ATOMS OF RECOGNITION IN HUMAN AND COMPUTER VISION

Ullman, Assif, Fetaya, Harari

Presenter: Enrico Mensa

INTRODUCTION

HUMANS AND SUBCONFIGURATIONS

- The human visual system makes highly effective use of limited information.
- It can **recognize consistently subconfigurations** that are severely reduced in size or resolution.
- The recognition of this subconfigurations is also important in order to deal with the **variability** of images pertaining the same class.

HUMANS AND SUBCONFIGURATIONS

A

B

MINIMAL RECOGNIZABLE CONFIGURATIONS

- A Minimal Recognizable Configuration (**MIRC**) is an image patch that:
  - Can be reliably recognized by humans observer.
  - Is minimal in that further reduction in either size or resolution makes the patch unrecognizable (below criterion).
- MIRCs are useful for **effective recognition** but they are also **computationally challenging** because each MIRC is nonredundant and therefore require the effective use of all available information.
- MIRCs can be used as sensitive tools to **identify fundamental limitations of existing models** of visual recognition and directions for essential extensions.

MIRC DISCOVERY

MIRC DISCOVERY

- To discover MIRCs, the authors conducted a large-scale psychophysical experiment for classification.
- The authors started from **10 greyscale images**, each showing an object from a different class, and tested a large hierarchy of patches at different positions and decreasing size and resolution.
- Each patch in this hierarchy **has five descendants**, obtained by either cropping the image or reducing its resolution. If an image patch was recognizable, we continued to test the recognition of its descendants by additional observers.
- A recognizable patch in this hierarchy is identified as a MIRC if **none of its five descendants reaches a recognition criterion**.

10 CLASSES

Airplane

Ship

Fly

Eagle

Horse

Bike

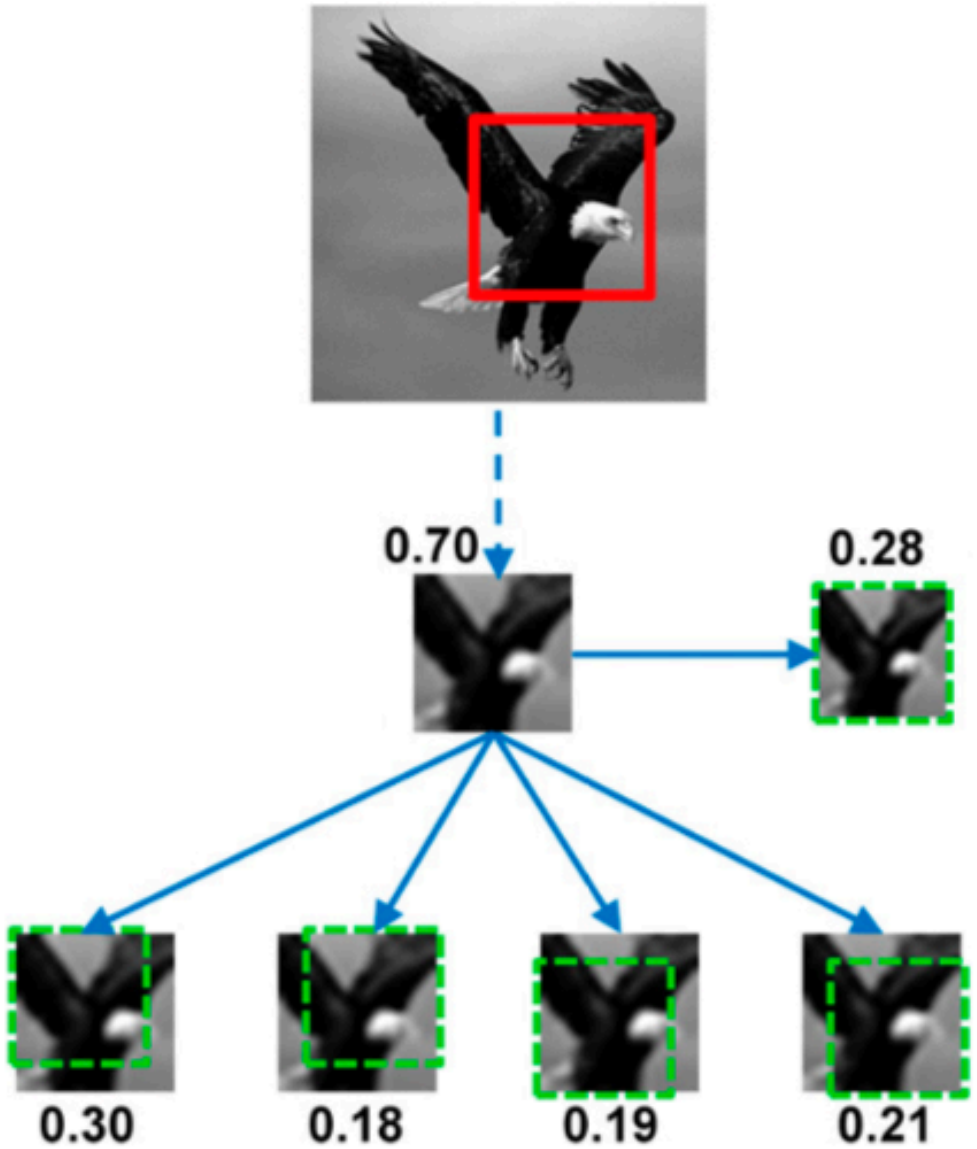
Car

Eye

Glasses

Suit

MIRC DISCOVERY EXAMPLE



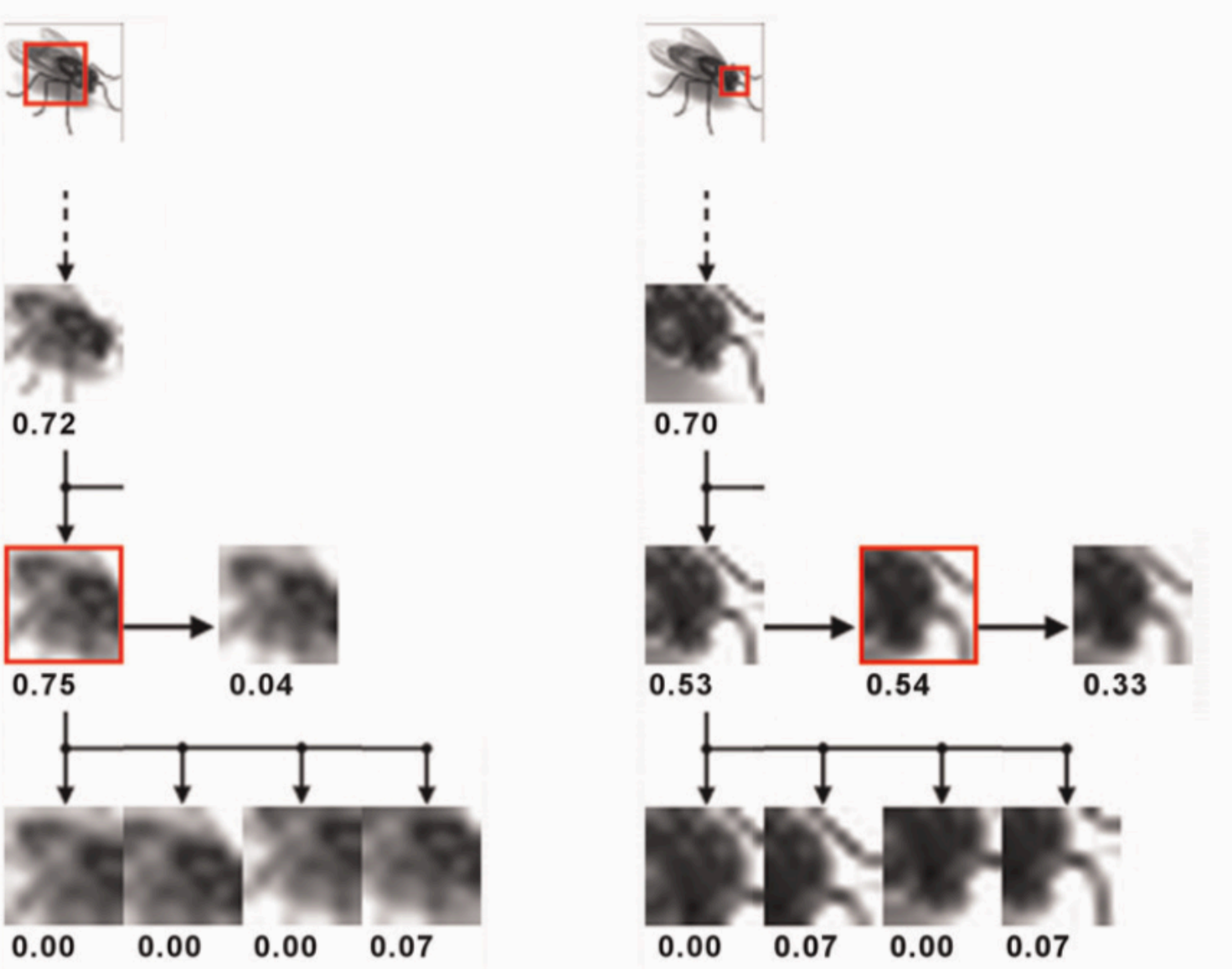
MIRC DISCOVERY EXPERIMENT

- 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) with about **14,000 subjects viewing 3,553 different patches**.
- The size of the patches was measured in in **image samples**.
- 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 descendants were obtained **by cropping** (by 20%) **at one corner**, and one was a **reduced resolution of the full patch**.

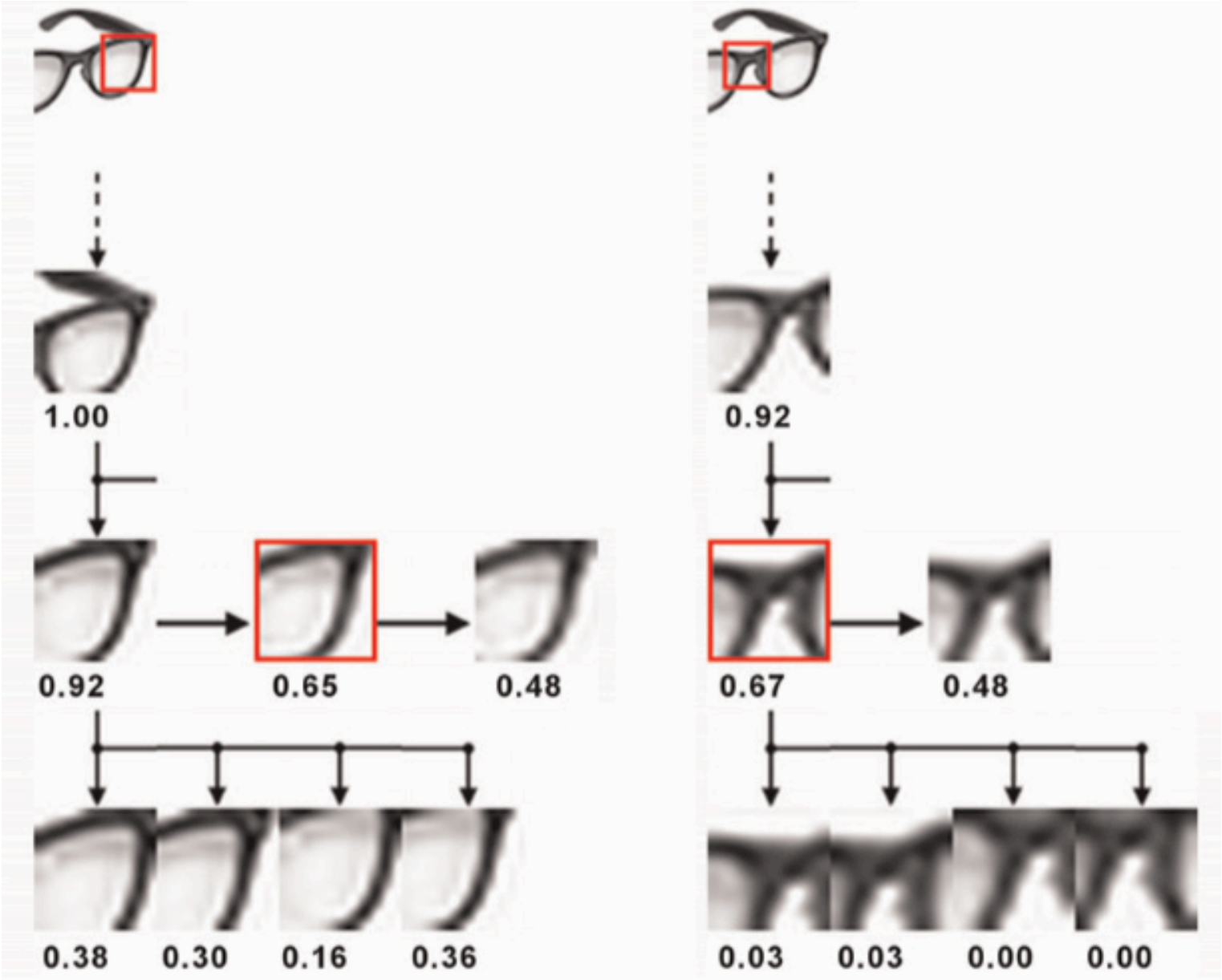
MIRC DISCOVERY EXPERIMENT

- A recognizable patch was **identified as a MIRC** if none of its five descendants reached a recognition criterion of 50%. 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).
- Each subject viewed a single patch from each of the 10 original images and one "catch" image (a highly recognizable image for control purposes).
- 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'."

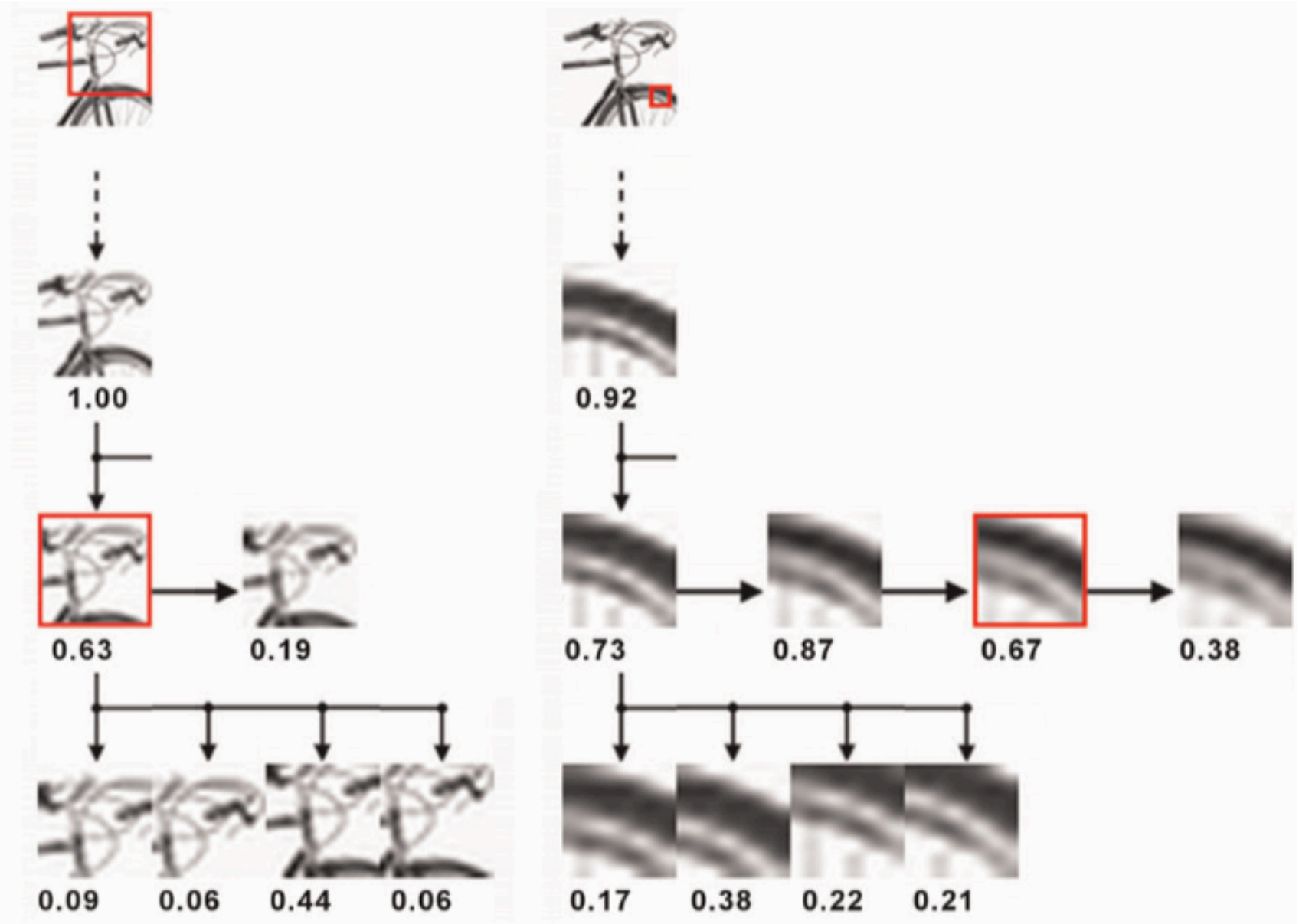
MIRC EXAMPLE (1)



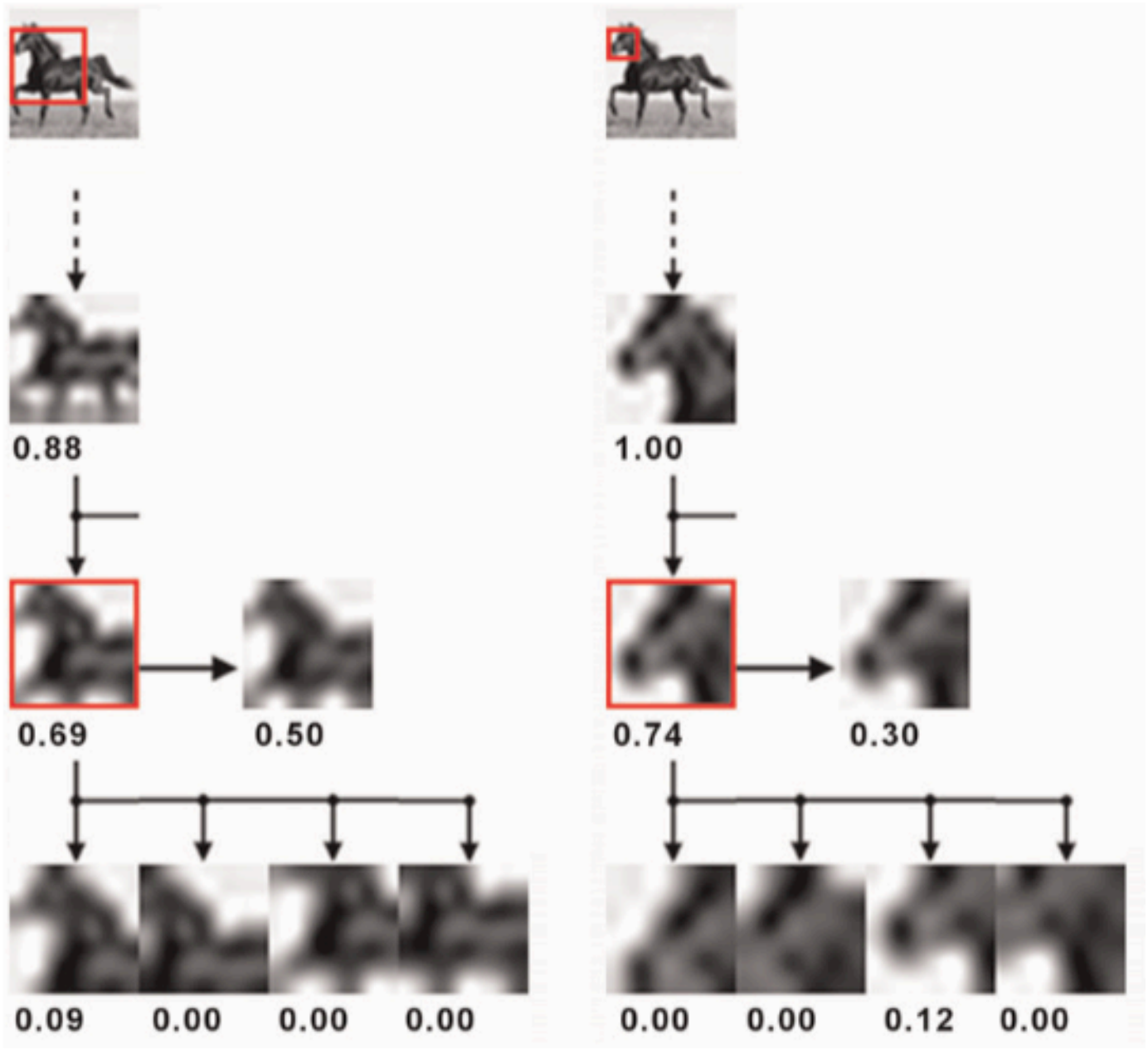
MIRC EXAMPLE (2)



MIRC EXAMPLE (3)

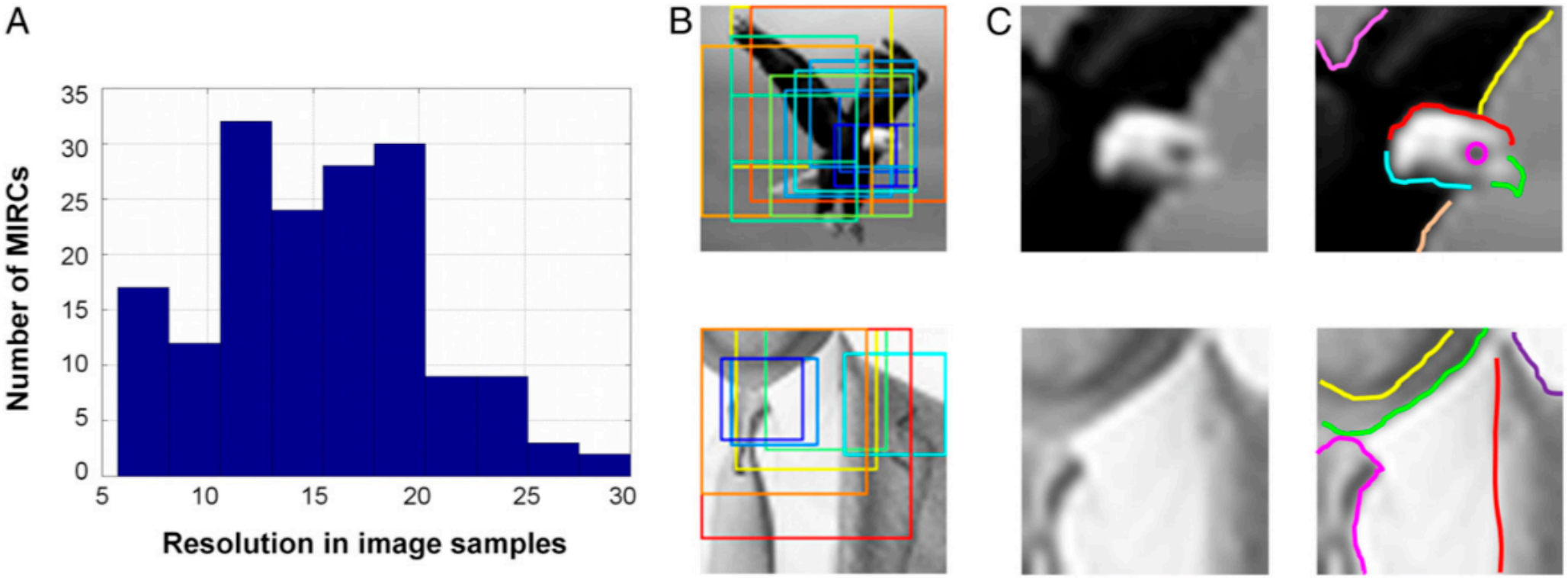


MIRC EXAMPLE (4)



RESULTS

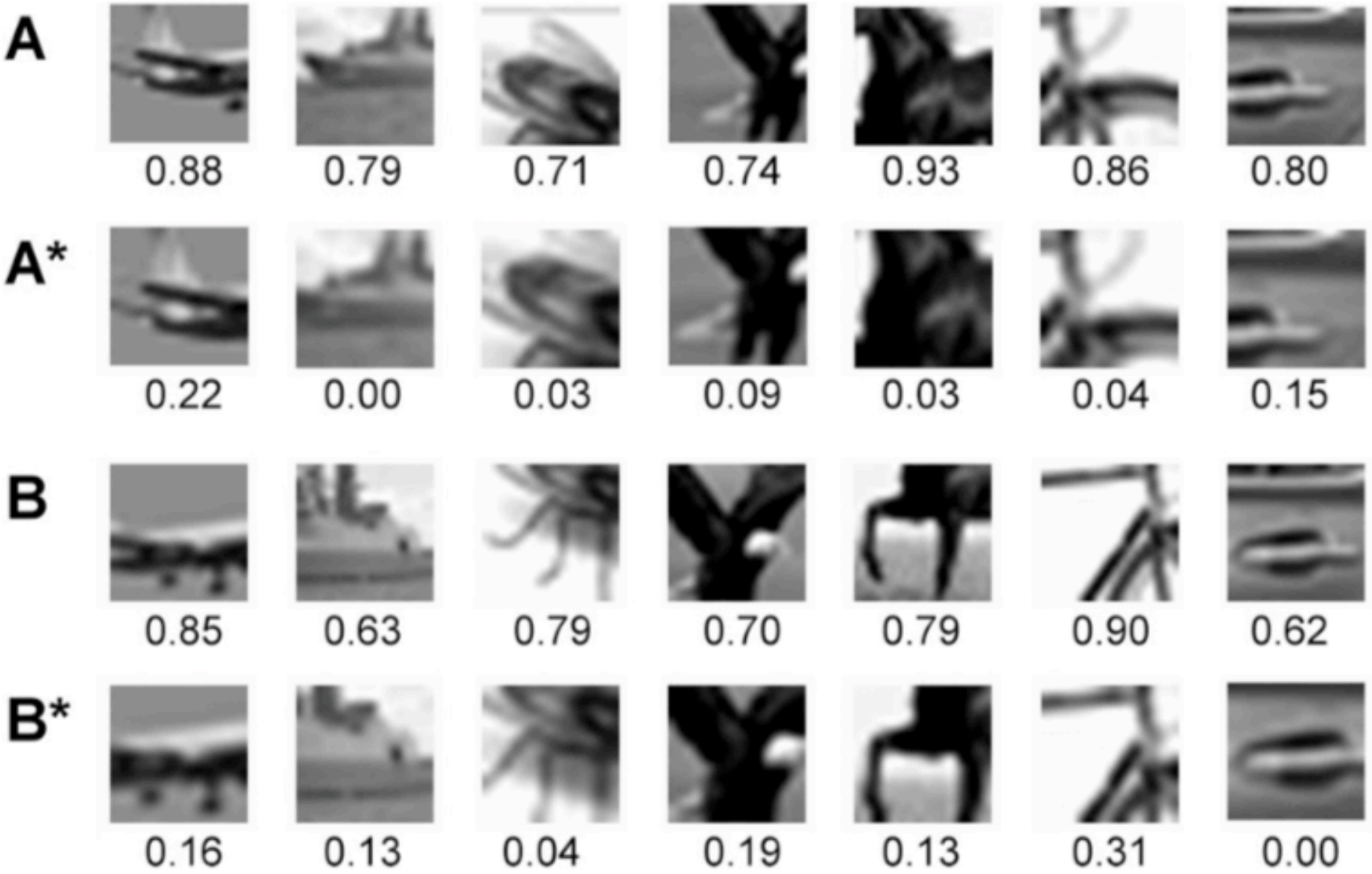
- Each of the 10 original images was covered by multiple MIRCs (15.1 ± 7.6 per image, excluding highly overlapping MIRCs) at different positions and sizes.



RESULTS – RECOGNITION RATE DROP

- The **transition** in recognition rate from a MIRC image to a nonrecognizable descendant 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 five descendants. The **average gradient was 0.57 ± 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).

RESULTS – RECOGNITION RATE DROP



THE RELEVANCE OF THE MIRC LEVEL

The changes adopted in the generation disrupt visual features to which the recognition system is sensitive; **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.

TRAINING MODELS ON FULL-OBJECT IMAGES

DO MODELS AND HUMANS ADOPT THE SAME VISUAL FEATURES?

By **comparing recognition rates of models at the MIRC and sub-MIRC levels**, authors were able to test computationally whether current models of human and computer vision extract and use similar visual features.

CONSIDERED MODELS

- The models considered by the authors are:
  - A high-performing biological model of the primate ventral stream (HMAX);
  - The Deformable Part Model (DPM);
  - Support Vector Machines (SVM) applied to histograms of gradients (HOG) representation;
  - Extended Bag-of-Words (BOW);
  - Deep Convolutional Networks (DNN).

TRAINING SETUP

- The first test was performed after **training with full-object images**.
- Each model was **trained by a set of class and non-class images** to produce a classifier that then could be applied to novel test images.
- For each of the 10 objects in the original images, authors used 60 class images (600 total) and an average of 727,000 non-class images (7,270,000 total).
- The class **examples showed full-object images** similar in shape and viewing direction to the stimuli in the psychophysical test.

EXAMPLE OF FULL-OBJECT IMAGES SIBLINGS

CLASSIFICATION RESULTS ON FULL-OBJECT IMAGES

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.

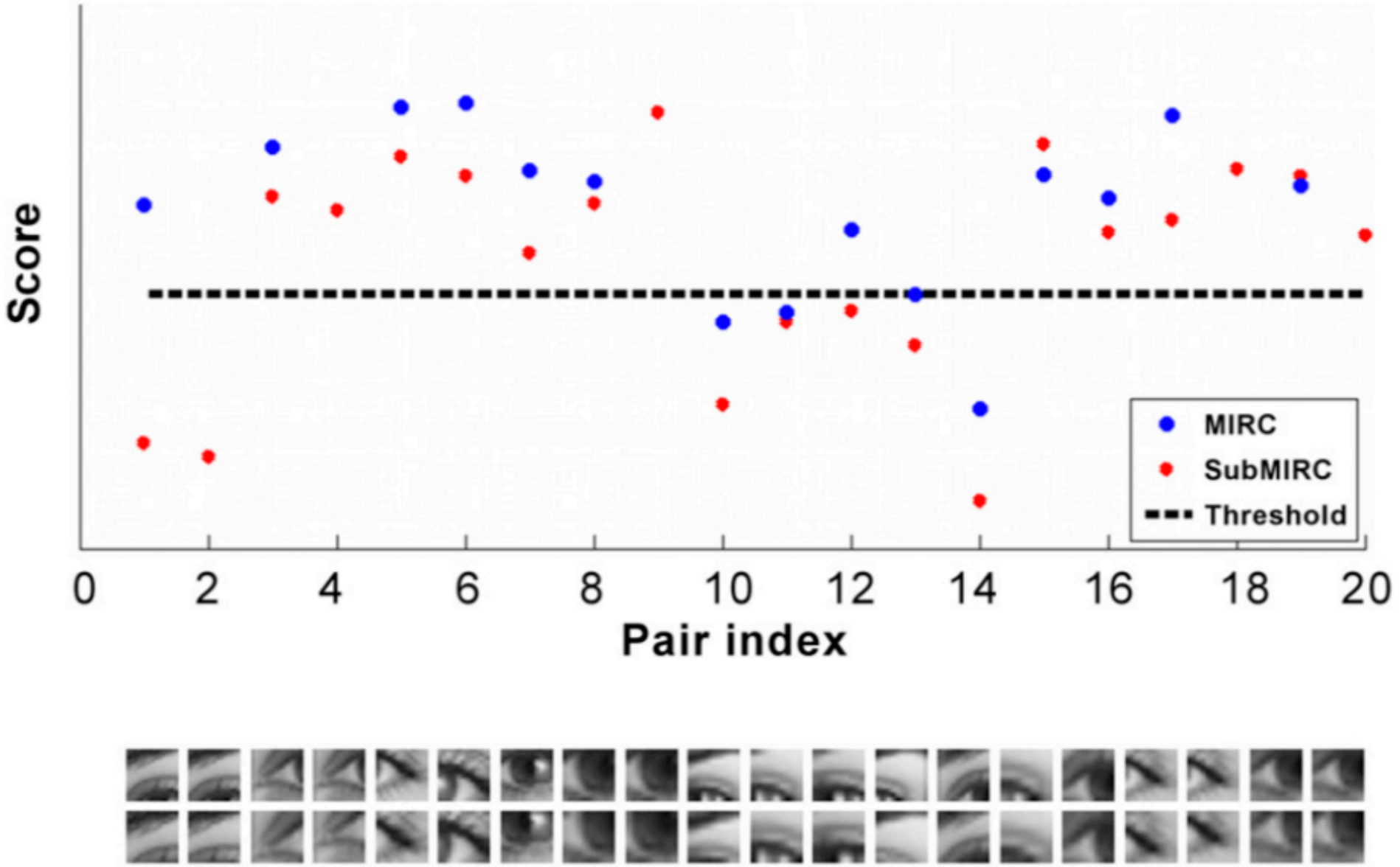
TESTING ON MIRC AND SUB-MIRC IMAGES: DATASET

- The trained classifiers were then **tested on MIRC and sub-MIRC images**, with the image patch shown in its original location and size and surrounded by an average gray image.
- An average of **10 MIRC 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 and were selected to have human recognition rates above 65% for MIRCs and below 20% for sub-MIRCs.

TESTING ON MIRC AND SUB-MIRC IMAGES: RECOGNITION GAP

- To obtain the classification results of a model, the model's classification score was compared against an **acceptance threshold** (AT), and scores above threshold were considered detections.
- The AT was set 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 AT of the model was set so that the model's recognition rate of MIRCs was 0.8.
- The recognition rate of the sub-MIRCs was found using this threshold.
- **The difference between the recognition rates of MIRCs and sub-MIRCs is the classifier's recognition gap**.

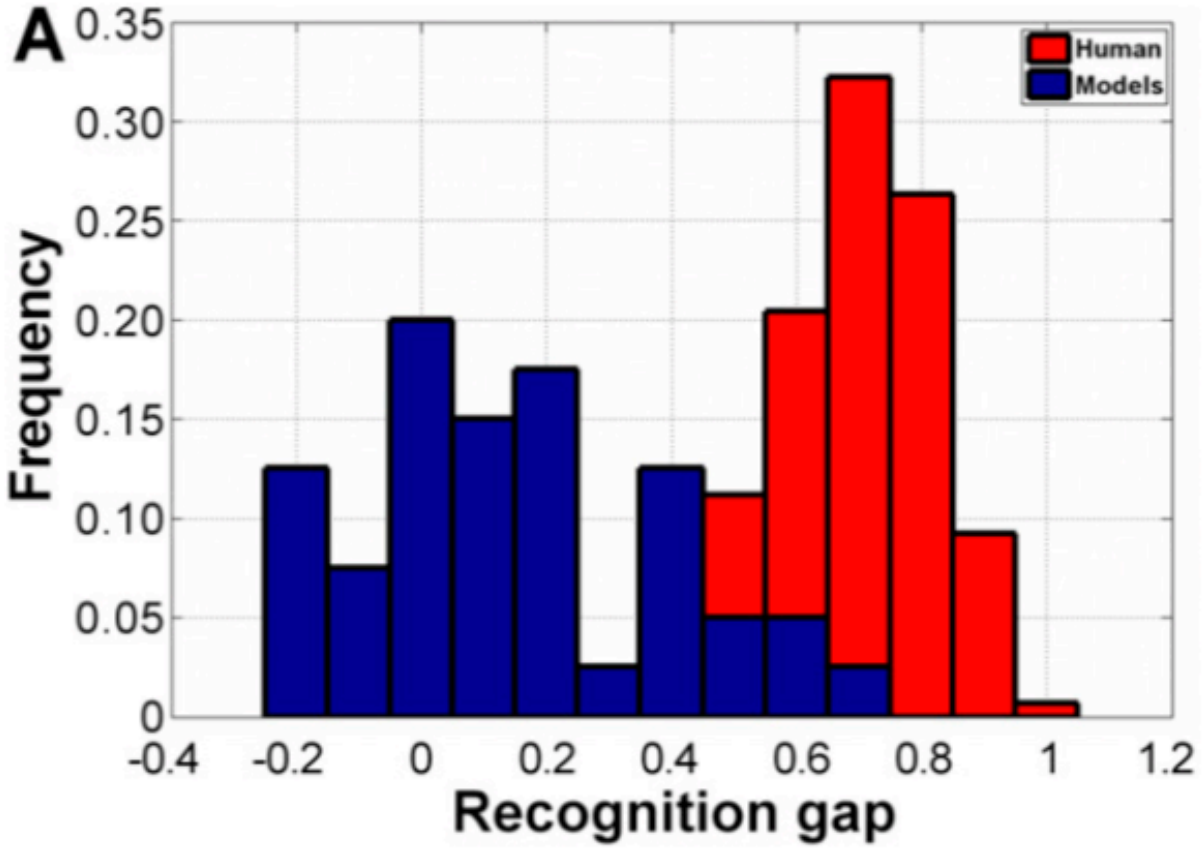
TESTING ON MIRC AND SUB-MIRC IMAGES: RECOGNITION GAP



RESULTS – RECOGNITION GAP

- Authors 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).
- The difference between the models' and human gaps was **highly significant** for all computer-vision 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 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.

RESULTS OF RECOGNITION GAP



ABOUT THE ACCURACY

- 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) 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.

FURTHER EXPERIMENTATIONS

- **Can the accuracy be improved by increasing the size of the model network or the amount of training data?**

  This possibility cannot be ruled out, but the author's further tests suggest that those improvements are likely to be insufficient.

- **Are models trained for binary decision whereas humans recognize multiple classes simultaneously?**

  The authors found that the gap is similar and somewhat smaller for multi-class recognition.

- **What about recognition in intermediate units of NNs?**

  The results are similar to the results of the network's standard top level output.

TRAINING MODELS ON IMAGE PATCHES

TRAINING SETUP & RESULTS

- The learning task is simplified: models are trained **directly** upon **images at MIRC level** instead of full-object images.
- Same classes as the previous experiment, an average of 46 examples per class.
- This time the models' accuracy in recognition was higher, but still low in absolute terms (**AP 0.74 ± 0.21**).
- 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).

CONCLUSIONS

DETAILED INTERNAL REPRESENTATION

- A 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, author's tests showed that **humans can consistently recognize multiple components internal to the MIRC**.
- 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.

CONCLUSIONS (1)

- 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.

CONCLUSIONS (2)

- 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** or by a **top-down process**, which currently is missing from the purely feed-forward computational models.
- 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.

CONCLUSIONS (3)

- 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.

EXTRA: HMAX

NEURAL BASIS OF VISUAL CATEGORIZATION (1)

- Visual perception is a dynamic process, which starts with a **coarse initial analysis** of a scene that gets **continuously refined** to reflect the infinite amount of details present in natural scenes: '*The more you look, the more you see*.'
- A large body of literature suggests that an initial coarse visual analysis relies on the extraction of relatively simple visual features via feature detectors that operate very rapidly and in parallel across the entire visual field.

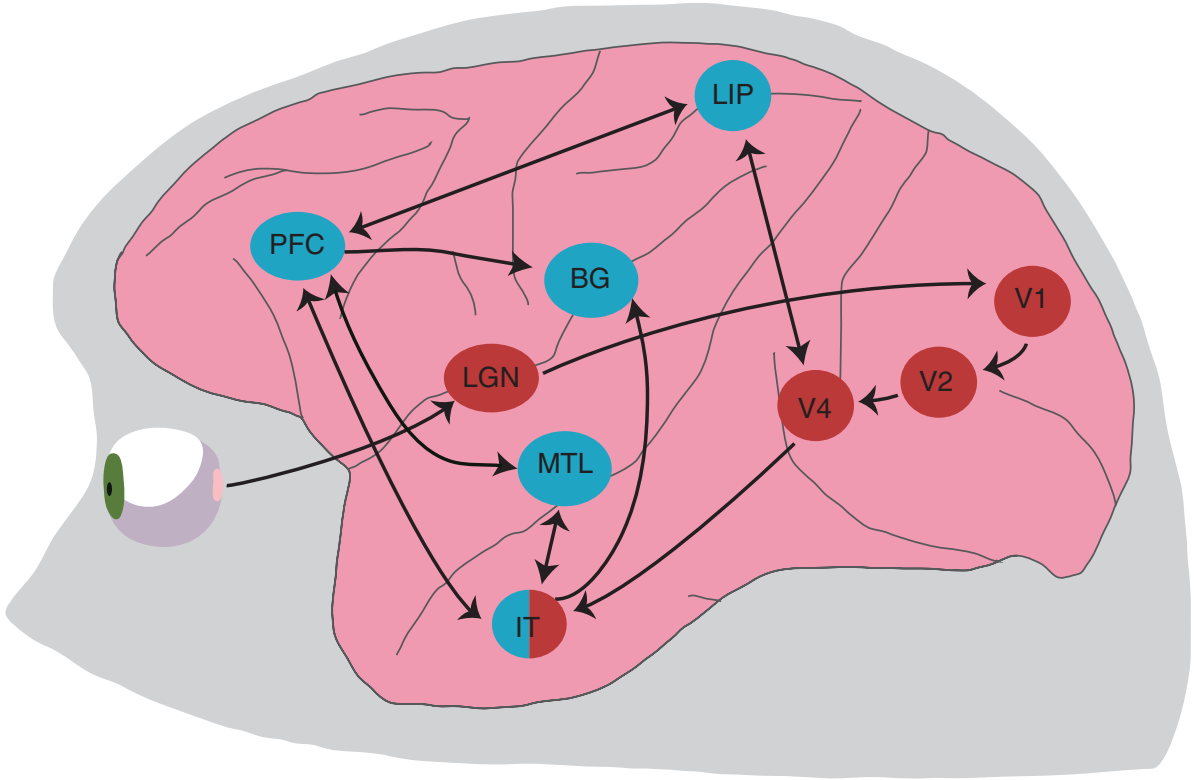
NEURAL BASIS OF VISUAL CATEGORIZATION

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

NEURAL BASIS OF VISUAL CATEGORIZATION

- At each stage of the processing hierarchy, the underlying visual **representation** becomes **increasingly complex** with cells becoming selective to increasingly more stimulus dimensions – from single orientations to image fragments and object views in higher visual areas. At the same time, the underlying visual representation becomes gradually **more tolerant to image transformations** (mainly changes in position and scale).
- Converging evidence suggests that the ventral stream of the visual cortex plays a key role in the encoding of object categories.

NEURAL BASIS OF VISUAL CATEGORIZATION

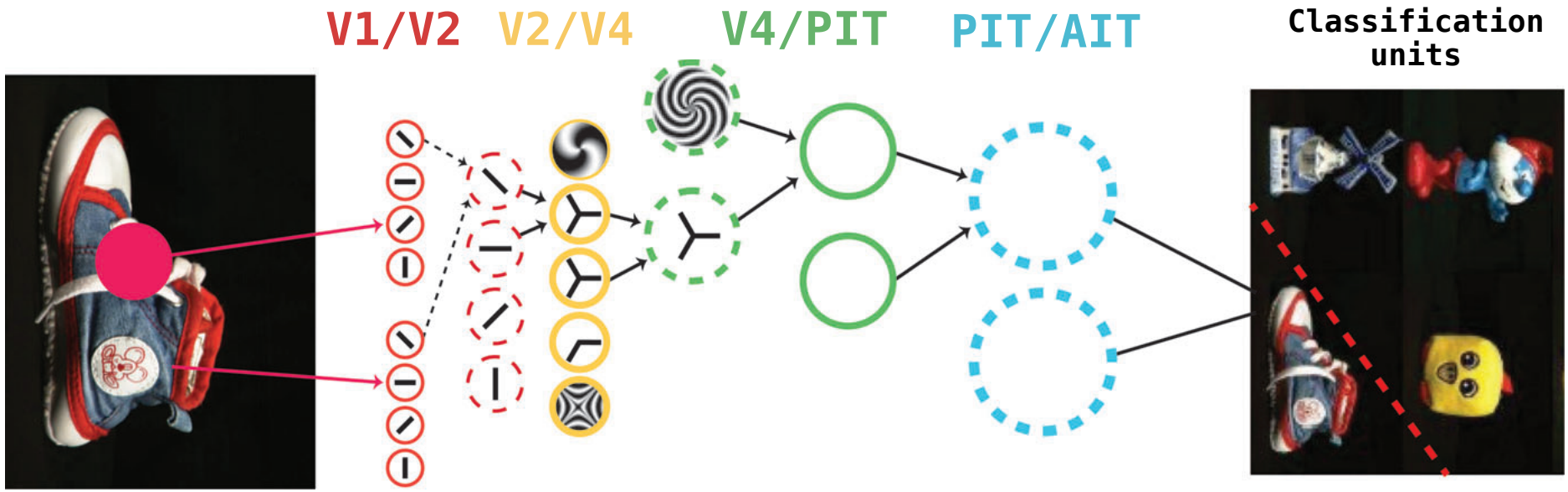


The neural basis of visual categorization. Shown are areas involved in visual categorization. Areas involved in the computation of visual features are shown in red and areas involved in categorization in cyan.

HMAX – BASICS

- The HMAX model is based on the idea of combining intermediate-level features in order to **obtain high complex features**.
- HMAX tries to **emulate** the main information processing stages across the entire ventral stream visual pathway and bridges the gap between multiple levels of understanding.
- This system-level model seems **consistent with physiological data** in nonhuman primates in different cortical areas of the ventral visual pathway, as well as human behavioral data during rapid categorization tasks with natural images.

HMAX



The model relies on two types of computations:

- A max operation (shown in the dashed circles, also called **invariance pooling**) over similar features at different position and scale to gradually build tolerance to position and scale.
- A bell-shaped tuning operation (shown in the plain circles, also called **selectivity pooling**) over multiple features to increase the complexity of the underlying representation.

HMAX VS DEEP LEARNING NETWORKS: TWO DIFFERENCES

- **Unsupervised learning**: HMAX unsupervised learning seems consistent with ITC recordings that have shown that the learning of position and scale invariance, for instance, is driven by the subject's visual experience and is unaffected by reward signals.
- **Parameters and layers:** HMAX parameters (receptive field sizes, invariance and other tuning properties, number of layers, etc) are constrained by available neuroscience data, while deep learning architectures do not try to imitate biology at such a level of detail. For instance, state-of-the-art deep learning architectures incorporate many more layers (more than 20 layers) than HMAX (7 layers).