Published in final edited form as:
Psychol Sci. 2016 June ; 27(6): 870–884. doi:10.1177/0956797616640237.

Searching for Category-Consistent Features: A Computational Approach to Understanding Visual Category Representation

Chen-Ping Yu[1], **Justin T. Maxfield**[2], and **Gregory J. Zelinsky**[1,2]
[1]Stony Brook University, Department of Computer Science
[2]Stony Brook University, Department of Psychology

Abstract

A generative model of category representation is introduced that uses computer vision methods to extract category-consistent features (CCFs) directly from images of category exemplars. The model was trained on 4800 images of common objects, and CCFs were obtained for 68 categories spanning subordinate, basic, and superordinate levels in a category hierarchy. Participants searched for these same categories. Targets cued at the subordinate level were preferentially fixated, but fixated targets were verified faster following a basic-level cue. The subordinate-level advantage in guidance is explained by the number of target category CCFs, a measure of category specificity that decreases with movement up the category hierarchy. The basic-level advantage in verification is explained by multiplying CCF number by sibling distance, a measure of category distinctiveness. With this model the visual representations of real-world object categories, each learned from the vast numbers of image exemplars accumulated throughout our everyday experience, can finally be studied.

Keywords

Category Representation; Categorization; Categorical Search; Categorical Features; Generative models; Category Hierarchies

Introduction

> Our categories make us who we are; they are the skeleton upon which grows the rest of our psychological being. Reflecting their diverse importance, categories have been studied from multiple perspectives: as a lens through which we perceive visual and acoustic objects in the world (Liberman, Harris, Hoffman, & Griffith, 1957; Regier & Kay, 2009) and the similarity relationships between these objects (Medin & Schaffer, 1978; Goldstone, 1994), and as the structure of concepts that organize our knowledge and define who we are (Kaplan &

Address correspondence to: Gregory J. Zelinsky, Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500, Phone: 631-632-7827, Gregory.Zelinsky@stonybrook.edu.

Author Contributions
All authors contributed to the study design and theoretical concept. J.T. Maxfield conducted and analyzed data from the behavioral experiment. C-P Yu implemented the model, conducted the simulations, and analyzed the computational results. G.J. Zelinsky advised on both of these efforts. All authors collaborated on drafts of the manuscript and approved the final version for submission.

Matlab code for the CCF model can be downloaded as a zip file from: https://github.com/cxy7452/bow-ccf.

Murphy, 2000; Pazzani, 1991; Murphy, 2002). Some approaches are also highly quantitative. The semantic organization literature uses formal methods from logic theory to understand the division of information into clusters of semantic nodes (Anderson, 1983; Collins & Quillian 1969), and the category learning literature models how corrective feedback about category membership can shape our categorization decisions (Anderson, 1996; Ashby & Maddox, 1993; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky; 1986; Nosofsky & Palmeri, 1997).

All of these approaches, however, have skirted a basic question of category representation—how might the visual features of common object categories be extracted from the many exemplar images of these objects that we encounter in our day-to-day lives?

Growing in parallel with these largely behavioral literatures has been another literature that may help answer this question. The field of computer vision is rich with operators and algorithms developed to detect members of object classes directly from pixels in images (Duda, Hart, & Stork, 2012). Moreover, these tools work with featurally-complex real-world categories, and their performance is evaluated using new or “unseen” category exemplars not used during model training. In contrast, behavioral work on category learning has placed less emphasis on real-world application and model prediction, focusing instead on how categories defined by a small number of simple features are learned from feedback (see Ashby & Maddox, 2005). A gap therefore exists in our fundamental understanding of categories; much is known about how simple features can be learned and used to discriminate one category from another, but little is known about the features composing the categories of common objects that populate our everyday experience. By bridging these different approaches we achieve a new understanding of categories. Our premise is that tools from computer vision can, and should, be exploited to characterize the feature representations of categories as they exist “in the wild”, formed simply from a lifetime of experience seeing category exemplars.

The Generative Modeling of Visual Categories

We adopt a generative modeling approach. Because generative models are usually unsupervised, they capture the implicit learning from exemplars that people and other animals use to acquire the within-category feature structure of visual object categories. Figure 1 helps to make this point. A generative model learns the features that are common among the objects of a category, much like the human visual system causes the perception of rectangles in this figure by finding common features among category exemplars grouped at the basic level.

Generative models can be contrasted with discriminative models, which use supervised error feedback to learn features that discriminate target from non-target categories (Ulusoy & Bishop, 2005). The vast majority of category learning studies adopt a discriminative modeling approach (Ashby & Maddox, 1993; Kruschke, 1992; Nosofsky; 1986; Nosofsky & Palmeri, 1997), which is appropriate given the heavy reliance on the artificial classification learning paradigm in this literature. A generative approach, however, is more appropriate when modeling data that do not reflect explicit classification decisions (Kurtz, 2015; Levering & Kurtz, 2014; see also Chin-Parker & Ross, 2004), such as the visual search data in the present study. Our position is that generative models better capture the features of a category used to construct visual-working-memory representations of search targets, similar to the features one might call to mind when forming a mental image of a target category. If searching for a Pekin duck one would probably look for a white, mailbox-sized object with orange at the top and bottom, despite these features potentially yielding poor discrimination from poodles and pumpkins.

Hierarchies of Categories

We evaluate our model within the context of a simple conceptual structure, a three-level hierarchy. Objects can be categorized at multiple levels in a conceptual hierarchy. A sea vessel powered by wind can be categorized as a sail boat (subordinate level), simply a boat (basic level), or more broadly as a vehicle (superordinate level). The basic-level superiority effect (BSE) refers to the finding that the acquisition and access of categorical information seems anchored around the basic level. It was first reported by Rosch and colleagues (1976) using a speeded category verification task, where they found that people were faster in judging a picture of an object as a member of a cued category when the cue was at the basic level. Subsequent work broadened the scope of the BSE by showing it to be the preferred level used in speech, and the first nouns generally learned and spoken by children (Mervis & Rosch, 1981; Rosch, 1978).

Explanation of the BSE has appealed to similarity relationships within and between categories. Basic-level categories are thought to maximize within-category similarity while simultaneously minimizing between-category similarity; subordinate or superordinate-level categories do one or the other, but not both (Rosch et al., 1976). Murphy and Brownell (1985) advanced this idea by theorizing that the BSE was a by-product of concurrent *specificity* and *distinctiveness* processes pulling categorization in opposing directions. Subordinate-level categories tend to have very specific features; Collies are medium-sized dogs with thin snouts, upright ears and white hair around their shoulders. However, these features overlap with other dog categories, making Collies sometimes challenging to distinguish from German Shepherds or Shelties. Superordinate-level categories have the opposite strengths and weaknesses. The features of animals overlap minimally with vehicles or musical instruments, making the category distinct. However, animal features are also highly variable, making superordinate categories lacking in specificity. The basic level strikes a balance between these opposing processes, and this balance is believed to underlie the BSE. Despite their variability in appearance, dogs have many features in common yet are still relatively distinct from ducks and dolphins and dinosaurs. The present work builds on this framework by making the processes of specificity and distinctiveness computationally explicit, and applying these principles directly to images of category exemplars.

Categorical Search

We evaluate the visual representation of common object categories using a *categorical search* task (Maxfield, Stadler, & Zelinsky, 2014; Schmidt & Zelinsky, 2009; Zelinsky, Adeli, Peng, Samaras, 2013; Zelinsky, Peng, Berg, & Samaras, 2013; Zelinsky, Peng, & Samaras, 2013). Categorical search differs from standard visual search in that targets are designated by category (e.g., the word “dog”) instead of by a picture pre-cue (e.g., an image of a specific dog), a situation that rarely exists outside the laboratory. Moreover, categorical search can be meaningfully divided into two epochs, one being the time between search display onset and first fixation on a target (*search guidance*) and the other being the time between first fixation on the target and the correct target-present judgment (*target verification*). Categorical search therefore embeds a standard category verification task within a search task, making it a powerful paradigm for studying the relationship between overt attention and categorization.

We introduce a method for quantifying the visual features of common object categories, and show that these features serve both to guide overt attention to a target and to categorize it after its fixation, with a BSE appearing during this latter target-verification epoch. The fact that our model captured these disparate behavioral measures provides converging evidence, within the context of a single categorical search task, that it can successfully identify the visual features used to represent common object categories. As such, this work creates a strong theoretical bridge between the attention (search guidance) and recognition (category verification) literatures.

Behavioral Methods

Participants

Twenty-six Stony Brook University undergraduates participated in a categorical search task. Sample size was determined based on a previous study using a similar method (Maxfield & Zelinsky, 2012). All participants reported normal or corrected-to-normal visual acuity and color vision, and that English was their native language. All also provided informed consent prior to participation in accordance with Stony Brook University's Committee on Research Involving Human Subjects.

Stimuli & Apparatus

Images of common objects were obtained from ImageNet (http://www.image-net.org) and various web sources. All images were closely cropped using a rectangular marquee to depict only the object and a minimal amount of background. Because object typicality can affect categorization and search (Murphy & Brownell, 1985, Maxfield et al., 2014), targets were selected to be typical members of their category at the subordinate, basic, and superordinate levels. We did this by having 45 participants complete a preliminary norming task in which 240 images (5 exemplars from each of 48 subordinate categories) were rated for both typicality and image agreement (Snodgrass & Vanderwart, 1980) at each hierarchical level using a 1 (high typicality/image agreement) to 7 (low typicality/image agreement) scale. The three most typical exemplars of each category were used as targets in the search task. Their mean typicality and image agreement was 2.29 and 2.31, respectively, and Table 1 lists these category names. In total there were 68 categories spanning 3 hierarchical levels; 4 superordinate-level categories, each having 4 basic-level categories, with each of these having 3 subordinate-level categories.

Eye position during the search task was sampled at 1000 Hz using an Eyelink 1000 eyetracker (SR Research) with default saccade detection settings. Calibrations were only accepted if the average spatial error was less than 0.5°, and the maximum error was less than 1°. Head position and viewing distance were fixed at 65 cm using a chinrest for the duration of the experiment. Stimuli were presented on a flat-screen CRT monitor set to a resolution of 1024 × 768 pixels and a refresh rate of 100 Hz. Text was drawn in 18-point Tahoma font and image patches subtended ~2.5° of visual angle. Trials were initiated using a button on the front of a gamepad controller and judgments were made by pressing the left and right triggers.

Search Procedure

A category name was displayed for 2500 ms, followed by a central fixation cross for 500 ms and finally a six-item search display (Figure 2). Items in the search display were image patches of objects arranged on a circle having a radius of 8°. There were 288 trials, half target-present and half target-absent. Target-present trials depicted a target and five distractor objects chosen from random non-target categories. Each participant saw one of the three selected exemplars for a given target category twice at each hierarchical level, with exemplars counterbalanced across participants. Half of the target-absent trials depicted six distractors; the other half depicted five distractors and one lure. Lures are needed to encourage encoding at the cued level (see Tanaka & Taylor, 1991). The lure was a categorical sibling of the cued target, drawn from target images one level above in the category hierarchy (e.g., a police car when cued with “taxi”, or a truck when cued with “car”). Lures at the superordinate level were drawn from other non-target categories, making them indistinguishable from the distractor objects.

Behavioral Results

Error rates differed between hierarchy conditions, $F(5,21) = 15.19$, $p < .001$, $\eta^2 = .378$. Post-hoc tests (LSD corrected) showed that accuracy on target-present trials at the superordinate level ($M = 84.9\%$, 95% CI [81.1, 88.7]) was lower than at the basic level ($M = 91.6\%$, 95% CI [89.5, 93.7]) and the subordinate level ($M = 92.3\%$, 95% CI [90, 94.6]), ps < .001. These additional misses are consistent with previous work (Maxfield & Zelinsky, 2012) and reflect participants occasionally failing to recognize a target as a member of the cued superordinate category (Murphy & Brownell, 1985). On target-absent trials, accuracy was lower at the subordinate level ($M = 89\%$, 95% CI [87, 91]) compared to the basic ($M = 96\%$, 95% CI [94.7, 97.3]) and superordinate ($M = 95.4\%$, 95% CI [93.2, 97.5]) levels, ps < .001. This increase in false positives was due to lures at the subordinate level being occasionally mistaken for the cued target category. Neither pattern of errors compromises our conclusions. Only correct trials were included in the subsequent analyses.

As in previous work (Castelhano, Pollatesk, & Cave, 2008, Schmidt & Zelinsky, 2009, Maxfield & Zelinsky, 2012), search performance was divided into target guidance and verification epochs and analyzed separately. Target guidance was defined in two ways: the time between search display onset and the participant’s first fixation on the target (time-to-target), and the proportion of trials in which the target was the first object fixated during search (immediate fixations). Target verification was defined as the time between a participant’s first fixation on the target and their correct target-present manual judgment.

Analyses of the initial guidance epoch of search revealed significant differences in time-to-target between conditions, $F(2,24) = 22.08$, $p < .001$, $\eta^2 = .508$. Targets cued at the subordinate level were fixated sooner on average than targets cued at the basic level, which were fixated sooner than targets cued at the superordinate level (*p*s ≤.021, Figure 3A, dark bars). This same trend held for immediate target fixations, $F(2,24) = 13.31$, $p < .001$, $\eta^2 = .456$ (Figure 3B, dark bars), a more conservative measure of guidance. Subordinate-level targets were first fixated more often than basic-level targets ($p < .001$), and basic-level targets were first fixated more often than superordinate-level targets ($p < .001$). Initial saccade latency did not reliably differ between cueing conditions ($p = .452$), suggesting that these differences were not due to a speed-accuracy tradeoff. Differences between conditions were also found during the verification epoch of search, $F(2,24) = 5.71$, $p = .006$, $\eta^2 = .215$. As shown in Figure 3C (dark bars), these differences took the form of a BSE; targets cued at the basic level were verified faster than those cued at the subordinate level ($p = .01$) and superordinate level ($p = .004$). These findings not only extend previous work in showing that the hierarchical level in which a target is cued differentially affects target guidance and verification processes (Maxfield & Zelinsky, 2012), they create a challenging guidance and verification behavioral ground truth against which our generative model of category representation can be evaluated.

Model Methods

Two distinct effects of category hierarchy were found in the behavioral data: a subordinate-level advantage in target guidance and a basic-level advantage in target verification. We explain both of these behavioral patterns using a single unsupervised generative model that extracts features from images of category exemplars and then reduces the dimensionality of this representation to obtain what we refer to as *Category-Consistent Features* (CCFs). Figure 4 provides an overview of this model.

Feature Extraction

Using the identical category hierarchy from the behavioral experiment, we built from ImageNet and Google Images an image dataset for model training. This consisted of 100 exemplars for each of the 48 subordinate-level categories (4,800 images in total; see Figure 1 for tiny views of these images), with each exemplar being an image patch closely cropped around the depicted object. Exemplars for basic-level and superordinate-level categories were obtained by combining the subordinate “children” exemplars under the “parent” categories. For example, the basic-level boat category had 300 exemplars consisting of 100 speed boats, 100 sail boats, and 100 cruise ships, and the superordinate-level vehicle category had 1,200 exemplars consisting of the 300 exemplars from each of the boat, car, truck, and plane siblings.

The first step in representing an object category is the extraction of features from exemplars. Two types of features were used: the Scale Invariant Feature Transform (SIFT) and a color histogram feature. SIFT features capture the structure of gradients in images using 16 spatially distributed histograms of scaled and normalized oriented-edge energy (Lowe, 2004). The color histogram feature (Van de Weijer & Schmid, 2006) captures the distribution of hue in an image, represented in the current implementation by 64 bins of Hue in 360° HSV color space. Using dense sampling (and discarding samples from uniform regions), we extracted 5 scales of SIFT descriptors from patches of 12×12, 24×24, 36×36, 48×48, and 60×60 pixels, and color histogram features from a fixed-size 20×20 pixel patch surrounding the center positions of every SIFT descriptor in each of the 4,800 exemplars. Color histograms were pooled over patches within exemplars to create a single 64-bin color histogram for each. However, to compare SIFT features between exemplars it is necessary to find a common feature space, and for this we used the Bag-of-Words (BoW) method (Csurka, Dance, Fan, Willamowski, & Bray, 2004). The SIFT features extracted from each exemplar were put into a metaphorical bag, and k-means clustering was performed on this bag to obtain a common vocabulary of 1,000 visual words (k = 1000). The 64 hue features from the color histogram were concatenated to this vocabulary, yielding a 1064-dimensional feature space in which each of the 4,800 exemplars could be represented as a BoW histogram, where the bins of the histogram correspond to the 1,064 visual word features and the height of each bin indicates the frequency of that feature in a given exemplar.

Category-Consistent Features (CCFs)

Having put all the category exemplars in a common feature space, the next step is to find those features that are most representative of each target category. This process begins by averaging over category the BoW exemplar histograms to obtain what might be called a proto-type for each category (Rosch, 1973), although we avoid using this theoretically-laden term so as not to associate a proto-type with a particular step in the computation of CCFs. Each averaged category histogram captures the mean frequency that each of the 1,064 features appeared in the category exemplars, along with the variance for each of these means (see Figure S2 in Supplemental Materials for a partial averaged histogram for the taxi category, and Figure S3A for a visualization of every complete histogram contributing to the taxi category averaged histogram).

Although methods abound in the computer vision literature for selecting features (e.g., Collins, Liu, & Leordeanu, 2005; Ullman, Vidal-Naquet, & Sali, 2002), most of these are tailored to finding features that discriminate between categories of objects for the purpose of classification. This makes them poorly aligned with our generative approach. Alternatively, feature selection under the CCF model is grounded in signal detection theory (Green & Swets, 1966). We assume that features having a high frequency and a low variance are more important than the rest, and use these simple measures to prune away the others. Specifically, features having a high mean frequency over the category exemplars are identified using the interquartile range rule: $X' = X > 1.5*(Q_3 (X)–Q_1 (X))$, where X is the average frequency of the 1,000 SIFT features or 64 color features (performed separately) for a given category histogram, and Q_1 and Q_3 are the first and third quartiles, respectively. For each of these frequently occurring features we then compute the inverse of its coefficient of variation by dividing its mean frequency by its standard deviation, a commonly used method for quantifying a scale-invariant signal-to-noise ratio (SNR; Russ, 2011). Finally, we weight each feature in the above set by its SNR, then perform k-means clustering, with k=2, on these feature weights to find a category-specific threshold to separate the important features from the less important features. The CCFs for a given category are those features falling above this threshold. CCFs are therefore the features that occur both frequently and reliably across the exemplars of a category, with each category having different CCFs in this 1064-dimensional feature space. These CCFs, and not the noisier category histogram formed by simply averaging exemplar histograms, are what we believe constitutes the learned visual representation of an object category (see Figure S3B for the CCFs from the taxi category, and how they compare to the corresponding averaged category histogram from Figure S3A).

Model Results

Can the CCF model capture the patterns of target guidance and verification observed in behavior? We show that these two very different patterns can be modeled as different properties of the same CCF category representations.

Target Guidance

The behavioral data showed that target guidance got weaker as targets were cued at higher levels in the category hierarchy. Guidance was strongest following a subordinate-level cue, weaker following a basic-level cue, and weakest following a superordinate-level cue. How does the CCF model explain target guidance, and its change across hierarchical level?

According to the CCF model, target guidance is proportional to the number of CCFs used to represent a target category. The logic underlying this prediction is straightforward. To the extent that CCFs are the important features in the representation of a category, more CCFs mean a better and more specific category representation (see also Schmidt, MacNamara, Proudfit, & Zelinsky, 2014). A target category having a larger number of CCFs would therefore be represented with a higher degree of specificity and, consequently, fixated more efficiently than a target having a sparser "template" (Schmidt & Zelinsky, 2009). As shown in Figure 5 (dark bars), the number of CCFs per category indeed varied with hierarchical level; the subordinate-level categories had the most CCFs, followed by the basic-level and finally the superordinate-level categories. This too was predicted. Subordinate-level categories have more details in common that can be represented and selected as CCFs, whereas at the higher levels greater variability between exemplars cause features to be excluded as CCFs, resulting in a smaller total number.

Figure 3 shows that the effect of hierarchical level on target guidance can be captured simply by the mean numbers of CCFs extracted for the 48 subordinate-level target categories, the 16 basic-level categories, and the 4 categories at the superordinate level. Specifically, the light bars in Figure 3A plot 1/CCF# to capture the increase in time-to-target with movement up the category hierarchy, while Figure 3B plots the raw numbers of CCFs to capture the downward trend in immediate target fixations. After linearly transforming the number of CCF data to put it into the same scales as the behavior, the model's behavior fell within the 95% confidence intervals surrounding all six of the behavioral means. This finding has implications for search theory. It suggests that the stronger target guidance reported for exemplar search (e.g., targets cued by picture preview) compared to categorical search (e.g., Schmidt & Zelinsky, 2009) may be due, not to a qualitative difference in underlying processes, but rather a quantitative difference in the number of "good" features in the target representation used to create the priority map that ultimately guides search (Zelinsky &

Bisley, 2015). Many strong guiding features can be extracted when the opportunity exists to preview the specific target exemplar, but strong guidance in a categorical search task requires a target category represented by many CCFs.

Target Verification

To the extent that more CCFs enable greater specificity in the target representation the converse is also true. Movement up the category hierarchy incurs a cost reflecting decreasing numbers of CCFs, with superordinate-level categories receiving the greatest cost, subordinate-level categories the least, and basic-level categories falling in between. We show that target verification can be modeled by combining this trend with a second and opposing trend, one based on the distance to neighboring categories.

Sibling Distance—In the context of a categorical search task, target verification refers to the time between first fixation on the target and the correct target-present judgment. The CCF model predicts that this time is proportional to the distance between the CCFs of the target category and the features of the target's categorical siblings, where siblings are defined as categories sharing the same parent (one level up in the category hierarchy). This logic is also straightforward. Verification difficulty should depend on the distance between the target category and the most similar non-target categories in a test set; as this distance increases, target verification should become easier. This follows from the fact that smaller distances create the potential for feature overlap between categories, and to the extent this happens one category might become confused with another. In the present context, these least distant and most similar non-target exemplars would be the categorical siblings of the target. If the target was a police car the non-target objects creating the greatest potential for confusion would be exemplars of race cars and taxis, with these objects largely determining the verification difficulty of the target. Indeed, these siblings were the same objects used as categorical lures in order to obtain our behavioral demonstration of a basic-level advantage.

To model the distance between a target and its categorical siblings we took the CCF histogram for each sibling and found the mean chi-squared distance between it and the BoW histogram for every exemplar under the parent category. We denote the full set of BoW features as F = {1,…,1064}, and the CCFs for target category k as F′, such that $k \in \{1,\ldots,68\}$ and F'_k is a subset of F, $F'_k \subseteq F$. Chi-squared distance is defined by:

$$\chi^2(x,y)=\frac{1}{2}\sum_i \frac{(\phi_i(x)-\phi_i(y))^2}{\phi_i(x)+\phi_i(y)}, \quad \text{Eq. 1}$$

where x and y are the two histograms to be compared, and φ_i is the value at the i^th^ bin of the 1064-bin feature histogram. Note, however, that following the dimensionality reduction that occurred in selecting the CCFs the sibling CCF histograms may no longer be in the same feature space as the BoW histograms for the exemplars. To compute the above-described distances we therefore must put the CCF and BoW histograms back into a common feature space, and we do this by adopting the following algorithm. For comparisons between a given CCF histogram of category k and its own BoW histogram exemplars, chi-squared distances were computed for only those bins in the BoW histograms for which there were corresponding bins in the CCF histogram, such that $i \in F'_k$. For comparisons between exemplar histograms from category *j* and histograms from a sibling category, *k*, chi-squared distances were limited to the feature space formed by the union of the two CCF histograms, such that $i \in \cup(F'_j, F'_k)$ for Eq. 1.

To clarify with an example, consider only two sibling categories, A and B, each having non-identical CCF bins ($F'_A \neq F'_B$) forming CCF histograms μ(A) and μ(B) based on exemplars A_n and B_n, where n describes all of the exemplars for a given category (either 100, 300, or 1200 for the subordinate, basic, or superordinate categories, respectfully, used in this study). We compute the chi-squared distances between μ(A) and the BoW histograms obtained for each of A's exemplars, A_n, for which there are corresponding bins in F'_A. If we denote this distance between the CCF histogram of A and all the A exemplar histograms as $d_{A,A}$, then $d_{A,A} = \chi^2(\mu(A), A_n \mid i \in F'_A))$. We also compute the chi-squared distances between μ(A) and the BoW histograms obtained for each of B's exemplars, $d_{A,B}$, with these comparisons now limited to the bins forming the union of the F'_A and F'_B CCF histograms, such that $d_{A,B} = \chi^2(\mu(A), B_n \mid i \in \cup(F'_A, F'_B))$. Doing the same for μ(B) and the BoW histograms of the B exemplars and the A exemplars (based on the union of CCFs F'_A and F'_B), gives us $d_{B,B}$ and $d_{B,A}$, respectively. Finally, taking the mean over the hundreds of distances in the sets $d_{A,A}$, $d_{A,B}$, $d_{B,B}$, and $d_{B,A}$, we obtain an estimate of the distance between sibling categories A and B, which we refer to as *sibling distance*.

To the extent that smaller sibling distances mean more difficult category verification decisions, the CCF model predicts a verification benefit for target categories designated at higher levels in the hierarchy. Computing sibling distances for all 64 target categories, then averaging within hierarchical level, we found that subordinate-level categories were closest to their sibling exemplars and that superordinate-level categories had the largest mean sibling distance (Fig 5, light bars). Verification times for race cars should therefore be relatively long due to the proximity of this category to taxi and police car exemplars, whereas shorter verification times are predicted for vehicles because of this category's greater mean distance to Oreo cookies and other sibling exemplars. The basic-level categories again fall between these two, enjoying neither a verification cost nor a benefit.

Basic-level Superiority Effect—Rather than the predicted speed-up in target verification times with movement up the hierarchy, we found instead the often-observed basic-level advantage; faster verification for targets cued at the basic level compared to the subordinate or superordinate levels. However, and consistent with early explanations (Murphy & Brownell, 1985), we explain this BSE as a tradeoff between two interacting processes, *specificity*, which we relate to the number of CCFs existing for a given category, and *distinctiveness*, which we relate to the distance between the CCFs of a given category and the features of its sibling exemplars. Indeed, the countervailing trends illustrated in Figure 5 reflect these opposing specificity and distinctiveness processes. To model the net impact of these interacting processes on target verification time we simply multiple one by the other. Specifically, for each target category we multiply its number of CCFs by its sibling distance to obtain a (unit-less) estimate of that category's verification difficulty. These results, averaged by hierarchical level and linearly transformed into the behavioral scale, are shown in Figure 3C (light bars). As was the case for target guidance, model estimates once again fell within the 95% confidence intervals surrounding the behavioral means. In a control experiment we also showed that randomly selecting the same numbers of visual word features failed to produce the BSE observed in behavior, thereby validating the CCF model —any features will not do, these features have to be CCFs (see Figure S4).

Although categories at the subordinate level have the most CCFs (a specificity benefit), they also have the smallest sibling distance (a distinctiveness cost). This results in an intermediate degree of verification difficulty. Superordinate-level categories have the opposite relationship, relatively few CCFs (a specificity cost) but a large sibling distance (a distinctiveness benefit). This, again, results in an intermediate degree of verification difficulty. The basic-level categories occupy a privileged position in the hierarchy that avoids these two extremes. They have a relatively high number of CCFs while also being relatively distant from their sibling exemplars. This favorable trade-off between distinctiveness and specificity produces the BSE, faster verification at the basic level relative to the levels above and below.

Predicting search behavior using the CCF model

The above-described analyses demonstrated that the CCF model captured trends observed in target guidance and verification across the superordinate, basic, and subordinate levels, but can this model also predict behavior occurring within each of these categorical levels? As a first step towards answering this question, we conducted a trial-by-trial analysis to predict how strongly the cued target category would guide search to an exemplar of that target. For each target-present trial, we computed the chi-squared distance between the CCF representation of the target category and the target exemplar appearing in the search display, then correlated these distances with the time-to-target measure of search guidance obtained for every trial. To evaluate the CCF model predictions we used the leave-one-out method to derive a Subject model, which indicates how well the mean target guidance of n-1 subjects predicts the guidance of the subject left out. This analysis provides an upper limit on the predictive success of the CCF model, as correlations higher than the Subject model would not be expected given subject variability in their guidance behavior. Figure 6 plots time-to-target correlations for the CCF model and the corresponding Subject model at each hierarchical level. Paired-group t-tests revealed that correlations did not reliably differ between the CCF and Subject model at the subordinate ($p = .078$) or basic ($p = .334$) levels, although correlations were significantly different at the superordinate level ($p < .001$). The poor correlation at the superordinate level is consistent with the absence of guidance reflected in the chance-level proportion of immediate target fixations at this level (Figure 3B). These findings suggest that the CCF model not only predicted the fine-grained search behavior occurring on individual trials, these predictions at the subordinate and basic levels were as good as could be expected given agreement in the participants' behavior.

Conclusion

Categories determine how we interact with the world. Understanding the forces that shape category representation is therefore essential to understanding behavior in every domain of psychological science. We introduce a computational model of category representation, one that accepts image exemplars of common object categories and finds the features appearing frequently and consistently within each category's exemplars—referred to here as *category-consistent features* (CCFs).

We validated the CCF model through comparison to behavior in a categorical search task. Categorical search is important, and has diverse applications. Each time a security screener searches for a weapon, or a radiologist searches for a tumor, they are engaging in categorical search. Categorical search is also unique in that this single task enables study of the representations used to guide attention to categorically-defined targets and the representations underlying the recognition of these objects as members of the target category. We manipulated the hierarchical level in which categorical targets were cued and found that these attention and recognition processes were expressed in very different behavioral patterns. One pattern was a subordinate-level advantage in target guidance; targets cued at the subordinate level were preferentially fixated compared to targets cued at the basic or superordinate levels. Another pattern was a basic-level advantage in target verification; fixated objects were verified faster as members of the target category following a basic-level cue compared to subordinate or superordinate-level cues.

Under the CCF model, both patterns depend on the number of CCFs extracted from exemplars at each hierarchical level. Target guidance weakens with movement up the category hierarchy due to exemplar variability at the higher levels restricting the formation of CCFs, resulting in less effective target templates for guiding search (Olivers, Peters, Roos, & Roelfsema, 2011). The CCF model advances existing search and visual working memory theory by making explicit the processes of extracting visual features from image exemplars of real-world categories and consolidating these features into lower-dimensional category representations (CCFs) that can be used to guide search. It also provides a theory for understanding effects of category hierarchy (Maxfield & Zelinsky, 2012) and target specificity (Schmidt & Zelinsky, 2009) on search behavior; search is guided more efficiently to targets specified lower in the category hierarchy because these objects would usually be represented using more CCFs. Target verification was modeled as a multiplicative interaction between CCF number and sibling distance—a measure of similarity between the CCFs of a target category and the features of its sibling exemplars. In doing this, the CCF model appealed to the core principles of specificity and distinctiveness that have been guiding categorization research for decades (Murphy & Brownell, 1985). The number of CCFs maps onto the idea of specificity. Subordinate-level categories are the most specific because they give rise to many CCFs. Sibling distance maps onto the idea of distinctiveness. Verification suffers with movement down the hierarchy because target representations start to share too many features with their closest categorical neighbors. The CCF model advances categorization theory by making these core principles computationally explicit and applicable to real-world object categories.

Of potentially even broader theoretical significance is the question of whether search and categorization share the same target representation; are the visual features used to guide overt attention to a categorical target in a search display the same as those used to categorize the target once it is fixated? The CCF model suggests that this is the case, and to the extent that this suggestion is supported through converging evidence (Zelinsky et al., 2013) a strong theoretical bridge will be built between the attention and categorization literatures. Future work will also strengthen the bridge to the computer vision and computational neuroscience literatures by attempting to learn CCFs using a deep convolutional neural network (CNN). Supervision is a powerful learning tool (Khaligh-Razavi & Kriegeskorte, 2014), and combining it with the generative extraction of features from exemplars may lead to significant advances in the understanding of category representation.

The CCF model makes possible the rigorous study of how visual object categories can be learned and represented from the vast numbers of diverse image exemplars accumulated throughout our everyday experience. Recent decades have seen scientific doors to the real world open for many psychological processes. The CCF model opens another such door into categorization.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Naomi Vingron, Richard Molander, and Ren-Ann Wang for their help with data collection, and professors Christian Luhmann, Dimitris Samaras, and Greg Murphy for invaluable conceptual input. GJZ was supported by a grant from the National Institutes of Health (R01-MH063748) and National Science Foundation grants IIS-1111047 and IIS-1161876.

References

Anderson JR. A spreading activation theory of memory. Journal of Verbal Learning and Verbal Behavior. 1983; 22:261–295.

Anderson JR. ACT: A simple theory of complex cognition. American Psychologist. 1996; 51(4):355.

Ashby FG, Maddox WT. Human category learning. Annu Rev Psychol. 2005; 56:149–178. [PubMed: 15709932]

Ashby FG, Maddox WT. Relations between prototype, exemplar, and decision bound models of categorization. Journal of Mathematical Psychology. 1993; 38:423–466.

Castelhano MS, Pollatsek A, Cave K. Typicality aids search for an unspecified target, but only in identification, and not in attentional guidance. Psychonomic Bulletin and Review. 2008; 15:795–801. [PubMed: 18792506]

Chin-Parker S, Ross BH. Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004; 30:216–226.

Collins RT, Liu Y, Leordeanu M. Online selection of discriminative tracking features. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2005; 27(10):1631–1643.

Collins AM, Quillian MR. Retrieval time from semantic memory. Journal of verbal learning and verbal behavior. 1969; 8(2):240–247.

Csurka G, Dance C, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision. 2004; 1:22.

Duda, RO.; Hart, PE.; Stork, DG. Pattern classification. John Wiley & Sons; 2012.

Goldstone RL. The role of similarity in categorization: Providing a groundwork. Cognition. 1994; 52(2):125–157. [PubMed: 7924201]

Green, D.; Swets, J. Signal detection theory and psychophysics. New York: Krieger; 1966.

Kaplan AS, Murphy GL. Category learning with minimal prior knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2000; 26:829–846.

Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology. 2014; 10(11):e1003915. [PubMed: 25375136]

Kruschke JK. ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review. 1992; 99(1):22. [PubMed: 1546117]

Kurtz KJ. Human category learning: Toward a broader explanatory account. Psychology of Learning and Motivation. 2015; 63:77–114.

Levering KR, Kurtz KJ. Observation versus classification in supervised category learning. Memory & Cognition. 2014; 43:266–282. [PubMed: 25190494]

Liberman AM, Harris KS, Hoffman HS, Griffith BC. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology. 1957; 54(5):358. [PubMed: 13481283]

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

Love BC, Medin DL, Gureckis TM. SUSTAIN: a network model of category learning. Psychological Review. 2004; 111(2):309. [PubMed: 15065912]

Maxfield JT, Stadler W, Zelinsky GJ. The Effects of Target Typicality on Guidance and Verification in Categorical Search. Journal of Vision. 2014; 13(9):524–524.

Maxfield JT, Zelinsky GJ. Searching through the hierarchy: How level of target categorization affects visual search. Visual Cognition. 2012; 20:10, 1153–1163.

Medin DL, Schaffer MM. Context theory of classification learning. Psychological Review. 1978; 85(3):207.

Mervis CB, Rosch E. Categorization of natural objects. Annual Review of Psychology. 1981; 32(1): 89–115.

Murphy GL, Brownell HH. Category differentiation in object recognition: Typicality constraints on the basic category advantage. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1985; 11:70–84.

Murphy, GL. The big book of concepts. Cambridge, MA: MIT Press; 2002.

Nosofsky RM. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General. 1986; 115(1):39–57. [PubMed: 2937873]

Nosofsky RM, Palmeri TJ. An exemplar-based random walk model of speeded classification. Psychological Review. 1997; 104(2):266. [PubMed: 9127583]

Olivers C, Peters J, Roos H, Roelfsema P. Different states in visual working memory: when it guides attention and when it does not. Trends in Cognitive Sciences. 2011; 15:327–334. [PubMed: 21665518]

Pazzani M. The influence of prior knowledge on concept acquisition: Experimental and computational results. Journal of Experimental Psychology: Learning, Memory & Cognition. 1991; 17(3):416–432.

Regier T, Kay P. Language, thought, and color: Whorf was half right. Trends in Cognitive Sciences. 2009; 13(10):439–446. [PubMed: 19716754]

Rosch EH, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognitive Psychology. 1976; 8:382–439.

Rosch, EH. Principles of categorization. In: Rosch, E.; Lloyd, BB., editors. Cognition and categorization. Hillsdale, NJ: Erlbaum; 1978. Reprinted in: Margolis, E. and Laurence, S. (Eds.) (1999). Concepts: Core readings. Cambridge, MA: MIT Press

Rosch EH. Natural Categories. Cognitive Psychology. 1973; 4(3):328–350.

Russ, JC. The image processing handbook. CRC press; 2011.

Schmidt J, Zelinsky GJ. Search guidance is proportional to the categorical specificity of a target cue. Quarterly Journal of Experimental Psychology. 2009; 62:1904–1914.

Schmidt J, MacNamara A, Proudfit GH, Zelinsky GJ. More target features in visual working memory leads to poorer search guidance: Evidence from contralateral delay activity. Journal of Vision. 2014; 14(3):8. [PubMed: 24599946]

Snodgrass JG, Vanderwart M. A standardized set of 260 pictures: Normed for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory. 1980; 6:174–215. [PubMed: 7373248]

Tanaka JW, Taylor M. Object categories and expertise: Is the basic level in the eye of the beholder? Cognitive Psychology. 1991; 23:457–482.

Ullman S, Vidal-Naquet M, Sali E. Visual features of intermediate complexity and their use in classification. Nature neuroscience. 2002; 5(7):682–687. [PubMed: 12055634]

Ulusoy I, Bishop CM. Generative versus discriminative methods for object recognition. In Computer Vision and Pattern Recognition. 2005; 2:258–265.

Van De Weijer J, Schmid C. Coloring local feature extraction. In European Conference on Computer Vision. 2006; 2:334–348.

Zelinsky GJ, Adeli H, Peng Y, Samaras D. Modelling eye movements in a categorical search task. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2013; 368(1628):20130058. [PubMed: 24018720]

Zelinsky GJ, Bisley JW. The what, where, and why of priority maps and their interactions with visual working memory. Annals of the New York Academy of Sciences. 2015; 1339:154–164. [PubMed: 25581477]

Zelinsky GJ, Peng Y, Berg AC, Samaras D. Modeling guidance and recognition in categorical search: Bridging human and computer object detection. Journal of Vision. 2013; 13(3):30. [PubMed: 24105460]

Zelinsky GJ, Peng Y, Samaras D. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. Journal of Vision. 2013; 13(14):10. [PubMed: 24338446]

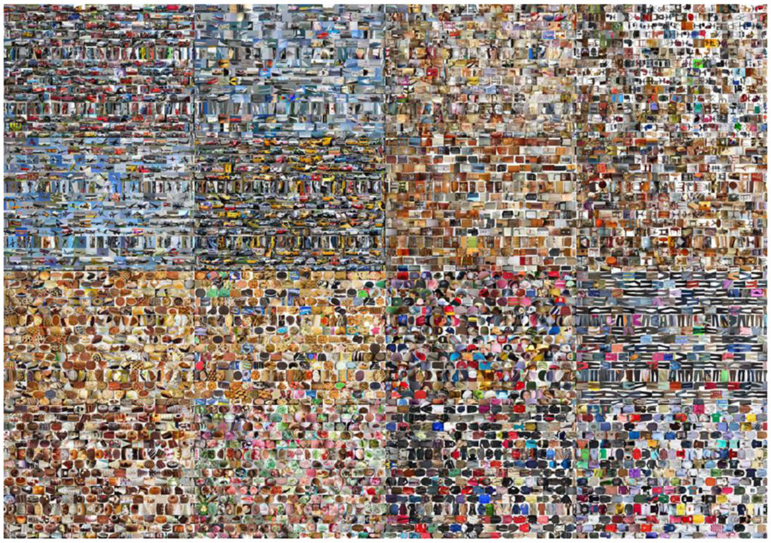

Figure 1.

Most of the 4,800 images used to train our model, grouped into the 16 basic-level categories used as stimuli. Images are shown as tiny thumbnails for illustration, but each was minimally 100×100 pixels and depicted a tightly cropped view of an object against a natural background. Common visual features among the 300 image exemplars of each category, and category-specific differences between these features, create the appearance of rectangles in this stimulus space. See Supplemental Materials for similar illustrations of exemplars grouped randomly (Figure S1A), at the superordinate level (Figure S1B), and at the subordinate level (Figure S1C).

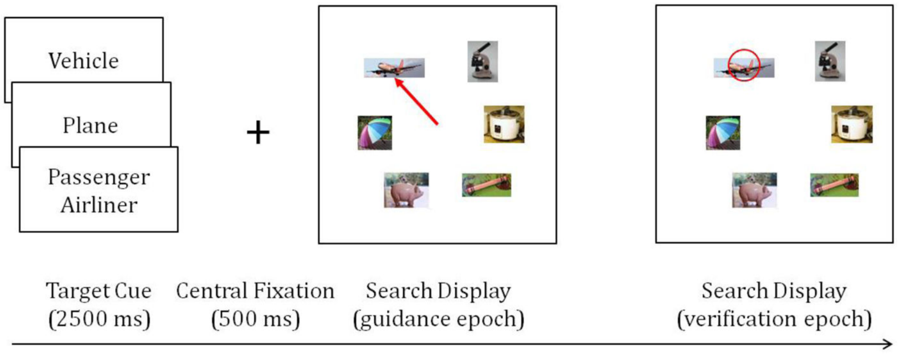
Vehicle
Plane
Passenger Airliner
+
Target Cue (2500 ms)
Central Fixation (500 ms)
Search Display (guidance epoch)
Search Display (verification epoch)

Figure 2.
Procedure for the categorical search task. A target was designated by category name at one of three hierarchical levels, followed after a delay by a six-item target-present/absent search display. The target guidance and verification epochs used for analysis are indicated by the red graphics (not shown to participants) superimposed over the search display.

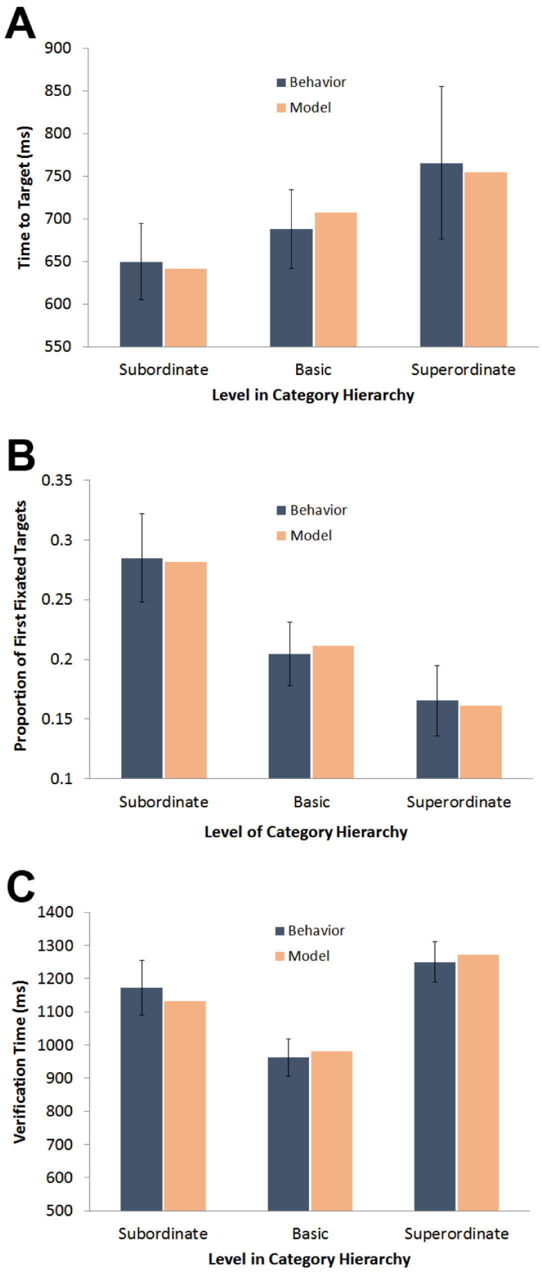

Figure 3.

Behavioral results (dark bars) for the categorical search experiment plotted with the CCF model output (light bars) for (A) time to the first fixation on the target, (B) proportion of immediate fixations on the target, and (C) time from first fixation on the target until the correct target-present button press decision. Error bars indicate 95% confidence intervals.

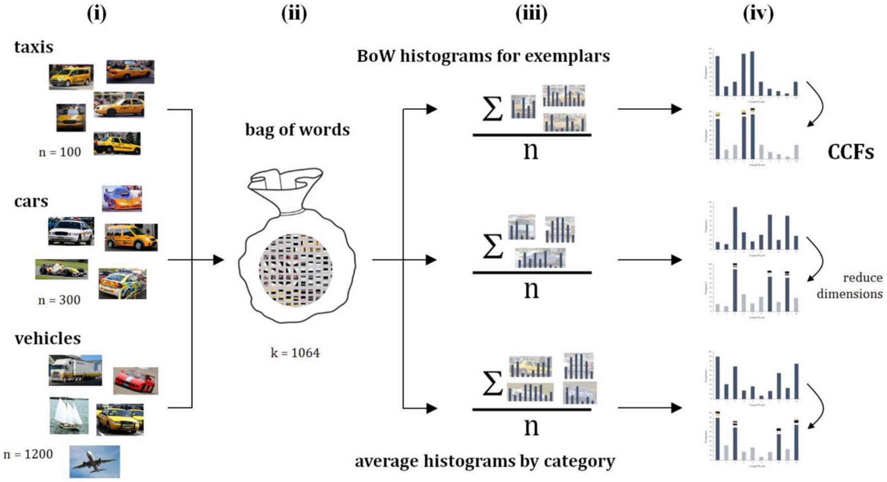

Figure 4.

An overview of the Category-Consistent Features Model. (i) 100 images of object exemplars were collected for 48 subordinate-level categories. These exemplars were combined to create 16 basic-level categories (each with 300 exemplars) and 4 superordinate-level categories (each with 1200 exemplars). (ii) SIFT and color histogram features were extracted from each exemplar and the Bag-of-Words (BoW) method was used to create from these a common feature space consisting of 1064 "visual words". (iii) 1064-bin BoW histograms were obtained for each exemplar, where the bins correspond to the visual words and bin height indicates the frequency of each feature in the exemplar image. BoW histograms were averaged by category to obtain 68 averaged histograms, each now having a mean frequency and variability associated with each visual word. (iv) Features in these averaged histograms having too low of a frequency or too high of a variability were excluded, resulting in a lower-dimensional feature representation of each category that we refer to as category-consistent features (CCFs)—those highly informative features that are present both frequently and consistently across the exemplars of a category.

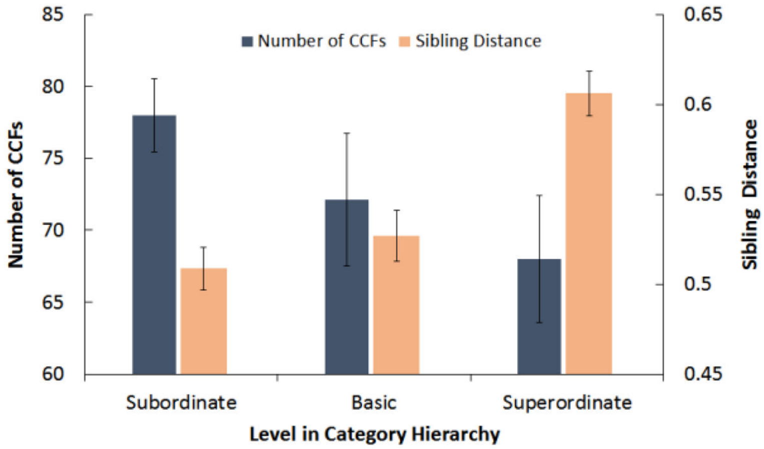

Figure 5.

Results from the CCF model showing the mean number of category-consistent features (dark bars) and mean sibling distances (light bars) by hierarchical level. Error bars indicate standard error of the mean, computed by treating the number of categories at each level as the number of sample observations (n).

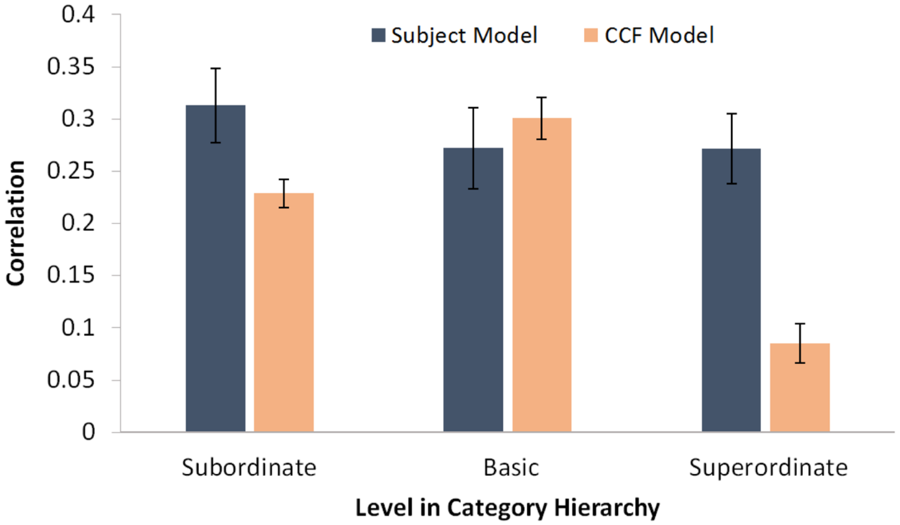

Figure 6.

Correlations between the trial-by-trial CCF model predictions and time-to-target (light bars), averaged by level (Fisher *z*-transformation) and plotted with correlations from a corresponding Subject model (dark bars) that captures agreement among participants in their guidance behavior. Data were from all correct trials in which the target was fixated. Error bars indicate one standard error.

Table 1

Object categories grouped by hierarchical level.

Superordinate	Basic	Subordinate
Vehicle	Car	Police Car
		Taxi
		Race Car
	Boat	Sail Boat
		Cruise Ship
		Speed Boat
	Plane	Passenger Airliner
		Biplane
		Fighter Jet
	Truck	18 Wheeler
		Fire Truck
		Pickup Truck
Furniture	Cabinet	Kitchen Cabinet
		Filing Cabinet
		China Cabinet
	Chair	Folding Chair
		Office Chair
		Dining Room Chair
	Bed	Twin Bed
		Canopy Bed
		Bunk Bed
	Table	Coffee Table
		Dining Room Table
		End Table
Clothing	Pants	Jeans
		Dress Pants
		Pajama Pants
	Shirt	Dress Shirt
		T-shirt
		Long Sleeve Shirt
	Hat	Baseball Hat
		Knit Cap
		Cowboy Hat
	Jacket	Winter Jacket
		Windbreaker
		Trench Coat
Dessert	Ice Cream	Chocolate Ice Cream
		Mint Choc. Chip Ice Cream
		Strawberry Ice Cream

Superordinate	**Basic**	**Subordinate**
	Pie	Pecan Pie
		Blueberry Pie
		Lemon Meringue Pie
	Cookie	Oreo
		Chocolate Chip Cookie
		Sugar Cookie
	Cake	Chocolate Cake
		Wedding Cake
		Bundt Cake