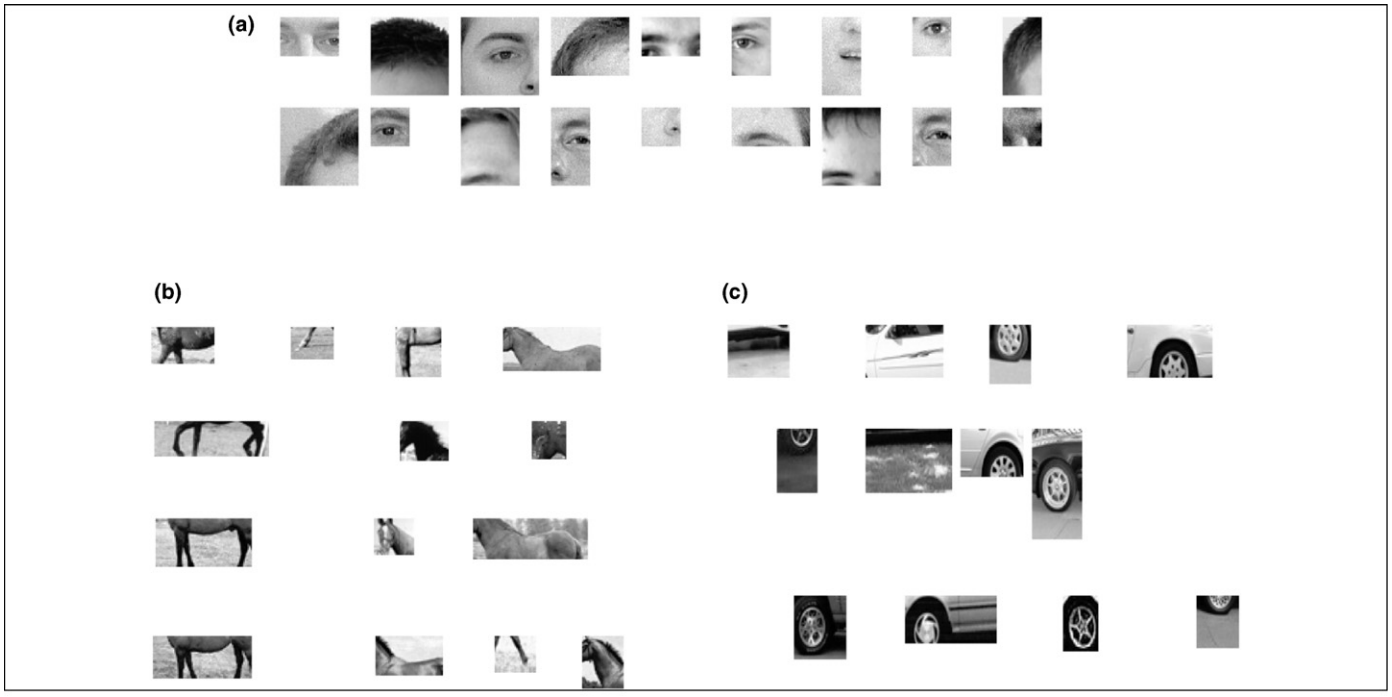

Figure 1. Highly informative class fragments for faces **(a)**, horses **(b)** and cars **(c)**.

Another fundamental advantage of the hierarchical representation is that it can be used for recognizing not only complete objects but also their parts. Computational experiments have shown that parts and sub-parts at multiple levels can be identified unambiguously by an efficient computation that combines bottom-up and top-down processes that are applied to the feature hierarchy (I. Lifshitz, MSc thesis, The Weizmann Institute of Science, 2005; see http://www.wisdom.weizmann.ac.il/~shimon/students.html).

Examples of hierarchical features for several object classes are shown in Figure 2. The full hierarchies were obtained automatically from image examples by a repeated application of the same feature-extraction process as that described above for the initial extraction of informative classification features [13]. In a biological implementation, such feature hierarchies can be imprinted initially from class examples and then refined with additional training. The lowest level of the hierarchy is composed of simple

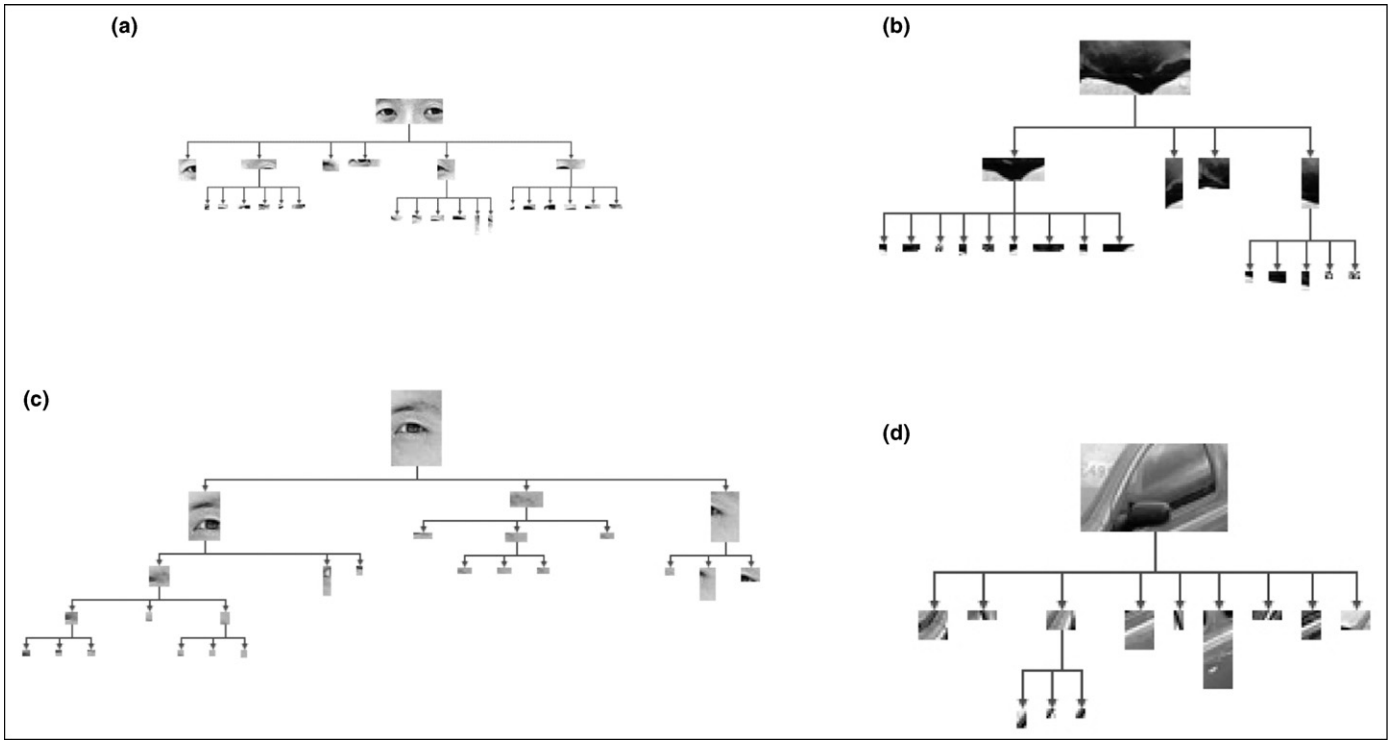

Figure 2. Features hierarchies from faces **(a,c)**, horses **(b)**, and cars **(d)**. Features at the top of the hierarchy are informative class fragments and features at the bottom are low-level atomic fragments. The hierarchy also encodes the relative positions between parts and sub-parts.

'atomic' fragments, which typically contain edges, lines or corners and cannot be decomposed further without losing information. During classification, only the atomic features are directly correlated with the input image, and their responses are combined to detect the higher level in the hierarchy.

For each fragment, a region of positional tolerance is determined, which can be considered as the receptive field (RF) of the feature. The amount of information that is delivered by a fragment depends on its RF location and size, which are learned adaptively during construction of the hierarchy. A comparison of hierarchies that are obtained from multiple object classes show that the lowest-level features are generic (i.e. they are useful for all natural objects), the highest-level features are specific to a class of visually similar objects (unlike generic hierarchies, such as those found in Ref. [4]) and intermediate features are shared by similar classes. This feature sharing promotes effective 'cross generalization' from one class of objects to related classes [14,15]. For example, after learning the appearance of a component such as a leg or tail for one class, the resulting representation will be used by other classes that share the same component.

In this article, I do not consider a specific neural model for constructing the hierarchies. However, network models have been used to optimize the information that is delivered by neuronal activity [16,17], and network models that incorporate the selection of informative fragments have also been described recently [18] (D. Levi, MSc thesis, The Weizmann Institute of Science, 2004; see http://www.wisdom.weizmann.ac.il/~danml/).

Abstract fragments for classification and recognition

The hierarchical representation that is described above can compensate effectively for local changes and distortions in the image. However, object components can also have multiple different appearances due to large changes in viewing conditions, such as view direction or shadows, or as a result of transformation of the component, such as an open versus closed mouth in face images. To deal with multiple appearances, it is natural to group together different appearances of the same component to create a higher-level, more abstract representation. Two plausible mechanisms can be used for identifying equivalent fragments that represent the same object components across changes in appearance. The first abstraction mechanism involves observing objects in motion: based on spatiotemporal continuity, a fragment can be tracked over time, and different appearances of the changing part can be grouped together. A neural model for association by spatiotemporal continuity (the 'trace' model) is described in Refs [19,20]; the extraction and use of motion-grouped fragments for invariant recognition is described in Refs [20–24].

The second abstraction mechanism is based on common context. If two fragments are interchangeable within a common context, they are likely to be semantically equivalent. For example, if multiple instances of the same face are observed with either a neutral or a smiling mouth, this provides evidence for the equivalence between the two appearances. A constellation of image fragments that co-occur with a fragment (F) supplies a context for F. If F can be replaced by another fragment (F') within a particular context, then F and F' are likely to represent two appearances of the same part [25]. Figure 3 shows examples of top-level abstract fragments – these are equivalence sets of fragments that depict different appearances of the same object part. All were obtained automatically using common-context abstraction. After abstraction, the components in the hierarchy that is described above will use abstract features rather than single-appearance fragments. Because the hierarchies of related classes share common components in their representation, the learning of invariant properties will generalize from a learned class to related classes with shared features [14,15].

Abstract features are particularly useful for recognizing individual objects under large changes in the viewing conditions. Using the abstract fragments, each object part is represented by a set of fragments that depict this part under different conditions, such as viewing directions, illumination and changes in the part itself. An abstract fragment is present in an image if one of its fragments is activated by the image. Thus, the abstract fragments form a view-invariant representation of objects within a general class. In this theory, invariant recognition at the object level is based on the observed invariance of selected components [21–24]. This is in contrast to recognition models in which invariant recognition is primarily based

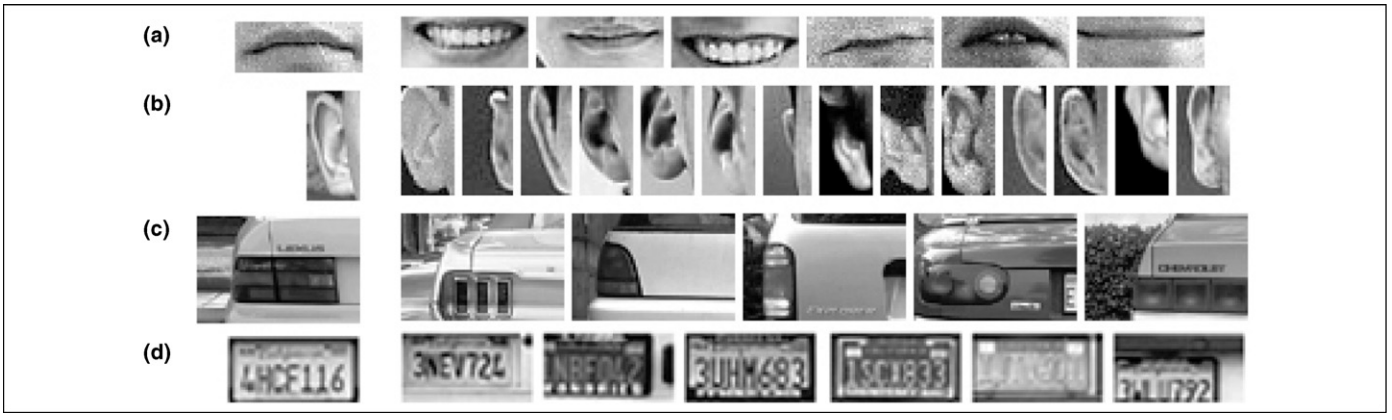

Figure 3. Abstract features that group together different appearances of the same object parts that have been extracted automatically from class examples [**(a)** mouths, **(b)** ears, **(c)** tail-lights and **(d)** license plates]. The features are semantically equivalent but highly variable in appearance.

on internal model manipulation, akin to mental rotation [26], or on the use of view-invariant features [5,9]. (However, more than one process could be involved.)

Combining segmentation and recognition

Figure–ground segmentation refers to the delineation of a region in an image as containing an object of interest. What is the relationship between segmentation and recognition? Segmentation has largely been viewed as a bottom-up process that precedes and facilitates recognition [27]. Bottom-up segmentation relies on the image-based criteria of 'Gestalt' and 'good continuation', such as color or texture uniformity of image regions, combined with the continuity of bounding contours. When applied to natural images, bottom-up segmentation is usually incomplete, due to unavoidable ambiguities that cannot be resolved without prior knowledge of the object class [28] (Figure 4).

Empirical evidence, reviewed in the next section, suggests that acquired knowledge about objects' shape has a crucial role in segmentation. But how can acquired knowledge of the shape of a highly variable class be used to segment a novel exemplar?

The hierarchical fragment representation enables acquired class-based information to guide the segmentation process and intimately integrate segmentation and recognition [29–32] (Figure 4). First, the figure–ground labeling of the stored fragments is learned from a collection of non-segmented image examples [30] (Figure 4b,c). Given a novel object from the learned class, the detected fragments produce a cover of the object in terms of stored fragments. The known figure–ground labeling of the fragments is then used to induce a top-down segmentation of the entire object (Figure 4d).

These top-down and traditional bottom-up segmentation processes have complementary advantages: the top-down process groups together image regions that belong to the same objects, despite region inhomogeneity and low-contrast boundaries, and the bottom-up process more accurately delineates the precise boundary locations. Therefore, the two processes can be combined naturally to obtain reliable and accurate segmentation of novel exemplars in cluttered scenes [32].

Perceptual and physiological implications

In summary, in the proposed approach, a hierarchy of abstract fragments that are continuously extracted from examples, based on delivered information and observed equivalence, combines classification, recognition and segmentation using a bi-directional interpretation process.

The main aspects of this approach are compatible with a substantial body of psychological and physiological evidence. Pertinent evidence is now reviewed, followed by predictions and questions for future studies. However, human object recognition is likely to use more than one process, and further studies are required to map the set of processes and the interactions that are involved.

The classification features that are used in the scheme are class specific, not generic, they are obtained through learning, based on their usefulness to classification, and they are pictorial, representing the image appearances of object components. Several psychophysical studies of humans and monkeys have shown that new features emerge in the visual system after categorization training [33,34]. Similarly, physiological studies support the view that class-specific features are acquired during the learning of new visual classes [35], with increased selectivity to features that support the classification task [36,37]. Several systematic studies of inferior temporal cortex (IT) neurons support the pictorial nature of these units, showing that the response selectivity of IT neurons indicates a

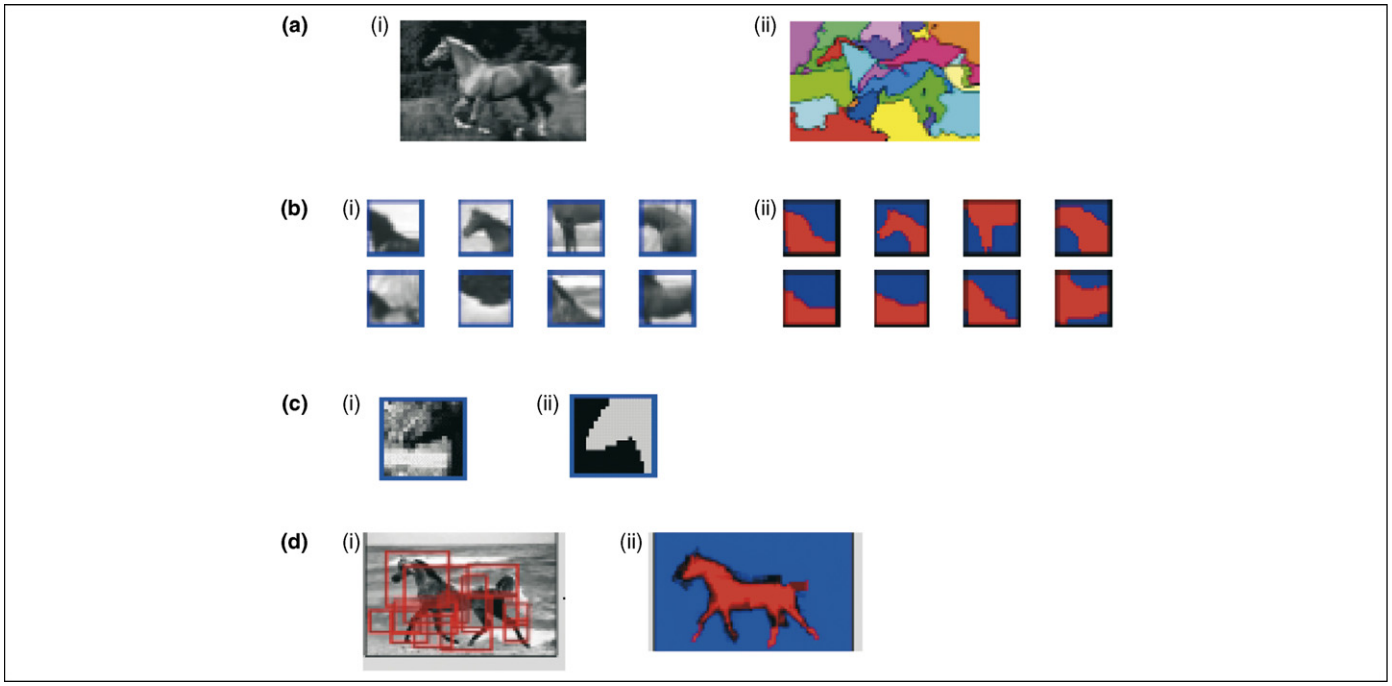

Figure 4. Top-down class-based segmentation. **(a)** Input image (i) and typical bottom-up segmentation (ii). **(b)** Informative class fragments (i) and their figure–ground segmentation (ii), which is obtained automatically from examples. **(c)** A single fragment (i) and its learned segmentation (ii). **(d)** An input image covered by class fragments (i). The class-fragments cover induces top-down segmentation (ii).

preference for specific image patterns or a configuration of patterns [38–40].

Regarding the hierarchical structure, maximizing information for recognition leads naturally to features hierarchies that have several basic properties in common with the known cortical hierarchy. Maximizing information for recognition produces a gradual increase in receptive fields at the higher level of the hierarchy, combined with a variable range of optimal sizes within each level; this is supported by empirical findings for receptive field sizes and variability [41]. For natural classes, the feature hierarchy has 3–5 useful levels. This is compatible with cortical structure and might relate to the statistical complexity of natural classes.

The use of abstract features in the model suggests that invariance is acquired for a class of objects from the observed equivalences of shared parts. Several psychophysical studies of recognition invariance support this notion. Studies of invariance to viewing direction [15] and position invariance [42,43] indicate that invariance to viewing conditions depends on training and that, following training, invariance is generalized across members in the training class. Evidence for linking relevant features by motion, as proposed by the notion of abstract fragments, comes from both psychophysical [44] and physiological studies [35]. The proposed mechanism of abstraction by common context remains open for future studies.

Psychophysical and physiological evidence indicates that in human and primate vision, figure–ground segmentation and object recognition proceed interactively and concurrently. Psychophysically, developmental performance [45], as well as adult performance [46,47], and fMRI evidence [48] suggest that acquired knowledge about objects' shape has a crucial role in segmentation. Evidence from neurophysiology shows that unit responses in low-level visual areas can depend on 'border ownership' [49] and on the overall figure–background relationships in the image [50].

The fragments hierarchy model raises predictions and questions for further empirical testing at the psychophysical, imaging and physiological levels. For example, in studies of human perception, the model suggests that the visual system uses fragment features that are informative for recognition. The relationship between this information measure of object fragments and human recognition performance can be tested in psychophysical studies. It is predicted that classification performance will increase with fragment-class mutual information. A similar prediction applies to brain imaging: the model predicts that there will be preferential activation in cortical object-related areas by informative, compared with less informative, object fragments. In addition, different levels in the hierarchy of object-related visual areas are expected to be preferentially activated by fragments that are extracted at different levels by the model hierarchy.

Physiological studies can examine whether tuning properties of units along the primate visual hierarchy in response to natural stimuli can be explained in terms of intermediate units in the abstract-fragments hierarchy, as described by the model. The lowest level of the computational hierarchy is composed of simple 'atomic' fragments, which typically contain edges, corners or lines. These features are similar to unit responses in the primary visual cortex (V1), but the model suggests that the set of features in V1 are richer than the standard model of this area.

Several basic questions and extensions remain open for future studies. For example, the biological implementation of informative fragment selection, and online learning: in a biological system, features must be continuously acquired from new examples, rather than using a fixed training period and limited training examples. Future extensions should also include the capacity to deal efficiently with a large number of classes and sub-classes and to distinguish reliably between closely similar classes. Two challenging goals for future studies are full-scene interpretation and dynamic recognition. Scene interpretation includes the recognition of multiple objects, their inter-relationships and their parts at different levels, within complex natural scenes. As for dynamic recognition, it remains to be seen whether principles similar to those discussed here will also be helpful for understanding the recognition of events and actions in scenes that include motion and changes over time.

Acknowledgements

Work reported here was supported by ISF grant 7-0369, IMOS grant 3-992 and EU IST grant FP6-2005-015803, and was conducted at the Moross Laboratory for Vision and Motor Control. Special thanks to B. Epshtein for his help.

References

1 Duda, R.O. *et al.* (2000) *Pattern Classification* (2nd edn), John Wiley & Sons
2 Mel, B.W. (1997) Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural. Comput.* 9, 777–804
3 Wiskott, L. *et al.* (1997) Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern. Anal. Mach. Intell.* 10, 775–779
4 Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025
5 Biederman, I. (1985) Human image understanding: recent research and a theory. *Comput. Vis. Graph. Image Process.* 32, 29–73
6 Sali, E. and Ullman, S. (1999) Combining class-specific fragments for object classification. *Proc. 10th British Machine Vision Conference* 1, 203–213
7 Ullman, S. and Sali, E. (2000) Object classification using a fragment-based representation. In *Lecture Notes in Computer Science: Biologically Motivated Computer Vision* (Vol. 1811) (Lee, S. *et al.*, eds), pp. 73–87, Springer
8 Ullman, S. *et al.* (2002) Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 1–6
9 Marr, D. (1982) *Vision,* W.H. Freeman
10 Schyns, P.G. *et al.* (2002) Show me the features! Understanding recognition from the use of visual information. *Psychol. Sci.* 13, 402–409
11 Turk, M. and Pentland, A. (1990) Eigenfaces for recognition. *J. Cogn. Neurosci.* 3, 71–86
12 Epshtein, B. and Ullman, S. (2006) Satellite features for the classification of visually similar classes. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2, 2079–2086
13 Epshtein, B. and Ullman, S. (2005) Feature hierarchies for object classification. *Proc. IEEE Int. Conf. Comput. Vis.* 1, 220–227
14 Bart, E. and Ullman, S. (2005) Cross generalization: learning novel classes from a single example by feature replacement. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 1, 672–679
15 Tarr, M.J. and Gauthier, I. (1998) Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition* 67, 73–110
16 Linsker, R. (1988) Self-organization in a perceptual network. *Comput.* 21, 105–128
17 Bell, A.J. and Sejnowski, T.J. (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159

18 Serre, T. *et al.* (2005) Object recognition with features inspired by visual cortex. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2, 994–1000
19 Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200
20 Rolls, E.T. and Milward, T.A. (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
21 Ullman, S. and Soloviev, S. (1999) Computation of pattern invariance in brain-like structures. *Neural Netw.* 12, 1021–1036
22 Wallis, G. and Bulthoff, H. (2001) Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4800–4804
23 Ullman, S. and Bart, E. (2004) Recognition invariance obtained by extended and invariant features. *Neural Netw.* 17, 833–848
24 Bart, E. *et al.* (2004) View-invariant recognition using corresponding object fragments. In *Lecture Notes in Computer Science: Computer Vision – ECCV 2004* (Pajdla, T. and Matas, J., eds), pp. 152–165, Springer
25 Epshtein, B. and Ullman, S. (2005) Identifying semantically equivalent object fragments. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 1, 2–9
26 Shepard, R.N. and Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science* 171, 701–703
27 Palmer, S. (1999) *Vision Science: Photons to Phenomenology,* MIT Press
28 Malik, J. *et al.* (2001) Contour and texture analysis for image segmentation. *Int. J. Comput. Vis.* 43, 7–27
29 Borenstein, E. and Ullman, S. (2002) Class-specific top-down segmentation. In *Lecture Notes in Computer Science: Computer Vision – ECCV 2002* (Heyden, A. *et al.*, eds), pp. 109–124, Springer
30 Borenstein, E. and Ullman, S. (2004) Learning to segment. In *Lecture Notes in Computer Science: Computer Vision – ECCV 2004* (Pajdla, T. and Matas, J., eds), pp. 315–328, Springer
31 Brady, M.J. and Kersten, D. (2003) Bootstrapped learning of novel objects. *J. Vis.* 3, 413–422
32 Borenstein, E. *et al.* (2004) Combining bottom-up and top-down segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, p. 46, IEEE Computer Society Digital Library (http://www.computer.org/portal/site/csdl/index.jsp)
33 Schyns, P.G. and Rodet, L. (1997) Categorization creates functional features. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 681–696
34 Sigala, N. *et al.* (2002) Visual categorization and object representation in monkeys and humans. *J. Cogn. Neurosci.* 14, 187–198
35 Logothetis, N.K. *et al.* (1995) Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563
36 Sigala, N. and Logothetis, N.K. (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415, 318–320
37 Baker, C.I. *et al.* (2002) Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.* 5, 1210–1216
38 Wachsmuth, E. *et al.* (1994) Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cereb. Cortex* 4, 509–522
39 Tanaka, K. (2003) Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* 13, 90–99
40 Brincat, S. and Connor, C.E. (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* 7, 880–886
41 DiCarlo, J.J. and Maunsell, J.H.R. (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J. Neurophysiol.* 89, 3264–3278
42 Nazir, T.A. and O'Regan, J.K. (1990) Some results on translation invariance in the human visual system. *Spat. Vis.* 5, 81–100
43 Dill, M. and Fahle, M. (1997) The role of visual field position in pattern-discrimination learning. *Proc. Biol. Sci.* 264, 1031–1036
44 Wallis, G. and Bulthoff, H. (2001) Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4800–4804
45 Needham, A. and Baillargeon, R. (1998) Effects of prior experience in 4.5-month-old infants' object segregation. *Infant Behav. Dev.* 21, 1–24
46 Peterson, M.A. and Gibson, B.S. (1993) Shape recognition contributions to figure–ground organization in three-dimensional displays. *Cognit. Psychol.* 25, 383–429
47 Zemel, R.S. *et al.* (2002) Experience-dependent perceptual grouping and object-based attention. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 202–217
48 Grill-Spector, K. and Kanwisher, N. (2005) Visual recognition: as soon as you know it is there, you know what it is. *Psychol. Sci.* 16, 152–160
49 Zhou, H. *et al.* (2000) Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611
50 Supèr, H. *et al.* (2001) Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nat. Neurosci.* 4, 304–310