HHS Public Access

Author manuscript

Int J Comput Vis. Author manuscript; available in PMC 2020 July 27.

Published in final edited form as:

Int J Comput Vis. 2011 January ; 91(1): 59–76. doi:10.1007/s11263-010-0376-0.

Measuring and Predicting Object Importance

Merrielle Spain, **Pietro Perona**

Abstract

How important is a particular object 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 a large number of object-related and image-related features. We validate our importance predictions on 2,841 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.

Keywords

Visual recognition; Object recognition; Importance; Perception; Keywording; Saliency; Rank aggregation; Amazon Mechanical Turk

1 Introduction

After an initial phase focused on detecting individual objects and categories (Lowe 1999, 2004; Viola and Jones 2001, researchers in visual recognition have moved on to detecting objects belonging to multiple classes (Grauman and Darrell 2005; Griffin et al. 2007; Lazebnik et al. 2006; Zhang et al. 2006). Recently, the problem of simultaneously detecting, localizing, and naming multiple objects in an image has become an active area of research (Everingham et al. 2008; Russell et al. 2007).

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 so useful. Indeed, change blindness experiments (Rensink et al. 1997) suggest that after looking at pictures of complex natural scenes, we only retain information about the overall gist of the scene and a handful of objects. The experiments show that we usually 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. Hence, most pictures are about a few important objects. If the goal is computing relationships between objects, the problem becomes more complex: we might face hundreds of irrelevant relationships in the description of a single picture. It is thus useful to select the most ‘meaningful’ of these objects and relationships. We explore whether it is possible to estimate the important objects in a given scene automatically and, as a result, produce a concise list that would facilitate image search and other applications.

M. Spain, 1200 E. California Blvd., MC 136-93, Pasadena, CA 91125, USA, spain@caltech.edu.

We face two main challenges: measuring importance, as perceived by viewers, and automatically predicting the importance of objects in a given image. Figure 1 depicts how these ideas fit together. Section 2 describes how we collect perceived importance information from viewers. Section 3 considers the problem of measuring importance by aggregating data collected from many viewers. Section 4 explains how to predict importance from bottom-up visual properties of an object. We discuss how subtle manipulation of the human task impacts importance in Sect. 5. Section 6 summarizes our main findings.

A preliminary version of this work was published in the proceedings of ECCV 2008 (Spain and Perona 2008). That version contained one of our methods for measuring importance and proposed our model of human object naming. In this work we provide analysis of the human data that justifies our model, a second measure of importance, and a more rigorous approach to predicting importance.

2 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 Sect. 3. For the moment we rely on the intuitive notion and explore ways of assessing which objects people notice most in a photograph.

2.1 Previous Work

Some previous research explores what people can recognize under extreme circumstances. Fei-Fei et al. (2007) examine how limited viewing time affects what viewers report. Torralba et al. (2008) investigate which objects people can name with limited image resolution.

The ESP game, by von Ahn and Dabbish (2004), 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* (Elazary and Itti 2008). In LabelMe (Russell et al. 2005) users name an object and outline its contour with mouse clicks. A user may annotate one or more 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 Sect. 3.3, viewers produce lists with inconsistent object order. Furthermore, the choice to outline an object is influenced by how easy the object is to outline (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 database of pedestrians.

2.2 Data Collection

We designed a method for collecting data on object importance in images with two criteria in mind: (a) the data should be collected independently from a large number of human viewers, (b) our annotators should not be driven by tasks/motivations that bias the data.

We collected ordered lists independently from 25 viewers for each image. Through Amazon Mechanical Turk, US viewers were instructed “Please look carefully at this image and name 10 objects that you see”. We asked for 10 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 all work on Mechanical Turk must be approved by the requester prior to payment. The complete instructions can be found in Appendix A.

Before analyzing the collected lists, we cleaned them in four steps. First, we eliminated empty lists, and lists that clearly 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 the most obvious synonym for each group of words. This step was necessary because the same word could have different 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.

2.3 Image Collection

We selected 97 pictures from Stephen Shore’s collections ‘American Surfaces’ and ‘Uncommon Places’ (Shore 2005; Shore et al. 2005). Shore took these pictures as a visual diary of his experience traveling in North America in the 70’s and 80’s. 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 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 the purpose of the experiment. 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 Mayer and Switkes (1985), by attaching a camera to a bicycle helmet and snapping one picture per minute automatically, the majority of photographs would be meaningless (e.g. the edge of an elevator door). So Shore’s photos are more

[1]See http://wordnet.princeton.edu.

objective than an object recognition dataset and more meaningful than randomly captured photographs.

2.4 Data Overview

Comparing Lists—Examples of 10-object lists produced by 5 viewers are displayed in Table 1. The number of objects that are present in an image may be estimated by considering the size of the union of the twenty-five 10-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 10-word lists vary. To understand the structure of the lists, we compare the lists generated by humans with chance lists. To generate the chance lists we consider the set of objects named in this image and randomly select 10 of them with uniform probability. We generate 25 chance lists per image.

First, we examine a pair of lists (generated by the same process) and count the number of objects that the lists share. Figure 3 shows that pairs of lists from viewers have a much larger intersection of objects than expected by chance (mean of 6.2 as opposed to 4.3).

Second, we look at two lists that share an object, we note the object's rank in each list 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 4 shows that an object's rank changes slightly less between human lists than expected by chance (mean of 2.5 vs. 3.1). For these two histogram comparisons, a Wilcoxon Rank Sum test rejects the null hypothesis that the distributions have the same median ($p = 0, 10^{-111}$).

Third, we look at all the lists for an image and count the number of viewers that name a particular object. Figure 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 per image.

Fourth, we count the number of objects named if we only consider the top k words in each list. Figure 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 number of objects that are associated with an image saturates after the first 4 objects are named, while the number of objects climbs much 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 make another object more or less likely to be named by that viewer? 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. (2007) who ask whether cars and tires appear in the same images. 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 coocurrence is consistent with independent naming. For a given object pair we find all the images that contain both objects and amass all the lists associated with these images. We apply the Pearson's chi-square test with the Bonferonni correction ($p \leq 0.05$/tests) and Yates' correction for continuity to assess the dependence. We only perform a test if each object is present/absent in at least 5 of these lists (4,224 of 15,043 pairs). The value of Pearson's chi-square test-statistic is

$$\chi^2_{Yates} = \frac{(|O_1 - E_1| - 0.5)^2}{E_1} + \frac{(|O_0 - E_0| - 0.5)^2}{E_0}, \quad (1)$$

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 otherwise.

We find that generally one object being named doesn't significantly influence the probability of another object being named. Only 19 of 4,224 tests (0.4%) show significant dependence. Table 2 enumerates the dependent object pairs; for all of these pairs the observed coocurrence is greater than expected coocurrence.

Failure to Name the Obvious—We noticed an interesting phenomenon: viewers sometimes fail to mention the most obvious object (Fig. 7). We identify the obvious object statistically 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 viewers name "person" or "house" very early, but others fail to mention them at all. These two objects account for almost all of the 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 Fig. 8 rules out this hypothesis. The frequency with which the obvious object is not reported does not increase as the viewer labels more images; it is the same on the 20th as it is on the 1st image labeled.

3 Measuring Importance

The observations that most objects are named independently (Sect. 2.4) and some objects are named early and often 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 the process of drawing balls from an urn without replacement (see Fig. 9). The urn contains one ball for each object category in the image. The balls are different sizes, affecting their probability of being chosen. Thus, a ball's size represents the importance of the corresponding object. We represent multiple viewers by repeatedly refilling the urn with the same set of balls and sampling.

This model is based on several assumptions. First, the draws are independent; this is reasonable because very few object pairs are dependent (Sect. 2.4). Second, everyone starts with the same urn; we don't see clusters of different viewer behavior in our data, as we discuss in Sect. 3.3. Third, balls can only be taken out of the urn if they are drawn. This last assumption is violated for some images. As discussed in Sect. 2.4, 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 balls as before, but the first ball may go unreported with a certain probability.[2]

Figure 10 shows the importance measured through Maximum Likelihood (ML), maximizing the likelihood of observing our data with the importance values as parameters (Sect. 3.1.1). 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 much more realistic than the urn's when the object is frequently forgotten.

3.1.1 Fitting the Model—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 (Fog 2008) or requires many balls of the same type (Manly 1974). Instead of using these approaches we measure importance by maximizing the likelihood of our observed data with respect to the object importances. To do this we need to calculate the probability of observing a set of sequences given the object importances $\boldsymbol{\pi}_i$.

Each sequence consists of 10 balls w_i^m, where w_i^m de notes the ith ball drawn in the mth sequence and is a variable that takes values $1, \ldots, N$ corresponding to object names. The w_i^m are drawn independently without replacement (out of N balls, where $N \gg 10$), so the probability of drawing a particular sequence of balls $(w_1^m, \ldots, w_{10}^m)$ is

$$\prod_{n=1}^{10} p(w_n^m \mid w_{n-1}^m, \ldots, w_1^m). \tag{2}$$

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

[2]So the rigorous definition of importance is the probability that a ball is drawn first, regardless of whether it is somehow forgotten.

$$p(w_n^m \mid w_{n-1}^m, \ldots, w_1^m) = \begin{cases} 0 & \text{if } \exists i \in [1, n-1] : w_i^m = w_n^m, \\ \dfrac{\pi_{w_n^m}}{1 - \sum_{i=1}^{n-1} \pi_{w_i^m}} & \text{otherwise,} \end{cases} \tag{3}$$

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

This model fits our observed data well with an exception: viewers sometimes fail to mention the most obvious object (as discussed in Sect. 2.4). Treating this phenomenon rigorously complicates the equations of the model, and the methods for fitting the probability parameters. Luckily, a simple approximation opens the way for an easy treatment: pretending the first ball is forgotten. Consider a sequence of balls where the first ball has been discarded (i.e. really drawn 1st, but considered undrawn); the ball is most likely $\operatorname{argmax}_{j:\, \forall i\, j \neq w_i^m} \pi_j$, the most important of the undrawn balls. In this case, π_j will likely be large. Whereas for a sequence of balls in which the first ball is not dropped, π_j will probably be small. Hence we can include the probability of the largest ball missing from the list, $\max_{\forall i\, j \neq w_i^m} \pi_j$, in the normalization. This results in little change when the first ball is not dropped and a mitigated impact on the probabilities when the first ball is dropped.

$$p(w_n^m \mid w_{n-1}^m, \ldots, w_1) = \frac{\pi_{w_n^m}}{\left(1 - \sum_{i=1}^{n-1} \pi_{w_i^m}\right) - \max_{\forall i\, j \neq w_i^m} \pi_j}. \tag{4}$$

Since we have 25 independent sequences, the likelihood of our observation is

$$p(\text{obs}) = \prod_{m=1}^{25} \prod_{n=1}^{10} \frac{\pi_{w_n^m}}{\left(1 - \sum_{i=1}^{n-1} \pi_{w_i^m}\right) - \max_{\forall i\, j \neq w_i^m} \pi_j}. \tag{5}$$

To measure importance $\pi_{w_i^m}$, we maximize the log-likelihood $\log(p(\text{obs}))$,

$$\sum_{m=1}^{25} \sum_{n=1}^{10} \log \pi_{w_n^m} - \log\left(\left(1 - \sum_{i=1}^{n-1} \pi_{w_i^n}\right) - \max_{\forall i\, j \neq w_i^m} \pi_j\right). \tag{6}$$

We can wonder if our 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. In order 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 ball to second place every

time it is drawn first. In our simulations this change does not decrease the estimated importance of this ball (Wilcoxon Rank Sum Test).

Optimization Note: There are as many parameters as objects mentioned. This number can get large, which results in poor convergence. However if we limit our optimization to the 10 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 100).

3.2 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 M. The system moves from state i to state j with probability M_{ij}. Reasonably behaved Markov chains eventually reach the stationary distribution, a unique fixed point where 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. (2001) 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 $\tau_1, \ldots, \tau_{25}$ containing i, then picking an object uniformly from the set of all objects j such that $\tau(j) \leq \tau(i)$.

Figure 11 gives an example of how the Markov Chain might act for the data in Fig. 9. Our intuition as to why the stationary distribution should approximate the 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 12 compares the MC importance with the forgetful urn ML importance. The right column in Fig. 10 shows the importance measured with the MC. The results are similar to the ML forgetful urn, except that the MC slightly underestimates the importance of objects that have a true importance of ≥0.3 in synthetic data.

3.3 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.

$$D(\sigma, \tau) = \sum_i |\sigma(i) - \tau(i)| . \quad (7)$$

This distance has already been applied in machine learning (Dwork et al. 2001; Lebanon and Lafferty 2002) to compare ranked lists.[3] However, since we want to penalize list pairs that share few items we need a different generalization to partial orderings than Dwork et al. (2001) 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 13 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.

We could assume that the human sequences cluster and if we select the one that is most similar to our held out sequence in the first k objects named, then this would improve our results. For 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 13 shows that the closest human doesn't become better than other methods as more objects are revealed, indicating that no substantial clustering exists.

Is the complexity of the ML or MC methods justified? One could estimate importance more simply by using the frequency with which words appear in the 25 lists, or perhaps the median rank that they have in the lists. We implemented such methods and compared them with the ML and MC. Figure 14 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.

4 Predicting Importance

Would it be possible to predict the importance of each 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 a number of 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 46 possible features, we select a small subset via regularized regression to maximize both the performance and interpretability of our model.

4.1 Object Outlines

Computing object importance requires that the image is 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 (Fowlkes et al. 2003;

[3]Kendall (1962) says that Spearman replaced the absolute value with the square.

Rabinovich et al. 2006) are not as detailed as the verbal responses. Figure 15 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, until automated segmentation reaches a sufficient level of performance, we have our images segmented by hand. We again use Mechanical Turk, but this time we ask 3 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 (2008). We generalize the common segmentation metric |intersection|/|union| (Stein et al. 2008; Everingham et al. 2008, or the Jaccard index (Jaccard 1901)), to evaluate the quality of these human segmentations. Our generalization of the criterion to three annotations is to compare the maximum of the 3 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 the accepted outlines belong to the object in our combined segmentation. In this way we obtain outlines for 2,841 named objects.

4.2 Features

We devise features to convey information about the photo's composition. Hopefully these features capture what makes a particular object important in a particular image. A more detailed description can be found in Appendix B.

First, we consider how to describe an object's position in the image. Figure 16 shows the distribution of objects over the photo. We take all the object masks (pixels are 1 if they contain the object, 0 otherwise) for all the images and sum them, creating an object map (Einhauser et al. 2008). We notice that the object map has a vertical symmetry axis, so we treat distances to the left and right of the midline the same. 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 the center point, horizontal midline, vertical midline, and 4 points that divide the image into thirds. We do this in order to produce features that encode where the object is in the image.

Second, we include an estimate of where people look. We use a Saliency Map (Itti et al. 1998) which is 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 (Walther and Koch 2006) to produce Saliency and Conspicuity Maps. We use the maximum, mean, and sum of Saliency or Conspicuity across the object. We also calculate the same values after modulating the Saliency map by multiplying it by a Gaussian window ($\sigma = 0.4$) to create a central bias.

Third, we consider an object's size; we use its area, log(area), and rank in terms of area.

Fourth, we consider what it overlaps with. How many objects overlap with the object and whether faces overlap with the object.

4.3 Regression

We approximate the function from features to importance as:

$$\log(\text{importance}) = \beta_0 + \sum_{j=1}^{p} (x_j \beta_j) \tag{8}$$

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 $\sum_{j=1}^{p} |\beta_j| \leq t$ in particular favors sparsity, additionally increasing interpretability (Tibshirani 1996). 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 (Hastie et al. 2009). 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 rather than human generated object lists.

Table 3 shows the chosen features and their coefficients. Of the 46 features, the only 15 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 the centrally biased version and the CMs are available. Area is not selected when log(area) is available.

Figure 18 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 6 objects per image, 2 objects per image, one object per image, and one object every three images respectively. We find that our prediction identifies high importance objects reasonably well; we find the area under the ROC curve to be 0.7, 0.78, 0.82, and 0.9 respectively. Figure 19 shows a scatter plot of the measured importance and normalized predicted 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 20 shows a few examples of our results; predicted importances are normalized so the importance in an image sums to one.

4.4 The Power of Features

One question we can ask is if we eliminate the 1-2-3 most important 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 3 features with the largest magnitude in Lasso. Figure 21 demonstrates that as one feature is excluded, another feature (or two) arise to replace it, indicating that our features are redundant.

Another question is how well a single feature, or only a few, can predict importance. Figure 22 shows that, using Stepwise Regression (adding features greedily), a few features can go a long way.

5 Generative & Discriminative Tasks

Up until now we have been considering the case that the viewer is asked for 10 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.

To compare the lists obtained from viewers performing either the Plain, Generative, or Discriminative tasks we can look at the measured importance values in Fig. 24. We can also compare the feature coefficients for predicted importance in Fig. 23. 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.

6 Conclusions

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; so we can often know that something is important without knowing what it is.

In order to study how humans perceive object importance, we asked a large number of English speaking observers to name objects they saw in photographs of everyday scenes. For each of 97 images, we collected 25 independent 10-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 balls from an urn without replacement. Furthermore some objects tend to be named earlier and more frequently, which we represent as the balls 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 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 my be necessary for 'perfect' prediction.

An unexpected phenomenon we observed was that our viewers sometimes failed to report the most obvious object in their 10-word list. This was very repeatable and had not been previously explored. Our urn model was easily modified to accommodate this phenomenon.

Our experiments show that it is not possible to isolate high importance objects with state of the art automatic object-level image segmentations. Progress in this area clearly has strategic value in machine vision. Semantic analysis of the image may also improve importance prediction.

Acknowledgments

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship, Office of Naval Research grant N00014-06-1-0734, and National Institutes of Health grant R01 DA022777.

Appendix

Appendix A: Instructions

Viewers were given the same detailed instructions for each task. These instructions are shown in Fig. 25. The three tasks differed only in the instructions present on the browser window where they completed the task.

Plain task:

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

Generative task:

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

Discriminative Task:

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

Appendix

Appendix B: Prediction Features

Table 4 is a complete list of the features used to predict importance. Figure 26 illustrates how the feature values were computed. The features fall into four general categories: distances, saliency, area, and overlapping.

Distances

We measure distances from the object mask to important positions in the image. For all distance measures we calculate the maximum, mean, and minimum distance between pixels in the object mask and the position in question. We measure the 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 box defined by the four points that divide the image into thirds.

Saliency

We use a Saliency Map (Itti et al. 1998) which is 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 the component Color CM, Intensities CM, Orientations CM, and well as the Saliency Map, a blurred Saliency Map (convolved with a 5×5 Gaussian window), and a Gaussian modulated Saliency Map (multiplying 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 measure falling under the mask of the object.

Area

We consider an object's size; we use its area, log(area) and rank in terms of area (in ascending and descending order).

Overlapping

We consider what an object overlaps with. The percent of the object that is overlapped by other outlined objects and the number of objects that overlap it pixel-wise. We also run a Viola-Jones face detector and take the output to be a mask of all faces in the image. 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.

References

Dwork C, Kumar R, Naor M, & Sivakumar D (2001). Rank aggregation methods for the web. In WWW (pp. 613–622).

Einhauser W, Spain M, & Perona P (2008). Objects predict fixations better than early saliency. Journal of Vision, 8(14), 1–26. URL: http://journalofvision.org/8/14/18/.

Elazary L, & Itti L (2008). Interesting objects are visually salient. Journal of Vision, 8(3:3), 1–15.

Everingham M, Van Gool L, Williams CKI, Winn J, & Zisserman A (2008). The PASCAL visual object classes challenge 2008 (VOC2008) results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

Fei-Fei L, Iyer A, Koch C, & Perona P (2007). What do we perceive in a glance of a real-world scene? Journal of Vision, 7(1), 1–29. URL: http://journalofvision.org/7/1/10/.

Fog A (2008). Calculation methods for Wallenius' noncentral hyper-geometric distribution. *Communications in Statictics*, Simulation and Computation, 37(2), 258–273.

Fowlkes C, Martin DR, & Malik J (2003). Learning affinity functions for image segmentation: combining patch-based and gradient-based approaches. In CVPR (2) (pp. 54–64).

Grauman K, & Darrell T (2005). The pyramid match kernel: discriminative classification with sets of image features. In ICCV (pp. 1458–1465).

Griffin G, Holub A, & Perona P (2007). Caltech-256 object category dataset (Tech. Rep. 7694). California Institute of Technology. URL: http://authors.library.caltech.edu/7694.

Hastie T, Tibshirani R, & Friedman JH (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed). New York: Springer.

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

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

Kendall MG (1962). Rank correlation methods. Charles Griffin and Company Limited.

Lazebnik S, Schmid C, & Ponce J (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In CVPR (2) (pp. 2169–2178).

Lebanon G, & Lafferty JD (2002). Cranking: combining rankings using conditional probability models on permutations. In ICML (pp. 363–370).

Lowe DG (1999). Object recognition from local scale-invariant features. In ICCV (pp. 1150–1157).

Lowe DG (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110.

Manly BFJ (1974). A model for certain types of selection experiments. Biometrics, 30(2), 281–294.

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

Rabinovich A, Belongie S, Lange T, & Buhmann JM (2006). Model order selection and cue combination for image segmentation. In CVPR (1) (pp. 1130–1137).

Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, & Belongie S (2007). Objects in context In ICCV (pp. 1–8). New York: IEEE.

Rensink RA, O'Regan JK, & Clark JJ (1997). To see or not to see: the need for attention to perceive changes in scenes. Psychological Science, 8, 368–373.

Russell BC, Torralba A, Murphy KP, & Freeman WT (2005). LabelMe: a database and web-based tool for image annotation (Tech. rep.)

Russell BC, Torralba AB, Liu C, Fergus R, & Freeman WT (2007). Object recognition by scene alignment. In NIPS.

Shore S (2005). Stephen Shore: American surfaces. Phaidon Press.

Shore S, Tillman L, & Schmidt-Wulffen S (2005). Uncommon places: the complete works. Aperture

Sorokin A, & Forsyth D (2008). Utility data annotation with amazon mechanical turk. In CVPR.

Spain M, & Perona P (2008). Some objects are more equal than others: measuring and predicting importance. In Proceedings of the European conference on computer vision (ECCV).

Stein AN, Stepleton TS, & Hebert M (2008). Towards unsupervised whole-object segmentation: combining automated matting with boundary detection. In CVPR.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B, 58(1), 267–288.

Torralba AB, Fergus R, & Freeman WT (2008). 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Transactions Pattern Analysis Machine Intelligence, 30(11), 1958–1970.

Viola PA, & Jones MJ (2001). Rapid object detection using a boosted cascade of simple features. In CVPR (1) (pp. 511–518).

von Ahn L, & Dabbish L (2004). Labeling images with a computer game. In CHI (pp. 319–326).

Walther D, & Koch C (2006). Modeling attention to salient proto-objects. Neural Networks, 19(9), 1395–1407. [PubMed: 17098563]

Zhang H, Berg AC, Maire M, & Malik J (2006). Svm-knn: discriminative nearest neighbor classification for visual category recognition. In CVPR (2) (pp. 2126–2136).

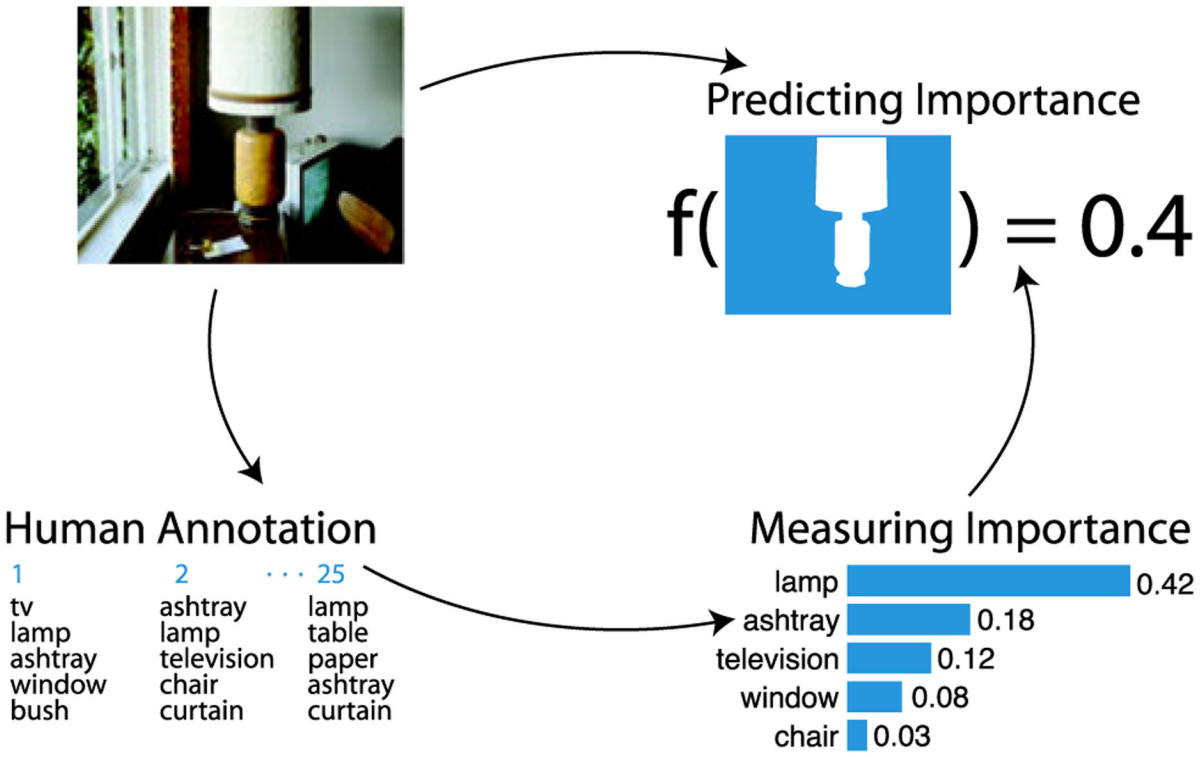

Fig. 1.

We wish to predict the importance of an object in a photo. In order to accomplish this, we must first produce a ground truth. We do so by combining the opinions of a large number of viewers (*bottom arrow*, Sect. 3). From this ground truth we may learn a function for predicting object importance from picture regions (*top* and *right arrows*, Sect. 4)

Fig. 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

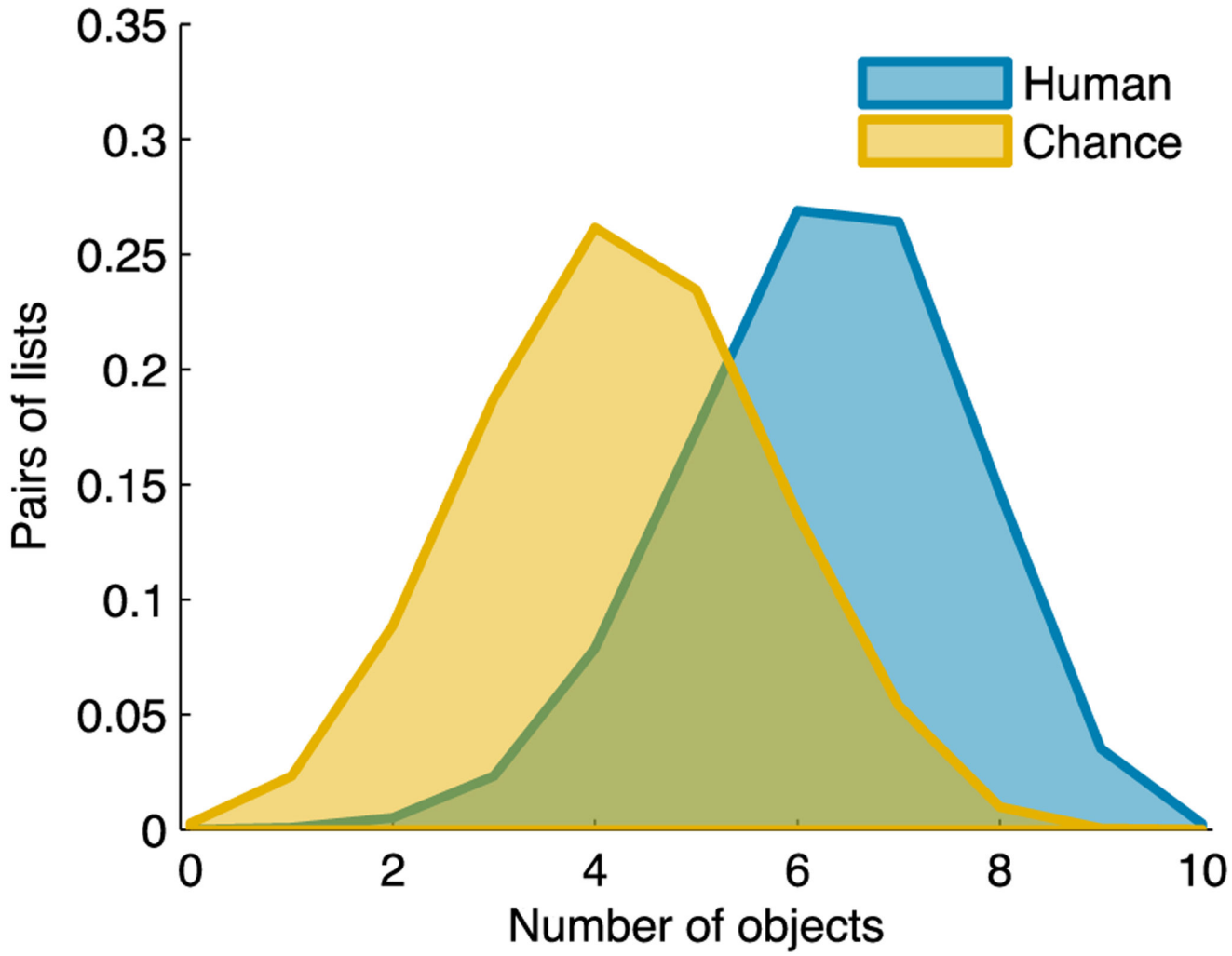

Fig. 3.

The number of objects shared by a pair of lists for the same image. Data collected from viewers (Human) is compared with random lists created by uniformly sampling from objects named for that image (Chance) in histogram form

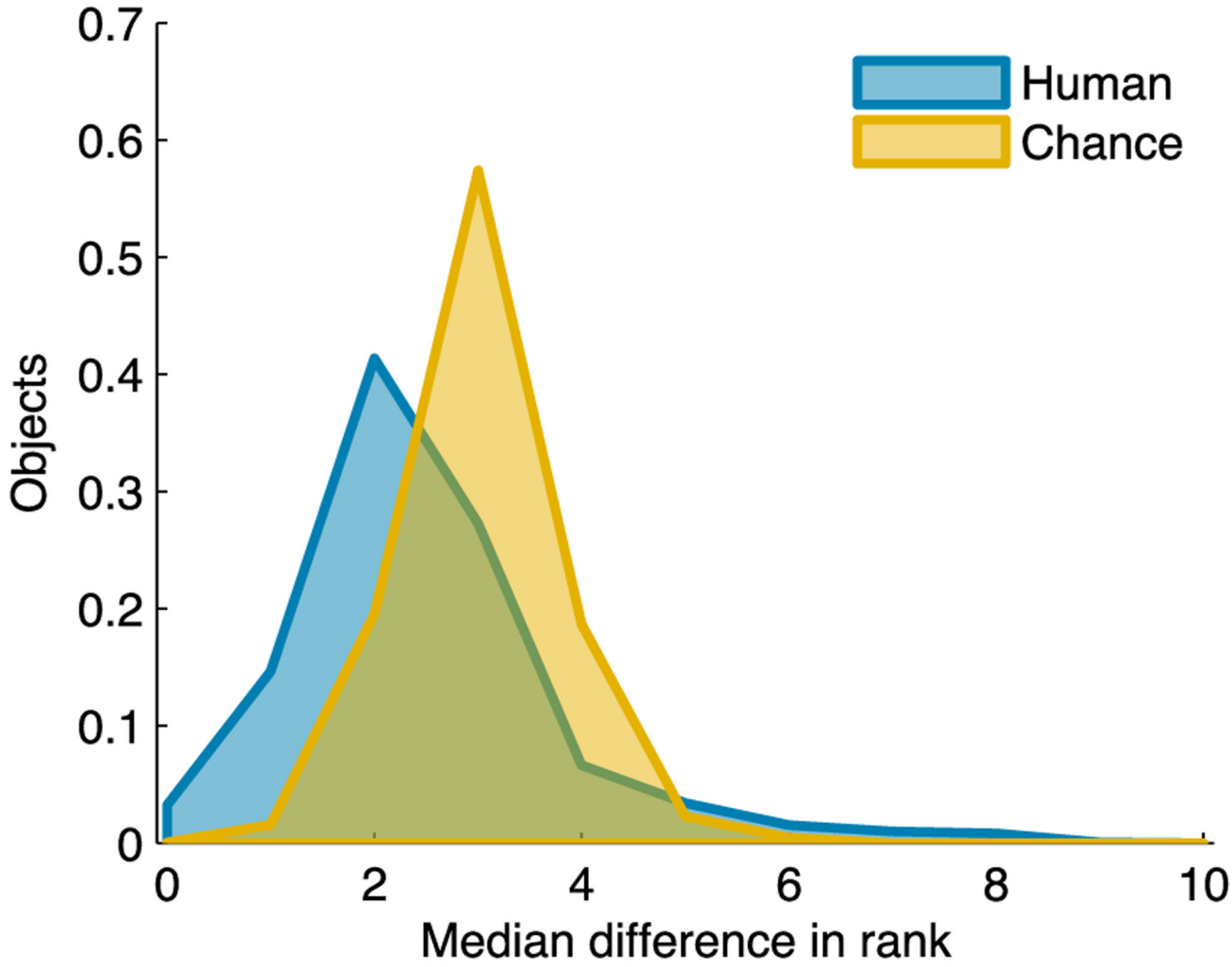

Fig. 4.

Given that an object appears on two lists for a particular image, how different is its rank on those lists? Each data point is the median of these differences for an object-image combination

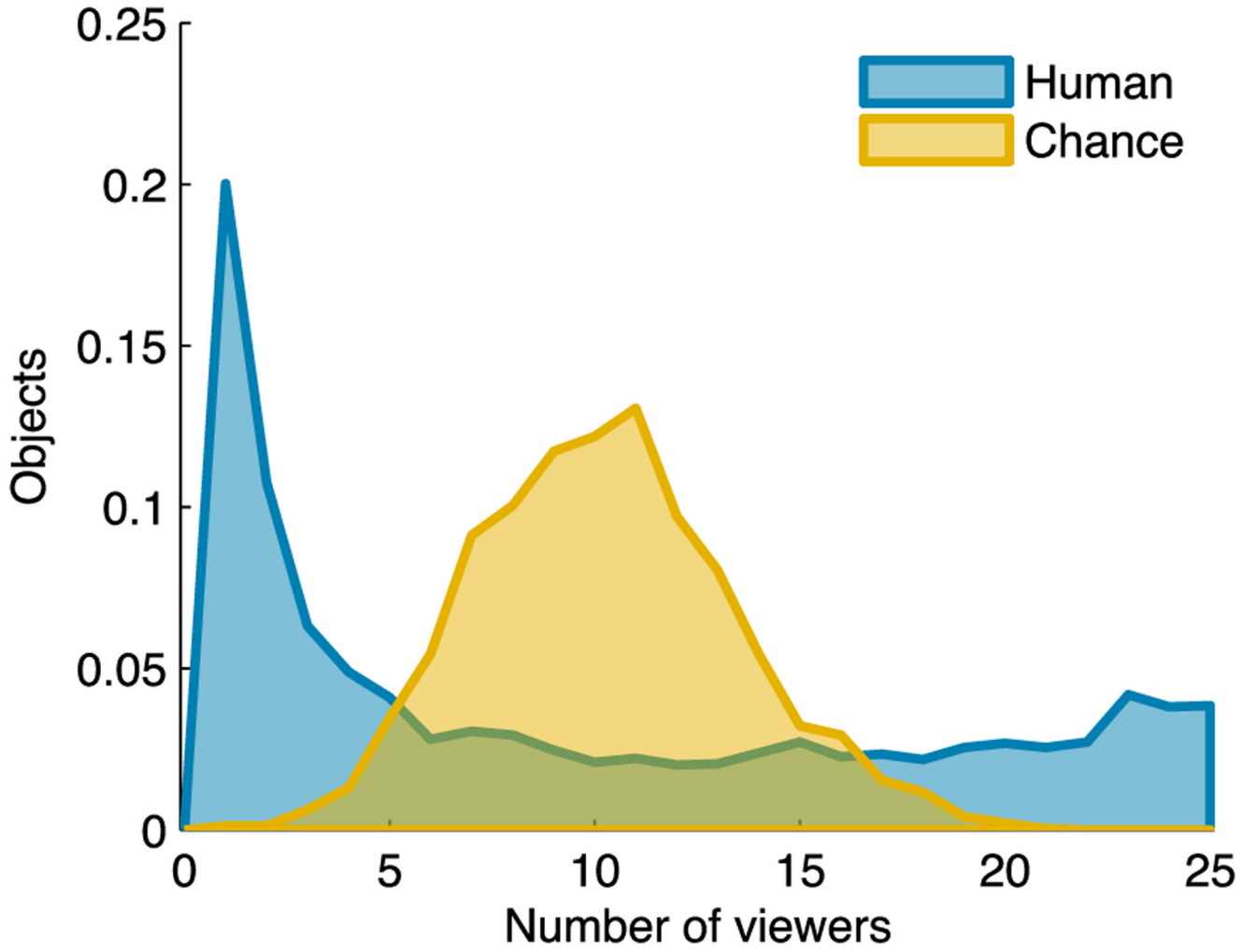

Fig. 5.

For a given image, how many viewers name a particular object? Each data point is the number of viewers that name that object in a specific image

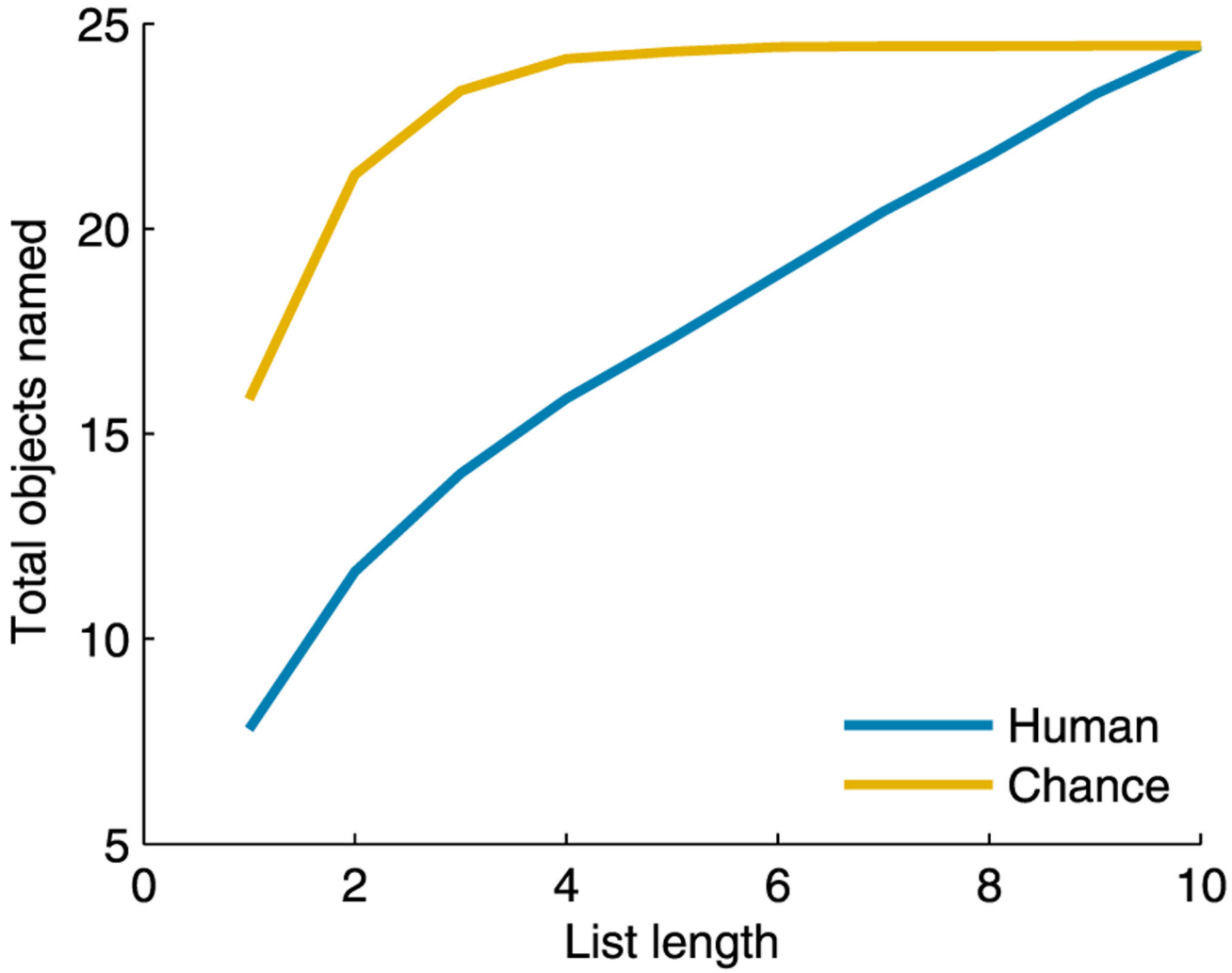

Fig. 6.

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

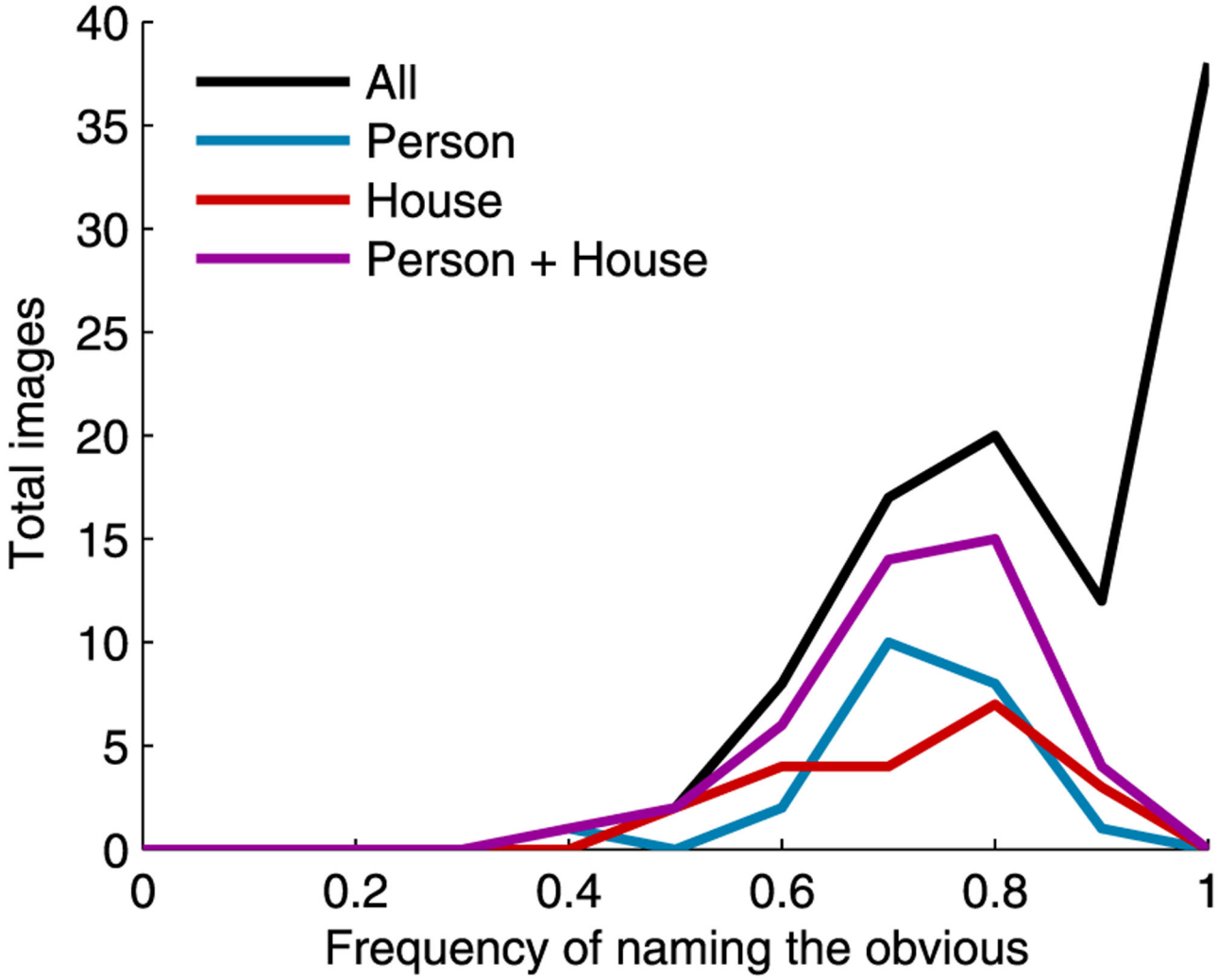

Fig. 7.

Some viewers fail to mention the obvious object. We define the ‘obvious object’ as the object named earliest (mean order), out of the most frequent half of objects. 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

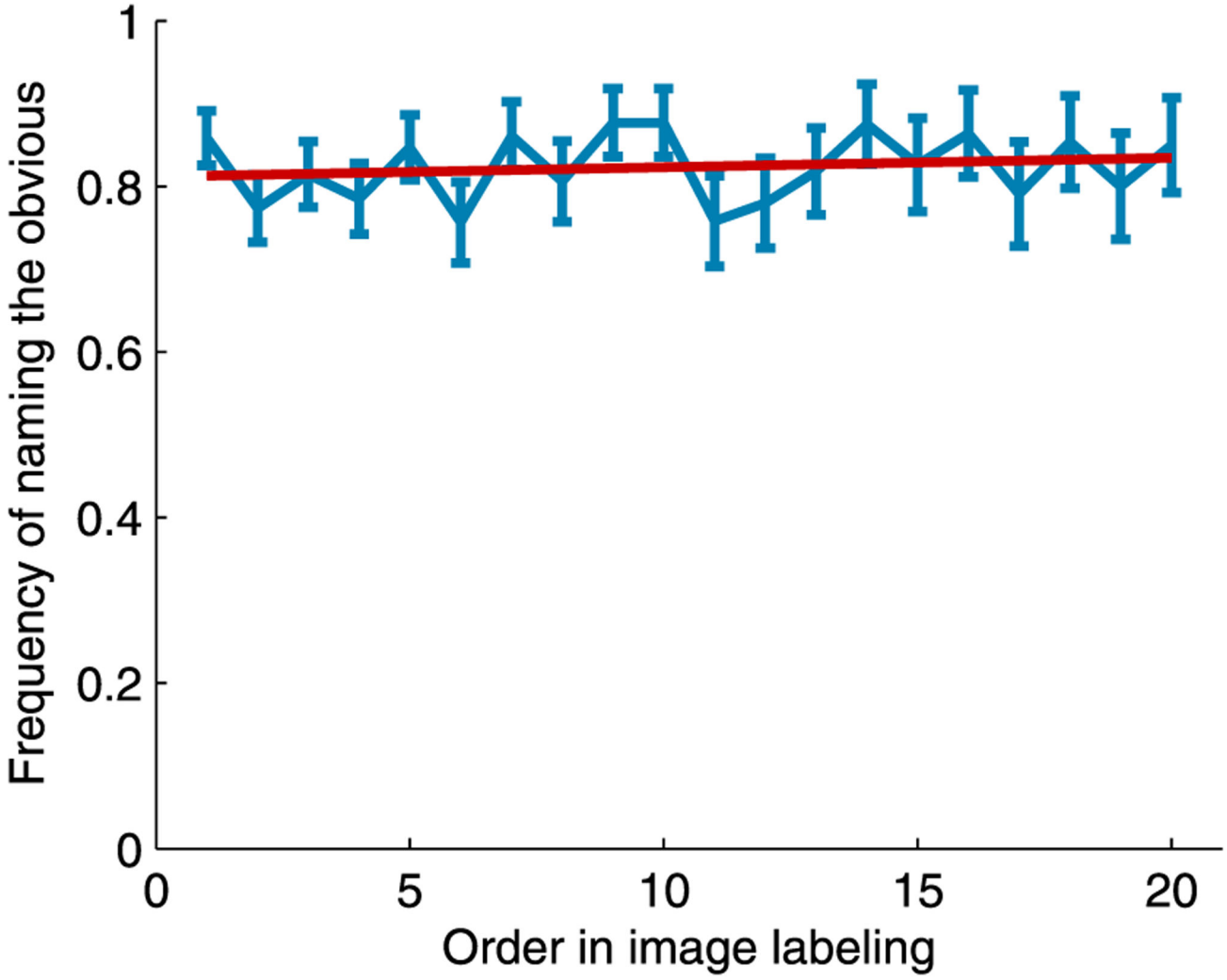

Fig. 8.
The frequency that the obvious object is named does not decrease as a viewer labels more images

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

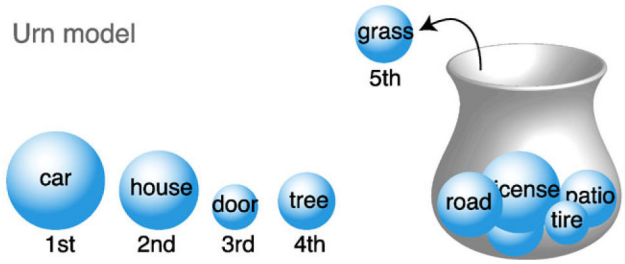

Fig. 9.

A photograph and corresponding lists generated by 5 observers. 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 which are more or less important in that image. A viewer names the objects one at a time until 10 objects are named. Similarly, an urn is filled with balls of different sizes, where larger balls are more likely drawn. 10 balls are removed from the urn, creating a sequence

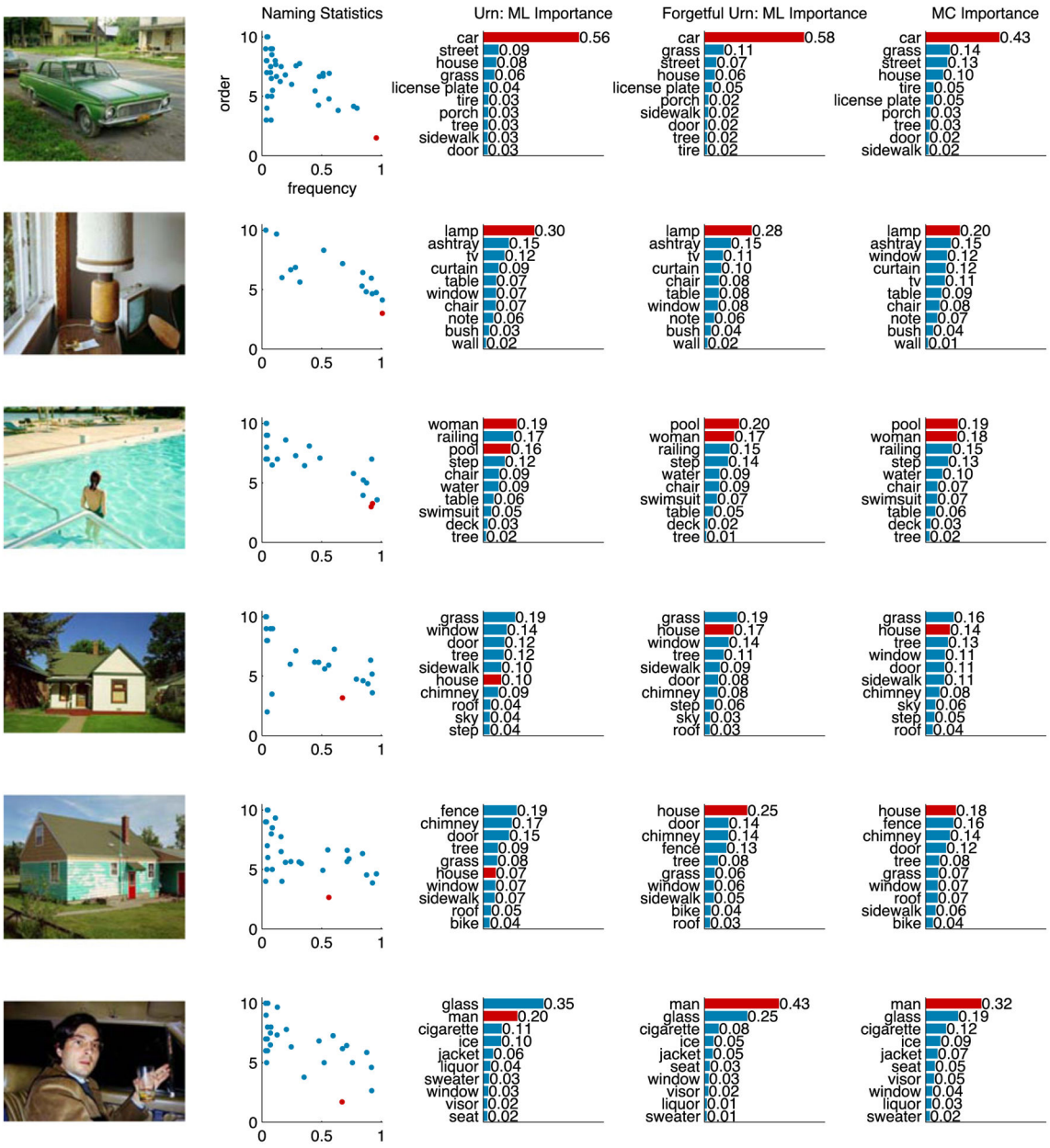

Fig. 10.

(Color online) Measured Importance. *2nd column*: For a particular image, we can calculate the proportion of lists that an object appears on (frequency) and it's mean order over the lists that mention it. 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 isn't missed the importance measurement is similar. The Markov Chain (*5th column*) arrives at similar results through a different approach

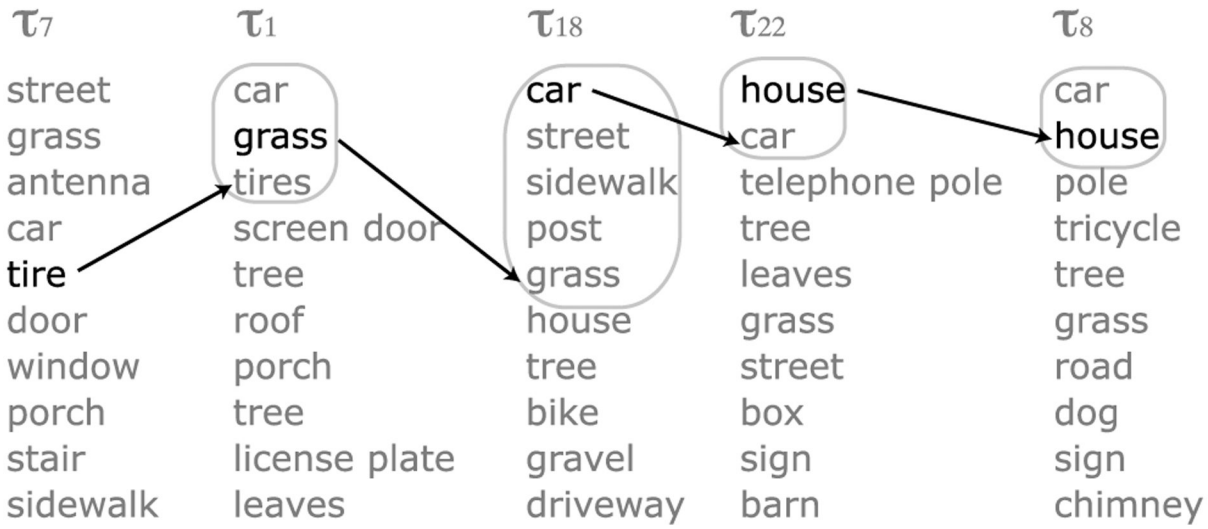

Fig. 11.

The Markov chain moves from object to object by 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

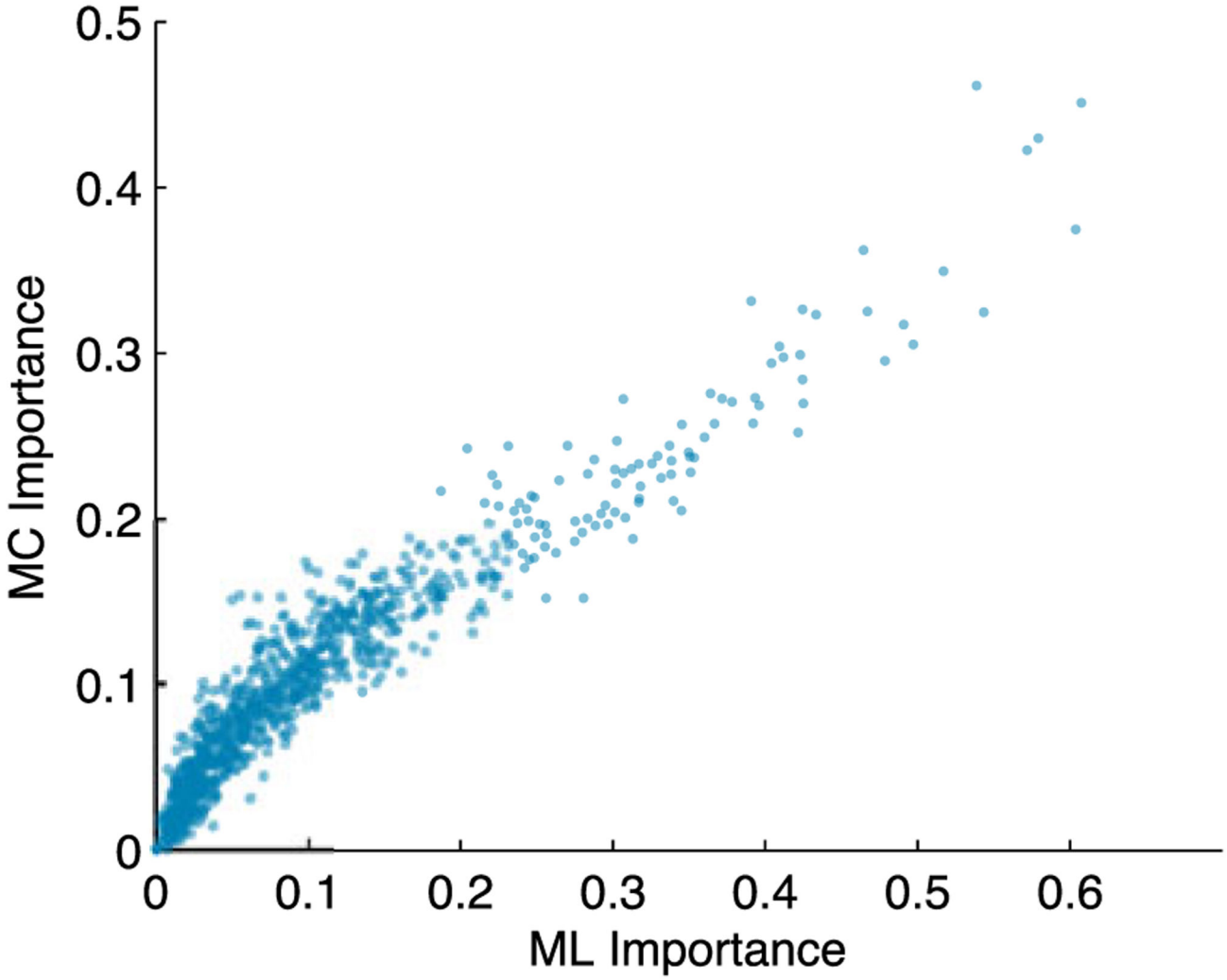

Fig. 12.
Forgetful urn Maximum Likelihood versus Markov chain measured importance

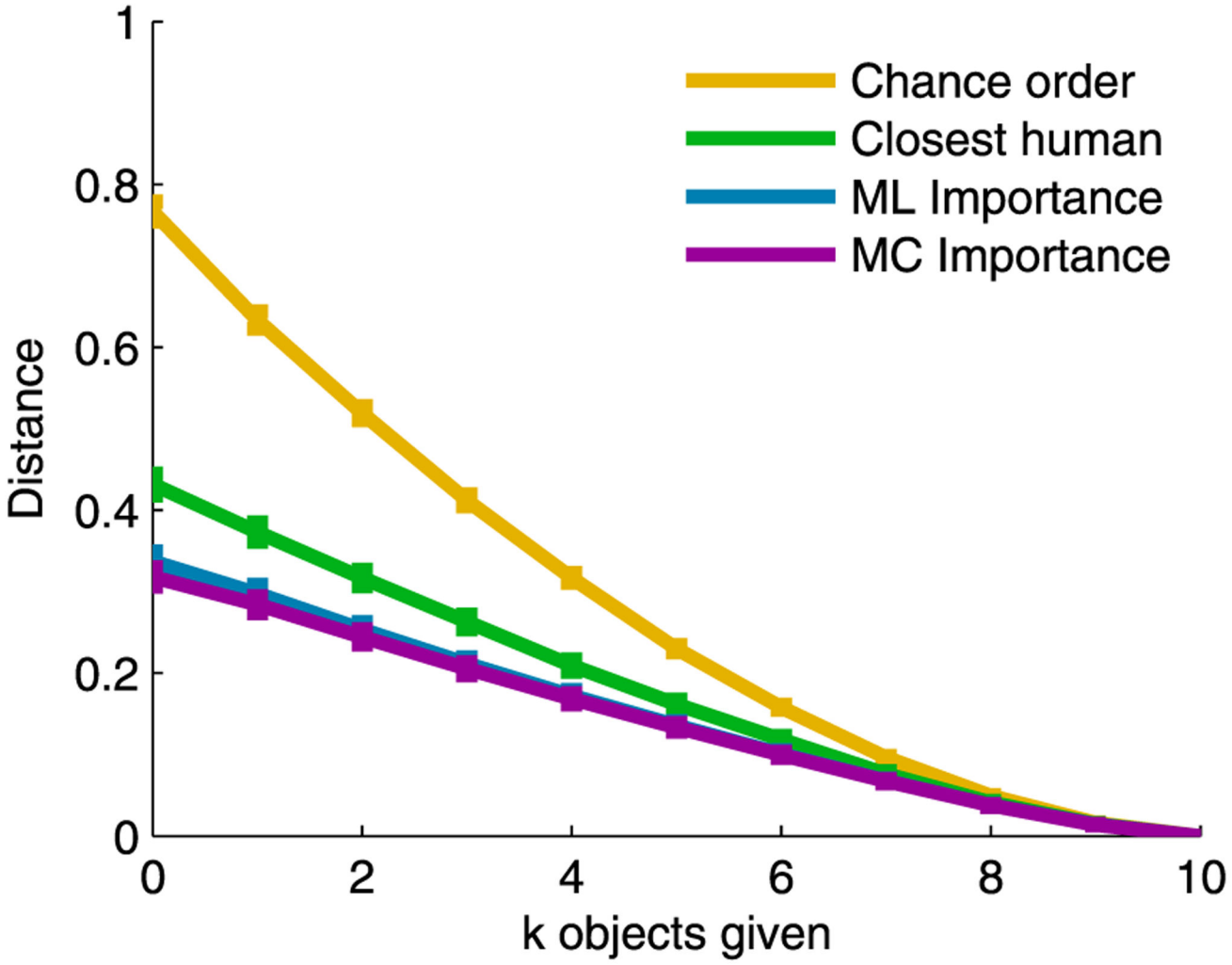

Fig. 13.

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

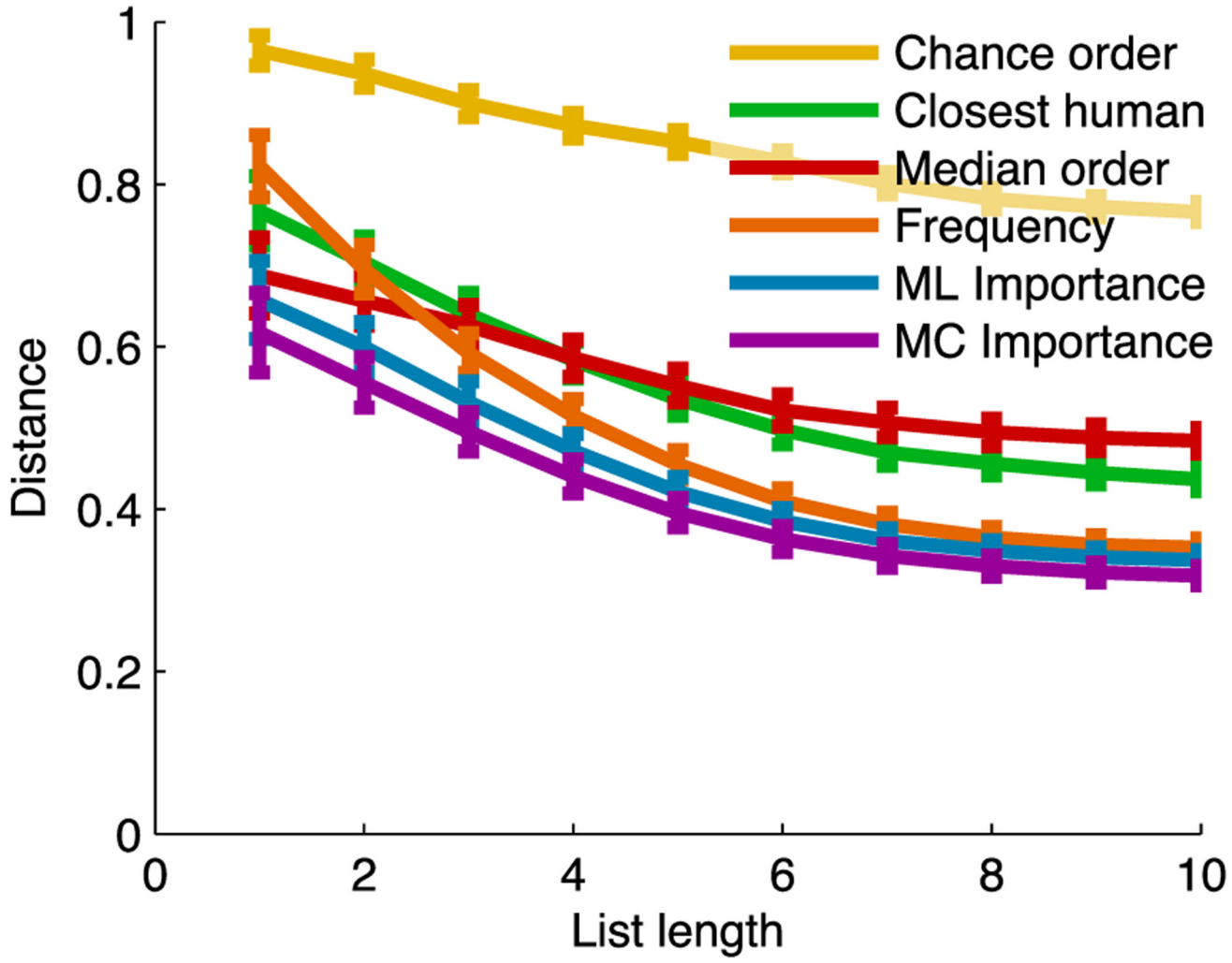

Fig. 14.

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

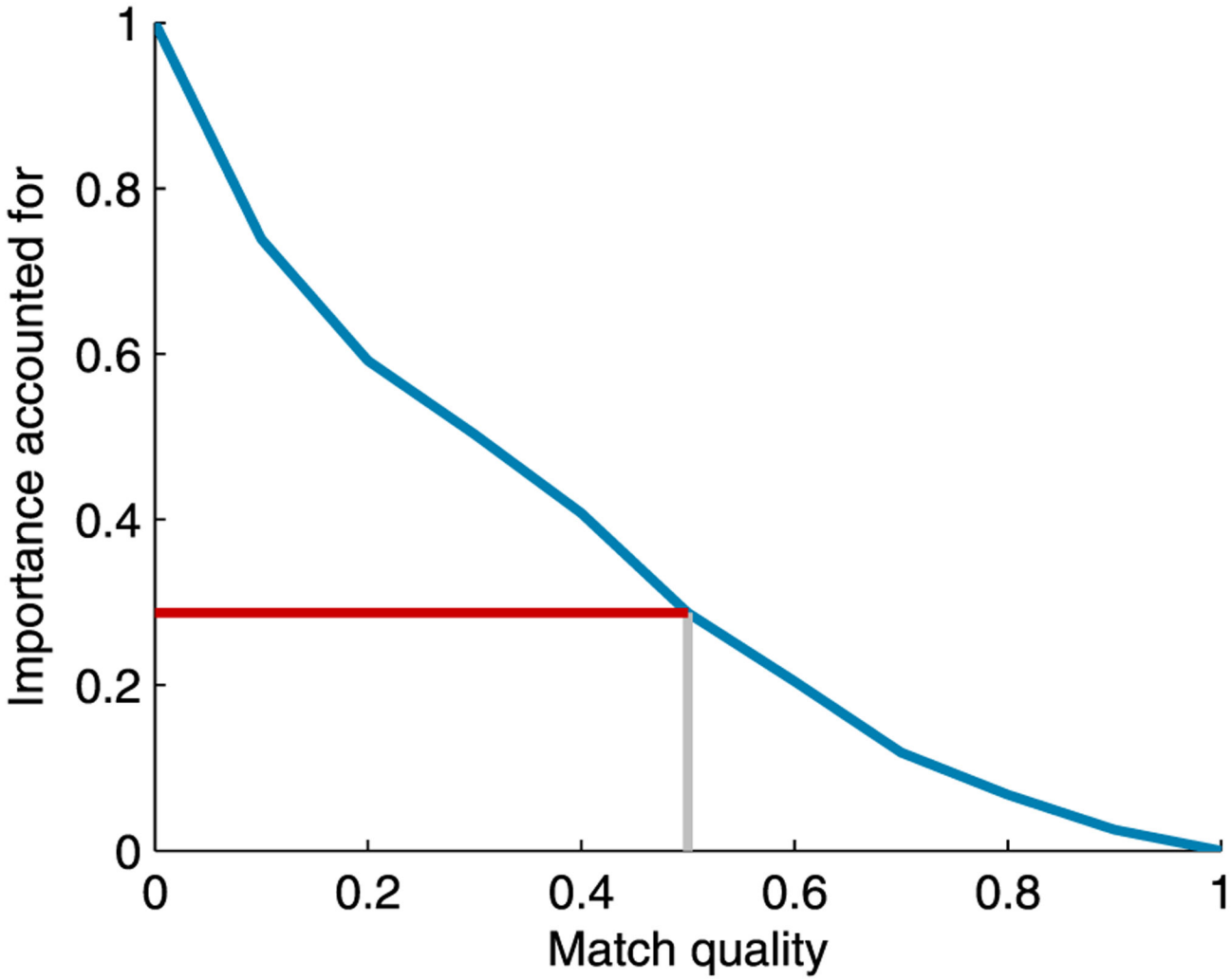

Fig. 15.

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 segmentation. We then sum the importance corresponding to the objects that meet a minimum match quality

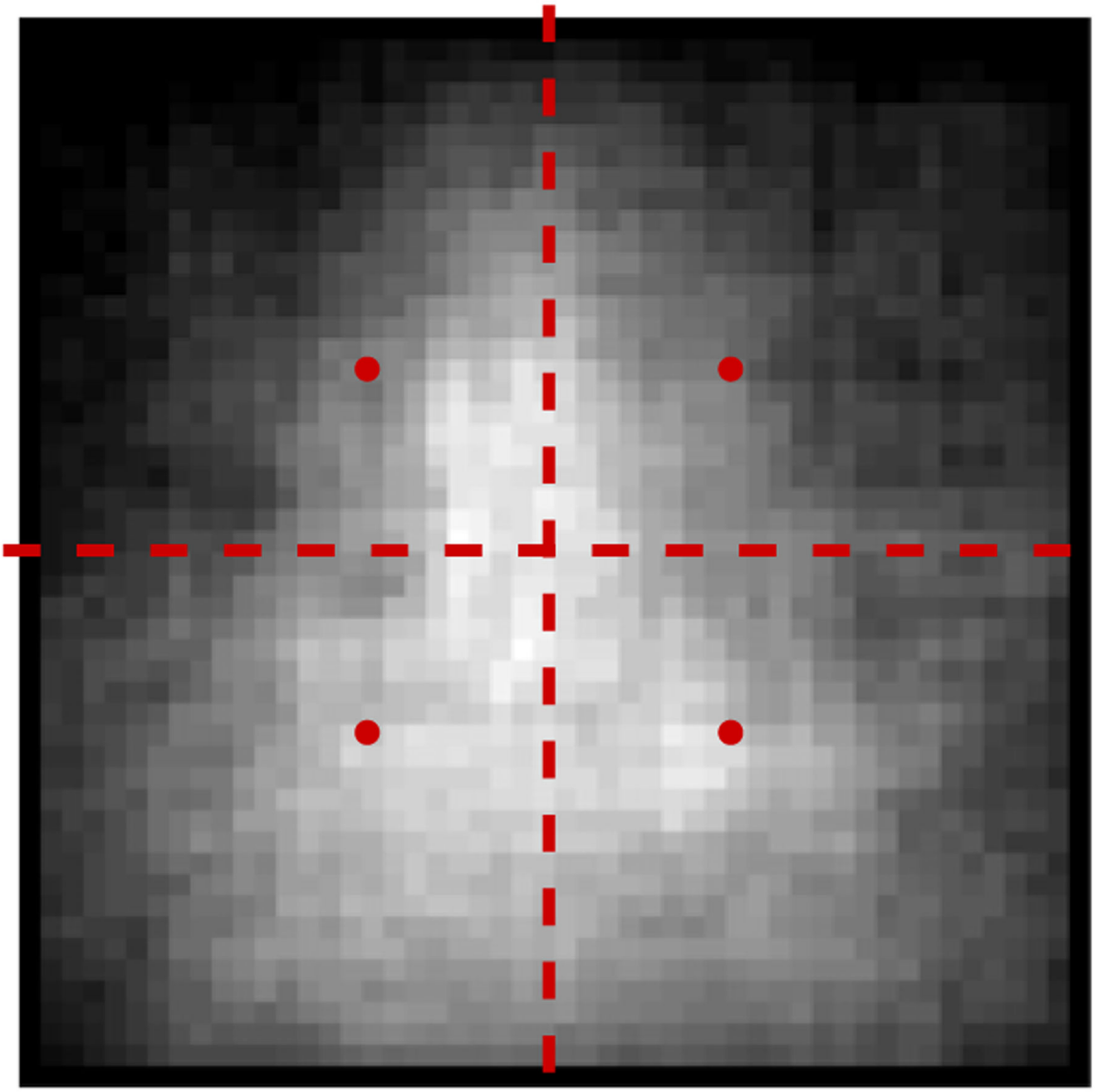

Fig. 16.

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

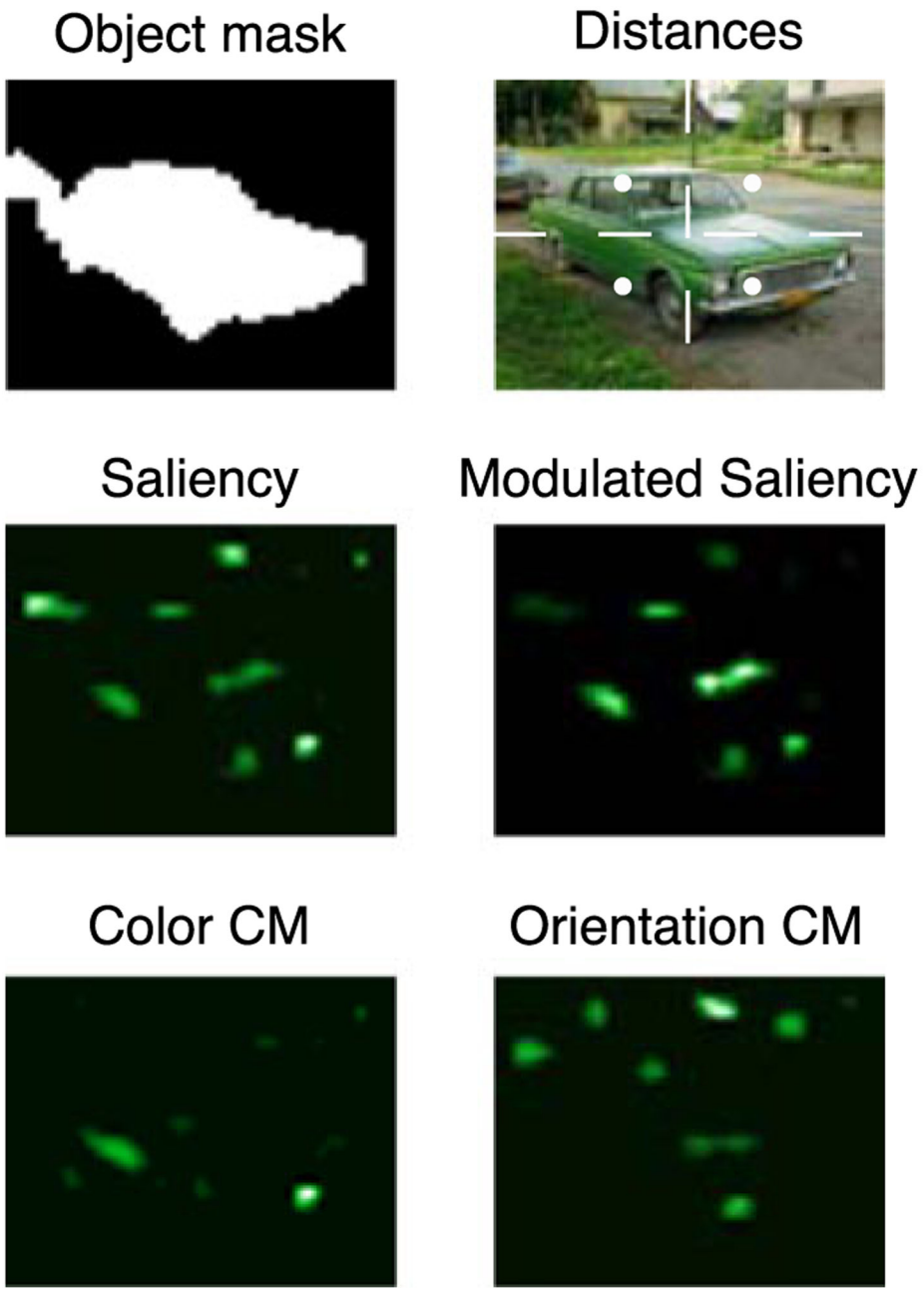

Fig. 17.

Feature Examples. We consider the mean, maximum, and minimum distances from the object mask to the image center (and *center lines*) and the points that divide the image into thirds. We look at the mean, maximum, and sum of values on the Saliency Map or Conspicuity Map (CM) that overlap with the object mask

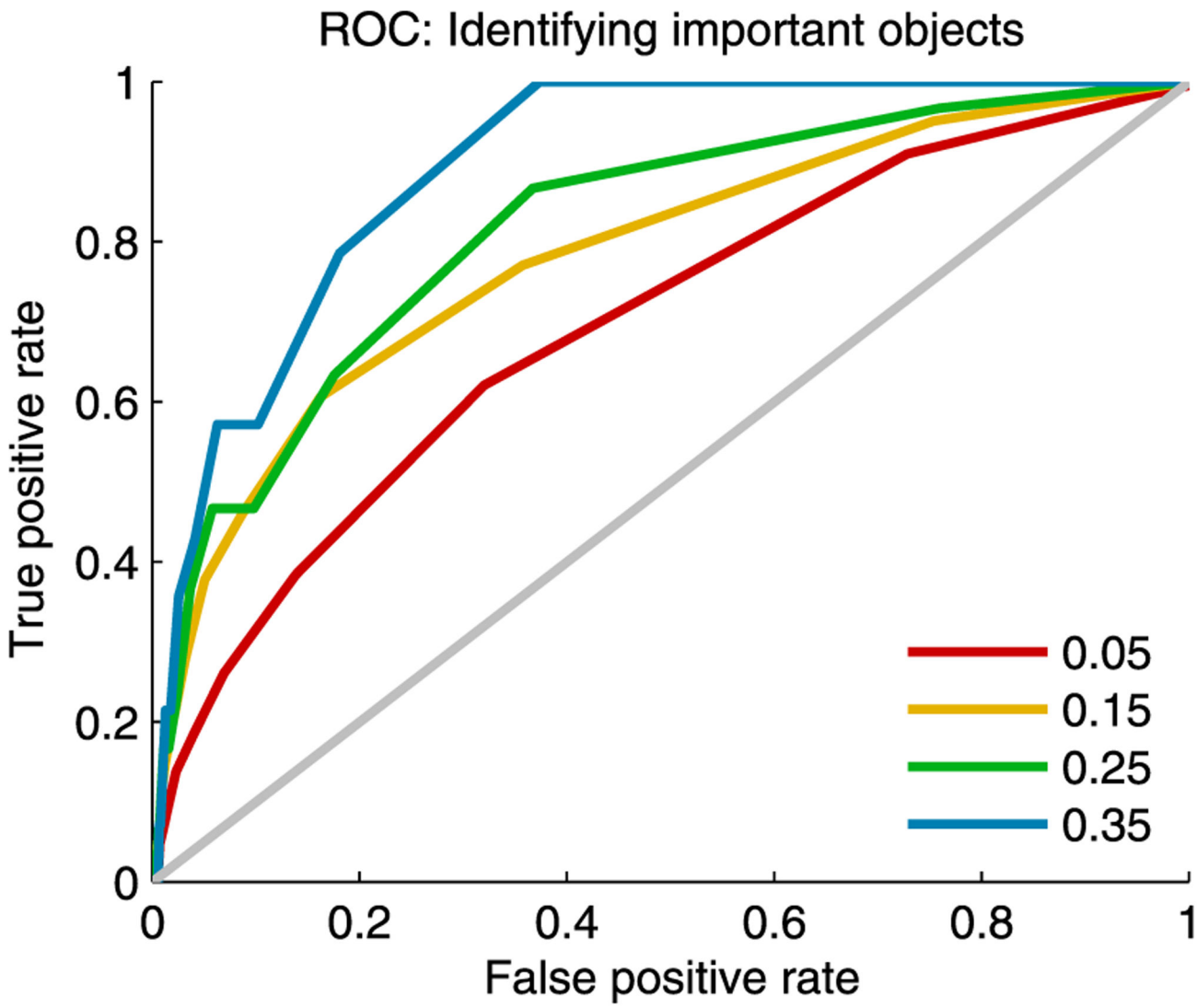

Fig. 18.

ROC curves for identifying important objects. We define an *important* object as having a measured importance ≥{0.05, 0.15, 0.25, 0.35} and move the threshold across the predicted importance

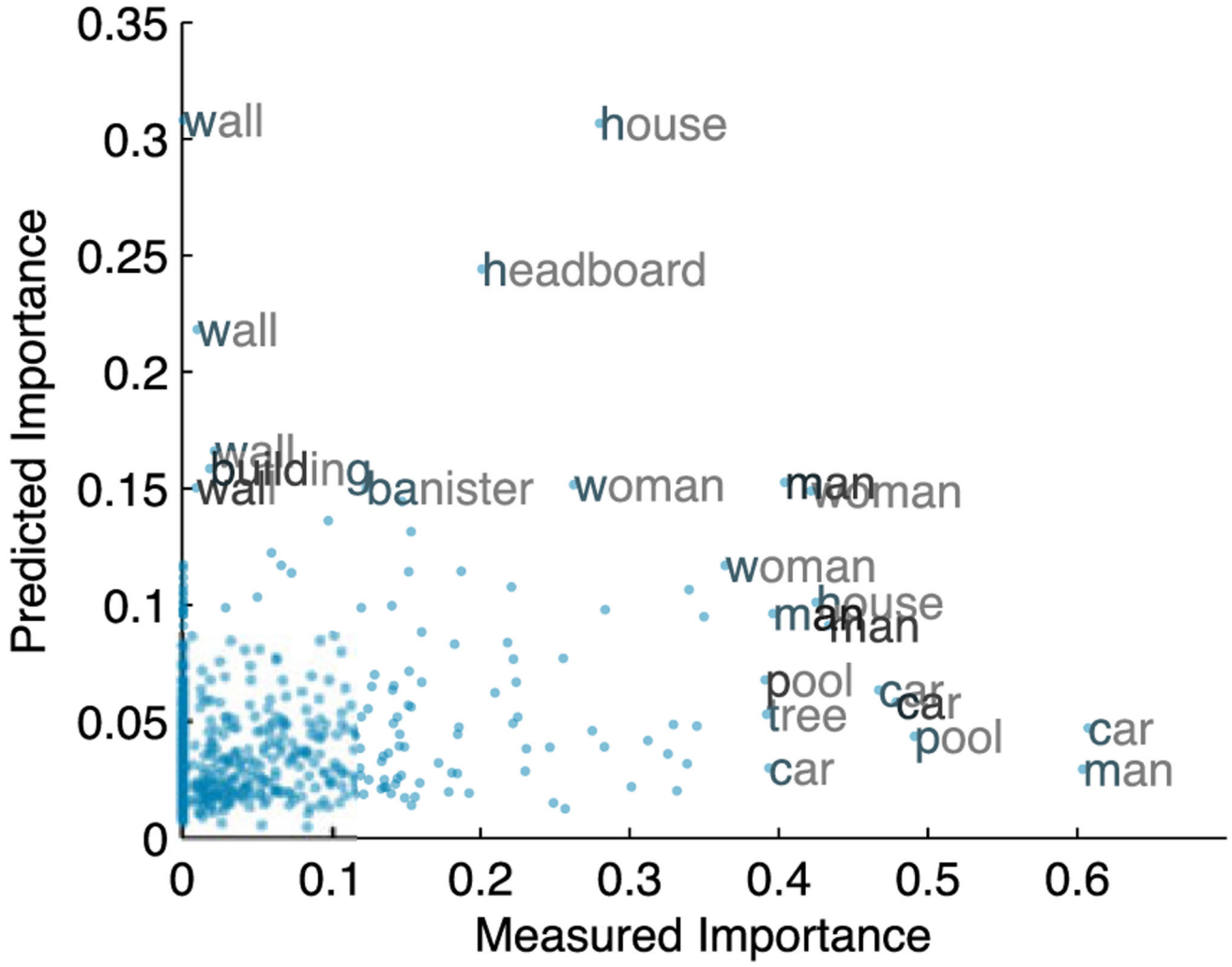

Fig. 19.
Scatter plot of predicted versus measured importance. Most objects have very low importances

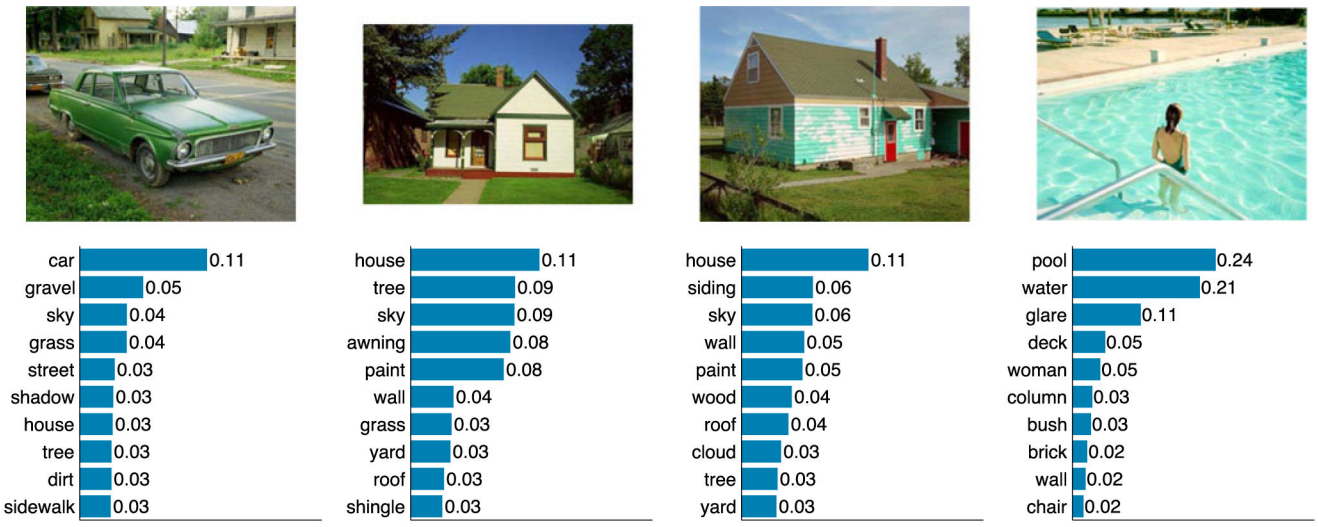

Fig. 20.

Predicted Importance. Importance predicted using the Lasso and simple image features. Notice that in the *4th 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

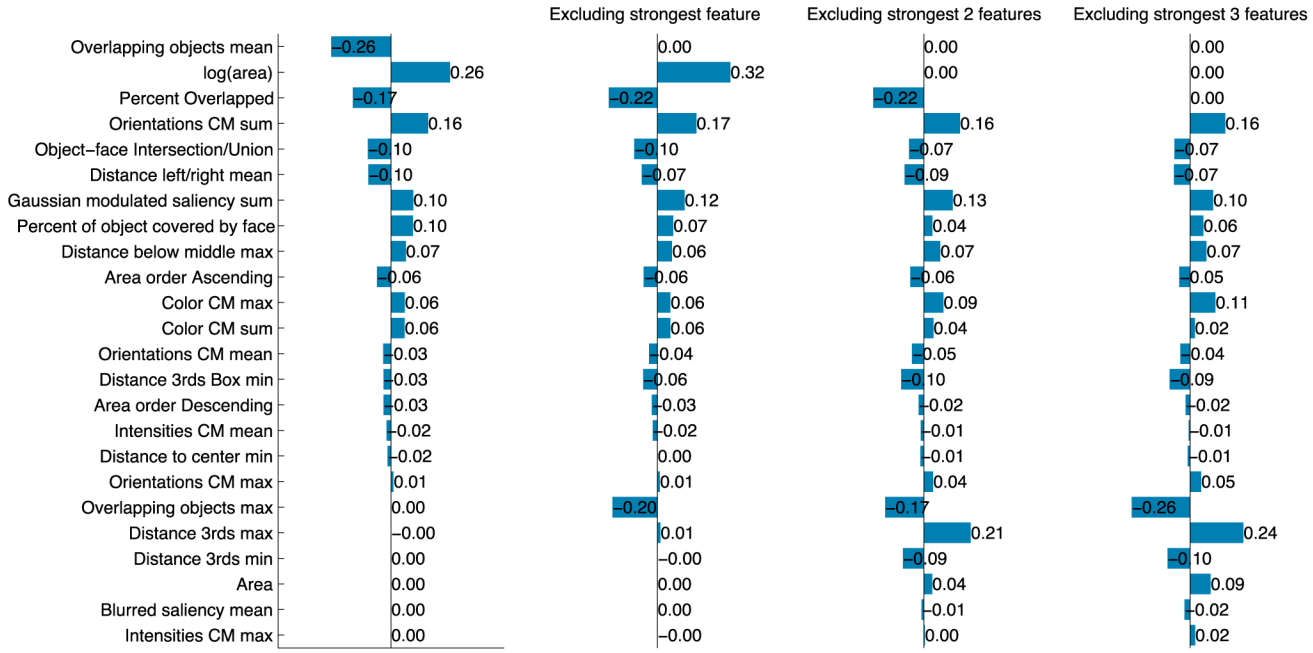

Fig. 21.

Excluding the features with the largest coefficients simply causes other features to replace them. The Residual Sum of Squares is only minimally affected

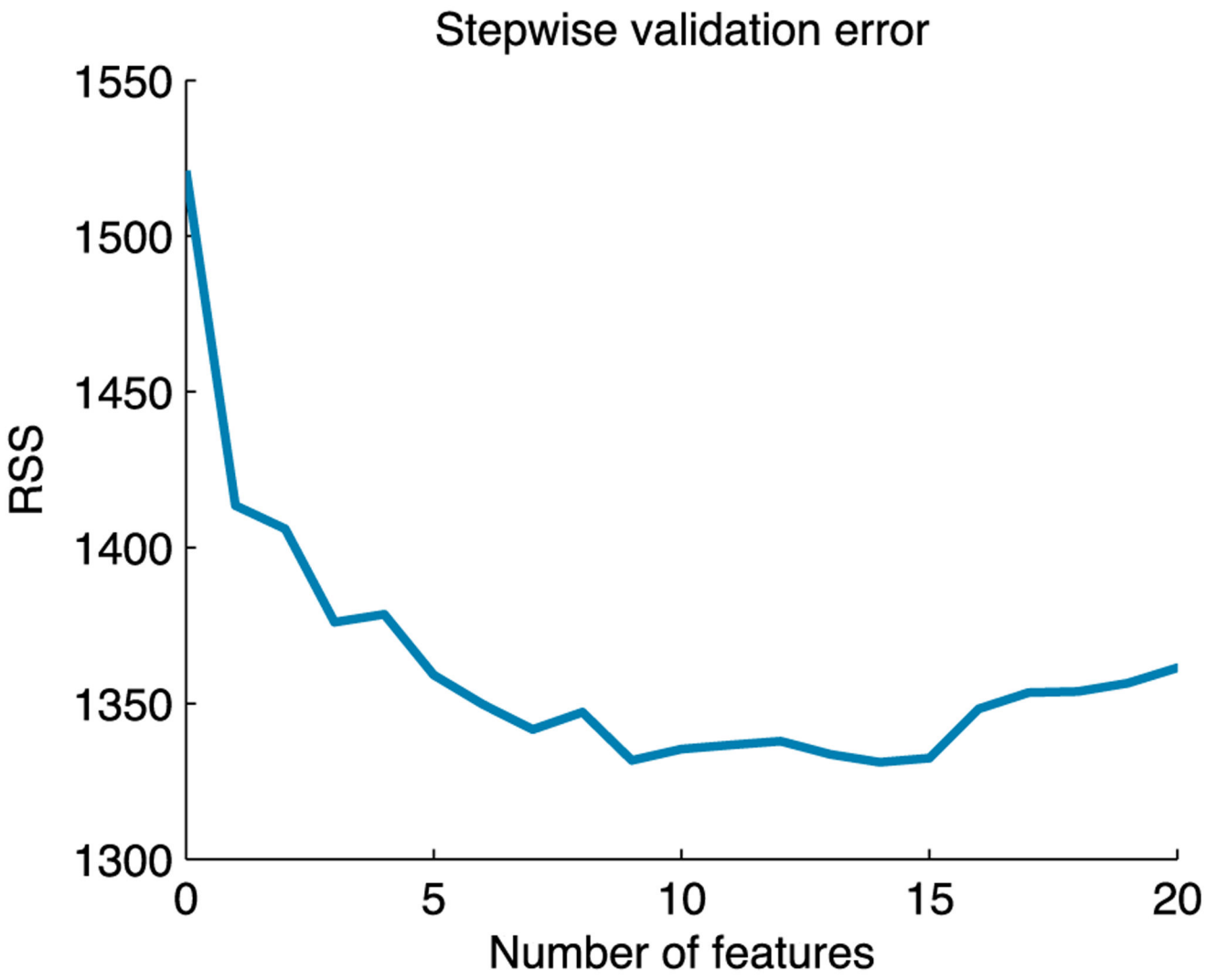

Fig. 22.

We see a diminishing return as we allow more features to be used in importance prediction by Stepwise regression

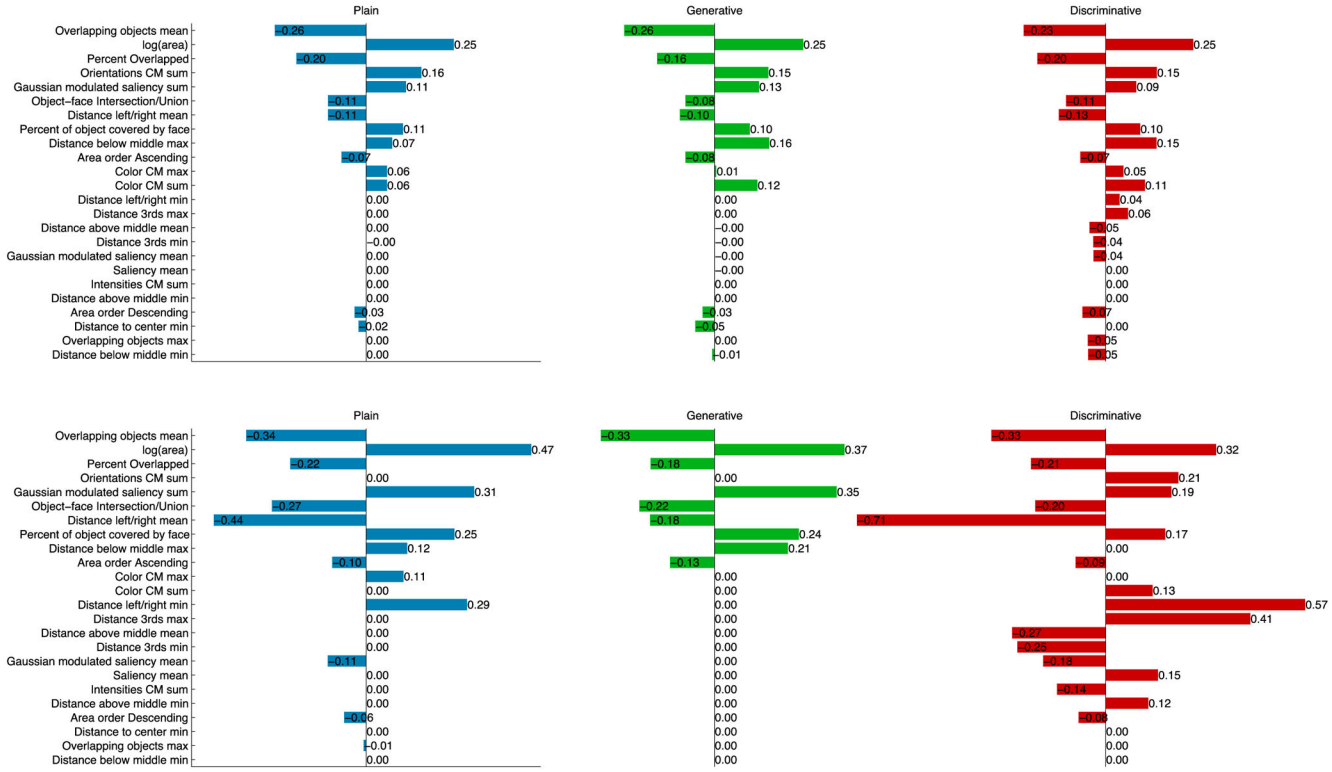

Fig. 23.

Coefficients for importance prediction. Data has 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*)

Fig. 24.

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

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

Fig. 25.
Detailed instructions given to all viewers

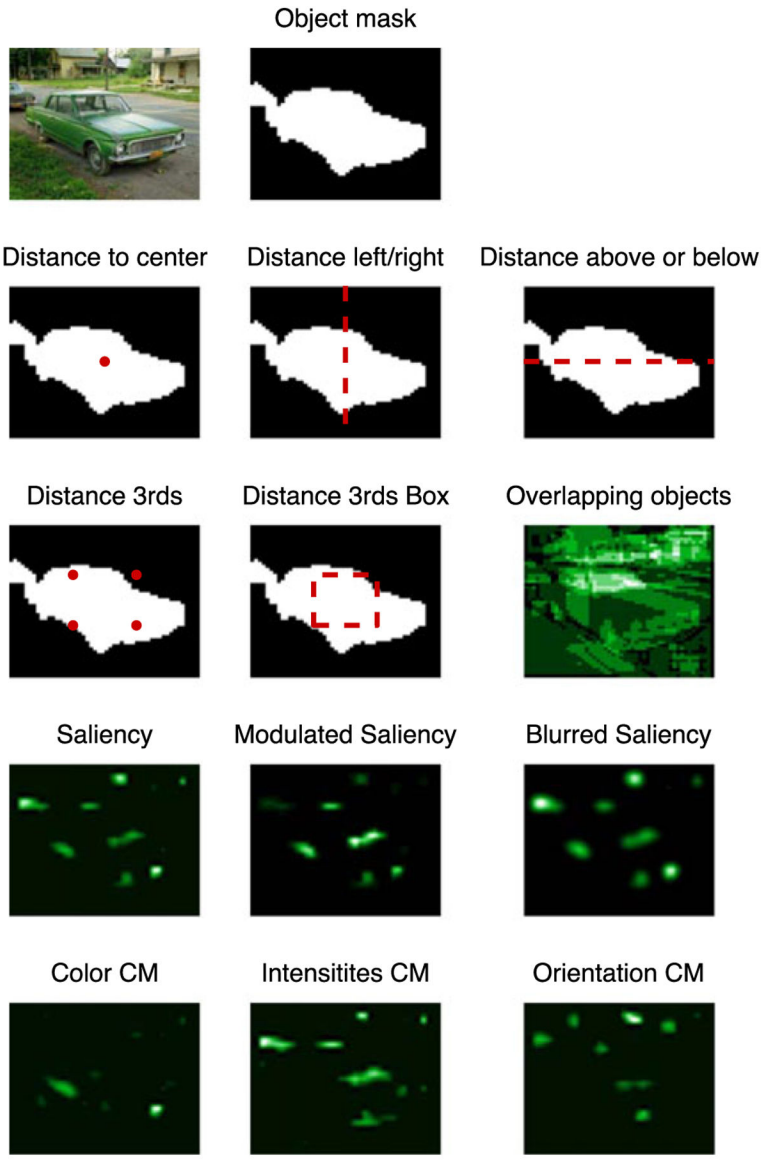

Fig. 26.

1st row: A photograph and example 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 our modifications. *5th row*: Conspicuity Maps

Table 1

Sample lists from 5 viewers of the first photo in Fig. 2

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

Table 2

Object naming is largely independent of other named objects. These are the only object pairs found to be dependent with Pearson chi-square test out of the 4,224 pairs

Word pair		***p* value**
Eye	nose	1.0e-05*
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

Table 3

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

Feature	Coefficient
Overlapping objects mean	−0.2645
log(area)	0.2605
Percent Overlapped	−0.1686
Orientations CM sum	0.1636
Object-face Intersection/Union	−0.103
Distance left/right mean	−0.1001
Gaussian modulated saliency sum	0.098
Percent of object covered by face	0.0969
Distance below middle max	0.0653
Area order Ascending	−0.0623
Color CM max	0.0609
Color CM sum	0.0602
Orientations CM mean	−0.0346
Distance 3rds Box min	−0.0337
Area order Descending	−0.0334

Table 4

List of all features used in importance prediction

Distance to center (max, mean, min)
Distance left/right max (max, mean, min)
Distance above middle (max, mean, min)
Distance below middle (max, mean, min)
Distance 3rds (max, mean, min)
Distance 3rds Box (max, mean, min)
Saliency (sum, max, mean)
Gaussian modulated saliency (sum, max, mean)
Blurred saliency (sum, max, mean)
Color CM (sum, max, mean)
Intensities CM (sum, max, mean)
Orientations CM (sum, max, mean)
Area
log(area)
Area order (Descending, Ascending)
Percent Overlapped
Number of Overlapping objects (max, mean)
Percent of face covered by object
Percent of object covered by face
Object-face Intersection/Union