Modeling and Predicting Object Attention in Natural Scenes

Thesis by

Merrielle Spain

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2011
(Defended May 13, 2011)

To Dad, who encourages questions

Acknowledgements

I thank my thesis adviser Prof. Pietro Perona for years of guidance, support, ideas, and inspiration. I am fortunate to have worked in his lab, and in the process I have learned a great deal from him about science and life.

Collaborating with a tenacious researcher, Wolfgang Einhäuser early in my graduate career taught me much about gathering, understanding, and explaining data. Prof. Yaser Abu-Mostafa, Prof. Serge Belongie, Prof. Christof Koch, and Prof. Shin Shimojo provided valuable feedback as my thesis committee.

I am grateful for the friendship and assistance of all my labmates in the Computational Vision Lab. Sitting next to Ryan generated both weighty discussions and levity. Peter collected bounding boxes and was a source of insightful conversations about Mechanical Turk. Alex, like a big brother, bestowed excellent advice. Greg and Marco gave me boundless computer assistance from hardware to obscure linux commands.

I would like to thank my wonderful friends and family for their love and encouragement: especially Angeline for her sage council, Ken for making every day brighter, and my dad for teaching me to: solve a similar problem, solve a simpler problem, or solve part of the problem.

My work was generously supported by a Graduate Research Fellowship from the National Science Foundation and a National Institute of Mental Health Training Grant.

Abstract

Humans automatically attend to certain objects in a scene. Better understanding this process could improve a computer's ability to parse scene images and convey information about them to humans. This thesis is arranged in three parts. The first part explores how important a particular object is in a photograph of a complex scene. We propose a definition of importance and present two methods for measuring object importance from human observers. Using this ground truth, we fit a function for predicting the importance of each object directly from a segmented image; our function combines many object-related and image-related features. We validate our importance predictions on a large set of objects and find that the most important objects may be identified automatically. We find that object position and size are particularly informative, while a popular measure of saliency is not.

The second part explores the relationship between object naming, eye movements, and saliency maps. Eye movements correlate with shifts in attention and are thought to be a consequence of optimal resource allocation for high-level tasks such as visual recognition. Saliency maps, are often built on the assumption that "early" features (*e.g.*, color, contrast, orientation, and motion) as opposed to objects themselves drive attention. We measure the eye position of humans viewing scenes and then ask them to recall objects that they saw in each scene. Weighted with recall frequency or maximum saliency, these objects predict fixations in individual images better than early saliency, suggesting that early saliency may have an indirect effect on attention, acting through detected objects.

The third part explores the problem of locating objects in a scene irrespective of category. We introduce the first benchmark for category-independent object detection. It is composed of a large public dataset of annotated high-resolution scene images and suitable metrics for performance evaluation. We demonstrate our benchmark by comparing three methods for generalized object detection against a baseline and an upper bound.

Contents

List of Figures

Chapter 1

Introduction

By summer 2011, Facebook will probably host 100 billion photographs [101]. Cheap cameras and memory have created an explosion of image data on the internet. But, how do you find the right image barring manual search? Two obstacles for a computer to recover the right image are: How does a human query the computer for an image? How does a computer parse the image contents?

There are straightforward solutions for single-object images. The human queries with an object category, such as “elephant,” or an instance, such as “Dumbo.” The computer performs category-level [43,44,70,149] or instance-level [79,80,142] recognition on images and returns likely candidates. However, most photos contain multiple objects. Recently, the problem of simultaneously detecting, localizing, and naming multiple objects in an image has become an active area of research [32, 116]. It is likely that we will eventually have software that can automatically list all the objects in an image. However, a laundry list containing dozens of object names might not be ideal for query. Certain objects might be more descriptive of the image than others.

Indeed, change blindness experiments [111] suggest that after looking at pictures of complex natural scenes, we retain information about only the overall gist of the scene and a handful of objects. The experiments show that we generally miss differences between two versions of the same picture, where differences have been introduced by photo editing, if changes are restricted to objects inessential to the overall meaning. This is related to a basic feature of human vision. We move our eyes about three times a second in a pattern of pause and rapid movement: fixation and saccade. Instead of an accident of evolution, this is believed to be a compromise between high visual acuity and rapid access to a large field of view. Visual acuity over our entire visual field equivalent to the fovea, the small central region of the retina, might require a ten-ton brain [39]. A sparse sampling suffices because the information content of the visual world is not uniform and humans are talented at sampling it. How do people know that a location is informative before they look at there? Different viewers tend to fixate similar regions of a given scene, although the sequence of fixations is highly variable [45, 85].

If you have viewed something "out of the corner of your eye," then you know that there is more to what you are looking at than where your eyes are pointed. The original notion was that *attention* "implies withdrawal from some things in order to deal effectively with others" [61]. It is generally believed that attention's main function is the allocation of processing resources to accomplish complex tasks. The two-process theory of detection, search, and attention differentiates between automatic detection and controlled search [119]. For instance, some search tasks are always fast, while ones that require conjunctions of features slow as the number of distractors increases [137]. If cheap tasks can identify where to attend, then expensive processing can be allocated effectively.

Pre-motor theory holds that spatial attention results from weaker activation of the same brain circuitry that drives saccadic eye movements [113]. This suggests that if an eye movement is made to a particular location, attention will arrive first and cannot be sent elsewhere. Indeed, the costs and benefits of attentional cueing, or indicating a search target's future location, can be eliminated by requiring saccades to other locations [54].

What guides attention? Although the concept of selective visual attention dates from the 19th century [61], the factors driving this selection process are still far from understood. First, what is the role of top-down factors (*e.g.*, task, observer idiosyncrasies) as compared with factors that can be inferred from the stimulus? Second, what is the role of low-level features such as contrast, color, orientation, flicker, or motion as compared with high-level stimulus structure such as objects or gist? We focus on the second question, with eye movements as correlates of attention [113]. Specifically, we ask whether fixations are driven directly by low-level or *early saliency*, or through correlations with high-level scene structure, such as the saliency of recognized objects. Is attention driven by mechanisms that are earlier than and independent of recognition, or is attention part of the recognition process itself?

Most attention models are based on a *saliency map* and a dynamical process for visiting saliency maxima [58, 67]. Filtering the input image with kernels reminiscent of early visual mechanisms generates feature maps at various spatial scales. These are then combined into a single saliency map, which encodes the probability that an image location will be attended. The saliency map is entirely based on early features and was originally designed to explain covert attention on simple stimuli. Saliency maps predict fixations in complex scenes to some extent [97, 99, 104, 127]. Some authors hope that, by progressively refining low-level models, human attention will eventually be modeled perfectly.

In this view, attention operates independent of object recognition and may be thought of as guiding scene analysis. This view has recently been challenged. Even if features of the saliency map, such as luminance-contrast, are good correlates of fixation [68,83,84,109] several authors have argued that they might not drive attention causally [9, 22, 129], but contingent on high-level statistics [25]. Rhesus monkeys preferentially fixate image regions with semantic content instead of noise regions with the same low-level statistics [63], and it has been suggested that objects, such as faces, may drive attention in a direct fashion [10, 51, 52], although there is some contrary evidence [141]. Along similar lines, the perceptual experience instead of the stimulus predominately influences eye movement behavior when viewing art that has ambiguous experiences [128]. Therefore, even without an explicitly formulated task, eye movements are largely influenced by high-level scene properties and scene interpretation.

The fact that the specifics of the task influence eye motions had been noticed as early as Buswell [8]. In his seminal study, Yarbus used a variety of tasks, including abstract interpretations, such as the judgment of social status [148]. In these cases, the task clearly dominates the fixation patterns, as it does in complex activities of daily living [69]. Recent studies suggest that during visual search, early saliency has little impact on fixation patterns [26, 48, 139], and the effect of a stimulus feature on fixation depends on its relation to the search target [102]. Models that modulate low-level channels attempt a mechanistic explanation for such top-down regulation [92, 107, 138].

Aside from task and stimulus-features, search in natural scenes is influenced by prior knowledge on the typical spatial location of the search target, as well as by contextual information. Modulating saliency map models with such priors improves their fixation prediction [135]. Such spatial priors may influence fixation behavior beyond search. The *central bias* of observers, the tendency for observes to fixate near the center of natural scene photographs, might reflect the expectation of interesting objects in this region [129]. In this view, spatial priors are believed to be a bottom-up function of scene statistics learned from experience and applied in a task-dependent top-down manner.

The *attentional bottleneck* view is that attention precedes recognition in the processing pipeline [91]. When different information is presented to each ear simultaneously and an observer attends to one, most words in the unattended ear cannot be recognized [6]. However, the fact that important stimuli, such as one's name, can be recognized [90] conflicts with the attentional bottleneck view. The precise relation of attention and recognition, is largely unresolved.

The extent to which overt, or eye movement associated, attention is needed to recall an item has been studied extensively. Unexpected items are fixated longer and recalled better [42]. In brief presentations, change detection requires close fixations [94]. There is an advantage for detecting changes in items fixated earlier and a correlation between the time spent fixating an item before the change and change detection [55]. Consistent with these results, information about an object's position is accumulated over fixations, but there is some evidence that information about object identity is not [130]. Besides, better memory for fixated items, it has been argued that changed items are also fixated earlier after the change [96]. This has been challenged with experiments embedding objects in a complex background, which find change detection restricted to a small region around the current fixation [50]. Although the details of the relation between fixation and memorization seem dependent on experimental paradigms, all these data suggest that there is some relation between the allocation of overt attention and the ability to recall certain properties of an item.

Attention-free feed-forward systems perform well on category recognition tasks when the scene is pre-segmented into regions containing a single object category [20, 34, 38, 72, 89, 112, 118, 145]. However, real-world objects generally occur in clutter not isolation, and may cover as little as 0.1% of the image area [117]. In clutter, attention may be a necessary preprocessing step for recognition [15], for learning new objects, and for hastening recognition [117]. In summary, while psychophysical evidence suggests that spatial attention is unnecessary for recognizing isolated objects or the gist of isolated scenes, attention most likely supports recognition in spatially and temporally cluttered settings. The interaction of attention and recognition in natural conditions is thus of interest for human and machine vision.

The literature suggests that saliency maps based on early visual features have some power in predicting eye movements and attention in natural complex scenes. This limit is likely intrinsic, and high-level visual properties of a scene will have to be considered to see a significant predictive improvement. While clearly objects such as faces have the power to draw attention, we are still far from a quantitative model that predicts eye movements from the configuration and visual properties of objects in a scene.

Eye movements indicate attention to locations, not objects. Moreover, there is evidence for separate object-based and space-based attention systems. Space-based attention is evidenced by negative priming for an object occupying the same space as the attended object in overlapping line drawings [134]. Object-based attention is supported by better performance and faster reaction times for moving attention within an object than the same distance across objects [17, 19]. Hence, fixation provides information about space-based attention but not object-based attention, especially when objects are crowded or hierarchical.

As we are interested in the high-level object-based attention, we wish to develop an object-specific correlate of eye motions. We consider object naming to be a way of discovering what observers are attending to. Dorsal simultanagnosiacs can only attend to one object at a time, and correspondingly can only see or recognize one object at a time, even when the objects occupy the same space [33]. While it is possible to detect animals and vehicles without attention [75], in the same paradigm subjects fail to identify (or localize) targets that they had correctly detected [30]. Also, several phenomena indicate that attention has a role in recognition, such as inattentional blindness [93,122], change blindness [111], repetition blindness [62], and the attentional blink [21, 30, 108]. In cluttered scenes, viewers must fixate an object in clutter to recognize it (*e.g.*, *Where's Waldo?*), indicating a need to separate an object from its surrounds to recognize it [74]. Hence, we consider object naming, which requires identification, indicative of object-based attention.

We explore the relation between attention and recognition in a natural setting, with semantically rich natural photographs [120, 121]. This thesis consists of three parts: First, we explore whether it is possible to identify the important objects in a given scene automatically and produce a concise list that would facilitate image search and other applications. To do this we analyze how a large group of viewers name objects in an image. Given certain observations we model how viewers name objects, and fit that model to measure object importance. Then, we use this measured object importance as a ground truth that we predict directly from an image and manually-segmented objects. This part is based on two published papers [124, 125].

Second, we explore which object properties drive attention. We test the hypothesis that the most meaningful object in an image attracts attention and, with this effect removed, raw saliency maps have little predictive power. To ensure alertness, while preserving natural viewing behavior, all observers are asked to aesthetically evaluate each picture. To investigate the effects of visual search on fixation statistics, half of the observers search for a verbally defined target object. In both conditions, after the image disappears and aesthetic evaluation, we ask observers to characterize scenes with keywords to measure which objects were seen and remembered as significant. For both conditions, we assess the mutual relation among three quantities: the locations our observers fixate, the locations of objects they recall, and the locations of highest saliency according to the Itti and Koch model [58]. This allows us to compare how well five quantities predict fixations: raw saliency, object saliency, an optimal combination of both measures, the mutual prediction of different observers, and general spatial biases. This part is based on published work with Wolfgang Einhäuser [27].

Third, we ask observers to identify objects with bounding boxes instead of naming. This creates a new measure of object importance that is instance-specific instead of category-specific. We combine this bounding box measure of importance with category-independent object detectors to a create system which can identify bounding boxes that likely contain important objects. We present the first benchmark for category-independent object recognition and importance prediction. Figure 1.1 shows boxes that an idealized category-independent object detector might output. These bounding boxes are actually an example of the ground truth in our dataset.

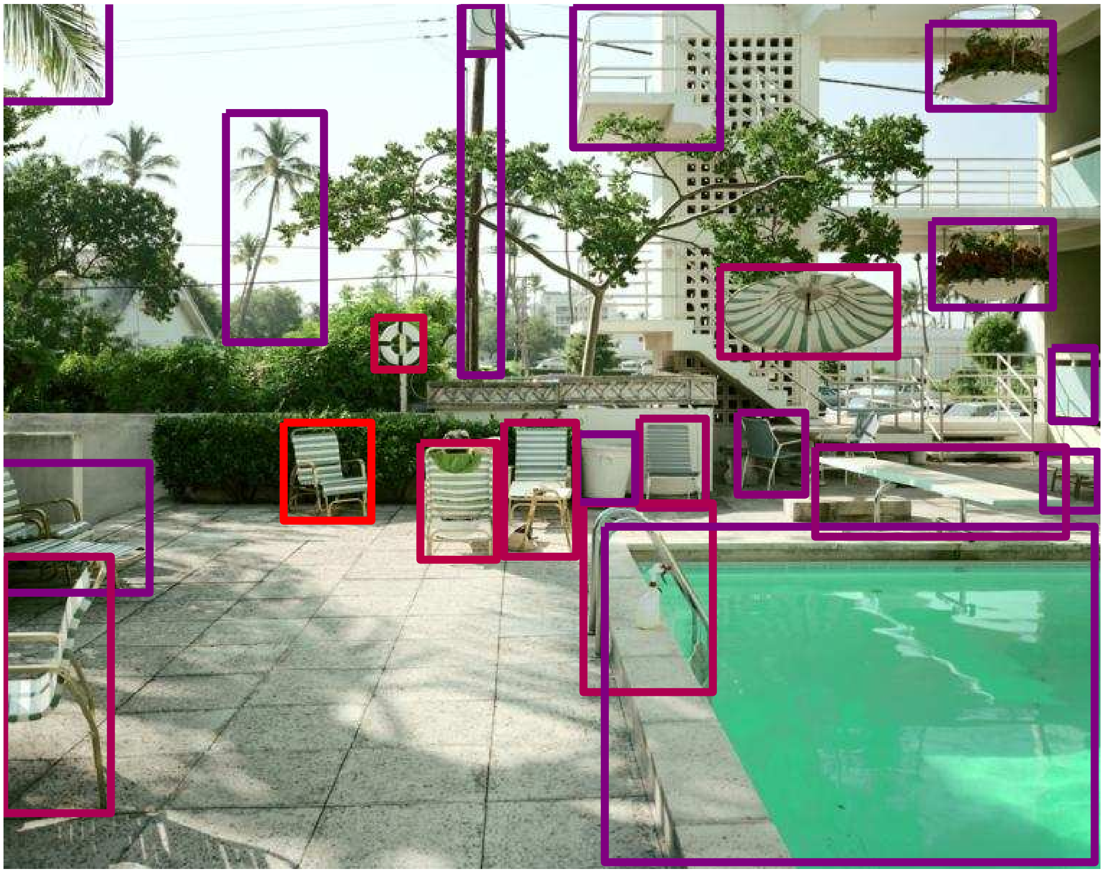


Figure 1.1: Example image annotated with boxed objects.

Part I

Object Importance

We face two main challenges: measuring importance, as perceived by viewers, and automatically predicting the importance of objects in a given image.

Figure 9.3 depicts how these ideas fit together. Chapter 2 describes how we collect importance information from viewers. Chapter 3 considers the problem of measuring importance by aggregating data collected from many viewers. Chapter 4 explains how to predict importance from bottom-up visual properties of an object. We discuss how subtle manipulation of the human task affects importance in Chapter 4.5. Chapter 4.5 summarizes our main findings.

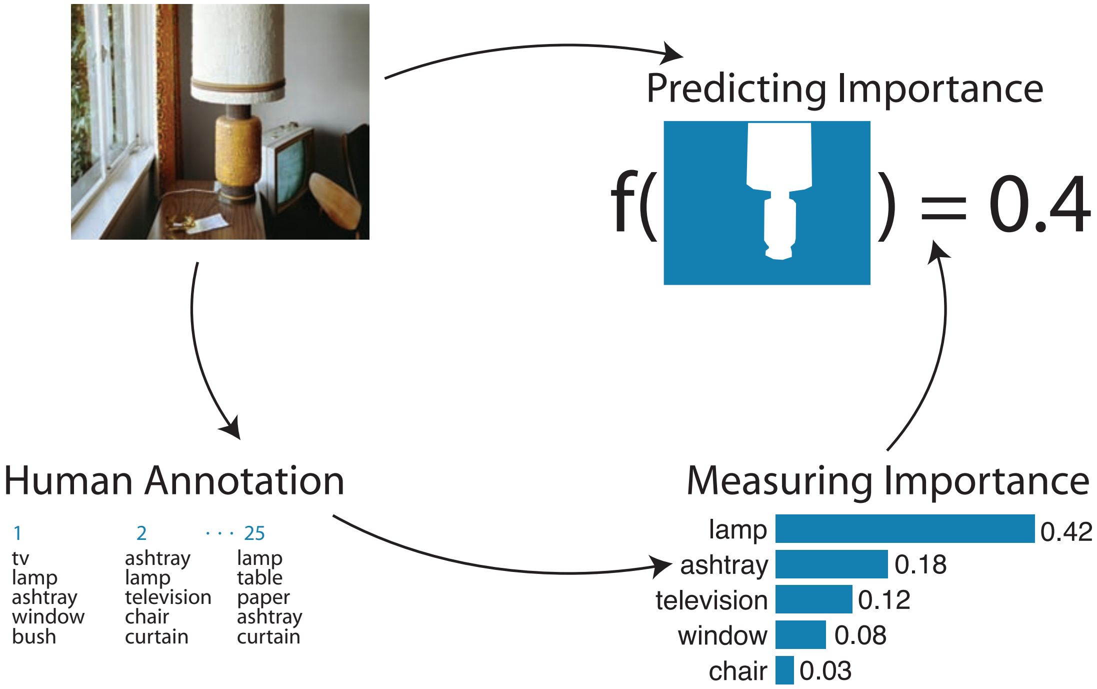


Figure 1.2: We wish to predict the importance of an object in a photo. To accomplish this, we must first produce a ground truth. We do so by combining the opinions of many viewers (bottom arrow; Chapter 3). From this ground truth we may learn a function for predicting object importance from image regions (top and right arrows; Chapter 4).

Chapter 2

Object Naming

2.1 Human Annotation

Our first step is to discover which objects humans consider important in a given image. We put off a formal definition of importance to Chapter 3. For the moment we rely on the intuitive notion and explore ways of identifying which objects people notice most in a photograph.

2.2 Previous Work

Some previous research explores what people can recognize under extreme circumstances. Fei-Fei *et al.* [36] examine how limited viewing time affects what viewers report. Torralba *et al.* [136] investigate which objects people can name with limited image resolution.

The ESP game, by Ahn and Dabbish [144], presents two players with an image. Each player types words independently. Their task is to produce a matching word in the fewest attempts. When the players produce a common word, the game ends, banning that word from future games. When multiple games are played on the same image, the resulting words form an ordered list. Intuitively, words associated with more important objects will tend to come up earlier. However, words are sometimes adjectives (*e.g.*, funny), word order is noisy since only two players play together, and players may develop strategies for reaching consensus quickly, for example naming the prevalent color in the image, or typing whatever text may be present.

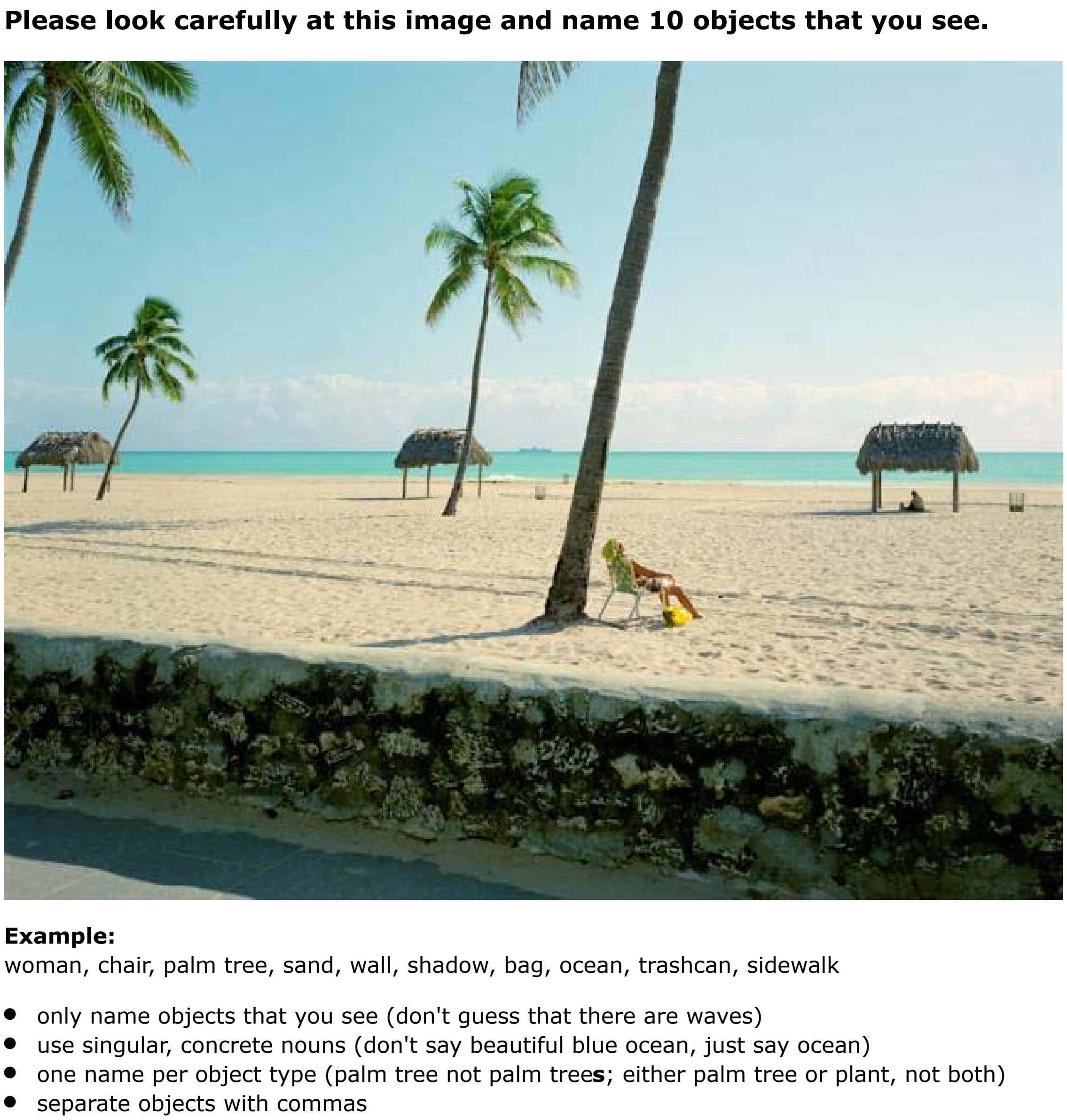
Elazary and Itti consider the order in which objects are named in LabelMe, a measure of object *interestingness* [28, 115]. In LabelMe users name an object and outline its contour with mouse clicks [115]. A user may annotate any objects in an image. Results from past users are visible to future users, so an object token can only be outlined once, producing a single list. This is problematic because, as we shall see in Chapter 3.6, viewers produce lists with inconsistent object order. Furthermore, the choice of object is influenced by the ease of outlining the object (*e.g.*, a window has a simple contour, while a tree in winter has a complex contour) and by the specific needs of the annotator, such as collecting a pedestrian database.

2.3 Data Collection

We designed a method for collecting object importance data with two criteria in mind: first, the data should be collected independently from many human viewers and second, our annotators should not be motivated by tasks that bias the data.

We collected ordered lists independently from 25 viewers for each image. Through Amazon Mechanical Turk, U.S. viewers were instructed "Please look carefully at this image and name 10 objects that you see." We asked for ten objects so that viewers wouldn't just name one or two. Each scene photograph was rescaled to a 600 pixel diagonal. Most viewers labeled fewer than 20 images, while a handful labeled all of them. We found that very few lists were empty or nonsense. Viewers received $0.10 per annotated image, and quality is encouraged as all work on Mechanical Turk must be approved by the requester prior to payment. Viewers were given the same detailed instructions for each task, which can be found in Figure 2.1.

Before analyzing the collected lists, we cleaned them in four steps. First, we eliminated lists that were empty or contained nonsense words. Second, we corrected misspellings with a spell checker. Third, we identified synonyms for each word in each list using WordNet [1]. Fourth, for each image we chose one synonym for each group of words. This step was necessary because the same word could have different



Please look carefully at this image and name 10 objects that you see.

Example:
woman, chair, palm tree, sand, wall, shadow, bag, ocean, trashcan, sidewalk

- only name objects that you see (don't guess that there are waves)
- use singular, concrete nouns (don't say beautiful blue ocean, just say ocean)
- one name per object type (palm tree not palm tree**s**; either palm tree or plant, not both)
- separate objects with commas

Figure 2.1: Detailed instructions given to all viewers on Mechanical Turk.

meanings in different images. For example "building" could mean house in a suburban picture or skyscraper in an urban one. The fourth step took the longest, requiring approximately 30 hours of manual labor.

2.4 Image Collection

We selected 97 pictures from Stephen Shore's collections "American Surfaces" and "Uncommon Places" [120, 121]. Shore took a photographic diary of his experience traveling in North America in the 1970s. Our collection of photos contains 22 bedroom scenes, 4 living room scenes, 5 pool scenes, 19 portraits, 35 suburban scenes, and 12 urban scenes. Figure 2.2 displays a representative sample of these photos. We picked these scenes because they are commonplace and represent the overall statistics of the collection. We did not include images that might have been disturbing or offensive to some viewers.

We chose to sample from the Shore collections because we needed an objective, representative, and meaningful set of scenes for our experiments. By objective, we mean that the choice of scenes should be as independent as possible from the experimenters and their goals. By representative, we mean that the collection of images should sample human visual experience broadly. By meaningful, we mean that the images should represent notable moments in a person's visual experience. If we collected objective and representative photos like Switkes [87], by attaching a camera to a bicycle helmet and snapping one picture per minute automatically, most photographs would be meaningless (*e.g.*, the edge of an elevator door). So Shore's photos are more objective than an object recognition dataset and more meaningful than randomly captured photographs.

2.5 Data Overview

Comparing lists Examples of ten-object lists produced by five viewers are displayed in Table 2.1. The number of objects that are present in an image may be

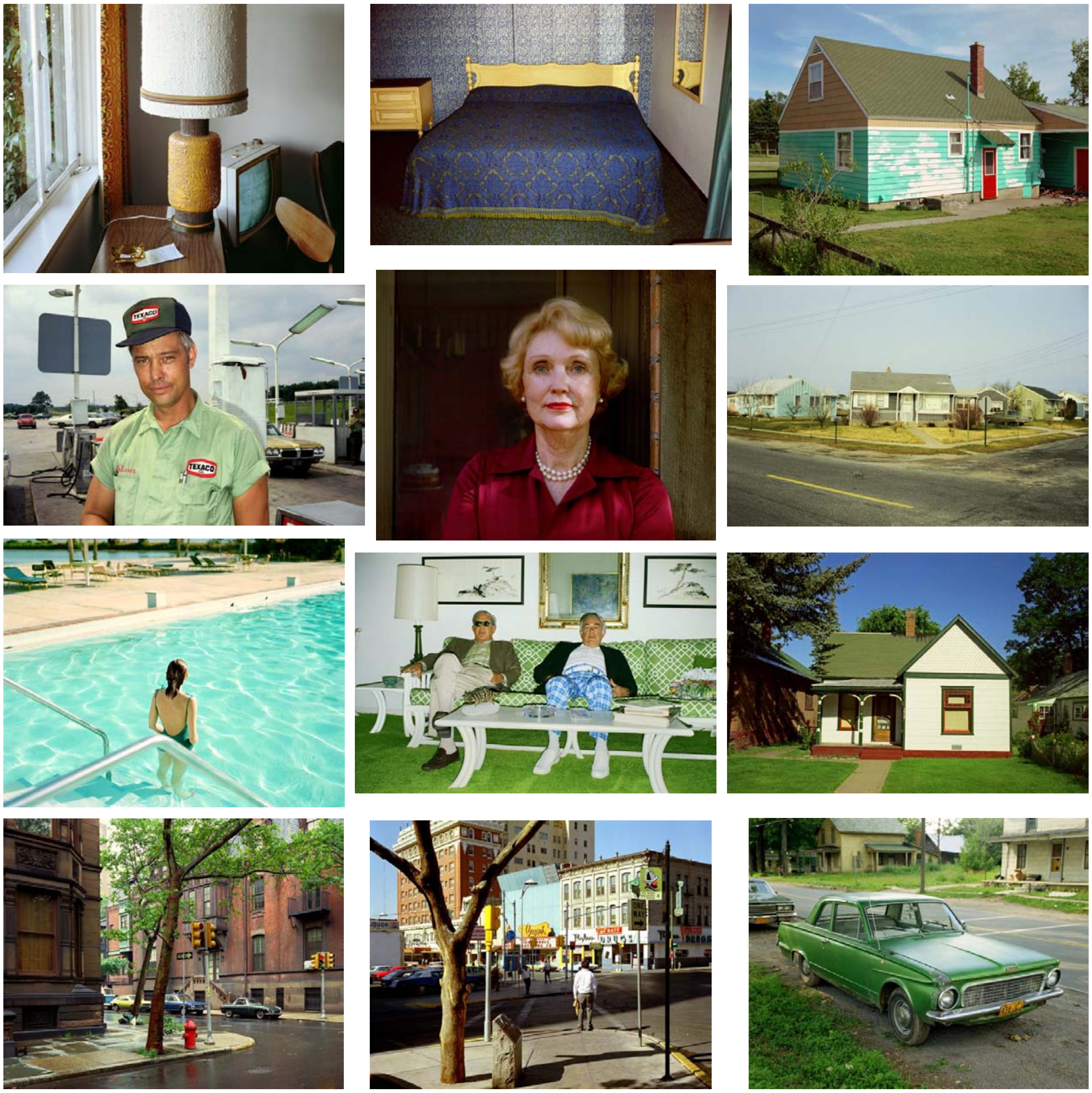


Figure 2.2: Representative sample of our images. These photos by artist Stephen Shore are a visual diary of arresting moments rather than a collection taken by a computer vision researcher for a particular purpose.

Table 2.1: Sample lists from five viewers of the first photo in Figure 2.2 as columns.

1	2	3	4	5
lamp	lamp	tv	ashtray	curtain
television	tv	lamp	lamp	table
chair	chair	ashtray	television	chair
ashtray	table	window	chair	cord
paper	ashtray	bush	curtain	lamp
table	matches	table	window	paper
curtain	paper	cigarette	paper	tree
window	window	paper	table	wall
wall	plant	chair	shade	window
shadow	curtain	curtain	latch	ash tray

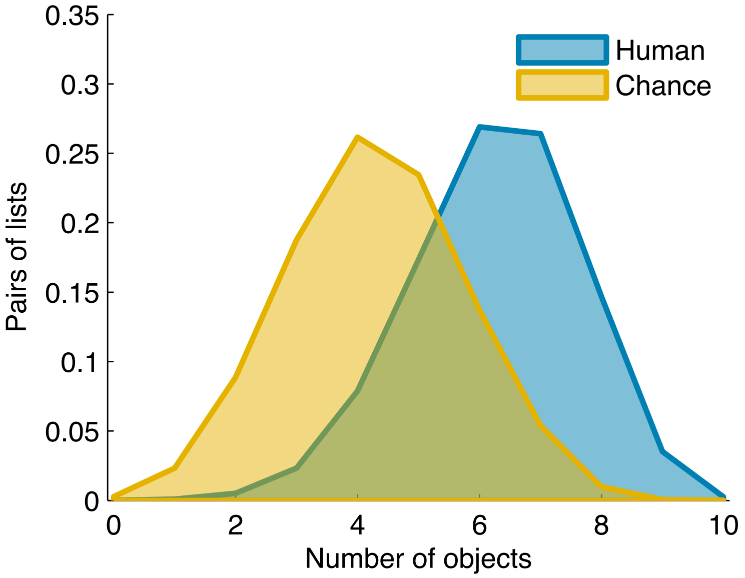


Figure 2.3: Histogram of the number of objects shared by a pair of lists for the same image. Data collected from viewers (blue) is compared with random lists created by uniformly sampling objects named for that image (yellow).

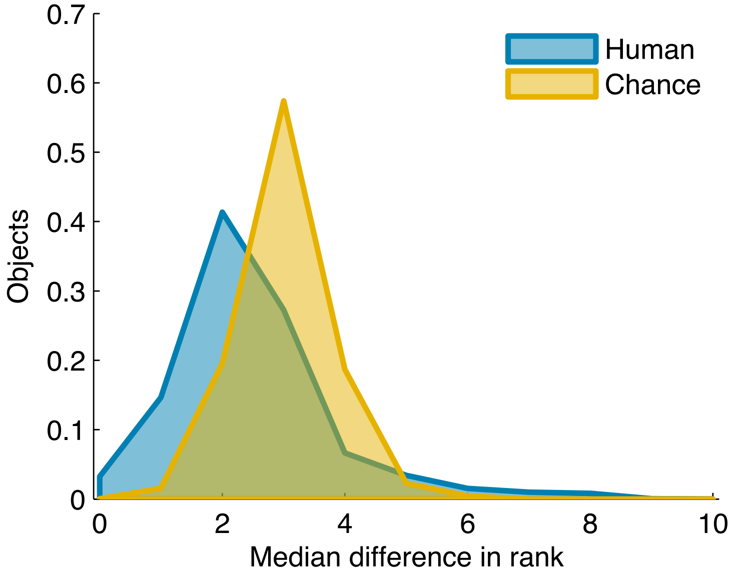


Figure 2.4: Histogram of an object's unsigned difference in the rank between two lists for the same image. Each data point is the median of these differences for an object-image combination.

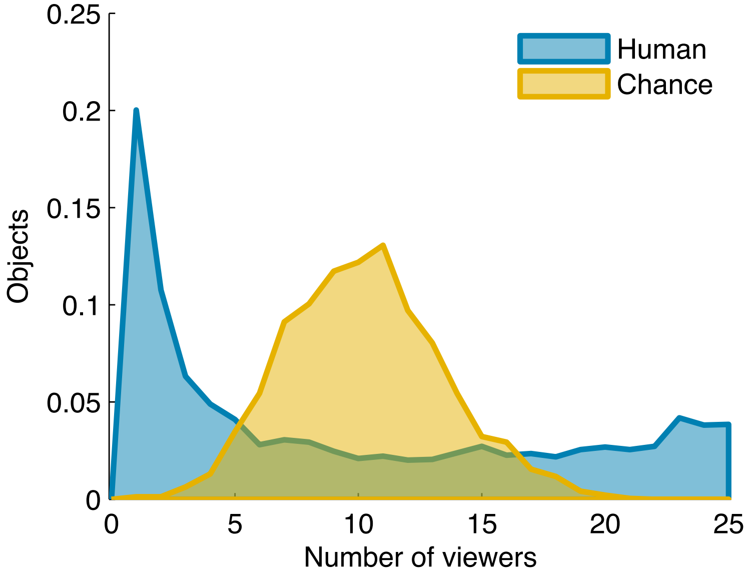


Figure 2.5: Histogram of the number of objects named by a particular number of viewers. Each data point represents an object in a specific image.

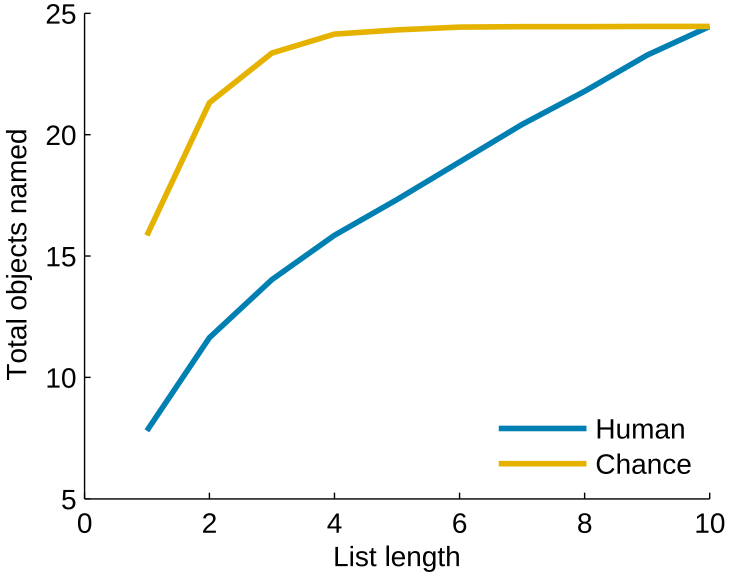


Figure 2.6: Total number of objects named per image as we consider longer lists. Lists of length k are obtained by selecting the top k elements of each list.

estimated by considering the size of the union of the 25 ten-word lists provided by our subjects for that image. We find that each image contains 16 to 40 (mean/median 24) objects. Correspondingly, both the composition and order of the word-word lists vary. To understand the structure of the lists, we compare these lists with chance lists. To generate the chance lists we consider the set of objects named in this image and randomly select ten objects with uniform probability. We generate 25 chance lists per image.

First, we examine a pair of lists (generated by the same process) and count how many objects the lists share. Figure 2.3 shows that pairs of lists from viewers have a much larger intersection of objects than expected by chance (mean 6.2 versus 4.3).

Second, we find a pair of lists that share an object, note the object's rank in both lists, and take the difference of those ranks. If the object appears in the same spot on both lists, then the difference in rank is 0, whereas if it appears first on one and last on the other, then the difference in rank is 9. We then take the median of the rank differences, so as not to double count objects. Figure 2.4 shows that an object's rank changes slightly less between human lists than expected by chance (mean of 2.5

versus 3.1). These distributions are statistically different for both list intersection and rank difference ($p = 0$ and $p = 10^{-111}$, Wilcoxon rank sum test).

Third, we look at all the lists for an image and count how many viewers name a particular object. Figure 2.5 shows that the number of viewers that name an object has a much larger variance than expected by chance. The lists generated by humans have many objects that are only named once.

Fourth, we count how many objects are named in the top k words in each list. Figure 2.6 shows that fewer objects appear at the top of the lists than would be expected by chance. Notice that for the chance lists, the object count for an image saturates after the first four objects are named, while the object count climbs more slowly for the human lists. This indicates agreement in the objects that viewers name early.

Naming Independence Another issue concerning list structure is whether object naming is independent. Will one object being named change the likelihood of another object being named? Given an image that contains both cars and tires, if someone says "car," does that make them more likely to say "tire?" Please note that this is a different concept than Rabinovich *et al.* who ask whether cars and tires appear in the same images [106]. We are not discussing the state of the world, but rather what people name, given the state of the world.

To answer this question we test whether the observed co-occurrence is consistent with independent naming. For a given object pair, we find all the images that contain both objects and gather the lists associated with these images. We perform a Pearson's chi-square test with the Bonferonni correction ($p \leq .05/tests$) and Yates' correction for continuity only if both objects are present/absent at least five times (4,224 of 15,043 list pairs). The value of Pearson's chi-square test-statistic is

$$\chi^2_{Yates} = \frac{(|O_1 - E_1| - .5)^2}{E_1} + \frac{(|O_0 - E_0| - .5)^2}{E_0}, \tag{2.1}$$

Where O is the observed count and E is the expected count given the marginal frequencies. The subscript 1 denotes that both objects are named and 0 denotes

Table 2.2: Object naming is largely independent of other named objects. These are the only object pairs found to be dependent out of the 4,224 pairs.

Word pair		p value
		1.0e-05 ×
eye	nose	0
door	window	0
head	skin	0
eye	hair	0
eyebrow	skin	0
hair	nose	0
shoulder	skin	0.002
mouth	nose	0.01
finger	nose	0.02
roof	window	0.06
finger	skin	0.07
eye	mouth	0.1
door	roof	0.3
neck	skin	0.3
nose	skin	0.3
chin	nose	0.3
eyebrow	shoulder	0.5
hair	hand	0.7
finger	neck	0.9

otherwise.

We find that generally one object being named does not significantly influence the probability of another object being named. Only 19 of 4,224 tests (0.4%) show significant dependence. Table 2.2 enumerates the dependent object pairs; for all of these pairs the observed co-occurrence is greater than expected co-occurrence.

Failure to name the obvious We noticed an interesting phenomenon: viewers sometimes fail to mention the most obvious object (Figure 2.7). We identify the *obvious object* as the object named early and often, the earliest in mean order of the more frequent half of objects. This criterion captures when an object is the main focus of an image. Interestingly, the frequency distribution of obvious objects is bimodal; many people fail to mention some obvious objects. For instance, most

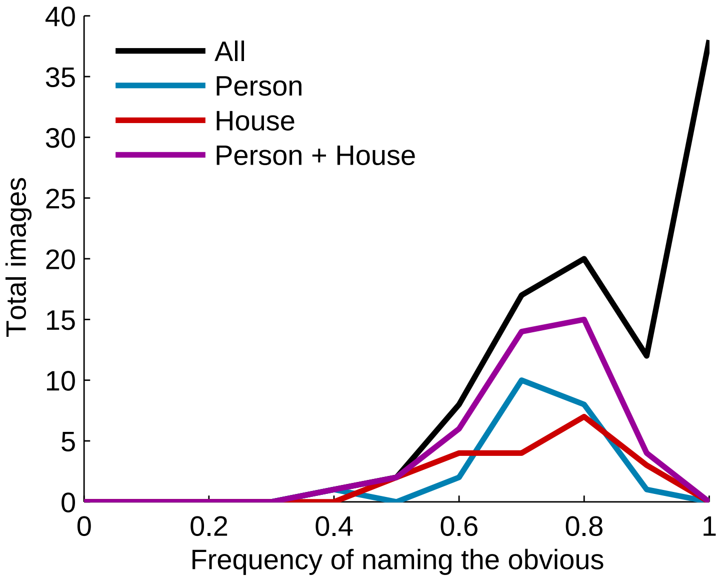


Figure 2.7: Some viewers fail to mention the obvious object. We histogram the number of images by the frequency that people mention the obvious object. While most viewers name “person” or “house” very early, others fail to mention them.

viewers name “person” or “house” very early, but others fail to mention them at all. These two objects account for nearly all images in which the obvious object is frequently missed. Because viewers often fail to name the obvious object, frequency is poor at identifying the most important object in an image. One possibility is that people become accustomed to the photos and stop naming things they have seen often. The data in Figure 2.8 rule out this hypothesis. How frequently the obvious object is skipped does not increase as the viewer labels more images; it is the same on the twentieth as it is on the first image labeled.

Westerners are better at ignoring context or frame and focusing on contents [66]. Hence, if the important object is large, viewers might consider it background and think of the other objects in terms of it.

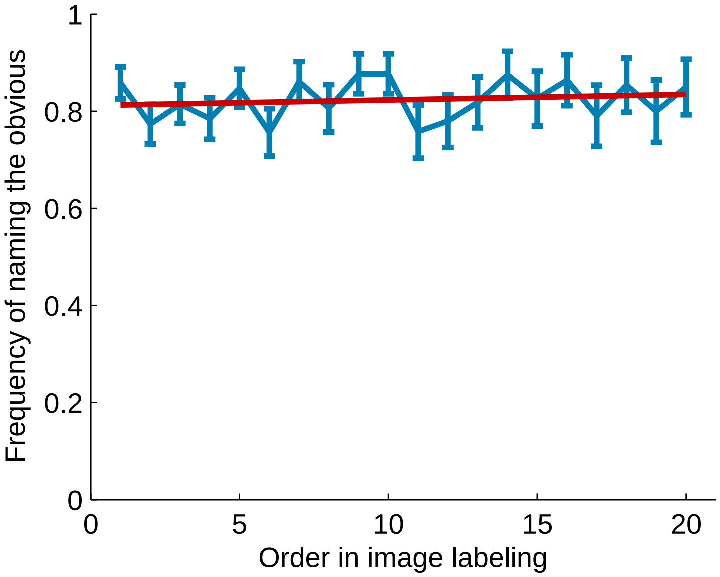


Figure 2.8: The frequency that the obvious object is named does not decrease as a viewer labels more images.

Chapter 3

Measuring Importance

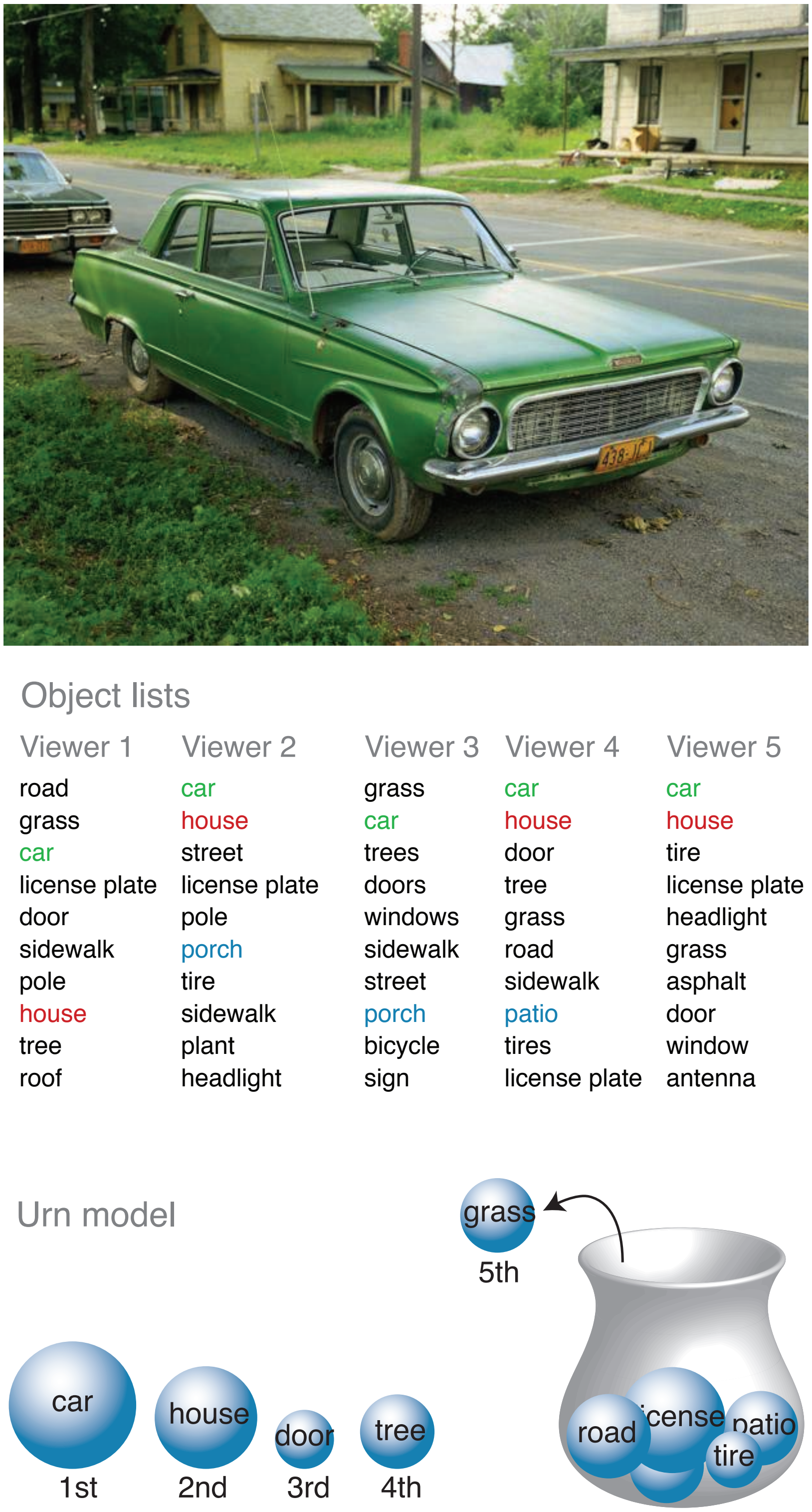
The observations that most objects are named independently and some objects are named early and often (Chapter 2.5) prompt us to formalize the concept of importance as

> An object's *importance* in a particular image is the probability that it will be mentioned first by a viewer.

In principle, we would need an extraordinary number of viewers to be able to directly calculate the importance of all the objects in a picture: some objects' importances may be less than 1%, and we would need hundreds of viewers to determine that. In this section we show that it is possible to measure an object's importance from fewer viewers by asking them to name more objects and creating models that take advantage of object order.

3.1 Urn Model

We model the naming of objects in an image with drawing marbles from an urn without replacement (see Figure 3.1). The urn contains one marble for each object category appearing in the image. The marbles are different sizes, affecting their probability of being chosen. Thus, a marble's size represents the importance of the corresponding object. We represent multiple viewers by refilling the urn with the same set of marbles and sampling.



Object lists

Viewer 1	Viewer 2	Viewer 3	Viewer 4	Viewer 5
road	car	grass	car	car
grass	house	car	house	house
car	street	trees	door	tire
license plate	license plate	doors	tree	license plate
door	pole	windows	grass	headlight
sidewalk	porch	sidewalk	road	grass
pole	tire	street	sidewalk	asphalt
house	sidewalk	porch	patio	door
tree	plant	bicycle	tires	window
roof	headlight	sign	license plate	antenna

Figure 3.1: A photograph and corresponding lists generated by five viewers. Words are color coded to facilitate perception of word order. The urn models how humans name sequences of objects. An image contains many object categories with varied importance in that image. A viewer names objects one by one until ten are named. Similarly, an urn is filled with marbles of different sizes, where larger marbles are more likely drawn. Ten marbles are removed from the urn, creating a sequence.

This model is based on several assumptions. First, the draws are independent; this is reasonable because very few object pairs are dependent (Chapter 2.5). Second, everyone starts with the same urn; we don't see clusters of different viewer behavior in our data, as we discuss in Chapter 3.6. Third, marbles can only be removed from the urn by being drawn. The third assumption is violated for some images. As discussed in Chapter 2.5, we find that obvious objects are named early or left unnamed. To model this we develop a variant of the urn model, which we call the forgetful urn. In this model, viewers draw marbles as before, but the first marble may go unreported with some probability.[1]

Figure 3.2 shows importance measured through maximum likelihood (ML), maximizing the likelihood of observing our data with the importance values as parameters (Chapter 3.1). The forgetful urn and the urn produce similar estimates of importance when the most obvious object is not often overlooked, but the forgetful urn's estimates are more realistic than the urn's when the obvious object is frequently skipped.

One possibility is that certain objects are named earlier and more often because they are more frequent in human speech and hence easier to access. Figure 3.3 compares object importance and naming frequency in our experiments with lexical frequency from the British National Corpus [73]. There is no pattern showing that lexical frequency is responsible for the observed naming behavior.

In the urn model that we just described, the probabilities of being drawn are what we are trying to measure from the data. Previous work on this problem uses complex numerical methods [40] or requires many marbles of the same type (we have only one) [82]. Instead of using these approaches, we measure importance by maximizing the likelihood or probability of observing a set of sequences given the object importances π_i.

Each sequence consists of 10 marbles w_i^m, where w_i^m denotes the ith marble drawn in the mth sequence and is a variable that takes values $1, ...N$ corresponding to object names. The w_i^m are drawn independently without replacement (out of N marbles,

[1] A rigorous definition of importance is the probability that a marble is drawn first, regardless of whether it is skipped.

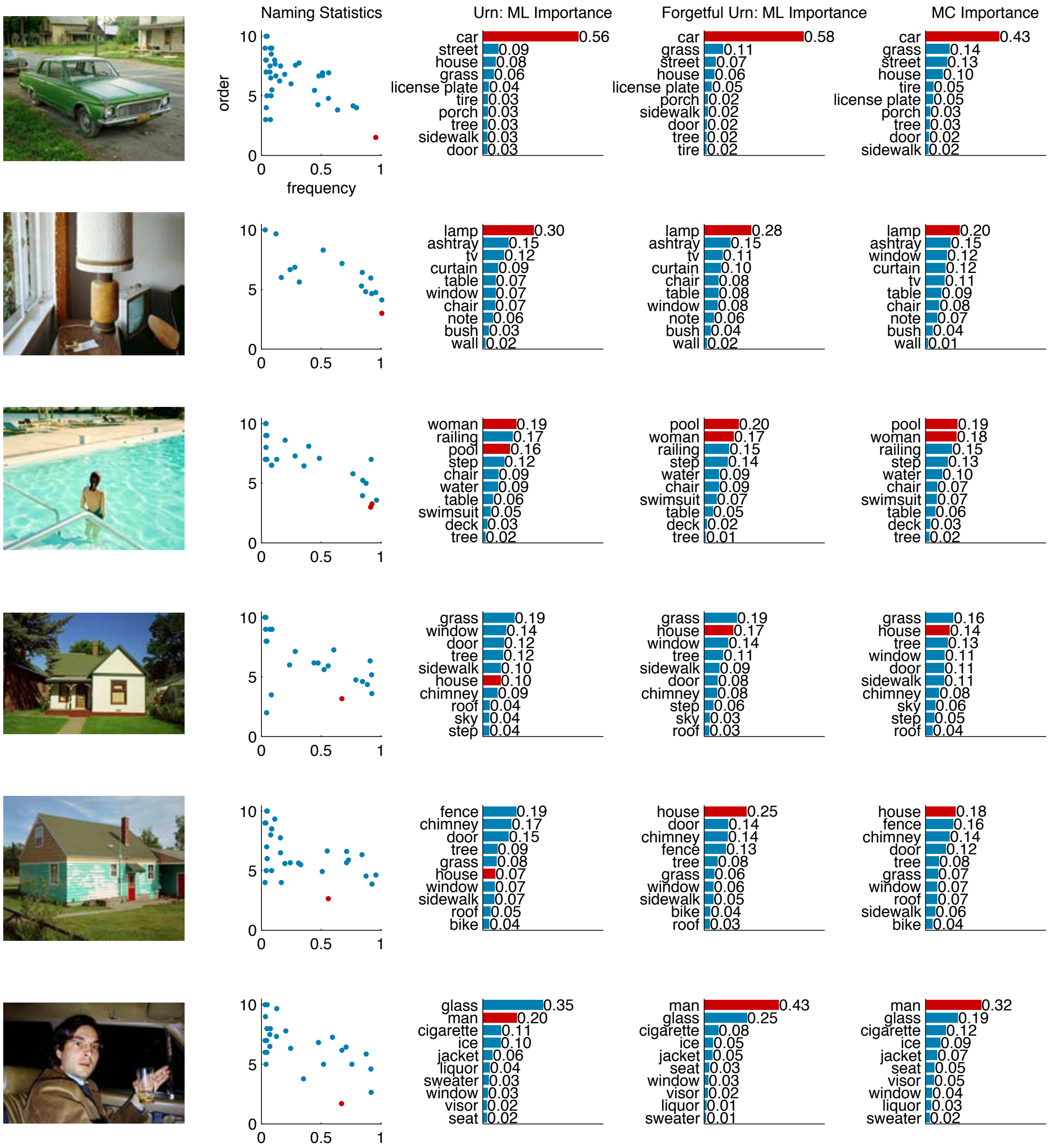


Figure 3.2: Measured Importance. Scatter plot of frequency that an object appears on lists and mean order over lists for an image (2nd column). A comparison of the mean order and frequency an object (dot) shows that in some images the obvious object (red) is sometimes not named at all. This violates our urn model, but we can compensate for this behavior and see an improvement in importance measurement in these cases for the Forgetful Urn (4th column) over the Urn (3rd column). In the cases where the obvious object is not skipped the importance measurement is similar. The Markov chain (5th column) arrives at similar results through a different approach.

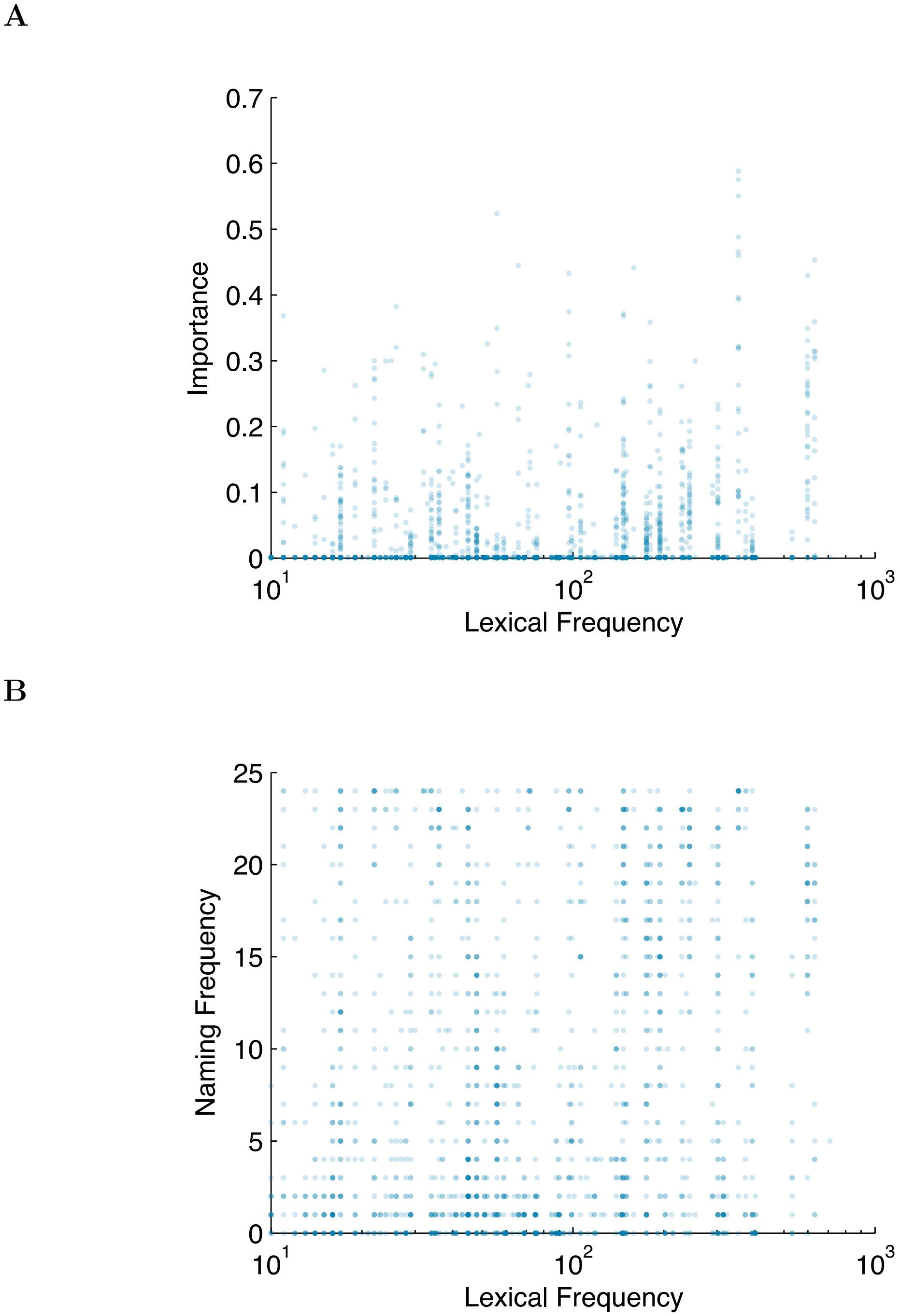


Figure 3.3: **A** Scatter plot of object lexical frequency and importance. Each dot represents an object in a particular image. **B** An object's lexical and within image naming frequency.

where $N >> 10$), so the probability of drawing a particular sequence of marbles $(w_1^m, ...w_{10}^m)$ is

$$\prod_{n=1}^{10} p(w_n^m | w_{n-1}^m, ...w_1^m) \ . \tag{3.1}$$

However, we are drawing marbles without replacement, so this equation is constrained by $w_i^m = w_j^m \implies i = j$. When we draw the nth marble of a sequence, $n-1$ marbles have already been removed from the urn, so we need to normalize the remaining importance to 1. The probability that the marble labeled w_n^m is the nth marble drawn is

$$p(w_n^m | w_{n-1}^m, ...w_1^m) = \begin{cases} 0 & \text{if } \exists i \in [1, n-1] : w_i^m = w_n^m \\ \frac{\pi_{w_n^m}}{1 - \sum_{i=1}^{n-1} \pi_{w_i^m}} & \text{otherwise,} \end{cases} \tag{3.2}$$

where π_i is the probability that marble i is drawn first (from a fresh urn) and $\sum_i \pi_i = 1$. The first case simply asserts that we are drawing marbles without replacement, so a marble cannot be drawn twice. If we assume that our data are valid then we are only concerned with the second case.

3.2 The Forgetful Urn

This model fits our observed data well with an exception: viewers sometimes skip the most obvious object (Chapter 2.5). Treating this phenomenon rigorously complicates the modeling equations and the methods for fitting the probability parameters. Luckily, a simple approximation opens the way for an easy treatment: pretending the first marble is forgotten. Consider a sequence of marbles where the first marble has been discarded (*i.e.*, really drawn first, but considered undrawn); the marble is most likely $\text{argmax}_{j:\forall_i j \neq w_i^m} \pi_j$, the most important of the undrawn marbles. In this case, π_j will likely be large, whereas for a sequence of marbles in which the first marble is included, π_j will probably be small. Hence we can include the probability of the largest marble missing from the list, $\max_{\forall_i j \neq w_i^m} \pi_j$, in the normalization

$$p(w_n^m|w_{n-1}^m, ...w_1^m) = \frac{\pi_{w_n^m}}{(1 - \sum_{i=1}^{n-1} \pi_{w_i^m}) - \max_{\forall_{ij \neq w_i^m}} \pi_j} . \tag{3.3}$$

This results in little change when the first marble is not skipped and a mitigated impact on the probabilities when the first marble is skipped. Since we have 25 independent sequences, the likelihood of our observation is

$$p(obs) = \prod_{m=1}^{25} \prod_{n=1}^{10} \frac{\pi_{w_n^m}}{(1 - \sum_{i=1}^{n-1} \pi_{w_i^m}) - \max_{\forall_{ij \neq w_i^m}} \pi_j} . \tag{3.4}$$

To measure importance $\pi_{w_i^m}$, we maximize the log-likelihood:

$$log(p(obs)) = \sum_{m=1}^{25} \sum_{n=1}^{10} \log \pi_{w_n^m} - \log((1 - \sum_{i=1}^{n-1} \pi_{w_i^m}) - \max_{\forall_{ij \neq w_i^m}} \pi_j) . \tag{3.5}$$

We can wonder if this definition of importance makes sense for objects that may never be named first. For instance in a photo of Batman and Robin, Robin may never be named first, yet he is important. In this example, Robin violates the independent draws assumption of our model, so the model considers Robin's subordinate position in the sequence accidental. To test whether this could significantly alter our estimates of importance, we can take data from the urn model and move the second most important marble to second place every time it is drawn first. In our simulations, this change does not decrease the estimated importance of this marble (Wilcoxon rank sum test).

Optimization Note There are as many parameters as objects mentioned. This number can get large, which results in poor convergence. If we limit our optimization to the ten most frequently named objects and set the importance of all other objects to 0.001, our convergence using `fmincon` in the Matlab Optimization Toolbox (with 100 repetitions after slight agitation of adding 0.5×rand and normalizing) is reasonable (it fails to converge one time in a hundred).

3.3 The Relaxed Urn

An alternate interpretation of skipping the obvious object is that the marble probabilities linearly relax with time, reaching a uniform distribution at draw $T \geq 10$. So the probability of drawing marble i on draw n is $\pi_i(n)$, a function of draw number and the original probability. The time-varying probability $\pi(\cdot)$ is only meant where marked with an argument. On the first draw $\pi_i(1) = \pi_i$, the importance of marble i. On draw n a marble's probability *with* replacement would be

$$\pi_i(n) = \frac{1}{N}\frac{n-1}{T-1} + \pi_i \frac{T-n}{T-1}, \tag{3.6}$$

where there are N marbles. As we require values *without* replacement we plug Equation 3.6 into Equation 3.2. This normalizes for drawn marbles, yielding the probability that the marble labeled w_n^m is the nth marble drawn is

$$p(w_n^m | w_{n-1}^m, ...w_1) = \frac{\frac{n-1}{N} + (T-n)\pi_{w_n^m}}{(T-1) - \frac{(n-1)^2}{N} - (T-n)\sum_{i=1}^{n-1} \pi_{w_i^m}}. \tag{3.7}$$

Figure 3.4 shows that under this model the human data are most likely when $T \approx 50$. This indicates that the best performance is when little relaxing occurs in the ten draws.

3.4 List Statistics and Urn Models

We can compare lists generated by the plain, forgetful, and relaxed urn models with the human lists from Section 2.5. The model lists were generated by Monte Carlo simulations using an urn model and marble probabilities fit with the relevant model. The forgetful urn had a 25% chance of skipping the first marble of a list. The relaxed urn had a setting of $T = 50$ for number of draws to reach uniform.

As earlier, we have several ways to compare list composition. First, we examine a pair of lists and count how many objects the lists share. Figure 3.5 shows that intersection size from the urn models is between viewers uniform chance. The relaxed

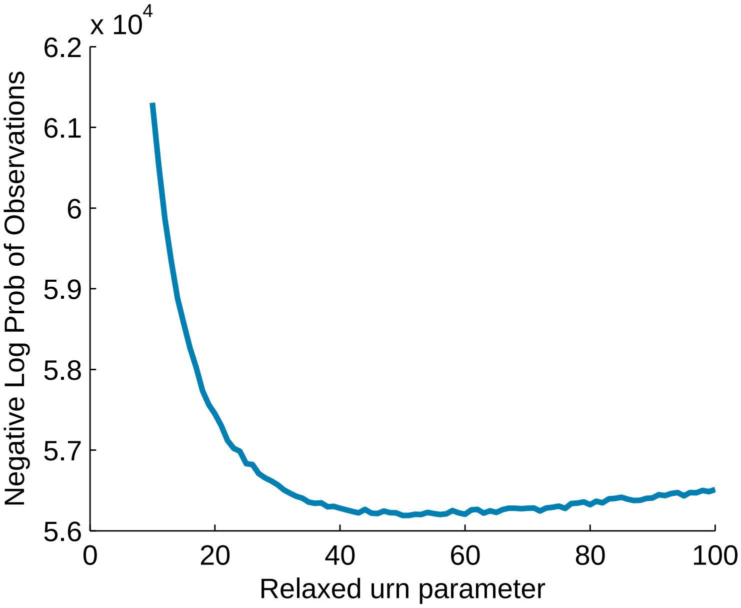


Figure 3.4: The marble probabilities of the relaxed urn relax to uniform in T draws.

urn with a setting of $T = 50$ generates the smallest intersection.

Second, we look at differences in object rank between lists. Figure 3.6 shows that an object's rank changes similarly in all three urn models, and they are all between human and chance data.

Third, we count how many objects are named by a certain number of viewers. Figure 3.7 shows that the number of viewers that name an object have similar distributions between the urn models. As with the other histogram comparisons, the urn models are all between human and chance data.

3.5 Markov Chain Method

It is also possible to approach importance estimation from a less formally motivated angle. We can use a Markov chain (MC) to calculate importance about a thousand times faster than the maximum likelihood approach, and always get a solution. A Markov chain is specified by a non-negative, stochastic transition matrix $\mathbf{M}$. The system moves from state i to state j with probability $\mathbf{M}_{ij}$. Aperiodic and irreducible Markov chains eventually reach a stationary distribution, a unique fixed point where

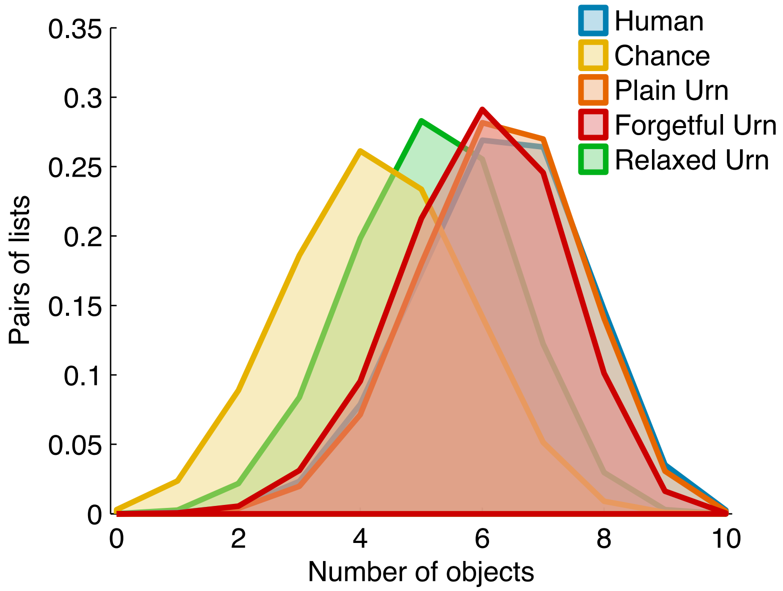


Figure 3.5: Histogram of the number of objects shared by a pair of lists for the same image. Data collected from viewers (blue) is compared with models (orange, red, green) and uniform chance (yellow).

the state distribution does not change. Conveniently, the stationary distribution is the principal left eigenvector of the transition matrix. We find the following Markov chain proposed by Dwork *et al.* [18] useful for measuring importance:

> If the current state is object i, then the next state is chosen by first picking a ranking τ uniformly from all lists $\tau_1, ..., \tau_{25}$ containing i, then picking an object uniformly from the set of all objects j such that $\tau(j) \leq \tau(i)$.

In our case $\mathbf{M}_{ij}$ is the number of lists on which object j appears earlier than or ties object i, divided by how many objects appear earlier than or tie object i on any list. The lists that do not contain object i are ignored for row i. This Markov chain is aperiodic by construction and empirically irreducible on our data, so we are guaranteed a stationary distribution. Figure 3.8 gives an example of how the Markov chain might act for the data in Figure 3.1. Our intuition as to why the stationary distribution should approximate importance is that the Markov chain is essentially running the urn backwards. So the stationary distribution is a smoothed version of the top of the lists.

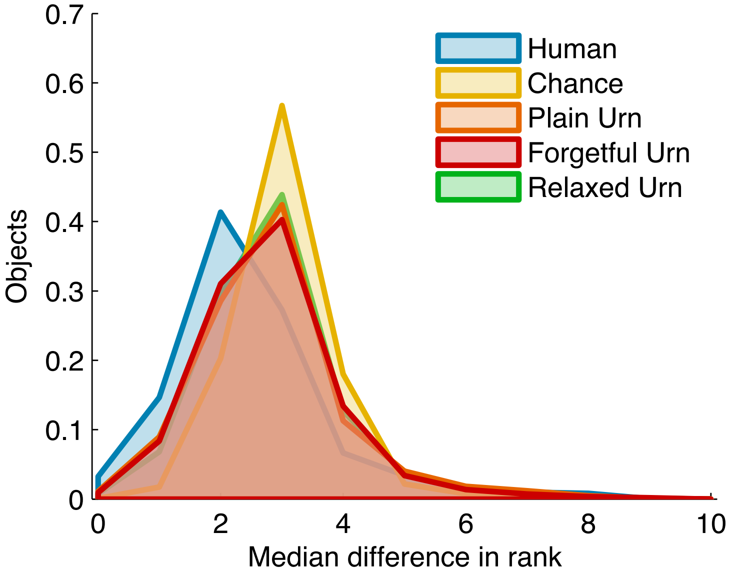


Figure 3.6: Histogram of an object's unsigned difference in the rank between two lists for the same image. Each data point is the median of these differences for an object-image combination.

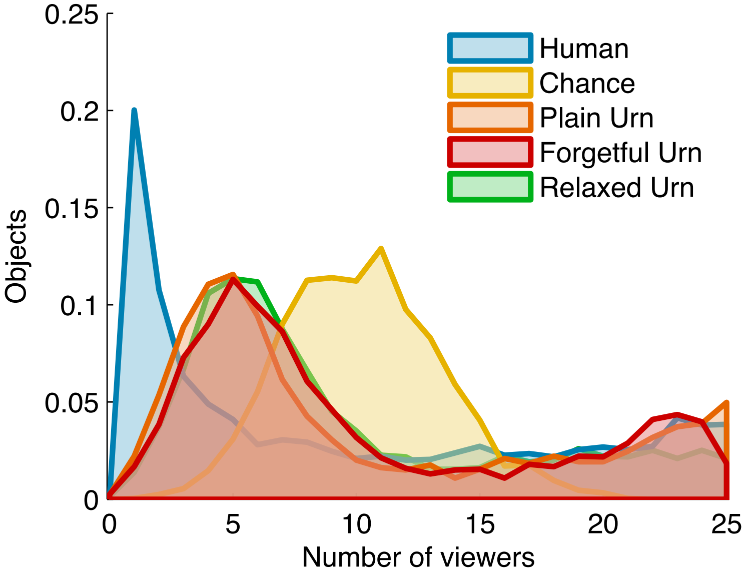


Figure 3.7: Histogram of the number of objects named by a particular number of viewers. Each data point represents an object in a specific image.

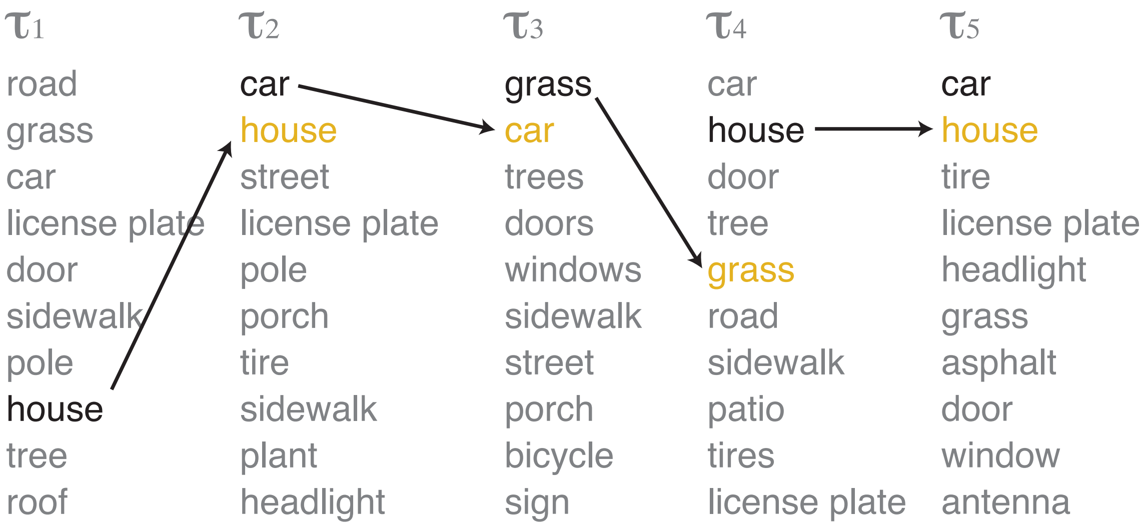


Figure 3.8: The Markov chain moves between objects selecting a list that contains the old object (arrow) and then choosing a new object (black) that was named earlier than or equal to the old object (yellow) on that list (τ). The asymptotic behavior of this Markov chain estimates importance.

Figure 3.9 compares the MC importance with the forgetful urn ML importance. The right column in Figure 3.2 shows the importance measured with the MC. The results are similar to the ML forgetful urn, except the MC slightly underestimates the importance of objects that have a true importance of ≥ 0.3 in synthetic data.

3.6 Left-out Object Sequence

One way to assess how much information about our human lists is captured by the importance values is to use 24 lists to measure importance and try to guess the left-out, 25th list. We do this by producing a most likely sequence based on the other human sequences. We use the Spearman footrule to measure the distance between two lists σ and τ, where $\sigma(i)$ is the rank assigned to object i in list σ.

$$D(\sigma, \tau) = \sum_i |\sigma(i) - \tau(i)| \tag{3.8}$$

This distance has already been applied in machine learning [18, 71] to compare

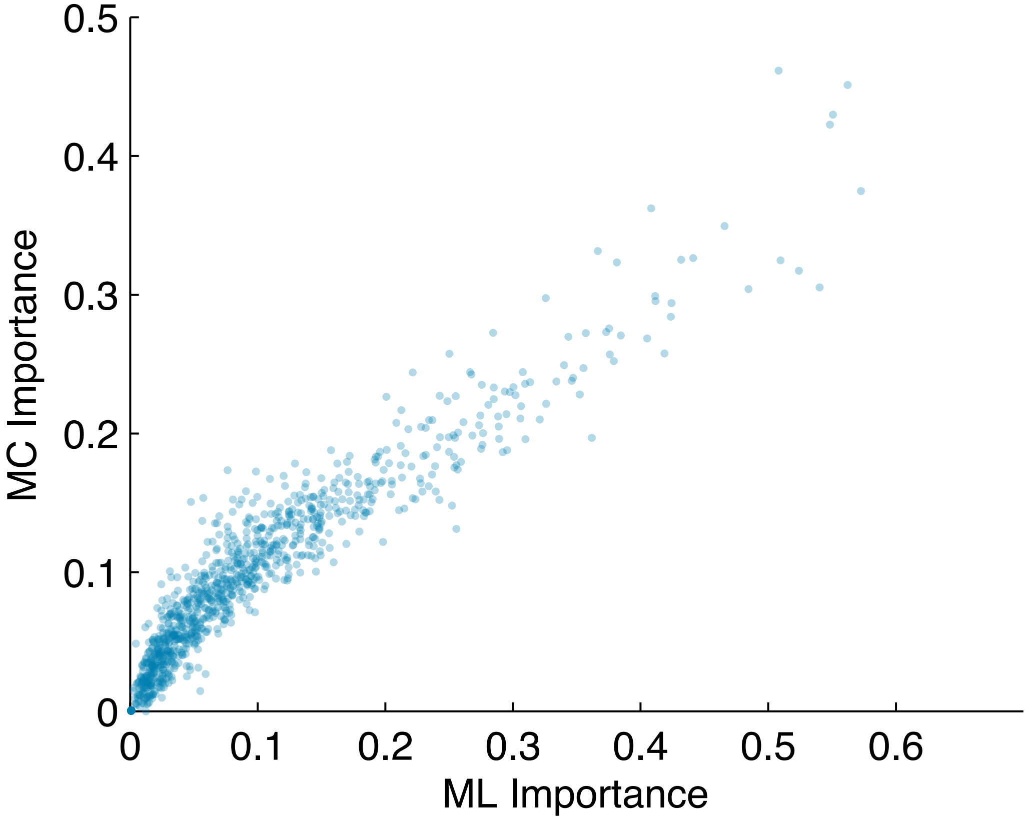


Figure 3.9: Scatter plot where a dot is plotted for each object at its forgetful urn maximum likelihood and Markov chain measured importance coordinates.

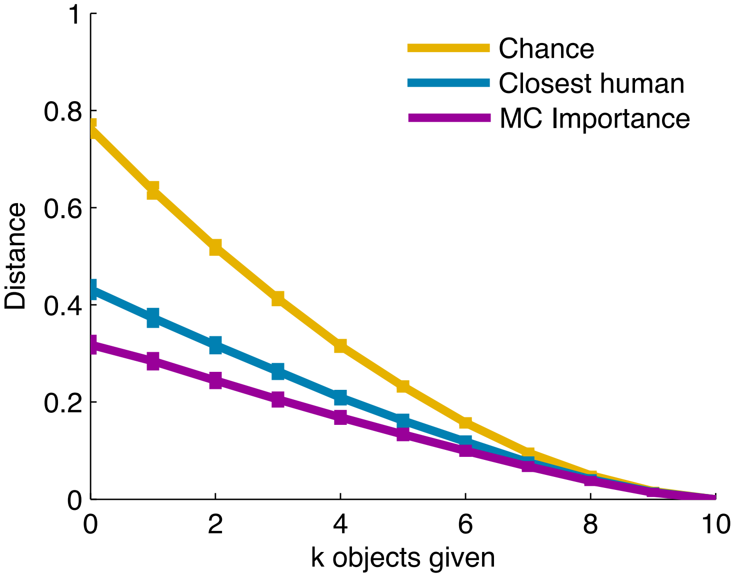


Figure 3.10: We measure the Spearman footrule distance between a left-out human list and a list generated from the other 24 human lists. To chose the closest human list, we consider the first k objects in our left-out list and choose the closest of the 24 lists. For a fair comparison, we force the first k objects in all lists to match the left-out list.

ranked lists.[2] However, since we want to penalize list pairs that share few items we need a different generalization to partial orderings than Dwork *et al.* [18] who disregard unmatched items. We do this by assigning every object missing from the list a rank of 11. This setting minimizes the variance of the distance as more objects are revealed on a list; however, other settings produce qualitatively similar results. We normalize by the maximum score attainable for each pair of lists.

We hide one of the human sequences and try to guess it using the remaining 24 sequences. We measure the performance of a given method by averaging the Spearman footrule distance between the guessed and the hidden list. Figure 3.10 shows that importance (both ML and MC methods) guesses sequences better than how one human sequence guesses another, which in turn is better than chance. Hence, the ML and MC importance estimates are a better summary of human data than another human list is.

[2] Kendall [64] states that Spearman replaced the absolute value with the square.

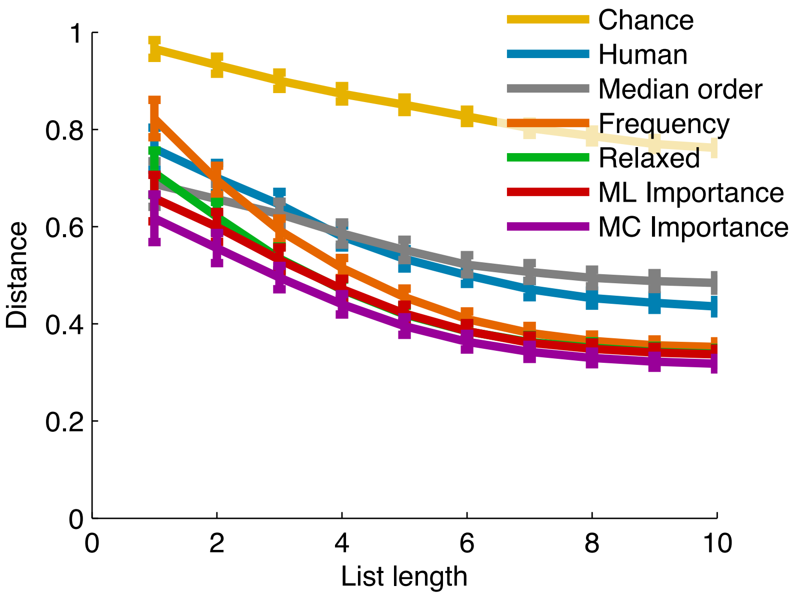


Figure 3.11: We measure the Spearman footrule distance between a left-out human list and a list generated from the other 24 human lists. We look at the distance between lists as the list length increases. Lists of length k are obtained by selecting the top k elements of each list.

We could assume that the human sequences cluster and if we select the most similar list to our held-out sequence in the first k objects named, then this would improve our results. For a fair comparison, we force the first k objects in all the guessed lists to match the hidden list and fill the other $10 - k$ entries with objects in the order of the guessed list. Figure 3.10 shows that the closest human doesn't become better than other methods as more objects are revealed, indicating that no substantial clustering exists.

One could wonder if the complexity of the ML or MC methods is justified. A simpler approach would be to estimate importance with an object's frequency or median rank across lists. We implemented such methods and compared them with the ML and MC. Figure 3.11 shows the leave-one-out guess distance as we change the list length from 1 to 10 objects. We see that median order guesses the beginning of the list better than the frequency. Importance does a good job overall.

Chapter 4

Predicting Importance

Is it possible to predict the importance of an object directly from a photograph without gathering object lists from humans? We explore a simple bottom-up approach where importance is predicted by the linear combination of many image features. We assume that in the near future there will be segmentation algorithms that can produce good object-level segmentations. Thus, we consider features that may be computed from the image once an outline of each object is available. Out of 49 possible features, we select a small subset via regularized regression to maximize both the performance and interpretability of our model.

4.1 Object Outlines

We assume that computing object importance requires that the image be segmented accurately into component objects. However, our scene photographs are large and complex, and, in our hands, segmentations produced by state of the art algorithms [41,105] are not as detailed as the verbal responses. Figure 4.1 shows that if we select the best segment for a particular object from multiple segmentations and discard objects for which a good segmentation cannot be found, most of the importance is thrown away. As a stop-gap measure, we had our images segmented by hand. We again use Mechanical Turk, but this time we ask three workers to outline all instances of a named object category in the image. Our user interface is based on flash code provided by Sorokin and Forsyth [123]. We generalize the common segmentation

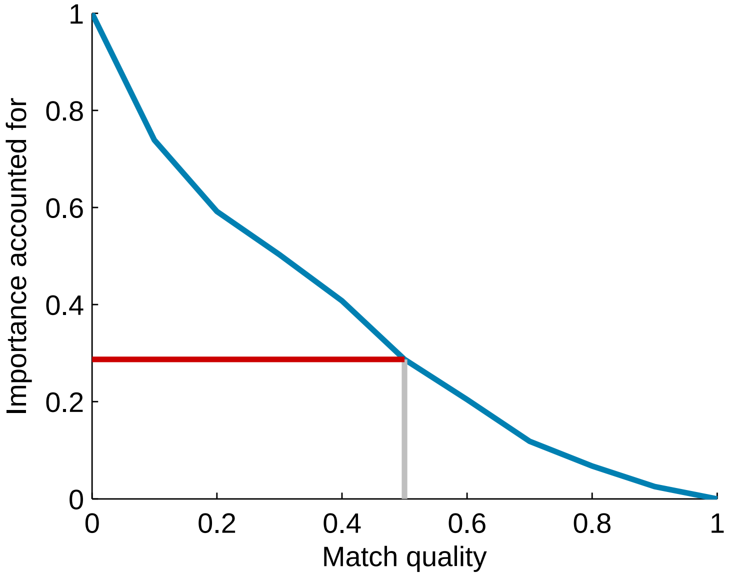


Figure 4.1: How well do state of the art segmentations match the human drawn segmentations? We measure the match quality as the intersection over union of the human and closest computer segment and then sum the importances of matched objects.

metric $|intersection|/|union|$ [32, 126] (a pixel-wise Jaccard index [60]) to evaluate the quality of these human segmentations. Our generalization of the criterion to three annotations is to compare the maximum of the three pairwise consistency values with 0.5. Outlines that do not satisfy the criterion are checked manually and rejected outlines are discarded. Pixels that are marked as the object in half or more of accepted outlines belong to the object in our final object mask. In this way, we obtain outlines for 2,841 named objects.

4.2 Features

We devise features to convey information about photographic composition. Conceivably, these features capture what makes a particular object important in a particular image. Table 4.1 is a complete list of the features used to predict importance. Figure 4.3 illustrates these features, which fall into four general categories: distances, saliency, area, and overlapping.

Table 4.1: List of all features used in importance prediction.

Feature	sum	max	mean	min
Distance to center		○	○	○
Distance left/right max		○	○	○
Distance above middle		○	○	○
Distance below middle		○	○	○
Distance 3rds		○	○	○
Distance 3rds box		○	○	○
Saliency	○	○	○	
Gaussian modulated saliency	○	○	○	
Blurred saliency	○	○	○	
Learned saliency	○	○	○	
Color conspicuity map (CM)	○	○	○	
Intensity CM	○	○	○	
Orientations CM	○	○	○	
Area				
log(area)				
Area order descending				
Area order ascending				
Percent overlapped				
Number of overlapping objects		○	○	
Percent of face covered by object				
Percent of object covered by face				
Object-face intersection/union				

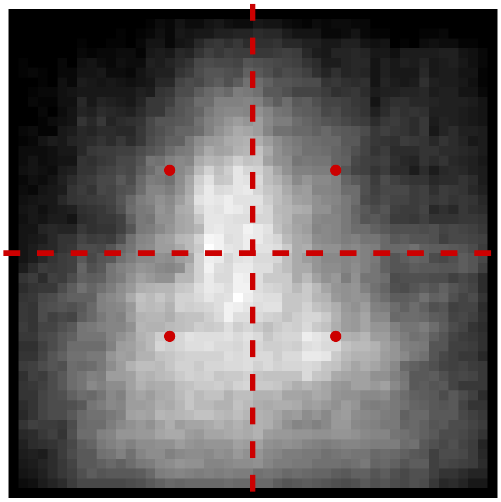


Figure 4.2: Density of named objects. If we look at the mean number of objects per image covering a particular pixel (photos resized to 50 × 50) we notice that the distribution is higher in the central third of the image. Furthermore, it is left-right symmetric, but not top-bottom symmetric. There appears to be a wider horizontal patch approximately one-third of the way from the bottom.

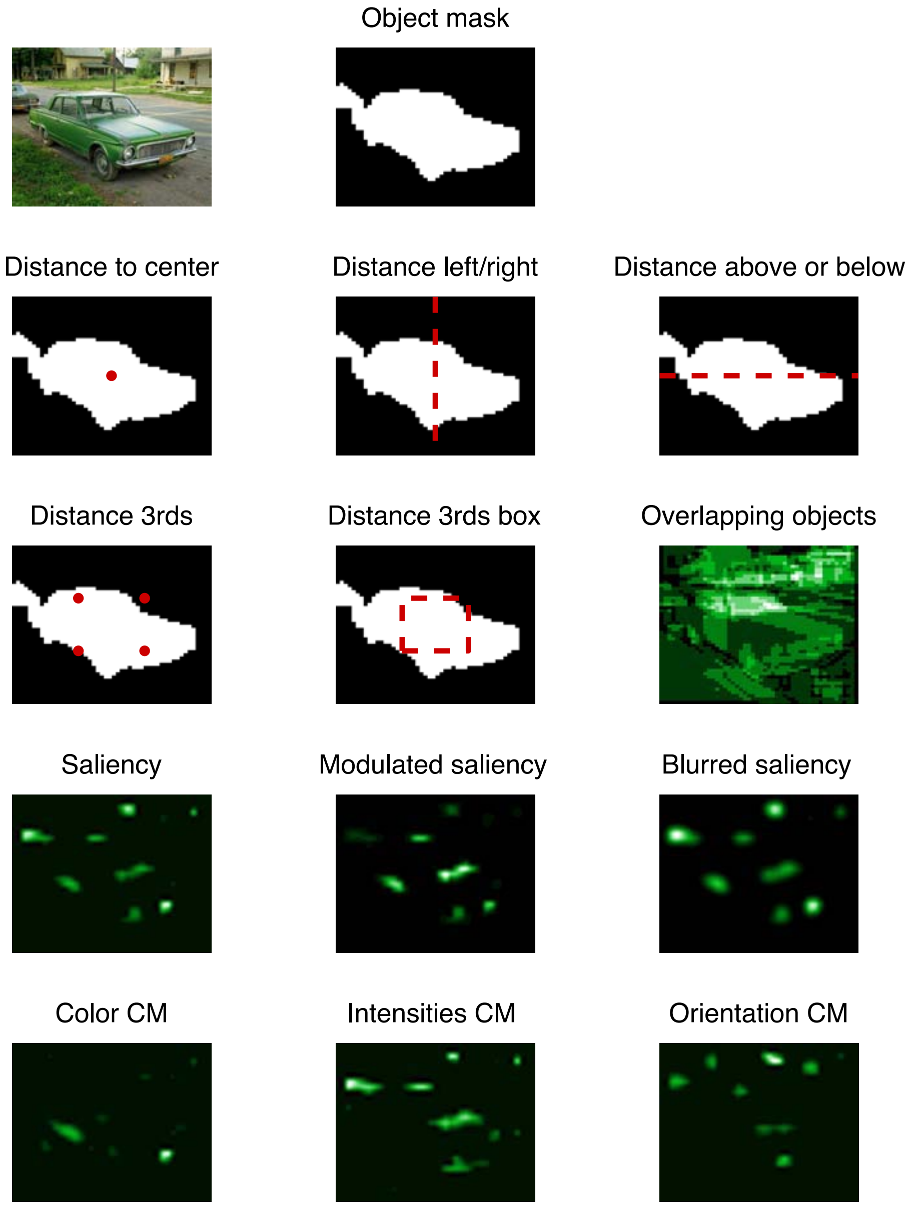


Figure 4.3: 1st row: a photograph and car object mask. 2nd row: Distances relating to center. 3rd row: Distances relating to the rule of thirds. Number of overlapping objects per pixel. 4th row: saliency map and modifications. 5th row: Conspicuity Maps.

Distances The central bias indicates that an object's position in an image influences attention [84, 127, 129]. Figure 4.2 shows the distribution of objects over a photo. We sum all object masks (pixels are 1 if they contain the object, 0 otherwise) for all images, creating an object map [27]. We notice that the object map has a vertical symmetry axis, so we treat distances to the left and right of the midline equivalently. However, the object map has no horizontal symmetry axis, so distances up and down are handled independently. We measure distances from the object mask to important positions in the image: distances to center, left/right of the vertical midline, above the horizontal midline, below the horizontal midline, to the four points that divide the image into thirds, and to the box defined by the four points that divide the image into thirds. For all distance measures we calculate the maximum, mean, and minimum distance between pixels in the object mask and the position in question.

Saliency We use a saliency map [59], a computational approach to describe how low-level features drive human eyes movements as a way to track the allocation of attention. Specifically the algorithm looks for regions that are conspicuous (or different from neighboring regions) in terms of color, intensity, or orientation, and then combines the conspicuity maps (CM) of these three channels. We use a publicly available implementation [146] to produce saliency and conspicuity maps. We use the color CM, intensity CM, orientations CM, as well as the saliency map. We introduce two modified versions of the saliency map: a blurred saliency map, which is convolved with a 5×5 Gaussian window, and a Gaussian modulated saliency map, which is multiplied by a Gaussian window ($\sigma = 0.4$) to create a central bias. For each of these measures, we took the sum, max, and mean of the saliency values covered by the object mask.

A recent approach uses fixations from an image set to create an optimal linear combination of color CM, intensities CM, orientations CM, and Viola and Jones face mask [142, 150]. We refer to this as learned saliency. Fixation data from 67 Shore images (Chapter 5) different from those used to measure object importance enabled us to fit $w \geq 0$ with constrained least squares. Vectorized color, intensities, and orientations CMs $\mathbf{c}$, $\mathbf{i}$, and $\mathbf{o}$ and a vectorized Viola and Jones face mask $\mathbf{f}$ predict a 2° std Gaussian smoothed fixation map $\mathbf{fix}$.

$$\arg\min_{w} ||[\mathbf{c}\ \mathbf{i}\ \mathbf{o}\ \mathbf{f}] \times w - \mathbf{fix}||^2 \tag{4.1}$$

On our images the learned saliency weights w were 0.02, 0.01, 0.0056, and 0.0001.

Area A larger object has more of a chance to be fixated randomly than a small object, so area is a natural feature. We use an object's area, log(area), and ascending and descending rank in terms of area.

Overlapping Parts might conceivably be less important than whole objects. How an object is overlapped indicates that it might be a part, so we include several related features: the percent of the object that is overlapped by other outlined objects and how many objects overlap it pixel-wise. We also run a Viola and Jones face detector and take the output to be a mask of all faces in the image [142]. We then look at the percent of the face mask that is covered by the object, the percent of the object covered by the face mask, and the intersection over union of the object and face masks.

4.3 Regression

We approximate the function from features to importance as:

$$\log(importance) = \beta_0 + \sum_{j=1}^{p}(x_j\beta_j) \tag{4.2}$$

where x_j is the value of the jth feature for an object and β_j is the coefficient of that feature.

Our two goals are maximizing prediction and interpretation; we don't want to overfit our data and we want to know which are the *useful* features. Limiting the magnitude of the βs (excluding β_0), called regularization or coefficient shrinkage, is a one popular way to improve prediction. The Lasso $\sum_{j=1}^{p} |\beta_j| \leq t$ specifically favors sparsity, additionally increasing interpretability [133]. We use a 1,455-object (50 image) training set and a 354-object (12 image) validation set to select the simplest Lasso model within one standard deviation of the lowest residual sum of squares (RSS) on the validation set. To compare β magnitudes, we standardize data to have mean 0 and standard deviation 1 before performing the Lasso [47]. We use RSS for validation only, not for test set evaluation. We do not use the footrule distance for evaluating predicted importance, because we have measured importance as our ground truth instead of human-generated object lists.

Figure 4.7 shows the Lasso chosen features and their coefficients. The only 17, of 49 features, with non-zero coefficients are log of area and ascending/descending rank of area, mean number of overlapping objects per pixel and percent of object overlapped by pixel, the intersection/union of object and face mask, percent of object covered by face, mean distance to the left or right of midline, maximum distance below the midline, minimum distance from the object to the box defined by the points that divide the image into thirds, sum of Orientations and Color CMs across object, maximum Color CM on the object, mean Orientations CM across object, sum of Gaussian modulated saliency. Plain saliency measures are not selected when a centrally biased version and CMs are available. Area is not selected when log(area) is available.

Figure 4.4A shows the quality of importance prediction on a 1,032-object (35 image) test set. We define an *important* object as having a measured importance $\geq \{0.05, 0.15, 0.25, 0.35\}$ and move the threshold across the predicted importance. These importance values correspond to the top six objects per image, two objects per image, one object per image, and one object every three images. We find that our prediction identifies high-importance objects reasonably well; the areas under the ROC curves are 0.7, 0.77, 0.81, and 0.89. This is in comparison with maximum learned saliency alone (Figure 4.4B). The corresponding areas under the ROC curves are 0.57, 0.72, 0.75, and 0.79.

Figure 4.5 shows a scatter plot of the measured importance and normalized pre-

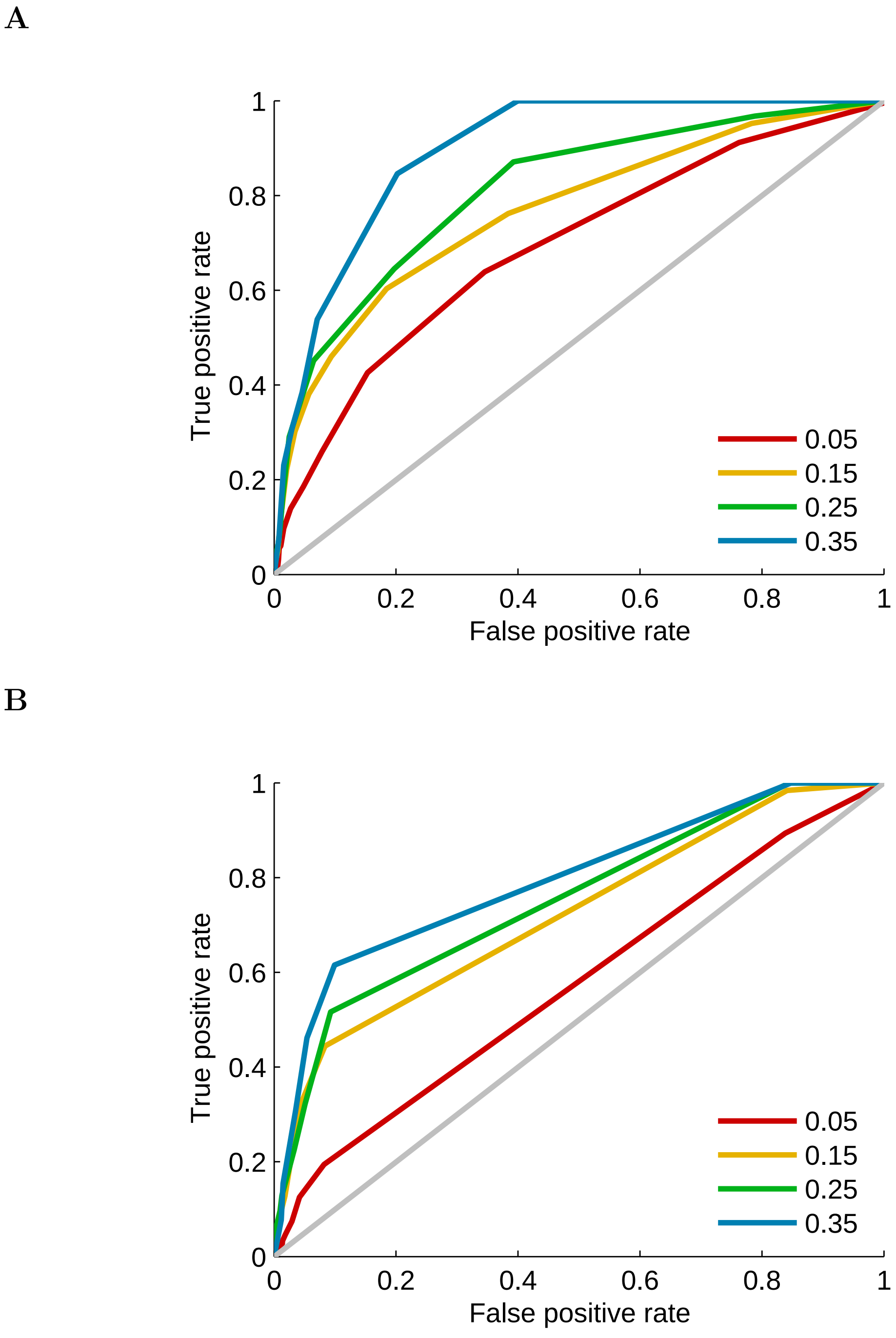


Figure 4.4: ROC curves for identifying important objects. We define an *important* object as having a measured importance $\geq \{0.05, 0.15, 0.25, 0.35\}$ and move the threshold across the predicted importance. **A** full model, **B** maximum learned saliency alone.

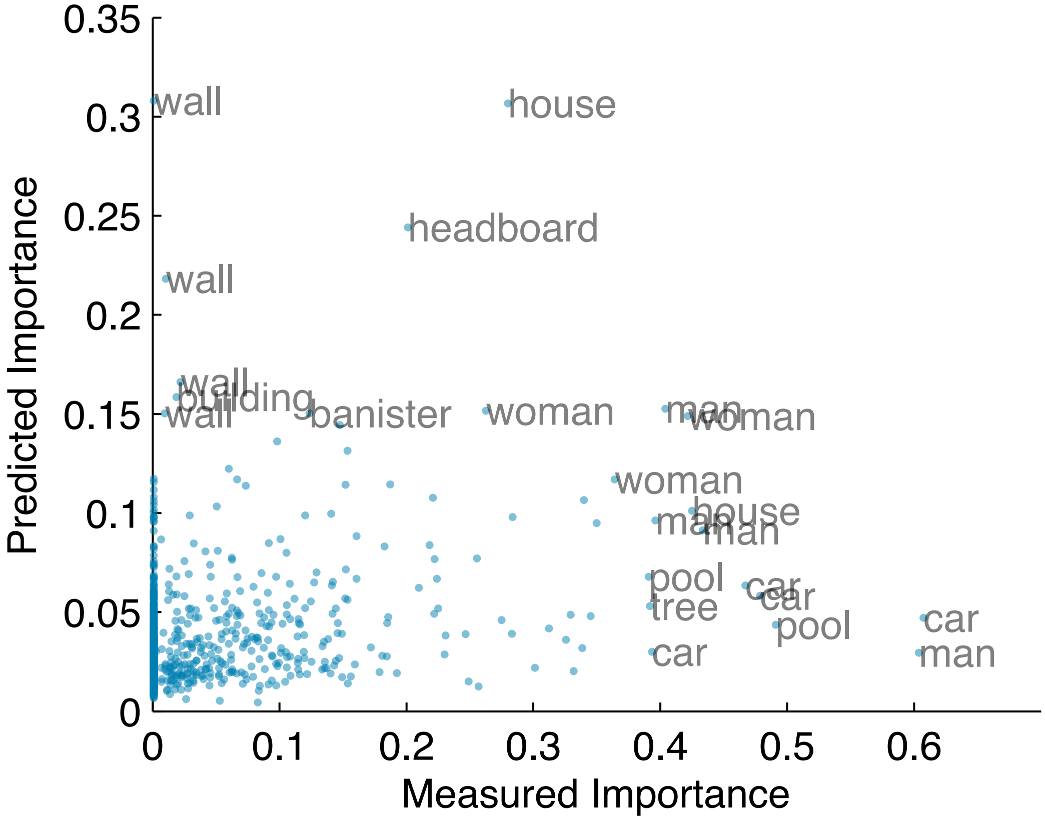


Figure 4.5: Scatter plot of predicted versus measured importance. Most objects have very low importances.

dicted importance (Pearson's correlation coefficient of 0.39). However, the scatter plot is difficult to interpret because most of the objects have very low importances. Figure 4.6 shows a few examples of our results; predicted importances are normalized so the importance in an image sums to 1.

4.4 The Power of Features

One question we can ask is, if we eliminate the largest valued features, does the prediction collapse? Actually, the RSS gracefully changes from 1,340 to 1,348 to 1,355 to 1,357 as we exclude the three features with the largest magnitude in Lasso. Figure 4.8 demonstrates that as one feature is excluded, another feature arises to

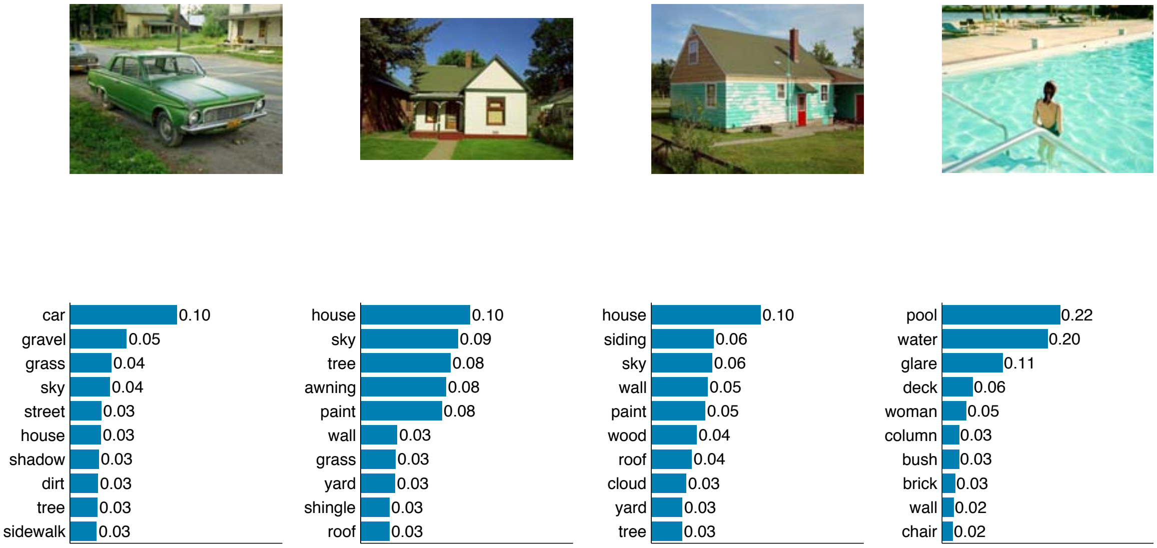


Figure 4.6: Predicted Importance. Importance predicted using the Lasso and simple image features. Notice that in the fourth image, pool and water are almost completely coincident, hence their importance estimate is almost identical. Our subjects consider water less important, and only a semantic analysis of the scene may resolve this issue.

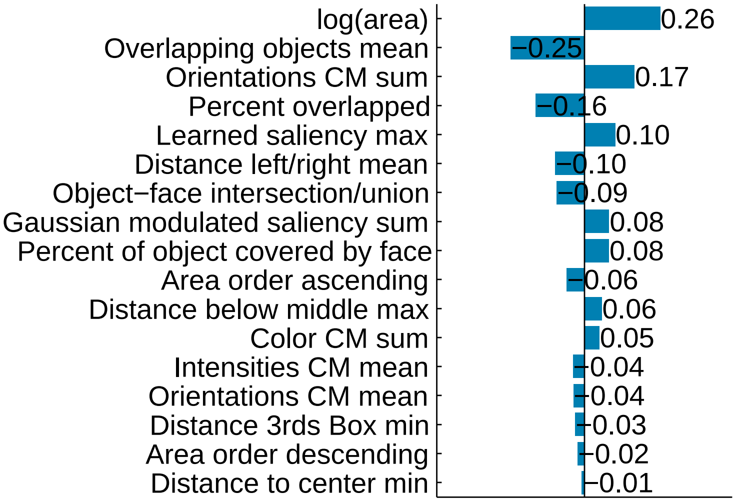


Figure 4.7: Lasso chosen features and their coefficients at $t/\sum_{j=1}^{p}|\hat{\beta}_j| = 0.14$.

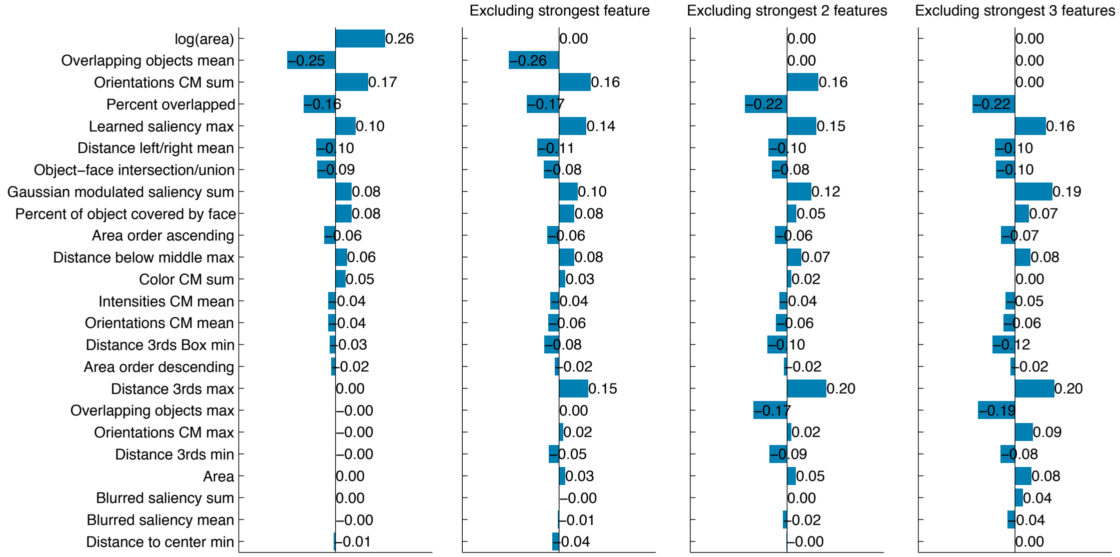


Figure 4.8: Excluding the features with the largest coefficients simply causes other features to replace them. The residual sum of squares is minimally affected.

replace it, indicating that our features are redundant.

Another question is how well a single feature, or only a few, can predict importance. Figure 4.9 shows that, adding features greedily with stepwise regression, a few features go a long way.

4.5 Generative and Discriminative Tasks

Earlier we considered the case that the viewer is asked for ten objects, but not told what exactly will be done with labels. We call this the Plain task:

> Please look carefully at this image and name 10 objects that you see.

Alternatively, we can give the viewer the Generative task:

> Name 10 objects in this image. Someone will use these words as search words to find similar images.

Or the Discriminative Task:

> Name 10 objects in this image. Focus on what distinguishes this image from similar-looking ones.

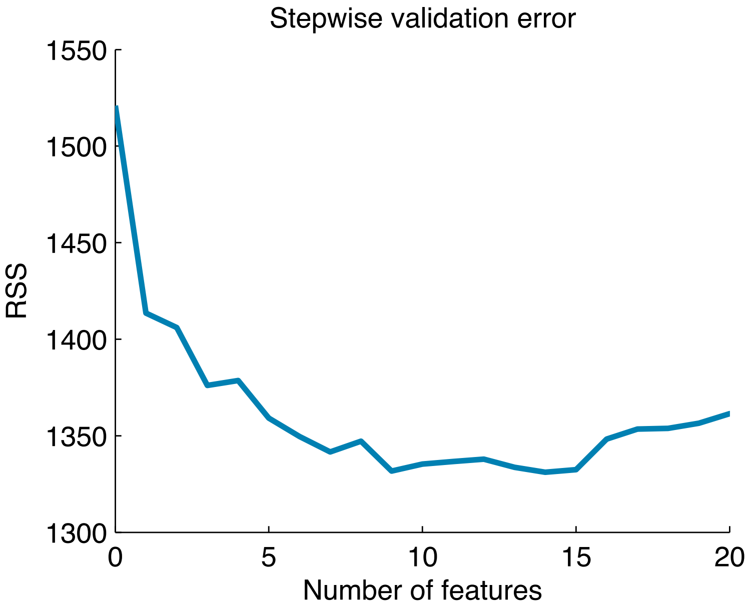


Figure 4.9: RSS error as features are included by stepwise regression.

The measured importance values in Figure 4.11 compare the lists obtained from viewers performing either the Plain, Generative, or Discriminative tasks. Figure 4.10 compares the feature coefficients for predicted importance. The values are similar for the Plain and Generative tasks when generated by the Lasso or stepwise regression. The Discriminative task produces different results from the other tasks with both methods. The most noticeable difference is that more weight is given to distance left/right. The overall differences are small, which tells us that viewers are performing a stable task.

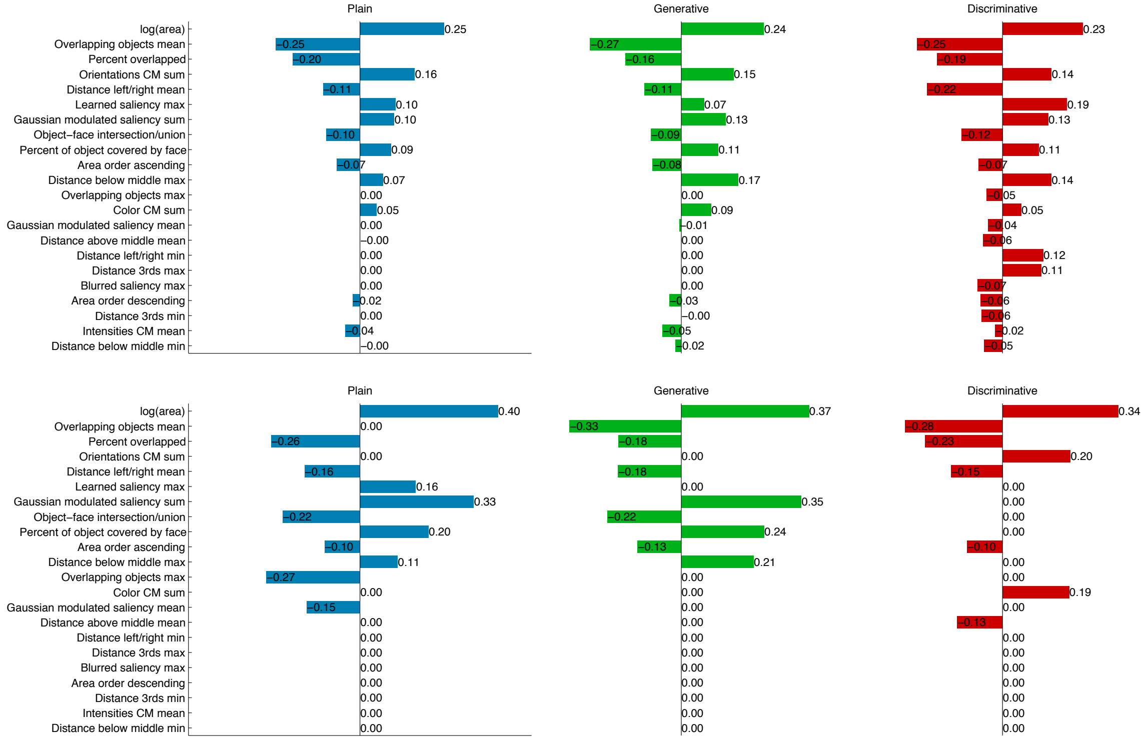


Figure 4.10: Coefficients for importance prediction. Data have been normalized so that coefficient magnitudes represent relative contribution. The values are similar for the Plain and Generative tasks when generated by the Lasso (top) or stepwise regression (bottom).

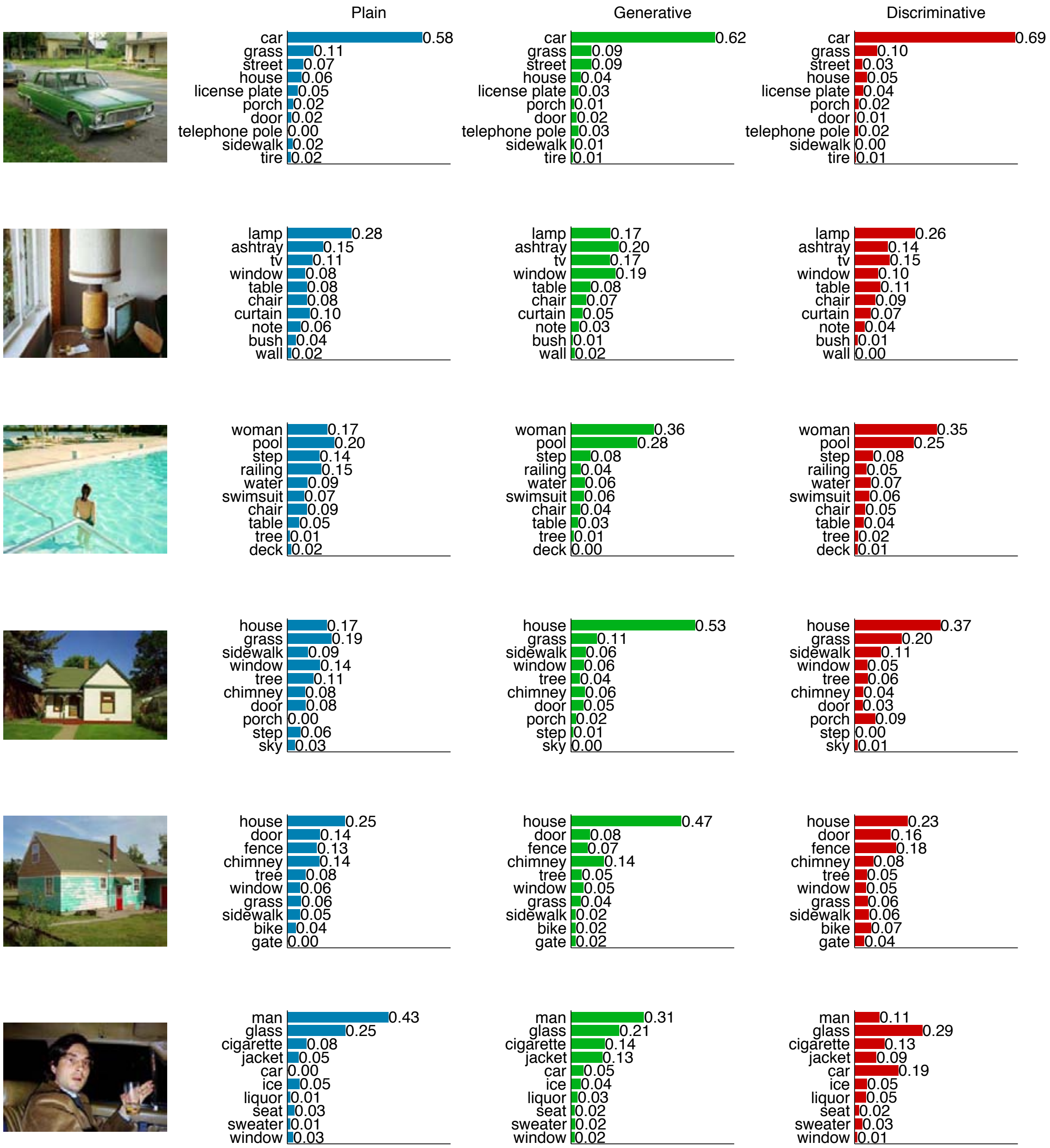


Figure 4.11: Measured importance for the Plain, Generative, and Discriminative tasks. The fact that these estimates are comparable despite different instructions suggest that our subjects are performing a stable and natural task.

Conclusion

We introduced the concept of object importance and showed how to estimate it once a high-quality object segmentation is available. Our estimator works without object identity: we can often know that an object is important without knowing what it is.

To study how humans perceive object importance, we asked a large group of English-speaking observers to name objects they saw in photographs of everyday scenes. For each of 97 images, we collected 25 independent ten-word lists. This data set allowed us to observe that objects are named quasi-independently. Thus, the process of naming objects in images is akin to drawing marbles from an urn without replacement. Furthermore, some objects tend to be named earlier and more frequently, which we represent as the marbles having different diameters, and thus different probabilities of being drawn. The urn model suggests that an object's importance should be defined as the *probability of being named first*. The urn model allowed us to estimate object importance using maximum likelihood applied to the word lists. We obtained similar results with a Markov chain approach.

We then turned to the question of whether it is possible to predict the importance of an object directly from an image. We used a simple regression model predicting importance from features that are measurable in the image. A side product of our Lasso regression was a ranking of how informative different object-related image features were for predicting importance. While position and size were quite useful, a plain saliency measure did not rank among the top features. We found that this bottom-up prediction will often select the most important objects in an image. However, information about the meaning of the scene may be necessary for perfect prediction.

An unexpected phenomenon we observed was that our viewers sometimes failed to report the most obvious object in their ten-word list. This was very repeatable and had not been previously explored. Our urn model was easily modified to accommodate this phenomenon.

Part II

Eye Movements

Chapter 5

Methods

5.1 Stimuli

The stimuli were 93 photographs from the artist S. Shore's collection "Uncommon Places" [121] (shown in Figure 5.1). The images were collected as a visual diary and come across as casual snapshots of everyday scenes. The images were presented on 20 inch CRT monitor, located in a dark room at 80 cm from the observer, and thus subtended 29×22 degrees of visual angle (°). The artist provided digitized high-resolution images. To fit the resolution and the aspect ratio of our presentation screen (1024×768 pixels), images were down-sampled and minimally cropped.

5.2 Experimental Conditions

We tested two experimental conditions, called "what" and "where." In both conditions, we instructed our observers to imagine that they were a judge for an art competition and to rate, on a scale from 1 to 5, how interesting each image was. Asking our observers to rate the images ensured careful observation. We ignored the ratings in our analysis. In both conditions, observers were asked to provide some (up to five) keywords to describe the scene. To avoid confounding the eye tracking data, the keywords were typed after the stimulus disappeared. In the "where" condition, observers additionally searched for a *target object*, an object that was specified in writing on the screen before image presentation. Observers were asked to decide as

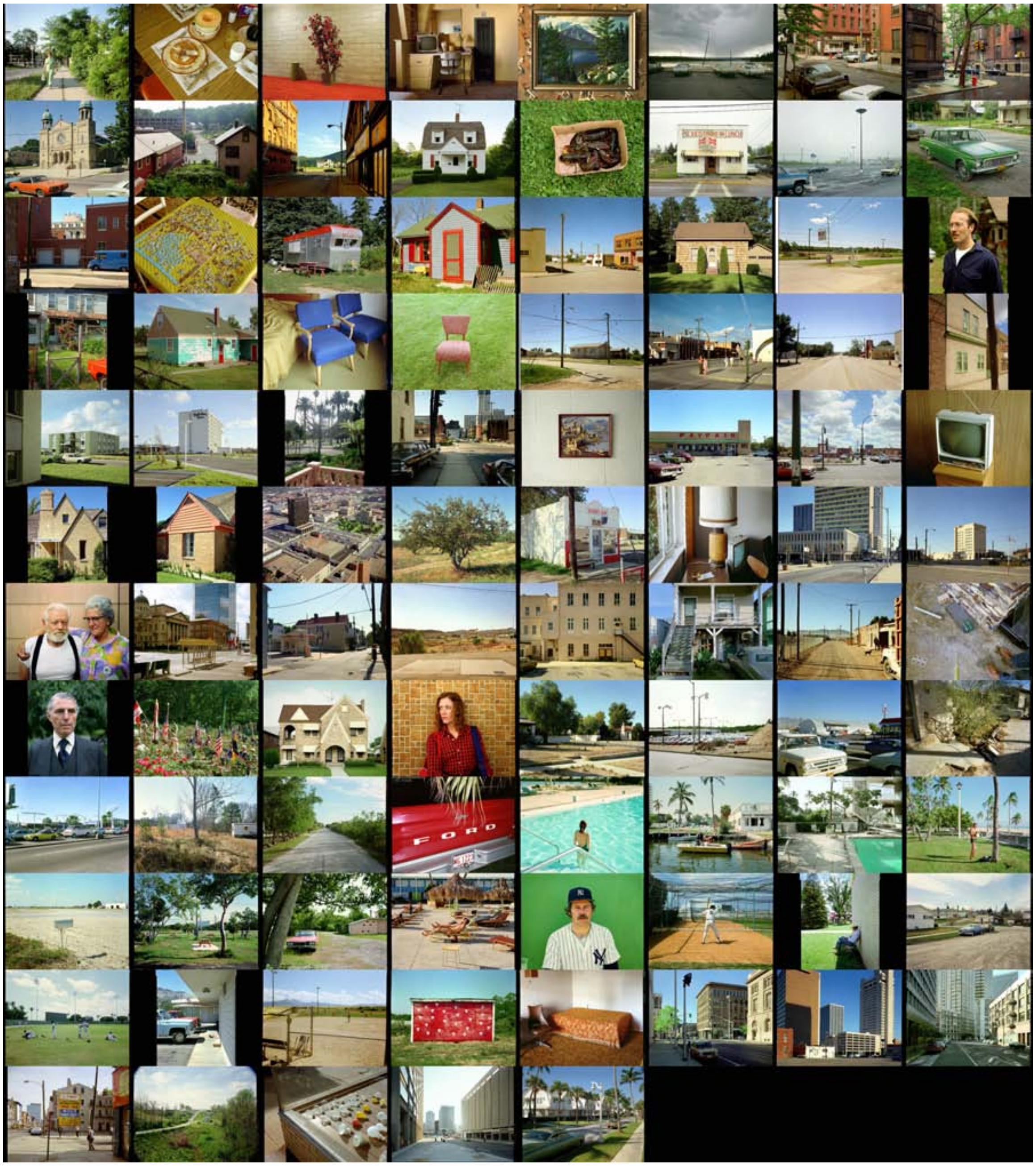


Figure 5.1: Ninety-three photographs of Stephen Shore's collection "Uncommon Places" were used as stimuli.

quickly as possible whether the target object was present in the scene. Target object selection made search difficult; target objects were either present but not obvious, or not present in the image but plausible for the scene and frequently named in other images (as established independently in an internet-based "what" condition without eye tracking). In the "what" condition an image was displayed for 3 seconds. In the "where" condition an image disappeared when observers pressed a key indicating target object presence or absence. After the disappearance of the image the observer rated its interestingness from 1 to 5, and then typed up to five keywords (Figure 5.2A).

5.3 Observers

Eight volunteers (six male, two female; mean age: 23) from the Caltech community participated for pay, four in each condition. All participants were native English speakers, had normal or corrected-to-normal vision, and normal color vision as assessed by Ishihara plates. None of the participants had any formal art training. All were naive to the experiment's purpose and had not previously seen the stimuli. All procedures conformed to national and institutional guidelines for experiments with human subjects and to the Declaration of Helsinki.

5.4 Recording Eye Position

Throughout the experiment a noninvasive infrared Eyelink-1000 (SR Research, Osgoode, ON, Canada) system monitored eye position at a 1000Hz sampling rate. Our analysis used only data recorded during stimulus presentation (Figure 5.2B). Chin and forehead rests stabilized observers' heads. The calibration of the eye tracker's gain was validated after each 10 trials and recalibrated as necessary. Linear drift of the eye tracker was controlled for before each trial onset and corrected when needed. The average validation error in a 13-point validation procedure was $0.56° \pm 0.10°$ (mean ± sd over subjects). This error is on the order of the saliency maps' resolution

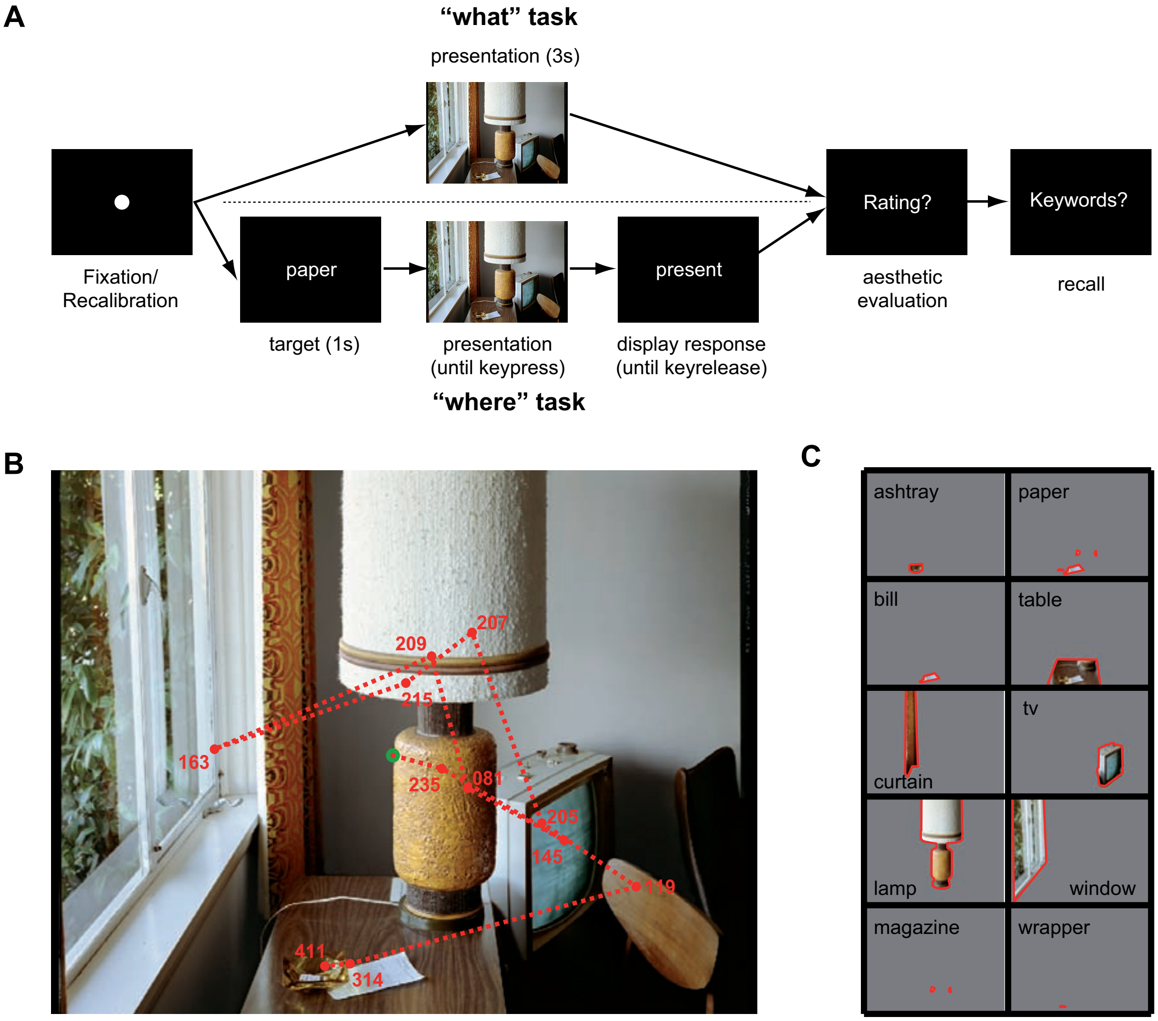


Figure 5.2: **A** Paradigm outline. In the "what" condition (top), observers see the image for 3s, are prompted for a rating and then for keywords. In the "where" condition (bottom) observers terminate presentation by deciding on the presence or absence of an item (search target) presented verbally at trial onset. **B** Example images superimposed with fixations of a single observer (MW). Numbers at fixations denote fixation duration in milliseconds. **C** Outlines of all named objects for images of panel B.

(1/16th of the image resolution, *i.e.*, 0.5°/bin) and smaller than the typical object size, which we coarsely estimate by the square root of the number of pixels covered by an object, yielding 223 pixels or 6.3° on average. Thresholds to detect saccades were set to a velocity of 35°/s and an acceleration of $9500°/\mathrm{s}^2$ as recommended by the manufacturer for the Eyelink-1000 device. There was no minimum duration for a fixation set, but 99.4% of the 7,318 fixations lasted longer than 50 ms and 97.6% longer than 100 ms (median: 251 ms; mean: 311 ms). The location of a fixation was defined as the mean eye position during this fixation. The maximum horizontal distance covered by the eye during a fixation was below 0.5° in 79.0% of cases, below 1° for 98.0% of fixations, for the vertical direction these values were 80.5% and 96.7%, respectively (mean: 0.37° and 0.38°; median: 0.32° and 0.31°). The standard deviation of a fixated location was on average 0.08° both in horizontal and in vertical direction. The typical variation of fixated location during a fixation is thus small compared to the absolute location accuracy of the eye tracker, the resolution of saliency maps and the typical size of objects. Presentation of stimuli, recording of eye position and analysis were implemented in Matlab using its psychophysics and eyelink toolbox extensions [5, 11, 98].

5.5 Object Annotation

For consistency of the main analysis, the authors marked the outlines of the objects named by the observers (Figure 5.2C). For analysis, we excluded terms describing the full image, objects not present, words other than concrete nouns, and repetitions (but counted them in the object naming order). Obvious synonyms were treated as the same object. The image annotation was blind with respect to the fixations, that is, only the keywords and the images were used during synonym determination and object outlining.

To obtain an independent set of labels, an additional observer outlined “all objects” in a subset of images. Since “all objects” is an ill-defined stopping criterion, we motivated this observer to label as many objects as possible by making payment proportional to the number of labeled objects plus a bonus for objects that occur in multiple instances in an image ($0.05 per object and $0.01 per additional instance).

5.6 Object Maps, Fixation Maps, and Saliency Maps

Our definition of a *saliency map* follows the model of Itti and Koch [58], with the authors' original implementation and parameters (http://ilab.usc.edu). The computed saliency map has a lower resolution than the original image i, and is linearly scaled up to image resolution to obtain the saliency map $S_i(x, y)$. Analogously, we define an *object map* $O_i(x, y)$: for each observer we count the number of objects overlapping with pixel (x,y) in image i. Then we sum these maps of all observers to obtain a single map for each image i. Note that in this default definition O_i depends on the frequency of recall; an object recalled by all observers is weighted 8 times stronger than an object named once. The term *object map* refers to this observer-weighted definition, unless stated otherwise. In addition, we consider *unweighted object maps* that count the number of objects overlapping with a given pixel irrespective of the number of observers recalling an object. Both maps are normalized to maximum 1 to ease comparison without affecting the relative ranking of pixels in each map. To test the consistency of observers' fixations, we define a *fixation map*: we assign each fixation to the nearest pixel and label the respective pixel as fixated. Due to the high resolution of the image, overlap between two fixations on the pixel level can be neglected and we obtain a binary map, with entry 1 for fixated pixels and 0 otherwise. This map is then smoothed with a $1°$ Gaussian kernel to obtain the fixation map. Figures 6.1 and 8.1 depict examples of object maps and saliency maps, Figure 6.3 shows an example of a fixation map.

We define the *total object saliency* of an object as the sum of saliency map values over the object's footprint divided by the sum across the whole image. Since this measure scales with object area, but the area cannot be factored out easily due to the sparseness of saliency maps, we consider an additional measure. We define *maximum object saliency* as the maximum saliency map value inside the object's outline. As the features of the saliency map are computed early in the visual hierarchy, we will refer to the saliency map values at a given location as *early saliency*.

5.7 Analysis Methods

Predicting fixations We compute how well the aforementioned maps predict fixations with a method proposed by Tatler *et al.* [127]. Each map pixel is either labeled 1 or 0 for fixated or non-fixated. We then computed the *hit rate* as the fraction of fixated pixels, where the map scored above a threshold, and the *false alarm rate* as the fraction of non-fixated pixels where the saliency map scored above the same threshold. We plotted hits versus false alarms while varying the threshold to obtain an ROC curve. The area under the ROC curve (AUC) measures a map's fixation prediction. Although other measures of fixation prediction have been proposed in the context of saliency maps (*e.g.*, normalized scanpath saliency [99]), our signal detection measure is invariant to monotonic scaling of maps. This is especially valuable for comparing the predictions of different maps.

Predicting recall The prediction of object recall cannot be tested directly, since objects that are not recalled by any observer remain unknown. Instead, we tested how fixated locations discriminate between *idiosyncratic objects*, or objects recalled by one observer, from objects recalled by multiple observers. We labeled the objects by how many observers recalled them, $l(o) = 1$ for idiosyncratic objects, $l(o) = 2+$ for objects recalled by two observers or more, $l(o) = 3+, ...l(o) = 8$. The fraction of fixations inside each object, pooled over all observers is used as measure $f(o)$. The objects with $f(o)$ above a threshold were false alarms if labeled 1 and hits if labeled n+. By varying the threshold we obtain an ROC, with the area under the curve (AUC) as a summary statistic. This AUC measures how fixations discriminate between objects recalled once from objects recalled twice or more. We performed the same analysis for objects recalled n times or more as compared to recalled once (objects recalled more than once but fewer than n times were excluded for this analysis). With the same analysis, we tested the recall prediction of the time of fixations on the objects, object area, length of the object's boundary, object saliency, and linear combinations of these measures.

Random reassignment baseline Fixation patterns are driven by both images and image independent spatial biases [129]. The *central bias*, the tendency to look straight ahead in head-fixed settings, is well-known (Figure 6.2 C and D). This bias predicts that centrally placed objects would be fixated more often in our experiments. If photographers place important objects centrally, then a double spatial bias would link important objects and fixation. Analogously, the relation between luminance-contrast and fixation partly results from such a double spatial bias [84, 127, 129].

We follow two strategies to assess the effect of these spatial biases: First, we directly measure the spatial biases of the feature under investigation (Figure 6.1 A and B). Second, we define a random reassignment baseline to measure how much of the prediction by a certain map can be explained by its image-independent spatial biases: We randomly reassign the object, saliency, and fixation map of one image to another image. Simultaneously we keep the fixations and object recall with the original image. On these surrogate data, analysis is performed identically to the actual data. Any effects arising from general biases in the feature are also reflected in this baseline, while any effects beyond the baseline are image-specific.

Chapter 6

Inter-Observer Consistency

6.1 Central Bias

In previous studies, fixation prediction could often be partly attributed to a double central bias [129]: Human observers tend to look straight ahead and images taken by human photographers tend to be centered on salient objects. We verified the photographer's bias in our sample, images by Stephen Shore, considering all 981 objects that were labeled by at least one of the eight observers. We define the center of an object as the center of mass of all its pixels. Half of the objects have their center in a circle of 6.1° radius around the image center, compared to the image width of 29°. That is, 50% of object centers fall within a central circle whose size constitutes 18.8% of the image area. This central bias occurs primarily in the horizontal direction: half of the objects are closer than ± 2.9° to the vertical midline of the image, a rectangle that corresponds to 20.2% of the image area. A similar result is observed when replacing the object's center its entire footprint, represented in the object maps: The average over all object maps exhibits its maximum horizontally in the image center, while the vertical peak is below the midline (Figure 6.1A). Hence there is a spatial bias on object location. Note that the spatial bias is enhanced by the facts that in artistic western photography objects are rarely cut off at image boundaries (and if so, pixels outside the image would be ignored), and that objects that span large parts of the scene necessarily have their center of mass close to the image center. Since the present study does not aim to understand the origin of this bias, it is considered a property

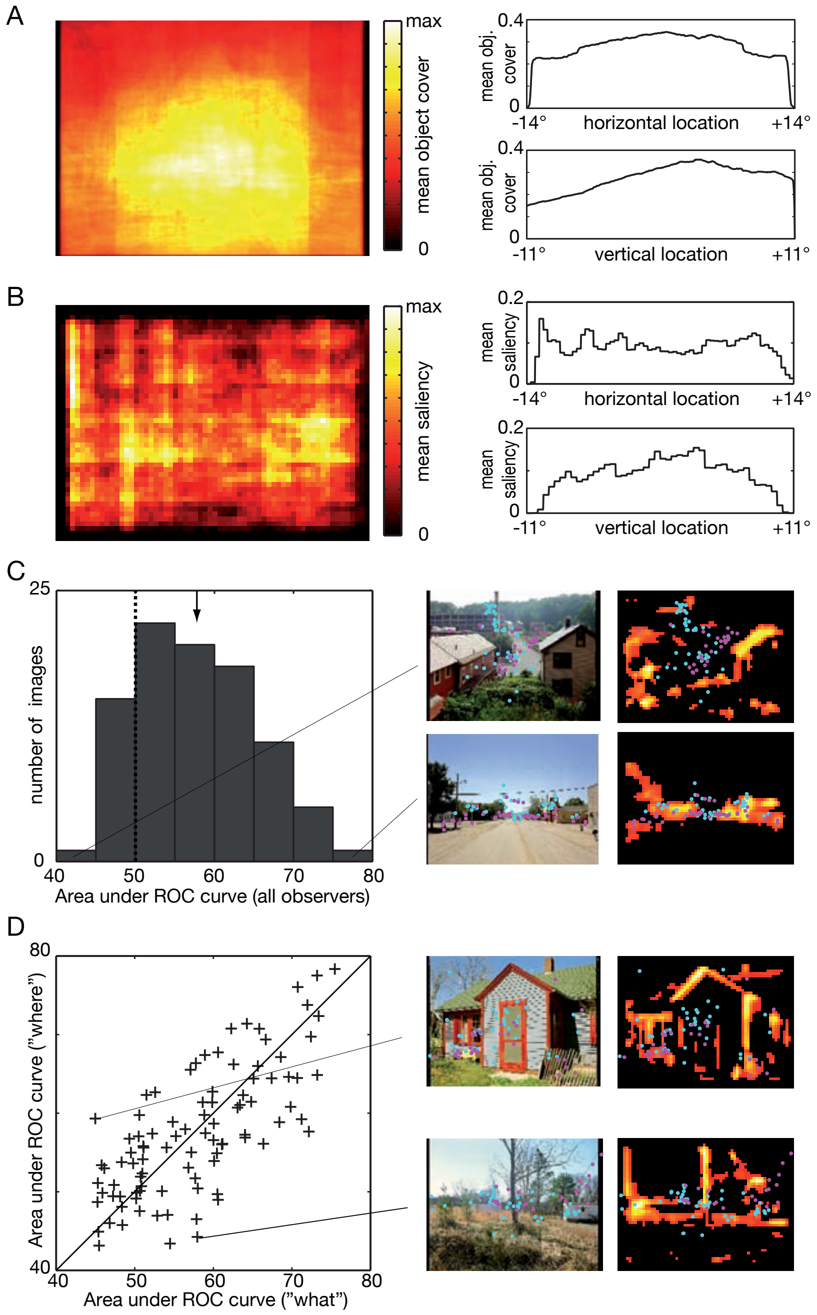


Figure 6.1: **A** Average object map, and its mean along cardinal axes. **B** Average saliency map and its mean. Note that there is no pronounced central bias to saliency. **C** Area under ROC curve for saliency map's prediction of fixated locations pooled over all observers in each image, histogram over images. Example images and saliency maps for images with best (bottom) and worst (top) fixation prediction. **D** Area under the curve separated for "what" and "where" observers. Each data point corresponds to one image; for points above the diagonal saliency map's prediction is better in "where" (45 images), below the diagonal in the "what" task (48 images). Example images and saliency maps for two data points.

of our stimulus material, in line with other stimulus sets used in the literature.

In contrast, the averaged saliency map does not exhibit a pronounced central peak. Instead, the saliency distribution is rather uniform if one ignores the boundaries where saliency is zero for technical reasons (Figure 6.1B). We conclude that, for our stimuli, saliency has no central bias.

6.2 Task and Fixation

Here we address the effect of task on fixation duration and location. During the 3s image presentation in the "what" task, observers make on average 10.0 ± 0.4 fixations (mean ± std across observers; all fixation counts exclude the initial 0th fixation). The mean is smaller (7.7 ± 2.0) for the "where" task, in which observers terminate each trial themselves, but there is a larger variation: The standard deviation across images is 1.9 ± 0.3 for "what," but 4.8 ± 2.1 for "where." As expected this high standard deviation arises from the fact that target-present trials have fewer fixations (7.0 ± 1.6) than target-absent trials (8.5 ± 2.6), and the inter-observer variation is substantial (Figure 6.2A).

Since trial duration differs between conditions, the fixation duration is of particular interest. In the "what" task a fixation takes 251 ms ± 134 ms (mean ± std across 3,716 fixations). In the "where" task a fixation takes 286 ms ± 153 ms (2,862 fixations), with no significant difference between target-present and target-absent trials ($p = 0.24$, t-test, Figure 6.2B). The fixation duration difference in the "what" and "where" tasks is highly significant ($p = 4 \times 10^{-23}$, t-test).

The spatial distribution of fixations shows a pronounced central bias (Figure 6.2C). Fixations in the "where" condition are spread more widely than in the "what" condition. The standard deviation of fixation location according to the Bienaymé formula quantifies this spread. It is larger in the "what" than in the "where" condition for fixations one to ten (Figure 6.2D). An alternative measure, the average distance between subsequent fixations, exhibits a similar time course (Figure 6.2E). In conclusion, duration and spatial distribution of fixations are task-dependent.

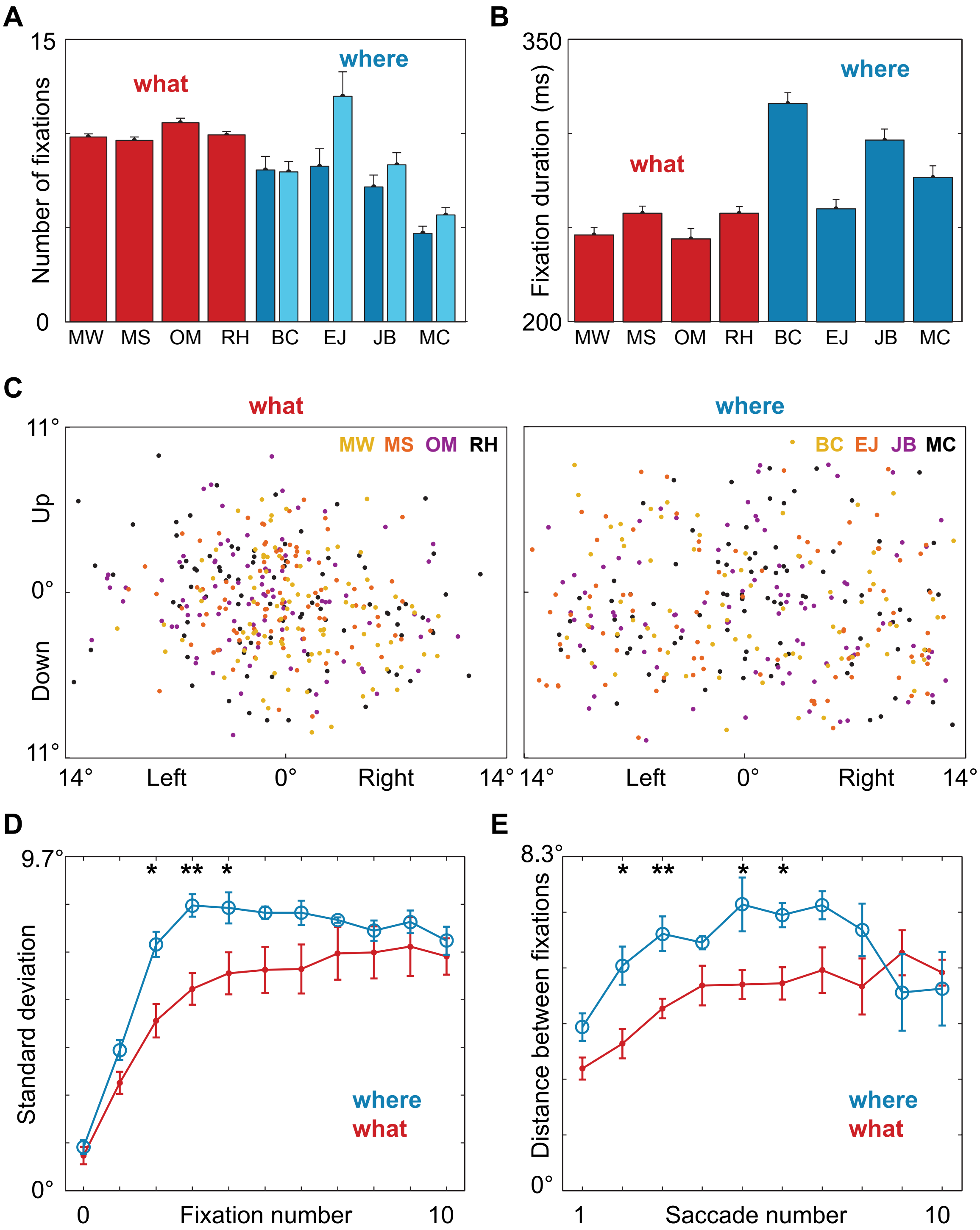


Figure 6.2: **A** Number of fixations mean and SE for individual observers. Red bars: "what" task observers (93 images), blue: "where" task, target present (51 images), light blue: "where" task, target absent (42 images). **B** Fixation duration mean and SE. **C** All third fixation locations on rectangle representing image for "what" and "where" tasks. **D** Standard deviation of fixation location (each y–tick mark corresponds to 50 pixels or 1.4°) by fixation number. Mean ± SE over observers in each task. 0 denotes initial fixation. **E** Distance between subsequent fixations. x–axis denotes fixation destination, so 1 denotes distance between initial and first fixations.

6.3 Fixation Consistency

To investigate inter-observer consistency, we compute a fixation map as described earlier except we leave out one observer and then predict that observer's fixations with the map (Figure 6.3A). The fixation map predicts fixations above chance (AUC $>$ 50%) in all images with the mean AUC over images ranging from 82.9% ± 8.4% (MW, mean ± sd) to 93.3% ± 5.4% (MC) with a 88.9% mean across observers (Figure 6.3B). The random reassignment baseline yields a range of 69.8% ± 11.0% to 79.8% ± 12.7% (mean: 75.7%, Figure 6.3B). This implies that a perfect model of average spatial distribution of fixations could predict up to 75.7% of fixations, without knowledge of the actual stimulus. Although this indicates that much inter-observer consistency is caused by common spatial biases, the actual data exceeds the random baseline significantly in all observers ($p_{max} = 4.8 \times 10^{-12}$, t-test). Consequently, there is a large image-specific component to inter-observer consistency. Limiting the map calculation to within-task slightly worsens predictions (Figure 6.3B), on average by 1.7% ("what" 1.9% ± 1.0%; "where" 1.6% ± 0.8%). This reduction is likely due to the smaller amount of data over which the map is computed, as the baseline shows a similar or larger drop ("what" 1.9% ± 0.3%; "where" 4.1% ± 0.9%). The significantly larger drop in the "where" condition (p=0.004, t-test), however, suggests that general spatial biases are slightly less relevant, as compared to image-specific effects, in the "where" condition. This is in line with the faster spread of fixations during search (Figure 6.2D,E).

Predicting fixations with a map from the other task is consistently worse than within-task prediction (Figure 6.3B) and significantly worse (at $p < 0.05$) in all but one observer (MW). This difference occurs even though the fixation map is based on four observers in the other task and only three in the same task. Nevertheless, even across-tasks, the prediction is above chance for all but one image (JB for the image of an isolated chair, Figure 5.1(4,4)). In the random reassignment baseline there is no difference between prediction within and across-tasks ($p > 0.05$ for all observers), such that we have no evidence for a task modulation of the generic component (or

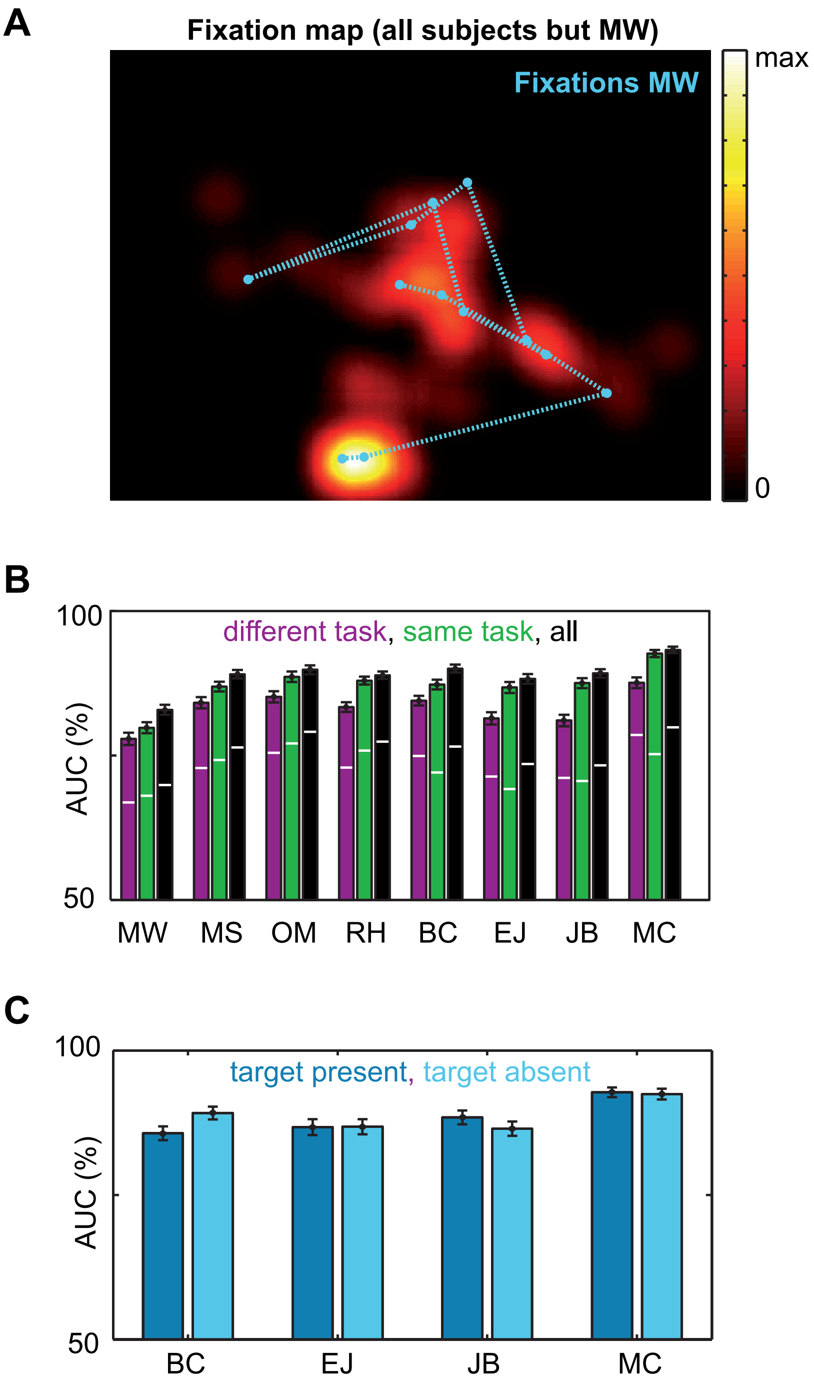


Figure 6.3: **A** Fixation map for image of Figure 5.2B with fixations of observer MW superimposed. **B** Area under ROC curve (AUC) for predicting fixations of one observer by the fixation map generated from other observers: 4 in the other task (left), 3 in the same task(middle), and 7 in both tasks (right). Mean and standard error across images for each observer. White lines denote results of random reassignment baseline. **C** Data for "where" observers of panel B separated by target present (dark gray) and target absent trials (light gray).

spatial bias) of inter-observer consistency. Within "where" observers prediction does not consistently depend on target presence (Figure 6.3C), ruling out that fixations on or close to the target dominate inter-observer consistency in the "where" task. In summary, there is enough inter-observer consistency to predict another individual's fixations, despite some task dependence.

6.4 Object Recall Consistency

In each of the 93 images, there were between 6 and 16 objects recalled by at least one observer and 10.5 ± 2.3 (mean ± sd) on average (Figure 6.4A). Across all images, the 8 observers recalled 981 individual objects (object categories are counted across images but once per image). Obvious synonyms were treated as the same object, while subcategories and parts were counted separately. Nearly half of the objects (457/981, 46.6%, Figure 6.4B) were recalled by only one individual, another 18.7% (183/981) only by two individuals. Analyzing the four observers in each task separately, the objects recalled by a single observer account for more than half of the objects recalled ("what" 338/590, 57.3%, Figure 6.4C; "where" 418/794, 52.6% Figure 6.4D). In all images there was at least one object recalled by at least four observers (Figure 6.4E). In 64/93 (68.8%) images, there was an object, which at least 7 observers recalled, and in 27/93 (29.0%) images, at least one object was recalled by all 8 observers (Figure 6.4E). This means that in most images, there is at least one *characteristic object*, an object that is recalled by most of the observers. This motivates the search for distinctive properties of these characteristic objects.

The order in which a given object is recalled presents an alternative measure how characteristic or important an object is for a scene. There is a highly significant correlation between recall frequency and recall order ($r = -0.31$, $p = 2\times10^{-23}$, Figure 6.4F, black). Actual recall ranks differ significantly from a baseline that corrects for having no lower limit on how many objects one recalls (Figure 6.4F, gray). Note that correlation values treat each of the 981 objects as individual datapoints, which is also used for the fits in the figures. Correlating naming frequency to mean values would result in higher correlation values ($r = 0.97$, $p = 6 \times 10^{-5}$). That is, individuals name frequently recalled objects earlier than idiosyncratic objects.

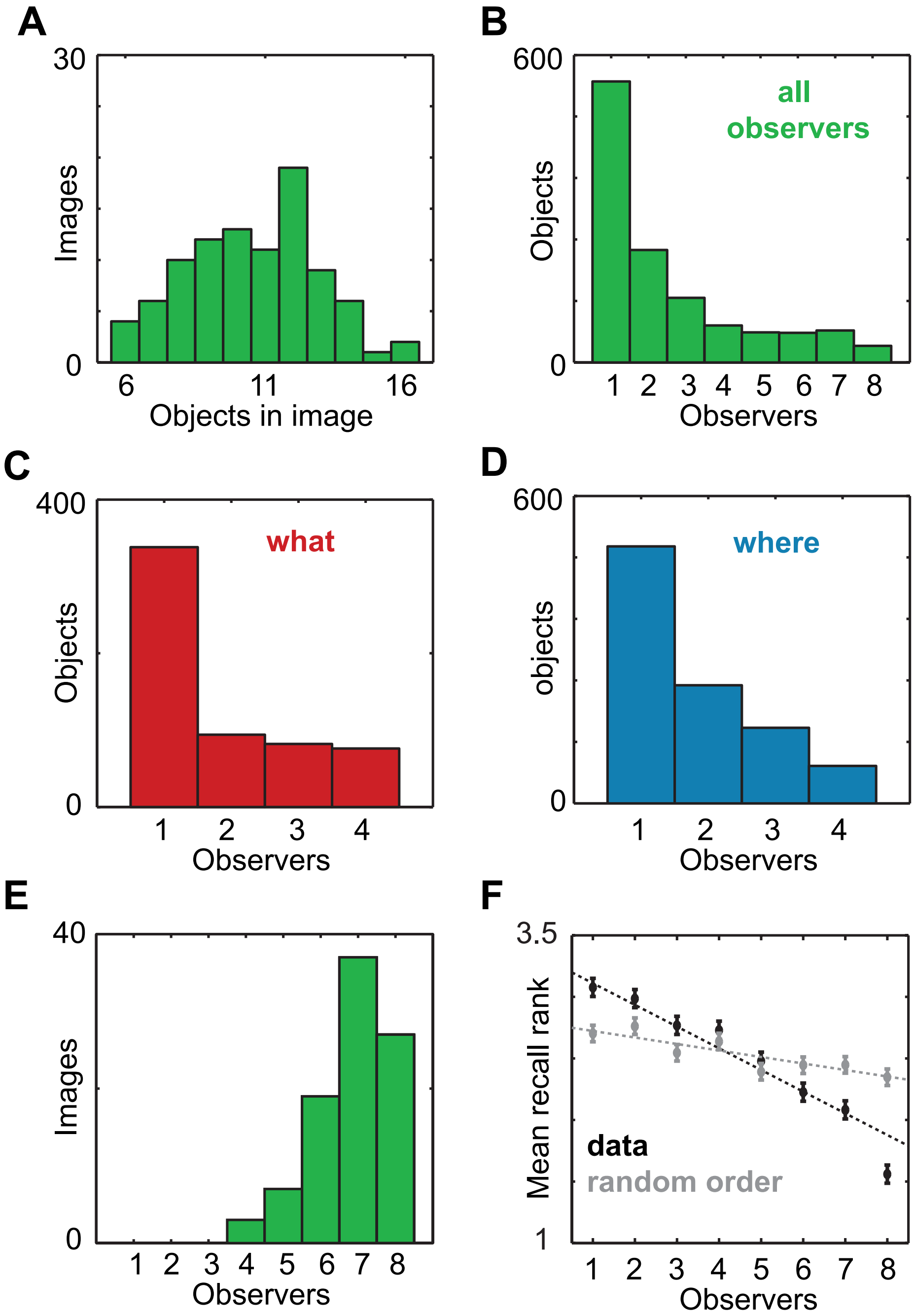


Figure 6.4: **A** Histogram of total objects recalled per image. **B** Histogram of number of observers recalling each object. **C** and **D** Same as B for "what" and "where" task observers only. **E** Histogram of the number of observers recalling the most frequently recalled object in an image. **F** Mean recall rank versus recall frequency. Black: mean ± SE across objects for human data. Gray: random baseline created by shuffling the objects a given observer recalled in an image. For extreme values of recall frequency the real data are significantly different from the baseline.

Chapter 7

Early Saliency and Fixations

7.1 Early Saliency Predicts Fixations Poorly

In this section, we assess how well saliency maps predict fixations. Basic fixation statistics, such as duration and spatial distribution, exhibit the expected dependence on task (Figure 6.2): in the "where" task, fixations last less time and are more widely spread. In some images saliency is an excellent predictor of fixated locations (Chapter 5), while in other images prediction is poor; the right panel of Figure 6.1C shows extreme examples of prediction performance. When pooling over all observers' fixations, the saliency map model's prediction is better than chance (50%) in 77/93 images. The mean area under the ROC curve is (57.8 ± 7.6)%, significantly different from chance ($p = 5 \times 10^{-16}$, t-test). To understand the meaning of this number, we compute the random assignment baseline as lower bound and the inter-observer prediction as upper bound.

To account for possible effects of spatial bias, we compute a *random reassignment* baseline (Chapter 5), as has been suggested earlier [84,127]. We superimpose fixations from one randomly chosen image on the saliency map of a different image. An effect resulting from generic biases would appear in this baseline. AUCs for this baseline reach (52.9 ± 5.7)%. Although this number is significantly larger than chance ($p = 3 \times 10^{-6}$), it is significantly exceeded by saliency's performance (57.8 ± 7.6)% ($p = 2 \times 10^{-6}$, t-test). Hence the prediction of fixations by saliency is not a consequence of general spatial bias alone.

As an upper bound we measure the fixation consistency of distinct observers. The fixations of one observer are predicted by a map generated from the fixations of all others with an average AUC of 88.9% (Figure 6.3). This number is far above the 57.8% obtained for saliency, which suggests that fixation prediction by saliency maps, albeit better than random, is far from optimal.

7.2 Task Independence of Saliency Map Predictions

Several recent studies [48, 139] suggest that saliency maps do not predict fixation in search tasks. As discussed above, we find a small amount of predictive power. Across our set of images, we do not find the prediction to be generally better for “what” than for “where”, although the differences in prediction performance can be substantial for individual images (Figure 6.1D). Therefore, across our set of object-rich images there is no evidence for saliency maps generally predicting fixations either better or worse in search tasks than in free-viewing for recall.

Chapter 8

Objects and Fixations

We now explore an alternative hypothesis: observers fixate objects instead of salient regions. If objects tend to be more salient than background then saliency maps would predict fixations indirectly, instead of driving fixation directly.

8.1 Predicting Fixations with Object Maps

To test how well objects predict fixations we define object maps in analogy to saliency maps (Chapter 5). The object map predicts fixated locations above chance in 83 images, with a mean AUC of 65.1% ± 10.6%, which significantly exceeds chance ($p = 5 \times 10^{-24}$, t-test, Figure 8.1A). This is not fully explained by general spatial biases, as it exceeds the random reassignment baselines of object maps and fixations (59.8% ± 10.7%) significantly ($p = 0.001$, t-test). When comparing the predictions of object map and saliency map for individual images, the object map outperforms the saliency map in 68 images, while the opposite is the case in only 25 images (Figure 8.1A). The image-specific ROCs are invariant to monotonic transformations, making direct comparison possible. A sign test shows that this fraction (68:25) is highly significant, even when ignoring the absolute size of the effect ($p = 9 \times 10^{-6}$). The default object map is weighted by the number of observers recalling an object. If instead the object map is based on the number of objects overlapping with a given pixel, the mean AUC drops to 61.9% ± 10.5%. This is significantly below the value for weighted maps ($p = 0.04$), but still significantly above the saliency maps' fixation prediction ($p = 0.003$). Image-by-image comparison shows that even the unweighted map outperforms raw saliency in 57/93 images, again a significant fraction (57:36, $p = 0.04$, sign test). Consequently, in most images, knowing the objects is more predictive of fixations than only knowing early saliency, even if the recall frequency of objects is unknown.

8.2 If Objects are Known, Early Saliency Contributes Little to Fixation Prediction

Object naming frequency predicts fixated locations in images. On average, this prediction is better than that of early saliency (Figure 8.1). Does saliency contribute any information beyond what objects tell us already? And vice versa? As first quantification, we ask how much a linear combination of maps can improve fixation prediction. Each pixel (x, y) in the image i has a value for the object map $O_i(x, y)$ and the saliency map $S_i(x, y)$. To account for their correlation, we treat O_i and S_i as dimensions of a plane, on which each original pixel is plotted. Note that the maps are normalized to the same dynamic range (0 to 1). For these data, we perform principal component analysis (PCA) and project the values on the principal axis. By reassigning the spatial coordinates, we obtain the linear combination of object and saliency maps that accounts for the most variance. Performing the signal detection analysis on this map yields a performance of 65.0% ± 11.6%, which is indistinguishable from the performance of the object map alone ($p = 0.995$, t-test), but significantly better than early saliency alone ($p = 10^{-6}$). The optimal linear combination of object and saliency maps is provided by Fisher's linear discriminant analysis (LDA).

Analogously we compute a map by projecting on the most discriminative dimension for separating fixated and non-fixated pixels. By construction, the prediction of this map for each image is better than the best of the individual maps. The average AUC over all images is 69.5% ± 8.2%, only 4.5% larger than the prediction by the object map alone. Hence the optimal linear combination of early saliency and object

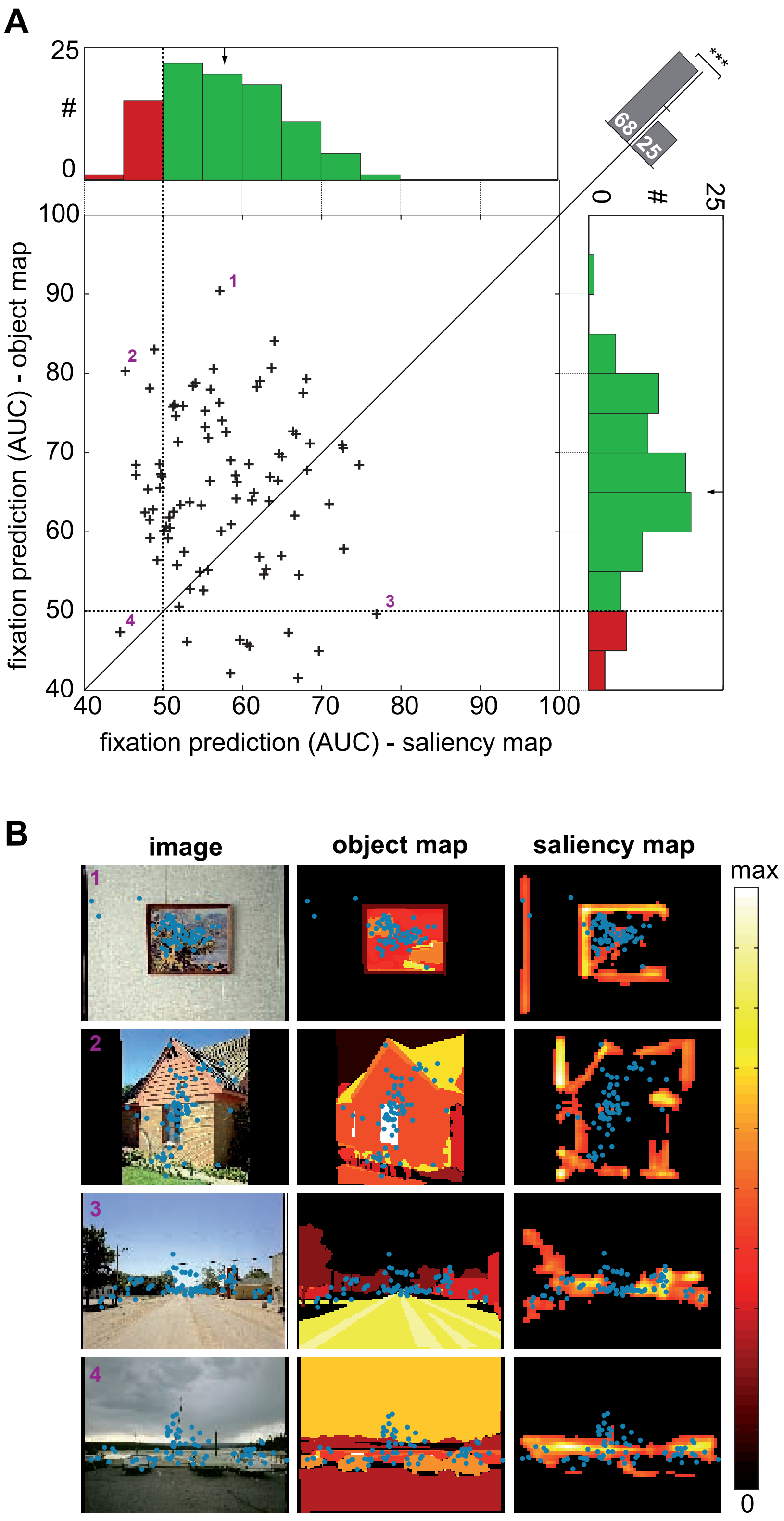


Figure 8.1: **A** Area under the curve (AUC) for fixations predicted by saliency and object maps. Each point corresponds to one image's AUC and histograms summarize performance. Purple numbers identify examples in panel B. **B** Examples of images, in which fixations are predicted **1** well by both, **2** well by the object map and poorly by saliency map, **3** well by saliency map and poorly by object map, and **4** poorly by both. Fixations of all observers in cyan.

map is only slightly better than the object map alone. Conversely, the optimal linear combination exceeds the AUC of saliency alone by 11.7%. This shows that early saliency does not add substantially to fixation prediction once recalled objects are known, while object maps are informative even when raw saliency is known. As we did not separate training and test set, 69.5% is an upper bound to the predictive power of the combined map on novel data. This is a strong indication that knowing saliency provides little extra information, once the objects are known.

8.3 Predicting Fixations with Object Saliency

Next we perform an alternative analysis to test whether saliency provides extra information on fixation probabilities beyond that already provided by object outlines. We combine object and saliency maps by computing four kinds of *object saliency maps* determined by two decisions: The first choice floods the object footprint with either the *maximum object saliency*, the maximum saliency map value inside the object, or the *total object saliency*, the sum of saliency map values inside the object. The second choice either weights the object with how many observers recalled it for *observer-weighted* or ignores recall frequency for *unweighted.* Fixation prediction with observer-weighted saliency maps is indistinguishable from object maps alone (65.1%): maximum object saliency results in 65.1% ± 10.9% AUC ($p = 0.98$, t-test), and total object saliency in 62.8% ± 12.0% (p=0.18). For unweighted maps, the numbers drop to 63.3% ± 11.4% and 62.3% ± 11.7%, respectively. These values fall between the results for weighted and unweighted object maps, but are indistinguishable from either (comparison to weighted object map: $p = 0.26$ and $p = 0.09$; comparison to unweighted: $p = 0.40$ and $p = 0.82$). This strengthens the result that–once the objects are known–saliency contributes little additional information to fixation prediction.

8.4 Predicting Recall with Fixations

Next we consider how well fixations predict object recall. For all analysis we split the objects into eight categories, depending on how many observers named the object.

First, we first pool fixations over all observers. The fraction of fixations that fall inside an object's boundary correlates with naming frequency (r=0.44, $p = 7 \times 10^{-49}$, Figure 8.2A) as does the relative time spent inside the object (r=0.43, $p = 3 \times 10^{-45}$). Frequently fixated objects are recalled more often. Using signal detection analysis, we compute how well the fraction of fixations inside an object discriminates objects recalled exactly once from objects recalled n or more times (Chapter 5). The fraction of fixations inside the object predicts whether an object is named twice or more (2+) compared to exactly once with an AUC of 70.3%. Objects named once are discriminated from objects named 8 times with an AUC of 90.4% (Figure 8.2B). This prediction is slightly better in the "where" task (67.2%, 76.3%, 81.1% for 1 vs. 2+, 1 vs. 3+, and 1 vs. 4) than in the "what" task (67.2%, 72.3%, 76.1%), but in general fixations predict recall well.

Since fixations are collected from the same individuals as recalled objects, one could argue that the relation between object maps and fixations just reflects the fact that fixated objects are recalled better. We test how well object maps obtained from a subset of observers predict the fixations of a different observer. As a baseline, we first predict fixations of each individual (instead of pooled fixations) by the full object map collected from all eight observers and average for each image over the eight resulting AUC values. As expected the mean AUC over observers is close to the pooled fixations (mean over images: 64.9% ± 10.8%, $p = 0.94$, t-test). More importantly, excluding the map of the observer, whose fixations are predicted, does not impair the result significantly (mean 64.5% ± 10.8%, $p = 0.77$). This shows that the predictive effect of object maps is not contingent on including a particular observer's fixations. To verify this further, we asked a single observer to label all objects in the ten images with best object map prediction. This observer was given unlimited time and paid based on the amount of labeled objects. Note that overlap

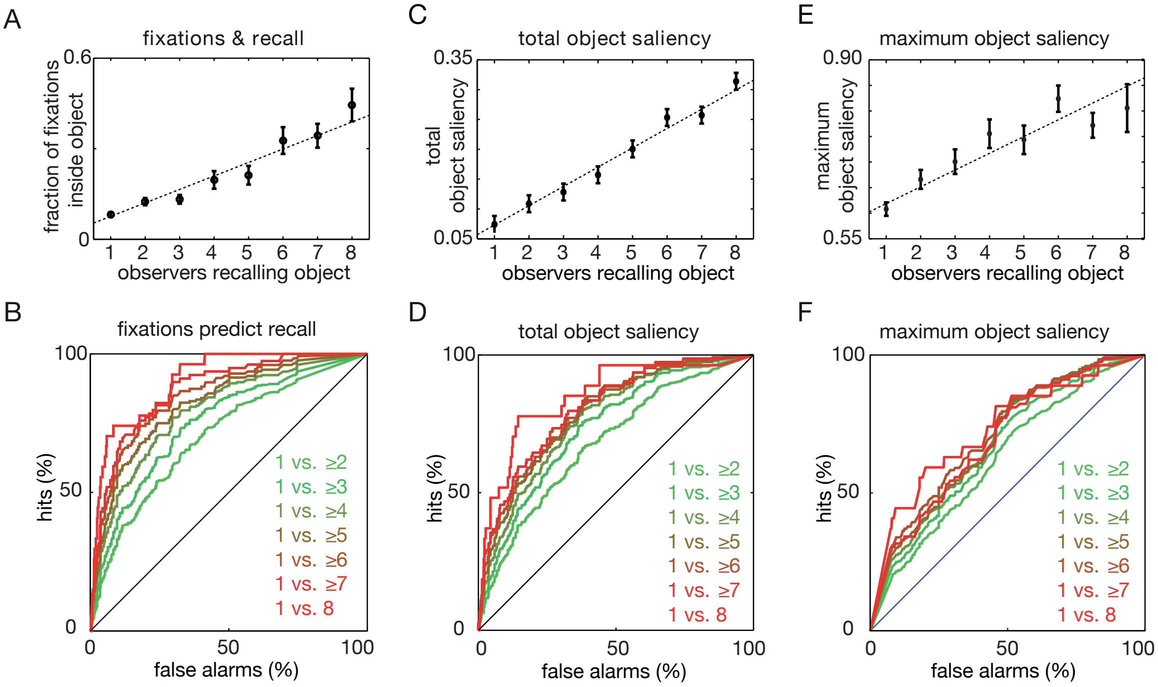


Figure 8.2: **A** Fraction of fixations inside object versus recall frequency. Mean ± SE across objects, fit treats all 981 objects individually. **B** ROC curves separating objects recalled once from objects recalled twice or more, three or more times, *etc*. using fraction of fixations on object. **C** Object saliency plotted versus recall frequency. **D** ROC curves in analogy to panel B, using object saliency. **E** As panel C for maximum object saliency. **F** As panel D for maximum object saliency.

of different objects prevents even the map of a single observer from being binary and from simply converging to uniformity. As expected, the prediction of this individual's object map is worse than the eight-observer object map for all ten images tested. However, the prediction of the individual's object map is still better than chance in 10/10 and better than that of the saliency map in 9/10 images. This shows that the predictive effect of the object maps is not contingent on the map resulting from the same observer or limits on labeling time.

We also look at the relationship between the fraction of fixations that fall on an object and the log-odds of an object being named versus not named. Here, we count an object as fixated if some fixation falls within 1° visual angle or 36 pixels of the object mask. We compare the fraction of fixations for named and unnamed objects. A higher fraction of a subject's fixations falls on named objects for all viewers (all subjects excluding MC $p \leq 10^{-11}$; MC $p = 0.03$) for a Wilcoxon rank sum test with a Holm-Bonferroni correction (Figure 8.3A). We also find a higher fraction of other viewer's fixation on named objects for seven out of eight subjects ($p \leq 10^{-5}$; MC $p = 0.11$). It might be possible to explain away this relationship as viewers naming objects they fixate, instead of fixating objects. However, a similar relationship is seen between one's own fixations and others' fixations on average (Figure 8.4) and for individuals (Figure 8.3B). This in turn could be explained by the correlation between different viewers' fixations. We take a step farther and consider the fixations of other observers on objects that were not fixated by the observer in question (Figure 8.5). As few (10-16%) named objects were not fixated by the viewers (excluding MC 30%) the individual tests were not significant, but pooling the data across subjects and excluding MC yielded a just significant result ($p = 0.49$). The recall of observer MC is not well predicted by the fixations of others (Figure 8.3B) and MC made fewer fixations than other viewers (Figure 8.6). Such a clean relationship is not seen between recall and the absolute number on fixations on an object or the distance of the nearest fixation to an object (Figure 8.7).

In summary, although fixations and recall are coupled, this effect is not observer-specific. Instead of recalling an object because of having fixated it, frequently recalled

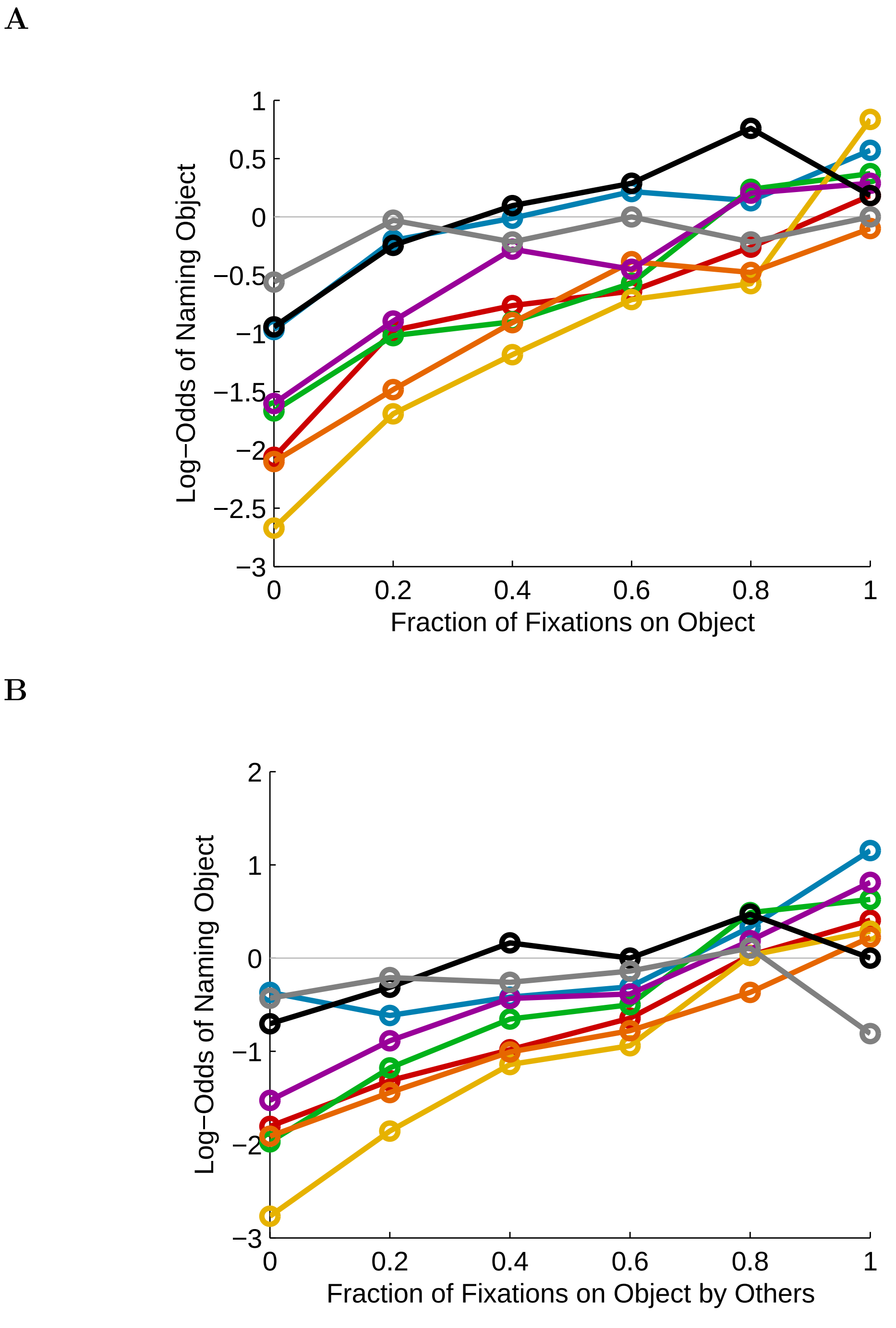


Figure 8.3: Log-odds of naming an object versus fraction of fixations on object by individual. **A** Fixations come from the same observer as recall. **B** Fixations come from the other 7 observers. Observer MC displays a different pattern than other observers (gray).

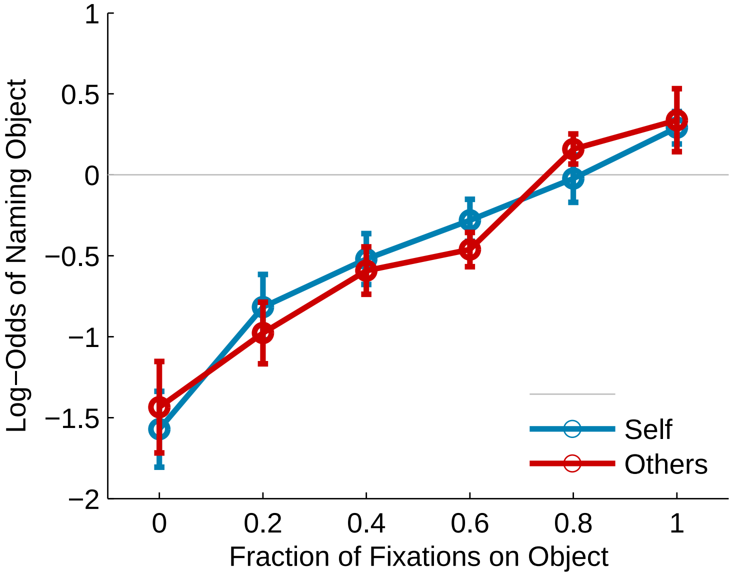


Figure 8.4: The log-odds of naming to not naming an object versus the fraction of fixations on an object. Averaging over viewers, we find similar relationships between one's own fixations and others' fixation with object naming.

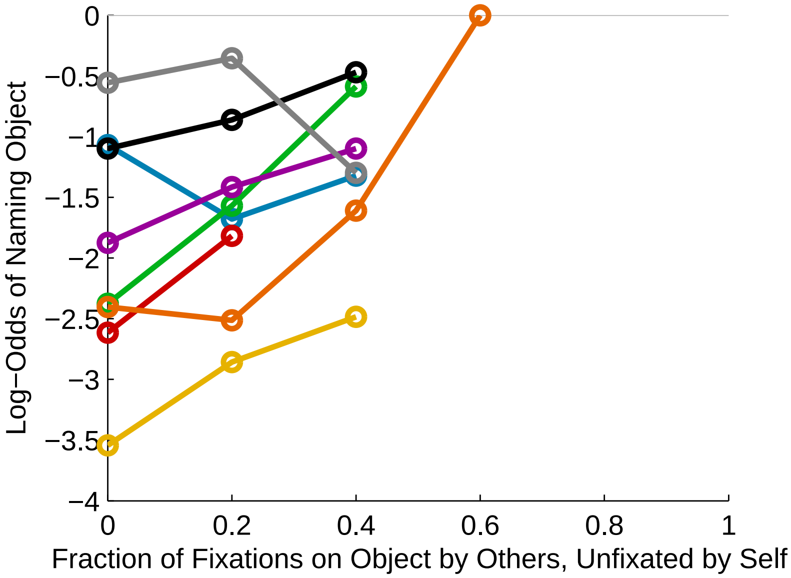


Figure 8.5: Log-odds of naming an unfixated object versus fraction of fixations on object by other viewers.

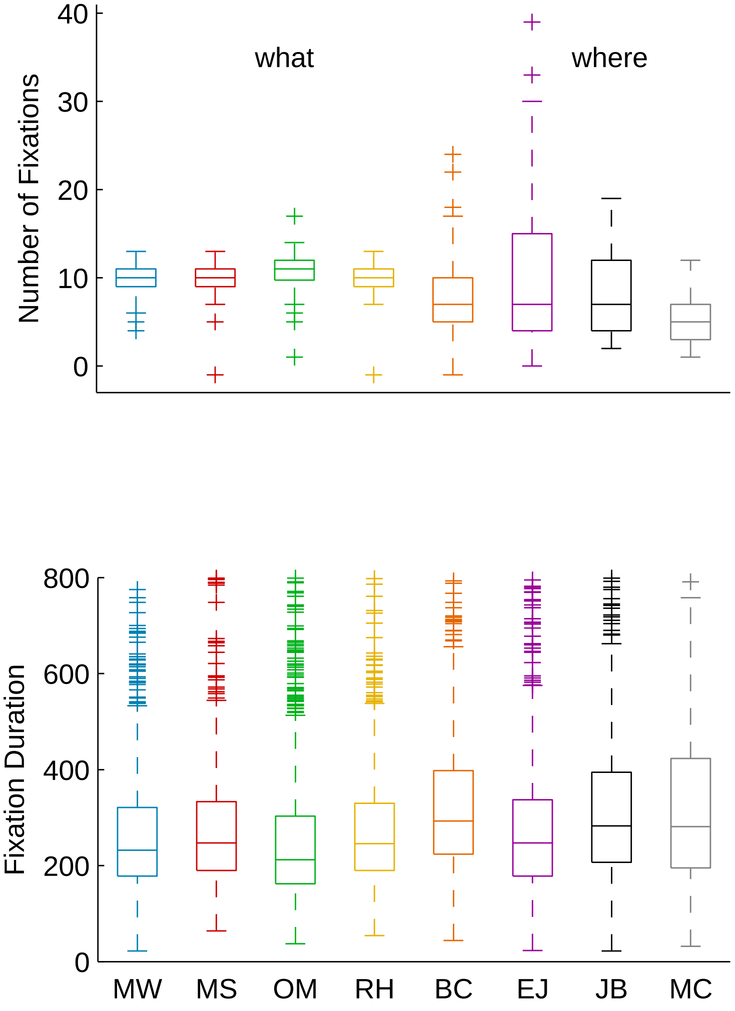
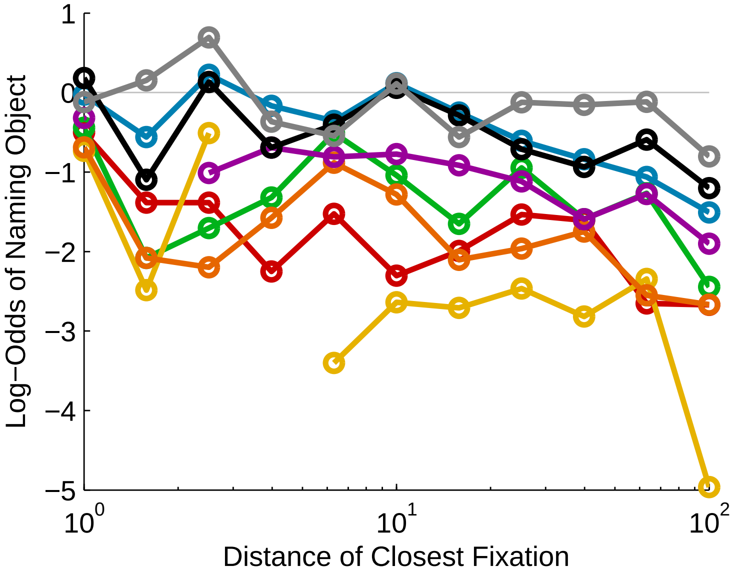


Figure 8.6: Median and quartile data for fixation count and duration by individual. MC shows similar fixation duration to other viewers, but fewer fixations.

A



B

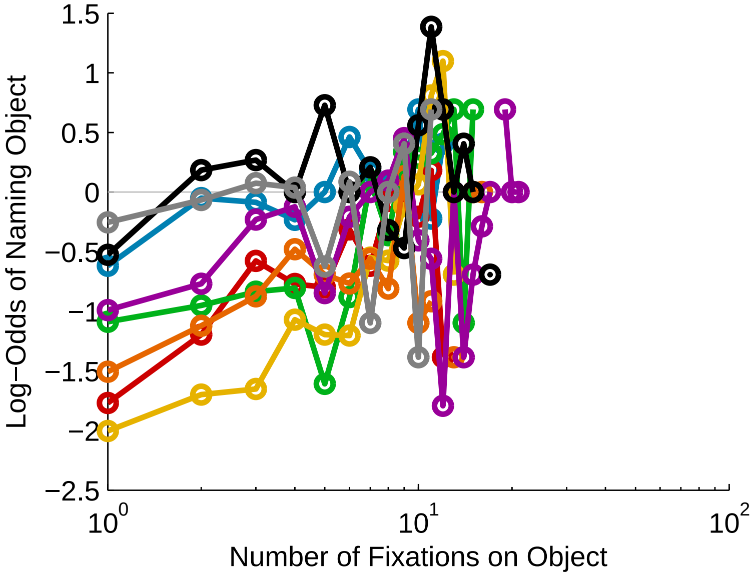


Figure 8.7: **A** Log-odds of naming an object versus distance of closest fixation to object in pixels. **B** Log-odds of naming an object versus absolute number of fixations on object.

objects are fixated frequently, even when fixations and recall come from different observers or the recalling observer does not fixate the object.

Chapter 9

Saliency and Object Recall

So far we have shown that saliency maps predict fixations to a limited extent, and frequently recalled objects are preferentially fixated. Next we aim at completing the argument that saliency maps predict objects and thus predict fixations indirectly. The missing part has recently been suggested (Elazary and Itti, 2008), but needs to be demonstrated for our data and conditions: How well do saliency maps predict object recall, how do their predictions compare with the predictions of other object properties, and do their predictions extend beyond fixation alone?

9.1 Object Saliency Predicts Recall Frequency

We assign each object a relative *total object saliency*, defined as the sum of saliency map values on the object divided by the sum across the whole image. Across all objects and observers, object saliency is highly significantly correlated to recall frequency ($r = 0.38$, $p = 2 \times 10^{-34}$, Figure 8.2C). Does this imply that object saliency predicts recall frequency on an object-by-object basis? As earlier, we perform signal detection analysis, testing how well objects named once can be discriminated from objects recalled more often. Based on object saliency, objects named by all observers are distinguishable from those named once with an AUC of 85.0%, and even objects named twice or more are distinguishable from those named once by an AUC of 68.2% (Figure 8.2D). This is only slightly worse than prediction by the fraction of fixations (Figure 8.2B); total object saliency predicts recall frequency nearly as well as the fraction of fixations on an object. For the "what" task, we find AUCs of 68.9%, 72.0%, and 76.6% for distinguishing an object that is named by exactly one observer from those named by two observers or more, named by three observers or more, and named by all four observers. The results for the "where" task are only slightly different and the differences do not have a consistent sign (AUC: 66.8%, 72.2%, 74.6%). This shows that object saliency's prediction of recall is not task-dependent.

Since total object saliency scales with object size, we also consider *maximum object saliency*, the maximum saliency map value inside an object. Although the correlation between maximum object saliency and recall frequency is lower than for total saliency ($r = 0.25$, Figure 8.2E), it is still larger than all other measures except object area, and is highly significantly different from 0 ($p = 1.2 \times 10^{-15}$). Similarly, prediction by maximum object saliency reaches AUCs from 62.3% (1 vs. 2+) to 72.9% (1 vs. 8), lower than for total object saliency, but still substantially above chance (Figure 8.2F).

Several measures other than object saliency suggest themselves for predicting object recall. For object location there is a highly significant correlation between the mean horizontal distance of an object to the image center and its recall frequency ($r = -0.20$; $p = 2 \times 10^{-10}$), whereas the vertical distance exhibits no significant correlation ($r = -0.04$, $p = 0.19$). Object size also seems intuitive for recall. Recall frequency is significantly correlated to the area covered by the object ($r = 0.32$, $p = 2 \times 10^{-24}$), and the length of the object's boundary ($r = 0.22$, $p = 8 \times 10^{-12}$). Although this indicates that observers preferentially recall large, central objects, the correlation between total object saliency and recall frequency (Figure 8.2C) exceeds all other measures tested. These object measures are correlated with object saliency measures and thus partly redundant in predicting object recall.

9.2 Combinations of Properties Predict Recall

Total object saliency combines the saliency of an object and its area to a common measure. So, both measures are tightly correlated ($r = 0.76$, $p = 3 \times 10^{-186}$); similarly, boundary length and area are trivially coupled, with larger area implying a longer boundary ($r = 0.63$, $p = 8 \times 10^{-110}$). As only parts of objects within the image are used to determine its center of mass, large objects are biased toward the center, reflected in a correlation between center distance and area ($r = -0.30$, $p = 4\times10^{-22}$). Maximum object saliency is correlated to all these measures, trivially to total object saliency ($r = 0.56$; $p = 1.5 \times 10^{-81}$) and to object area ($r = 0.39$,$p = 3.4 \times 10^{-36}$). The latter correlation can partly be understood as a consequence of the sparsity of saliency maps: peaks are rare, while low values occur frequently; hence, larger objects have a slightly better chance to capture a peak.

How well does a linear combination of these properties predict recall frequency? Combining area and total object saliency by performing discrimination along the first principal axis of all data yields slightly better results than either measure alone: AUCs range from 69.2% (named once versus named twice or more, Figure 9.1) to 86.2% (once versus eight times). Similar unsupervised inclusion of the other measures or combining more than two measures does not yield better prediction performance (Figure 9.1).

Hence, total object saliency is a better predictor of recall than all other measures combined, and including other measures only marginally improves prediction. As object saliency enables good prediction of how often an object is recalled, how redundant are fixations? Figure 9.2A depicts the relation of total object saliency and fraction of fixations inside an object. Combining the measures along the principal axis of all data slightly improves the discrimination of rarely named objects from others, but does not improve already good discrimination (Figure 9.2B). Similarly, maximum object saliency and fixations on an object are related in predicting recall (Figure 9.2C), but fixation does not add much to object saliency (Figure 9.2D). Interestingly, combining maximum object saliency with object area does not reach the levels of total object area, which argues against the effect of total object saliency resulting from its correlation with object area. This implies that frequently named objects are distinguished from rarely named objects by virtue of maximum or total object saliency and knowing the fixations provides little extra information.

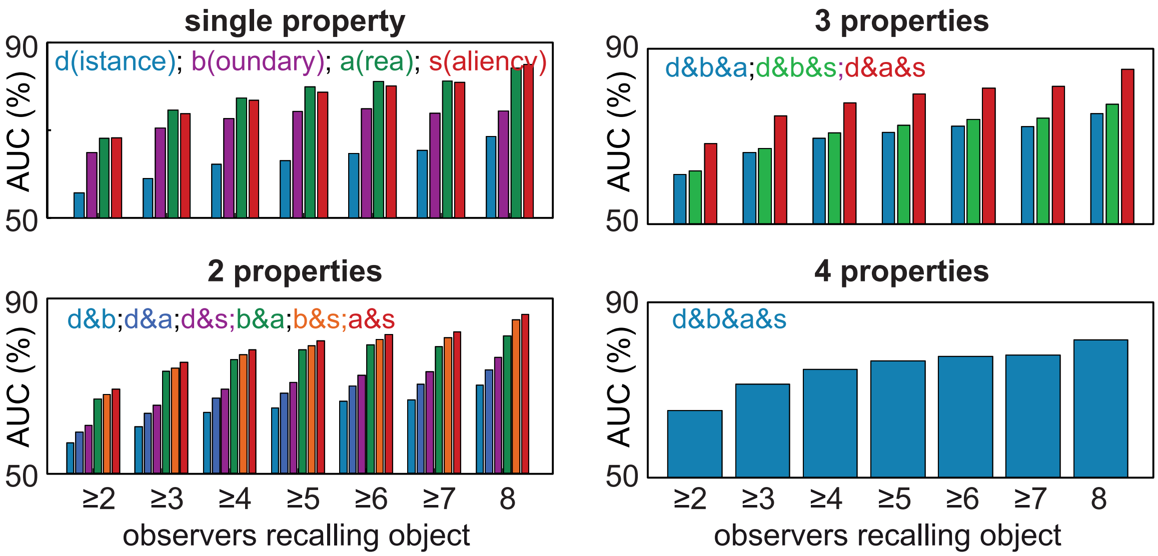


Figure 9.1: Recall prediction by various object properties and their linear combination. AUC for prediction of recall using combinations of four image properties: distance from center, boundary length, fractional area, total object saliency. Bar colors denote measures as given in panel legend. Measures including saliency or area outperform the other measures.

Interestingly, for idiosyncratic objects the fraction of fixations inside the object is consistently low. Less than 25% of such objects have a fixation fraction above 8.9%. In contrast, for objects recalled by all observers the middle half of data extends from 15.8% to 70.4%. In general, objects recalled by many observers received a wider range of fraction of fixations, than objects recalled by few (Figure 9.2E). A similar tendency is observed for total object saliency (Figure 9.2F).

It is tempting to speculate that objects recalled by many observers do not require a fixation to be recalled, while a fixation is necessary to recall objects that are recalled by few. In this view, objects recalled frequently would be named because they are diagnostic for a scene or consistent with its general context, while lesser named objects are primarily recalled as a consequence of fixation. If this hypothesis holds true, the probability to fixate an infrequently recalled object should be larger for the observers recalling it, or *recalling observers*, than for the *non-recalling observers*. This difference should be less pronounced for more frequently named objects. Of the 457 idiosyncratic objects, the recalling observer fixated 188 (41.1%). This compares to

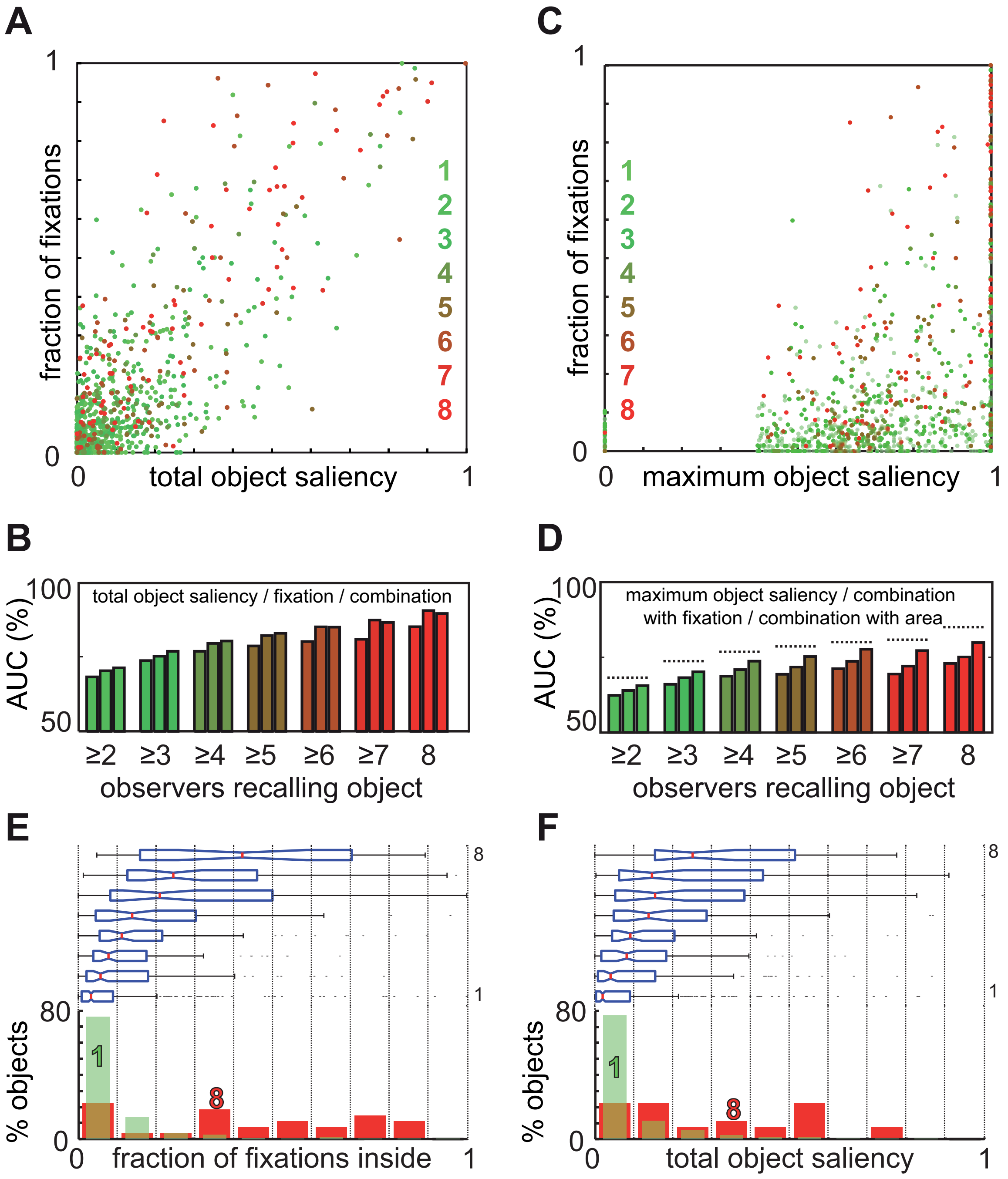


Figure 9.2: **A** Fraction of fixations on an object plotted against total object saliency, color denotes recall frequency. **B** AUC for recall prediction on the basis of total object saliency alone (left bars), fixations alone (middle bars), or the combination of both (right bars). **C** As panel A for maximum object saliency. Note that maximum saliency frequently takes extreme values. **D** AUC for prediction by maximum object saliency (left), maximum object saliency combined with fixations (middle), and maximum object saliency combined with object area (right). Dotted lines replicate the results for total object saliency from panel B. **E** Normalized histograms of fraction of fixations inside object boundary for objects recalled by 1 (green) or all (red) observers; Boxplots of fixations inside object for objects recalled by all (top) to 1 (bottom) observers. **F** As panel E for total object saliency with same horizontal axis.

33.6% of non-recalling observers fixating the same objects. Hence, for idiosyncratic objects, recalling observers are about 22.5% more likely to fixate the recalled object than non-recalling observers. The symmetric situation is constituted by the 52 objects that have seven recalling and one non-recalling observers. Here 78.3% of the recalling observers fixated the object, compared to 73.1% for non-recalling observers. Hence for frequently recalled objects, recalling observers are only 7.1% more likely to fixate the object than non-recalling observers. Similar patterns arise if the fraction of fixations inside the object is considered instead of binary fixated/non-fixated split: For idiosyncratic objects, the fractions are 10.5% for recallers, compared to 7.9% for non-recallers, an increase of 32.9%. For objects recalled by seven observers, the increase is merely 7.9% (34.9% compared to 32.3%). The increase in fixation fraction from non-recallers to recallers is anti-correlated with the overall number of observers recalling the object (1,...7) ($r = -0.85$, $p = 0.02$). Consequently and consistent with the hypothesis, the relative benefit of fixation for recall reduces with increasing number of recalling observers.

9.3 Saliency Predicts a Scene's Most Characteristic Object

As described before, all object saliency measures are correlated with other object properties such as area. To estimate the effectiveness of saliency in identifying relevant objects in a scene, a more direct approach is to ask whether saliency can predict the most frequently recalled or *characteristic object*. The object with the highest total object saliency is among those named most frequently in 34/93 images. The most frequently named object is unique in 77/93 images (in all but one image, no more than two objects share the highest naming count). In 28/77 of these images, the object most frequently named has the highest total object saliency (Table 9.1). For comparison, we measure the probability of obtaining this result through random selection. By performing 10,000 simulations of this drawing process, we estimate the expected numbers to be 11.0/93 and 7.7/77, more than threefold below the actual values. The maxima obtained across these 10,000 simulations are 25/93 and 19/77. This indicates that the probability of obtaining the actual numbers of 34/93 and 28/77 at random is far below 1/10,000 ($p \ll 10^{-5}$). Hence, total object saliency predicts the most frequently named object significantly better than uniform random selection. Since total object saliency factors in object area, we also tested maximum object saliency. The object with highest maximum object saliency is among the most frequently selected in 35/93 and 26/77 images ($P_{93}(X \geq 35) \ll 10^{-5}$;$P_{77}(X \geq 26) \ll 10^{-5}$). Remarkably, 9/26 (13/35) of these objects were not selected by total object saliency (Table 9.1).

How does the object saliency measure compare to other measures in predicting the most frequently named object? The largest object is among the most frequently named in 22/93 (16/77) images, which is still significantly better than chance (simulations: $P_{93}(X \geq 22) = 0.001$; $P_{77}(X \geq 16) = 0.004$) but more than 50% exceeded by the 34/93 and 28/77 of saliency. Similarly, proximity to the image center is not as predictive (23/93, 18/77; $P_{93}(X \geq 23) = 0.0006$, $P_{77}(X \geq 18) = 0.0006$) as saliency. Choosing the object with the longest boundary is indistinguishable from random selection (13/93, 7/77; $P_{93}(X \geq 13) = 0.30$, $P_{77}(X \geq 7) = 0.65$). This shows that although object size and central bias contribute to object selection, they are exceeded by both total and maximum object saliency.

The characteristic object is among the largest, most central, and longest boundary objects in 36/93 (27/77) cases, making this combination of properties comparable to or worse than object saliency (Table 9.1). In only 10/59 (7/49) images for which the characteristic object is not the most salient do the other measures predict this object. Hence, the other measures provide additional information to object saliency in only a few images. So, object saliency predicts the most frequently named object better than any other tested measure or combination of them. Furthermore, other measures do not add much, once object saliency is known. In summary, object saliency best predicts which object is most frequently recalled in each image.

Table 9.1: Out of the 77 images, which have a unique characteristic object (2nd column), this object has the highest total object saliency in 28 images (4th column), the highest maximum object saliency in 26 (5th column), is the largest in 16 (7th column), the closest to the center in 18 (8th column), and the one with largest boundary in 7 (9th column). The maximum of the saliency map falls on the most frequently recalled object in 22 images (6th column), even if the fraction of image covered by this object may be as small as 4% (number in 6th column).

		Times	Highest	Highest	Saliency	Largest	Closest	Longest
Image	Most Named	Named	Total Saliency	Max Saliency	Peak	Area	to Center	Boundary
9	Church	8		X	X (.30)			
16	Car	8	X	X		X	X	
24	Man	8	X			X		X
26	House	8	X	X		X	X	
27	Chair	8	X	X	X (.36)	X	X	
31	Road	8				X		
40	TV	8	X	X	X (.25)		X	
46	Lamp	8					X	X
48	Building	8	X	X	X (.14)			
59	House	8	X	X	X (.49)	X	X	
65	Car	8	X					
69	Pool	8		X	X (.68)			
71	Pool	8	X			X		
73	Mailbox	8					X	
84	Shed	8	X				X	
85	Bed	8	X	X	X (.20)		X	
91	Lightbulb	8		X	X (.12)			
5	Painting	7	X	X	X (.50)	X	X	
12	House	7	X					
15	Parking Lot	7	X					
18	Puzzle	7					X	
19	Trailer	7	X				X	
20	House	7	X	X	X (.63)	X	X	X
22	House	7	X	X	X (.30)		X	
25	House	7	X	X	X (.26)	X		
41	House	7	X	X	X (.42)	X	X	X
46	Parking Lot	7						X
54	House	7	X	X	X (.67)	X	X	
60	Woman	7	X	X	X (.31)			
63	Car	7	X			X		X
80	Car	7					X	
81	Team	7		X				
90	House	7					X	
4	Chair	6		X				
14	Cafe	6		X				
29	House	6	X					
53	Building	6	X	X	X (.73)	X		
68	Ford-Sign	6		X	X (.04)			
74	Tree	6	X			X		X
77	Man	6	X	X	X (.31)			
49	Companion	5	X	X	X (.56)	X		
58	Flag	5		X	X (.07)			
64	Bush	4		X	X (.31)			
83	Field	4	X					
Sum			28	26	22	16	18	7

9.4 Saliency Peak Falls on Frequently Named Objects

A complementary way of analyzing how well saliency predicts named objects is to ask whether the *peak* or maximum value of the saliency map is located within a recalled object and, if so, is it within the characteristic object. Note that this is different from the maximum object saliency analysis before, as the peak is determined over the whole image and there is a possibility that the maximum is not covered by an object. The baseline for this analysis is the probability that the peak falls on the object at random, which equals the object area divided by the image area. The peak falls on a named object in 78/93 images (83.4%), compared to the mean over all images for the baseline value (mean object coverage) of 77.0% ± 18.7%. To assess significance we compare the mean of the baseline values to the fraction of images in which the maximum is located within the object boundary (one value for the set), and find them to be significantly different ($p = 6.4 \times 10^{-4}$, t-test). The most frequently named object encloses the maximum of the saliency map in 29/93 images (31.2%), which is again significantly larger than the baseline (22.6% ± 19.4%) of area covered by the most frequently recalled object(s) ($p = 4.6 \times 10^{-5}$). Restricting analysis to the 77 images with a unique characteristic object, the maximum is in this object in 22/77 images (28.6%) compared to the 20.6% ± 17.8% of area covered by these objects on average ($p = 1.9 \times 10^{-4}$). Table 9.1 (6th column) provides a list of these objects with the respective baseline values. In summary, these data show that saliency maps, even without any further knowledge of object content, can be used to pick an image region containing a relevant object better than chance. This reinforces the interpretation of saliency maps as measures of (possibly pre-attentive) scene content.

Conclusion

The present study reconciles two apparently conflicting views of attention: On one hand, theoretical models use early features [58, 67] and presuppose saliency computation in early visual areas [76]. On the other hand, there is mounting physiological evidence that saliency is computed later in the visual hierarchy: frontal areas, such as the frontal eye fields [131] are known to represent saliency. Furthermore, recent microstimulation experiments [3] suggest a direct link from FEF to saliency representation in visual area V4, which is a prime physiological candidate for saliency computation [4, 88, 95]. In light of our results, these views are not conflicting (Figure 9.3). Early saliency is computed in early visual areas V1 and V2, but alone has only a small effect on attention guidance (57.8% AUC; Figure 9.3 black pathway). Instead, early saliency in combination with other object properties models the probability of an object being recalled. The location of characteristic objects then predicts attention (65.1% AUC) better than early saliency alone. Furthermore, adding early saliency information to object location contributes little predictive power. LDA analysis shows that an upper bound on the linear combination of object footprints and early saliency does not substantially exceed object recall alone (69.5% versus 65.1%; green versus red). In this view, early saliency does not drive attention directly (Figure 9.3 black pathway), but through its correlation with object properties (red pathway). In other words, the prediction of objects by saliency, together with the prediction of fixations by objects, explain away the prediction of fixations by saliency. Based on the aforementioned physiological evidence, we may speculate that areas high in the ventral stream, such as V4 or IT, serve as an integration site of object recognition and early saliency. Regardless of cortical site, our data indicate that the computation

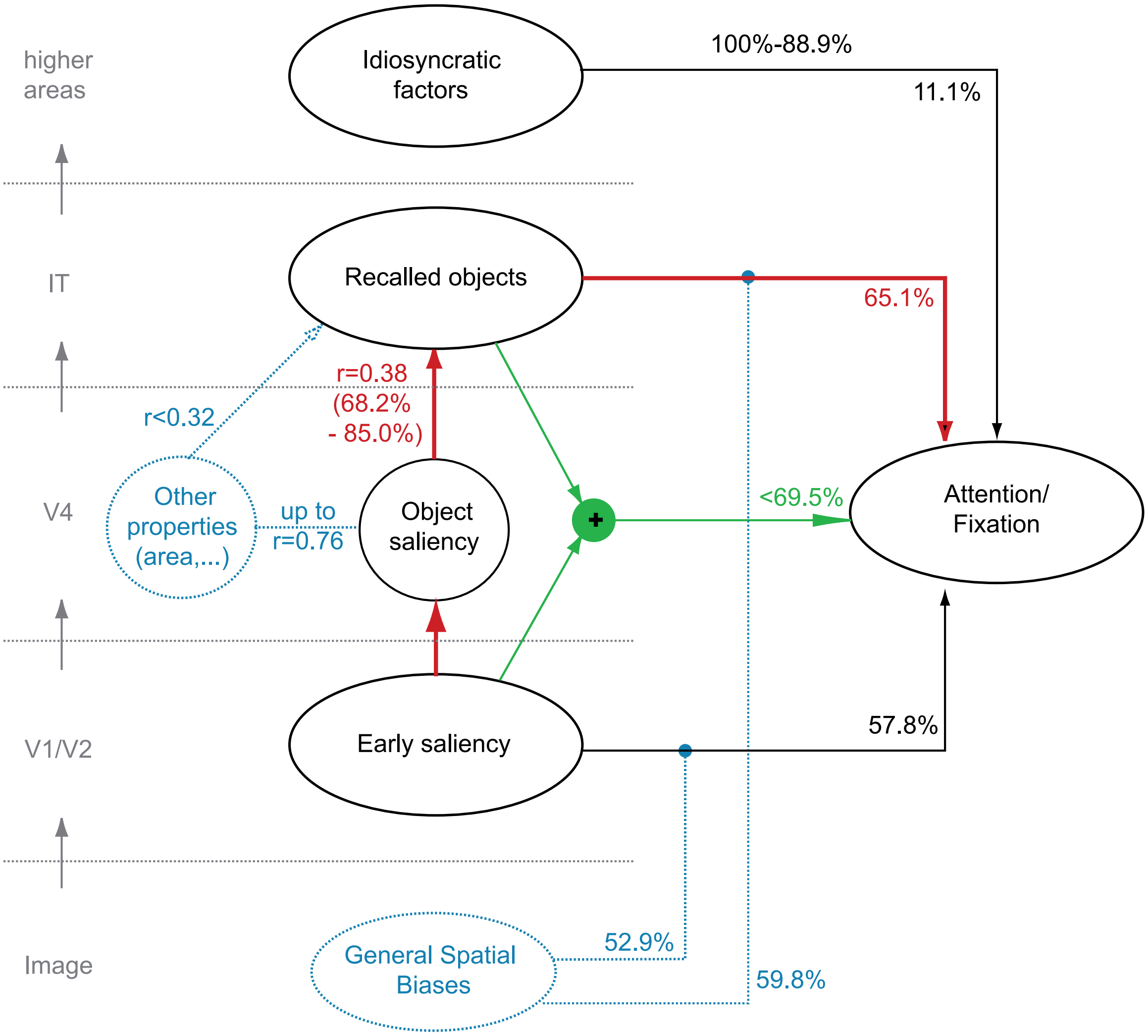


Figure 9.3: Right: Although early saliency predicts fixations to some extent (57.8% average AUC), this prediction is mostly explained through correlations with object recall (red). Object saliency is a measure of saliency within an object's boundary, which is highly correlated with recall frequency. The resulting object map predicts fixations (65.1% AUC) slightly below the upper bound for an optimal linear combination with saliency (69.5%; green). The random reassignment baseline reveals that some of the results are accounted for by general spatial biases, which are not specific to individual images (blue). Idiosyncratic factors include everything not explained by the mutual prediction of different observers (88.9% AUC). Left: putative brain areas for computation of the individual steps: early saliency is based on early visual mechanisms, while object representations follow in higher ventral areas. This is consistent with the prime site of saliency computation being in V4 or IT.

of attention-driving saliency may be distributed, but has a component late in visual processing that is relevant for natural scene perception.

The suggestion of the present data that saliency drives attention indirectly through predicting interesting objects reconciles earlier findings: saliency map features do not need to drive attention [9, 22, 129] despite saliency's undisputed correlation with fixations during free viewing [99]. However, we do not argue that saliency maps fully answer how interesting objects are selected, or that saliency map features causally drive object recognition. Further research by targeted manipulations of object properties is needed to analyze which stimulus features drive attention, and how they relate to features that make an object interesting, characteristic, or diagnostic for a scene and to different types of recall (tokens, types, scene gist, object positions, *etc.*). However, our data suggest that the allocation of attention is preceded by some pre-attentive scene understanding. This is in line with the data [13, 14], showing that the even the earliest guidance of attention and fixation depends on whether an object is semantically plausible in a scene. The minimum requirement for such a decision is a coarse pre-attentive recognition of the scene context, or gist, and some form of pre-attentive figure-ground segmentation. Taken with our present data, this strongly suggests that attention cannot be understood as a mere preprocessing step for recognition, but both need to be handled in a common framework.

In earlier studies of eye movements in natural scenes, prediction by saliency maps could often be partly attributed to generic spatial biases in fixation and saliency. Human photographers typically center objects in images (Figure 6.1A) and we prefer to look straight ahead, so this double central bias can artificially enhance fixation prediction [83, 127, 129]. Here, we find that the influence of such double biases is substantial, but does not fully explain the observed relations (Figure 9.3, blue). Furthermore, the bias must be represented in the brain and have adapted to stimulus statistics. Hence, even a stronger bias than the one observed would not invalidate the conclusions regarding the neural computation of attention guidance.

We do not find task dependence of saliency's predictive power for fixation. Two differences from previous studies may be responsible: first, our targets are verbally defined preventing observers from knowing their features in advance; second, the locations of our targets are difficult to predict from context, which plays an important role in search [135]. So our results do not necessarily conflict with these studies. The effect of observer idiosyncrasies (*e.g.*, memories or cognitive preferences) is low for our stimuli and tasks, as reflected by the high inter-observer consistency in mutual fixation prediction (88.9% AUC). It is well conceivable that this number, which bounds the possible performance of bottom-up models, may drop substantially for different tasks or stimuli. This, however, would only strengthen the conclusion that higher sites are important for driving attention. We stress that the interaction of top-down and bottom-up is not topic of the present study. Instead, we focus on the bottom-up aspect, in evaluating the relation between early saliency and object saliency.

In any study of overt attention or object recognition, stimulus choice is critical. Stimulus category influences the prediction performance of saliency maps and other attention models [20,97,99,103,104]. For our photographs of complex everyday scenes, fixation prediction (58% AUC) is within the range of similar paradigms, which extends from 53% for the foliage images of Einhauser *et al.* [23] to the 68% of Peters *et al.* [99]. In terms of relating saliency maps to fixations, our images are typical.

The result that interesting objects are often accompanied by high saliency values was independently observed in a recent analysis [28] with a complementary approach: while we used a controlled setting, these authors used the large LabelMe database annotated remotely by often unknown observers [115]. The fact that Elazary and Itti arrive at similar conclusions about the prediction of objects by saliency maps supports that this finding is not a consequence of our specific setting, tasks, or image material. Our data confirm the findings of Elazary and Itti on the relation between saliency and object naming in a controlled subject population and adds the direct measurement of fixations. However, neither our data nor Elazary and Itti's prove a causal link between saliency and object recall. Saliency might merely be a correlate instead of a guiding principle of where objects are in natural scenes. The extent to which low-level features, such as those of the saliency model, guide object recall remains an interesting issue for further research.

Object saliency's prediction of object recall suggests that models of attention may characterize object properties in natural scenes. This opens several further lines of research. First, can attention models not only predict free recall, but also recognition performance under difficult conditions? Evidence suggests that a Bayesian model of surprising events not only predicts attention allocation [57], but also predicts human errors during natural scene recognition [24]. Second, can we adapt low-level models of object recognition to predict attention allocation? Walther *et al*. [147] have proposed an architecture that shares features between attention and recognition. Third, can manipulating scene statistics dissociate attention and recognition? While these questions are beyond the scope of this paper, our data indicate that investigating the coupling of attention and recognition will be fruitful for understanding human vision under natural conditions and for modeling attention and recognition in real-world scenes.

Although frequently named objects are generally more fixated and more salient, the number of fixations on an object shows a larger variation for frequently named objects than for rarely named ones. In addition, if only one observer recalls a particular object, she has a slight tendency for a larger fraction of fixations on that object. Since we ask for keywords, we may have biased observers to name scene-diagnostic objects. It is therefore possible that rarely named objects could still be remembered, if they were specifically queried. In this view, less expected objects need more fixations, or more salience, to be named. This is in line with the idea that surprising or unexpected events draw attention, whether they deviate statistically [57] or semantically [42]. Indeed, implausible objects (*i.e.*, objects that conflict with scene gist) tend to be recalled better [100], although they are recognized worse [12] and their effective field of view is smaller [13]. Whether semantically implausible objects are fixated earlier or even pop out [78] has remained controversial. Recent studies that use more complex scenes than Loftus and Mackworth [78] and control the saliency of the critical item, typically do not find an early preference to fixate implausible objects. Instead, they find that implausible objects are fixated longer, more likely to be fixated again [49], and are fixated earlier than plausible objects only after prolonged viewing [140] or if they appear while saccadic suppression suppresses bottom-up attention capture [7]. Under some experimental conditions the recall of an item is improved by increased numbers of fixations on the object [55], although this effect can be restricted to certain aspects of the item and depend on query methods [130]. The effect different object properties (saliency, object properties, fixation frequency, naming frequency, *etc.*) have on the ability of an observer to recall an item when queried will be an interesting issue for further investigation. The diversity of findings stresses that the querying for keywords in the present study and the unknown motivation of LabelMe participants in Elazary and Itti [28] may yield substantially different results from other tasks, such as change detection or item recall.

In conclusion, we provide evidence that interesting objects, not early features, guide human attention. In this view, saliency maps model the saliency of an object instead of the saliency of a location. We stress that this does not deny the usefulness of saliency maps. On the contrary: saliency maps are reinterpreted as unsupervised models of characteristic objects in a scene, irrespective of whether their features causally drive object recall. Some high-level scene interpretation is rapidly available to the visual system [75, 114, 132], potentially faster or with less effort than low-level concepts [53, 75]. With the present data, this suggests another interesting speculation: eye movements or spatial attention are by-products of object-based attention or object recognition.

Part III

Generalized Object Detection

Category-independent object detection, or detecting objects without explicit class labels, could facilitate unsupervised learning about novel objects [117], enable smarter image retargeting, and increase speed in recognizing objects in scenes. Also, there is evidence that humans must separate an object from its surround prior to recognizing it (*e.g.*, Where's Waldo?) [33, 74]. In 2010, alternative algorithms for category-independent object detection were proposed by Alexe *et al.* [2] and Endres and Hoiem [29].

Category-independent object detection could determine which regions are likely object candidates so that expensive operations are not wasted on obviously non-object regions. Object recognition approaches founded on sliding windows [143] or segmentation [46] could directly benefit from this. Specifically, category-independent object detection could reduce false positives and increase detection speed.

There is currently no appropriate benchmark to compare category-independent object detection methods. Challenging well-designed benchmarks have a history of driving progress in computer vision. The Berkeley Segmentation Dataset [86], the Caltech 101 [35] and PASCAL VOC [31] object categorization datasets, and the Caltech Pedestrian Dataset [16], helped direct research efforts toward successful approaches in their fields. In this vein, our aim is to provide a suitable benchmark so researchers can compare performance on appropriately annotated, challenging images. The images in this dataset are rich scenes captured by a professional photographer and are not biased, since they are collected independently of machine vision research.

Chapter 10

Previous Work

The layout of the paper is as follows: Section 10.1 discusses previous work toward the goal of detecting objects independent of category membership. Section 10.2 describes the shortcomings of the datasets used for evaluation in the previous work. Section 11.1 describes the Shore dataset that we will release to address these shortcomings. Section 11.2 compares the metrics that have been used to evaluate category-independent object detection. Section 11.3 evaluates the major approaches toward category-independent object recognition on both the Shore and PASCAL VOC datasets. Section 11.4 discusses our findings.

10.1 Methods

Work on category-independent object detection has focused on detecting either a single object per image or multiple objects per image. Single-object approaches are relevant, but unsuitable for our multiple-object images. Segmentation plays a key role in both the single-object and multiple-object cases.

Detecting a Single Object

Liu *et al*. focused on the problem of localizing a single salient object in a photo [77]. They combined a collection of features, such as RGB chi-square distance and multi-scale contrast. They mentioned applying inhibition of return to detect multiple salient objects, but provided neither method nor analysis.

	Alexe [2]	Endres [29]	Hou [56]
Saliency	○		○
Segmentation	○	○	
Boundaries	○	○	
Color distance	○	○	
Texture distance		○	
Occlusion		○	

Table 10.1: Comparison of visual cues used by Alexe *et al.*, Endres and Hoiem, and Hou and Zhang.

Kim and Torralba considered the related problem of finding regions of interest (ROIs) in cluttered, unlabeled web images [65]. They iteratively chose exemplars across a dataset and refined ROIs with respect to the exemplar set. The authors assumed that each photo had exactly one ROI and that categories were repeated frequently across the dataset. These assumptions preclude detecting multiple objects or unseen object categories in an image.

Detecting Multiple Objects

Ma and Zhang took the fuzzy grown connected components of a contrast-based saliency map to be detections [81]. The results were judged qualitatively without comparison to baselines or other methods.

Rutishauser *et al*. segmented a region in a feature map with adaptive thresholding [117]. They evaluated recognition instead of detection.

Hou and Zhang defined a spectral residual saliency map [56]. Drawing inspiration from the scale invariance observation in natural image statistics, Hou and Zhang equated spectral residual saliency with bumpiness of the log Fourier spectrum. They thresholded a saliency map and took connected components to be detections. Therefore, this method cannot detect large, hierarchical, or overlapping objects (*e.g.*, Figure 10.1).

Alexe *et al*. formulated the problem as measuring how likely it is for an arbitrary bounding box to cover an object of any class [2]. To calculate an objectness score

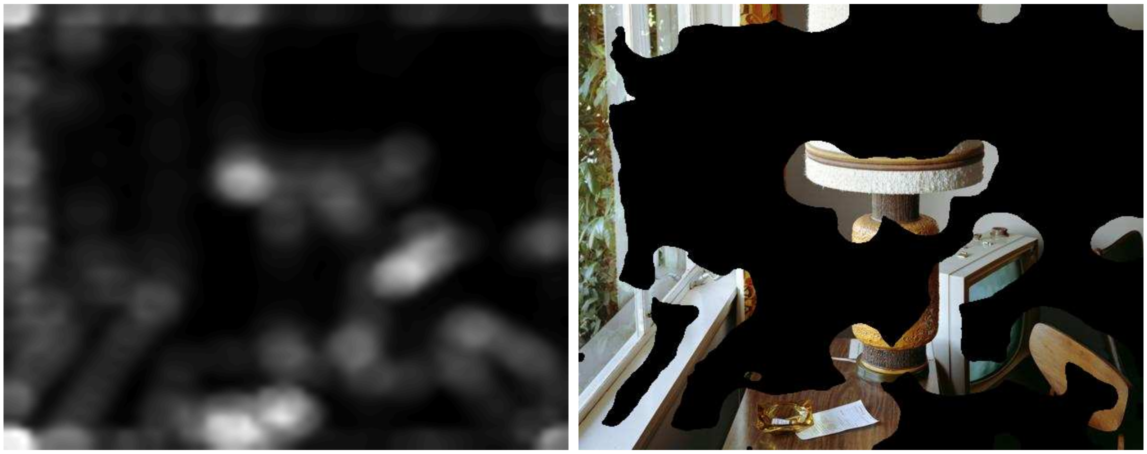


Figure 10.1: The spectral residual saliency map (left) results in candidate object regions (right) when thresholded.

they combined spectral residual saliency, LAB chi-square distance, and Canny edge density near the detection boundary. However, the feature that they found most useful was superpixel straddling, a measure of how well a bounding box aligned with superpixels [110]. They used training data to learn the eight parameters of their cues, which in turn produced an objectness value for a bounding box.

Endres and Hoiem [29] approached the category-independent object detection problem with two steps. First, they generated a set of segments grounded in hierarchical segmentation. Second, they ranked the segments aiming to rank one clean segment of each object object highly. Features they used for the ranking include color chi-square distance, texture chi-square distance, boundary strength, and occlusion information. They use training data to both propose and rank objects.

We evaluated three of the methods described in this subsection, but not the fourth, Ma and Zhang [81], as it is nearly redundant with Hou and Zhang [56]. Table 10.1 compares the visual cues used by the methods we evaluated.

Segmentation as the Crux

In the methods reviewed earlier, segmentation provides important information. Here, we explain how.

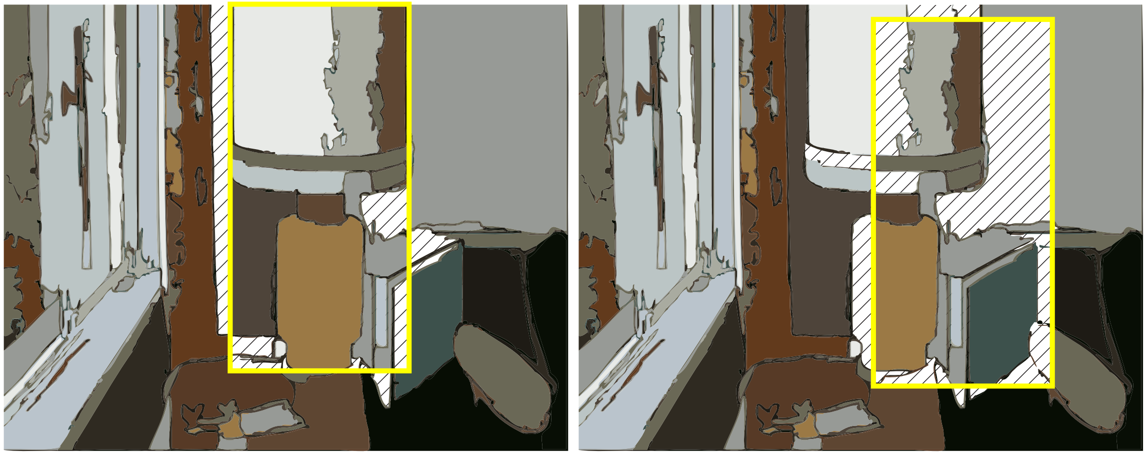


Figure 10.2: The box aligned with the lamp disconnects fewer pixels from their segments, reducing the striped area.

Kim and Torralba observed that superpixels tend to be contained within objects instead of crossing boundaries. This indicates that bounding boxes should contain entire segments. Starting with an oversegmentation, they generated detections as minimum bounding boxes on segment combinations.

Alexe *et al.* used segmentation to evaluate instead of to propose bounding boxes. They oversegmented an image with Felzenszwalb and Huttenlocher segmentation [37], and reasoned that a good window should not break superpixels in half. Figure 10.2 shows two examples of this measure in action. A well-aligned window results in less orphaned area (the smaller portion of a broken superpixel). However, this treats object and background segments equivalently.

Endres and Hoiem proposed regions as likely object candidates based on segmentation. Each detection was grown by appending superpixels likely to belong to the same object as a subregion of the image assumed to be foreground.

10.2 Datasets

The methods discussed in the last section were all tested on different datasets, so their relative performance is unknown. Here, we discuss why these datasets are inappropriate for evaluating category-independent object detection. Figure 10.3 displays

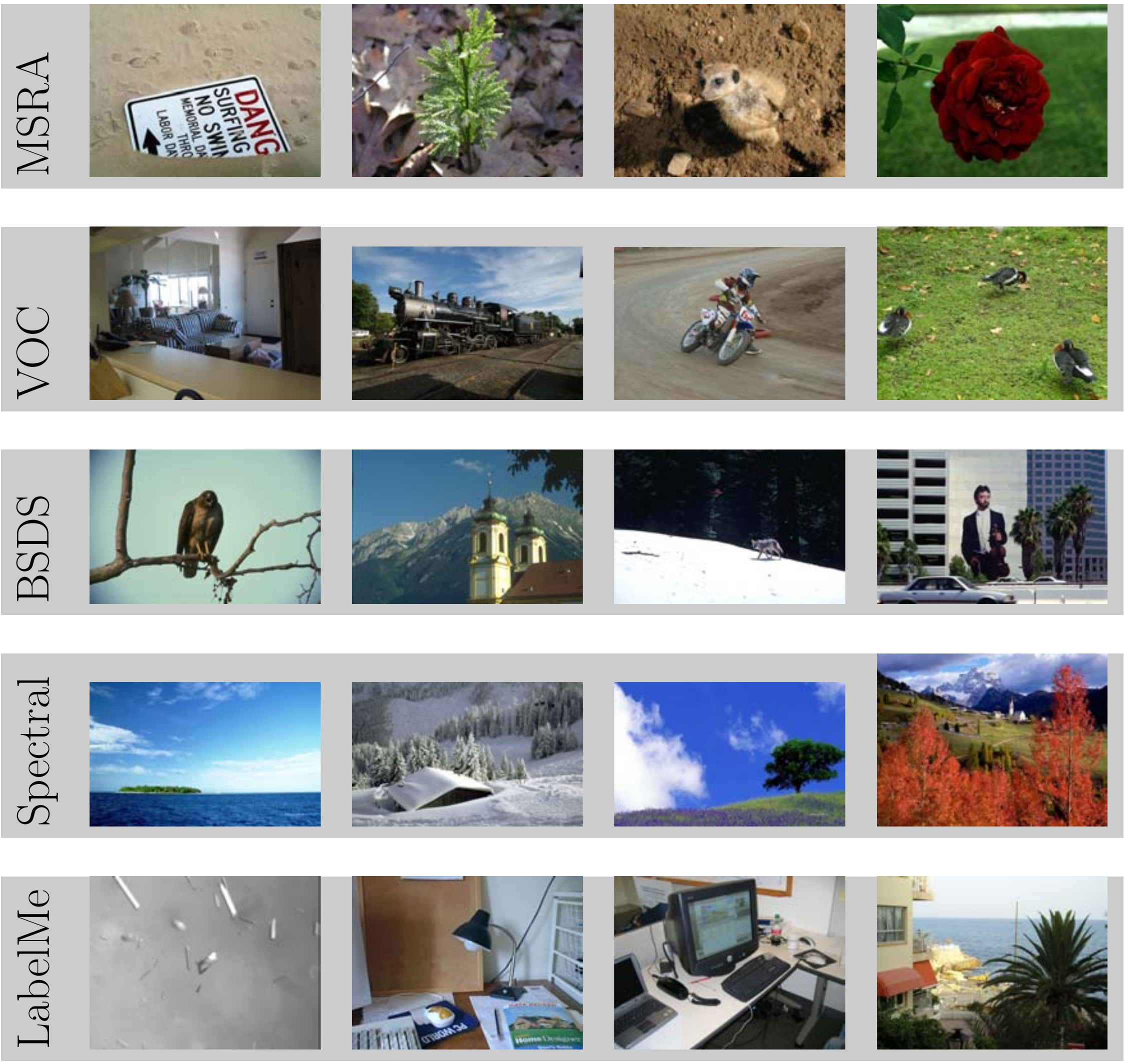


Figure 10.3: Comparison of image datasets used for category-independent object detection. Our dataset is shown in Figure 11.1.

Dataset	Images	Objects per Im.	Image max side	Annotators per Im.	Categories	Used by
MSRA	20,840	1	400	3	All	Liu *et al.* [77]
VOC2007 Test	4,952	2.4	500	1+1	20	Alexe *et al.* [2], Kim and Torralba [65]
VOC2008 Seg.Val.	512	2.3	500	1+1	20+2	Endres and Hoiem [29]
BSDS Test	100	2.2	481	3+	All	Endres and Hoiem [29]
Spectral	62	1.6	256-1280	4	All	Hou [56]
LabelMe	30,369	3.3	Any	1+	All, parts	No relevant work
Shore	536	27.0	800 x 600	25	All, parts	New Benchmark
VOC2007 MTurk	200	24.8	500	25	All, parts	New Annotation

Table 10.2: Existing datasets contain smaller images with one or two objects annotated per image. LabelMe is a dynamic dataset so numbers are from a specific date [115].

example images, and Table 10.2 contains the statistics from all the datasets discussed in this section.

The **MSRA** Salient Object dataset addresses the problem of detecting a single salient object in a photo [77]. So, the images generally contain a single annotated object.

The **PASCAL VOC** image set [31] consists of 20 object categories labeled over a large collection of photos. Each photo was annotated by one human and checked by another. Objects of categories other than the chosen 20 are not annotated; we find that the annotated objects only account for one-tenth of the visible objects (Table 10.2).

The **PASCAL VOC Segmentation** Taster marks objects in a subset of VOC images pixel-wise instead of by bounding box. As with the standard VOC image set, object categories other than a select 20 are unmarked. Additionally, since pixels belong to a single-object, parts cannot be objects.

The **Berkeley Segmentation Dataset (BSDS)** [86] enabled quantitative evaluation of segmentation and boundary detection. Objects are not annotated in the dataset, so Endres and Hoiem added annotations that agreed with the original boundaries. The annotations were generated by the researchers without a formal procedure to distinguish objects from non-objects.

The **Spectral Residual** dataset is a small set of images with annotations summed across viewers [56]. Hou and Zhang motivated their method as exploring the properties of backgrounds. Correspondingly, these images have large homogeneous backgrounds and small non-overlapping objects.

LabelMe allows anyone to upload photos or annotate objects of any type with a polygon and a keyword [115]. However, there are possibilities of bias and inconsistency: Machine vision scientists upload images for particular tasks. There is no redundancy in annotation, and annotators provide quality control themselves.

Chapter 11

Benchmark

11.1 New Dataset

As existing datasets have simple images, small images, limited categories of annotation, or inconsistent annotations, we introduce a new dataset. We describe our method of collecting bounding boxes and our high-resolution, object-rich image set, which was not collected for category recognition.

Annotation Collection

To generate a dataset of category-agnostic objects requires discussion of what an object is. Alexe *et al*. give this definition of an object:

> Objects are standalone things with a well-defined boundary and center, such as cows, cars, and telephones, as opposed to amorphous background stuff, such as sky, grass, and road.

This is a clear, straightforward operational definition. Unfortunately, the human concept of an object is inherently flexible such that parts are often valid objects. Is the door of a house an object, as well as a part? What about a nose? At what scale? Alexe *et al*. avoid this issue, because none of the VOC objects are parts of other VOC objects.

Another issue is that while most objects have well-defined physical boundaries, they do not always have well-defined image boundaries, as anyone working in segmentation knows far too well. Also, people will spontaneously group a V formation of geese into an object, although it lacks physical boundaries.

Our solution is pragmatic: ask users to provide bounding boxes for objects in scene photographs without telling them what an object is. Humans readily and consistently carry out this task without a definition of object. Creating a suitable benchmark may help clarify the task of category-independent object detection. We believe that generic object detection is as well-defined as other vision tasks such as boundary detection, texture classification, and shape recovery. Traditionally machine vision has studied many visual tasks that are not goal-directed.

For each image, we collected up to ten object bounding boxes each from 25 Amazon Mechanical Turk workers. Amazon Mechanical Turk workers were instructed:

> Draw a tight-fitting rectangle around each object you see. You may stop at **10** or when you run out of objects, whichever comes first. If you skip objects and provide fewer than **10** rectangles, your HIT will be rejected. Big objects, small objects, and parts of objects are valid.

We encouraged viewers to annotate many objects, but we used consistent boxes only. A *consistent* box is one that matches a box from another viewer as determined by the PASCAL VOC criterion: the overlap ratio a_o between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 0.5 (50%) by the formula

$$a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{11.1}$$

We found 87% of provided boxes to be consistent after removing three spammers.

Viewers received \$0.10 per image for accepted work. The image content determines whether a single box is an adequate description. Hence, the bar for acceptance was half as many consistent boxes as the bottom quartile for that image. So if 75% of viewers produce at least three consistent boxes for an image, then acceptance requires two consistent boxes.

Ground truth boxes were generated in the following way: All the consistent boxes for an image were clustered with a single linkage cutoff of $a_o = 0.5$. A median box

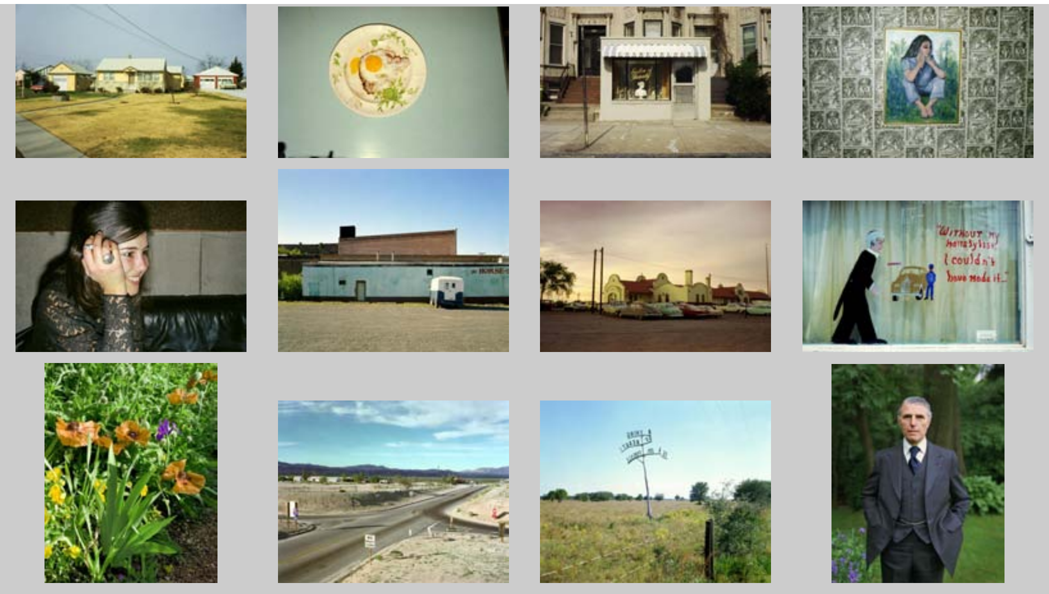


Figure 11.1: Random sample of our images. These photos by artist Stephen Shore are a sampling of experiences rather than images gathered for segmentation or object recognition.

represented each cluster.

Images

We ran our protocol on a new dataset as well as 200 images randomly sampled from VOC2007 training. We introduce the VOC2007 images to verify that the results are not specific to the Shore images, and that many object categories are not in the 20 categories. Annotation for all images can be found in the supplementary material.

Our collection of 536 photographs comprises Stephen Shore's collections "American Surfaces" and "Uncommon Places" [120, 121]. Shore created a visual diary of his experience traveling across America in the 1970s. The images in this dataset are object-rich and large: maximum height of 600 or width of 800 pixels. While there are fewer images than the VOC2007 test, there are ten times as many objects consistently annotated per image, so the object count is equivalent.

Table 11.1 shows how these images divide into scene categories. Figure 11.1 displays a random sample of these photos. We used all the images except for six that

Category	Counts
City	49
Closeup other	42
Food	27
Indoor	70
Nature	52
Outdoor other	30
Person	81
Picture	21
Storefront	35
Street	54
Suburb	75
All	536

Table 11.1: The Shore dataset contains a rich variety of images.

might be considered offensive: three cadavers and three nudes.

Stephen Shore explained his sampling method:

> I was photographing almost every meal I ate, every person I met, every waiter or waitress who served me, every bed I slept in, every toilet I peed in. But also, I was photographing streets I was driving through, buildings I would see.

These photos are particularly well suited for category-independent object detection because they depict many objects and scenes. Furthermore, Shore's goals are removed from the biases of computer vision.

11.2 Metrics

Approaches to category-independent object detection have been evaluated under different metrics, on different datasets. Here, we describe these metrics and analyze their suitability.

The widespread **PASCAL criterion** (Equation 11.1) defines a match in these evaluations. When multiple detections match one ground truth object, the situation becomes more complex than counting matches. Detection benchmarks often constrain this problem by specifying that each ground truth bounding box may be matched at most once, starting with highest confidence first [16, 31]. Repeat detections count as false positives, encouraging methods to perform non-maximal suppression.

We wish to compute the maximum number of objects and detections that can be matched one-to-one. The minimum between matched objects and matched detections provides an upper bound on one-to-one matches. This upper bound is only different from the true value in pathological situations: multiple objects must align well enough to match the same detection, but poorly enough so that multiple other detections will not match enough of the objects. We ignore this rare case.

So **precision** is the fraction of detections that match unique objects, while **recall** is the fraction of objects that match detections one-to-one. As detections are usually more bountiful than objects, recall is unlikely to be less than the fraction of detected objects after the first few detections.

A different approach, taken by Alexe *et al.*, allows multiple matches and defines signal to noise and detection rate. **Signal to noise** is the fraction of detections covering ground truth objects. **Detection rate** is the fraction of ground truth objects covered by detections. So signal to noise and detection rate are analogous to precision and recall.

However, signal to noise rewards repeat detections, resulting in bizarre situations. Correctly finding one object out of dozens can result in detection rate $= 1$ and signal to noise ≥ 0.5. Uniformly sampling the entire image, interleaving the known object location with the sampled windows, yields this result. The more often we insert the known detection, the higher the signal to noise will become. Repeat detections should not be rewarded, but whether they should be punished or ignored depends on the application. Alexe *et al.* apply this metric to uniformly sampled windows only. Pathological situations are prevented if the sampling is sparse enough, but signal to noise cannot be used generally.

The approach taken by Endres and Hoiem compares recall with the mean number of **detections per image**. If we want to detect as many objects as possible and don't differentiate between repeat detections and false detections (*i.e.*, punish repeat detections), this is an intuitive measure. However, if larger images tend to contain more objects, this measure is not invariant to image size.

Another way we can look at category-independent object detection is to consider how much object importance has been recalled (Chapter 3).

The **Average Precision** is a summary statistic that approximates the area under the precision-recall curve. The VOC2007 measures this as the mean precision at recall $[0, 0.1, ...1]$ [31]. Another alternative is the F-Measure or the harmonic mean of precision and recall.

11.3 Benchmarking Previous Methods

We evaluate generic object detectors on the new Shore database and PASCAL VOC2007 [31]. We compare these methods with several baselines and an upper bound. Ground truth consists of bounding boxes around 15-30 objects per image and 20 object categories.

Algorithms

The **Segment Baseline (BL)** proposes bounding boxes from segments alone. Felzenszwalb and Huttenlocher [37] segments and pairs of any two segments define minimum bounding boxes. No bounding boxes are allowed to overlap with ones already in the collection by $a_o > 0.8$. This segmentation software works quickly on large images and provides reasonable results on the Shore images. Boxes are scored by summing the values they cover on a Gaussian window the size of the image. Hence, large boxes and central boxes score highly.

The **Segment Upper Bound (UB)** provides an upper bound on segment performance; it is generated by ordering the segment baseline bounding boxes so that n objects are matched by the first n detections. This is the best possible performance with segment pairs.

The **Hou** algorithm thresholds a saliency map and places minimum bounding boxes on connected components.

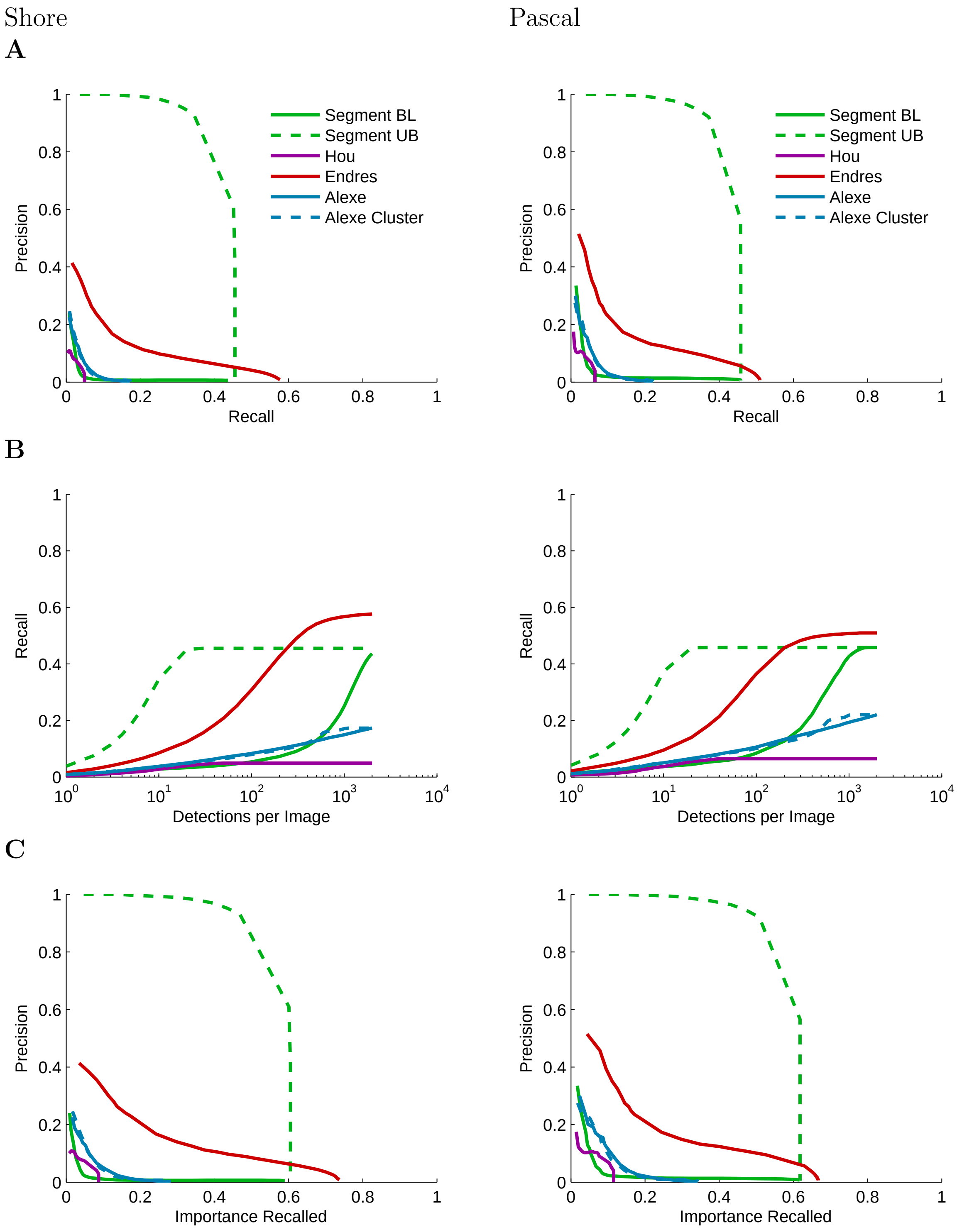


Figure 11.2: Methods evaluated on new benchmark. **A** Precision and recall suggest that while Endres and Hoiem outperform other methods, there is significant room for improvement. **B** Comparison between recall and detections per image indicates that Alexe *et al*. usually perform no better than baseline. **C** Replacing recall with importance recalled gives a measure of how well the prominent objects have been detected.

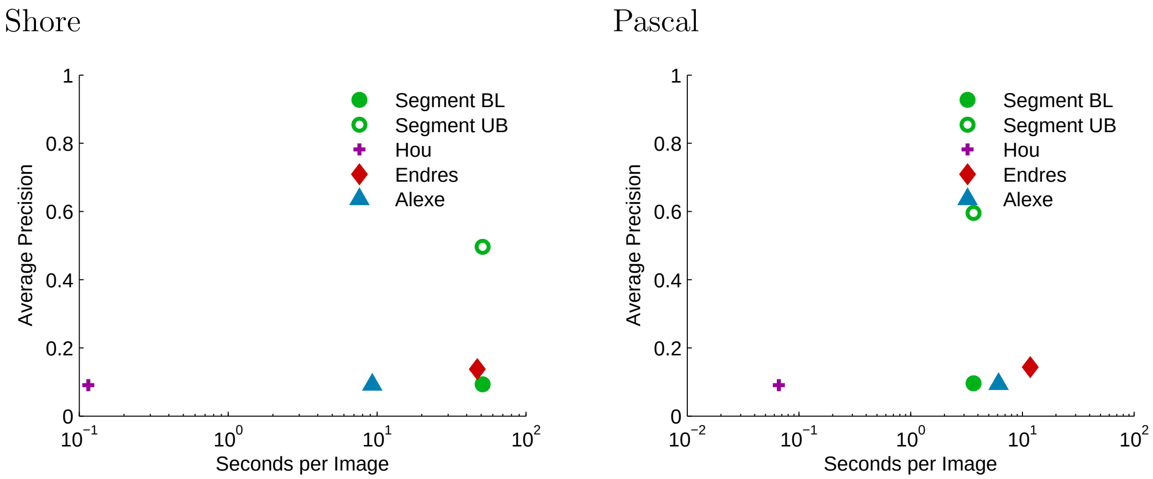


Figure 11.3: Algorithm compute time shows that there is no algorithm that is both fast and precise.

The **Endres** algorithm generates proposal segments and then ranks them. As with the Hou algorithm, we use the tightest bounding box as the ground truth consists of bounding boxes.

The **Alexe** algorithm combines features to measure the objectness of an arbitrary bounding box. We use their sampling method to generate windows.

The **Alexe Cluster** approach reduces repeat-detections. Alexe *et al.* use category-specific scores to perform non-maximal suppression. As we do not have these scores, we combine overlapping detections with complete linkage clustering and a cutoff of $a_o = 0.8$.

Analysis

Figure 11.2A shows the precision-recall curves for one-to-one matching between objects and detections. Category-specific detection results on PASCAL yield high precision in low-recall regimes but zero precision above 0.5 recall [31]. As was found with category-specific detections, category-independent detections have low precision in the high-recall regime. None of the methods reach high precision, even with low-recall. The Endres and Hoiem method dominates others over most recall values. Hou is never better than the baselines. Precision drops rapidly as a function of recall for

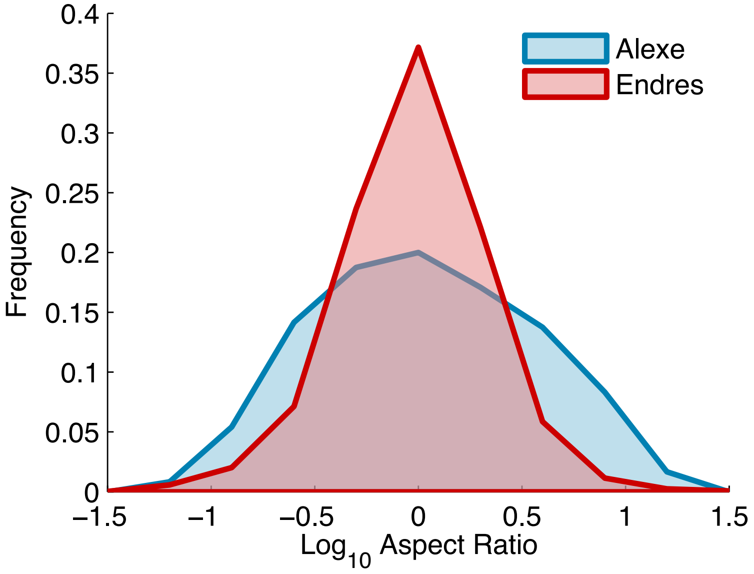


Figure 11.4: Histogram of object aspect ratios that Alexe *et al*. detect but Endres and Hoiem miss (blue) and vice versa (red).

all methods.

Good performance in Figure 11.2B is indicated by fewer detections being able to recall more objects. Currently, hundreds of detections are required to saturate recall. This metric is useful in trying to determine if a module is useful for an object recognition system, as it measures how many detection are needed to reach a certain recall. Precision does not convey that information. Endres and Hoiem actually surpass the segmentation upper bound when more than 200 detections per image are used.

Figure 11.2C replaces recall with object importance recalled where each object has an importance in $[0, 1]$ and they sum to 1 (Chapter 3). We measure these values with order information about the annotations. This measures whether the objects that people attend to have been detected.

Figure 11.2D compares performance against compute time, or the average precision against seconds per image that each method requires. Notably, there is no method that attains a high average precision with a low compute time.

Figure 11.4 shows that the strength of Alexe's method is long, skinny objects whereas Endres and Hoiem's method finds square objects that Alexe *et al.* misses. Endres and Hoiem detect parts, which according to the MTurk annotators are valid objects.

11.4 Predicting Box Importance

The Shore images were divided into three sets: 100 training images, 36 validation images, and 400 test images. There was an additional test set of 200 Pascal images.

We trained and tested boxes generated by the same technique. Each box was labeled with the importance of the closest ground truth box, given a match (Equation 11.1) and otherwise zero. If a technique did not produce many boxes, but the ones that it did were easily ranked by importance, a great ROC was obtained (*e.g.*, Hou in Figure 11.6). Endres and Hoiem do the best job detecting objects (Figure 11.2) and their object candidates are somewhat amenable to importance prediction.

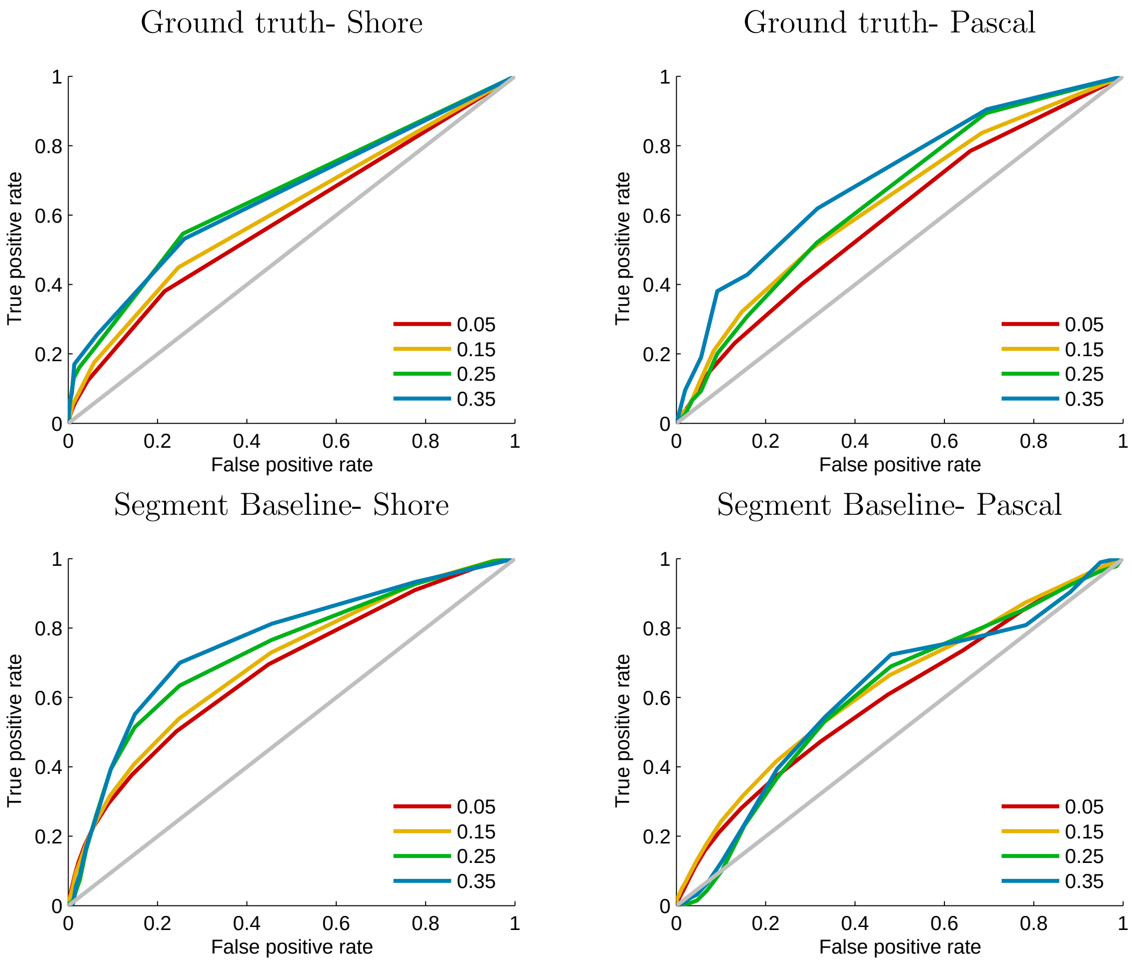
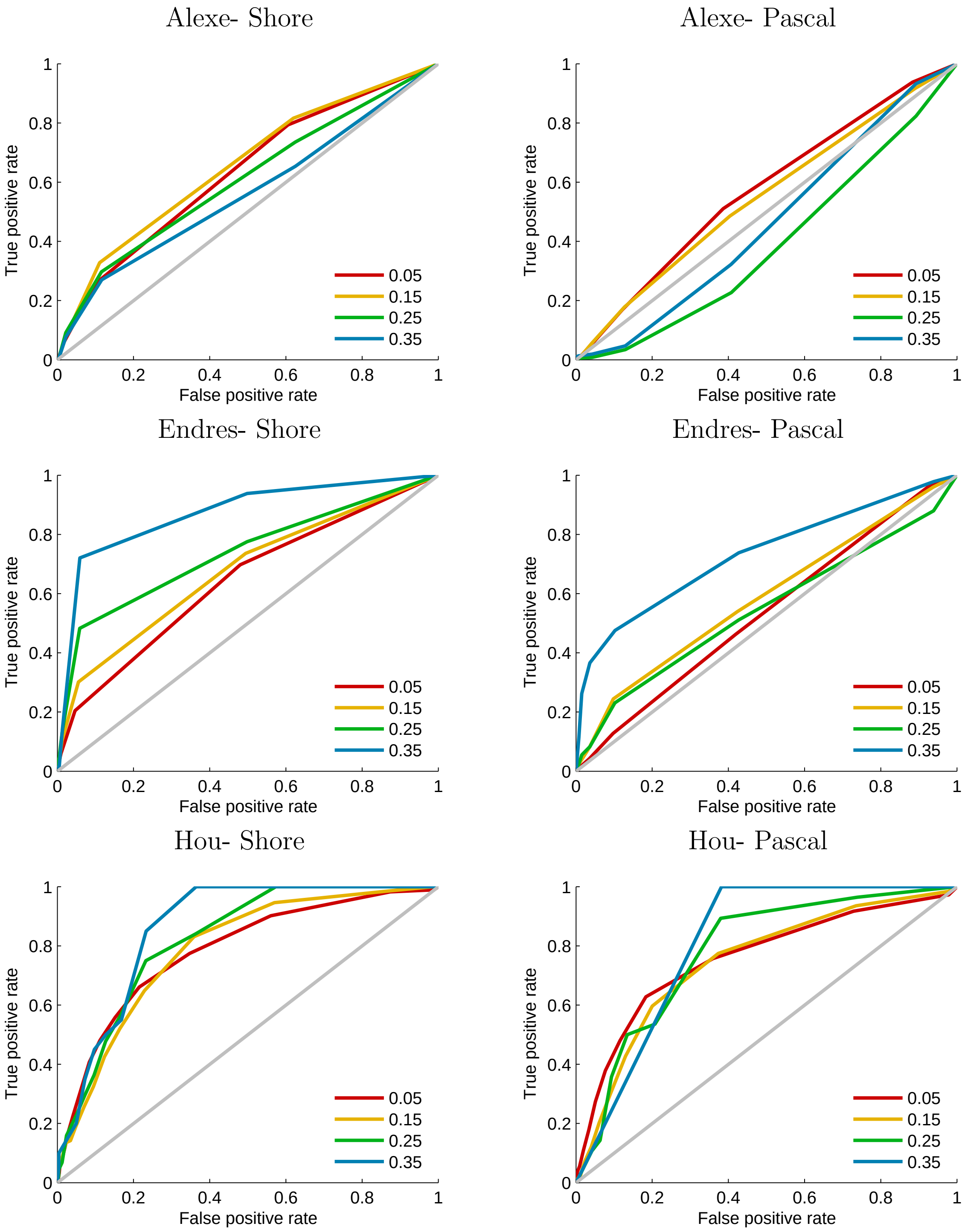


Figure 11.5: Importance prediction on test data.


Alexe- Shore

Alexe- Pascal

Endres- Shore

Endres- Pascal

Hou- Shore

Hou- Pascal

Figure 11.6: Importance prediction on test data continued.

Conclusion

The Shore dataset will provide a challenging test for category-independent object detection. As a 536-scene dataset, selected and annotated for this problem, it is more suitable for testing this problem than any other public dataset. This dataset is so challenging that the best performer finds a new object only once every ten detections.

Endres and Hoiem outperform other methods. This is true for both datasets and all metrics.

An ideal detector would be both fast and accurate. Current methods are fast or somewhat accurate, but not both. Endres and Hoiem outperform Alexe *et al.* on these datasets, but the latter is much faster than the former (Figure 11.2). Alexe *et al.* proved the usefulness of an imperfect detector in greatly reducing the number of windows considered in an object recognition task.

Segmentation is a driving force in category-independent object detection. While state-of-the-art approaches harness segmentation, segmentation is not the same thing as category-independent object detection. Category-independent object detection is the unknown process that boosts segment baseline performance to segment upper bound performance (Figure 11.2).

Bibliography

[1] Wordnet.

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

[3] K. Armstrong, J. Fitzgerald, and T. Moore. Changes in visual receptive fields with microstimulation of frontal cortex. *Neuron*, 50:791–798, 2006.

[4] N. Bichot, A. Rossi, and R. Desimone. Parallel and serial neural mechanisms for visual search in macaque area v4. *Science*, 308:529–534, 2005.

[5] D. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

[6] D. E. Broadbent. The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, 44:51–55, 1954.

[7] J. R. Brockmole and J. M. Henderson. Prioritizing new objects for eye fixation in real-world scenes: Effects of object-scene consistency. *Visual Cognition*, 16:375–390, 2008.

[8] G. Buswell. *How people look at pictures. A study of the psychology of perception in art.* Chicago, IL: The University of Chicago Press, 1935.

[9] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46:4333–4345, 2006.

[10] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing (NIPS)*, 20:241–248, 2008.

[11] F. Cornelissen, E. Peters, and J. Palmer. The eyelink toolbox: Eye tracking with matlab and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers*, 34:613–617, 2002.

[12] J. L. Davenport and M. C. Potter. Scene consistency in object and background perception. *Psychological Science*, 15:559–564, 2004.

[13] P. De Graef. *Eye guidance in reading and scene perception*, chapter Prefixational object perception in scenes: Objects popping out of schemas. 1998.

[14] P. De Graef. *Cognitive processes in eye guidance*, chapter Semantic effects on object selection in real-world scene perception. Oxford: Oxford University Press, 2005.

[15] S. Dickinson, H. Christensen, J. Tsotsos, and G. Olofsson. Active object recognition integrating attention and viewpoint control. *Computer Vision and Image Understanding*, 63:239–260, 1997.

[16] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.

[17] J. Duncan. Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, 113(4):501–517, 1984.

[18] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[19] R. Egly, J. Driver, and R. D. Rafal. Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123:161–177, 1994.

[20] W. Einhäuser, J. Hipp, J. Eggert, E. Körner, and P. König. Learning viewpoint invariant object representations using a temporal coherence principle. *Biol. Cybern*, 93(1):79–90, 2005.

[21] W. Einhäuser, C. Koch, and S. Makeig. The duration of the attentional blink in natural scenes depends on stimulus category. *Vision Research*, 47:597–607, 2007.

[22] W. Einhäuser and P. König. Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, 17(5):1089–1097, 2003.

[23] W. Einhäuser, W. Kruse, K. Hoffmann, and P. König. Differences of monkey and human overt attention under natural conditions. *Vision Research*, 46(8-9):1194–1209, 2006.

[24] W. Einhäuser, T. Mundhenk, P. Baldi, C. Koch, and L. Itti. A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition. *Journal of Vision*, 7(10):1–13, 2007.

[25] W. Einhäuser, U. Rutishauser, E. Frady, S. Nadler, P. König, and C. Koch. The relation of phase-noise and luminance-contrast to overt attention in complex visual stimuli. *Journal of Vision*, 6:1148–1158, 2006.

[26] W. Einhäuser, U. Rutishauser, and C. Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 6, 2008.

[27] W. Einhauser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26, 2008.

[28] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3:3):1–15, Mar 2008.

[29] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010.

[30] K. K. Evans and A. Treisman. Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31:1476–1492, 2005.

[31] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2), 2010.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[33] M. Farah. *Visual Agnosia.* MIT Press, 2nd edition, 2004.

[34] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[35] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4), 2006.

[36] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *J. Vis.*, 7(1):1–29, 1 2007.

[37] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2), 2004.

[38] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[39] J. M. Findlay and I. D. Gilchrist. *Active Vision: The Psychology of Looking and Seeing*. Oxford Universtiy Press, 2003.

[40] A. Fog. Calculation methods for wallenius' noncentral hypergeometric distribution. *Communications in Statictics, Simulation and Computation*, 37(2):258–273, 2008.

[41] C. Fowlkes, D. R. Martin, and J. Malik. Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches. In *CVPR*, pages 54–64, 2003.

[42] A. Friedman. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108(3):316–355, 1979.

[43] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005.

[44] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[45] R. Groner, F. Walder, and M. Groner. *Looking at Faces: Local and Global Aspects of Scanpaths*, pages 523–533. Elsevier Science Publishers B.V., North-Holland, 1984.

[46] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, pages 1030–1037, 2009.

[47] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second edition, February 2009.

[48] J. Henderson, J. Brockmole, M. Castelhano, and M. Mack. *Eye Movement Research: Insights into Mind and Brain*, chapter Visual Saliency does not account for Eye-Movements during Visual Search in Real-World Scenes. Elsevier, 2006.

[49] J. Henderson, P. Weeks Jr., and A. Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1):210–228, 1999.

[50] J. Henderson, C. Williams, M. Castelhano, and R. Falk. Eye movements and picture processing during recognition. *Perception and Psychophysics*, 65(5):725–734, 2003.

[51] O. Hershler and S. Hochstein. At first sight: a high-level pop out effect for faces. *Vision Research*, 45(13):1707–1724, 2005.

[52] O. Hershler and S. Hochstein. With a careful look: Still no low-level confound to face pop-out. *Vision Research*, 46(18):3028–3035, 2006.

[53] S. Hochstein and M. Ahissar. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002.

[54] J. E. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, 57(6):787–795, 1995.

[55] A. Hollingworth and J. M. Henderson. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology - Human Perception and Performance*, 28(1):113–136, 2002.

[56] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[57] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems (NIPS 2005)*, 19:1–8, 2006.

[58] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–506, 2000.

[59] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[60] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(547–579), 1901.

[61] W. James. *Principles of Psychology.* New York, Holt, 1890.

[62] N. Kanwisher. Repetition blindness: type recognition without token individuation. *Cognition*, 27(2):117–143, 1987.

[63] C. Kayser, K. Nielsen, and N. Logothetis. Fixations in natural scenes: Interaction of image structure and image content. *Vision Research*, 46(16):2535–2545, 2006.

[64] M. G. Kendall. *Rank Correlation Methods.* Charles Griffin and Company Limited, 1962.

[65] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *NIPS*, 2009.

[66] S. Kitayama, S. Duffy, T. Kawamura, and J. Larsen. Perceiving an object and its context in different cultures: a cultural look at new look. *Psychol Sci.*, 14(3):201–6, 2003.

[67] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[68] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche. Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, 13:201–214, 2000.

[69] M. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41:3559–3565, 2001.

[70] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[71] G. Lebanon and J. D. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *ICML*, pages 363–370, 2002.

[72] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[73] G. Leech, P. Rayson, and A. Wilson. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman, London, 2001.

[74] D. M. Levi. Crowding – an essential bottleneck for object recognition: a mini-review. *Vision Res.*, 48(5):635–54, 2008.

[75] F. F. Li, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99:9596–9601, 2002.

[76] Z. Li. A saliency map in primary visual cortex. *Trends in Cognitive Science*, 6:9–16, 2002.

[77] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR*, 2007.

[78] G. Loftus and N. Mackworth. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4):565–572, 1978.

[79] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[80] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[81] Y.-F. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*, 2003.

[82] B. F. J. Manly. A model for certain types of selection experiments. *Biometrics*, 30(2):281–294, 1974.

[83] S. Mannan, K. Ruddock, and D. Wooding. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10(3):165–188, 1996.

[84] S. Mannan, K. Ruddock, and D. Wooding. Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26(8):1059–1072, 1997.

[85] S. Mannan, K. Ruddock, and D. Wooding. Fixation sequences made during visual examination of briefly presented 2d images. *Spatial Vision*, 11(2):157–78, 1997.

[86] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, July 2001.

[87] M. Mayer and E. Switkes. Spatial frequency taxonomy of the visual environment. *Investigative Ophthalmology and Visual Science*, 26(280), 1985.

[88] J. Mazer and J. Gallant. Goal-related activity in v4 during free viewing visual search. evidence for a ventral stream visual salience map. *Neuron*, 40:1241–1250, 2003.

[89] B. Mel. Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804, 1997.

[90] N. Moraya. Attention in dichotic listening: Affective cues and the influence of instructions. *The Quarterly Journal of Experimental Psychology*, 11(1):56 – 60, 1959.

[91] K. Nakayama. *Vision: Coding and efficiency*, chapter The iconic bottleneck and the tenuous link between early visual processing and perception. Cambridge University Press, 1990.

[92] V. Navalpakkam and L. Itti. Search goal tunes visual features optimally. *Neuron*, 15(4):605–617, 2007.

[93] U. Neisser and L. Becklen. Selective looking: attending to visually specified events. *Cognitive Psychology*, 7:480–494, 1975.

[94] W. Nelson and G. Loftus. The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, 6(4):391–399, 1980.

[95] T. Ogawa and H. Komatsu. Neuronal dynamics of bottom-up and top-down processes in area v4 of macaque monkeys performing a visual search. *Experimental Brain Research*, 173(1):1–13, 2006.

[96] R. Parker. Picture processing during recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2):284–293, 1978.

[97] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

[98] D. Pelli. The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10:437–442, 1997.

[99] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.

[100] K. Pezdek, T. Whetstone, K. Reynolds, N. Ashkari, and T. Dougherty. Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4):587–595, 1989.

[101] Pixable. Facebook photo trends, April 2011.

[102] M. Pomplun. Saccadic selectivity in complex visual search displays. *Vision Research*, 46:1886–1900, 2006.

[103] C. Privitera, T. Fujita, D. Chernyak, and L. Stark. On the discriminability of hrois, human visually selected regions-of-interest. *Biological Cybernetics*, 93(2):141–152, 2005.

[104] C. Privitera and L. Stark. Alogrithms for defining visual regions-of-interest: Comparision with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.

[105] A. Rabinovich, S. Belongie, T. Lange, and J. M. Buhmann. Model order selection and cue combination for image segmentation. In *CVPR*, pages 1130–1137, 2006.

[106] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, pages 1–8, 2007.

[107] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard. Eye movements in iconic visual search. *Vision Research*, 42:1447–1463, 2002.

[108] J. Raymond, K. Shapiro, and K. Arnell. Temporary suppression of visual processing in an rsvp task: an attentional blink? *Journal of Experimental Psychology. Human Perception and Performance*, 18:849–860, 1992.

[109] P. Reinagel and A. Zador. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10:341–350, 1999.

[110] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, pages 10–17, 2003.

[111] R. Rensink, J. O'Regan, and J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373, 1997.

[112] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12(2):162–168, 2002.

[113] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umilta. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25:31–40, 1987.

[114] G. Rousselet, M. Fabre-Thorpe, and S. Thorpe. Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, 5:629–630, 2002.

[115] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. Technical report, 2005.

[116] B. C. Russell, A. B. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[117] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR*, pages 37–44, 2004.

[118] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on pattern analysis and machine intelligence*, 29(3):411–426, 2007.

[119] R. M. Shiffrin and W. Schneider. Controlled and automatic human information processing: Ii. perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84(2):127–190, 1977.

[120] S. Shore. *Stephen Shore: American Surfaces*. Phaidon Press, 2005.

[121] S. Shore, L. Tillman, and S. Schmidt-Wulffen. *Uncommon Places: The Complete Works*. Aperture, 2005.

[122] D. Simons. Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, 4:147–155, 2000.

[123] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *CVPR*, 2008.

[124] M. Spain and P. Perona. Some objects are more equal than others: measuring and predicting importance. In *ECCV*, 2008.

[125] M. Spain and P. Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 2010.

[126] A. N. Stein, T. S. Stepleton, and M. Hebert. Towards unsupervised whole-object segmentation: Combining automated matting with boundary detection. In *CVPR*, 2008.

[127] B. Tatler, R. Baddeley, and I. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643–659, 2005.

[128] B. Tatler, N. Wade, and K. Kaulard. Examining art: dissociating pattern and perceptual influences on oculomotor behaviour. *Spatial Vision*, 21(1-2):165–184, 2007.

[129] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17, 2007.

[130] B. W. Tatler, I. D. Gilchrist, and M. F. Land. Visual memory for objects in natural scenes: From fixations to object files. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, 58(5):931–960, 2005.

[131] K. Thompson and N. Bichot. A visual salience map in the primate frontal eye field. *Progress in Brain Research*, 147:251–262, 2005.

[132] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

[133] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

[134] S. Tipper. The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology*, 37A:571–590, 1985.

[135] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113:766–786, 2006.

[136] A. B. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[137] A. M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[138] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–547, 1995.

[139] G. Underwood, T. Foulsham, E. van Loon, L. Humphreys, and J. Bloyce. Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology*, 18(3):321–343, 2006.

[140] G. Underwood, E. Templeman, L. Lamming, and T. Foulsham. Is attention necessary for object identification? evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, 17(1):159–170, 2008.

[141] R. VanRullen. On second glance: Still no high-level pop-out effect for faces. *Vision Research*, 46(18):3017–3027, 2006.

[142] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001.

[143] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[144] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI*, pages 319–326, 2004.

[145] G. Wallis and E. T. Rolls. A model of invariant object recognition in the visual system. *Progress in Neurobiology*, 51:167–194, 1997.

[146] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[147] D. Walther, T. Serre, T. Poggio, and C. Koch. Modeling feature sharing between object detection and top-down attention. *Journal of Vision*, 5(8), 2005.

[148] A. Yarbus. *Eye movements and vision.* New York: Plenum Press, 1967.

[149] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, pages 2126–2136, 2006.

[150] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3):1–15, 2011.