Discovering the visual features and representations used by the brain to recognize objects is a central problem in the study of vision. Recently, neural network models of visual object recognition, including biological and deep network models, have shown remarkable progress and have begun to rival human performance in some challenging tasks. These models are trained on image examples and learn to extract features and representations and to use them for categorization. It remains unclear, however, whether the representations and learning processes discovered by current models are similar to those used by the human visual system. Here we show, by introducing and using minimal recognizable images, that the human visual system uses features and processes that are not used by current models and that are critical for recognition. We found by psychophysical studies that at the level of minimal recognizable images a minute change in the image can have a drastic effect on recognition, thus identifying features that are critical for the task. Simulations then showed that current models cannot explain this sensitivity to precise feature configurations and, more generally, do not learn to recognize minimal images at a human level. The role of the features shown here is revealed uniquely at the minimal level, where the contribution of each feature is essential. A full understanding of the learning and use of such features will extend our understanding of visual recognition and its cortical mechanisms and will enhance the capacity of computational models to learn from visual experience and to deal with recognition and detailed image interpretation.

**Methods**

A MIRC is defined as an image patch that can be reliably recognized by human observers and which is minimal in that further reduction in either size or resolution makes the patch unrecognizable (below criterion) \((Methods)\). To discover MIRCs, we conducted a large-scale psychophysical experiment for classification. We started from 10 greyscale images, each showing an object from a different class (Fig. S3), and tested a large hierarchy of image interpretations. Each patch in this hierarchy has five descendants, obtained by either reduction in either size or resolution. Effective recognition of each of the 10 original images was covered by multiple MIRCs \((15.1 \pm 7.6 \text{ per image, excluding highly overlapping MIRCs})\) \((Methods)\) at different positions and sizes (Fig. 3 and Figs. S1 and S2). The resolution (measured in image samples) was small \((14.92 \pm 5.2 \text{ samples})\) \((Methods)\), with some correlation \((0.46)\) between resolution and size \((Methods)\). Each human subject viewed a single patch from each image with unlimited viewing time and was not tested again. Testing was conducted online using the Amazon Mechanical Turk (MTurk) \((Methods)\) with about 14,000 subjects viewing 3,553 different patches combined with controls for consistency and presentation size \((Methods)\). The size of the patches was measured in image samples, i.e., the number of samples required to represent the image without redundancy \((Methods)\). For presentation to subjects, all patches were scaled to \(100 \times 100\) pixels by standard interpolation; this scaling increases the size of the presented image smoothly without adding or losing information.

**Results**

**Discovered MIRCs.** Each of the 10 original images was covered by multiple MIRCs \((15.1 \pm 7.6 \text{ per image, excluding highly overlapping MIRCs})\) \((Methods)\) at different positions and sizes (Fig. 3 and Figs. S1 and S2). The resolution (measured in image samples) was small \((14.92 \pm 5.2 \text{ samples})\) \((Methods)\), with some correlation \((0.46)\) between resolution and size \((Methods)\). Each human subject viewed a single patch from each image with unlimited viewing time and was not tested again. Testing was conducted online using the Amazon Mechanical Turk (MTurk) \((Methods)\) with about 14,000 subjects viewing 3,553 different patches combined with controls for consistency and presentation size \((Methods)\). The size of the patches was measured in image samples, i.e., the number of samples required to represent the image without redundancy \((Methods)\). For presentation to subjects, all patches were scaled to \(100 \times 100\) pixels by standard interpolation; this scaling increases the size of the presented image smoothly without adding or losing information.

The transition in recognition rate from a MIRC image to a nonrecognizable descendant (termed a “sub-MIRC”) is typically sharp: A surprisingly small change at the MIRC level can make it unrecognizable \((Methods)\). The drop in recognition rate was quantified by measuring a recognition gradient, defined as the maximal difference in recognition rate between the MIRC and a recognizable MIRC.

**Significance**

Discovering the visual features and representations used by the brain to recognize objects is a central problem in the study of vision. Recent successes in computational models of visual recognition naturally raise the question: Do computer systems and the human brain use similar or different computations? We show by combining a novel method (minimal images) and simulations that the human recognition system uses features and learning processes, which are critical for recognition, but are not used by current models. The study uses a “phase transition” phenomenon in minimal images, in which minor changes to the image have a drastic effect on its recognition. The results show fundamental limitations of current approaches and suggest directions to produce more realistic and better-performing models.
its five descendants. The average gradient was $0.57 \pm 0.11$, indicating that much of the drop from full to no recognition occurs for a small change at the MIRC level (the MIRC itself or one level above, where the gradient also was found to be high). The examples in Fig. 4 illustrate how small changes at the MIRC level can have a dramatic effect on recognition rates. These changes disrupt visual features to which the recognition system is sensitive (6–9); these features are present in the MIRCs but not in the sub-MIRCs. Crucially, the role of these features is revealed uniquely at the MIRC level, because information is more redundant in the full-object image, and a similar loss of features will have a small effect. By comparing recognition rates of models at the MIRC and sub-MIRC levels, we were able to test computationally whether current models of human and computer vision extract and use similar visual features and to test the ability of recognition models to recognize minimal images at a human level. The models in our testing included HMAX (10), a high-performing biological model of the primate ventral stream, along with four state-of-the-art computer vision models: (i) the Deformable Part Model (DPM) (11); (ii) support vector machines (SVM) applied to histograms of gradients (HOG) representations (12); (iii) extended Bag-of-Words (BOW) (13, 14); and (iv) deep convolutional networks (Methods) (15). All are among the top-performing schemes in standard evaluations (16).

Training Models on Full-Object Images. We first tested the models after training with full-object images. Each of the classification schemes was trained by a set of class and nonclass images to produce a classifier that then could be applied to novel test images. For each of the 10 objects in the original images we used 60 class images and an average of 727,000 nonclass images (Methods). Results did not change by increasing the number of training class images to 472 (Methods and SI Methods). The class examples showed full-object images similar in shape and viewing direction to the stimuli in the psychophysical test (Fig. S5).

After training, all classifiers showed good classification results when applied to novel full-object images, as is consistent with the reported results for these classifiers [average precision (AP) = 0.84 ± 0.19 across classes]. The trained classifiers then were tested on MIRC and sub-MIRC images from the human testing, with the image patch shown in its original location and size and surrounded by an average gray image. The first objective was to test whether the sharp transition shown in human recognition between images at the MIRC level and their descendant sub-MIRCs is reproduced by any of the models (the accuracy of MIRC detection is discussed separately below). An average of 10 MIRC level patches per class and 16 of their similar sub-MIRCs were selected for testing, together with 246,000 non-class patches. These MIRCs represent about 62% of the total number of MIRCs and were selected to have human recognition rates above 65% for MIRCs and below 20% for sub-MIRCs (Methods). To test the recognition gap, we set the acceptance threshold of the classifier to match the average human recognition rate for the class (e.g., 81% for the MIRC-level patches from the original image of an eye) (Methods and Fig. S6) and then compared the percentage of MIRCs vs. sub-MIRCs that exceeded the classifier’s acceptance threshold (results were insensitive to threshold setting over the range of recognition rates 0.5–0.9).

We computed the gap between MIRC and sub-MIRC recognition rates for the 10 classes and the different models and compared the gaps in the models’ and human recognition rates. None of the models came close to replicating the large drop shown in human recognition (average gap 0.14 ± 0.24 for models vs. 0.71 ± 0.05 for humans) (Fig. S7A). The difference between the models’ and human gaps was highly significant for all computer-version models ($P < 1.64 \times 10^{-4}$ for all classifiers, $n = 10$ classes, df = 9, average 16 pairs per class, one-tailed paired $t$ test). HMAX (10) showed similar results (gap 0.21 ± 0.23). The gap is small because, for the models, the representations of MIRCs and sub-MIRCs are closely similar, and consequently the recognition scores of MIRCs and sub-MIRCs are not well separated.

It should be noted that recognition rates by themselves do not directly reflect the accuracy of the learned classifier: A classifier can recognize a large fraction of MIRC and sub-MIRC examples by setting a low acceptance threshold, but doing so will result in the erroneous acceptance of nonclass images. In all models, the accuracy of MIRC recognition (AP 0.07 ± 0.10) (Fig. S7B) was low compared with the recognition of full objects (AP 0.84 ± 0.19) and was still lower for sub-MIRCs (0.02 ± 0.05). At these low MIRC recognition rates the system will be hampered by a large number of false detections.

**Fig. 1.** Reduced configurations. (A) Configurations that are reduced in size (Left) or resolution (Right) can often be recognized on their own. (B) The full images (Upper Row) are highly variable. Recognition of the common action can be obtained from local configurations (Lower Row), in which variability is reduced.
A conceivable possibility is that the performance of model networks applied to minimal images could be improved to the human level by increasing the size of the model network or the number of explicitly or implicitly labeled training data. Our tests suggest that although these possibilities cannot be ruled out, they appear unlikely to be sufficient. In terms of network size, doubling the number of levels (see ref. 17 vs. ref. 18) did not improve MIRC recognition performance. Regarding training examples, our testing included two network models (17, 18) that were trained previously on 1.2 million examples from 1,000 categories, including 7 of our 10 classes, but the recognition gap and accuracy of these models applied to MIRC images were similar to those in the other models.

We considered the possibility that the models are trained for a binary decision, class vs. nonclass, whereas humans recognize multiple classes simultaneously, but we found that the gap is similar and somewhat smaller for multiclass recognition (Methods and SI Methods). We also examined responses of intermediate units in the network models and found that results for the best-performing intermediate layers were similar to the results of the network’s standard top-level output (Methods).

**Training Models on Image Patches.** In a further test we simplified the learning task by training the models directly with images at the MIRC level rather than with full-object images. Class examples were taken from the same class images used in full-object learning but using local regions at the true MIRC locations and approximate scale (average 46 examples per class) that had been verified empirically to be recognizable on their own (Methods and SI Methods). After training, the models’ accuracy in recognizing MIRC images was significantly higher than in learning from full-object images but still was low in absolute terms and in comparison with human recognition (AP 0.74 ± 0.21 for training on patches vs. 0.07 ± 0.10 for training on full-object images) (SI Methods, Training Object on Image Patches and SI Methods, Human Binary Classification Test). The gap in recognition between MIRC and sub-MIRC images remained low (0.20 ± 0.15 averaged over pairs and classifiers) and was significantly lower than the human gap for all classifiers ($P < 1.87 \times 10^{-4}$ for all classifiers, $n = 10$ classes, df = 9, one-tailed paired t test) (Methods and SI Methods).

**Detailed Internal Interpretation.** An additional limitation of current modeling compared with human vision is the ability to perform a detailed internal interpretation of MIRC images. Although MIRCs are “atomic” in the sense that their partial images become unrecognizable, our tests showed that humans can consistently recognize multiple components internal to the MIRC (Methods and Fig. 3C). Such internal interpretation is beyond the capacities of current neural network models, and it can contribute to accurate recognition, because a false detection could be rejected if it does not have the expected internal interpretation.

**Discussion**

The results indicate that the human visual system uses features and processes that current models do not. As a result, humans are better at recognizing minimal images, and they exhibit a sharp drop in recognition at the MIRC level, which is not replicated in models. The sharp drop at the MIRC level also suggests that different human observers share similar visual representations, because the transitions occur for the same images, regardless of individual visual experience. An interesting open question is whether the additional features and processes are used in the visual system as a part of the cortical feed-forward process (19) or by a top-down process (20–23), which currently is missing from the purely feed-forward computational models.

We hypothesize based on initial computational modeling that top-down processes are likely to be involved. The reason is that detailed interpretation appears to require features and interrelations that are relatively complex and are class-specific, in the sense that their presence depends on a specific class and location (24). This application of top-down processes naturally divides the recognition process into two main stages: The first leads to the initial activation of class candidates, which is incomplete and with limited accuracy. The activated representations then trigger the application of class-specific interpretation and validation processes, which recover richer and more accurate interpretation of the visible scene.

A further study of the extraction and use of such features by the brain, combining physiological recordings and modeling, will extend our understanding of visual recognition and improve the capacity of computational models to deal with recognition and detailed image interpretation.

---

**Fig. 3.** (A) Distribution of MIRCs’ resolution (measured in image samples), average 14.92 ± 5.2 samples. (B) MIRCs’ coverage. The original images are covered by multiple MIRCs at different positions, sizes, and resolutions. Each colored frame outlines a MIRC (which may be at a reduced resolution). Because each MIRC is recognizable on its own, this coverage provides robustness to occlusion and transformations. (C) Detailed internal interpretation labeled by subjects ($n = 30$) (Methods). Suit image parts: tie, shirt, jacket, chin, neck. Eagle image parts: eye, beak, head, wing, body, sky.
Methods

Data for MIRC Discovery. A set of 10 images was used to discover MIRCs in the psychophysics experiment. These images of objects and object parts (one image from each of 10 classes) were used to generate the stimuli for the human tests (Fig. S3). Each image was of size 50 × 50 image samples (cutoff frequency of 25 cycles per image).

Data for Training and Testing on Full-Object Images. A set of 600 images was used for training models on full-object images. For each of the 10 images in the psychophysical experiment, 60 training class images were obtained (from Google images, Flickr) by selecting similar images as measured by their HOG descriptor. Additional deep-network models tested were a model developed with a single mode. For RCNN we used a pretrained network (15), which uses the last feature layer of the deep network trained on ImageNet (17) as a descriptor. Additional deep-network models tested were a model developed for recognizing small (32 × 32) images (29), and Very Deep Convolutional Network (18), which was adapted for recognizing small images. HMAX (10) used the implementation of Cortical Network Simulator (CNS) (30) with six scales, a buffer size of 640 × 640, and a base size of 384 × 384.

Fig. 4. Recognition gradient. A small change in images at the MIRC level can cause a large drop in the human recognition rate. Shown are examples of MIRCs (A and B) and corresponding sub-MIRCs (A* and B*). The numbers under each image indicate the human recognition rate. The average drop in recognition for these pairs is 0.67.

Model Versions and Parameters. The versions and parameters of the four classification models used were as follows. The HOG (12) model used the implementation of VLFeat version 0.9.17 (www.vlfeat.org), an open and portable library of computer vision algorithms, using cell size 8. For BOW we used the selective search method (26) using the implementation of VLFeat with an encoding of VLAD (vector of locally aggregated descriptors) (14, 27), a dictionary of size 20, a 3 × 3 grid division, and dense SIFT (28) descriptor. DPM (11) used latest version (release 5, www.cs.berkeley.edu/~bvg/dpm) with a single mode. For RCNN we used a pretrained network (15), which uses the last feature layer of the deep network trained on ImageNet (17) as a descriptor. Additional deep-network models tested were a model developed for recognizing small (32 × 32) images (29), and Very Deep Convolutional Network (18), which was adapted for recognizing small images. HMAX (10) used the implementation of Cortical Network Simulator (CNS) (30) with six scales, a buffer size of 640 × 640, and a base size of 384 × 384.

MIRCs Discovery Experiment. This psychophysics experiment identified MIRCs within the original 10 images at different sizes and resolutions (by steps of 20%). At each trial, a single image patch from each of the 10 images, starting with the full-object image, was presented to observers. If a patch was recognizable, five descendants were presented to additional observers; four of the descendents were obtained by cropping (by 20%) at one corner, and one was a reduced resolution of the full patch. For instance, the 50 × 50 original image produced four cropped images of size 40 × 40 samples, together with a 40 × 40 reduced-resolution copy of the original (Fig. 2). For presentation, all patches were resized to 100 × 100 pixels by image interpolation so that the size of the presented image was increased without the addition or loss of information. A search algorithm was used to accelerate the search, based on the following monotonicity assumption: If a patch P is recognizable, then larger patches or P at a higher resolution will also be recognized; similarly, if P is not recognized, then a cropped or reduced resolution version also will be unrecognized.

A recognizable patch was identified as a MIRC (Fig. 2 and Fig. S4) if none of its five descendents reached a recognition criterion of 50%. (The acceptence threshold has only a small effect on the final MIRCs because of the sharp gradient in recognition rate at the MIRC level.) Each subject viewed a single patch from each image and was not tested again. The full procedure required a large number of subjects (a total of 14,008 different subjects; average age 31.5 y; 52% males). Testing was conducted online using the Amazon MTurk platform (3, 4). Each subject viewed a single patch from each of the 10 original images (i.e., class images) and one “catch” image (a highly recognizable image for control purposes, as explained below). Subjects were given the following instructions: “Below are 11 images of objects and object parts. For each image type the name of the object or part in the image. If you do not recognize anything type ‘none.’” Presentation time was not limited, and the subject responded by typing the labels. All experiments and
procedures were approved by the institutional review boards of the Weizmann Institute of Science, Rehovot, Israel. All participants gave informed consent before starting the experiments.

In comparative studies M Turk has been shown to produce reliable repeatable behavior data, and many classic findings in cognitive psychology have been replicated using data collected online (4). The testing was accompanied by the following controls. To verify comprehension (4), each test included a highly recognizable image; responses were rejected if this catch image was not correctly recognized (rejection rate <1%). We tested the consistency of the responses by dividing the responses of 30 subjects for each of 1,419 image patches into two groups of 15 workers per group and compared responses across groups. Correlation was 0.91, and the difference was not significant (n = 1,419, P = 0.29, two-tailed paired t test), showing that the procedure yields consistent recognition rates. A laboratory test under controlled conditions replicated the recognition results obtained in the field. Recognition rates directly from the human test, and the second was by direct multiclass classification. This multiclass task can lead to cases in which classification results temporally. This multiclass task can lead to cases in which classification results were the same as for the full-object image test, repeating in five-folds, each using 37 patches for training and nine for testing. Before the computational testing, we measured in psychophysical testing the recognition rates of all the patches from all class images to compare human and model recognition rates directly on the same image (see examples in Fig. S8). After training, we compared the recognition rates of MIRCs and sub-MIRCs by testing our multitask models and their recognition accuracy, as in the full-object image test.

We also tested intermediate units in a deep convolutional network (18) by selecting a layer (the eighth of 19) in which units had the markings. One side of a contour was colored red, and subjects produced two labels for the two sides of the contour (e.g., ship and sea). In both alternatives the subjects were asked to name the object they saw in the image (without the markings).

ACKNOWLEDGMENTS. We thank Michal Wolf for help with data collection and Guy Ben-Yosef, Leyla Isik, Ellen Hildreth, Elias Issa, Gabriel Kreiman, and Tomaso Poggio for discussions and comments. This work was supported by European Research Council Advanced Grant “Digital Baby” (to S.U.) and in part by the Center for Brains, Minds and Machines, funded by National Science Foundation Science and Technology Centers Award CCF-1232116.


