Comparing machines and humans on a visual categorization test

François Fleuret[a,b,1], Ting Li[c], Charles Dubout[a,b], Emma K. Wampler[d], Steven Yantis[d], and Donald Geman[c]

[a]Idiap Research Institute, 1920 Martigny, Switzerland; [b]École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [c]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218; and [d]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218

Edited* by David Mumford, Brown University, Providence, RI, and approved September 8, 2011 (received for review June 13, 2011)

Automated scene interpretation has benefited from advances in machine learning, and restricted tasks, such as face detection, have been solved with sufficient accuracy for restricted settings. However, the performance of machines in providing rich semantic descriptions of natural scenes from digital images remains highly limited and hugely inferior to that of humans. Here we quantify this "semantic gap" in a particular setting: We compare the efficiency of human and machine learning in assigning an image to one of two categories determined by the spatial arrangement of constituent parts. The images are not real, but the category-defining rules reflect the compositional structure of real images and the type of "reasoning" that appears to be necessary for semantic parsing. Experiments demonstrate that human subjects grasp the separating principles from a handful of examples, whereas the error rates of computer programs fluctuate wildly and remain far behind that of humans even after exposure to thousands of examples. These observations lend support to current trends in computer vision such as integrating machine learning with parts-based modeling.

abstract reasoning | human learning | pattern recognition

Image interpretation, effortless and instantaneous for people, remains a fundamental challenge for artificial intelligence. The goal is to build a "description machine" that automatically annotates a scene from image data, detecting and describing objects, relationships, and context. It is generally acknowledged that building such a machine is not possible with current methodology, at least when measuring success against human performance.

Some well-circumscribed problems have been solved with sufficient speed and accuracy for real-world applications. Almost every digital camera on the market today carries a face detection algorithm that allows one to adjust the focus according to the presence of humans in the scene; and machine vision systems routinely recognize flaws in manufacturing, handwritten characters, and other visual patterns in controlled industrial settings.

However, such cases usually involve a single quasi-rigid object or an arrangement of a few discernible parts and thus do not display many of the complications of full-scale "scene understanding." Moreover, achieving high accuracy usually requires intense "training" with gigantic amounts of data. Systems that attempt to deal with multiple object categories, high intraclass variability, occlusion, context, and unanticipated arrangements, all of which are easily handled by people, typically perform poorly. Such visual complexity seems to require a form of global reasoning that uncovers patterns and generates high-level hypotheses from local measurements and prior world knowledge.

In order to go beyond general observation and speculation, we have designed a controlled experiment to measure the difference in performance between computer programs and human subjects. The Synthetic Visual Reasoning Test (SVRT) is a series of 23 classification problems involving images of randomly generated shapes; see Fig. 1. Whereas many factors affect the performance of both machines and people in analyzing real images, the SVRT is designed to focus on one in particular—abstract reasoning. As a result, we have purposely removed many of the subtasks and complications encountered in parsing images acquired from natural scenes: There is no need to recognize natural objects or to account for volume, illumination, texture, shadow, or noise. Moreover, being planar and randomly generated, the shapes are "unknown" to humans, which ameliorates our advantage over machines due to extensive experience with everyday objects and a three-dimensional world.

For each problem there are two disjoint "categories" of images. Fig. 1 displays one example from each category for eight of the 23 problems. Classification is at the level of relationships, not individual shapes; the difference between the two categories boils down to a compositional "rule." Several of these rules are illustrated in Fig. 1, including "inside," "in between," and "same." For each category in each problem we can generate as many examples of images as desired. Formally, in fact, each category is defined by a probability distribution over images (see *Methods*), and generating an image from the category means calling a computer program to sample from the corresponding distribution.

Assessing how well machines can perform is less straightforward than with people. Computer vision supports a wide variety of competing paradigms (see *Discussion*). One approach is supervised and unstructured machine learning: a computer program whose input is a given set of images together with their true labels and whose output is a decision rule for labeling a new image (see *Methods*). This approach accounts for many of the success stories in computer vision (e.g., cell phone face detectors). Other prominent strategies for building image interpretation machines include constructing stochastic, generative image models, likelihood-based statistical inference, and designing biologically-inspired hierarchical models, as well as many hybrids of such models and machine learning (see *Discussion*).

Our goal here is to see what "off-the-shelf" machine learning technology can do—namely, methods that do not require customized tuning for the SVRT. We are interested in both accuracy and learning efficiency, meaning the number of training examples necessary to either "grasp the rule" (for humans) or reach a given level of accuracy (machines).

SVRT

For each of the 23 problems the objective is to assign an observed 128×128 binary image I to one of two categories. From a human perspective, the SVRT is designed so that the two categories can be perfectly separated once the underlying rule is understood. See *Methods* for the technical definitions of the categories in the context of statistical learning.

Author contributions: F.F., S.Y., and D.G. designed research; T.L., C.D., and E.K.W. performed research; F.F., E.K.W., and D.G. analyzed data; and F.F., S.Y., and D.G. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed: E-mail: francois.fleuret@idiap.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1109168108/-/DCSupplemental.

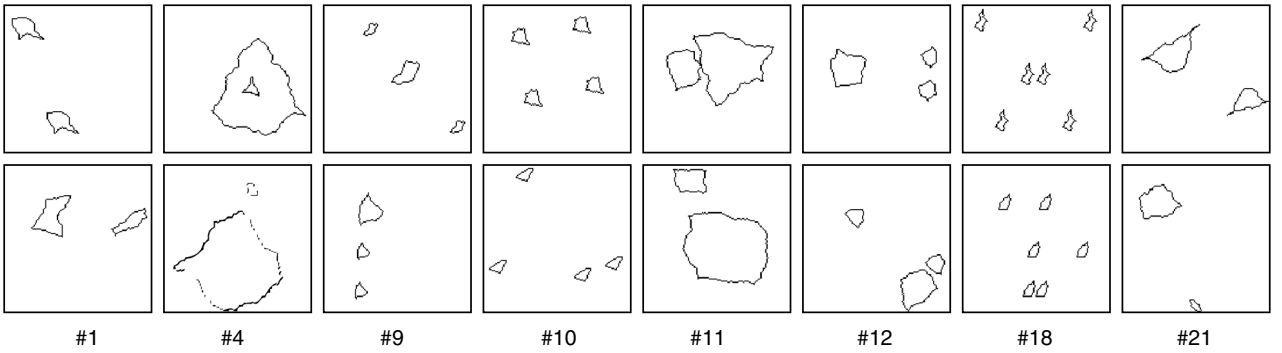

Fig. 1. A selection of visual categorization problems. One instance from Category 1 (top) and one instance from Category 2 (bottom) is shown for each of eight different problems. Each instance is a binary image of resolution 128×128 pixels. In problem #1 both categories are represented by two randomly generated and randomly positioned shapes; the difference is that the two shapes are identical in Category 1. The underlying differentiating "rule" in problem #4 is "outside vs. inside," whereas in problem #9 the largest of the three shapes is "in between" the two smaller ones in Category 1 but not in Category 2 and in problem #21 one of the shapes in category one can be made to coincide with the other one by translating, scaling, and rotating. The difference between the two categories in each of the other four problems can also be "explained" in terms of concepts such as distance, symmetry, and reflection. Multiple instances of each of the 23 problems can be found in the *SI Appendix*.

In each case, what distinguishes the two categories is some gestalt-like property of the global spatial arrangement of parts, which are randomly generated, highly irregular closed contours (see Fig. 1). The number of possible parts is very large, and brute-force memorization of those already seen serves no purpose. Indeed, the categories cannot be separated based on the appearance, spatial positioning, or any other geometric or topological property of individual parts. Separation must be "holistic" in the sense of discovering the principles that determine how the parts are combined into a global pattern. A complete description of the SVRT, including illustrations and an expanded discussion of the correspondences between problems and concepts, appears in *SI Appendix*.

Needless to say, these parts and arrangements represent a gross oversimplification of the natural world. In addition to the absence of intensity variations, pixel-level noise and other properties of natural images, real physical components such as limbs, leaves, handles, and windows are individually recognizable and help us to identify the categories to which they belong. Here, by design, the individual shapes are not meaningful. Nonetheless, parsing visual data also involves detecting organizational principles similar to those underlying the SVRT (proximity, similarity, symmetry, etc.); indeed, the same "part" may appear in many different objects, and parts themselves are typically composed of subparts that may not be so easy to recognize except in the context of other parts. Many would argue (see *Discussion*) that the ability of humans to annotate scenes with words derives at least partially from the ability to evaluate the plausibility of arrangements of parts at many scales and levels of semantic resolution (1–4).

Hence the simplicity of the images in the SVRT, and the fact that the parts are very weakly informative about the category, necessitates that whatever "reasoning" is to occur must take into account the types of rules listed above, which are involved in most challenging computer vision problems.

Results

Human Experiments. Each participant completed the same 23 problems in a random order. For each problem, they were shown one instance at a time selected randomly with equal probability from either a set of instances that satisfied the current rule (i.e., one category) or a set that did not satisfy the rule (i.e., from the other category). See Fig. 1 for examples and *Methods* for a more detailed description of the stimuli. The participant assigned the instance to one of the two categories. Feedback was provided after each response, and all instances viewed so far for that problem remained on the screen, clustered according to their correct categorization, so they could be used as the problem progressed to help learn the rule (see Fig. 2 and *Methods* for more details).

Fig. 3 summarizes human performance on this task. The mean number of instances required to learn each rule (see *Methods*) is plotted in Fig. 3*A* against the number of subjects (out of 20) who failed to learn that rule; these two measures were highly correlated ($r = 0.929, p < 001$). Performance on four of the problems was categorically poorer than on the rest; these "hard" problems correspond to the four points clustered in the upper right of Fig. 3*A* and are identified in *SI Appendix*. Participants viewed an average of 6.27 ± 0.85 instances before successfully learning each rule. Fig. 3*B* shows the frequency of number of instances viewed before learning the rule for the 397 successfully learned rules (23 rules $\times$ 20 subjects—63 failures). Seventeen of the 20 participants successfully learned 19 or more of the 23 problems.

Machine Experiments. We have used two popular machine learning algorithms in our experiments: boosting with the standard Adaboost procedure (5) and a support vector machine with a Gaussian kernel (6). See *Methods* for the experimental settings, including parameter choices and image preprocessing (feature design). We observed lower error rates with boosting, and the

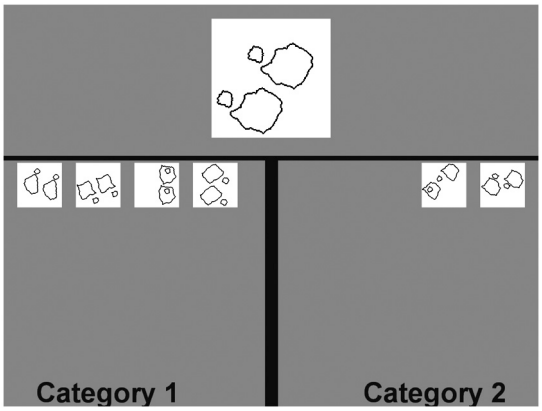

Fig. 2. Example screen shot of the interface for the human experiments. This participant is working on problem #13, has already classified six training instances (whether right or wrong), and is considering the seventh one, displayed at the top center. The previous instances are shown correctly classified on the left (Category 1) and right (Category 2). Responses were unspeeded. A session for a given problem and individual terminates following either a success (7 correct responses in a row) or failure (35 instances categorized without success).

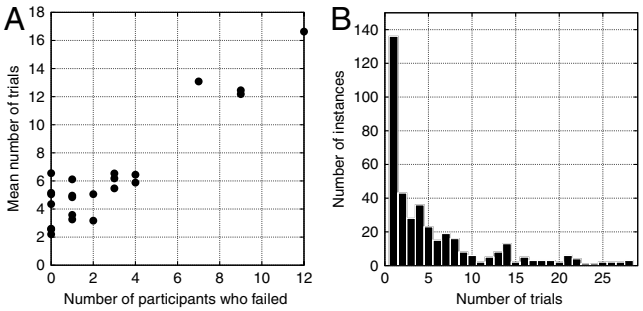

Fig. 3. Summary of human performance. There were 20 participants and 23 problems for a total of 460 attempts. Of these, 63 were not successful. (*A*) Given successful learning, the mean number of trials required to learn the rule plotted against the number of participants who failed to learn the rule. Each point is a problem. (*B*) Distribution of the number of instances required for participants to successfully learn the rule (i.e., correctly categorize seven subsequent instances without error) over the 397 successful attempts.

results reported in this section were obtained with this algorithm; additional results with the support vector machine are given in *SI Appendix*.

Performance varies considerably depending on the problem and the number of training examples, with prediction rates spanning the full range from 0% to 50%, as shown in Fig. 4. However, some trends are evident. Performance strictly increases with both the number of training samples (Fig. 4*A*) and the complexity of the image processing in terms of the richness of the features extracted from the raw binary images prior to machine learning (Fig. 4*B*).

With only 10 examples of each category for training, the error remains virtually at 50% for every problem. Some problems could not be solved with even 10,000 training examples, with the error rate remaining above 25% for six problems (Fig. 4*A*). This is in sharp contrast with human performance (see *Discussion*). As for the effect of the choice of features, for multiple problems the error rate dropped from values above 30% to values below 5% when moving from the simplest image processing scheme to the most complex one, in some cases because adding Fourier features exposed symmetries (see Fig. 4*B* and problems #1, #16, and #22 in *SI Appendix*).

Discussion

The SVRT exhibits patterns that are easy to spot and characterize for humans and extremely difficult to learn for generic machine learning systems. Humans solved the problems after seeing fewer than 20 examples in most cases. After seeing at most a few tens of examples, and usually many fewer, more than 90% of the participants solved 14 of the 23 problems, and another group of five problems was solved by 75%. In contrast, even with 10,000 examples for training and complex image preprocessing, the boosting machine learning algorithm was only able to solve 11 of the problems at an error rate below 10% and another five with an error rate below 25% (see Fig. 5 for comparison with human performance). If the number of training examples is of the same order as for human learning, the machine performance amounts to random guessing.

Still, it has been well-known for a long time from the theory of nonparametric inference that even naive machine learning techniques, such as nearest-neighbor classification, can achieve optimal performance in the large-sample limit (7, 8). And we do observe a marked improvement as the number of training examples increases, albeit on a log scale. However, the results are still far from optimal (zero error rate) even after 10,000 examples.

People tend to characterize a category in phrases such as: "the two shapes are in contact," "the two halves of the picture are symmetric," "the shapes are aligned with the large one between two small ones." Many of the rules that distinguish instances in each category of a given problem instantiate one or more of the Gestalt principles of perceptual organization (9, 10), including proximity, similarity, symmetry, inclusion, collinearity, and others. They are higher-order (nonlocal) configural properties of the displays that biological visual systems have evolved to perceive effortlessly as part of scene understanding. However, we are not drawing conclusions about natural visual recognition from the performance of humans on the SVRT. In addition to the fact that the SVRT images are nothing like natural images, it is not clear whether the observed performance of humans on the SVRT would be maintained by young children, under brief time exposures, or without batch learning or any training in logical reasoning.

Due to the black box nature of the computer algorithms, which learn very high-dimensional decision boundaries, it is difficult to measure the extent to which the resulting classifiers "understand" the categories. In particular, the machine learning methods we have used cannot directly extract and process information about the overall geometry of the scene. They must "learn" solely from elementary, statistical local measurements (see *Methods*). They only have direct access to information of the sort "there is an elongated dark area," "the black pixels are spread out," "there is a patch with edges all in the same direction." In particular, there is no platform for learning concepts like "symmetric" or "aligned"—i.e., no hard-wired mechanism for constructing an

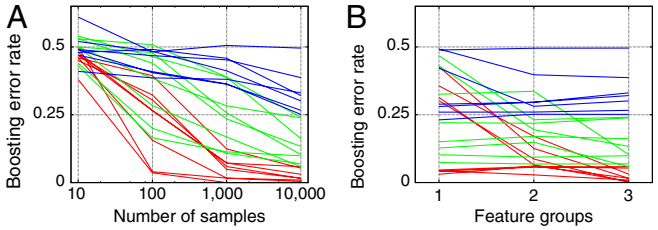

Fig. 4. Summary of machine performance. Both graphs show error rates on the 23 problems, organized in three arbitrary subgroups ranked by difficulty. The blue group contains all the problems for which the machine learning reached a final error rate greater than 25%, the red group contains the problems for which the best error rate was lower than 6%. The graph (*A*) shows the error rate as a function of the number of examples available for training, using all the image features. The error rate fluctuates around 50% when only 10 examples are used, and for almost all problems, the error rate decreases sharply when the number of samples increases. Graph *B* shows the error rates with 10,000 training examples as a function of the complexity of the image features used. The features in group 1 compute the number of black pixels over rectangular areas of varying sizes, those in group 2 are based on edge statistics, and those of group 3 are related to the spectral properties of the image (Fourier and wavelet coefficients).

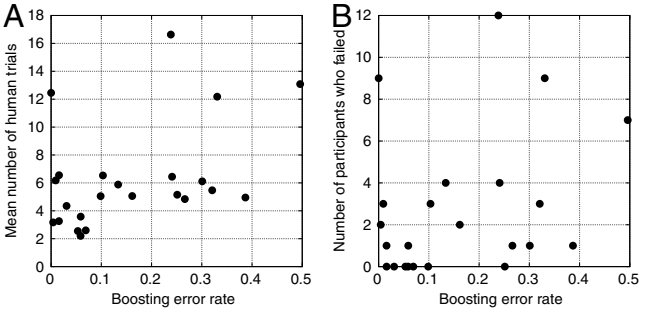

Fig. 5. Comparison of human and machine performance in terms of learning efficiency and error rate. In both plots, the horizontal axis is the error rate of the predictor trained with boosting and 10,000 instances per problem and the most complete feature set. The vertical axis in *A* is the average number of samples necessary for human participants to learn the rule, and the vertical axis in *B* is number of participants (out of 20) failing to learn the concept after 35 examples.

abstract, category-specific model of the arrangements of shapes. Instead, such methods rely on observing a great many sample configurations in each category in order to reach even modest error rates.

There is one aspect of the machine learning that can be deconstructed by observing the features selected by the boosting algorithm: an exploitation of statistical cues that may at first seem irrelevant to the actual structure of the problem. Whereas we attempted to eliminate gross intensity differences between the categories (e.g., equalizing the average number of black pixels), many "tells" slipped through. For instance, samples of problem #8 are composed of two closed shapes of different size, with the small one enclosed by the large one in the first category but not in the second category. As a result, the categories can be separated with a very crude test on the variance of the black pixel locations, these being more "spread out" in category two. In fact, simply thresholding the standard statistical measure of variance yields an error rate of only 9%. The same "trick" applies to several other problems, and similar tells can be used to detect symmetry with respect to a centered axis, because the distribution of black pixels is more dispersed horizontally and more peaked vertically. In the end, a few global, exact geometrical properties are perceived through a multitude of cues reflecting small statistical differences between the distribution of mass in the two categories.

In computer vision there is a long history of variations on the intuitive strategy of decomposing complex entities into their constituent parts in order to facilitate recognizing common objects in complex natural scenes (11). For instance, a car is composed of wheels, doors, windows, and other components that are individually recognizable and that come together with a preferred geometry. Whereas there has been considerable progress in object recognition based on explicitly compositional models (1, 2, 4), as well as some biologically inspired ones (12), the most popular techniques until recently (13–15) were surprisingly closer to pure machine learning without explicitly introducing either compositions or invariance to geometric deformations. Such methods do not accommodate variability other than changes in illumination and local deformations. Only multilayer neural networks have been leveraging more complex models, which can be seen as parts and composition of parts (16).

More recently, machine learning methods have evolved progressively toward the part-based techniques, either by combining simple part characterizations with absolute constraints on their locations (17) or by introducing latent variables related to the location of parts (3, 18, 19). In fact, methods currently considered state-of-the-art on canonical benchmarks belong to this family and combine discriminative part detectors with simple models of arrangements of parts (20). Also, training procedures for multilayer neural networks have been improved to leverage large sets of unsupervised data, which allows one to discover richer latent structures (21).

In summary, we have demonstrated the poor performance of model-free machine learning, both in absolute terms and relative to humans, on visual tasks designed to require abstract reasoning about scene constituents. People learn far faster and perform far better than machines on the SVRT, and machines appear to lack the proper representations to handle abstract reasoning. Whereas these observations lend support to current trends in computer vision to merge tabula rasa machine learning with hierarchical image models, it is still doubtful that any current method could match human performance on the SVRT, namely near-perfect categorization with at most tens of examples.

Methods

Human Experiments. Twenty members of the Johns Hopkins University community (14 women and 6 men, ages 18–21) each participated in a one-hour session and received partial course credit. All participants had either normal or corrected-to-normal vision. Each participant signed an informed consent form and participated under a protocol that was approved by the JHU Homewood Institutional Review Board.

For each pattern classification problem, binary images containing configurations of shapes (see Fig. 1 for examples) were displayed one at a time and classified as one of two categories. Images from category one were considered to satisfy some discriminating rule. Participants had to learn the problem-dependent classification rule by trial and error. Stimuli were presented and responses collected using a custom script written with the PsychToolbox extension of MATLAB on a PC.

For each problem, the participant first saw an instance subtending 6.7 ° of visual angle in the upper center of the screen (Fig. 2); it remained on the screen until the participant responded. The participant pressed one key to indicate that the exemplar belonged to category one (satisfied the rule) and another key to indicate that it did not—i.e., belonged to category two (did not satisfy the rule). The responses were unspeeded. Following their response, feedback text appeared (either "Correct!" or "Incorrect"), and the current instance then appeared (along with previous instances from that problem) within a box on the lower left of the computer screen if it was in fact an instance that satisfied the current rule, or lower right if not. These previously seen stimuli (subtending 3.4 ° of visual angle) remained on the screen throughout the rest of the current problem, so the participant could refer to them as they worked on the current problem. The feedback text remained on the screen for 0.9 sec before the next instance appeared.

Participants continued classifying images until they made seven correct responses in a row (counted as a "success") or until they had seen a total of 35 instances without success (counted as a "failure"). They received feedback for the problem ("Good job!" or "Nice try") and were then prompted to press the space bar on the keyboard when they were ready to move on to the next problem.

Each participant completed the same 23 classification problems. Participant 1 completed the problems in a random order, and participant 2 completed the problems in the reverse order; participant 3 completed the problems in a new random order, and participant 4 in the reverse order; and so forth. A large pool of instances from each category was randomly generated for each problem using the algorithm described in the text. Each instance was shown only once in the entire experiment.

Machine Experiments. Each problem is represented by two probability distributions P_1 and P_2 over binary images. These distributions define the two categories: If $P_1(I) > 0$ (resp., $P_2(I) > 0$) then image I belongs to category one (resp., category two), and no image satisfies both positivity conditions. In the language of statistical learning, the Bayes error rate is zero. Sampling from P_1 and P_2 is very simple, and consequently one can generate as many independent instances as desired in order to assess the effect of the number of training examples—previously seen correctly labeled instances—on the ability of either a human or machine to correctly classify a new sample.

Machine learning techniques combine two modules addressing complementary aspects of the problem. The first module is called a feature extractor. It is hand-designed and remains unmodified during learning. The purpose is to compute numerical properties of the raw image data that may be useful in discriminating between the two categories. A useful property is then one whose typical values are appreciably different from one subpopulation of images to another in a statistical sense.

The second module is the machine learning algorithm per se. The input is the list of numerical values computed by the feature extractor (the "feature vector"), and the output is the predicted category for the feature vector. The decision rule is characterized by a very large number of parameters (weights, synaptic coefficients, etc.), and training consists in optimizing these parameters so that predictions on the training data are as consistent as possible with the known labels of the images. As a result, the nature of decision-making is usually difficult to describe in ordinary language ("black boxes").

We used two standard learning methods: boosting of stumps and SVM with a Gaussian kernel. Each method was trained with three different groups of features of increasing complexity. We did not use features that require training. The features of group 1 compute the number of black pixels in a rectangular subregion of the image for a large number of such regions; those in group 2 also gather information about the distribution of edges (sharp local transitions) in the image; and those of group 3 add spectral properties of the image (Fourier and wavelet coefficients). All of these features are generic and are not dedicated or tuned to the types of images or category differences.

The boosting method is standard Adaboost with feature sampling. During training, it iteratively selects 1,000 "stumps," each defined by a feature, a threshold, and a signed weight. For each stump, we sample 100 features and compute for each the optimal threshold and weight. Because the features are organized into families, we sample each feature by first picking

a family at random uniformly, and then a feature at random in this family, also uniformly. This procedure ensures a uniform sampling among families of features, despite their strong difference of cardinality. During testing, each stump votes for one of the two categories depending on the observed value of feature relative to the learned threshold. Hence, for instance, if a particular feature usually has higher values for images in the first category, the boosting algorithm may select it and assign it a positive weight.

The second technique we used is a support vector machine with a Gaussian kernel. It associates a weight with each training example and classifies a test image depending on its similarity to the training samples and their weights. The measure of similarity between images depends on the features. For computational reason, we do not work with all the features but randomly sample 10,000 of them. In addition to the choice of features, the SVM depends critically on two parameters: the penalty constant C and the variance σ of the Gaussian kernel. Both choices were optimized through fivefold cross-validation. No prefiltering of the features was performed for the SVMs. This may explain in part the limited improvement in performance with more complex feature groups when compared with boosting.

ACKNOWLEDGMENTS. F.F. was supported by European Union Grant Agreement 247022—MASH; C.D. by the Swiss National Science Foundation under Grant 200021-124822—VELASH; S.Y. by National Institutes of Health Grant R01-DA013165; and D.G. by Office of Naval Research Contract N00014-07-1-1002 and National Science Foundation Grant 0427223.

1. Ommer B, Buhmann JM (2010) Learning the compositional nature of visual object categories for recognition. *IEEE Trans Pattern Anal Mach Intell* 32:501–516.
2. Zhu S, Mumford D (2006) A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, (Now Publishers), 2, pp 259–362.
3. Amit Y, Trouvé A (2007) POP: Patchwork of parts models for object recognition. *Int J Comput Vis* 75:267–282.
4. Chang LB, Jin Y, Zhang W, Borenstein E, Geman S (2010) Context, computation, and optimal roc performance in hierarchical models. *Int J Comput Vis* 93:117–140.
5. Freund Y, Schapire RE (1999) A short introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence* 14:771–780.
6. Vapnik VN (1995) *The Nature of Statistical Learning Theory* (Springer, New York).
7. Duda R, Hart P (1973) *Pattern Classification and Scene Analysis* (John Wiley & Sons, New York).
8. Grenander U (1980) *Abstract Inference* (John Wiley & Sons, New York).
9. Wertheimer M (1923) Laws of organization in perceptual forms. *Psycologische Forschung* 4:301–350 (in German) Translated excerpts published in Ellis, WD, ed. (1999) *A Source Book of Gestalt Psychology* (Routledge, London), pp 71–88.
10. Metzger W (1936) *Gesetze des Sehens [Laws of seeing]* (MIT Press, Cambridge, MA), Translated and reprinted in 2006.
11. Fischler MA, Elschlager RA (1973) The representation and matching of pictorial structures. *IEEE Trans Comput* 22:67–92.
12. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025.
13. Rowley HA, Baluja S, Kanade T (1998) Neural network-based face detection. *IEEE Trans Pattern Anal Mach Intell* 20:23–28.
14. Viola P, Jones MJ (2004) Robust real-time face detection. *Int J Comput Vis* 57:137–154.
15. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp 886–893 .
16. LeCun Y, Bengio Y (1994) Word-level training of a handwritten word recognizer based on convolutional neural networks. *Proceedings of the 12th IAPR International Conference on Computer Vision and Image Processing*, Vol. II pp:88–92.
17. Grauman K, Darrell T (2005) The pyramid match kernel: Discriminative classification with sets of image features. *IEEE International Conference on Computer Vision* pp 1458–1465.
18. Li F, Fergus R, Perona P (2003) A Bayesian approach to unsupervised one-shot learning of object categories. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2 p:1134.
19. Huttenlocher D, Felzenszwalb P (2005) Pictorial structures for object recognition. *Int J Comput Vis* 61:55–79.
20. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32:1627–1645.
21. Hinton G, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18:1527–1554.