Stony Brook University

Modeling guidance and recognition in categorical search: bridging human and computer object detection

A Thesis Presented

by

Yifan Peng

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Master of Science

in

Department of Computer Science

Stony Brook University

December 2012

Stony Brook University

The Graduate School

Yifan Peng

We, the thesis committee for the above candidate for the

Master of Science degree, hereby recommend

acceptance of this thesis.

Dimitris Samaras
Associate Professor, Department of Computer Science

Gregory Zelinsky
Associate Professor, Department of Psychology

Tamara Berg
Assistant Professor, Department of Computer Science

This thesis is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Thesis

Modeling guidance and recognition in categorical search: bridging human and computer object detection

by

Yifan Peng

Master of Science

in

Department of Computer Science

Stony Brook University

December 2012

Although various object detection methods have been widely studied, state-of-the-art performance of object detectors still lag far behind human performance. Humans can perform object detection tasks on various object categories hundreds of times a day in an effortless manner. The main effort in computer vision community is aiming at improving the performance of object detectors, while on the other side only little research has been done on understanding how humans perform in the object detection process.

In this thesis, we analyze the relationship between human behaviors and object detection methods in computer vision on both guidance and recognition task. In our experiment, human observers searched for a categorically-defined teddy bear or butterfly target among non-targets rated as having HIGH, MEDIUM or LOW visual similarity to target classes. Actual targets show very strong search guidance, measured by the first fixated objects. Also guidance to non-targets objects are in proportion to their visual similarity to the target; high-similarity objects were first fixated the most and low-similarity objects the least. We design several computational experiments:

First, we propose a computational model that uses C2 features and SVMs in the context of Target Acquisition Model (TAM), to model human behavior in an object detection task. Eye movement behavior of our computation model matched human behavior almost perfectly, showing strong guidance to targets and same pattern of first fixation on target-similar objects. We conclude that categorical search is guided, and that driving this guidance are visual similarity relationships that can be quantified in terms of distance from a SVM classification boundary.

Second, we train and evaluate computational vision models for object category recognition and compare their output to the human behavior. Some algorithms do well at predicting which object humans will fixate first, but there are differences between which features perform best for classification and which predict human behavior most closely. This is a critical question for developing visual search algorithms that produce

perceptually meaningful results.

In additional, we demonstrate that the information available in the fixation behavior of subjects is often sufficient to decode the category of their search target--essentially reading a person's mind by analyzing what they look at using a technique that we refer to as behavioral decoding. Our results show we can predict an observer's search target based on their fixation pattern using two SVM-based classifiers, especially when one of the distractors were rated as being visually similar to the target category. These findings have implications for the visual similarity relationships underlying search guidance and distractor rejection, and demonstrate the feasibility in using these relationships to decode a person's task or goal.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Object detection has been widely studied in computer vision research for years. It involves not only verifying the presence of object, but also giving localization information of that object. Object detection has numerous applications in computer vision areas, such as image retrieval and video surveillance. The detection of a particular category, such as face detection and pedestrian detection, has been well-studied for years and a lot of excellent models are proposed[1, 2]. However, these models usually add lots of categorical-based information to boost performance of detection on a single class. Therefore they are not generalized for various object categories. Since objects in rich categories have significant variability, due to illumination, view point variation, occlusion, change in scales, deformable shapes, and intra-class variability, how to perform object detection with high accuracy in a short time is still a big challenge in the computer vision field.

The main effort on object detection study in the computer vision research community aims at improving the performance of detectors. However, despite recent advances in computer vision, state-of-the-art object detection methods still lag far behind human performance. Human observers perform object detection task in real scenes everyday with little error and in an effortless manner. That brings the importance of study the process of object detection of human both to the behavioral and computational vision study. Behaviorally, such search tasks are performed hundreds of times each day, and are widely believed to engage attention processes that underlie much of human behavior. Computationally, it is still unknown how this task can be accomplished with such robustness and efficiency.

In the mean time, behavioral community has studied human eye fixation for decades. During a typical search task, a human observer is asked to look for a target object in a given scene and press a button upon detecting a target. These experiments are measured in term of human's speed and accuracy in pressing the button. But this mechanism focused on the end of detecting process rather than process leading up to detect the object. Consequently, there has been fierce debate among behavioral theories of search [3--7], with part of this debate arising from the failure of button press responses to directly measure the hypothetical movements of attention believed to underlie detection.

In recognition of this problem, some efforts have been focusing on the eye movement in the analysis of human behaviors. Eye fixations are captured by eye trackers when showing a search scene to a human observer. There are some works on bottom-up saliency model, which predict the fixation sequence or fixation regions of human observers by defining a saliency map using low-level features. The most recent works focus on adding top-down information to the model, to predict human's fixation region when given a target-specified detection task. These works show a close fit to human behavioral on regions of interested, but giving no sequence of eye movement during the detection task.

The sequence of eye movements would provide a detailed description on how detection process is guided. Not only which object is fixated in the experiment, but also the sequences of fixation and how fixation is guided is shown in the detection process. Such eye movement measures have revealed a wealth of information about search, including descriptions of how the spatio-temporal distribution of attention changes as search converges on a target [8--10], how search can be segregated into distinct target guidance and target verification stages [11--13] and how visual similarity relationships cause target-similar objects to attract more early eye movements than objects less visually similar to a target [14--16]. This latter characterization of the visual confusability between objects is particularly important to understand search, irrespective of whether the searcher is a human or a computer vision system.

Studying eye movements allows human visual search to be decomposed into (at least) two distinct stages [11--13]: one a very efficient guidance of the high-resolution fovea to an object suspected of being the target, and the other the actual recognition of that object as a target or non-target. Efficient guidance requires estimates of visual similarity relationships between the features of the target class and those of the objects appearing in an image [16]; the better this match, the stronger the guidance signal. To some extent guidance and recognition have access to different information about an object. Guidance, by definition, applies to objects that have not yet been fixated, meaning that they are being viewed in peripheral vision and are therefore *blurred*. Recognition is thought to occur after an object is fixated, and therefore has access to a non-blurred view of an object.

In this paper, we study human behavior during visual search for object category targets. We show how well our computational models agree with the human behavior, focusing on whether similar objects are found to be confusing by people and by the computer vision models.

In the following chapter, we provide a survey on various methods of object detection in Section 2.1. We provide a survey on models that simulate human fixation in Section 2.2. In Chapter 3, we describe our behavior experiment on object similarity and visual guidance. We use a large-scale web-based experiment to collect visual similarity estimates between random objects and these two target classes (see also [16]), then construct from these similarity estimates search displays consisting of objects with known similarity relationships to the target classes in Section 3.1.We show how eye tracking data are collected during visual search for categorical targets in Section 3.2.

We design several computational experiments in 4. First we propose an eye movement model using C2

feature and color feature with probabilistic SVMs for object detection combine with the TAM for fixation prediction in 4.2. Second, we train and evaluate multiple computer vision models for object recognition and compare these models with human behavior for confusing non-target objects in 4.3.

In addition we are interested in if it is possible to infer what a person is thinking or doing by analyzing their patterns of eye fixations. We introduce the idea of *behavioral decoding*, the use of fine-grained measures of a person's behavior to infer their thoughts, goals, or mental states. Like neural decoding, behavioral decoding assumes that the execution of a complex cognitive behavior requires the coordination and use of more elemental cognitive operations that are expressed during the performance of the more complex task. To the extent that this is true, it may be possible to isolate and decode from these elemental operations the higher level task or goal. Given an unknown target category, is it possible to read a person's mind to reveal the target of their search by decoding their eye movements?

We combined behavioral and computational techniques to answer this question. Behaviorally human subjects are performed separate bear and butterfly search tasks during which their eye movements were recorded. On trials in which no target was presented, the bear and butterfly search tasks were identical (i.e., the same distractors in the same locations). Our behavioral experiment is described in 3. Computationally, we trained two discriminative models to recognize objects from two target classes, teddy bears and butterflies/moths. We show we can classify whether a person was searching for a bear or a butterfly in 4.4 based on the distractor objects fixated by subjects in these target-absent displays.

Chapter 2

Related work

2.1 Object detection models

In this part we first provide a survey on local features to represent images. Then we describe three popular models in object detection field: bag-of-feature model, part-based model, and boosting-based detector.

2.1.1 Local Features and Global Features

Local and global features have played an important role in improving the performance of object recognition methods. These features are typically invariant to illumination changes and small deformations. Stable local feature detection and representation are very important to many image registrations and object recognition algorithms.

2.1.1.1 Histogram of Gradient

Dalal and Triggs proposed histogram of gradient (HOG) feature in [17]. To compute HOG feature, an image is partitioned into 8×8 blocks. A histogram of gradient is computed for each block. Then HOG features are computed in different resolution as in a pyramid. The array of weighted feature from a detector window is sent to Linear SVM to train a detector. They use a single filter based on HOG features to represent an object category. For detecting object, they use a sliding windows approach to apply this filter to all positions in an image under different scales. For each window, the filter will output a score based on matching. They test the detector on MIT pedestrian detection dataset [18] and get nearly perfect separation between non-pedestrian/pedestrian.

2.1.1.2 SIFT descriptor

SIFT feature (Scale invariant feature transform) is introduced by Lowe in [19] and has been widely used in computer vision field. The detector and descriptor are designed carefully as a combination with excellent performance shown in many literatures. The detector part uses Difference of Gaussian filter (DoG), a scale invariant detector. An image is scaled into different octaves. In each octave, one image is blurred with Gaussian filters with different kernel size. Then difference of Gaussian image is generated from difference between two neighboring Gaussian blurred images in the same octave. Keypoints are extracted from DoG images as local minima/maxima of DoG image across scales. By comparing each pixel in the DoG image to its eight neighbors at the same scale and nine corresponding neighboring pixel in its neighboring scales, only if the point is the minimum or maximum among all neighboring pixels, the point will be selected as key point candidates. These candidates are processed by keypoints localization step, which removes unstable keypoints from candidates. First all points selected from DoG image are projected back to the corresponding location of image by interpolation of nearby data. Then keypoints with low contrast are removed and responses along edges are eliminated. Then finally each keypoint is assigned with an orientation.

A gradient orientation histogram is computed in the neighborhood of the keypoint (using the Gaussian image at the closest scale to the keypoint's scale). The contribution of each neighboring point is given with Gaussian weights. The peak of the orientation histogram is selected as the main orientation of the keypoints. For each keypoint, SIFT descriptors are extracted as following: A 4×4 pixel neighbors is selected around the keypoint and a feature descriptor is computed as a set of orientation histograms. The magnitude is also weighted by a Gaussian weight with regarding to distance to the keypoint. Since there are 8 bins in the orientation histogram, the final descriptor has $4\times4\times8 = 128$ dimensions. SIFT is a very robust feature descriptor that it is invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint.

Lowe also use SIFT feature for shape matching in [20]. In this work SIFT feature is extracted for a query image and comparing the object models in the model dataset. Given the large dataset and high-dimensional features, they provided Best Bin First (BBF), a variant of k-d tree search algorithm that accelerates indexing in higher-dimension dataset. They integrated a fully developed recognition system to detect object fast and in complex scene.

There are some extension works of local feature based on SIFT feature:

SURF (Speeded Up Robust Feature) is a fast version of SIFT feature. It is introduced by Bay et al. in [21]. It is a fast version of SIFT which also has the detector and descriptor part. It uses an integer approximation to the determinant of Hessian blob detector to find keypoints. For representation of a feature descriptor, it uses the sum of the Harr wavelet response around keypoints. Since both detection part and feature extraction part can be accelerated by using integral map, SURF is several times faster than SIFT feature. And the SURF feature is more robust against different image transformations.

Ke and Sukthankar proposed PCA-SIFT in [22]. They improved SIFT by examining the feature descriptor of

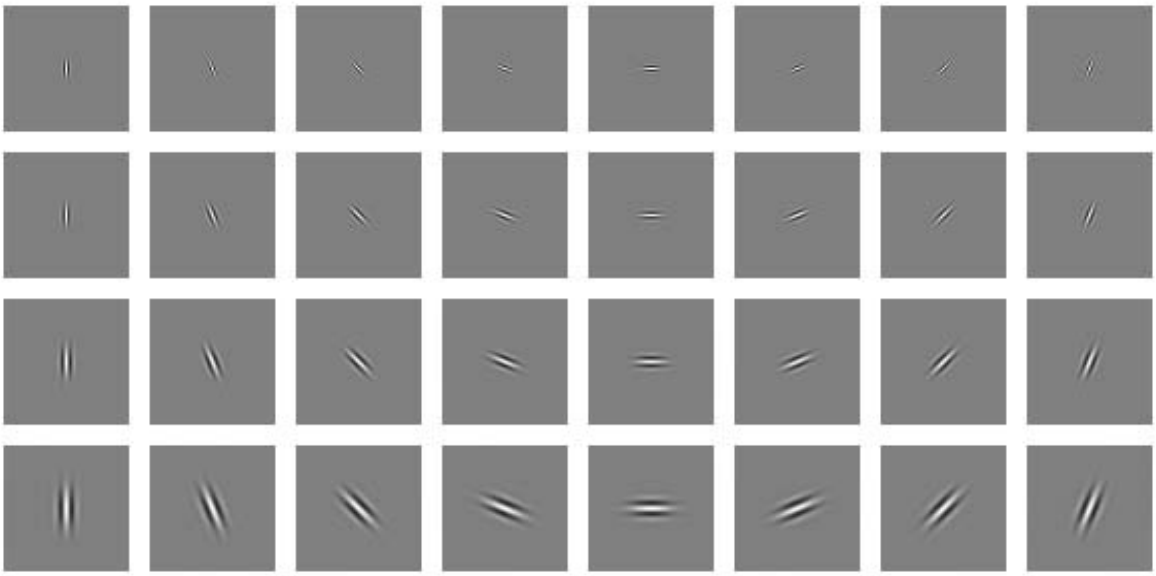


Figure 2.1: A set of Gabor filters with different envelop sizes σ and different orientations θ.

SIFT. They apply Principal Component Analysis to the normalized gradient patch in SIFT feature extraction part. This results in a more compact feature vector. They show that PCA based local descriptors are more distinctive and robust to image deformation than standard SIFT descriptor. Using PCA-SIFT provides increased accuracy and faster matching in an image retrieval task.

2.1.1.3 Gabor filter

Gabor filters have been used in computer vision for decades since Gabor introduced elementary Gabor function in [23]. In [24] Daugman found that simple cells in the visual cortex of mammalian brains can be modeled by 2D Gabor function. Thus image analysis based on Gabor functions is similar to perception in the human visual system. Bovik et al. [25] used Gabor filters for texture segmentation. And Okada et al. [26] use Gabor filter for face recognition.

$$\begin{aligned} G(x,y) &= exp(-\frac{X^2+\gamma^2Y^2}{2\sigma^2}) \times \cos(\frac{2\pi}{\lambda}X+\psi) \\ X &= x\cos\theta + y\sin\theta \\ Y &= -x\sin\theta + y\cos\theta \end{aligned} \tag{2.1}$$

A two dimensional Gabor Filter is defined in Equation 2.1, with orientation θ, size of Gaussian envelop σ, wavelength λ, aspect ratio γ and phase offset ψ.

A set of Gabor filters shown in Figure 2.1 with different frequencies and different orientations are usually used for feature extraction from image. In computer vision, Gabor filters are popularly used for representing texture from image. In the recent years there is still a lot of work on using Gabor function for feature extraction in object recognition, such like C2 feature and V1 feature.

C_2 Feature

C_2 feature is introduced by Serre et al. in [27]. This feature is extracted by a hierarchical system that designed based on the organization of visual cortex. The approach to calculate C_2 features is using a hierarchical system obtained by a bank of Gabor filters, and then apply a template matching and a maximum pooling operation. The process of calculating this feature can be described as follows.

Given a gray-level image, we calculate C_2 feature using a hierarchical system modeling cells in human brain. First two layers are corresponding to the simple cell (S_1) and complex cell (C_1) in visual cortex.

S_1 (Gabor filter) Layer: Apply a battery of Gabor filters to the images. The battery of Gabor filters consists of filters with 16 scales and 4 orientations, arranging in 8 bands. Each band is consisted by filters from 2 neighboring scales, with 4 orientations per scale. The outputs of these filters are 64 S_1 layers. For each band, there are 8 S_1 layers with 4 orientation and 2 scales.

C_1 (local invariance) Layer: For each band, the maximum over scales and position of S_1 layers is taken by sub-sampling over a grid. Since there are 8 S_1 maps with 4 orientations and 2 scales for each band, these 8 maps are sub-sampled using a grid cell of 8×8 and the maximum value of each cell is taken. Therefore we have 8 subsampled maps with 2 scales and 4 orientations per band. The maximum is taken again in the correpsonding cell-grid of 2 scales, for each orientation. Then we have 4 C_1 layers for each orientation per band.

S_2 (intermediate feature) Layer: We extract K prototyped patches $P_{i=1\ldots K}$ of various sizes and all four orientations from C_1 maps that randomly selected from positive training images. Notice each patch contains 4 C_1 layers with different orientation. S_2 layers are computed by comparing C_1 layers of input images to K training patches using distance function: $Y = exp(-\gamma||X - P_i||^2)$, in which X is the patches sampled from all possible position on C_1. Thus we get K$\times$8 S_2 layers from all prototyped patches and all bands.

C_2 (global invariance) response: For a prototype patch P we have 8 S_2 layers from all bands, and a maximum value over all positions and bands is taken as a C_2 reponse. Since we have K$\times$8 S_2 layers, by taking max only over the 8 S_2 layers corresponding to the same training patch, we get K C_2 reponses.

In the training step, we compute C_1 layers for each training samples, and extracted K patches randomly from C_1 layers of positive training samples. Then we calculate all C_2 features for all training samples. These C_2 features are passed to a classifier for the final classification.

They show that C_2 feature framework exhibits excellent recognition performance and outperforms several state-of-the-art systems on MIT-CBCL datasets and Caltech 101 dataset[28]. In [29] they compare C_1 and C_2 standard feature model (SMF) in object detection and scene recognition on StreetScene Dataset. Also they use C_1 and C_2 SMF in object recognition of texture objects and pixel-wise object segmentation. Their experiments show that C_1 feature is more suitable for shape-based objects which have clear segmentation, and C2 feature is more suitable for texture-based objects and unsegmentated objects.

Jhuang et al. in [30] extends this model to action recognition in video stream. They extend the model by adding two more layers: time invariant S_3 and C_3 unit. Since C_2 layer in [29] is limited to recognition in an image, adding S_3 and C_3 enables the model to recognize action in a video sequence.

V1 feature

Pinto et al. introduced a simple V1 feature in [31]. They show that V1-like model outperforms state-of-the-art object recognition systems on a standard, ostensibly natural image recognition test. Their method of constructing V1 feature is simple. First they normalized images so that each image has zero mean and unit variance and a similar scale with size H$\times$W . Then they perform local division normalization on for each 3$\times$3 neighborhood. They apply a bank of N Gabor Filter of different orientations and frequencies to the normalized image, resulting in a H$\times$W$\times$N matrix. This matrix is normalized again with local divisive normalization. PCA is performed on feature vectors to reduce the dimension of features. The authors show that V1-like model outperforms state-of-the-art method on Caltech 101, but fails to deal with more complicated object recognition scenes, like rotations and viewpoint change.

2.1.1.4 Summary of features

We summarized the features we have been discussed this section in Table 2.1.

Table 2.1: Comparsion between computer vision features.

Feature	Type	Keypoint Detection	Feature Representation	Invariance	Biologically-plausible
HOG	Local	Evenly Grid	Histogram of gradient	none	no
SIFT	Local	DoG Filter	Histogram of gradient	scale, rotation, noise small deformation	no
SURF	Local	Hessian Blob Detector	Harr wavelet responses	scale, rotation	no
PCA-SIFT	Local	DoG Filter	PCA projection of image patch	scale	no
C2	Global	N/A	Maximum from template matching responses	scale	yes
V1	Local	N/A	Normalized reponses from Gabor filters outputs	scale	yes

2.1.2 Bag-of-feature model

Bag-of-feature model origin from research in texture recognition and nature language processing. In texture recognition, texton is used to describe basic elements that repeatly form texture. A universal texton dictionary is built and one texture pattern can be described by a histogram of texton. In the research of natural language processing, bag-of-word model is proposed in [32] to represent an orderless document by frequencies of words from a dictionary. In [33] Csurka et al. introduce bag-of-word model into computer vision field. Features are extracted from image instead of word in the natural language processing field.

The main structure of bag-of-feature of model is: (1) Extract features/image patches from image. (2) Learn ‘‘visual vocabulary’’ from features/image patches used for training. (3) Assign each feature/image patch to the closest visual vocabulary in the dictionary. (4) Each image is represented by the frequency of ‘‘visual word’’ in a histogram. (5) Histograms are used as feature vectors to train classifiers and make classification decisions.

In the feature extraction part, evenly sampled gird is used in [34, 35]. Some works used interested point detectors to select feature [34, 36]. Harris affine detector is used in [33]. SIFT feature is the most common used feature to describe image patches [33, 34].

To learn ‘‘visual vocabulary’’, an unsupervised clustering process is applied to find the center of K vocabulary centers. Each cluster center selected by clustering algorithm, such as K-means clustering and will be used as ‘‘codevector’’. Each feature from one image is assigned to the closest ‘‘codevector’’ and a histogram of frequencies on K bins is obtained. In [33] vocabulary histogram is passed into two types of classifiers: Naive Bayes and SVM. The main advantages of bag-of-feature model are that it is simple, computationally efficient

and intrinsically invariant.

In [34] Fei-Fei and Perona use a Bayesian hierarchical model to learn the categories. The training images are unlabeled. They use bag-of-feature model to generate a histogram of vocabulary for each image, then use the Bayesian model to learn ‘‘category’’ for each trained image.

Grauman and Darrell proposed Pyramid Match Kernel in [37]. Pyramid Match Kernel is a new kernel function that is based on implicit correspondences which maps unordered feature sets to multi-resolution histogram and computes a weighted histogram intersection in this space. The design of this kernel function is to approximate the similarity of the best partial matching between the feature sets. A Harris detector is used to find interested point and several local descriptors (SIFT, JET, patches) are used to compute the features in an image. This kernel is tested with SVM on Caltech 101 for object recognition performance and yields 43% recognition accuracy. Since this method seeks best correspondence between partial feature sets of image, it can handle unsegmented, cluttered data well.

The main disadvantage of bag-of-feature model is that spatial information of features is not considered in the model. It is hard to capture shape information or to segment the object from background. In order to make up this, Lazebni et al. proposed spatial pyramid matching in [38] base on pyramid matching kernel in [37]. By partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-regions, this method builds up a histogram from different scale level of image which contain spatial information. This method exceeds the state-of-the-art on Caltech 101 dataset and achieves high accuracy on a large database of fifteen natural scene categories.

2.1.3 Part-based model

The disadvantage of bag-of-feature model is that spatial information between features is not included in the model. Part-based model treats object as set of N parts and use geometry information between different parts. Part-based model is first introduced by Fischler and Elschlager in 1973 [39]. They proposed ‘‘part and structure model’’ and apply the model on face recognition.

Fergus et al. use a generative model for object recognition in [40]. In this model, objects are modeled as flexible constellations of parts. A probabilistic representation is used for all aspects of the object: shape, appearance, occlusion and relative scales which are modeled by probability density functions. For each image, N interesting features are found with locations X, scales S, and appearances A. A generative model is learned for each category that with P parts and parameter θ. Saliency regions are found both in scale and location, which provides information on location X and scale S. Each saliency region is cropped and rescaled to the same size. PCA is performed on each patch and only the vector within the top principal components is used to describe appearance A for a patch. In the learning part, expectation-maximization (EM) algorithm is used to learn the parameters in the model. The parameters are learned using maximize likelihood estimation. This model is tested on 6 diverse object categories with less than 10% error rate. However the framework is

heavily dependent on whether feature detectors pick up useful features on the object.

Fei-Fei et al. in [28] extend Fergus' model in [40] by using a Bayesian model instead of maximum likelihood model to incrementally learn the parameters in Fergus' model. Bayesian model outperforms maximum likelihood methods on small dataset, and incrementally learning is significantly fast in learning process. This model is tested on Caltech 101 dataset, and this method is significantly better than Fergus' model when training images are less than 10. The maximum likelihood method matches Bayesian model method when training images is around 15.

Felzenszwalb and Girshick present their work based on mixtures of multi-scale deformable part models in [41]. In their work, they combined Dalal and Triggs's work on HOG filter[17] with part-based model. They extend the model using a star-structure part-based model defined by a root filter (global filter) that covers the whole object plus a set of part filters which have higher resolution than the root filter. Each filter is HOG filter from Dalal and Triggs's work and will outputs a score based on convolution. For each image the algorithm computes HOG features, and then applies both root filter and part filters to compute score. Part filters are applied to feature map that have twice spatial resolutions comparing to the root filter. The final score is defined as combination of score by each filter at their relative location and minus a deformation cost that depends of each part with respect to the root. Therefore the score combines both data term and spatial prior as well as bias term. To detect objects in an image, for each root location an overall score is computed to find the best placement of all parts. In addition they use a mixture model for each object category. Each object category is presented by M mixture components, in which each component has several parts. The computation of score for mixture model can be expressed by a dot product between a vector of model parameters and a vector of single component score.

Since the training data only have bounding box of object for each image, they use latent variable formulation of MI-SVM (Latent SVM) to learn model structure, filters and deformation cost. Latent variables are added into the cost function. The problem of training is reduced to Latent SVM training process and latent variable is learned during the process.

Later Felzenszwalb and Girshick extend their model to a cascade classifier. In [42] they describe a general method to build a cascade classifier from part-based deformable models. They provide a simple algorithm based on partial hypothesis pruning which speeds up object detection without sacrificing detection accuracy. They introduced the notion of probably approximately admissible (PAA) thresholds. Such thresholds provide theoretical guarantees on the performance of the cascade method and can be computed from a small sample of positive examples. Then they extend the pruning methods to a general class of model which could have mixture deformable model instead of tree-structured pictorial structures.

2.1.4 Boosting-based object detection

Viola and Jones proposed in [1, 43] a machine learning approach based on Adaboost for rapid detection of visual objects which performs detection in real time with high accuracy. They introduced a new image representation method as ''Integral image'' that allowing rapid feature computing. The integral image is similar to the summed area table used in graphics. In the integral image, each point presents sum of the pixel (feature) above and to the left of the pixel. So for any given rectangular, the sum of pixel (feature) inside the rectangular is computed from four array references in constant time. In this detection model, Harr-like features are used and about 180,000 features are extracted from each sub-regions using integral image.

In addition, Viola proposed a variant of AdaBoost which both selects a small set of features and trains the classifier. In the original form of AdaBoost, the algorithm is used to boost the classification performance using a set of weak classifiers with linear combination of weights. However, there could be more than 180,000 features associated with each rectangular window. To train a classifier using Adaboost on such large set of weak classifiers is too computational expensive. In Viola's variant Ababoost, only one weak classifier that has best classification performance is added into the final classifier in each iteration. Therefore the dimensionality of feature space is largely reduced.

Also a method for combining successively more complex classifiers in a cascade structure is introduced to improve the speed of detector. This cascade structure is used to rapidly determine if there is an object in the image and quickly reject false sub-windows. Only sub-windows that are not rejected by initial classifiers are passed into the next stage. The structure of this detection process is similar to a degenerate decision tree.

This method is applied to face detection and reaches real time detection speed which is roughly 15 times faster than previous method. The face detector is tested on MIT+CMU test set and outcomes the state-of-art performance of face detector at the time. However, a large training image set is used for training the detector, which contains 4916 training faces and 10,000 non-face sub-windows. This method might not be as generalized as to common object categories, which typically only has hundred of images each category in a common object detection dataset. Viola and Jones extend this detection framework to pedestrian detection in [44]. Motion features and appearance features are integrated and added into the model to enable detecting pedestrian from video sequences.

2.2 Survey on human model

The patterns of fixations and saccades made during search also comprise a rich data set for the purpose of modeling, with each of these fixations, and their serial order, being a behavior requiring explanation. Although there are several excellent models of eye movements during search and scene viewing [45--50], one of the most comprehensive is the Target Acquisition Model, or TAM[14]. In the following we will survey several bottom-up attention models in 3.1, and several task-specified search models in 3.2, and we

will describe TAM in 3.3.

2.2.1 Saliency model

How to predict where human observers are looking at in the real world scene? In visual search theory, bottom-up information is based on image itself and low-level features without prior knowledge. Top-down information is task-specified information prior to visual search. There is lots of research done for the bottom-up saliency model of visual attention. Wolfe in [51] demonstrated that basic visual features can capture and guide our attention in simple displays. Koch and Ullman [52] proposed that a set of basic feature can be extracted and combined to form a saliency map. Itti and Koch [53, 54] extends this idea into a computational model which generates a saliency map based on search image, and simulates a sequence of human eye movements based on a saliency map. For a given image, low-level features (color, intensity, orientations) are extracted from multiple scales. Then all maps are combined into one saliency map by weighted linear combination. The saccade sequences are generated by winner-take-all strategy and after each saccade the fixation point will be inhibited to generate next saccade. This model provides a way to generate saccade sequences that fitting human observers' behavioral pattern in a free-viewing behavioral task.

Bruce and Tsotsos define bottom-up saliency based on maximum information sampling in [55] Information is computed as Shannon's self-information as $-logp(F)$ where F is a vector of visual features at a point of image. The statistic distribution is estimated from a neighborhood of the point, or from the entire image. ICA filters are used to extract feature from the image.

In [56] Zhang et al. presents a model using bottom-up information and define saliency using natural statistic. The model is based on Bayesian framework on statistic. Saliency is defined by the probability of a target at every location given the visual features observed. When there is no information about the target, saliency can be derived from low-level feature information. The statistic on feature information is learned from a training image dataset of natural scenes. The feature they used in the model is generated in two methods: DoG filter and Linear ICA filters. Both filters are biological plausible in some sense. They show the linear ICA filters are better than DoG since it doesn't assume independence between features. And their saliency model works better than Itti's model [53] on more complicated scenes. Comparing to Bruce and Tsotsos's work, the statistic on feature information is learned from a training image dataset of natural scenes which is more robust, instead of from one image in [55].

2.2.2 Categorical-search model

In the study of saliency model, only bottom-up information which is inspired by low-level features from image is considered. Bottom-up information is not relevant to any search target. How to predict human's fixation when a target is specified? The recent study on behavioral theory shows that, not only preview

of search target, but also categorical information helps search guidance in human behavioral experiment. Visual search studies usually assume the target is very precise to guide search as a picture of image, but in the real world visual search is usually defined categorically. Schmidt and Zelinsky show that search guidance is proportional to the categorical specificity of a target cue in [57]. In their experiment, 5 target preview conditions are provided for a visual search: a picture of the target, an abstract textual description of the target, a precise textual description, an abstract + color textual description, or a precise +color textual description. The experiment shows that not only a picture of target, but also categorical information will help search guidance.

Zhang and Zelinsky present a computational model of human eye movement in an object class detection task in [58]. The model uses an object detection method (SIFT feature trained using Adaboost) with a biological plausible model of human eye movement. The eye model will produce a sequence of eye fixations when searching for a target. The performance of model is comparing to human behavioral experiment of the same search task in several measures, such as detection accuracy, number of fixations, cumulative probability of fixating the target, and length of scan path. The model shows a close match to human behavior for the same task.

Eckstein et al [59] study how predictive cues help observers search for objects in the real scene. Their behavioral experiment shows that accuracy of first saccade during search for objects was significant higher when the target appears at an expected location than an unexpected location. They present two computational models: Differential-Weighting Model (Bayesian Priors) uses a Bayesian framework to weight the evidence of target presence at each location by the prior probability of target co-occurring with a highly visible cue. Limited-Attentional-Resources Model deploys attentional resources at likely target locations, which are cued by other highly visible objects. The model generates a saccade to the location with highest likelihood ratio. They compared the distance of first saccades between human observers and two models, under target present/absent at expected or unexpected location. Their Differential-Weighting Model shows a closer fit to human behavioral.

Torralba and Oliva [48] show how contextual information helps to guide eye movement. They model contexture information based on a Bayesian framework. The model has two pathways: one computes local features as bottom-up saliency, while the other one computes global feature that would provide scene recognition and contextual information of the search target. By combining bottom-up saliency with the top-down contextual information the model would predict the image regions that likely to be fixated by human observers when performing a search task. The features they extracted in both local and global pathway are generated from a bank of steerable pyramid filters.

Ehinger in [60] extends the work in [48] to a person object detection task. In their experiment they recorded 14 observers' eye movement when searching for person in 900 outdoor scenes. The regions of eye movement are highly consistent between observers even when the target is absent from the scene. They use eye movements to evaluate computational models of search guidance from saliency, target feature and scene context. The combination of these guidance cues predicted 94% of human observers' fixation regions.

Similarly, Kanan and Zhang present SUN framework in [49, 56] which also make use of Bayesian model. In [49] they added top-down knowledge information based on appearance to the model. They trained classifiers for people, mug, painting using LabelME dataset[61]. Top-down knowledge is obtained from probabilistic SVMs and added into model. The model is also able to predict where people are looking at when looking for a target class. However this model is not able to predict anything about human eye movement sequences.

2.2.3 The Target Acquisition Model

Target Acquisition Model is introduced by Zelinsky in [14]. TAM is relatively unique in that it has a simulated foveated retina, without which eye movements would be unnecessary, and generates a sequence of saccades that align this fovea with a target. It does this by correlating the responses of Gabor-like filters to the target and search images to obtain a target map indicating visual evidence for the target in the search scene. Importantly, the quality of this evidence reflects acuity limitations resulting from a peripheral viewing of the target. Following an iterative pruning process, an eye movement is programmed to the location on the target map offering the most evidence for the target. If the fixated pattern is determined not to be the target, this false target is inhibited and the cycle begins anew with the selection of a new target candidate for inspection. Processing stops when the target match exceeds a high detection threshold, which often occurs only after the target has been fixated by the simulated fovea. TAM therefore makes eye movements during search for the same reason that people do, to offset retinal acuity limitations that prevent very confident search decisions.

TAM has been applied to several search tasks and can explain a range of overt search behaviors. These include the number of fixations needed to locate targets in realistic scenes, the expression of center-of-gravity fixations in the context of simple scenes, and many benchmark search patterns, such as set size effects, search asymmetry effects, eccentricity effects, and target/non-target similarity effects. But TAM also has many weaknesses, with perhaps the greatest being the need for precise knowledge of the target's appearance in the search scene. This luxury is rarely afforded in real world search, where variability in targets is commonplace. The extreme example of this is categorical search, when the target can be any member of a target class, and the actual search target on a given trial was never before seen. Previous work has shown that search is guided even to such categorically-defined targets [57, 62], but to date no model of eye movements during search exists to describe this behavior. In the following chapters, we introduce the first major modification to TAM, one that lifts this fundamental limitation and enables TAM to search for categorical targets. We do this by using techniques borrowed from computer vision to create the target map, keeping the model's other dynamics the same. We evaluate this change by comparing TAM's eye movements to those of human observers, with the goal being to emulate the human ability to guide search to categorical targets, and to capture the preferential direction of gaze to non-target objects that are visually similar to a target class.

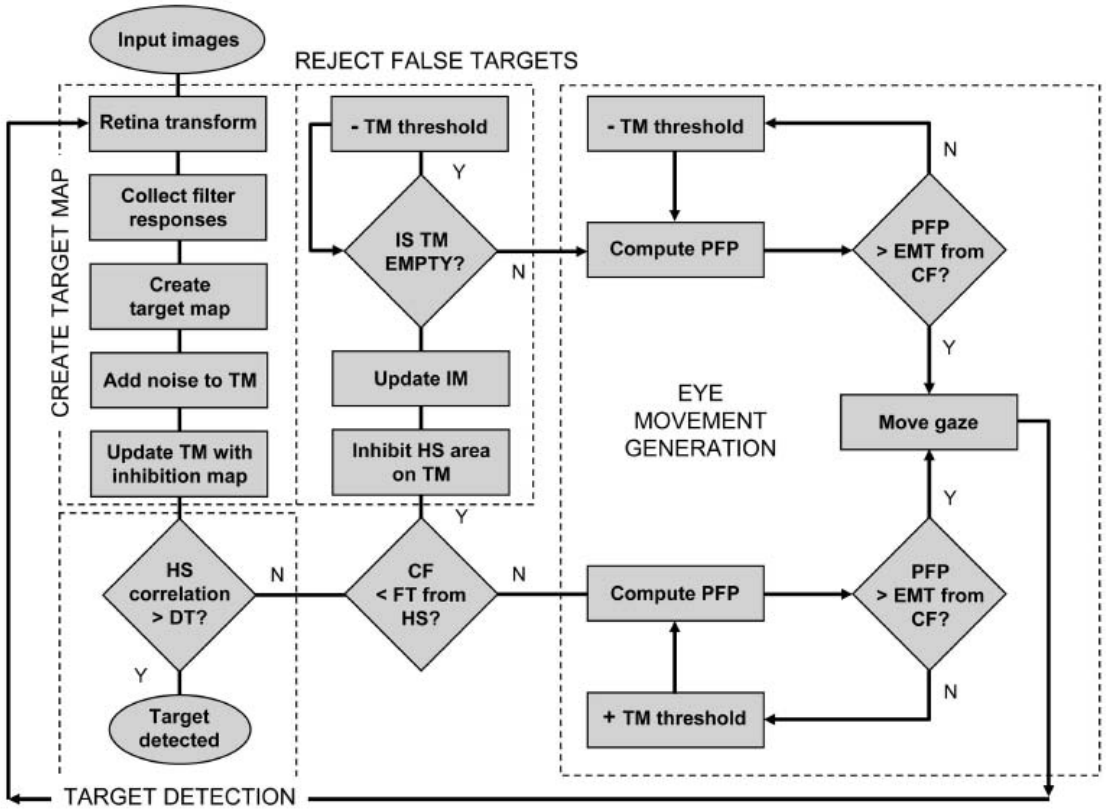


Figure 2.2: Flow processing of TAM [14]. Dashed boxes indicate four key conceptual stages: TM creation, target detection, false-target rejection, and eye-movement generation. Rectangular symbols indicate computational processes, diamond symbols indicate decision processes, and oval symbols indicate processing input and termination. TM: target map; HS: hotspot; DT: detection threshold; IM: inhibition map; CF: current fixation; FT: fixation threshold; PFP: proposed fixation point; EMT: eye-movement threshold; Y: yes; N: no.

2.2.4 Summary of human model

We summarized the current methods on computational human models in Table 2.2.

Table 2.2: Comparing between several computational human models. These models are comparing in the objective of behavioral task: search for nothing in the image, search for categorical object, or search for one specific object with image preview. Also they are comparing in the output of computational model: fixation points only, fixations sequences, or fixation regions.

Methods	Behavorial Task	Model Prediction Results
Itti and Koch[53]	Saliency Points	Saccade sequences
Eckstein[59]	Categorial Search	First saccade
Torralba and Oliva[48]	Categorial Search	Fixation regions
Zhang and Zelinsky[58]	Categorial Search	Saccade sequences
Zelinsky[14]	Object Search	Saccade sequences
Zhang[56]	Saliency Points	Fixation points
Kanan[49]	Categorial Search	Fixation regions
Ehinger[60]	Categorial Search	Fixation regions

Chapter 3

Behavioral Experiment

Figure 3.1a illustrates a core problem facing humans as they search. Suppose the task is to determine whether a teddy bear target is present in this object array. Assuming that gaze is located at the center (blue dot), note that all of the objects are slightly blurred due to retinal acuity limitations. This blurring necessarily reduces one's confidence that a member of the teddy bear class is present. Fortunately, humans can offset the impact of retinal blur and boost the confidence of their search decisions by making eye movements to objects. Upon fixating the bear (Figure 3.1b) this object would no longer be blurred, allowing its confident recognition. Much of the efficiency of human search can be attributed to the fact that visual similarity relationships to the target are used to guide these eye movements. This is clearest when a target actually appears in the search display, but this holds even when the search objects are all non-targets. Presumably, the bow tie in Figure 3.1c was fixated first due to its similarity to a butterfly, which was the target on this trial. Our behavioral experiment was conducted to explore the effects of these similarity relationships on eye movements in a categorical search task.

3.1 Similarity Ranking

Two target classes were explored in this study, teddy bears and butterflies. A pilot study using completely random objects as search distractors yielded only weak evidence for categorical guidance to targets. This was to be expected as most randomly selected real-world objects are minimally similar to either teddy bears or butterflies, making target classification impossible due to the absence of discriminative features. We therefore used the results from a web experiment to ensure that some of the search displays (depending on condition) had at least one distractor that was perceptually similar to the target category. Visual similarity relationships between random objects (from a broad range of categories) and the two target classes were obtained from this research [16].

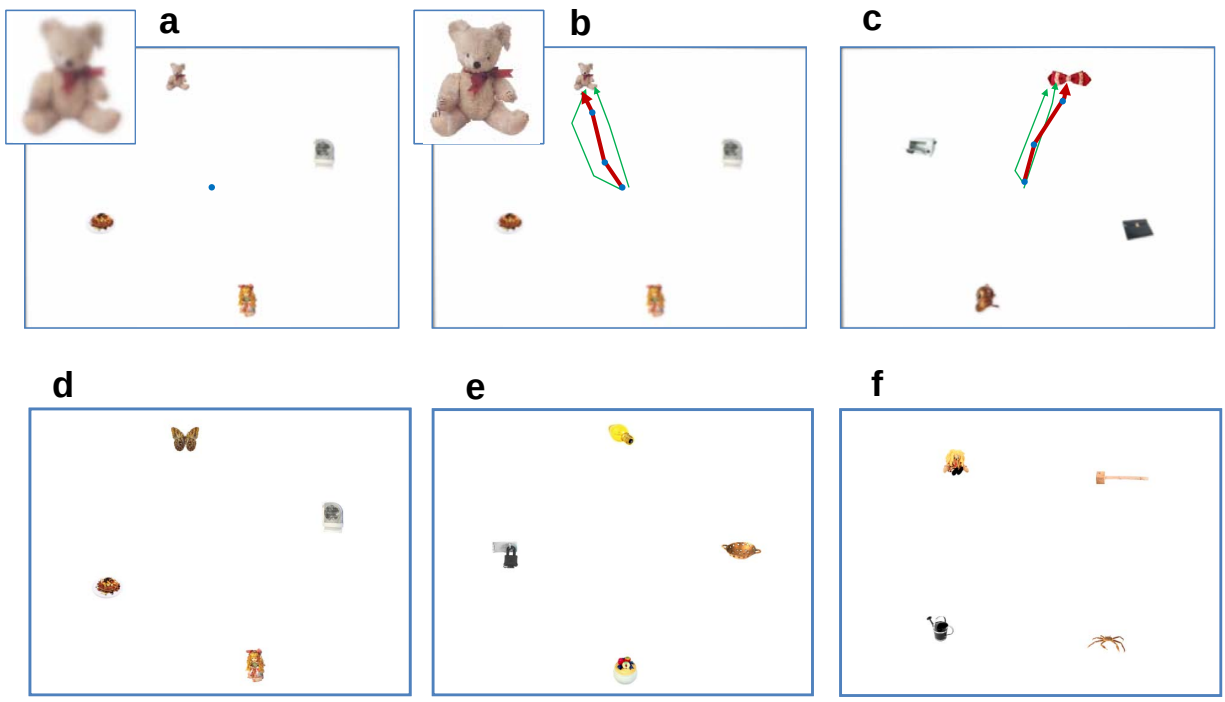


Figure 3.1: Example of four type of search displays. (a) Example of a target present search display as viewed initially from a central fixation position. Note the blurred target, shown enlarged in the offset. (b) The target is no longer blurred after fixation. Thin green lines show eye movements from two observers, the heavy red line shows eye movements from the model. (c) Example of a first object fixation in a target absent search display(butterfly search task, HIGH-MED-LOW similarity condition). (d) Example of target present trial (butterfly search task, corresponding to (a)). (e) Example of RANDOM condition. (f) Example of HIGH-MED similarity condition

In that study, each of the 142 participants were shown groups of five random objects (500 out of a total set of 2000 objects for each participants) and asked to rank order them by their visual similarity to either bears or butterflies. All objects were selected from the Hemera object database. Based on these 71,000 similarity estimates, we divided the objects into the following groups: bear-similar objects, bear-dissimilar objects, butterfly-similar objects, butterfly-dissimilar objects, and random objects that were rated either inconsistently by participants or as having intermediate visual similarity to both target classes. In the following discussion we will refer to these objects as *high*, *medium*, and *low* similarity objects, with the understanding that these similarity estimates are specific to either the bear or butterfly target categories. Figure 3.2 shows examples of high and low similarity objects for both target classes, and also examples of bear and butterfly targets.

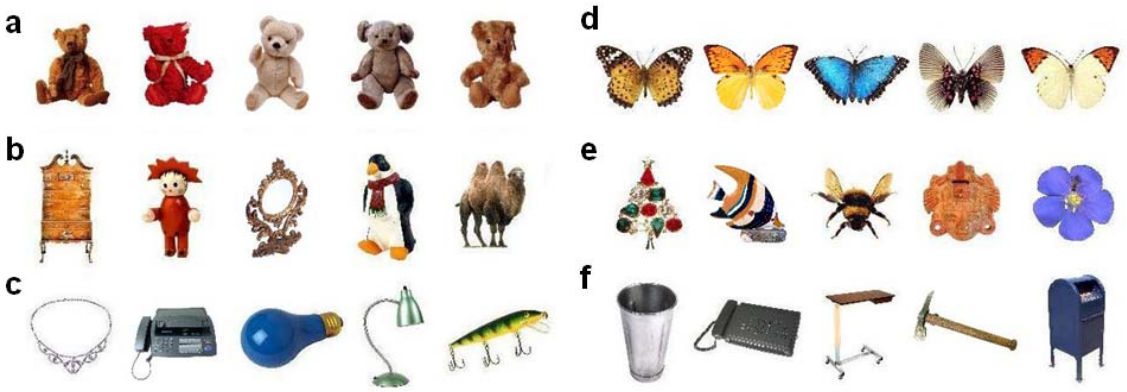


Figure 3.2: Examples of objects by group. (a) bear targets. (b) high-similarity to bears. (c) low-similarity to bears. (d) butterfly targets. (e) high-similarity to butterflies. (f) low-similarity to butterflies.

3.2 Construct search displays and collect fixation data

From these groups of similarity-rated objects we constructed four types of visual search displays (Table 3.1). In the *target-present* (TP) condition there was either a teddy bear or a butterfly target presented with three medium distractors that were ranked as having intermediate visual similarity to both target classes (Figure 3.1(a)(b)(d)). There were also three target-absent conditions. In the *high-medium* (TA-HM) condition there was one high-similarity distractor and three medium-similarity distractors (Figure 3.1(f)). In the *high-medium-low* (TA-HML) condition there was one low-similarity distractor, one high-similarity distractor, and two medium-similarity distractors (Figure 3.1(c)). Finally, in the *random* condition all four distractors were medium in visual similarity to the target (Figure 3.1(e)). The four objects in each search display were arranged on an imaginary circle (8.9 degree radius) around a point corresponding to starting gaze position (Figure 3.1).

There were 16 participants, half of whom searched for a bear and the other half a butterfly. The target class was designated via instruction; no specific target preview was shown prior to each search display. Participants

Table 3.1: The types of objects and search displays used in the behavioral experiment.

Display type	# trials	Type of objects			
TP	64	Target	Medium	Medium	Medium
TA-HML	44	Bear-similar	Butterfly-similar	Bear-dissimilar	Butterfly-dissimilar
TA-HM	40	Bear-similar	Butterfly-similar	Medium	Medium
Random	44	Medium	Medium	Medium	Medium

in the bear and butterfly search tasks viewed the identical search displays, except for the substitution of each bear or butterfly with an object from the other target category in target present trials. None of the target or non-target objects repeated throughout the experiment. Each of the 192 trials began with the observer fixating the center of the display and pressing a button. A search display then appeared and the task was to press one of two other buttons indicating the presence or absence of a target as quickly as possible while maintaining accuracy. Eye position was sampled at 500 Hz using an Eyelink II eye tracker.

The above-described similarity conditions served two goals. First, we wanted some of the displays (TA-HML and TA-HM) to depict both bear-similar and butterfly-similar objects, thereby giving subjects a choice as to which object they prefer to look at first or fixate the longest when searching for a particular category of target. To the extent that search is guided to categorical targets, we expected subjects looking for teddy bears to first fixate the bear-similar distractors and subjects looking for butterflies to first fixate the butterfly-similar distractors. Second, we wanted to vary the strength of this guidance signal. By including both target-similar and target-dissimilar objects in TA-HML displays, this condition offered the greatest potential for categorical guidance. A weaker guidance signal was expected in the TA-HM displays due to the replacement of the target-dissimilar objects with medium-similarity distractors, and little or no guidance was expected in the random displays where all of the distractors were random objects having weak or inconsistent similarity relationships to teddy bears or butterflies.

Importantly, search displays were crafted so as to have these similarity relationships apply to both target classes. As an example, on a TA-HM trial one object was rated as similar to a bear and the other three as medium, but among the medium similarity objects one of these would be rated as similar to a butterfly while the bear-similar object may be rated as having medium butterfly similarity. The same logic applied to the TA-HML condition. Except for the identity of the target object on target-present trials, subjects participating in the bear and butterfly search tasks therefore viewed the same search displays--the same distractor objects appearing in the same display locations. The fact that the identical target absent displays were used for the two search tasks is critical to the decoding goals of this study, as differences in bear and butterfly classification rates could not be attributed to differences in the composition of the search displays.

3.3 Behavioral Results

3.3.1 Search Accuracy

Subjects were both accurate and efficient in their categorical search for teddy bears and butterflies. Table 3.2 shows accuracy and reaction time averaged across subjects and grouped by target category and trial type. Bear and butterfly targets were correctly detected on 96% and 98% of the trials respectively, and false positive rates averaged less than 3%. Manual response times (RTs) averaged about 800 msec for correct trials, and was significantly shorter in the target present condition compared to the target absent conditions for both bears and butterflies, $t(7) \geq 3.68$, $p \leq .008$. RTs were slowest in the TA-HM conditions, faster in the TA-HML conditions, and fastest in the Random condition, although none of these target-absent comparisons were reliable, $p \geq .08$. This was true for both bears and butterflies, and reflects the fact that multiple objects are usually inspected before concluding that a target does not appear in a search display.

Table 3.2: Accuracy and reaction time of subjects searching for categorically-defined teddy bears and butterflies, grouped by display condition.

		Bear search				Butterfly search			
		TA-HML	TA-HM	Random	TP	TA-HML	TA-HM	Random	TP
Accuracy (%)	Mean	98.0	97.8	99.4	95.7	98.6	96.2	98.9	98.4
	SEM	0.90	0.74	0.37	1.24	0.74	0.96	0.86	0.73
RTs (msec)	Mean	806	857	786	675	921	938	872	672
	SEM	61.2	64.8	44.2	34.7	102.9	105.1	85.4	38.3

3.3.2 Fixation preferences

Two fixation preferences are explored in this section: the object that was first fixated during each search trial (first-fixated object, FFO) and the object that was fixated the longest during each search trial (longest-fixated object, LFO). FFO and LFO are not always the same object in each trials. Figure 3.3 shows example of human search path on target-absent displays that FFO and LFO are different.

Table 3.3 summarizes fixation behavior for the target and target-similar distractors. Trials in which there were errors (2.3%) or no fixated objects (3.2%) were excluded. Data are shown for both the FFO and LFO selection measures. Both measures showed a pronounced preference for the target object in target present trials---when a target was present in the display it was highly likely to be both fixated first (79.4% bear, 71.4% butterfly) and fixated the longest (98.9% bear, 99.6% butterfly) compared to the 25% rate expected by chance. More interesting are the target absent data, where a very similar preference was found for distractors rated as being visually similar to the target. Target similar distractors in the TA-HM and TA-HML conditions

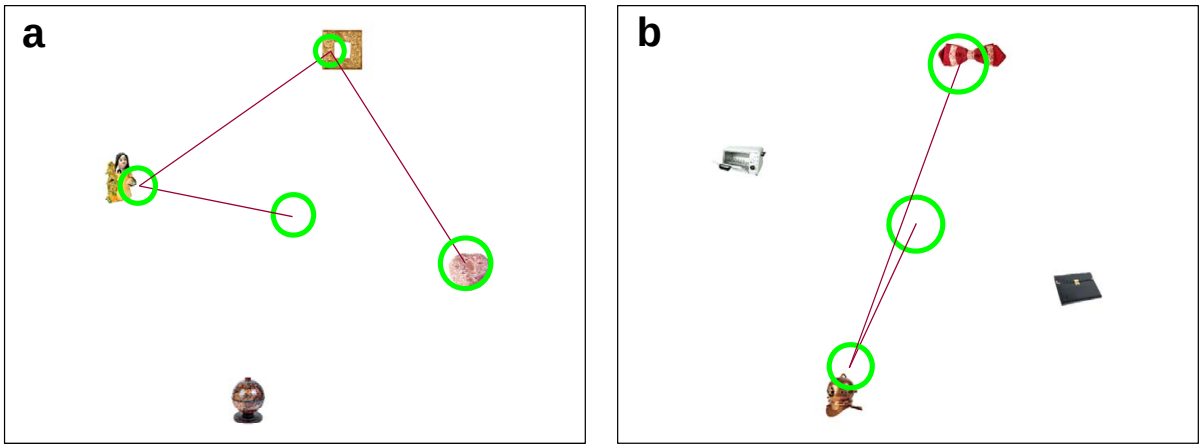


Figure 3.3: Representative target-absent displays showing superimposed scanpaths illustrating typical eye movement behavior. (a) A trial in which a bear-similar object was fixated first but not fixated longest. (b) A trial in which a butterfly-similar object was fixated longest but not fixated first.

were preferentially fixated well above chance for both the FFO and LFO measures and both search tasks, all $p < .001$. We also found significantly smaller preferences in the TA-HM conditions relative to the TA-HML conditions ($t(7) \geq 3.85$, $p \leq .006$), except in the case of the butterfly search data using the FFO measure, p = 48. This suggests that the presence of a target-dissimilar distractor in the TA-HML displays resulted in better search guidance to the target-similar object (FFO) and easier rejection of this object as a distractor after its fixation (LFO), probably due to the target-dissimilar object competing less for attention than medium objects.

Despite evidence for fixation preferences in both eye movement measures, comparison of the two clearly suggests that these preferences are better reflected in the LFO measure. Subjects were far more consistent in looking longer at targets or target-similar distractors than in looking first to these objects. On average, the target or target-similar object was 26% more likely to be predicted by the FFO measure compared to FFO. The effect of adding a target-dissimilar object to the display was also more consistent using the LFO measure. Averaging over bear and butterfly searches, the boost in preference found for the TA-HML displays appeared in 94% of the subjects using the LFO measure, but only in 67% of the subjects using the FFO measure. This suggests that more information about the target category is available from the LFO measure; to the extent that behavioral decoding of a search target is possible, we therefore expect it to be strongest using the longest fixated distractor.

To determine whether target-similar objects attracted attention we analyzed the type of object that was first fixated during search on correct trials. Figure 3.4 (blue bars) shows the percentages of these immediate

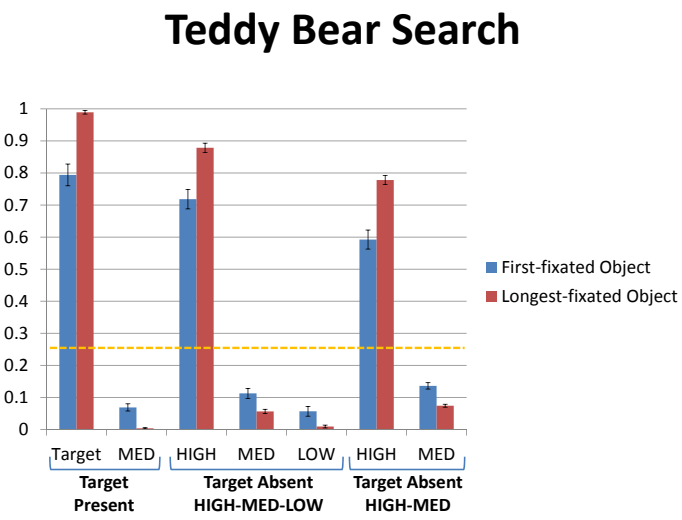
Table 3.3: Percentage of trials in which the target (TP) or target-similar distractor (TA) was preferentially fixated. First fixated object (FFO) and longest fixated object (LFO) refers to the methods of selecting objects used to decode the target category.

		Bears: display type			Butterflies: display type		
		TP	TA-HML	TA-HM	TP	TA-HML	TA-HM
First Fixated Object	Mean	79.4	71.8	59.2	71.4	42.5	44.4
	SEM	3.4	3.0	3.0	7.1	3.1	2.6
Longest Fixated Object	Mean	98.9	87.9	77.8	99.6	83.2	76.1
	SEM	0.6	1.4	1.4	0.3	2.1	2.0

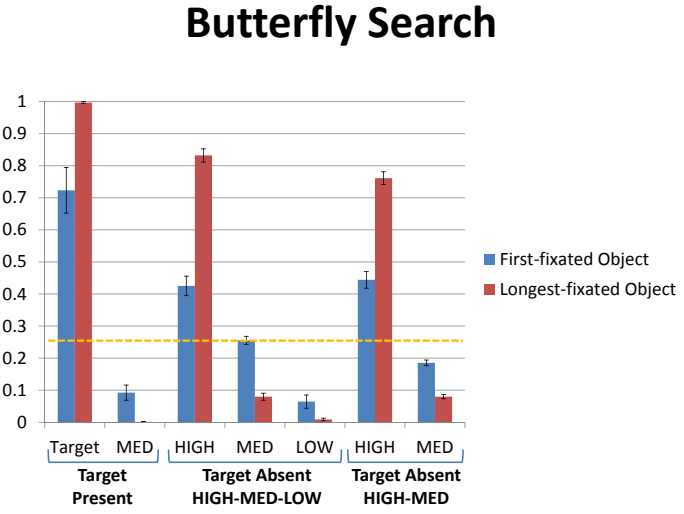
fixations by condition and object type.[1] When an actual target appeared in the display observers were far more likely to fixate it immediately compared to either chance or the medium similarity objects (all $p < .001$). This is strong evidence for search being guided to a member of a target class. More interesting are the target absent trials, where we find a similar pattern for the high-similarity objects. In the HIGH-MED condition the percentage of immediate fixations on target-like objects was again well above both chance and the immediate fixation rate on medium-similarity objects (all $p<.001$), but significantly less than immediate fixations on the actual targets ($p = .002$, bears; $p = .001$, butterflies). In the HIGH-MED-LOW condition we again found strong guidance to high-similarity objects, which in the case of the bear search task was not significantly different from guidance to the actual target ($p = .10$). This pronounced search guidance to target-similar objects was well above chance and stronger than guidance to either the medium or low-similarity objects (all $p \leq .004$). Moreover, guidance to low-similarity objects was weaker than guidance to objects rated medium in similarity to the target category ($p = .05$, bears; $p<.001$, butterflies). This suggests that search was not only guided to target-similar objects, it was also guided away from objects that did not look like targets.

In summary, we found that search was guided to non-target objects in proportion to their similarity to the target class; objects that were highly similar to the target attracted the most immediate fixations, and objects that were low in target similarity attracted the least, with initial looks to medium-similarity objects falling somewhere between.

[1] Note that values were adjusted to correct for the multiple instances of medium similarity objects in a display; 25% therefore reflects chance object fixation across conditions, but this adjustment results in object fixations within a condition not summing to 100%.



(a)



(b)

Figure 3.4: Percentages of first fixated objects (adjusted for chance) grouped by object type (targets, high-similarity, medium-similarity, or low-similarity to targets) and search condition (target present, HIGH-MED-LOW target absent, or HIGH-MED target absent). Error bars on the behavioral data indicate a 95% confidence interval, and the dashed lines indicate chance.

Chapter 4

Computational Experiment

The behavioral patterns describe in Chapter 3 constitute a rigorous test of a model of search. Such a model should not only be able to recognize both bear and butterfly targets at human levels, but to do so based on initially blurred views of these objects. It should also capture the graded effects of target/non-target similarity observed in the behavioral data; actual targets should be fixated first most frequently (target present trials), following by high-similarity objects, medium-similarity objects, and finally low-similarity objects (target absent trials).

In this chapter, we design several computational experiments. First, we present an eye model using C2 and color features with a probabilistic SVM to simulate human's fixation. Second, we train and evaluate several computer vision models and compare their outputs with human's during object detection task. Third, we propose a classification method to predict which target a human observer is looking for during visual search based on fixation patterns.

In Section 4.1 we introduced the features used in our computational methods. In Section 4.2 we introduced an eyemodel based on target acquisition model (TAM) to simulate eye movement during object detection. In Section 4.3 we evaluate performance of models using different features comparing to human's behaviors. In Section 4.4 we propose our behavior decoding methods to predict a human observer's search target.

4.1 Features

C2: C2 features, introduced by Serre et al [27], reflect the initial feed-forward visual processing known to be performed by simple and complex cells in primary visual cortex. In the basic four-layer model, the responses of simple cells, approximated by a bank of Gabor filters applied to an image, are pooled by complex cells (C1) using a local maximum operation, allowing limited invariance to changes in position and scale. Prototype

patches are sampled from the C1 responses in training images. The max response in a window for each C1 prototype forms the C2 feature for the window. In our implementation we used a bank of Gabor filters with 16 scales and 8 orientations, and extracted random C1 patches from positive training samples for use as prototypes.

SIFT and SIFT+SPM Features: SIFT features introduce by Lowe [63] represent the structure of gradients in a local patch of image with 16 spatial distributed histograms of oriented edge energy carefully scaled and noramlized. We follow the procedure laid out by Lazebnik *et al.* [38]. Specifically we use a vocabulary of 200 visual words, and either a single histogram over the bounding box or a two layer spatial pyramid in our implementation.[1]

V1 Features: V1 feature is introduced by Pinto [31] which is very similar to C1 feature in C2 model. First, each image is normalized so that each image has zero mean and unit variance and was downsampled to size H$\times$W. In our experiment we use H = 30 and W = 30. Then a local division normalization is performed on for each 3$\times$3 neighborhood. Then a bank of N Gabor Filter of different orientations and frequencies is applied to the normalized image, resulting in a H$\times$W$\times$N matrix. In our experiment, we use N = 96 filters from 16 different orientations and 6 spatial frequencies. This matrix is normalized again with local divisive normalization. In the end we have almost 90,000 dimension of vector for a single image. PCA is performed on feature vectors to reduce the dimension of features. We obtain a vector of 635 dimension after PCA for each image.

The standard C2, V1 and SIFT features do not represent color information, but color is known to be an important feature for guiding search [47, 64, 65]. We therefore implement a simple color histogram feature defined in the DKL color space [66]. This color space has been shown to closely approximate the sensitivity of short, medium, and long-wavelength cone receptors, and also captures the luminance and color opponent properties of double opponent cells. The color histogram feature used 10 evenly spaced bins for each channel (Luminance, Red-Green, Blue-Yellow), where each dimension was normalized to fit in [0, 1]. These color histograms are used in place of Gabor responses in the above described procedure to produce a ‘‘C2-like’’ feature for color.

COLOR: The procedure for computing color features is as follows. First, all images were converted from RGB space to DKL space using Equation 4.1 from [67]. Next, we built an image pyramid using 3 scales per image, and from each layer we sampled 24$\times$24 pixel image patches, with each patch separated by 12 pixels. A color histogram was computed for each sampled patch. We then randomly selected patches from 3-layer pyramids created for the positive training images and used these as prototypes. The max response to each prototype over a window is used as the color feature for that window. In our experiment we use 250 color patches.

[1] Note that we report results using a linear classifier in place of the histogram intersection kernel (HIK) used by [37, 38]. This is for consistency with experiments on other features, and while HIK does produce better classification accuracy for the SIFT+SPM feature, the overall pattern of results is similar.

$$\begin{pmatrix} DKL\_RG \\ DKL\_BY \\ DKL\_L \end{pmatrix} = \begin{pmatrix} -1.3546 & 1.6569 & -0.3023 \\ 0.5410 & 2.3910 & -2.9312 \\ 0.7425 & 3.3068 & 1.5957 \end{pmatrix} \times \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{4.1}$$

4.2 Simulation of eye movement during object detection

4.2.1 Methods

The present work builds on TAM [14], introduced earlier. Core subroutines in TAM pre-process an image to approximate the acuity limitations arising from a foveated retina at a given fixation location, compute a target map based on this pre-processed image, and select the next fixation by iteratively maximizing an objective function based on the current fixation's target map. In this paper we generalize the target map to categorical search, removing the previously limiting requirement that an exact image of the target object be made available. We refer the reader to [14] for details on the temporal dynamics of TAM, so as to focus here on the features and methods used to quantify the visual similarity relationships underlying categorical search.

Visual similarity relationships between objects and a target category are embodied in our model by the target map, a map of target probabilities computed for each point in the search image. We obtain these probabilities by mapping the output of a classifier trained to recognize the target category. The classifier is a linear kernel based SVM with probabilistic estimation on C2 features and color features computed in a sliding window around each point in the image.

In this experiment we used 500 C2 features and 250 color features concatenated into a 750 dimensional feature to train two linear SVM classifiers [68], one to separate bears from non-bears and the other to separate butterflies from non-butterflies. To generate the target maps for each search task, we passed a circular sliding window over each search image (the identical images shown to our behavioral observers) and found the target probability for each point. Target probabilities were based on the distance from the decision boundary in a linear SVM trained using the probability estimation method described in [69]. Note also that TAM requires that this be done for every fixation made during search, so as to accurately reflect the impact of retinal acuity limitations on the calculations of the target probabilities. Using Figure 4.1 as an example, panel (d) shows that the target map computed for the initial fixation position would use the blurred image shown in panel (a). However, panel (e) shows that the target map for the fourth (and final) fixation on that trial would be based on the image shown in panel (b), which depicts a foveated non-blurred bear among highly blurred non-targets. This differential blurring affects the target probabilities. Whereas the blurred bear from the first fixation did not yield a target probability exceeding a high recognition threshold, as shown in panel (d), the non-blurred bear from the fourth fixation did, as shown in panel (e). TAM searches for a target until a probability value on the target map exceeds a detection threshold (which might occur on any fixation), or until all of the objects have been inspected and rejected, at which point it concludes that the target is not present in the search image.

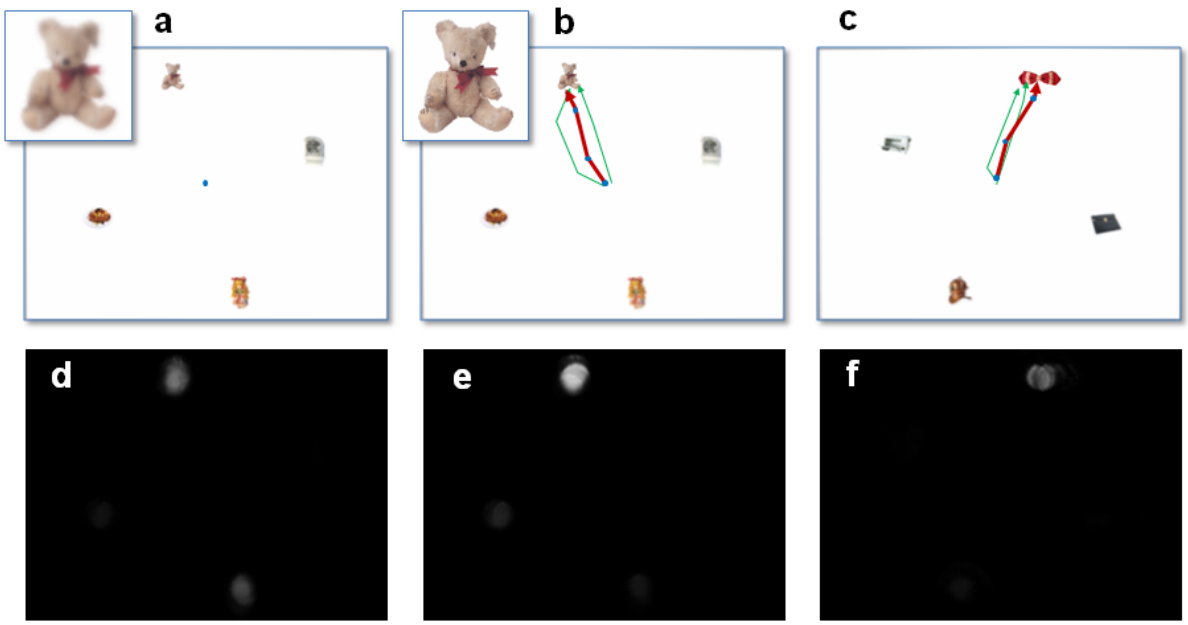


Figure 4.1: (a) Example of a target present search display as viewed initially from a central fixation position. Note the blurred target, shown enlarged in the offset. (b) The target is no longer blurred after fixation. Thin green lines show eye movements from two observers, the heavy red line shows eye movements from the model. (c) Example of a first object fixation in a target absent search display(butterfly search task, HIGH-MED-LOW similarity condition). (d-f) Target maps corresponding to fixations 1, 4, and 4 in panels a-c, respectively.

This detection threshold was set at .8 probability for both the bear and butterfly detection tasks, based on the selection of a threshold minimizing a weighted 0-1 loss function on our validation dataset.

4.2.2 Result Analysis

As in the case of human observers, this model is capable of making false negatives and positives (Table 4.1). These errors on target absent trials were $< 5\%$ for both the bear and butterfly search tasks. Like our observers, this model did not often mistake a non-target object for a member of the target class. However, errors on target present trials were higher than what we found in our behavioral data, suggesting that the model currently lacks a human ability to recognize objects as bears or butterflies even after they have been fixated (i.e., non-blurred). We attribute these unrealistically high miss rates to an insufficient number of positive training samples (90 bears and 90 butterflies); had we had more bears and butterflies to use for training (this number was limited by object consistency constraints imposed by the behavioral experiment), we would expect recognition failures to drop down to human levels.

Table 4.1: Error rates and response times (RT) by search task and condition for observers and our model

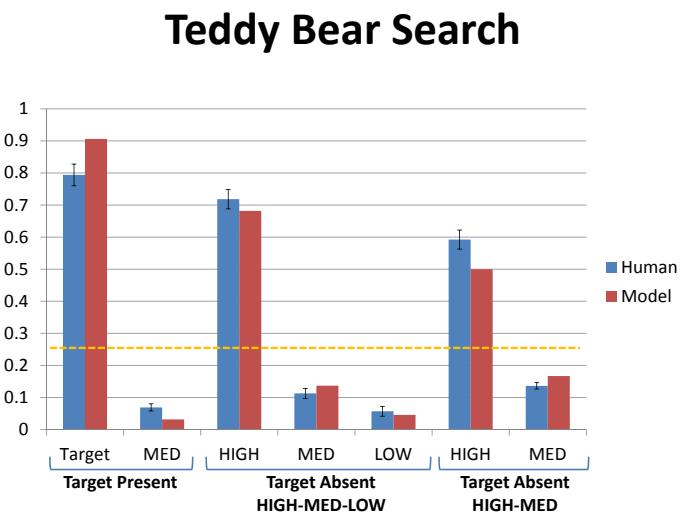
	Bear Seach				Butterfly Search			
	Target Present		Target Absent		Target Present		Target Absent	
	Errors(%)	RT(ms)	Errors(%)	RT(ms)	Errors(%)	RT(ms)	Errors(%)	RT(ms)
Human	4.7	684	2.1	818	2.0	684	2.5	929
Model	10.9	N/A	1.2	N/A	15.6	N/A	3.6	N/A

The primary aim of this study was to describe eye movements to categorically-defined targets, and the similarity relationships affecting this behavior. Towards this end, we found the first object fixated by the model on each of the 148 search displays, exactly as we did for our human observers. These data are shown by the red bars in Figure 4.2. Turning first to the bear search task, on target present trials the model captured the very strong guidance to categorical targets that we observed in human behavior. Indeed, it reflected this human tendency a bit too well, fixating the target immediately in each case. It is unclear, however, whether this difference in immediate fixation rates is meaningful. Although humans are clearly capable of fixating a bear target directly, one would not expect this to happen on every trial; on some trials their motivation would inevitably falter, resulting in imperfect guidance. Human guidance therefore has a ceiling, a level that aggregated behavior will not exceed, whereas a model has unflagging motivation and no such limitation. In the case of the target absent trials the model's behavior is less open to interpretation; in each condition it matched human behavior almost perfectly, well within the respective 95% confidence intervals surrounding the behavioral means. The model looked directly at high-similarity objects far more frequently than either medium-similarity or low-similarity objects, with this difference appearing in both the HIGH-MED and HIGH-MED-LOW conditions. Even more subtle behavioral patterns were captured, such as the greater immediate fixation rate on high-similarity objects in the HIGH-MED-LOW condition compared to high-similarity objects in the HIGH-MED condition, as well as the anemic difference in immediate fixations between the medium-similarity and low-similarity objects. Such agreement is impressive given that no parameters of the model were adjusted to fit the behavioral data; the classification boundaries used by the model naturally captured the similarity relationships used by humans to guide their eyes to categorical targets.

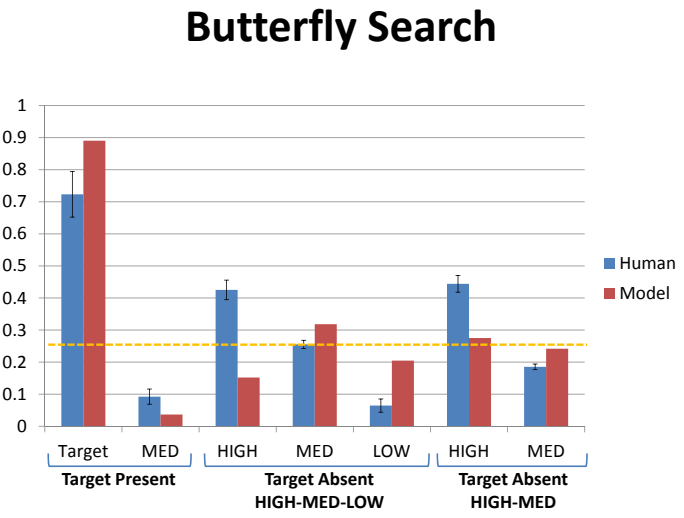
Table 4.2: Percentage of observers fixating (or not fixating) the high-similarity object first (HSO) given first fixation (or not fixation) by the model. Note that values do not sum to 100% due to cases in which observers failed to fixate any object.

	Model fixates HSO		Model does not fixate HSO	
	Observers Fixating HSO	Observers Not Fixating HSO	Observers Fixating HSO	Observers Not Fixating HSO
Bear	73%	22%	42%	52%
Butterfly	55%	42%	38%	59%

The model's behavior in the butterfly search task was nearly as impressive. Again turning first to the target present trials, human guidance to butterfly targets was slightly lower than guidance to bears, suggesting that blurred butterflies may be harder for humans to discriminate from non-targets than blurred bears. The model



(a)



(b)

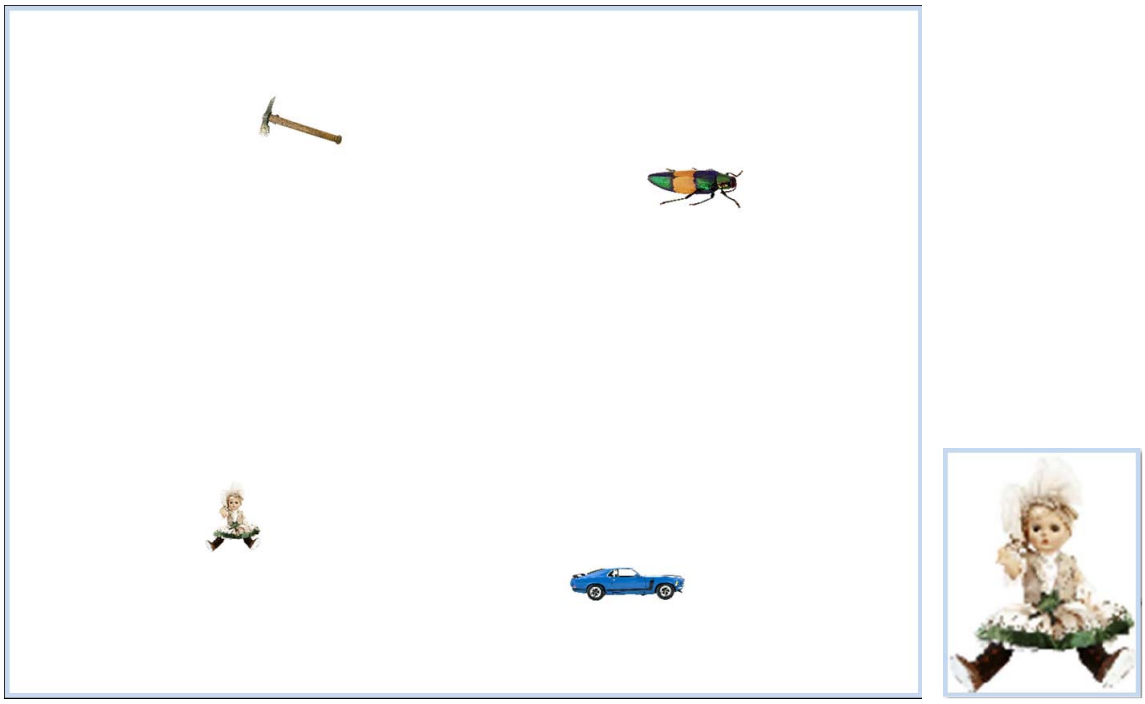
Figure 4.2: Percentages of first fixated objects (adjusted for chance) grouped by object type (targets, high-similarity, medium-similarity, or low-similarity to targets) and search condition (target present, HIGH-MED-LOW target absent, or HIGH-MED target absent). Error bars on the behavioral data indicate a 95% confidence interval, and the dashed lines indicate chance.

also reflected this pattern, now making only 86% of its immediate fixations on the target. This imperfect immediate fixation rate, although numerically higher than the human fixation rate, was still within the 95% confidence interval of the behavioral mean. The model also captured the main pattern in the target absent behavioral data; high-similarity objects were fixated first more often than medium-similarity objects in HIGH-MED condition. However the model fails on the HIGH-MED-LOW condition, suggesting features in our model might not be the best presentative features for butterfly cateogory search. These differences between search tasks are consistent with the speculation that search is guided less efficiently to butterflies than teddy bears, possibly due to greater feature variability in the butterfly category.

The above analyses told us that the model captured quite well the average effects of target similarity on categorical search guidance, but stronger still would be a demonstration that these effects exist on a trial-by-trial basis. Perhaps the trials on which the model's eye was strongly guided to high-similarity objects were not the same trials on which this happened for humans. To address this possibility, we segregated the target absent data depending on whether the model made an immediate fixation on the high-similarity object or not, then for each trial in these two groups we found the percentage of observers who also looked first to the high-similarity object. A complementary analysis was conducted for trials in which the model failed to look directly at the target similar object. Both analyses are shown in Table 4.2. On trials in which the model made an immediate fixation on a target-similar object, 73% of our bear observers and 55% of our butterfly observers also fixated that object first. However, when the model failed to first fixate the high-similarity object, these percentages dropped to 42% and 38% for the bear and butterfly tasks, respectively. Although we can only speculate as to the reason why guidance was stronger for some objects than others, the implications of this pattern are clear; our observers and model tended to agree on the specific objects that were most and least effective in guiding categorical search. In Figure 4.3 and 4.4 we shows two typical cases that search of target category is guided to HIGH similarity object in both human and model experiment. In Figure 4.5 we show one search display that model doesn't fix the HSO as well as the majority of human subjects.

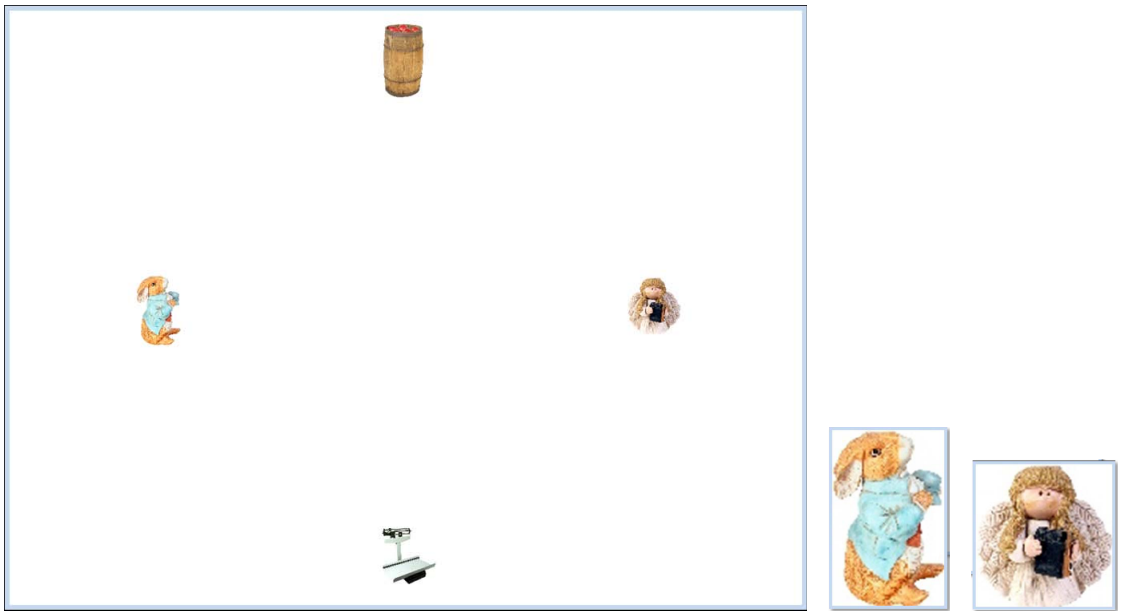
(a) (b)

Figure 4.3: HIGH-similarity object shows strong guidance in butterfly search display. During the search for a butterfly in 4.3(a), the model and 7/8 of the observers first fixated the butterfly-similar brooch object in Figure 4.3(b).



(a) (b)

Figure 4.4: HIGH-similarity object shows strong guidance in teddy bear search display. During the search for a teddy bear 4.4(a) , the model and 8/8 of the observers first fixated the bear-similar doll object in Figure 4.4(b)



(a) (b) (c)

Figure 4.5: Similar mistake make by human and model in a teddy bear search display. During the search for a teddy bear in Figure 4.5(a), the model and 6/8 of the observers did not fixate the object rated as bear-similar (the rabbit doll in Figure 4.5(b)). Instead, the model and 6/8 of the observers first fixated the object rated as butterfly-similar (the angel doll, in Figure 4.5(c)).

4.3 Using computer vision models to predict human confusion

Instead of studying all fixations simulated by a computational model during visual search task, we are more interested to discover the visual similarity relationship between the first fixated object with target in a visual search task. Also we are exploring the role of blurness during a visual search for object detection methods. We train and analyze how well computer vision models agree with the human behavior, focusing on whether similar objects are found to be confusing by people and by the computer vision models.

4.3.1 Computational Models

For the computational models we trained sliding window detectors using classifiers on top of various descriptors. We chose to use C2 features [27] as a representative for biologically inspired features that have performed very well in object recognition. We also evaluated using histograms of vector quantized SIFT descriptors computed in a uniform grid over bounding boxes, perhaps the most commonly used descriptors for object recognition. Following common practice, we included evaluations using spatial pyramid descriptors (SPM) [38] on top of the quantized SIFT descriptors. Color features were computed using a variant of the C2 feature approach, and optionally combined with each of the features. Linear classifiers were trained using

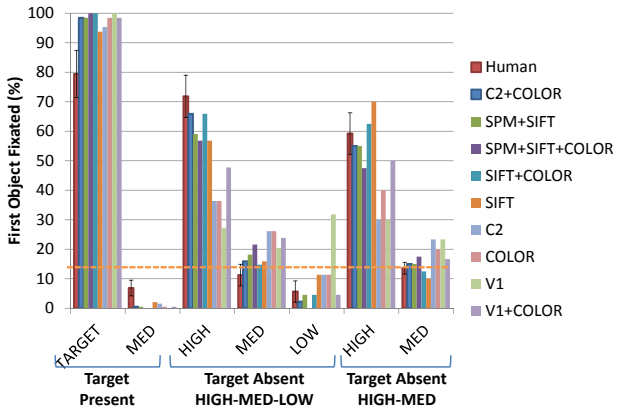
training data completely disjoint from the test data.

In our experiments, we are exploring how combinations of features affect performance of object detection and predict similarity to human behavior. We add a color feature into standard SIFT and C2 features; we use 200 SIFT features, 1000 SPM-SIFT features, 1000 C2 features, each concatenated with 250 color features that encode color information into our model. For each feature set, we trained two linear SVM classifiers [68], one to separate teddy bears from non-bears and the other to separate butterflies from non-butterflies. Each classifer is trained using 136 target images and 500 distractors. To detect an object, we apply a sliding window over each search display (identical to the displays shown to human observers). For each object in the scene, the detector returns the local maximum response around it. These responses are probabilities based on the distance from the decision boundary in a linear SVM trained using the probability estimation method described in [69]. Our computational model considers the object with the maximum response from the detector as the first-fixated object.

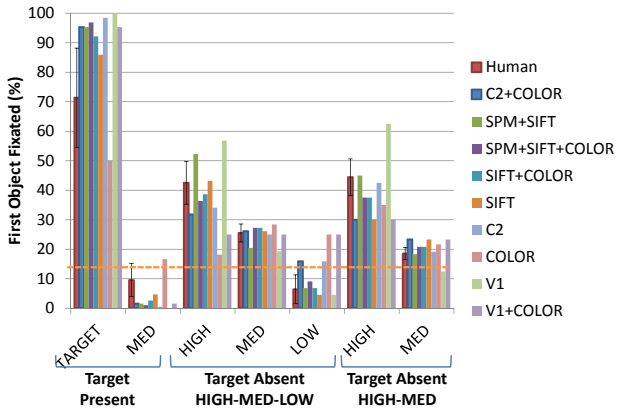
In order to compare the models' behavior with human behavior, we training and test under 3 different conditions: (1) Training using nonblurred images and testing on nonblurred displays (UB-UB) as in Figure 4.1(a). (2) Training using nonblurred images and testing on blurred displays (UB-B) as in Figure 4.1(b). (3) Training using blurred images and testing on blurred displays (B-B). Condition (1) is similar to object detection in computer vision. Condition (2) is simulation of the human detection task. In human vision, the resolution of images decreases with increasing distance from the fovea. These blurred versions of the objects aim to simulate the retinal effect in humans[70] when the eyes are fixated at the center of the image; they are generated by a gaussian pyramid under a fixed σ parameter (chosen to reflect the degree of blurring actually perceived by humans). So in our blurred version of the test displays, all objects have the same level of blur due to the same distance to the image center. We also train a blurred version classifier in Condition (3), using a training set that has same level of blur as objects in the blurred test displays.

4.3.2 Computational Results

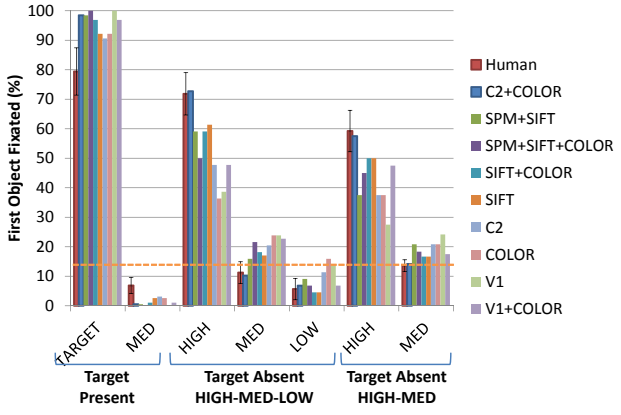
The primary aim of this study was to model eye movements to categorically-defined targets, and the similarity relationships affecting this behavior. Towards this end, we found the first object fixated by the model on each of the 148 search displays, exactly as we did for our human observers. We evalaute the results of the object detectors in two ways: Based on accuracy in detecting the target when present (and reporting it's absence otherwise) and based on agreement with human behavior both when the target is present and when it is absent. We conducted experiments under 18 experimental conditions: 2 object categories $\times$ 3 train/test sets $\times$ 3 target conditions. The categories are bears and butterflies, the training/testing sets are non-blurred/nonblurred(UB-UB) nonblurred/blurred(UB-B) and blurred/blurred (B-B) and the target conditions as described in Sec. 3.2 are Target Present (TP), (HIGH-MED-LOW) Target Absent and (HIGH-MED) Target Absent. For each of the 18 conditions we tried the following 9 descriptor/feature combinations: C2, SIFT, COLOR, SPM-SIFT, V1, C2+COLOR, SIFT+COLOR, SPM-SIFT + COLOR, V1+COLOR. We compute detection accuracy as an f-score and we report agreement between humans and the model for each



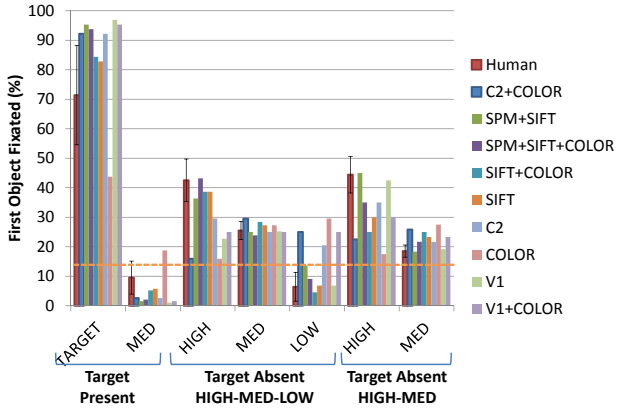
(a) Teddy Bear UB-UB



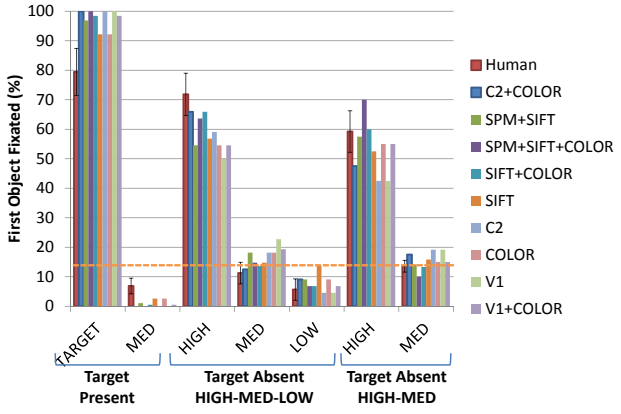
(b) Butterfly UB-UB



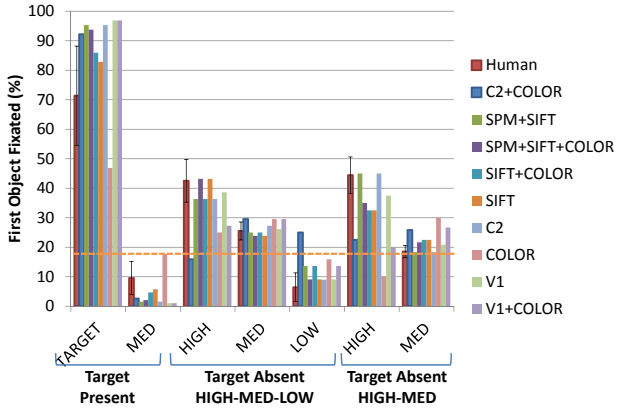
(c) Teddy Bear UB-B



(d) Butterfly UB-B

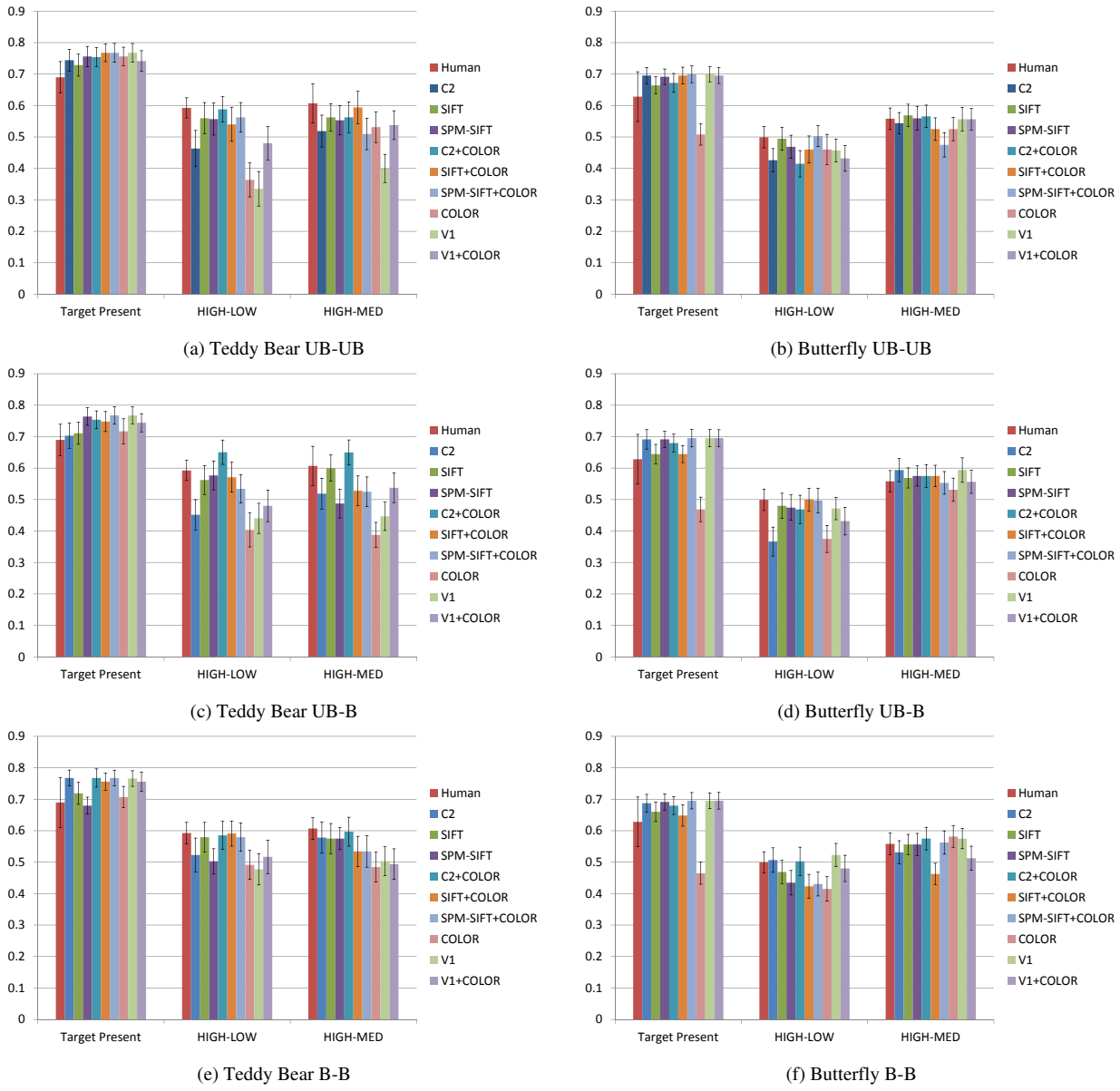


(e) Teddy Bear B-B



(f) Butterfly B-B

Figure 4.6: Percentages of first fixated objects (adjusted for chance) grouped by object type (targets, high-similarity, medium-similarity, or low-similarity to targets) and search condition (target present, HIGH-MED-LOW target absent, or HIGH-MED target absent) for humans and models under different training and test condition. Error bars on the behavioral data indicate a 95% confidence interval, and the dashed lines indicate chance (25%).



(a) Teddy Bear UB-UB

(b) Butterfly UB-UB

(c) Teddy Bear UB-B

(d) Butterfly UB-B

(e) Teddy Bear B-B

(f) Butterfly B-B

Figure 4.7: Agreement scores between first fixated objects predicted by models and human behavior grouped by condition (target present, HIGH-MED-LOW target absent, or HIGH-MED target absent). Human agreement score is computed by averaging scores for each subject compared to the rest of the subjects. Error bars on human data indicate deviation between subjects, and error bars on model data indicate standard errors.

Table 4.3: F-scores and agreement scores from each method under different training and testing condition for teddy bear search task. The bold numbers are the best scores for each particular evaluation score, for each of the 3 train/test conditions.

	Bear Seach				
	Method	F-score	Target Present Agreement Score	HIGH-MED-LOW Agreement Score	HIGH-MED Agreement Score
Human		0.9499	0.6897	0.5925	0.6071
UB-UB	C2	0.8976	0.7441	0.4631	0.5188
	SIFT	0.7722	0.7285	0.5597	0.5625
	SPM-SIFT	0.9291	0.7559	0.5568	0.5531
	C2+COLOR	0.9457	0.7539	**0.5880**	0.5625
	SIFT+COLOR	0.9173	**0.7676**	0.5398	**0.5938**
	SPM-SIFT+COLOR	0.9244	**0.7676**	0.5625	0.5094
	COLOR	0.8759	0.7559	0.3636	0.5313
	V1	**0.9677**	**0.7676**	0.3350	0.4000
	V1+COLOR	0.9077	0.7558	0.4233	0.4812
UB-B	C2	0.7619	0.7031	0.4517	0.5188
	SIFT	0.7606	0.7109	0.5653	0.600
	SPM-SIFT	**0.9291**	0.7637	0.5767	0.4875
	C2+COLOR	0.8189	0.7539	**0.6505**	**0.6500**
	SIFT+COLOR	0.8676	0.7480	0.5710	0.5281
	SPM-SIFT+COLOR	0.9091	**0.7676**	0.5341	0.5250
	COLOR	0.7101	0.7168	0.4034	0.3875
	V1	**0.9449**	**0.7676**	0.4403	0.4469
	V1+COLOR	0.8718	0.7441	0.4801	0.5375
B-B	C2	0.8448	**0.7676**	0.5227	0.5781
	SIFT	0.7656	0.7188	0.5795	0.5750
	SPM-SIFT	0.9104	0.7441	**0.5994**	**0.6250**
	C2+COLOR	0.8943	**0.7676**	0.5852	0.5969
	SIFT+COLOR	0.8780	0.7559	0.5909	0.5344
	SPM-SIFT+COLOR	0.9242	**0.7676**	0.5795	0.5343
	COLOR	0.7273	0.7070	0.4915	0.4844
	V1	**0.9764**	**0.7676**	0.4772	0.5031
	V1+COLOR	0.8444	0.7559	0.517	0.4938

of the target presence scenarios. Overall the blurred/blurred experiments perform in similar manner as the nonblurred/blurred ones.

Table 4.3 and Table 4.4 present the method with the best accuracy and agreement scores for each experimental condition.

From Figure 4.6 we know that several models do an excellent job in capturing the patterns of confusions shown by humans doing object detection; target-similar objects are fixated more than medium-similarity objects, and these are fixated more than target-dissimilar objects. From Figure 4 we know that this excellent fit also extends to agreement between humans and the model on a trial-by-trial basis; those trials in which humans tended to look initially at the target-similar object were the same trials in which the model tended to

Table 4.4: F-scores and agreement scores from each method under different training and testing condition for butterfly search task. The bold numbers are the best scores for each particular evaluation score, for each of the 3 train/test conditions.

	Butterfly Seach				
	Method	F-score	Target Present Agreement Score	HIGH-MED-LOW Agreement Score	HIGH-MED Agreement Score
Human		0.9655	0.6282	0.4992	0.5580
UB-UB	C2	0.8906	0.6953	0.4261	0.5438
	SIFT	0.6732	0.6641	0.4943	**0.5688**
	SPM-SIFT	0.9180	0.6914	0.4688	0.5594
	C2+COLOR	0.9008	0.6719	0.4148	0.5656
	SIFT+COLOR	0.7907	0.6953	0.4602	0.5250
	SPM-SIFT+COLOR	0.9280	**0.6992**	**0.5028**	0.4750
	COLOR	0.6250	0.5078	0.4375	0.5250
	V1	**0.9683**	**0.6992**	0.4574	0.5562
	V1+COLOR	0.9687	0.6953	0.4403	0.5375
UB-B	C2	0.8906	0.6914	0.3665	**0.5938**
	SIFT	0.7321	0.6445	0.4801	0.5688
	SPM-SIFT	0.9344	0.6914	0.4744	0.5750
	C2+COLOR	0.8421	0.6797	0.4688	0.5750
	SIFT+COLOR	0.7576	0.6445	**0.5000**	0.5750
	SPM-SIFT+COLOR	0.9194	**0.6953**	0.4972	0.5531
	COLOR	0.5039	0.4688	0.3750	0.5313
	V1	**0.9457**	**0.6953**	0.4716	**0.5938**
	V1+COLOR	0.8926	**0.6953**	0.4318	0.5563
B-B	C2	0.8361	0.6875	0.5057	0.5313
	SIFT	0.7523	0.6602	0.4688	0.5563
	SPM-SIFT	0.9365	0.6914	0.4347	0.5563
	C2+COLOR	0.8421	0.6797	0.5028	0.5750
	SIFT+COLOR	0.7193	0.6484	0.4233	0.4625
	SPM-SIFT+COLOR	**0.9516**	**0.6953**	0.4205	0.5625
	COLOR	0.5385	0.4648	0.4128	**0.5813**
	V1	0.9323	**0.6953**	**0.5227**	0.5750
	V1+COLOR	0.9016	0.6953	0.4801	0.5125

look initially at these objects.

We examine the agreement between the human observers themselves. These data are shown by the red (leftmost) bars in Figure 4.7. The other bars show the trial-by-trial agreement of the various computational models with the human observers. The agreement scores are computed by comparing one model's prediction to all 8 subjects and taking an average of accurracy from fitting subject's behaviors on trial by trial basis for each trial condition (Target present, target absent HIGH-MED, target absent HIGH-MED-LOW). Human agreement is computed by comparing each one subject behavioral with the other 7 subjects.

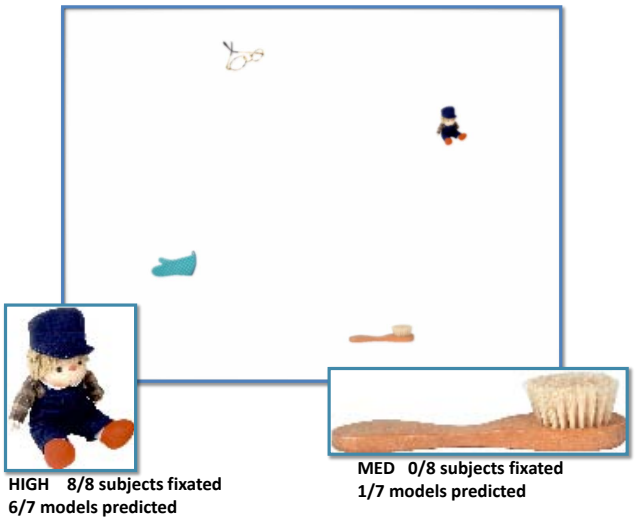
In Figure 4.7 we observe that our computational models for the most part agree with human observers when the target is present. What is remarkable is that two of them, C2+COLOR and SPM-SIFT+COLOR fit human

behavior very well in both scenarios when the target is absent. This suggests that we are able to simulate human behavior not only in terms of accuracy of prediction but also in terms of modeling the confusion of humans when the target is absent by predicting the object that they will fixate first in that case.

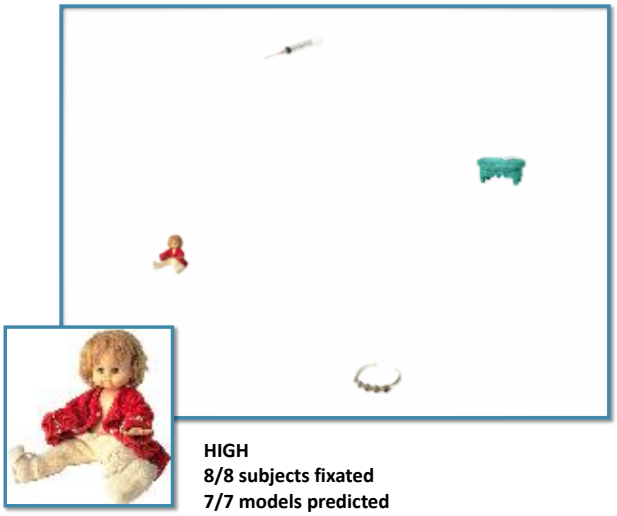
Results in Figures 4.6 and 4.7 suggest that computer vision models are able to capture human object detection at a very fine grain; describing not only detection performance, but also the patterns of confusions that determine human search efficiency.

4.3.3 Example of search displays

We show some examples of search displays in HIGH-MED-LOW and HIGH-MED condition during target search. For the teddy bear search task, Figure 4.8 shows examples where the models perform similarly to human subjects, and Figure 4.9 shows examples where the models fail to predict the first fixated object. For the butterfly search task, Figure 4.10 shows examples where the models perform similarly to human subjects, and Figure 4.11 shows examples where the models fail.

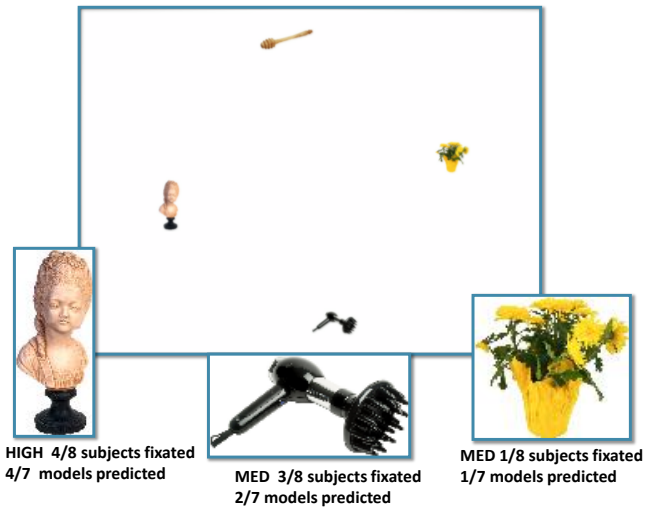


(a) HIGH-MED search display for teddy bear

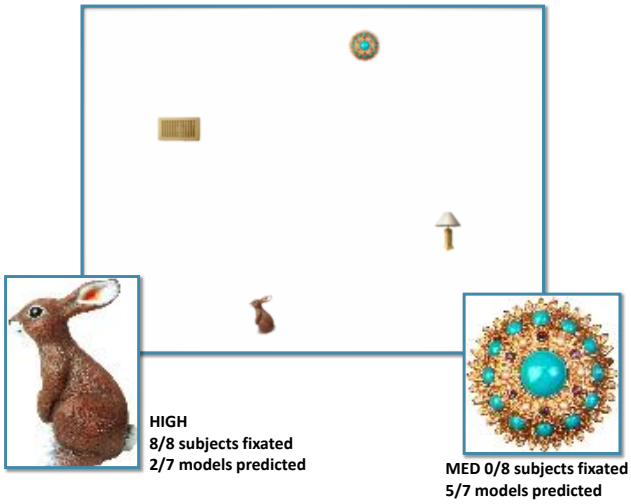


(b) HIGH-MED-LOW search display for teddy bear

Figure 4.8: Example of model predictions that match with human subjects on a teddy bear search task. Subfigure (a) shows an example of a HIGH-MED-LOW search display with two first fixated objects. When searching for a teddy bear as the target, 8 of 8 subjects fixated on the HIGH similarity object(doll, enlarged in the bottom left). Similarly, our models predict the doll as the first fixated object, except the COLOR model. The COLOR model predict the brush (MED) as the first fixated object, probably because the brush has similar color as a brown teddy bear. Subfigure (b) shows an example of HIGH-MED search display and the first fixated object. When searching for a teddy bear, all human subjects fixated on the doll(enlarged) first. All of our models predict the doll as the first fixated object.

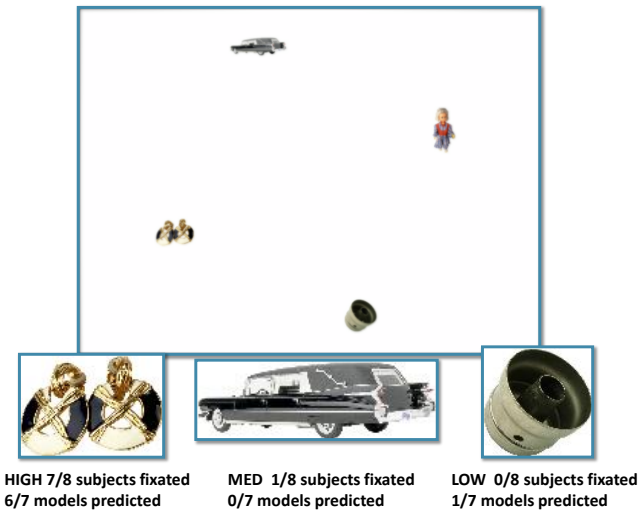


(a) HIGH-MED search display for teddy bear

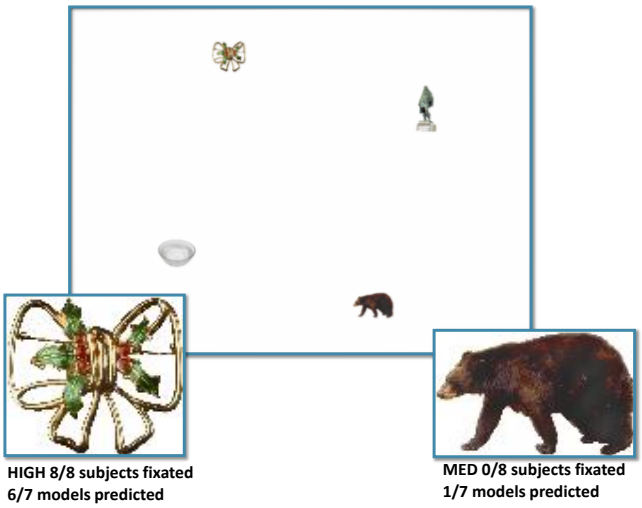


(b) HIGH-MED-LOW search display for teddy bear

Figure 4.9: Example of model predictions that do not match with human subjects on a teddy bear search task. Subfigure (a) shows an example of HIGH-MED-LOW search display with three first fixated objects. 4 out of 8 human subjects first fixated on the figurine (HIGH), 3 out of 8 subjects fixated on the hair tool(MED) and 1 out of 8 subjects fixated on the flower pot(MED). Our C2+COLOR, SIFT, SIFT+COLOR, SPM-SIFT+COLOR models predict the figurine as the first fixated object. Our SPM-SIFT and COLOR models predict the hair tool as the first fixated object, while the C2 model predicts the flower pot. Subfigure **??** shows an example of HIGH-MED search display and two first fixated objects. When searching for a teddy bear, all human subjects fixated on the rabbit(HIGH-similarity object) first. Only the C2+COLOR and SIFT model predict the rabbit as the first fixated object, while all other models predict the broach (MED-similarity object) as first fixated.

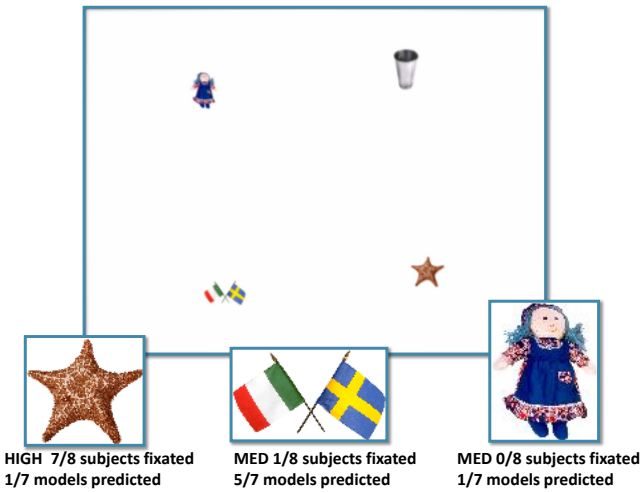


(a) HIGH-MED search display for butterfly

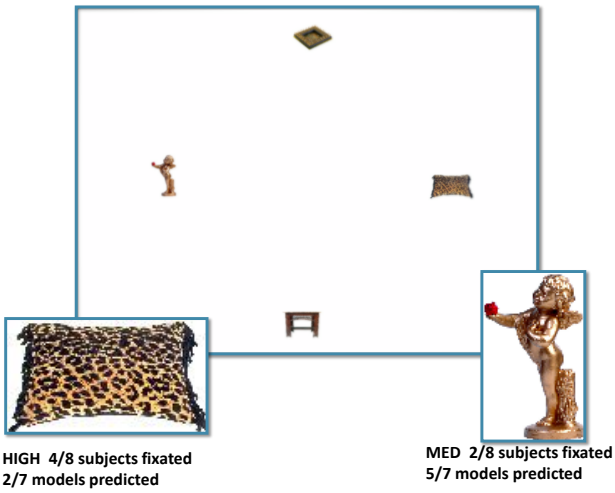


(b) HIGH-MED-LOW search display for butterfly

Figure 4.10: Example of model predictions that match with human subjects on a butterfly search task. Subfigure (a) shows an example of HIGH-MED-LOW search display with three first fixated objects. 7 of 8 human subjects first fixated on the golden earring(HIGH) and 1 of 8 human subjects fixated on the car (MED) first when searching for a butterfly. Our models except for the SIFT model predict the golden earring, while the SIFT model predicts the pot-like object(LOW). Subfigure (b) shows a example of HIGH-MED search display and first fixated object. All human subjects first fixated on broach (HIGH similarity object) during searching for a butterfly. Our models except the COLOR model predict the broach as the first fixated object. The COLOR model predicts the bear(MED).



(a) HIGH-MED search display for butterfly



(b) HIGH-MED-LOW search display for butterfly

Figure 4.11: Example of model predictions that do not match with human subjects on a butterfly search task. Subfigure (a) shows an example of a HIGH-MED-LOW search display with three first fixated objects. 7 of 8 human subjects first fixated on the starfish (HIGH) and 1 of 8 human subjects fixated on the flags (MED) first when searching for a butterfly. Only the C2+COLOR model predicts the starfish (HIGH), and the COLOR model predicts the doll(MED). All other models predict the flags (MED). Subfigure (b) shows an example of HIGH-MED search display and two first fixated object. 4 out of 8 human subjects first fixated on the pillow (HIGH) object when searching for a butterfly. Only the C2 and C2+COLOR predict the pillow(HIGH). Our SIFT, SPM-SIFT, SIFT+COLOR and SPM-SIFT+COLOR models predict the angel figurine (MED).

4.4 Decoding observers' search target from gaze fixations

Another computational experiment we design is to predict a human observer's target from fixation pattern. In this study we focus on first-fixated object (FFO) and longest-fixated object (LFO) during each search trial.

4.4.1 Computional Methods

The computational approach can be divided into training and classification stages. Two linear-kernal SVM classifiers were trained [68], one to separate bears from non-bears and the other to separate butterflies from non-butterflies. Consistent with standard object recognition methods, classifiers were trained to find a feature-based description for each object category using positive and negative training samples. Positive samples were 136 images of teddy bears and butterflies, negative samples were 500 images of random objects unrated for visual similarity to the target categories. The same negative samples were used for training the bear and butterfly classifiers. Negative and positive samples were selected from the same databases of images (teddy bears were adapted from [71], butterflies and distractors were from Hemera), although the training and testing sets obviously didn't overlap.

These classifiers used two types of features to obtain probabilistic estimates of targets. We combined the SIFT feature [63] with the spatial pyramid matching procedure described in [38] to create what we are calling a SIFT+SPM model. Specifically, we use a vocabulary of 200 visual words, and a two-layer spatial pyramid to obtain a 1000 dimensional ''SIFT+SPM'' feature. In order to make use of color, we concatenated the SIFT+SPM feature with a color histogram feature [67]. We then used these histograms as prototypes for classification, similar to the method used by HMAX [27].

Our classification method differed from standard methods in one key respect; rather than attempting to recognize positive from negative samples of a target class, our model estimated the visual similarity of non-target objects relative to the teddy bear or butterfly categories learned from training. Critically, the objects for which we obtained these similarity estimates, each based on the distance from the respective SVM classification boundary, were those selected by the human subjects as they searched. Two methods of selection were explored in this study: the object that was first fixated during each search trial (first-fixated object, FFO) and the object that was fixated the longest (longest-fixated object, LFO). For every trial from every subject, first-fixated and longest-fixated objects were input to each classifier and bear and butterfly similarity estimates were obtained. From these behavior-based estimates, we then attempted to decode whether the search target was a teddy bear or a butterfly.

4.4.2 Classification Performance

Is it possible to decode the target of a person's search from their looking behavior? To answer this question we attempted to classify each subject as searching for a teddy bear or a butterfly based on their object fixations in a given condition. We did this by grouping the data by subject, and determining the distance of the preferentially fixated object (using both the FFO nad LFO methods) from the SVM decision boundary for the bear and butterfly classifiers[2]. We then subtracted the butterfly distance from the bear distance to obtained a distance difference score, and calculated a median difference score for all the trials in a given display condition (a median was used because it is less sensitive to outliers than a mean, resulting in a more stable similarity estimate). The target of a subject's search was classified using this median difference score, with a negative value indicating a butterfly classification and a positive value indicating a bear classification.

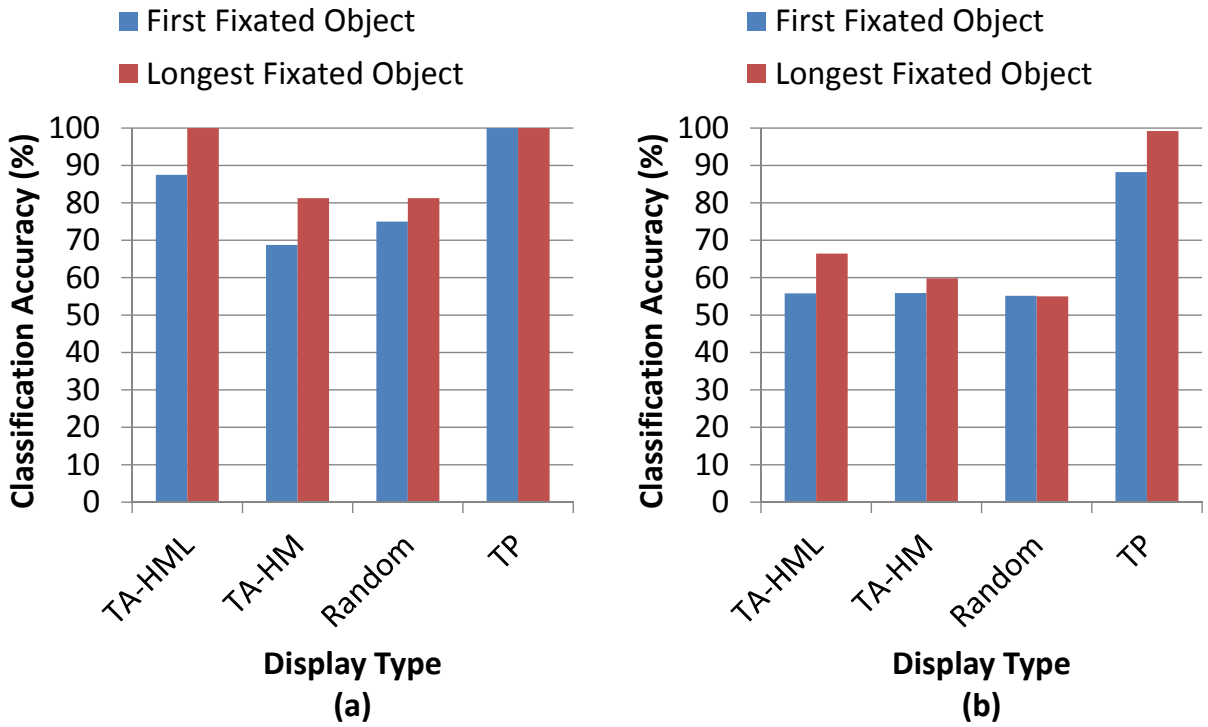


Figure 4.12: Accuracy of classification (a) by subject (% of subjects in which the target category was successfuly decoded), and (b) by trials (% of trials in which the target category was successfully decoded).

These by-subject classification rates are shown in Figure 4.12(a) for both the FFO and LFO selection methods.

[2]In pilot work we determined that the bear classifier was more discriminative than the butterfly classifier, producing a bias to classify a random object as a butterfly. To estimate and correct for this bias we evaluate both classifiers using a validation set consisting of 600 medium images. We found the median distance from each classifier to the objects in this set, and adjusted each classifier by these values to make them unbiased. All of the reported classification rates reflect this bias adjustment.

As expected, classification was trivial when the target actually appeared in the display, leading to perfect accuracy in the target-present conditions regardless of selection method. Less trivial are the target absent searches, where classifications were based solely on distractor fixations. Three patterns should be noted. First, classification accuracy was overall quite high, ranging from 69-100%, and was significantly better then the 50% level predicted by chance in each of the conditions (all $p < .001$). Indeed, using the LFO method we were able to decode the search target for each of our 16 subjects in the TA-HML condition---reading their minds by analyzing the distractor that they looked at the longest. This shows that subjects preferentially fixated the target-similar distractor in these displays, and the classification of these distractors was sufficient to decode the subject's target category. Second, the longest-fixated distractor yielded consistently better classification than the first-fixated distractor. This follows from the behavioral data. Subjects were more consistent in fixating the target-similar distractor longer (rather than first, Table 3.3), thereby making more of these highly informative objects available for classification. Third, good classification (>75%) was even achieved using random distractors having inconsistent similarity relationships. Information useful for classification therefore exists, not only in objects that were explicitly rated as being target similar, but also in the seemingly random objects that normal people choose to fixate as they search.

Table 4.5: Classification accuracy for individual trials, grouped by target and display condition.

Selection	Bear Search				Butterfly Search			
	TA-HML	TA-HM	Random	TP	TA-HML	TA-HM	Random	TP
FFO	64.00	63.88	59.73	92.78	53.62	47.88	50.59	83.67
LFO	71.38	67.89	57.05	100	61.45	51.79	52.94	98.41

The above-described by-subject classification results pooled information from fixated objects over all the trials from a given subject and condition---would it still be possible to decode the target category based on only a single fixated object from a single trial? To answer this question we discarded information about subject and display type, classifying each trial independently from all others. These results are shown in Figure 4.12b collapsed across target category, and in Table 4.5 broken down by bear and butterfly targets. As expected, this far more challenging decoding task resulted in much lower classification rates. Classification was again very high for target present trials, a result again following from the behavioral data. Targets were very likely to be fixated first and longest, so performance in this condition essentially validated the bear and butterfly classifiers. However, for the more interesting target-absent conditions these rates were often at or near chance. This was especially true for the butterfly targets, where accuracy was above chance only in the TA-HML condition ($p < .01$). Accuracy was above chance in each of the bear conditions ($p < .01$). In general, these target-absent classification rates tended to mirror the difficulty in behaviorally selecting the target-similar distractor. Classification was better for bears than butterflies because bear-similar distractors were fixated more consistently by subjects. Likewise, classification was best when the display contained both a target-similar and a target-dissimilar distractor (TA-HML), was worse when only a target-similar distractor was present (TA-HM), and was only slightly better than chance in the control condition consisting only of random objects. This graded pattern of classification performance was clearest in the case of the LFO method, which we again attribute to the more consistent expression of target similarity effects in the durations of

distractor fixations. However, the difference between the two selection methods was smaller here compared to the by-subject classification. We attribute this to the fact that on 67.84% of the trials the first fixated object was also the longest fixated object, thereby confining the expression of a difference between these methods to relatively few trials.

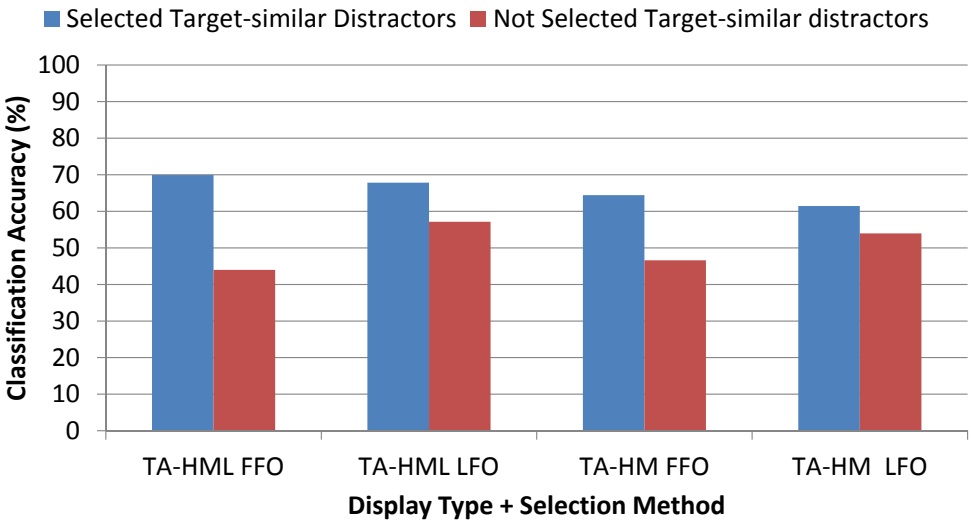


Figure 4.13: Classification rates conditionalized on whether the target-similar distractor was selected (blue bars) or not (red bars).

The previous results showed good classification accuracy when the target-similar objects were selected by subjects with their gaze, but what classification rates would be expected if only the target-similar objects were selected? To answer this question we grouped the data into cases in which the target-similar object was selected and cases in which the target-similar object was not selected and show these conditional classification rates in Figure 4.13. The blue bars indicate a sort of upper bound on classification rates in this taskâĂŤhow well the models were able to classify the target-similar distractors as either bears or butterflies. Although these optimal classification rates are quite high, 60-70%, they also suggest an upper limit on the level of by-trial classification success reported in Figure 4.12b---classification success would not be expected to exceed these levels. Trials in which the target-similar distractor was not selected (red bars) produced classification rates at or near chance when averaged over selection method, as expected.

Chapter 5

Conclusions and Future Work

In this thesis, we analyze the relationship between human behaviors and object detection methods in computer vision on both guidance and recognition task. Our behavioral studies show that targets show very strong search guidance, measured by the first fixated objects. Also guidance to non-targets objects is in proportion to their visual similarity to the target; high-similarity objects were first fixated the most and low-similarity objects the least. We design several computation experiments to compare object detection performance of our computational models with human observers' behaviors during visual search: First, we extend a previously successful model of eye movements during search (TAM) to the task of object class detection and this model shows similar confusion pattern during object detection comparing to human behavior. Second, we train and evaluate computational vision algorithms for object category recognition in order to compare their output to the human behavior. Some algorithms do well at predicting which object humans will fixate first, but there are differences between which features perform best for classification and which predict human behavior most closely. Also we show an observer's search target can be predicted based on their fixation pattern using two SVM-based classifiers, especially when one of the distractors in the search display were rated as being visually similar to the target category.

5.1 Conclusion

The behavioral and computational findings from this study make a profound contribution to our understanding of categorical search. First, they tell us that search guidance to categorical targets can be accomplished using purely visual information. This had not been established, and indeed categorical guidance had even been used as an argument for the existence of semantic features [72]. Our model serves as an implementation proof that visual features alone are sufficient to describe this fundamental human behavior, at least for the target categories explored in this study. Second, our behavioral data tell us that categorical guidance is subject to the same visual similarity relationships that are known to guide search to specific visual targets [14, 51]. This

too was not known. Our data show that search objects are rank ordered based on their visual similarity to a target class, and that gaze is sent to the object offering the most evidence for being the target. Third, the fact that our model was able to capture these similarity relationships using biologically-plausible features and a SVM is informative, and arguably unexpected. SVM forms decision boundaries to distinguish targets from non-targets; it was not at all certain whether the same boundaries used for classification would also separate high- and low-similarity non-targets. The fact that this was the case suggests that distance from a decision boundary in a SVM might be used to study visual similarity relationships more broadly.

We evaluate several computer vision models and compare their outputs with human behaviors. Our results show that computer vision models are able to capture human object detection at a very fine grain; describing not only detection performance, but also the patterns of confusions that determine human search efficiency. This is important in cases where we are interested in algorithms that return results that are relevant to human users of search. Our results also show that among all computer vision models we evaluated, while accuracy in object detection / recognition is quite high for all of them, there are significant differences in how well the models predict what objects are confusing to human.

In our behavior decoding study we demonstrate that the information available in the fixation behavior of subjects as they search is often sufficient to decode the category of their search target. By analyzing the distractors that were preferentially fixated during search, we found that the search target could be decoded perfectly when one of the distractors was rated as being visually similar to the target category. Even with completely random distractors, the target category could still be decoded for 75-80% of the subjects. The much harder task of decoding the target on individual trials (from a single distractor fixation) resulted in much lower classification rates, although targets were still decoded above chance. Two methods of preferential fixation were explored, the object first fixated during search and the object fixated the longest. Although both methods carried information about target category, the LFO method was used most consistently by subjects and resulted in better classification success. This suggests that the information related to target verification and distractor rejection is more useful for the behavioral decoding of search targets than information about preferential fixation typically associated with search guidance.

5.2 Future work

In future work we will further explore computational models' behavior in increasingly challenging contexts, such as manipulating the number of objects in the search display, embedding these objects in complex backgrounds, and increasing variability within the target class. This latter goal might be accomplished in one of two ways, by having observers and the model search simultaneously for two targets (either a teddy bear or a butterfly), or by specifying the target at different levels in the categorical hierarchy (Monarch butterfly, butterfly, or flying insect). We will also test whether the model generalizes to other target classes. In additional to the first fixated objects, we will also extend our work to study the whole fixation patterns and visual behaviors. On the topic of behavior decoding, we will extend our target classes to multiple classes.

Also the fixation patterns can be taken as features instead of FFO and LFO methods.

References

[1] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:511, 2001.

[2] Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, pages 878--885, 2005.

[3] M.P. Eckstein. The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, 9(2):111--118, 1998.

[4] J.T. Townsend. Serial vs. parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, 1(1):46--54, 1990.

[5] A. Treisman. Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3):652--676, 1991.

[6] J. Palmer. Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, 4(4):118--123, 1995.

[7] T.S. Horowitz and J.M. Wolfe. Visual search has no memory. *Nature*, 394(6693):575--577, 1998.

[8] X. Chen and G.J. Zelinsky. Real-world visual search is dominated by top-down guidance. *Vision Research*, 46(24):4118--4133, 2006.

[9] G.J. Zelinsky, R.P.N. Rao, M.M. Hayhoe, and D.H. Ballard. Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, 8(6):448--453, 1997.

[10] B.C. Motter and J. Holsapple. Saccades and covert shifts of attention during active visual search: Spatial distributions, memory, and items per fixation. *Vision Research*, 47(10):1261--1281, 2007.

[11] G.L. Malcolm and J.M. Henderson. Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, 10(2):1--11, 2010.

[12] G.L. Malcolm and J.M. Henderson. The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, 9(11):1--13, 2009.

[13] M.S. Castelhano, A. Pollatsek, and K.R. Cave. Typicality aids search for an unspecified target, but only in identification and not in attentional guidance. *Psychonomic Bulletin & Review*, 15(4):795--801, 2008.

[14] G.J. Zelinsky. A theory of eye movements during target acquisition. *Psychological Review*, 115(4):787--835, 2008.

[15] B.C. Motter and E.J. Belky. The guidance of eye movements during active visual search. *Vision Research*, 38(12):1805--1815, 1998.

[16] R.G. Alexander and G.J. Zelinsky. Visual similarity effects in categorical search. *Journal of Vision*, 11(8):9,1--15, 2011.

[17] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. pages 886--893, 2005.

[18] *MIT Pedestrian Database MITP*. Online, 2000.

[19] D.G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150--1157. Ieee, 1999.

[20] Jeffrey S. Beis and David G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. pages 1000--1006, 1997.

[21] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision--ECCV 2006*, pages 404--417, 2006.

[22] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:506--513, 2004.

[23] D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429--441, 1946.

[24] J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2:1160--1169, 1985.

[25] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(1):55--73, 1990.

[26] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg. The bochum/usc face recognition system and how it fared in the feret phase iii test. 1998.

[27] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE CVPR 2006*, volume 2, pages 994--1000. IEEE.

[28] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59--70, 2007.

[29] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, pages 411--426, 2007.

[30] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. 2007.

[31] N. Pinto, D.D. Cox, and J.J. DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27, 2008.

[32] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[33] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22. Citeseer, 2004.

[34] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524--531. Ieee, 2005.

[35] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. *Image and Video Retrieval*, pages 1950--1950, 2004.

[36] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *IEEE International Conference on Computer Vision*, 2005.

[37] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. 2005.

[38] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169--2178. Ieee, 2006.

[39] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 100(1):67--92, 1973.

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

[41] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, pages 1627--1645, 2009.

[42] P.F. Felzenszwalb, R.B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 2241--2248. IEEE, 2010.

[43] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137--154, 2002.

[44] P. Viola, M.J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153--161, 2005.

[45] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107--123, 2002.

[46] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205--231, 2005.

[47] U. Rutishauser and C. Koch. Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision*, 7(6):1--20, 2007.

[48] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113(4):766--786, 2006.

[49] C. Kanan, M.H. Tong, L. Zhang, and G.W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6):979--1003, 2009.

[50] K.A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6):945--978, 2009.

[51] J.M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202--238, 1994.

[52] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219--27, 1985.

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

[54] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194--203, 2001.

[55] N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in neural information processing systems*, 18:155, 2006.

[56] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[57] J. Schmidt and G.J. Zelinsky. Search guidance is proportional to the categorical specificity of a target cue. *The Quarterly Journal of Experimental Psychology*, 62(10):1904--1914, 2009.

[58] W. Zhang, H. Yang, D. Samaras, and G. Zelinsky. A computational model of eye movements during object class detection. *Advances in Neural Information Processing Systems*, 18:1609, 2006.

[59] M.P. Eckstein, B.A. Drescher, and S.S. Shimozaki. Attentional cues in real scenes, saccadic targeting, and bayesian priors. *Psychological Science*, 17(11):973, 2006.

[60] K.A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945, 2009.

[61] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157--173, 2008.

[62] H. Yang and G.J. Zelinsky. Visual search is guided to categorically-defined targets. *Vision research*, 49(16):2095--2103, 2009.

[63] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91--110, 2004.

[64] A.D. Hwang, E.C. Higgins, and M. Pomplun. A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, 9(5):1--18, 2009.

[65] L.G. Williams. The effect of target specification on objects fixated during visual search. *Attention, Perception, & Psychophysics*, 1(5):315--318, 1966.

[66] A.M. Derrington, J. Krauskopf, and P. Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357(1):241--265, 1984.

[67] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11--32, 1991.

[68] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. 2001.

[69] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advanceds in Large Margin Classifiers*, pages 61--74. MIT Press, 1999.

[70] W.S. Geisler and J.S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. In *Proc. SPIE*, volume 3299, pages 294--305, 1998.

[71] P. Cockrill. *The teddy bear encyclopedia*. DK Publishing, Inc., 2001.

[72] M.W. Becker, H. Pashler, and J. Lubin. Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance*, 33(1):20--30, 2007.