Title:

Atoms of recognition in human and computer vision

Author affiliations and footnotes:

Shimon Ullman[a,b,1,2], Liav Assif[a,1], Ethan Fetaya[a], Daniel Harari[a,c,1]

[a] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, 234 Herzl Street, Rehovot 7610001, Israel.

[b] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

[c] McGovern Institute for Brain Research, Cambridge, MA 02139

[1] S.U., L.A. and D.H. contributed equally to this work.

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

Author contributions:

S.U., D.H., L.A. and E.F. designed research; D.H. and L.A. performed research; D.H. and L.A. analyzed data; S.U., D.H., and L.A. wrote the paper.

The authors declare no conflict of interest.

Keywords:

Object recognition, minimal images, visual perception, visual representations, computer vision, computational models.

Citation:

S. Ullman, L. Assif, E. Fetaya, D. Harari (2016). “Atoms of recognition in human and computer vision”. *Proceedings of the National Academy of Sciences*, 113(10): 2744-2749. doi: 10.1073/pnas.1513198113. Url: http://www.pnas.org/content/113/10/2744.abstract

Abstract

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, which starts to rival human performance in some challenging tasks. These models are trained on image examples, learn to extract features and representations, and use them for categorization. It remains unclear, however, whether the representations and learning processes discovered by current models are similar to the ones 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 not used by current models, and which are critical for recognition.

We found by psychophysical studies that at the level of minimal recognizable images, a minute change of the image can have a drastic effect on recognition, 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 enhance the capacity of computational models to learn from visual experience and deal with recognition and detailed image interpretation.

Significance Statement

Discovering the visual features and representations used by the brain to recognize objects is a central problem in the study of vision. The successes of recent 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 not used by current models, which are critical for recognition. The study uses a 'phase transition' phenomenon in minimal images, where minor changes to the image make 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.

\body

Introduction

The human visual system makes highly effective use of limited information (1, 2). As shown below (Fig. 1, 3, 4B, S1, S2), it can recognize consistently severely reduced sub-configurations in terms of size or resolution. Effective recognition of reduced configurations is desirable for dealing with image variability: images of a given category are highly variable, making recognition difficult, but this variability is reduced at the level of recognizable but minimal sub-configurations (Fig. 1B). Minimal recognizable configurations (termed 'MIRCs') are useful for effective recognition, but as shown below, they are also computationally challenging because each MIRC is non-redundant and therefore requires the effective use of all available information. We use them here as sensitive tools to identify fundamental limitations of existing models of visual recognition and directions for essential extensions.

A minimal recognizable configuration is defined as an image patch that can be reliably recognized by human observers, and which is minimal in the sense that further reduction by 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. S4), and tested a large hierarchy of patches at different positions and decreasing size and resolution. Each patch in this hierarchy has 5 descendants, obtained by either cropping or reduced resolution (Fig. 2). If an image patch was recognizable, we continued to test its descendants by additional observers. A recognizable patch in this hierarchy is identified as a MIRC if none of its 5 descendants reach a recognition criterion (50% recognition, results are insensitive to criterion, Methods, Fig. S3). 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 (3, 4) 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, which is the number of samples required to represent the image without redundancy (twice the image frequency cutoff (5)). For presentation to subjects, all patches were scaled to 100×100 pixels by standard interpolation; this increases the size of the presented image smoothly without adding or losing information.

Results

Discovered minimal recognizable configurations. Each of the 10 original images was covered by multiple MIRCs (15.1±7.6 per image, excluding highly overlapping MIRCs, Methods) at different positions and sizes (Fig. 4B, S1, S2). The resolution (measured in image samples) was small (14.92 ± 5.2 samples, Fig. 4A), with some correlation (0.46) between resolution and size (fraction of the object covered). Since each MIRC is recognizable on its own, this coverage provides robustness to occlusion and distortions at the object level, as some MIRCs may be occluded and the overall object may distort and still be recognized by a subset of recognizable MIRCs.

The transition in recognition rate from a MIRC image to a non-recognizable descendant (termed 'sub-MIRC') is typically sharp: a surprisingly small change at the MIRC level can make it unrecognizable. The drop in recognition rate was quantified by measuring a 'recognition gradient', defined as the maximal difference in recognition rate between the MIRC and its 5 descendants; average gradient was 0.57±0.11. This indicates 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 was also found to be high). Examples (Fig. 3) illustrate how small changes at the MIRC level can have a dramatic effect on recognition rates. These changes disrupt visual features that the recognition system is sensitive to (6–9), which are present in the MIRCs but not the sub-MIRCs. Crucially, the role of these features is revealed uniquely at the MIRC level, since in the full-object image, information is more redundant and a similar loss of features will have a small effect. This allowed us to test computationally whether current models of human and computer vision extract and use similar visual features, along with their ability to recognize minimal images at a human level, by comparing recognition rates of models at the MIRC and sub-MIRC levels. The models in our testing included HMAX (10), a high-performing biological model of

the primate ventral stream, along with 4 state-of-the-art computer vision models ('CV' below), Deformable Part Model (11), support vector machines (SVM) applied to histograms of gradients (HOG) representations (12), extended Bag-of-Words (13, 14) and deep convolutional networks (15) (Methods), all 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 non-class images, to produce a classifier that can then 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 non-class images (Methods). Results did not change by increasing the number of training class images to 472 (Methods, SI Methods). The class examples showed full-object images similar in shape and viewing direction to the stimuli in the psychophysical test (Fig. S5).

Following training, all classifiers showed good classification results when applied to novel full-object images, consistent with reported results for these classifiers (average precision (hence: AP) = 0.84±0.19 across classes). The trained classifiers were then tested on MIRC and sub-MIRC images from the human testing, showing the image patch in its original location and size surrounded by an average grey 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 (accuracy of MIRC detection is discussed separately below). An average of 10 MIRC level patches and 16 of their similar sub-MIRCs were selected for testing per class, together with 246,000 non-class patches. These represent about 62% of the total number of MIRCs, selected to have human recognition rate above 65%, and for sub-MIRCs, below 20% (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, 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 thresholds 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 models and human gaps. None of the models came close to replicating the large drop shown in human recognition (average gap for models 0.14±0.24, for humans 0.71±0.05*, Fig S7A,). The difference between the models and human gaps were highly

significant for all CV models ($p < 1.64\times10^{-4}$ for all classifiers, n=10 classes, df=9, average 16 pairs/class, 1-tailed paired t-test). HMAX (10) showed similar results (gap 0.21±0.23). The reason for the small gap is that 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 reflect directly 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 this will also result in the erroneous acceptance of non-class 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 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.

A conceivable possibility is that the performance of model networks applied to minimal images could be improved to human level by increasing the size of the model network or the number of explicitly or implicitly labeled training data. Our tests suggest that while these possibilities cannot be ruled out, they appear unlikely to be sufficient. In terms of network size, doubling the number of levels ((17) vs. (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 recognition gap and accuracy of these models applied to MIRC images were similar to the other models.

We considered the possibility that the models are trained for a binary decision, class vs. non-class, while humans recognize multiple classes simultaneously, but found that the gap is similar and somewhat smaller for multi-class recognition (Methods, 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 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/class), verified empirically to

be recognizable on their own (Methods, Fig. S8). Following training, the accuracy of the models in recognizing MIRC images was significantly higher than learning from full object images, but still low in absolute terms, and in comparison with human recognition (SI Methods sections: Training object on image patches, Human binary classification test ; AP 0.74±0.2, training on patches vs. 0.07±0.10, training on full-object images). The gap in recognition between MIRC and sub-MIRC images remained low (0.20±0.15 averaged over pairs and classifiers), and significantly lower than the human gap for all classifiers ($p<1.87\times10^{-4}$ for all classifiers, n=10 classes, df=9, 1-tailed paired t-test, Methods, 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 (Fig. 4C, Methods). Such internal interpretation is beyond the capacities of current neural network models, and it can contribute to accurate recognition, since a false detection could be rejected if it does not have the expected internal interpretation.

Discussion

The results indicate that the human visuals system uses features and processes, which 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, since the transitions occur for the same images, regardless of individual visual experience. An interesting open question is whether the additional features and processes are employed in the visual system as a part of the cortical feed-forward process (19), or by a top-down process (20–23), which is currently 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 inter-relations, which are relatively complex and are class-specific, in the sense that their presence depends on a specific class and location (24). This 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.

Methods

Data for MIRC discovery. A set of 10 images were used to discover MIRCs (minimal recognizable configurations) in the psychophysics experiment. These images of objects and object parts, from ten object classes (one image from each class), were used to generate the stimuli for the human tests (Fig. S4). 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 ten images in the psychophysical experiment, 60 training class images were obtained (from Google images, Flickr) by selecting similar images (measured by their HOG (12), Histogram of Gradients, representations; examples in Fig. S5). The images were of full objects (e.g. car side-view rather than the door only). These provided positive class examples, on which the classifiers were trained, using 30-50 images for training, the rest for testing (different train/test splits yielded similar results). We also tested the effect of increasing the number of positive examples to 472 (split into 342 train, 130 test) on 3 classes (horse, bicycle, airplane), for which large data sets are available in PASCAL (16) and ImageNet (Russakovsky O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575*, 2014). For the convolutional neural network (CNN) multi-class model used (15), the number of training images was 1.2 million from 1000 categories, including 7 of the 10 classes used in our experiment.

To introduce some size variations, two sizes differing by 20% were used for each image. The size of the full-object images was scaled such that the part used in the human experiment (e.g. car-door) was 50×50 image samples (with 20% variation). For use in the different classifiers, the images were interpolated to match the format used by the specific implementations (e.g. 227×227 for RCNN (15)). The negative images were taken from the PASCAL VOC 2011

(http://host.robots.ox.ac.uk/pascal/VOC/voc2011/index.html), an average of 727,440 non-class image regions per class extracted from 2,260 images in training and 245,970 image regions extracted from a different set of 2,260 images for testing. The number of non-class images is larger than class images in training and testing, since this is also common under natural viewing conditions of class and non-class images.

Data for training and testing on image patches. These were image patches taken from the same 600 images used in full-object image training, but using local regions at the true location and size of MIRCs and sub-MIRCs (called the 'siblings' data set, Fig. S8). Patches were scaled to a common size for each of the classifiers. An average of 46 image patches from each class (23 MIRC and 23 sub-MIRC siblings) were used as positive class examples, together with a pool of 1,734,000 random non-class patches of similar sizes taken from 2,260 non-class images. Negative non-class images during testing were 225,000 random patches from another set of 2,260 images.

Models versions and parameters. Versions and parameters of the four classification models used: The HOG (12) model used the implementation of VLFeat, an open and portable library of computer vision algorithms (version 0.9.17, http://www.vlfeat.org/), cell size 8. For BOW (Bag of visual Words), we used the selective search method (25) using the implementation of VLFeat, with an encoding of VLAD (14, 26) (vector of locally aggregated descriptors), a dictionary of size 20, 3×3 grid division and dense SIFT (27) descriptor. DPM (11) used latest version (release 5, http://www.cs.berkeley.edu/~rbg/latent/) with a single mode. For RCNN we used a pre-trained 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 CIFAR, which is a model developed for recognizing small (32×32) images (Krizhevsky A., Learning Multiple Layers of Features from Tiny Images, *University of Toronto Technical Report*, 2009), and Very Deep Convolutional Network (18), adapted for recognizing small images. HMAX (10) used the implementation of CNS (Mutch J., Knoblich U., Poggio T., CNS: a GPU-based framework for simulating cortically-organized networks. *MIT-CSAIL-TR-2010-013/CBCL-286*, 2010), with 6 scales, buffer size of 640×640 and base size of 384×384.

MIRCs discovery experiment. This psychophysics experiment identified minimal recognizable configurations 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, was presented to observers (starting with the full-object image). If a patch was recognizable, 5 descendants were presented to additional observers: 4 descendants were obtained by cropping (by 20%) at one corner, and one by a reduced resolution of the full patch. For instance, the 50×50 original image produced 4 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 re-scaled to 100×100 pixels by image interpolation; this increases the size of the presented image without adding or losing 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 will also be unrecognized.

A recognizable patch was identified as a MIRC (Fig. 2, S3) if none of its 5 descendants reach recognition criterion of 50%. (The acceptance threshold has only a small effect on the final MIRCs, due to 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; 52% males). Testing was conducted on-line using the Amazon Mechanical Turk (3, 4) (MTurk) platform. Each subject viewed a single patch from each of the 10 original images (called 'class images') and one 'catch' image (a highly recognizable image for control purposes as explained below). Instructions to subjects were: 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 Weizmann Institute of Science, Rehovot, Israel. All participants gave informed consent before starting the experiments.

MTurk has been shown in comparative studies 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 (< 1%). We tested consistency of the responses by dividing responses of 30

subjects for each of 1419 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=1419 p=0.29, 2-tailed paired t-test), showing that the procedure yields consistent recognition rates. A laboratory test under controlled conditions replicated recognition results obtained in the on-line study: recognition rates for 20 MIRCs/sub-MIRCs in the on-line and laboratory studies had correlation of 0.84, and all MIRC sub-MIRC pairs, which were statistically different in the on-line study were also statistically different in the laboratory study. Since viewing size cannot be accurately controlled in the on-line trials, we verified in a laboratory experiment that recognition rates do not change significantly over 1-4 degrees of visual angle.

Subjects were excluded from the analysis if they failed to label all 10 class image patches or failed to label correctly the catch image (2.2%). The average number of valid responses was 23.7 per patch tested. A response was scored as '1' if it gave the correct object name and '0' otherwise. Some answers required decisions regarding the use of related terms, e.g. whether ‘bee’ instead of ‘fly’ would be accepted. Decision was based on the WordNet hierarchy (28); we allowed sister terms that have the same direct parent (hypernym) or two levels up. For instance, a 'cow' was an accepted label for 'horse', but 'dog' or 'bear' were not. Part-names were accepted if they labeled correctly the visible object in the partial image (e.g. ‘wheel’ in bicycle, ‘tie’ for suit image, ‘jet engine’ for airplane part); descriptions that did not name specific objects (such as ‘cloth’, ‘an animal part’ ‘wire’) were not accepted.

Training models on full-object images. Training was done for each of the classifiers using the training data, except for the multi-class CNN classifier (15), which was pre-trained on 1000 object categories based on ImageNet (26). Classifiers were then tested on novel full images using standard procedures, followed by testing on MIRC and sub-MIRC test images.

Detection of MIRCs and sub-MIRCs: An average of 10 MIRC level patches and 16 of their sub-MIRCs were selected for testing per class. These represent about 62% of the total number of MIRCs, selected by their recognition gap and images similarity between MIRCs and sub-MIRCs. (Human recognition rate above 65% for MIRC level patches, and below 20% for their sub-MIRCs, similarity measured by overlap of image contours; the same MIRC could have several sub-MIRCs). The tested patches were placed in their original size and location on a grey

background; for example, an eye MIRC with a size of 20×20 samples (obtained in the human experiment) was placed on gray background image, at the original eye location.

Computing the recognition gap: To obtain the classification results of a model, the model's classification score is compared against an acceptance threshold (29), and scores above threshold are considered detections. After training a model classifier, we set its acceptance threshold to produce the same recognition rate of MIRC patches as the human recognition rate for the same class. For example, for the eye class, the average human recognition rate of MIRCs was 0.81; the model threshold was set so that the model's recognition rate of MIRCs was 0.8. We then found the recognition rate of the sub-MIRCs using this threshold. The difference between the recognition rates of MIRCs and sub-MIRCs is the classifier's recognition gap. (Fig S5). We found in an additional test that the results were insensitive to the setting of the models' threshold, by testing the gap while varying the threshold to produce recognition rates in the range 0.5 - 0.9. For the computational models, the scores of sub-MIRCs were intermixed with the scores of MIRCs, limiting the recognition gap between the two, compared with human vision.

Multi-class estimation: The computational models are trained for a binary decision, class vs. non-class, while humans recognize multiple classes simultaneously. This can lead to cases where classification results of the correct class may be overridden by a competing class. The multi-class effect was evaluated in two ways. The first was by simulations, using statistics from the human test, and the second by direct multi-class classification, using the CNN multi-class classifier (15). The mean rate of giving a wrong-class response (rather than producing the 'none' label) in the human experiments ranged from 37% for the lowest recognition rates to 4% at highest recognition rates. The effect of multi-class decision on the binary classifier was simulated by allowing each tested MIRC or sub-MIRC to be overridden by a different class than the tested category, with a probability that varies linearly between 4% for the highest scoring results and 37% for the lowest scoring results in each class. The gap between MIRC and sub-MIRC recognition is computed as before, but with the additional misclassifications produced by the simulated multi-class effect. Average recognition gap (between MIRCs and sub-MIRCS) was 0.11±0.16 for multiclass vs. 0.14±0.24 for binary classification. The multi-class effect was expected to be small since the scores in the models for the MIRCs and sub-MIRCs were highly intermixed. Multi-class was also tested directly using the CNN model that was trained previously on 1,000

categories (15), including 7 of our 10 classes. Given a test image, the model produces the probability that this image belongs to each of the network categories. The score for each MIRC and sub-MIRC is the probability of the tested class given the test image (e.g. the probability of the 'airplane' class given an airplane MIRC or sub-MIRC). The average gap for the 7 classes was small (0.14±0.35) with no significant difference between MIRCs and sub-MIRCs.

Classification accuracy: was computed by the average precision (AP) of the classifier, the standard evaluation measure for classifiers (16). To compare the AP between the full object, the MIRC and the sub-MIRC detection tasks, we normalize the results to the same number of positive and negative examples across the three test sets.

Intermediate units: In training and testing the HMAX model (10), we examined whether any intermediate units in the network developed specific response to a MIRC image during training. Following full-object image training, we tested the responses of all units at all layers of the network to MIRC patches and non-class patches. We identified the best performing unit at each level of the network (up to the C3 output unit) in terms of its precision at recognizing a particular MIRC type. On this set, AP at the network output was 94±9% for full-object images, and 19±19% for MIRCs. For units with best AP across the network, results were low (still higher than the single C3 output unit: AP = 40±24% S2 level, 44±27% C2 level, 39±21% S3 level).

Training models on image patches. The same classifiers as the ones used in the full-object image experiment were trained and tested (for the RCNN model (15), the test patch was either in its original size or scaled up to the network size of 227×227). In addition, the CIFAR (Krizhevsky A., 2009) and Very Deep deep-network models (18), adapted for recognizing small images, were also tested. Training and testing procedures were the same as for the full-object image test, repeating in 5-folds, each using 35 patches for training and 9 for testing. Prior to the computational testing, we measured in psychophysical testing the recognition rates of all the patches from all class images, to compare directly human and model recognition rates on the same image; (examples in Fig. S8). Following training, we compared the recognition recall rates of MIRCs and sub-MIRCs by the different 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 (8th out of 19) where units' receptive field sizes best approximated the size of MIRC patches. The

activation levels of all units in this layer were used as an input layer to an SVM classifier, replacing the usual top-level layer. The gap and accuracy of MIRC classification based on the intermediate units were not significantly changed compared with the results of the networks' tested top-level output.

Internal interpretation labeling. Subjects (n = 30) were presented with a MIRC image where a red arrow pointed to an image location (e.g. 'beak' of the eagle) and were asked to give a label for the pointed location. Alternatively, one contour was colored red and subjects produced two labels, for the two sides of the contour (e.g. 'ship, sea'). For either of the alternatives, the subjects were also asked to name the object they see in the image (without the markings).

Acknowledgments

The work was supported by European Research Council (ERC) Advanced Grant “Digital Baby” (to S.U.), and in part by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We thank Michal Wolf for help with data collection, Guy Ben-Yosef, Leyla Isik, Ellen Hildreth, Elias Issa, Gabriel Kreiman, and Tomaso Poggio for discussions and comments.

References

1. Torralba A, Fergus R, Freeman WT (2008) 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans Pattern Anal Mach Intell* 30(11):1958–70.

2. Bourdev L, et al. (2009) Poselets : Body Part Detectors Trained Using 3D Human Pose Annotations. *Int Conf Comp Vis*:2–9.

3. Buhrmester M, Kwang T, Gosling SD (2011) Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspect Psychol Sci* 6(1):3–5.

4. Crump MJC, McDonnell J V, Gureckis TM (2013) Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8(3):e57410.

5. Bracewell RN (1999) *The Fourier Transform and Its Applications* (McGraw-Hill, Singapore). 3rd Ed.

6. Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci* 4(8):2051–2062.

7. Fujita I, Tanaka K, Ito M, Cheng K (1992) Columns for visual features of objects in monkey inferotemporal cortex. *Nature* 360:343–346.

8. Gallant JL, Braun J, Van Essen DC (1993) Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science (80-)* 259:100–103.

9. Kourtzi Z, Connor CE (2011) Neural representations for object perception: structure, category, and adaptive coding. *Annu Rev Neurosci* 34:45–67.

10. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2(11):1019–25.

11. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans Pat Anal Mach Intel* 32(9):1627–1645.

12. Dalal N, Triggs B (2005) Histograms of Oriented Gradients for Human Detection. *Proc Comp Vis Pat Rec*:886–893.

13. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. *Work Stat Learn Comp Vis*:1–22.

14. Arandjelovic R, Zisserman A (2013) All about VLAD. *Proc Comp Vis Pat Rec*:1578–1585.

15. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc Comp Vis Pat Rec*:580–587.

16. Everingham M, Gool L, Williams CKI, Winn J, Zisserman A (2009) The Pascal Visual Object Classes (VOC) Challenge. *Int J Comp Vis* 88(2):303–338.

17. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. *Proc NIPS*:1–9.

18. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. *Int Conf Learn Rep*:1–13.

19. Yamins DLK, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24.

20. Bar M, et al. (2006) Top-down facilitation of visual recognition. *PNAS* 103(2):449–54.

21. Zylberberg A, Dehaene S, Roelfsema PR, Sigman M (2011) The human Turing machine: A neural framework for mental programs. *Trends Cog Sci* 15(7):293–300.

22. Gilbert CD, Li W (2013) Top-down influences on visual processing. *Nat Rev Neurosci* 14(5):350–363.

23. Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis* 20(7):1434–1448.

24. Geman S (2006) Invariance and selectivity in the ventral visual pathway. *J Physiol Paris* 100(4):212–224.

25. Uijlings J, Sande K van de, Gevers T, Smeulders AWM (2013) Selective search for object recognition. *Int J Comp Vis* 104(2):154–171.

26. Jegou H, Douze M, Schmid C, Perez P (2010) Aggregating local descriptors into a compact image representation. *Proc Comp Vis Pat Rec*:3304–3311.

27. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comp Vis* 60(2):91–110.

28. Miller GA (1995) WordNet: a lexical database for English. *Comm ACM* 38(11):39–41.

29. Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* (Robert E. Krieger Publishing Co., Huntington, NY).

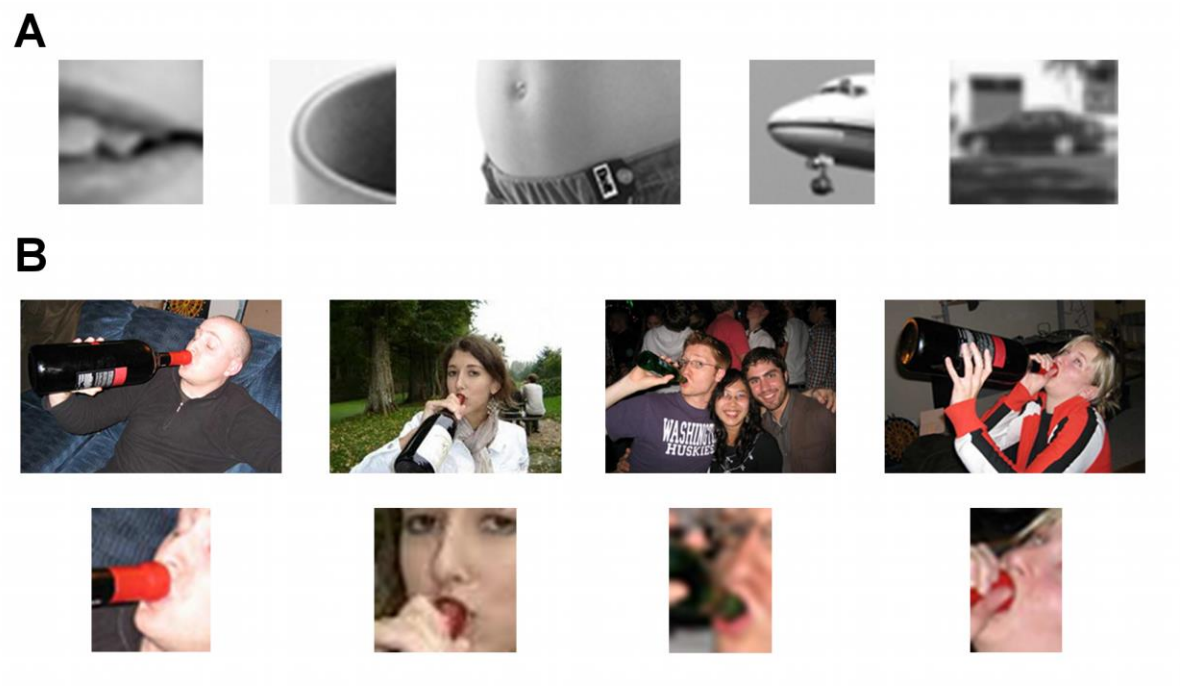

Figure 1: Reduced configurations. (**A)** Reduced configurations in size (e.g. top left) or resolution (top right) can often be recognized on their own. (**B)** The full images (top) are highly variable. Recognition of the common action can be obtained from local configurations (bottom), in which variability is reduced.

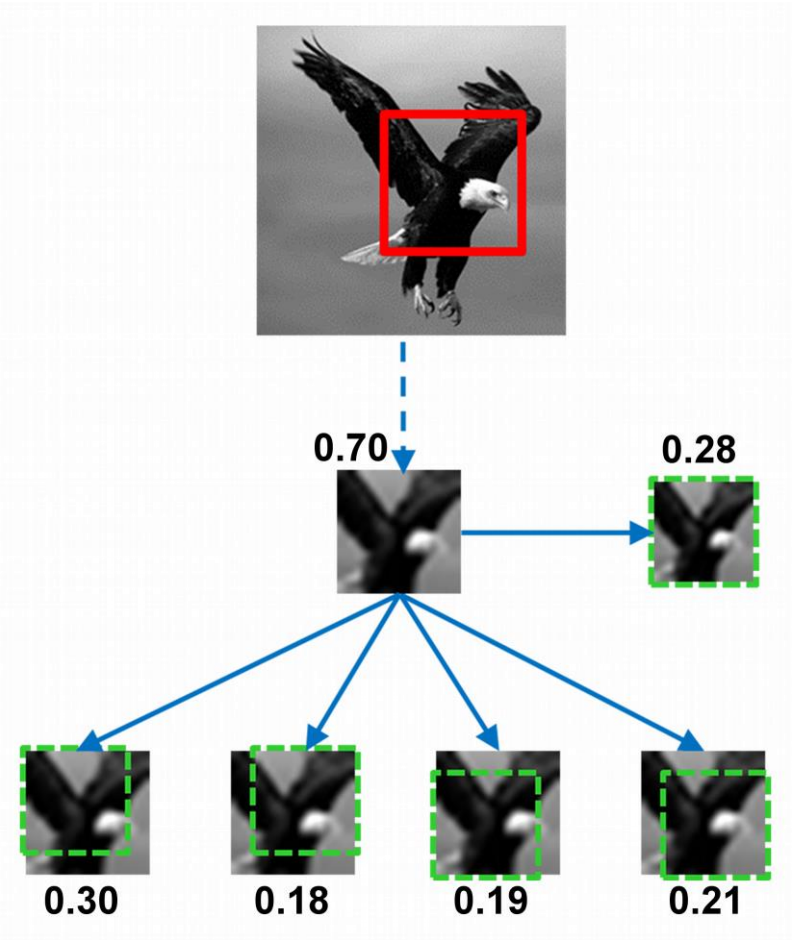

Figure 2: MIRCs discovery. If an image patch was recognized by human subjects, 5 descendants were presented to additional observers: 4 were obtained by 20% image crop (bottom) and one by 20% reduced resolution (right); the process repeated on all descendants until none of the descendants reach recognition criterion (50% , detailed examples in Fig. S3). Numbers next to each image indicate fraction of subjects that correctly recognized the image.

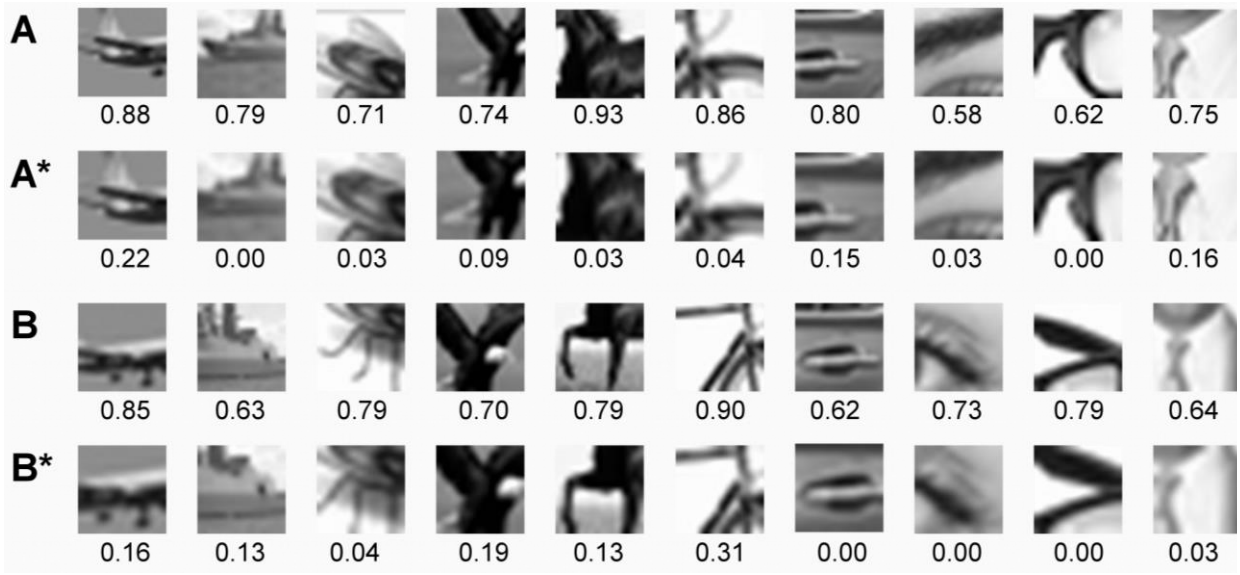

Figure 3: Recognition gradient. A small change in images at the MIRC level can cause a large drop in human recognition rate. Examples of MIRCs **(A, B)** and corresponding sub-MIRCs **(A*, B*)**. Numbers under each image indicate the human recognition rate. The average recognition drop for these pairs is 0.67.

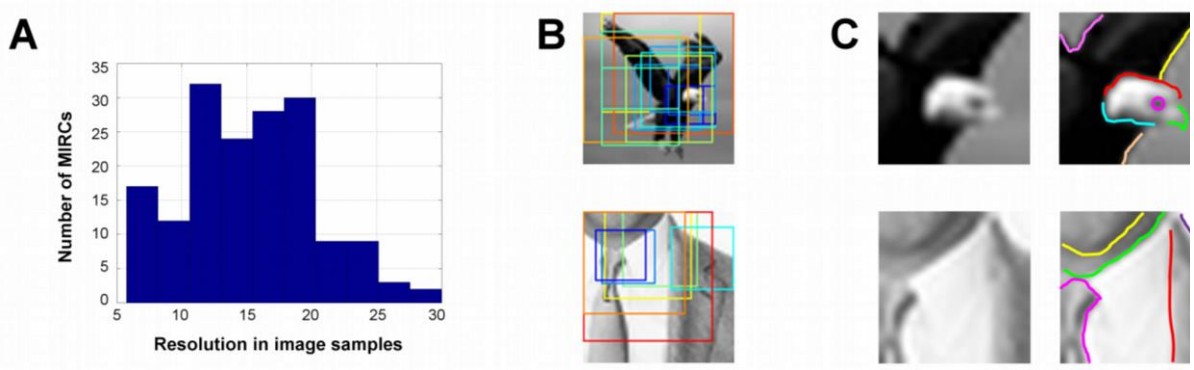

Figure 4: (**A**) **Distribution of MIRCs resolution** (measured in image samples), average 14.92 ± 5.2 samples. (**B**) **MIRCs coverage.** The original images are covered by multiple MIRCs at different positions, sizes and resolutions. Each colored frame outlines a MIRC (which may be at a reduced resolution). Since 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.

Supplementary Methods:

Training models on full-object images. The human average MIRC recall was 0.81 and sub-MIRC recall 0.10. The models' average MIRC and sub-MIRC recall was 0.84 and 0.70 respectively. The HMAX model showed similar results, MIRC and sub-MIRC recall rates 0.84 and 0.63 respectively, with a recognition gap of 0.21±0.23.

The difference between the human and model recognition gaps were highly significant for all the models tested (n=10 classes, df=9, 1-tailed paired t-test: DPM: $p<1.05\times10^{-5}$, BOW: $p<1.64\times10^{-4}$, HOG: $p< 4.2\times10^{-5}$, RCNN: $p< 3.88\times10^{-6}$, HMAX: $p<6.89\times10^{-5}$).

In terms of accuracy, we computed the equal error rate (EER) in the ROC of the computational models, the error (average across models) was 0.02 for full-object images, but high for the MIRCs (0.23) and sub-NIRCs (0.23). Similarly for HMAX, EER was 0.03, 0.33 and 0.39 for the full object, MIRC and sub-MIRC, respectively.

Training with more images: The MIRC vs. sub-MIRC recognition gap remained small compared with human recognition (models gap of 0.01±0.18 vs. human gap of 0.7±0.06, BOW: $p< 0.046$, HOG: $p< 0.002$, RCNN: $p< 0.006$; n = 3 classes, df=2, 1-tailed paired t-test). The CNN multi-class model has been trained on 1.2 million images from 1000 categories (including 7 categories we use). For this model too, the recognition gap was small and recognition accuracy was low: recognition gap of 0.14±0.35; AP was 0.36, 0.01, 0.01 for full-object image, MIRC and sub-MIRC respectively, and EER in the ROC curve was 0.03, 0.31 and 0.35 respectively.

Training models on image patches. In terms of recognition gap, none of the models produces a recognition gap that was comparable to the human gap. The human gap was higher and the differences between each of the models and human results were all highly significant (n=10 classes, df=9, 1-tailed paired t-test: DPM: $p<1.87\times10^{-4}$, BOW: $p<3.75\times10^{-5}$, HOG: $p<1.3\times10^{-6}$, RCNN: $p<1.71\times10^{-7}$, HMAX: $p<4.62\times10^{-8}$). In terms of recognition accuracy, the average precision of MIRC recognition across classifiers was 0.74±0.21 (lower for the HMAX model, 0.38). All additional deep-network models we tested (very deep CNN (18) and CIFAR (Krizhevsky A., 2009)), gave similar results.

Human binary classification test. We noted that models often produced false MIRC detections that appear unacceptable to humans. We therefore compared the distribution of errors made by

humans and the HMAX model in recognizing minimal images. Humans (n=30) were tested in 12 trials, each using 60 image patches, 30 positive class examples, and 30 non-class images. The positive set included MIRC patches from the siblings' dataset above (Fig. S8). These images were similar to one of the discovered MIRCs, depicting the same object part (e.g. horse-torso) at the same image resolution, and were recognizable when tested on human subjects in a free classification task. The 30 negative image patches were automatically selected by the following procedure. A DPM (Deformable Part Model) classifier (11) was trained on separate positive examples together with a large number of randomly selected patches, as described in training models on image patches above. We then used the 30 top-scoring non-class patches as 'hard negatives' for testing.

All 60 image patches were presented on the screen randomly ordered in 5 rows (12 patches per row). Subjects were asked to tag each image patch as a positive or negative example of the object category (e.g. 'ship'). The experiment consisted of 12 trials in total, one trial per each of the 10 object categories, except the eye (2 patches), horse (3 patches), with different object parts, and the car (not tested). Out of the 360 subjects, we discarded responses that failed to label one or more images, leaving 275 complete responses.

For comparing human results with a biological model applied to the same images, the HMAX model (10) was trained on image patches as described above, and applied to the same 60 image patches that were presented to the human subjects in each of the 12 trials.

We tested whether the HMAX model response vector to the 60 images was a likely response given the distribution of human responses, or an outlier. We measured the Euclidean distance between the response vectors of human subjects to the ground truth, and found that the distance of the response vector of the model to the ground truth is unlikely to come from the same distribution. The test was a two-sample, tailed t-test with the null hypothesis that the distance between the HMAX response vector and the ground-truth vector in each class (X), and the distance between human response vectors and the ground-truth response in each class (Y), are independent random samples from normal distributions with equal means, and unequal, unknown variances (Welch's t-test using MathWorks MATLAB *ttest2* function). The null hypothesis was rejected, ($p=9.41\times10^{-5}$, n1=12, n2=275, df=12.19).

Humans were also significantly better at MIRC recognition compared with the model. We compared the classification accuracy of the test images by humans vs. the HMAX model. For humans, we calculated the classification score for each test image as the fraction of positive responses out of the total number of responses for the image. We computed the receiver operation characteristic (ROC) graphs for humans and the HMAX model for each of the 12 classes and used the equal error rates (EER) for the comparison. The average human EER was significantly lower (humans: 0.75% error $\pm 13.6\times10^{-3}$, model: 15.9% error $\pm 8.27\times10^{-2}$, $p=1.30\times10^{-6}$, df=22, 1-tailed paired t-test).

Figure S1: Minimal Recognizable Configurations (MIRCs). Discovered MIRCs for each of the 10 original images (10 object classes) are ordered from large image coverage to small, within each class. Below each MIRC are the recognition rate (left) and size in image samples (right).

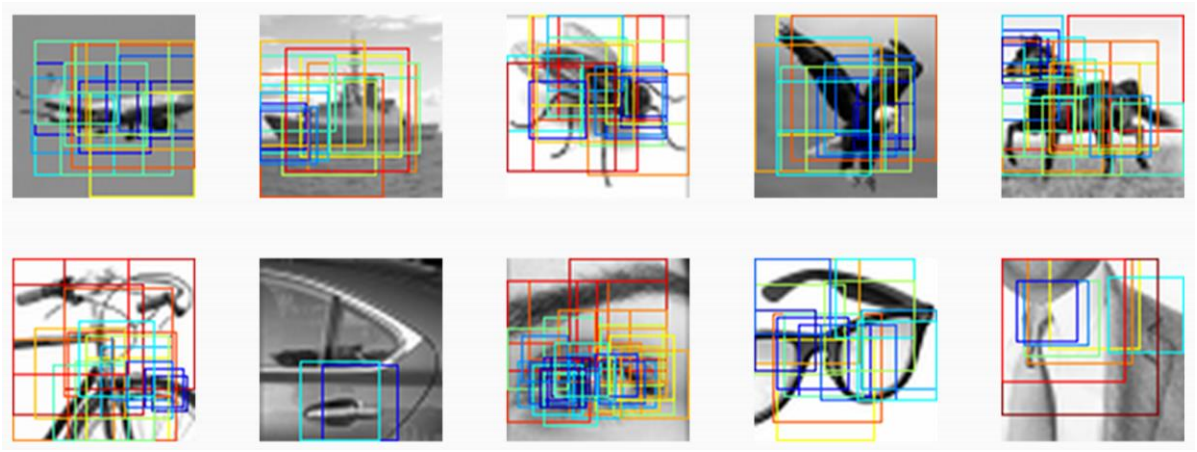

Figure S2: MIRCs coverage. Each colored frame outlines a MIRC (which may be at a reduced resolution). Together, they provide a redundant representation since recognition can be obtained from a single MIRC. Warmer colors of the MIRC frame outline larger coverage.

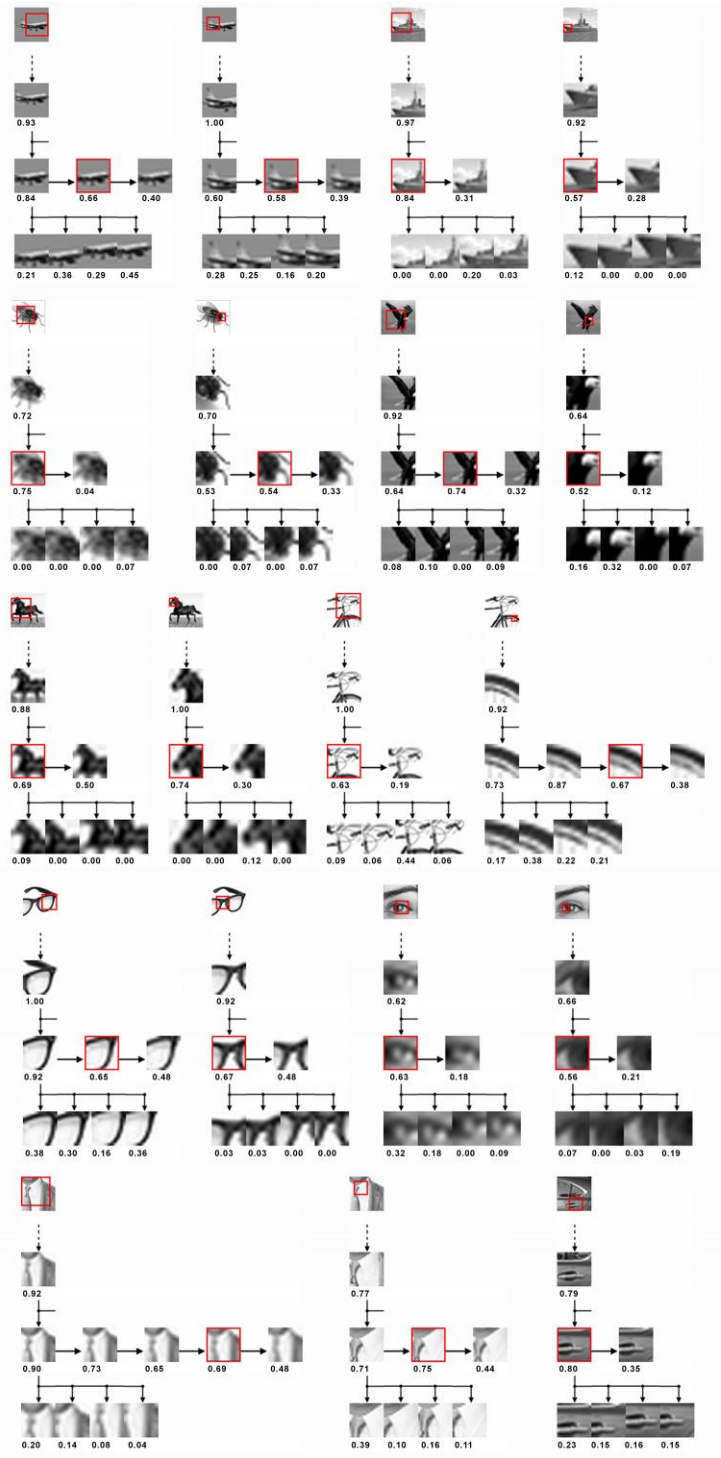

Figure S3: MIRC hierarchical trees. Examples of MIRCs (in red boxes) and their hierarchical trees, including sub-image descendants (sub-MIRCs) and super-image ancestors (super-MIRCs). At the top of each tree is a depiction of the MIRC's position in the original image marked in a red bounding box. The human recognition rate is shown below the image patches.

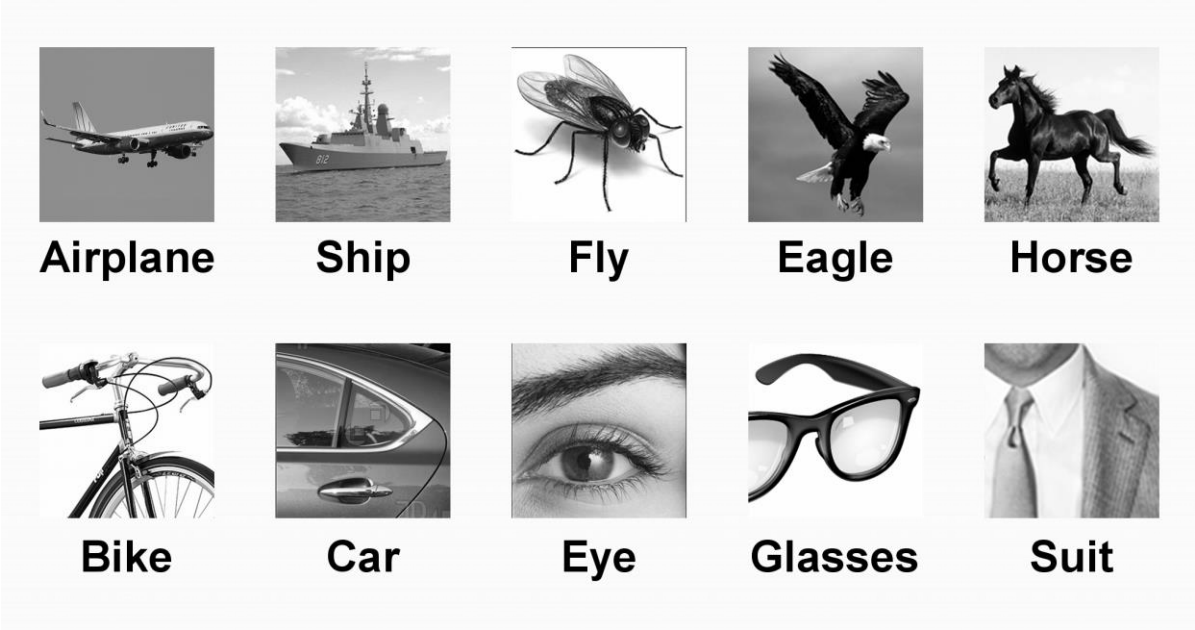

Figure S4: Original images used in the human study. The image stimuli in the human study were extracted from these 10 ‘original’ images (10 object categories)**.** In the experiment, the size of each original image was 50×50 image samples, or a cutoff spatial frequency of 25 cycles per image.

Figure S5: Full-object image siblings. 60 class (‘Airplane’ in this example) images obtained from the web (Google images, Flickr) by selecting similar images (using HOG similarity) to the corresponding ‘original’ image of the same class, that was used in the human psychophysics experiment.

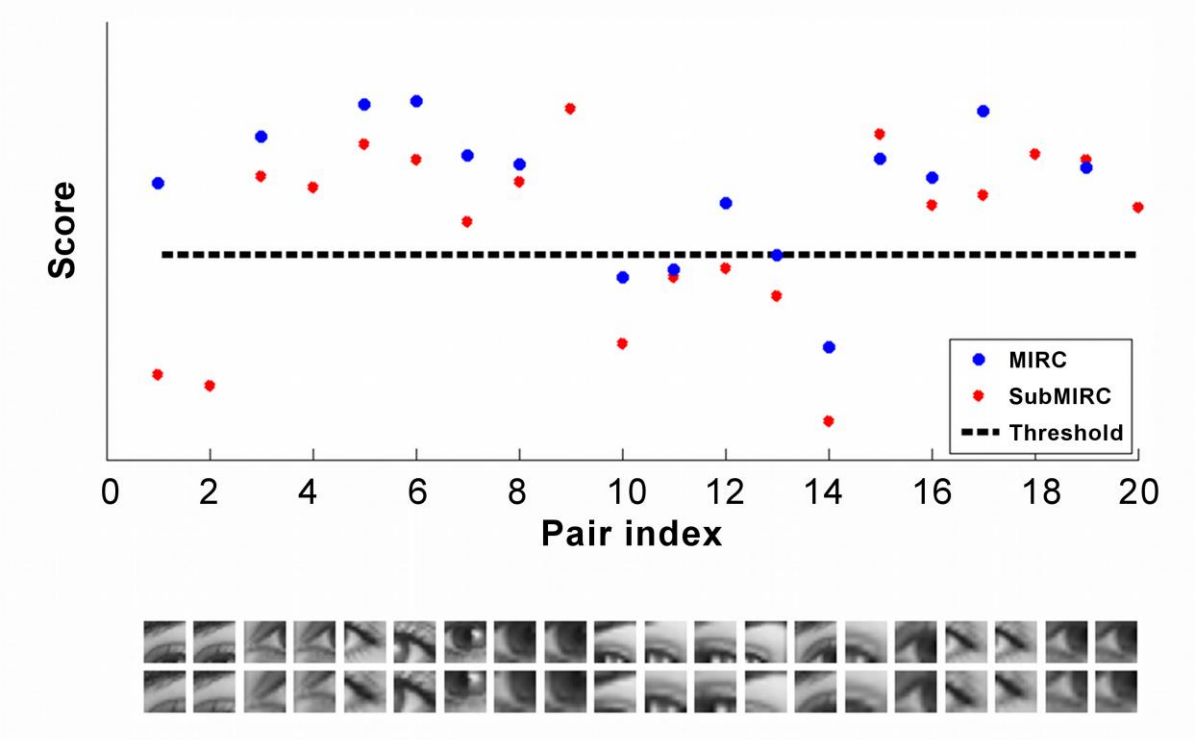

Figure S6: Acceptance thresholds for models. A visualization of the procedure for determining the models acceptance thresholds (in this example the RCNN model (15) applied to the Eye class). The human recognition rate for the MIRCs was 0.81. The threshold is set so that the model recognition rate will match the human recognition rate (12 out of 15 MIRCs exceed the threshold). For this threshold, the model recognition rate for the sub-MIRCs is 0.65. The MIRC/sub-MIRC pairs are shown at the bottom; several pairs have the same MIRC, since a single MIRC has more than one sub-MIRC.

To test the sensitivity of the models recognition gap to the threshold setting, we first set the threshold to produce recognition rate for MIRCs of 0.50 (instead of 0.80). This yields a recognition gap of 0.23. When setting the recognition rate to 0.90, the recognition gap is 0.18. On average across classes and models, the mean recognition gap for this range of threshold setting is 0.13, indicating that the models recognition gap was insensitive to threshold setting.

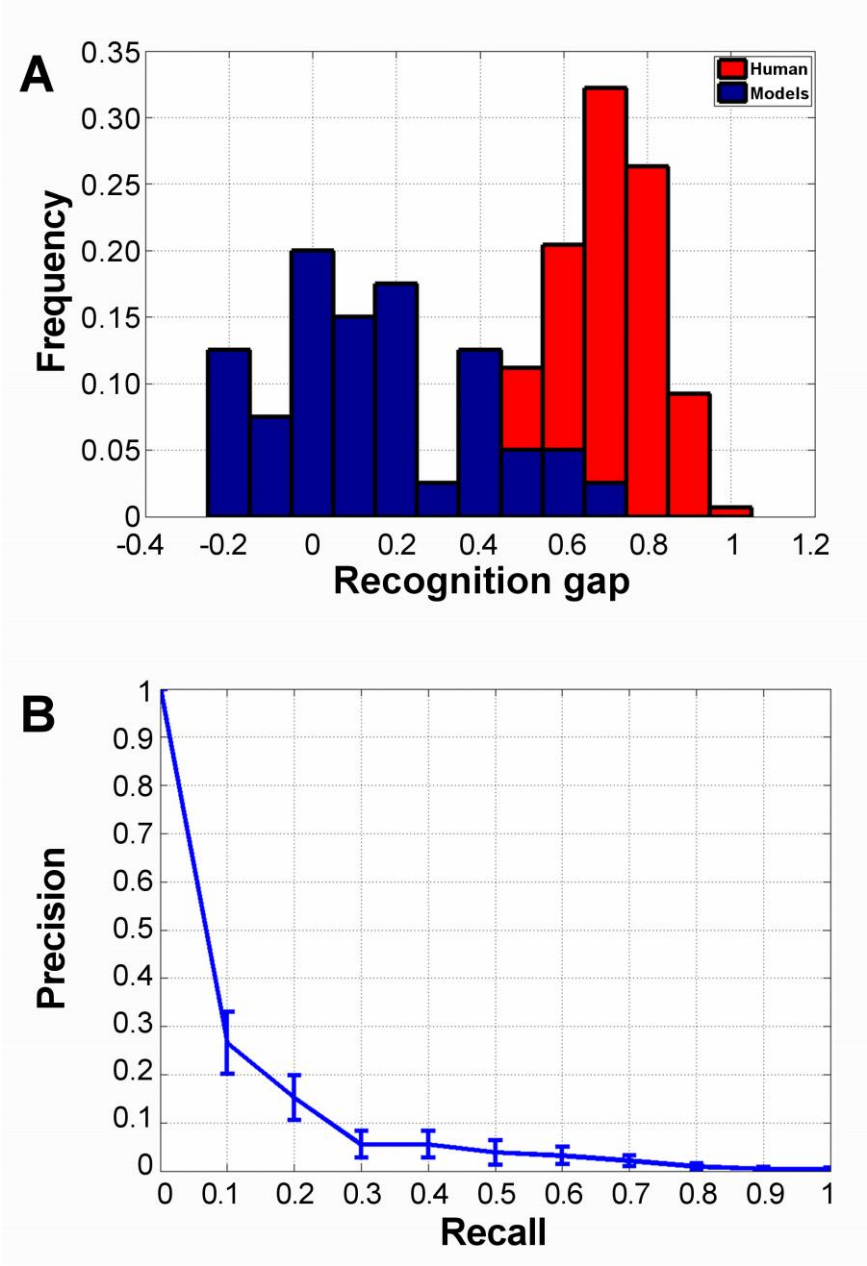

Figure S7: Models recognition gap and performance. (**A**) Distributions of the recognition gap (between MIRCs and their similar but unrecognized sub-MIRCs), by humans and by computational models (average gap over all MIRC and sub-MIRC pairs of the same class). (**B**) Models performance: average precision-recall curve of the computational models, training of full-object images. The error bars show the standard deviation from the average precision for each recall rate.

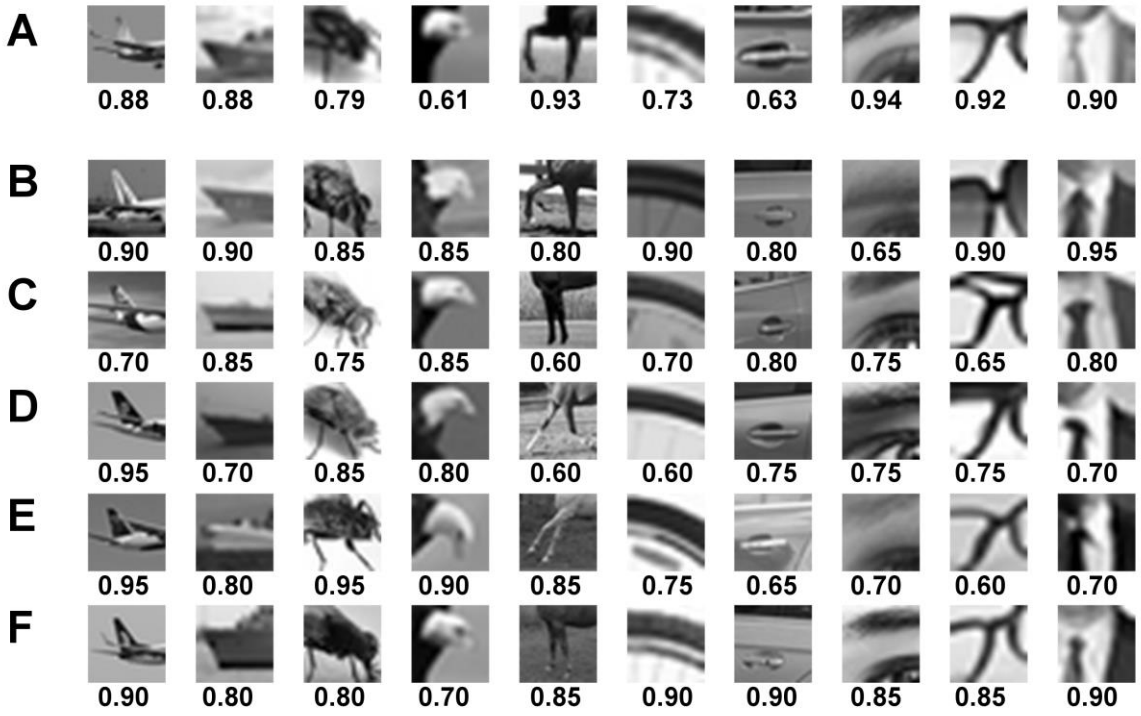

Figure S8: MIRC siblings. (A) Discovered MIRCs, one from each of the 10 original images. **(B-F)** Five examples of extracted image patches from the full-object image siblings (Fig. S5), at a similar position and size to the discovered MIRCs in (A). Below each image is its human recognition rate.