Acquiring Visual Classifiers from Human Imagination

Carl Vondrick, Hamed Pirsiavash, Aude Oliva, Antonio Torralba
Massachusetts Institute of Technology
{vondrick,hpirsiav,oliva,torralba}@mit.edu

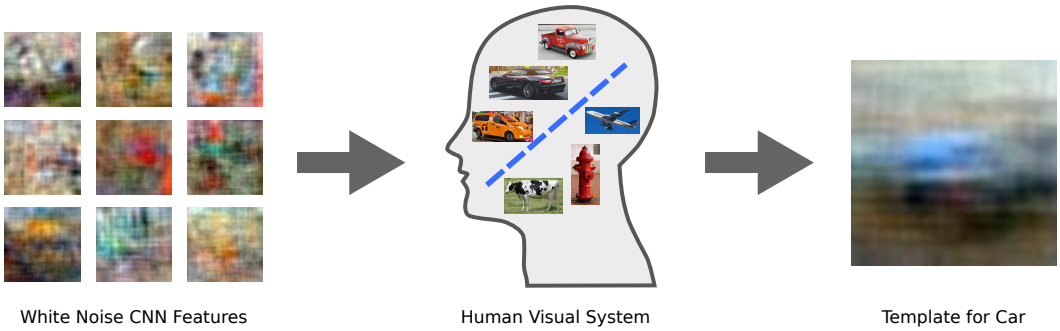

Figure 1: Although all image patches in (a) are just noise, when we show thousands of them to online workers and ask them to find ones that look like cars, a car emerges in the average, shown in (c). This noise-driven method is a well known approach in human psychophysics that extracts the decision function that the human visual system uses for recognition. We are adopting this method to build object recognition systems that use classifiers printed from the human mind.

1. Introduction

Computers routinely beat the human brain on challenges with logic and calculation speed. But, when it comes to object recognition, humans are still the state-of-the-art. What is the key difference between human recognition and machine recognition?

One answer is that the best object recognition systems today, despite incredible progress, are unable to imagine objects that they have never encountered. Yet, the human mind is able to effortlessly imagine objects that it has never seen, touched, or heard. And remarkably, humans can do this in any color, orientation, deformation, put upside down, in and out of context, all in vivid detail.

In this project, we *print* the mental images of what a human can imagine into an object recognition system. We combine the strengths of two approaches: state-of-the-art features in computer vision with a method in human psychophysics [1] that estimates the decision boundary that humans use for recognition.

Consider what may seem like an odd experiment: we sample a random point in a visual feature space from a standard normal distribution. What is the chance that this sample is a car? Fig.1a visualizes some samples [2] and, as expected, we see noise. But, let us not stop there. We next generate one hundred thousand points from the same distribution, and ask workers on Amazon Mechanical Turk to classify each sample as a car or not. Fig.1c shows the average of visual features that workers believed were cars. Although our dataset consists of only noise, a car emerges!

While sampling noise may seem unusual to computer vision researchers, a similar procedure, named *classification images*, has gained popularity in human psychophysics [1] for estimating the template the human visual system internally uses for recognition. In the procedure, an observer looks at random noise and indicates whether they perceive a target category. After many trials, psychophysics researchers can apply basic statistics to extract the internal template the observer used for recognition. We found that the same approach can build object recognition systems that use classifiers acquired from the human visual system.

Motivated by the observation that human visual system is a rich source of information, this paper investigates the scientific question: **can we extract visual classifiers from the human visual system?** We show how to use classification images to estimate boundaries from the human mind, but in a feature space that is compact and discriminative for computers. To our knowledge, we are the first to build classification images in computer vision feature spaces. Our experiments are promising, and take the first steps towards printing visual classifiers from the human mind.

2. Classification Images

The goal of classification images is to estimate the decision function that the human visual system uses to discriminate between two classes A and B. Suppose we have images $a \in A$ and $b \in B$. If we sample white noise ϵ and ask an observer to indicate the class label for $a + \epsilon$, most of the time the observer will answer class A. However, there is a chance that ϵ might manipulate a to cause the observer to mistakenly label $a + \epsilon$ as class B. The key insight is that, if we perform a large number of trials, then we can use basic statistics to estimate a decision function $f(\cdot)$ that discriminates between A and B, but makes the same mistakes as the observer. Since $f(\cdot)$ makes the same errors, it provides

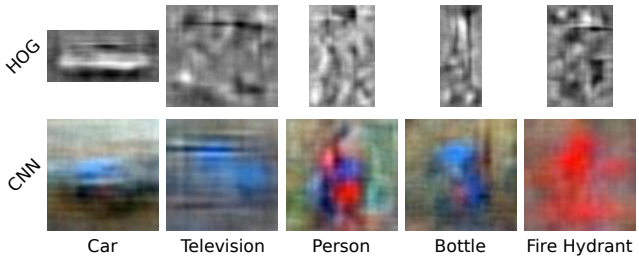

Figure 2: **Classifiers from the Mind:** We visualize some decision boundaries acquired from the MTurk workers' minds. The car classification image captures a darker road below the car, and a lighter sky towards the top. The television shows a rectangular structure, the person mimics a pedestrian, and the valves can be seen in the fire hydrant.

a good model for the internal decision function that the observer uses. For a more in-depth review, please see [1].

We discovered that this psychophysics approach can be modified to estimate classification images in computer vision feature spaces. Instead of sampling white noise ϵ in pixel space, we sample white noise in feature space from a zero mean, identity covariance Gaussian distribution. We then invert the noise features back to an image with [2] and ask observers to classify the inversion. After averaging the features that workers positively labeled on MTurk, we capture the visual template $c \in \mathbb{R}^d$ that the human visual system uses to recognize objects in a feature space. Fig.2 visualizes some classification images that we have extracted.

Since the classification image is estimated from the human visual system, we expect it to capture good biases about the visual world. We incorporate this bias into SVMs by constraining the hyperplane w to be oriented at most $\cos^{-1}(\theta)$ degrees away from the classification image c:

$$\min_{w,b,\xi} ||w||_2^2 + \lambda \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0$$
$$\theta \leq \frac{w^T c}{||w||_2 ||c||_2}$$

We rewrite the above as a conic program, which is convex by construction, and optimize it with off-the-shelf solvers.

3. Experiments

Our experiments suggest that it is possible to print classifiers from the human visual system. Since classification images do not depend on real images, Fig.3 shows it is possible to classify objects without training on any real images, which can be useful in situations where it is difficult to collect data, such as underwater or outerspace. As classification images are estimated only with noise, they tend to be biased towards the human visual system, which Fig.4 suggests is a favorable bias. Moreover, everyone does not necessarily share the same classification image, and Fig.5 reveals there is a cultural bias in the human visual system. Overall, these results hint that human imagination can be a futuristic resource for recognition systems.

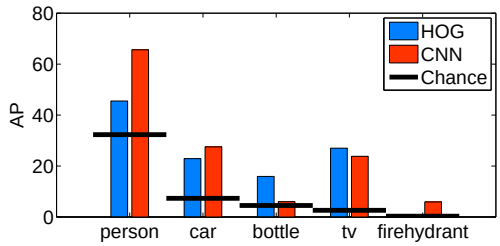

Figure 3: **Recognition without Data:** Even though the classification image was created without a dataset, it performs significantly above chance for object classification on PASCAL VOC 2011. If a machine learning algorithm were trained without data, the best it could do is chance.

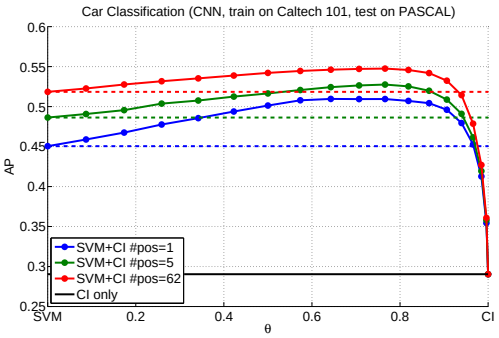

Figure 4: **Favorable Biases:** Since classification images are estimated only by humans looking at noise, it is biased towards the human visual system, which we suspect is a good bias. We train an SVM+CNN to classify cars on Caltech 101 and constrain it towards the classification image. When we evaluate it on PASCAL VOC 2011, generalization performance improves, suggesting this bias is favorable.

(a) India (b) United States

Figure 5: **Mental Images:** We discovered people do not share the same mental image of objects inside their head. We asked workers to classify CNN noise as a sports ball or not, and created a classification image by country. Indians seem to imagine a red ball, which is the color for a cricket ball and the predominant sport in India. Americans seem to imagine a brown or orange ball, which could be an American football or basketball, both popular U.S. sports.

[1] R. F. Murray. Classification images: A review. *Journal of Vision*, 2011.

[2] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: Visualizing Object Detection Features. *ICCV*, 2013.