Advanced Review

Models of visual categorization

Thomas Serre*

Visual categorization refers to our ability to organize objects and visual scenes into discrete categories. It is an essential skill as it allows us to distinguish friend from foe or edible versus poisonous food. Understanding how the visual system categorizes objects and scenes is a challenge because it requires bridging the gap between different levels of understanding–from the level of neural circuits and neural networks to the level of information processing and, ultimately, behavior. Computational models have become powerful tools for integrating knowledge across these levels of analysis. We review recent progress in our understanding of the computational mechanisms underlying visual categorization and discuss some of the remaining challenges.

How to cite this article:
WIREs Cogn Sci 2016, 7:197–213. doi: 10.1002/wcs.1385

INTRODUCTION

Categorization is arguably one of the most critical tasks that must be solved by our visual system: It is essential to survival because it allows an animal to make inferences regarding an object's properties by generalizing from other category members.[1] Human and nonhuman primates excel at visual categorization tasks and can reliably categorize objects embedded in complex natural visual scenes with only a glimpse (see Refs 2,3 for recent reviews).

Formally, visual categorization is the process by which a set of visual stimuli $\mathbf{x}_i$ get associated with class labels y_i to form $(\mathbf{x_i}, y_i)$ exemplar-label pairs (Figure 1). It requires learning a decision function f that best represents the mapping between the inputs and the corresponding outputs $f(\mathbf{x_i}) \approx y_i$.

Work in biological and machine vision has traditionally focused on the study of perceptual representations: How does the visual system extract diagnostic visual features to build robust visual representations $\mathbf{x}_i$ that are tolerant with respect to the many factors that affect the appearance of natural object categories? Work in cognitive psychology, on the other hand, has focused on the mechanisms underlying the categorization process, i.e. what is the nature of the space of decision functions f and how are these decision functions learned from training examples?

Historically, perceptual representations and categorization processes have been studied with little overlap.[4,5] Understanding visual categorization will necessarily require bringing together multiple disciplines from computer vision and machine learning to cognitive psychology and systems neuroscience. The goal of this review is thus to integrate key pieces of the literature from relevant disciplines and to provide an overview of the current state of the field.

In the following, we start with a brief description of the organization of the visual cortex and review the neural basis of visual categorization. We then survey existing computational models of visual perception—highlighting recent developments in (deep) learning of visual representations. We proceed with an overview of formal models of categorization from the perspective of computational learning theory and cognitive psychology. We conclude with open questions and highlight promising future directions for research.

THE NEURAL BASIS OF VISUAL CATEGORIZATION

Visual perception is a dynamic process, which starts with a coarse initial analysis of a scene that gets continuously refined to reflect the infinite amount of details present in natural scenes[6]: 'The more you look, the more you see.' A large body of literature

*Correspondence to: Thomas_Serre@brown.edu

Cognitive, Linguistic & Psychological Sciences Department, Institute for Brain Sciences, Brown University, Providence, RI, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

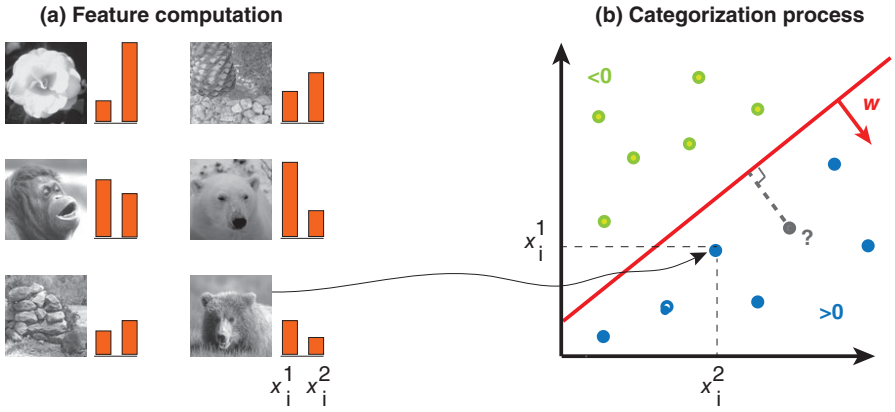

FIGURE 1 | Computational models of visual categorization. Visual categorization has traditionally been described as a two-stage process: (a) Visual features must be computed to build a visual representation of an input stimulus $\mathbf{x}_i$. It is desirable for the representation to be both tolerant to the many factors that can affect the appearance of an object and also selective enough to capture subtle differences between exemplars across the category boundary. Different computational models of feature computation vary in their degree of invariance and specificity. For illustration purposes, two features x_i^k are being computed (superscripts are used as feature indexes and subscripts as stimulus indexes) but more generally, the total number of features N used to represent visual stimuli can be quite large ($N \approx 10^2 - 10^4$). More generally, visual stimuli can be thought of as N-dimensional feature vectors (also called data points) $\boldsymbol{x_i} = (x_i^1, \ldots, x_i^k, \ldots, x_i^N)$ in this representational space whereby the kth coordinate of $\mathbf{x}_i$ corresponds to the response of the kth feature detector x_i^k. (b) A categorization process associates these data points $\mathbf{x}_i$ to category labels y_i through a learned function f such that $f(\mathbf{x}_i) \approx y_i$. Here, we consider a binary classification task with a positive (target) and a negative (distractor) category label ($y_i = \{-1, 1\}$). Shown in red is a linear classification function f that separates the positive and negative examples. This function is parametrized by the vector $\mathbf{w} = (w_1, w_2)$, which is the vector normal to the underlying decision boundary. In practice, these functions are learned from training examples. For instance, supervised learning algorithms learn this mapping from the presentation of $(\mathbf{x}_i, y_i)$ exemplar-label pairs. After learning, the algorithm tries to predict the category label of a new stimulus $\mathbf{x}_*$ by considering whether the stimulus projected in the feature space falls on the right or left side of the boundary. This can be done by computing the dot-product between the input stimulus and the normal vector and subtracting off a fixed threshold θ: $f(\boldsymbol{x}_*) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_* - \theta) = \text{sign}\left(\sum_k w^k x_*^k - \theta\right)$.

suggests that an initial coarse visual analysis relies on the extraction of relatively simple visual features via feature detectors that operate very rapidly and in parallel across the entire visual field.[7] Experiments using artificial search arrays of stimuli have demonstrated that simple image features can be processed pre-attentively and in parallel while more complex feature combinations require a serial attentional process.[8]

Our visual system appears to be surprisingly well adapted to our natural environment. Studies conducted using natural visual scenes have demonstrated the incredible speed and accuracy of the visual system for some of the most challenging visual recognition tasks (such as animal vs. non-animal categorization) in the near absence of attention[9] (see Ref 2 for review).

The underlying visual representation remains limited to relatively coarse shape information as human observers frequently fail to localize targets (the *where* task) that they had correctly detected[10] (the *what* task). Such results seem inconsistent with vision theories that rely on explicit encoding of spatial relationships between features[11] and suggest instead that rapid visual categorization may rely on a dictionary of *unbound* visual features.[12,13] However, this leaves open the question of where these visual features are computed.

Visual processing consists of a series of neurally interconnected stages (Figure 2), starting at the level of the retina, and proceeding through the Lateral Geniculate Nucleus (LGN) of the thalamus to the primary visual cortex (V1). The primary visual cortex, in turn, projects to extra-striate visual areas along the ventral stream of the visual cortex from area V2 and V4 to the inferotemporal cortex (ITC).[16,17] The ITC constitutes the final stage between visual cortices, on the one hand, and the limbic systems and frontal areas,[18] on the other hand, effectively linking perception to memory and action.

The notion of a visual hierarchy began with the groundbreaking work of Hubel and Wiesel in the striate cortex (see Ref 19 for review), who first suggested how tuning for orientation and tolerance to small shifts in position in V1 could originate from a hierarchy of selective pooling mechanisms. It is now relatively well established that these types of hierarchical pooling mechanisms extend beyond V1 to the

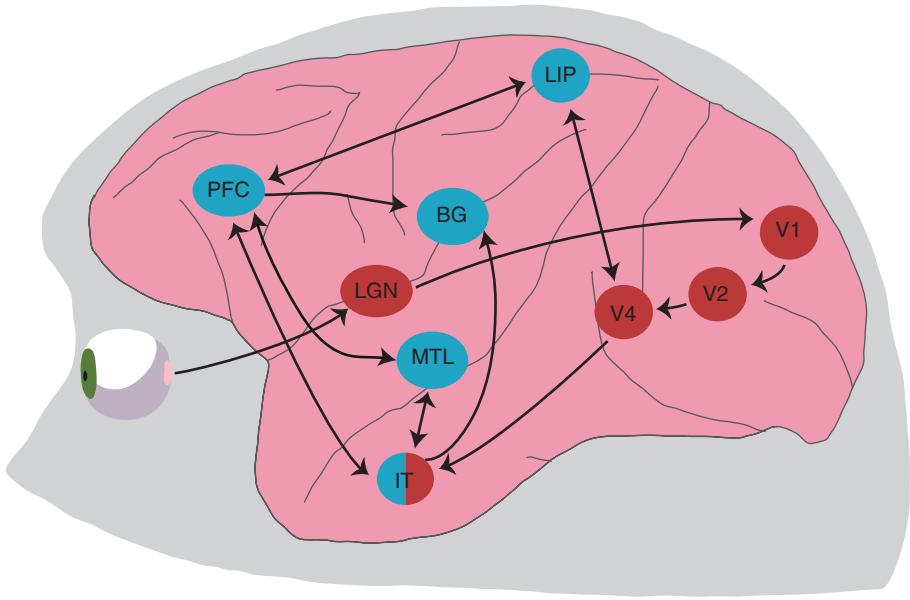

FIGURE 2 | The neural basis of visual categorization. Shown are areas involved in visual categorization. Areas involved in the computation of visual features are shown in red and areas involved in categorization in cyan. Some subcortical areas known to play a role in categorization are not shown including the striatum.[14] (Adapted from Ref 15)

entire ventral stream:[16,17] At each stage of the processing hierarchy, the underlying visual representation becomes increasingly complex with cells becoming selective to increasingly more stimulus dimensions – from single orientations to image fragments and object views in higher visual areas. At the same time, the underlying visual representation becomes gradually more tolerant to image transformations (mainly changes in position and scale).

Converging evidence suggests that the ventral stream of the visual cortex plays a key role in the encoding of object categories.[16,17,20] For instance, category information has been found in V4[21] and entire clusters of ITC neurons are selectively tuned to ecologically important categories of stimuli such as faces and body parts[22] as well as objects of expertise in humans.[23] Interestingly, semantic information encoded in the human homologue of the ITC is remarkably similar to that found in monkeys.[24,25] Several studies have also reported that natural object categories can be read out from ventral stream neural activity.[26–28] Furthermore, category learning introduces local changes in the ITC[29] (see also Ref 30) as well as distributed effects throughout the ventral stream of the visual cortex (see Ref 31 for review).

Whether the ventral stream is directly involved in the categorization process per se, as opposed to building a robust visual representation, remains, however, a matter of debate. For instance, it has been shown that ITC neurons often do not completely generalize among category members and remain selective for the perceptual similarity between individuals.[20] Category signals in the prefrontal cortex (PFC), one of the main projections of the ITC, tend to be stronger with shorter latencies compared to ITC.[20] The category information found in the ITC could thus reflect top-down signals[32] from the PFC (but see also Refs 33,34) or other memory-related areas.

Outside the ventral stream, category selectivity has also been reported in the lateral intra-parietal (LIP)[20] as well as the hippocampus, amygdala and entorhinal cortex.[35] Selectivity for threatening stimuli (snakes) has been observed in the pulvinar within ≈50 ms post-stimulus onset.[36] Selectivity for animate object categories has been demonstrated in the human amygdala.[37] This has lead some researchers to argue for the possibility of a 'low-road' subcortical pathway for visual recognition that would bypass the aforementioned 'high-road' ventral stream cortical pathway (see Ref 38 for a review). However, the observed latencies are relatively slow (>300 ms) compared to the fastest reaction times observed during rapid categorization tasks[2] (<300 ms). In comparison, object category information can be decoded from the ventral stream of the visual cortex much faster (within ≈100 ms post stimulus onset) in both humans[39] and monkeys.[26,27]

A very recent study[28] has shown that category signals in the ventral stream of the visual cortex co-

vary with monkey behavioral responses during a rapid categorization task, offering a more direct evidence for the 'high-road' (cortical) hypothesis. At the same time, the existence of direct projections between the pulvinar and intermediate areas of the ventral stream leaves open the possibility that the category selectivity found in this study originates in subcortical areas with the ventral stream simply relaying the information to downstream areas.[40]

MODELS OF FEATURE COMPUTATION

One of the main challenges associated with visual categorization stems from the need to build a representation that achieves a difficult trade-off between invariance and selectivity.[41] On the one hand, our visual system must build a visual representation that is tolerant to the many factors affecting the appearance of an image such as changes in the position, scale, illumination or viewpoint of an object in our field of view. On the other hand, the underlying visual representation must remain selective enough so as to maintain an ability to judge subtle differences between similar object categories. Different visual categorization tasks may require different trade-offs between selectivity and invariance, and computational models of feature computation form a continuum from low-level to intermediate and higher-level visual representations, which we review below (see Table 1 for an overview).

Lower-Level Visual Features

Some of the simplest image features that have been proposed are those based on linear filter responses. These include filters modeled after the center-surround receptive fields found in the LGN.[42,43] For instance, a simple parametric distribution (*Weibull* function) computed over the response of such filters has been proposed as a model of natural scene recognition and was shown to be a good predictor of brain EEG responses during rapid visual presentations.[43]

Visual representations based on the output of oriented filters such as Gabor functions, Gaussian derivatives or other steerable filters have been exceedingly popular in the past two decades. In computational models of the visual cortex, these filters aim to mimic processing by cortical cells tuned to different orientations and spatial frequencies as found in the primary visual cortex.[19] *Gabor Jets*, for instance, have been used to model face recognition[44] and shown to account well for face similarity measurements derived from psychophysical data.[54] A few years ago, closely related computational models of the early visual cortex were shown to compete with state-of-the-art computer vision algorithms.[55,56]

The *Gist* algorithm[45] is another popular algorithm based on relatively low-level oriented filters. The model has been shown to perform well on a variety of scene categorization tasks.[57,58] Unlike the *Gabor Jets* described above, the *Gist* algorithm relies on local spatial pooling mechanisms to build a coarse scene representation (and avoid an explosion in the number of visual features used). A computational model based on the *Gist* algorithm was shown to account well for behavioral data on the effect of context during rapid scene categorization.[59]

Visual Features of Higher Complexity

How do simple local visual representations, as found in early cortical areas, yield the more complex visual features found in higher visual areas of the visual cortex and that are known to be optimal for visual categorization?[60,61] Currently, one of the most prominent proposals is based on the notion of a visual hierarchy whereby another (and possibly several) filtering-like processing stages can be applied in cascade.

TABLE 1 | Representative Models of Feature Computation, Corresponding Receptive Field (RF) Types and Complexity as well as Associated Key References

Models	RF types	Complex	Key references
Weibull	Single layer, center-surround	+	42,43
Gist, Gabor jets, V1 model	Single layer, oriented	+	44–46
Textons, TextSynth	Two-layer, summary statistics	++	47–49
Shape descriptors	Fourier, boundary moments, medial axis	++	50,51
Neocognitron, HMAX, deep learning	Multi-layer, selectivity and invariance pooling	+++	41,52,53

The level of complexity is indicated with '+' signs from lowest (+) to intermediate (++) and highest (+++). Many more visual features have been proposed in the context of computer vision, and we here only focus on those that are relevant to biological vision.

For instance, *textons* (the term was first coined by Julesz to describe the fundamental atoms of pre-attentive vision[62]), corresponding to combinations of oriented linear filters via a two-stage cascade, were shown to approach the level of accuracy of human observers during the categorization of visual scenes.[47] Beyond scene categorization, one of the best computer vision systems for the detection of contours in natural images[63] is also based on textons.

Another algorithm for computing intermediate-level features, *TextSynth*, was derived from a successful approach to texture synthesis.[64] The *TextSynth* algorithm not only takes into account first order statistics (i.e. feature count or mean feature responses over a small region of space as with most features described above) but also higher-order statistical dependencies found in natural textures (including kurtosis and other higher-order moments). This approach was shown to predict human performance in crowding experiments[48,49] as well as the strength of neuronal and fMRI responses in intermediate visual area V2.[65]

Much as features of intermediate complexity can be obtained by pooling together simpler ones, features of higher complexity can be obtained by combining intermediate-level ones. This idea is a key feature of many visual architectures including the Neocognitron,[52] deep learning networks[66] and other models of object recognition[12,41,60,67–71] (see Ref 72 for a recent review).

The HMAX model[12,41,70] shown in Figure 3 constitutes a representative example of the class of feedforward hierarchical models. The model combines mechanisms for the hierarchical build-up of invariance and selectivity inspired by the Neocognitron[52] with view-based theories of 3D object recognition.[73] HMAX tries to emulate the main information processing stages across the entire ventral stream visual pathway and bridges the gap between multiple levels of understanding:[74] This system-level model seems consistent with physiological data in nonhuman primates in different cortical areas of the ventral visual pathway,[70] as well as human behavioral data during rapid categorization tasks with natural images[12,70,75] (but see also Ref 76–79 and later discussion).

In recent years, a number of HMAX extensions have been proposed. For the most part, these extensions have focused on the learning of visual representations in intermediate stages of the model. One prominent example includes the work by Masquelier et al. who incorporated biologically plausible learning mechanisms in the HMAX based on temporal continuity in video sequences,[80] evolutionary

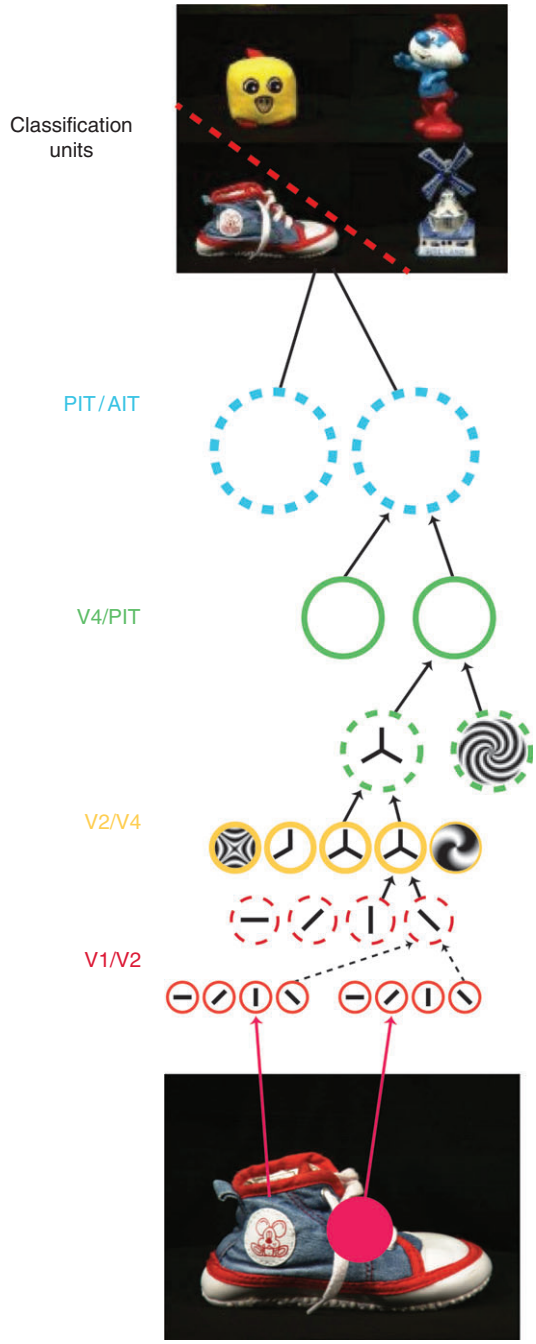

FIGURE 3 | Sketch of the (HMAX) hierarchical model of visual processing: Acronyms: V1, V2, and V4 correspond to primary, secondary and quaternary visual areas, PIT and AIT to posterior and anterior inferotemporal areas, respectively (tentative mapping with areas of the visual cortex shown in color, some areas of the parietal cortex and dorsal streams not shown). The model relies on two types of computations: A max operation (shown in the dashed circles, also called invariance pooling) over similar features at different position and scale to gradually build tolerance to position and scale and a bell-shaped tuning operation (shown in the plain circles, also called selectivity pooling) over multiple features to increase the complexity of the underlying representation, see Ref 12,70 and text for details.

algorithms[81] as well as spike-timing dependent-based learning rules.[71,82]

Deep Learning Networks

Nearly all hierarchical models of object recognition described above learn invariant visual representations without any supervision using (Hebbian-like) *unsupervised* learning rules. These models learn visual features that are common in natural images irrespective of their underlying diagnosticity for particular categorization tasks. This type of learning seems consistent with ITC recordings that have shown that the learning of position and scale invariance, for instance, is driven by the subject's visual experience[83,84] and is unaffected by reward signals.[85]

In recent years, however, a class of neural networks called deep learning architectures have brought about a revolution in machine learning. These networks have pushed the state of the art on a range of categorization problems ranging from speech and music to text, genome and image categorization (see Ref 66 for a very recent review). There are two fundamental differences between these deep learning architectures and hierarchical models of the visual cortex such as HMAX and the Neocognitron described above. First, learning across processing stages is fully supervised: It uses the back-propagation algorithm (see Ref 66 for a history) which propagates an error signal from superficial (categorization) layers toward deeper (perceptual) processing stages. Thus only visual features that are diagnostic for the trained categorization tasks will be learned.

Second, unlike HMAX, whose parameters (receptive field sizes, invariance and other tuning properties, number of layers, etc) are constrained by available neuroscience data, deep learning architectures do not try to imitate biology at such a level of detail. For instance, state-of-the-art deep learning architectures incorporate many more layers (more than 20 layers[86,87]) than the aforementioned hierarchical models of the visual cortex (e.g. seven layers for HMAX) – possibly also incorporating not just one visual architecture but entire ensembles of deep networks for a given categorization task.[86,87]

Recent innovations in training methods have yielded deeper networks with improved accuracy. Of course, an increase in the number of layers and units in the network comes at the expense of sample complexity (see Section *Models of Classification*) as the number of parameters to be learned increases together with the complexity of the corresponding classification function. Unsurprisingly, a very significant effort has been dedicated in recent years to building increasingly large annotated image and video datasets (the ImageNet Large Scale Visual Recognition Challenge[88] contains more than 1 million images and 1000 categories) enabling the training of increasingly large networks (compare with the 2010 PASCAL VOC challenge[89] with less than 20,000 images and 20 categories).

Despite the absence of neuroscience constraints on modern deep learning architectures, recent work has shown that these architectures are better able to explain ventral stream neural data than alternative models.[78,79,90,91] In addition, these networks outperform all other models by a large margin[90] and are starting to match human-level accuracy for difficult object categorization tasks.[87]

MODELS OF CLASSIFICATION

A key question for cognitive science is to understand the processes that link visual stimuli with category labels. Our visual system must learn these associations and, beyond rote memorization, it must learn to generalize to previously unseen exemplars. After a brief mathematical description of the category learning problem, we review existing cognitive models of categorization and survey different kinds of classification problems.

Learning from Examples

Formally, visual categorization is the process by which a set of visual stimuli $\mathbf{x}_i$ ($i = 1 \ldots m$) gets associated with a category label y_i to form ($\mathbf{x}_i$, y_i) exemplar-label pairs. In general, $\mathbf{x}_i$ is a feature vector in a possibly high dimensional space (see Section *Models of Feature Computation* and Figure 1). For instance, this could be N pixel intensities from an $\sqrt{N} \times \sqrt{N}$ input image or the response of an N-dimensional array of photoreceptors or any other feature detectors $\mathbf{x}_i = \left(x_i^1, \ldots x_i^k, \ldots x_i^N\right)$ to the presentation of stimulus $\mathbf{x}_i$.

For binary classification tasks, it is customary to consider positive and negative training samples (i.e. y_i is a binary variable taking on values {–1, 1}, but in the general multiclass categorization case, y_i could take on any integer value). Multiclass categorization problems, on the other hand, require choosing the category (out of k possible) to which a visual stimulus belongs. Formally, these can be described as multiple binary classification problems[a]. Learning to categorize visual stimuli requires learning a function,[92] called the classification function or

classifier f, that best represents the relation between the inputs $\mathbf{x}_i$ and the corresponding outputs y_i such that $f(\mathbf{x}_i) \approx y_i$.

One can distinguish between different kinds of learning scenarios: In supervised learning, a teacher provides input–output pairs to the learner. This is closely approximated in a lab setting where, for instance, a participant is being shown sets of examples from two classes with corresponding labels in order to learn to discriminate between these two categories. At the other extreme of the continuum, in unsupervised learning scenarios, the learner is only provided with training examples $\mathbf{x}_i$, and the y_i labels along with their associations with training examples have to be 'guessed.' There also exist approaches that are hybrids between supervised and unsupervised learning schemes, such as the semi-supervised learning approach, whereby some but not all labels are provided to the learner. More natural learning scenarios inspired by early behavioral psychology work on conditioning include reinforcement learning algorithms whereby the learner receives a reward/punishment following its actions or decisions. The stimuli and labels are never explicitly given as pairs but can nonetheless be learned indirectly from the reward signal. These types of reinforcement learning scenarios are extensively used in robotic applications (for instance, for robots to learn to walk).

One needs to distinguish between the function $f(\mathbf{x}) \approx y$ that tries to predict the label y of an unknown input stimulus $\mathbf{x}$ and the algorithm used to learn the function f itself. There exist a multitude of algorithms for learning classification functions that have been proposed in the past decades of research in machine learning. Some of the most common algorithms in visual categorization include the perceptron learning algorithm and extensions to neural networks with multiple layers, as well as support vector machines (SVMs) and closely related regularization networks, boosting and many others (see Refs 93,94 for general overviews). All these algorithms aim at selecting one of usually many possible classification functions. These algorithms typically try to achieve a difficult trade-off between minimizing the classification error on the training set (i.e. fitting the training data well) and finding a function that is sufficiently smooth (to prevent overfitting to the training set and provide some guarantee to generalize well to future examples, see below).

To achieve this trade-off, a learner needs to make explicit assumptions (this is related to the so-called *inductive bias* in cognitive psychology) about the nature of the categorization problem in order to generalize to new exemplars that it has not encountered so far. Different learning algorithms rely on different assumptions.[93,94] The maximum margin assumption, for instance, which is used in the formulation of SVMs, posits that the best classification function will be the one that maximizes the distance between the closest exemplars from the two classes and the corresponding classification boundary.[95] Nearest neighbor algorithms typically assume that exemplars that fall within the same neighborhood (i.e. that are near each other) belong to the same class. In addition, there are many possible types of functions[93,94] (or kernels) that can be used to parameterize the space of classification functions used by the learner (which is independent of the learning algorithm). For instance, the same classification problem can be solved with a linear SVM or a nonlinear SVM (including polynomial functions, radial basis functions RBFs, etc). We next review some of the main algorithms that have been proposed in the past decades and that are of relevance to cognitive psychology.

Machine Learning and Cognitive Models

One of the very first computational models of categorization was the *perceptron*.[96] It extends McCulloch and Pitts' earlier model of an artificial neuron[97] with the ability to adjust its synaptic weights via a simple learning rule in order to learn new categories. The underlying circuit used for categorization is quite general, as it can describe any linear classification function learned by modern learning algorithms.

An idealized classification unit receives its inputs from an N-dimensional population of feature detectors $\mathbf{x} = (x^k)_{k=1\ldots N}$ with corresponding synaptic strengths $\mathbf{w} = (w^k)_{k=1\ldots N}$ (Figure 4) that are learned from a training set of exemplar-label pairs. After learning, the unit predicts the category label of a new stimulus $\mathbf{x}$ by comparing its output value given by the sum of its inputs weighted by the corresponding synaptic weights with a fixed threshold θ: $y = \text{sign}(\sum_k w^k x^k - \theta)$. This is equivalent to evaluating whether the stimulus falls on the right (positive) or left (negative) side of a corresponding categorization boundary (Figure 5).

In such boundary-based classification models, the raw classifier output (before thresholding) is a linear function of the distance of the data point to the decision boundary. A closely related cognitive psychology model is the General Recognition Theory (GRT) model and its extensions[98] whereby a diffusion process accumulates information along each feature dimension independently until a decision boundary is reached.

One of the key limitations of the original perceptron learning algorithm is that it is limited to linearly separable classes. A significant extension of the linear perceptron is the multilayer perceptron (MLP). The architecture is built by stacking multiple perceptrons on top of each other such that one stage feeds into the next. These can be seen as precursors of the feedforward hierarchical models and deep learning architectures described in Section *Models of Feature Computation*.

The MLP in effect implements a complex nonlinear function f between its input and output. In fact, it can be shown that under relatively general assumptions,[99] certain architectures called universal approximators like the MLP or the RBF discussed below can approximate any arbitrary input–output function (and hence categorization tasks of arbitrary difficulty). However, learning to solve the problem may require a prohibitively large number of training samples and hence, all else being equal, it is always desirable to consider the 'simplest' classification function sufficient to solve the categorization problem (this principle is known as Occam's razor).

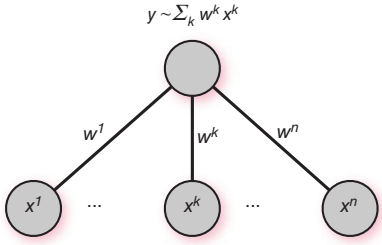

FIGURE 4 | Simple categorization unit. A perceptron-like categorization unit reweights the response of individual feature detectors x^k – from a population of N feature detectors $\mathbf{x} = (x^1, \ldots x^k, \ldots x^N)$ – by the corresponding vector of synaptic strength $\mathbf{w} = (w^1, \ldots w^k, \ldots w^N)$ before summing them up ($\sum_k w^k x^k$) and subtracting off a threshold θ. This is followed by a rectification stage to obtain a binary class label {–1, 1}. Formally this model unit would be able to implement the classification boundary described in Figure 1.

A popular class of neural network models, which are also universal approximators, are regularization networks based on radial basis functions (so-called RBF networks). In the RBF scheme, a number of 'templates' $\mathbf{x}_i$ are stored in memory by individual model units. These templates may correspond to individual exemplars previously encountered by the network or 'prototypes' learned via clustering of the training data. During categorization, each unit matches the input stimulus $\mathbf{x}$ against its template using a particular kernel function $K(\mathbf{x}_i, \mathbf{x})$. For RBF networks, the kernel function takes the form of a Gaussian function $K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}\|^2)$, where γ is a constant. A classification unit then computes the sum of the response of all template units $K(\mathbf{x}_i, \mathbf{x})$ weighted by their associated synaptic weights c_i. Formally, the categorization process can be described with the following equation:

$$y = \sum_i c_i K(\mathbf{x}_i, \mathbf{x}) = \sum_i c_i \exp\left(-\gamma||\mathbf{x}_i - \mathbf{x}||^2\right).$$

This approach was shown to successfully learn to synthesize novel object views from 2D templates stored in memory.[100] This so-called view-based theory of pose-invariant object recognition offered an alternative to models that rely on explicit 3D CAD-like representations of objects[11] and motivated subsequent monkey psychophysics and electrophysiology

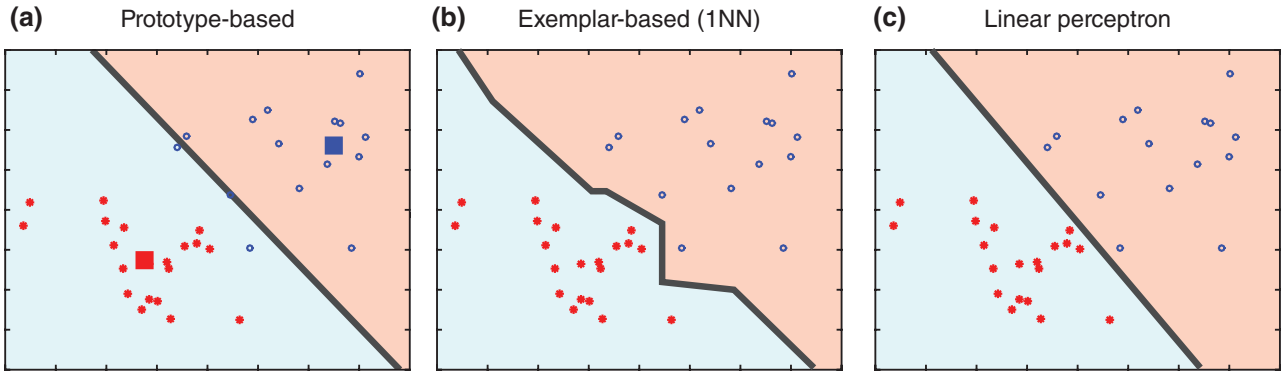

FIGURE 5 | Decision boundaries implemented by three different cognitive models of categorization. (a) Prototype-based, (b) exemplar-based and (c) linear perceptron. Note that only the perceptron learning algorithm explicitly computes a decision boundary. A decision boundary can, however, be recovered for instance-based algorithms by assigning a class-label to every point of the feature space by computing the distance between each point and the closest prototype (obtained by computing the mean of all exemplars for each class) as in the prototype-based approach or the closest exemplar as in the exemplar-based approach.

studies.[101] A refinement of the framework lead to the *Chorus of Prototypes* model, which was shown to generate correct predictions for behavioral, electrophysiological and imaging experiments.[102]

RBF networks fall under the general family of kernel methods. Many modern machine learning algorithms correspond to different choices of the kernel function described in Equation (3). This mathematical formalism is quite general and, indeed, many of the cognitive models of categorization described in the past decades (see Table 2 for an overview and Ref 115 for a more in-depth review of these models) can be formally shown to be special cases of Equation (3) for a particular form of the kernel function.[116] Representative examples include the generalized context model (GCM), which assumes that category exemplars are being stored in memory for later retrieval during the categorization of new stimuli[117] and a connectionist variant called ALCOVE.[110]

Both the GCM and the ALCOVE models belong to the general class of instance-based algorithms, which also include popular machine learning algorithms such as the nearest neighbor algorithm (or the related prototype-based models). The k-nearest neighbor (k-NN) algorithm, for instance, is often referred to as a 'lazy learning' algorithm because learning with this model only involves storing exemplar-label pairs. The classification function is only approximated locally (around training examples) and the actual crux of the computation is deferred until classification. A category label is obtained by associating the label of the k-nearest neighbors to an unfamiliar example. An example of the decision boundary associated with a 1-NN classifier is shown on Figure 5. Unlike the other models described above, k-NN does not re-weight dimensions according to their diagnosticity for the task (which is akin to stretching perceptual representations along relevant dimensions and shrinking them along irrelevant ones).

Extensions of the GCM include the exemplar-based random walk (EBRW) model by Nosofsky and Palmeri.[111] Categorization proceeds by sampling stored exemplars sequentially, which pushes an integrator toward one of two boundaries according to the exemplar's identity. The rate at which an exemplar is retrieved increases with its similarity to the stimulus and the memory strength. The model was shown to provide an account for the effects of familiarity and similarity on the categorization process.

Prototype-based models constitute an alternative to the family of instance-based models described above. Category prototypes are learned (usually by clustering) and used in lieu of exemplars during categorization, thus reducing the number of category instances to be stored (Figure 5). A related model is the norm-based model of face perception (oftentimes called the 'average face' model) whereby faces are categorized by considering the distance along prespecified directions in the space coding for face identities with respect to a single average face prototype.[118]

A Note on Representational Complexity

One way to compare the quality of different visual representations for a given categorization problem (and a given stimulus set) is to consider their representational complexity. The representational complexity of a given visual representation can be thought of as the complexity of the *simplest* classification function necessary to reach a particular level of categorization accuracy.

For instance, let us consider least-square regression as our learning algorithm and different classes of decision functions (or kernels), say linear, quadratic and Gaussian (one could also vary the regularization parameter to restrict the complexity of the resulting decision function). Similarly, one could consider neural networks with increasing numbers of hidden units. The number of 'wiggles' on these functions provides a coarse approximation of their degree of freedom. With more free parameters, these functions can exhibit sharper peaks and valleys, which in turn, allow them to solve increasingly complex categorization problems. The number of wiggles can thus be taken as a hand-wavy estimate of the relative

TABLE 2 | Representative Cognitive Models of Categorization

Models	Categorization process	Key references
RULEX, COVIS	rule	103–105
GRT, perceptron, MLP	boundary	106
Prototype, SUSTAIN	prototype (single or clusters)	107–109
ALCOVE, GCM, ERBW, RBF	exemplar	100,110–112

GRT, General Recognition theory; MLP, Multilayer Perceptron; GCM, Generalized Context Model; ERBW, Exemplar-Based Random Walk; RBF, Radial basis function (network).
Many more models exist and we refer the readers to Refs 113,114 for a more complete survey of these models.

complexity of these functions (linear < quadratic < Gaussian).

In practice, beyond simple considerations on the number of free parameters, one has to worry about the *capacity* of a classification function which needs to take into account the expressive power or flexibility of the function. Several measures such as the VC dimension have been proposed to make this notion of capacity more rigorous.[95] Measures of the capacity of a classifier are important because they provide bounds on the number of samples (also called sample complexity) that are needed for proper generalization of the corresponding classification function.[b] There are exceptions to this general rule, in particular, committees of perceptrons can give better generalization than a single perceptron. The issue of why this occurs has not been resolved.[c]

Let's now turn to three hypothetical visual representations shown in Figure 6. With no prior assumption on the class of functions f to be learned, the simplest classification function that can *accurately* separate the data (i.e. without any classification error) in panel (a) (linear) is simpler that the simplest classification function that can separate the data in panels (b) and (c) (quadratic and Gaussian kernel, respectively). Hence, the complexity of the representation in panel (a) is much lower than that of the representation in panel (b), which is itself lower than that of the representation in panel (c).

The argument above offers a measure of the relative effectiveness of different visual features to solve a particular categorization problem because representations of lower complexity will yield classification functions that require fewer examples to train or equivalently, provides the best guarantee of generalization to novel exemplars (see Ref 121 for a more in-depth treatment). In practice, the complexity of visual representations (such as when comparing a perceptual model to populations of cortical neurons) has to be estimated empirically.[77] It has been shown experimentally that the representational complexity of deep learning architectures tends to be much lower than that of other architectures.[77]

OPEN QUESTIONS AND FUTURE DIRECTIONS

In recent years, progress in our understanding of the computational mechanisms underlying visual categorization has been significant. Face detection systems are now readily available on consumer-grade digital cameras, and automated face identification algorithms are being integrated in digital photo library suites. Automated pedestrian detection and computer systems for driver assistance are already available in luxury vehicles and will become standard equipment on most models in the near future. Below, we identify some of the remaining challenges in our understanding of the biological and computational mechanisms underlying visual categorization.

Categorization Across Taxonomic Levels: One or Multiple Computational Mechanisms?

One may consider natural object categories at three distinct levels of abstraction[122] (see Table 3): The

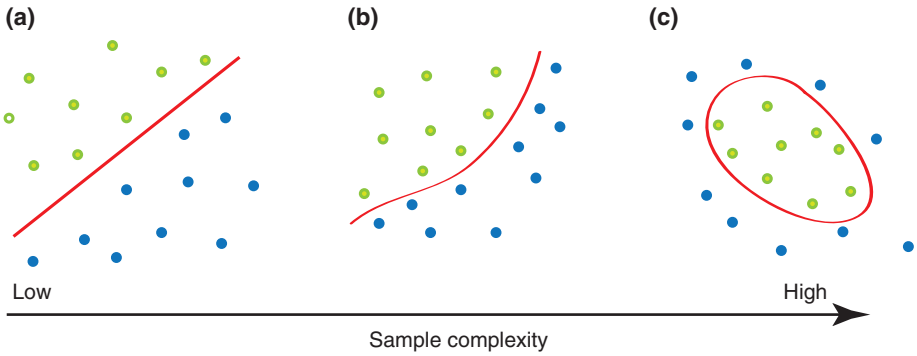

FIGURE 6 | Representational complexity. Not all feature representations are created equal. Here we compare three hypothetical visual representations for the same set of stimuli and how they impact the subsequent classification processes. Individual category exemplars are shown as dots (blue and green corresponding to each of the two classes) and classification functions as a red line. Representation (a) is the *best* representation for the categorization problem considered because the two classes can be separated by one of the simplest classification functions (i.e. a linear function). The complexity of the corresponding classification function increases from left to right. Representation (b) and (c) will, in principle, require more training examples to properly generalize to new stimuli or equivalently will tend to under-perform Representation (a) in regimes where relatively small numbers of training examples are available.

TABLE 3 | Examples of Category Taxonomies[122]

Superordinate	Basic	Subordinate
Furniture	Chair	Kitchen chair
		Living-room chair
	Table	Kitchen table
		Dining-room table
	Lamp	Floor lamp
		Desk lamp
Tree	Oak	White oak
		Red oak
	Maple	Silver maple
		Sugar maple
	Birch	River birch
		White birch

superordinate level corresponds perhaps to the most popular task in both machine and biological perception, and is often referred to as the 'detection' task. This includes categorizing a target object category (e.g. animals, vehicles, people or faces) against distractor stimuli that do not contain the target object category (often times natural scenes or other object categories). The basic level (e.g. dog vs. non-dog animals) has been described as the level for which members are first categorized, whereby additional processing may be needed for other levels[122] (but see later discussion). The subordinate level (e.g. dalmatian vs. non-dalmatian dogs) is also commonly referred to as the fine-level categorization in computer vision (e.g. which type of dog is it?). Object 'recognition' usually refers to classification tasks at the same basic level (e.g. cats vs. dogs or animals vs. people) while 'identification' usually corresponds to discriminating between one specific versus other instances of an object category (e.g. my dog Rusty vs. other dogs).

The past decade of research has shown that there exist systematic differences in participants' behavioral responses across taxonomic levels, e.g. Ref 123–125. For instance, observers' subordinate-level categorization of natural object categories tends to be slower and less accurate than basic-level categorization.[124] Participants' basic-level categorization of natural scene categories tends to be slower and less accurate than superordinate-level categorization.[125,126]

These systematic differences have often been taken as suggestive evidence for separate computational mechanisms governing different categorization tasks. For instance, it is often assumed that face detection and people recognition rely on separate computational mechanisms.[127,128] Similarly, semantic theories of categorization assume an underlying hierarchical organization of the categorization processes with some categorization tasks taking precedence over others. For instance, 'entry-point' theories typically assume that memory-related factors, including typicality, affect which taxonomic level will act as an entry-point for categorization.[129] Another prominent theory is the 'global-to-specific' theory, whereby categorization at more global stages needs to be completed before processing at finer levels can begin. Greene and Oliva[130] suggested that the categorization of global scene properties (e.g. 'is the scene open or closed?') is a necessary first step for more specific categorization tasks including basic-level categorization. A similar explanation has been proposed to explain the 'superordinate advantage' in scene categorization characterized by participants' higher accuracy and faster reaction times for categorization at the superordinate versus basic level of categorization.[125,126]

Interestingly, from the computational point-of-view, however, classification tasks at different levels of categorization are equivalent (Figure 7, see Ref 73 for a discussion). The aforementioned categorization problems can all be described within the general classification framework described in Section *Models of Classification*: Different categorization problems simply correspond to multiple associations of the same $\mathbf{x}_i$ stimuli with different class labels y_i^t where the t reflects different categorization tasks. Indeed, very recent work[131] has shown that, for the case of rapid scene categorization, the differences in behavioral responses observed across different tasks reflect natural variations in perceptual discriminability (computed using a model of categorization as reviewed in Section *Models of Classification* trained using a very large database of scene images). Furthermore, the study showed that by properly selecting visual stimuli using the proposed computational model, it is possible to reverse the 'superordinate advantage.' It remains an open question whether similar models and ideas also extend to other classes of stimuli beyond scenes such as animals[132] or faces.[124]

On the Need to Integrate Models of Feature Computation and Categorization

As stated throughout this review, studies on visual representations and categorization have been conducted somewhat independently.[4,5] One of the key assumptions in most previous work is that the brain

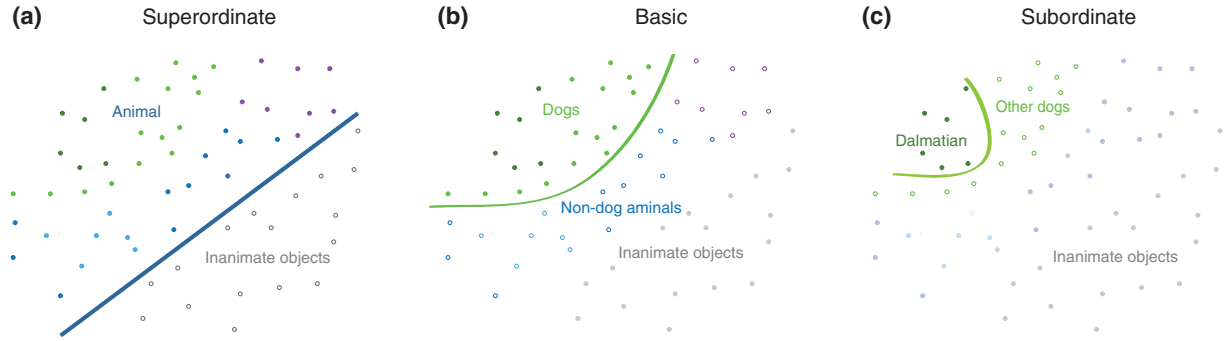

FIGURE 7 | Levels of categorization. One can distinguish between three levels of categorization. Shown are hypothetical examples related to the categorization of animal stimuli and many alternatives are possible. (a) The superordinate level corresponds to categorization between animal/animate versus non-animal/inanimate objects (i.e. any visual scene that does not contain an animal). (b) The basic level (also referred to as the generic level of categorization) requires discrimination between various species (e.g. dog vs. non-dog animals). (c) Last, the subordinate level requires discrimination between various dog types (e.g. dalmatian vs non-dalmatian, etc). Of course this classification is not unique and many other classification types can be performed, such as cat versus dogs, etc. Binary classification tasks are very general and it can be shown formally that any multiclass classification task (e.g. what animal is it? dog vs. cat vs. bird, etc) can be decomposed as multiple binary classification problems.

mechanisms responsible for the computation of visual features are somewhat distinct and independent from those responsible for categorization. As discussed in Section *The Neural Basis of Visual Categorization*, a growing number of electrophysiology studies have started to show the limits of this dichotomy. For instance, although the ventral stream is commonly thought of as primarily responsible for the extraction of visual features, significant category signals can already be found in intermediate[133] and low[134] visual areas.

It is very unlikely that category signals found within early ventral stream areas arise from purely bottom-up visual inputs (as predicted by feedforward hierarchical models and most models described in this review). Such category signals are likely to emanate from cortical feedback from higher level visual areas (e.g. Ref 135–137) involved in categorization tasks. Understanding the neural mechanisms used for combining bottom-up sensory-driven information with top-down attention and memory-driven processes will require the development of computational models which integrate feature computation and categorization processes.

Beyond Categorization: Visual Reasoning

Most modern vision architectures including the ones described in Section *Models of Feature Computation* are pre-attentive vision architectures. They perform well in visual categorization tasks that do not seem to require attentional mechanisms such as those based on the detection of diagnostic visual features.[13] At the same time, they are expected to be quite limited in handling categorization tasks that require explicit geometric relationships between features to be represented. This lack of explicit representation of geometric relations is consistent with results from psychophysics experiments using rapid presentation paradigms that have shown that participants often fail to report the location of a target object that was otherwise correctly detected.[10]

Overall, it is expected that these models will fail to capture the scope of human visual categorization capabilities beyond our very glimpse of a visual scene. Indeed, a recent study compared human observers and state-of-the-art machine vision systems for the classification of several artificial object categories.[138] The study demonstrated that modern (pre-attentive) visual architectures related to those surveyed in Section *Models of Feature Computation* perform well in visual categorization tasks that involve relatively rigid objects or objects with diagnostic visual features (assuming enough training examples). However, these same architectures were quite limited in handling categorization problems defined by rules rather than individual shapes as when categorization boils down to a compositional 'rule' (see Ref 138 for details on the tasks).

Challenging tasks for models that are effortless for human participants include the notion of 'sameness' ('when all shapes are the same vs. at least one is different'), 'insideness' (as when a smaller shape is included vs. not included in a larger one) and 'in-

betweeness' ('when the unique shape is in-between the other two vs. not'). It is worth emphasizing that these categorization tasks were not speeded and that images were presented until the participant responded. It is thus very likely that attention is needed to solve these types of reasoning tasks.

While it is becoming increasingly clear that attentional mechanisms will be needed to solve many of these 'reasoning' tasks, a key question for computational neuroscience will be to understand how attention may mechanistically operate on these early preattentive representations to enable more complex inferences derived from compositional rules (see Ref 139 for one proposal exploiting multilevel contextual constraints and Ref 140 for one proposal on how attention can be implemented in a deep recurrent network).

NOTES

[a] There are two main strategies to solve multiclass categorization problems. The one-vs-all approach considers one binary classification for each category versus the rest: k classifiers are considered for each of the k classes and a classification label is obtained by considering the classifier with the largest output. The all-pairs approach (also called one-vs-one) considers one binary classification for each pair: $n(n-1)/2$ classifiers thus need to be considered and a classification label is obtained by voting across all pairs.

[b] Note that the argument above has to do with the ability of a particular classifier to generalize to novel data, which is quite different from the perspective of fitting these models to human behavioral data (see Ref 119 for a treatment of the latter case).

[c] Indeed, Leo Breiman[120] has called this 'the most important unsolved problem in machine learning.'

ACKNOWLEDGMENTS

The author would like to thank Dr. Imri Sofer for helpful comments on the manuscript. This work was supported by NSF early career award [grant number IIS-1252951]. Additional support was provided by DARPA young faculty award [grant number YFA N66001-14-1-4037] and ONR [grant number N000141110743].

REFERENCES

1. Rosch E. Natural categories. *Cogn Psychol* 1973, 4:328–350.
2. Fabre-Thorpe M. The characteristics and limits of rapid visual categorization. *Front Psychol* 2011, 2:243.
3. Potter MC. Recognition and memory for briefly presented scenes. *Front Psychol* 2012, 3:32.
4. Schyns PG, Goldstone RL, Thibaut JP. The development of features in object concepts. *Behav Brain Sci* 1998, 21:1–17, discussion 17–54.
5. Palmeri TJ, Gauthier I. Visual object understanding. *Nat Rev Neurosci* 2004, 5:291–303.
6. Hegdé J. Time course of visual perception: coarse-to-fine processing and beyond. *Prog Neurobiol* 2008, 84:405–39.
7. Treisman A, Gelade G. A feature-integration theory of attention. *Cogn Psychol* 1980, 136:97–136.
8. Wolfe JM, Horowitz TS. What attributes guide the deployment of visual attention and how do they do it? *Nat Rev Neurosci* 2004, 5:495–501.
9. Li F, VanRullen R, Koch C, Perona P. Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci USA* 2002, 99:9596–601.
10. Evans KK, Treisman A. Perception of objects in natural scenes: is it really attention free? *J Exp Psychol Hum Percept Perform* 2005, 31:1476–92.
11. Biederman I. Recognition-by-components: a theory of human image understanding. *Psychol Rev* 1987, 94:115–47.
12. Serre T, Oliva A, Poggio T. A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci USA* 2007, 104:6424–9.
13. VanRullen R. The power of the feed-forward sweep. *Adv Cogn Psychol* 2007, 3:167–176.
14. Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. A neuropsychological theory of multiple systems in category learning. *Psychol Rev* 1998, 105:442–81.
15. Thorpe SJ, Gegenfurtner KR, Fabre-Thorpe M. Detection of animals in natural images using far peripheral vision. *Eur J Neurosci* 2001, 14:869–876.
16. Ungerleider LG, Bell AH. Uncovering the visual 'alphabet': advances in our understanding of object perception. *Vision Res* 2011, 51:782–99.
17. Dicarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? *Neuron* 2012, 73:415–434.
18. Miyashita Y. Inferior temporal cortex: where visual perception meets memory. *Annu Rev Neurosci* 1993, 16:245–263.
19. Hubel DH, Wiesel TN. Early exploration of the visual cortex. *Neuron* 1998, 20:401–12.

20. Seger CA, Miller EK. Category learning in the brain. *Annu Rev Neurosci* 2010, 33:203–19.
21. Mirabella G, Bertini G, Samengo I, Kilavik BE, Frilli D, Della Libera C, Chelazzi L. Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. *Neuron* 2007, 54:303–18.
22. Pinsk MA, DeSimone K, Moore T, Gross CG, Kastner S. Representations of faces and body parts in macaque temporal cortex: a functional MRI study. *Proc Natl Acad Sci USA* 2005, 102:6996–7001.
23. McGugin RW, Gatenby JC, Gore JC, Gauthier I. High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proc Natl Acad Sci* 2012, 109:17063–17068.
24. Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 2008, 60:1126–1141.
25. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. *Nat Neurosci* 2014, 17:455–62.
26. Hung C, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. *Science* 2005, 310:863–866.
27. Kreiman G, Hung C, Kraskov A. Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron* 2006, 49:433–445.
28. Cauchoix M, Crouzet SM, Fize D, Serre T. Fast ventral stream neural activity enables rapid visual categorization. *Neuroimage* 2016, 125:280–290.
29. Sigala N, Logothetis NK. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 2002, 415:318–20.
30. Folstein JR, Palmeri TJ, Gauthier I. Category learning increases discriminability of relevant object dimensions in visual cortex. *Cereb Cortex* 2013, 23:814–23.
31. de Beeck H, Baker C. The neural basis of visual object learning. *Trends Cogn Sci* 2010, 14:1–18.
32. Bar M. Top-down facilitation of visual recognition. *Proc Natl Acad Sci* 2006, 103:449–454.
33. Minamimoto T, Saunders RC, Richmond BJ. Monkeys quickly learn and generalize visual categories without lateral prefrontal cortex. *Neuron* 2010, 66:501–7.
34. Swaminathan SK, Freedman DJ. Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nat Neurosci* 2012, 15:315–20.
35. Kreiman G, Koch C, Fried I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* 2000, 3:946–953.
36. Van Le Q, Isbell LA, Matsumoto J, Nguyen M, Hori E, Maior RS, Tomaz C, Tran AH, Ono T, Nishijo H. Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. *Proc Natl Acad Sci USA* 2013, 110:19000–5.
37. Mormann F, Dubois J, Kornblith S, Milosavljevic M, Cerf M, Ison M, Tsuchiya N, Kraskov A, Quiroga RQ, Adolphs R, et al. A category-specific response to animals in the right human amygdala. *Nat Neurosci* 2011, 14:1247–1249.
38. Cauchoix M, Crouzet SM. How plausible is a subcortical account of rapid visual recognition? *Front Hum Neurosci* 2013, 7:1–4.
39. Liu H, Agam Y, Madsen JR, Kreiman G. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 2009, 62:281–290.
40. Pessoa L, Adolphs R. Emotion and the brain: multiple roads are better than one. *Nat Rev Neurosci* 2011, 12:425–425.
41. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci* 1999, 2:1019–25.
42. Ghebreab S, Smeulders A, Scholte H, Lamme VAF. A biologically plausible model for rapid natural image identification. In: *Advances in Neural Information Processing System*. NIPS Proceedings; 2009, 1–9.
43. Scholte H, Ghebreab S, Waldorp L, Smeulders A, Lamme VAF. Brain responses strongly correlate with Weibull image statistics when processing natural images. *J Vis* 2009, 9:1–15.
44. Wiskott L, Fellous JM, Krger N, von der Malsburg C. Face recognition by elastic bunch graph matching. *IEEE Trans Pattern Anal Mach Intell* 1997, 19:775–779.
45. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vis* 2001, 42:145–175.
46. Pinto N, Doukhan D, DiCarlo JJ, Cox DD. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol* 2009, 5:e1000579.
47. Renninger L, Malik J. When is scene identification just texture recognition? *Vision Res* 2004, 44:2301–2311.
48. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis* 2009, 9:1–18.
49. Freeman J, Simoncelli E. Metamers of the ventral stream. *Nat Neurosci* 2011, 14:1195–1201.
50. Kimia BB. On the role of medial geometry in human vision. *J Physiol Paris* 2003, 97:155–90.
51. Feldman J, Singh M. Bayesian estimation of the shape skeleton. *Proc Natl Acad Sci USA* 2006, 103:18014–9.

52. Fukushima K. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans Syst Man Cybern* 1983, 13:826–834.
53. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell* 2007, 29:411–26.
54. Yue X, Biederman I, Mangini MC, Malsburg CVD, Amir O. Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Res* 2012, 55:41–6.
55. Shan H, Cottrell GW. Looking around the backyard helps to recognize faces and digits. *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK: IEEE; 2008.
56. Pinto N, DiCarlo J, Cox D. How far can you get with a modern face recognition test set using only simple features? *2009 I.E. Conference Computer Vision Pattern Recognition*, 2591–2598, 2009.
57. Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition. *Prog Brain Res* 2006, 155:23.
58. Xiao J, Hays J, Ehinger K. Sun database: large-scale scene recognition from abbey to zoo. *IEEE Conference Computer Vision Pattern Recognition*, 3485–3492, IEEE, 2010.
59. Mack ML, Palmeri TJ. Modeling categorization of scenes containing consistent versus inconsistent objects. *J Vis* 2010, 10:11.1–11.
60. Ullman S. Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn Sci* 2007, 11:58–64.
61. Lerner Y, Epshtein B, Ullman S, Malach R. Class information predicts activation by object fragments in human object areas. *J Cogn Neurosci* 2008, 20:1189–1206.
62. Julesz B. Textons, the elements of texture perception, and their interactions. *Nature* 1981, 290:91–97.
63. Arbelaez P, Maire M, Fowlkes C, Malik J. Contour detection and hierarchical image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2010, 33:898–916.
64. Portilla J, Simoncelli EP. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int J Comput Vis* 2000, 40:49–71.
65. Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. A functional and perceptual signature of the second visual area in primates. *Nat Neurosci* 2013, 16:974–81.
66. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015, 521:436–444.
67. Wallis G, Rolls ET. A model of invariant recognition in the visual system. *Prog Neurobiol* 1997, 51:167–194.
68. Mel BW. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput* 1997, 9:777–804.
69. Wersing H, Koerner E. Learning optimized features for hierarchical models of invariant recognition. *Neural Comput* 2003, 15:1559–1588.
70. Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T. A quantitative theory of immediate visual recognition. *Prog Brain Res* 2007, 165:33–56.
71. Masquelier T, Thorpe SJ. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol* 2007, 3:e31.
72. Serre T., Hierarchical models of the visual system. Encyclopedia of Computational Neuroscience, 2014.
73. Riesenhuber M, Poggio T. Models of object recognition. *Nat Neurosci* 2000, 3:1199–204.
74. Serre T, Poggio T. A neuromorphic approach to computer vision. *Commun Assoc Comput Mach* 2010, 53:54.
75. Crouzet SM, Serre T. What are the visual features underlying rapid object recognition? *Front Psychol* 2011, 2:326.
76. Ghodrati M, Farzmahdi A, Rajaei K, Ebrahimpour R, Khaligh-Razavi S-M. Feedforward object-vision models only tolerate small image variations compared to human. *Front Comput Neurosci* 2014, 8:74.
77. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol* 2014, 10:35.
78. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA* 2014, 111:8619–24.
79. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol* 2014, 10: e1003915.
80. Masquelier T, Serre T, Poggio T. Learning complex cell invariance from natural videos: a plausibility proof. Technical Report. Massachusetts Institute of Technology, Cambridge MA, 2007.
81. Ghodrati M, Khaligh-Razavi S-M, Ebrahimpour R, Rajaei K, Pooyan M. How can selection of biologically inspired features improve the performance of a robust object recognition model? *PLoS One* 2012, 7: e32357.
82. Kheradpisheh SR, Ganjtabesh M, Masquelier T. Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition. *CoRR*, abs/1504.0, 2015.

83. Li N, DiCarlo JJ. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 2008, 321:1502–1507.
84. Li N, DiCarlo JJ. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 2010, 67:1062–1075.
85. Li N, Dicarlo JJ. Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *J Neurosci* 2012, 32:6611–20.
86. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. *CoRR*, abs/1409.4, 2014.
87. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. arXiv:1502.01852, 2015.
88. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. *CoRR*, abs/1409.0 2014, 43.
89. Everingham M, Van Gool L, Williams C, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *Int J Comput Vis* 2010, 88:303–338.
90. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol* 2014, 10:e1003963.
91. Guclu U, van Gerven MAJ. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J Neurosci* 2015, 35:10005–10014.
92. Poggio T, Smale S. The mathematics of learning: dealing with data. *Not Am Math Soc* 2003, 50:537–544.
93. Bishop C. *Pattern Recognition and Machine Learning*. New York: Springer; 2007.
94. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 7nd ed. New York: Springer; 2013.
95. Vapnik V. *Statistical Learning Theory*. New York, NY:John Wiley & Sons Inc.; 1998.
96. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958, 65:386–408.
97. McCulloch W, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 1943, 5:115–133.
98. Ashby F. A stochastic version of general recognition theory. *J Math Psychol* 2000, 44:310–329.
99. Haykin S. *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Macmillan Publishing Company; 1994.
100. Poggio T. A theory of how the brain might work. *Cold Spring Harb Symp Quant Biol* 1990, 55:899–910.
101. Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. *Curr Biol* 1995, 5:552–63.
102. Edelman S. *Representation and Recognition in Vision*. Cambridge, MA: MIT Press; 1999.
103. Nosofsky R, Palmeri T, McKinley S. Rule-plus-exception model of classification learning. *Psychol Rev* 1994, 101:53–79.
104. Nosofsky R, Palmeri T. A rule-plus-exception model for classifying in continuous-dimension spaces. *Psychon Bull Rev* 1998, 5:345–369.
105. Ashby FG, Waldron EM. On the nature of implicit categorization. *Psychon Bull Rev* 1999, 6:363–378.
106. Maddox WT, Ashby FG. Comparing decision bound and exemplar models of categorization. *Percept Psychophys* 1993, 53:49–70.
107. Posner MI, Keele SW. On the genesis of abstract ideas. *J Exp Psychol* 1968, 77:353–363.
108. Anderson JR. The adaptive nature of human categorization. *Psychol Rev* 1991, 98:409–429.
109. Love BC, Medin DL, Gureckis TM. SUSTAIN: a network model of category learning. *Psychol Rev* 2004, 111:309–32.
110. Kruschke JK. ALCOVE: an exemplar-based connectionist model of category learning. *Psychol Rev* 1992, 99:22–44.
111. Nosofsky RM, Palmeri TJ. An exemplar-based random walk model of speeded classification. *Psychol Rev* 1997, 104:266–300.
112. Nosofsky R, Palmeri T. An exemplar-based random-walk model of categorization and recognition. In: Busemeyer J, Townsend J, Wang Z, Eidels A, eds. *Oxford Handbook Computational and Mathematical Psychology*. Oxford, UK: Oxford University Press; 2015, 142–164.
113. Ashby FG, Maddox WT. Human category learning 2.0. *Ann N Y Acad Sci* 2011, 1224:147–61.
114. Love BC. Category learning, computational perspectives. In: Pashler H, eds. *Encyclopedia of the Mind*. London: Sage; 2013.
115. Richler JJ, Palmeri TJ. Visual category learning. *WIREs Cogn Sci* 2014, 5:75–94.
116. Jäkel F, Schölkopf B, Wichmann FA. Generalization and similarity in exemplar models of categorization: insights from machine learning. *Psychon Bull Rev* 2008, 15:256–271.
117. Nosofsky RM. Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen* 1986, 115:39–61.

118. Valentine T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A* 1991, 43:161–204.

119. Myung JI, Tang Y, Pitt MA. Evaluation and comparison of computational models. *Methods Enzymo*. 2009, 454:287–304.

120. Breiman L. Population theory for boosting ensembles. *Ann Stat* 2004, 32:1–11.

121. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput* 1992, 4:1–58.

122. Rosch E. Principles of categorization. In: Rosch E, Lloyd BB, eds. *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum; 1978, 27–48.

123. Tanaka JW. The entry point of face recognition: evidence for face expertise. *J Exp Psychol Gen* 2001, 130:534–43.

124. Grill-Spector K, Kanwisher N. Visual recognition as soon as you know it is there, you know what it is. *Psychol Sci* 2005, 16:152–160.

125. Loschky LC, Larson AM. The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Vis Cogn* 2010, 18:513–536.

126. Kadar I, Ben-Shahar O. A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *J Vis* 2012, 12:1.

127. Liu J, Harris A, Kanwisher N. Stages of processing in face perception: an MEG study. *Nat Neurosci* 2002, 5:910–916.

128. Tsao DY, Livingstone MS. Mechanisms of face perception. *Annu Rev Neurosci* 2008, 31:411–437.

129. Jolicoeur P, Gluck MA, Kosslyn SM. Pictures and names: making the connection. *Cogn Psychol* 1984, 16:243–275.

130. Greene MR, Oliva A. The briefest of glances: the time course of natural scene understanding. *Psychol Sci* 2009, 20:464–72.

131. Sofer I, Crouzet S, Serre T. Explaining the timing of natural scene understanding with a computational model of perceptual categorization. *PLoS Comput Biol* 2015, 11:e1004456.

132. Macé MJ-M, Joubert OR, Nespoulous J-L, Fabre-Thorpe M. The time-course of visual categorizations: you spot the animal faster than the bird. *PLoS One* 2009, 4:e5927.

133. Kilavik BE, Frilli D, Libera CD, Mirabella G, Bertini G, Chelazzi L, Samengo I, Dellalibera C. Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. *Neuron* 2007, 54:303–318.

134. Hsieh P-J, Vul E, Kanwisher N. Recognition alters the spatial pattern of FMRI activation in early retinotopic cortex. *J Neurophysiol* 2010, 103:1501–1507.

135. Bullier J. Integrated model of visual processing. *Brain Res Brain Res Rev* 2001, 36:96–107.

136. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A* 2003, 20:1434–1448.

137. Bar M. Visual objects in context. *Nat Rev Neurosci* 2004, 5:617–629.

138. Fleuret F, Li T, Dubout C, Wampler EK, Yantis S, Geman D. Comparing machines and humans on a visual categorization test. *Proc Natl Acad Sci USA* 2011, 108:17621–5.

139. Chang L, Jin Y, Zhang W, Borenstein E, Geman S. Context, computation, and optimal ROC performance in hierarchical models. *Int J Comput Vis* 2011, 93:117–140.

140. Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention. arXiv:1412.7755, 2014.

