Atten Percept Psychophys (2017) 79:2021–2036
DOI 10.3758/s13414-017-1359-9

How do targets, nontargets, and scene context influence real-world object detection?

Harish Katti[1] · **Marius V. Peelen**[2] · **S. P. Arun**[1]

Published online: 28 June 2017

Abstract Humans excel at finding objects in complex natural scenes, but the features that guide this behaviour have proved elusive. We used computational modeling to measure the contributions of target, nontarget, and coarse scene features towards object detection in humans. In separate experiments, participants detected cars or people in a large set of natural scenes. For each scene, we extracted target-associated features, annotated the presence of nontarget objects (e.g., parking meter, traffic light), and extracted coarse scene structure from the blurred image. These scene-specific values were then used to model human reaction times for each novel scene. As expected, target features were the strongest predictor of detection times in both tasks. Interestingly, target detection time was additionally facilitated by coarse scene features but not by nontarget objects. In contrast, nontarget objects predicted target-absent responses in both person and car tasks, with contributions from target features in the person task. In most cases, features that speeded up detection tended to slow down rejection. Taken together, these findings demonstrate that humans show systematic variations in object detection that can be understood using computational modeling.

Keywords Categorization · Scene Perception · Object Recognition

Electronic supplementary material The online version of this article (doi:10.3758/s13414-017-1359-9) contains supplementary material, which is available to authorized users.

✉ Harish Katti
harish2006@gmail.com

[1] Centre for Neuroscience, Indian Institute of Science, Bangalore 560012, India

[2] Center for Mind/Brain Sciences, University of Trento, 38068 Rovereto, Italy

Detecting objects such as cars or people in natural scenes is effortless for us (Li, VanRullen, Koch, & Perona, 2002; Thorpe, Fize, & Marlot, 1996), but extremely challenging for computer algorithms (Everingham et al., 2014). It is challenging because natural scenes vary not only in target appearance but also in the surrounding objects and overall scene layout, all of which can potentially facilitate detection. To illustrate these three different properties, consider the scene of Fig. 1a. Finding a person in this scene could be slow due to the unusual view, but could be aided by objects such as a bag or dustbin that co-occur near people, and by the coarse scene layout typical of people-associated market scenes. Likewise, finding a car in Fig. 1b may be aided by many cars, informative objects such as traffic lights and by the coarse layout typical of car-associated road scenes. The goal of our study was to characterize how these three information channels (targets, nontargets, and coarse scene features) influence target detection in both tasks. Below we review the existing literature in relation to this overall goal.

Object detection in humans

The general approach taken to understand target features used by humans has been to characterize how performance changes with various image manipulations. Object detection, for

Fig. 1 Representative scenes and experiment design. **a** *Left:* Example image in the person detection task. *Right:* Targets, nontarget objects (pram, stand, bag), and coarse scene structure. All these properties can potentially influence the eventual detection of the target. **b** Example image from the car detection task, with targets, nontargets, and coarse scene information. **c** Stimuli used in the two tasks. Subjects performed either a car detection task or a person detection task consisting of 1,300 scenes. Scenes unique to each task and common to both tasks are shown. (Color figure online)

instance, is unaffected by removal of color or Fourier phase (Harel & Bentin, 2009; Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009; Morrison & Schyns, 2001), but depends on coarse object features (Fabre-Thorpe, 2011; Mohan & Arun, 2012) that may vary in their spatial frequency (Harel & Bentin, 2009). It is generally thought that object detection is driven by category-diagnostic features of intermediate complexity (Delorme, Richard, & Fabre-Thorpe, 2010; Reeder & Peelen, 2013; Ullman, Vidal-Naquet, & Sali, 2002).

There is also evidence that, apart from using target features to perform object detection, humans can exploit scene regularities to efficiently search for target objects (Castelhano & Heaven, 2010; Malcolm, Nuthmann, & Schyns, 2014; Torralba, Oliva, Castelhano, & Henderson, 2006) and use these predictions to facilitate object recognition (Auckland, Cave, & Donnelly, 2007; Bar, 2004; Bar & Ullman, 1996; Biederman, Mezzanotte, & Rabinowitz, 1982; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008; Zimmermann, Schnier, & Lappe, 2010). Informative scene layouts also make it possible to detect targets at coarse resolutions (Barenholtz, 2013), and coarse scrambling of surrounding context has been shown to impair face detection (Lewis & Edmonds, 2003). However, the surrounding context can consist of either nontargets similar to the target, dissimilar but co-occurring nontargets, or informative scene layouts—all of which might influence search for a target. There have been very few attempts to disentangle these factors, with two notable exceptions. The first is the finding that nontargets similar to the target attract gaze (Zelinsky, Peng, & Samaras, 2013a) and more generally, influence search difficulty (Duncan & Humphreys, 1989; Vighneshvel & Arun, 2013). The second is the finding that fixation locations of subjects searching for a person in a scene can be predicted to some extent using local salience and target features, but best of all by coarse scene features (Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009).

While most studies of object detection have naturally focused on target-present scenes, human performance on target-absent scenes is relatively less understood. Are target-absent responses systematic, and, if so, what makes a scene easy or hard to reject? In general, targets are harder to find or reject with increasing clutter (Neider & Zelinsky, 2011; Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011), and fixations in target-absent scenes are driven by target and coarse scene features (Ehinger et al., 2009; Zelinsky et al., 2013a). However, the influence of targets, nontargets, and coarse scene features on target rejection is poorly understood.

Object detection in machines

Although machine vision systems do not yet match humans in performance, they nonetheless indicate the detection performance possible with each class of features. For instance, models of object detection based on even rudimentary target features such as configurations of gradient-based whole object detectors (Dalal & Triggs, 2005) or containing representations of the whole object and its parts (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010) have been reasonably successful in object detection. These suggest that object detection can be driven primarily by target features alone without relying on nontarget objects or context. Hierarchical models with biologically plausible operations (Serre, Oliva, & Poggio, 2007) have been used to categorize objects with some success. Recently, convolutional neural networks have outperformed all other detectors (Krizhevsky, Sulskever, & Hinton, 2012), although our understanding of their features remains elusive. While object features have typically been modeled using intensity gradients, contextual information has been modeled using coarse Gabor features computed across the entire scene. Such coarse scene information has been used to classify scenes, constrain object detection, and even identify incongruent portions of scenes (Choi, Torralba, & Willsky, 2012; Oliva & Torralba, 2001, 2008; Torralba, 2003).

Overview of the current study

To summarize, studies of object detection in humans and machines have shown that target features and scene context can guide detection, but the relative influences of target, nontarget, and coarse scene features in object detection remain unclear. To address this issue, we performed two experiments in which participants performed person and car detection. In each experiment, we measured human performance on real-world scenes and used this data to train computational models.

There were two distinctive aspects of our approach: First, we measured human performance on detecting two different objects (people or cars) on the same set of scenes. This allowed us to measure task-specific and task-independent contributions on object detection. Second, we trained computational models on three distinct information channels: isolated target object features, presence/absence of other objects, and features extracted from blurred scenes. This allowed us to characterize how each type of information in a scene (targets, nontarget objects, and coarse scene structure) influences object detection. We sought to explain the scene-by-scene variation in the average response time of subjects performing the person or car detection tasks. However, we fit separate models on target-present and target-absent scenes based on the premise that they may be qualitatively different: for example, a fast response on a present scene may occur for a scene with an easy target, whereas a fast response on an absent scene may occur for a scene where it is easy to reject a target. Our main finding is that different channels contribute to target-present and target-absent responses in a task-dependent manner. These results yield insights into the relative importance of three types of scene information during rapid object detection and the underlying mechanisms.

Method

Participants

In all, 30 subjects (20–30 years old, five female) participated in the person detection task, and 31 subjects (20–30 years old, 11 female) participated in the car detection task. All subjects had normal or corrected-to-normal vision and gave written informed consent to an experimental protocol approved by the Institutional Human Ethics Committee of the Indian Institute of Science.

Person task stimuli

A total of 1,300 full color real-world scenes (measuring 13.5° × 10.1°, native resolution 640 × 480 pixels) were selected from a wide range of environments, from natural to urban, with varying numbers of people. Half of these scenes (n = 650) contained no people or cars and were common to both person detection and car detection tasks. Of the 650 person-present scenes, 225 also contained cars (common to both tasks), and the remaining 425 contained only people (unique to this task). Thus, there were 1,725 unique scenes that were used across both tasks (650 + 225 common and 425 scenes unique to each task). A large fraction of these scenes were from the LabelMe dataset (Russell, Torralba, Murphy, & Freeman, 2008) and were used in a previous fMRI study (Peelen, Fei-Fei, & Kastner, 2009), and the rest were from a personal collection of one of the authors (M.V.P).

The 425 person-present scenes varied widely in their field of view and included buildings, streets, residential neighborhoods, shop-fronts, parks, sports venues, lakes, rivers, mountainous terrains, and so on. The 225 scenes that contained both cars and people were relatively less variable in their scene properties because both objects tend to co-occur in scenes such as neighborhoods with streets and parking areas. Although there was plenty of overlap of scene types between the 650 person-present and 650 target-absent scenes, there was greater variability in terms of field of view and scene depth for the absent scenes. Target-absent scenes also contained scenes such as highways that are associated more with cars than people.

Car task stimuli

The stimuli consisted of 1,300 full-color real-world scenes with cars in a variety of scene locations, poses, and environments. Of these, 650 scenes contained one or more cars, and the remaining 650 contained neither cars nor people and were common to both tasks. Of the 650 car-present scenes, 225 also contained people and were common to both tasks. Car-present scenes had a greater incidence of highways, parking lots, and urban streets.

Procedure

Subjects were seated ~60 cm from a computer monitor under control of custom programs written in PsychToolbox (Brainard, 1997; Pelli, 1997). Each trial began with a fixation cross (measuring 0.45° square) presented for 500 ms, followed by a scene that flashed briefly for 83 ms, and then by a noise mask (measuring 13.5° × 10.1°) that stayed on for 450 milliseconds and was replaced by a blank screen until a response was made. Noise masks were created by superimposing naturalistic textures on a mixture of white noise at different spatial frequencies (Walther & Fei-Fei, 2007). Subjects in the person (car) detection task were instructed to respond as quickly and accurately as possible using a key press to indicate whether they saw one or more people (cars) in the scene, using the key "Y" for present and "N" for absent. The trial timed out after 4.5 seconds. Trials with incorrect or no responses were repeated after a random number of other trials. In all, we collected 1,300 correct responses from each subject in each task. Although repeated scenes may have elicited qualitatively different responses (due to memory or priming effects), in practice they were too few in number to analyze because subjects were highly accurate (>92% correct in both tasks; 76 incorrect trials for 1,300 correct trials for the person task; 103 incorrect for 1,300 correct trials for the car task). Their accuracy was high, even upon considering the first responses to all scenes (average accuracy: 94% for person; 92% for car task). Excluding these repeated scenes from our analyses yielded extremely similar results. All results reported in the main text are (for simplicity) based on analysis of correct trials that included responses to repeated scenes.

Noise ceiling estimates

To estimate an upper bound on model performance for the first approach (generalization across scenes), we reasoned that no model can exceed the reliability of the data itself. Although this number can be measured using the split-half correlation between two halves of the data, it underestimates the true reliability of the data, which is the correlation that would have been obtained if there were two data sets of the same size. Since the cross-validated models are trained on RT data from all subjects, we accordingly applied a Spearman-Brown correction (with data-split parameter = 2), which estimates the true reliability of the data. This is calculated as $rd = 2rs/(rs + 1)$, where rs is the split-half correlation and rd is the data reliability or noise ceiling. The mean and standard deviation of the noise ceiling shown in the corresponding figures was obtained by calculating the corrected split-half correlation over many two-way splits of the data. Noise ceiling estimates for the second approach (generalization across subjects) were calculated as the average correlation between each held-out subject's response times across scenes with the average response times across the remaining subjects.

Model fitting

We sought to explain the scene-by-scene variation in average response time and fit separate models on target-present and target-absent scenes. To evaluate the contributions of targets, nontargets, and context, we fit separate models on each group of features using linear regression. For instance, to predict responses on target-present scenes using target features, we compiled the average response times across all 650 scenes into a single response vector, **y**. The corresponding target features for each scene were compiled into a matrix, **X**, containing 650 rows and columns equal to the number of target features. We then asked whether the response time could be predicted using a linear combination of target feature activations: in other words, whether $\mathbf{y} = \mathbf{Xb}$ for some unknown activation weights **b**. The weights **b** were found using standard linear regression. This was done likewise for other groups of target features. In addition to fitting the behavioural data to each channel separately, we also fit composite models in which the feature vector for each scene corresponded to the concatenated feature vectors from individual channels.

A channel with more features could yield a better model solely because it has more degrees of freedom. We therefore reduced the dimensionality of each channel using principal component analysis performed across all scenes and projected the features in each group along their first 20 principal components. These principal components captured more than 85% of variance across scenes for all three channels, indicating that they provide a compact description of the features extracted in each channel.

Model evaluation

We used two different approaches to evaluate model performance. While these two approaches are quantitatively different, the results were qualitatively highly similar and are detailed below.

The first approach was to use the model to predict the responses to novel scenes. Here, we evaluated how well a

model trained on the average reaction times across subjects on one set of scenes can predict reaction times on another, independent set of scenes. To this end, we used a five-fold cross-validation procedure: the scenes were split into five parts, and for each split the predicted reaction time on this 20% of the data was obtained from a model trained on the remaining 80% of the data. This yielded predicted response times for all 650 scenes, where the response to each scene is based a model that is never exposed to that scene during training. We then reported model performance as the correlation between the predicted and observed reaction times obtained in this manner across 5,000 random 80-20 splits of the data. This approach has the advantage that it captures the group-averaged variation in reaction times and allows for generalization across scenes.

The second approach was to use the model to predict the responses of novel subjects. Here we evaluated model performance using a leave-one-subject-out approach. We calculated the degree to which a particular subject's reaction times could be predicted by a model trained on the remaining subjects. This approach has the advantage that each subject is an independent sample that can be directly used as degrees of freedom in statistical tests.

Choice of features in each channel

For each scene, we extracted image features corresponding to each channel (targets, nontarget objects, and scene layout). The features of each channel were designed to be distinct from each other, and are summarized briefly below with details in the Supplementary Material (Section S1).

Target features

We carefully selected target features that minimized any contamination from the surrounding context features.

Histograms of oriented gradients (HOG), are commonly used in object detection (Felzenszwalb et al., 2010). We chose a total of 31 features optimized for person/car detection learned from an independent set of close-cropped person and car images with very little context. These templates contain information about the coarse as well as fine grained structure of people and cars (Fig. 2) and have been trained on large-scale publicly available data sets distinct from those used in this study. Six templates each for canonical appearances or cars as well as people were obtained automatically by the training algorithm (Felzenszwalb et al., 2010). The degree of match to these HOG templates was obtained by convolving the car/person appearance templates in an exhaustive manner across image locations, at multiple scales. Strong matches to the reference templates typically result in hits (Figs. 2g–h) and weak or partial matches typically result in false alarms (Figs. 2i–j). Since HOG-based templates can yield a different number of matches, match-locations, scales, and detector confidence on each scene, we summarized these statistics into a 31 dimensional feature vector. These features are detailed in Section S1 and included (1) the weighted sum of detector confidence with detected box area for high-confidence matches (i.e., potential hits); (2) weighted sum of detector confidence with box area for low-confidence matches (i.e., potential false alarms). This method of aggregating target confidence scores for each scene was motivated by the faster detection of larger objects in real-world scenes (Wolfe et al., 2011) and by participant feedback indicating ease of detecting conspicuous and larger targets. Since there are six appearance templates each for car and person, we computed the weighted sum of hits and false alarms for each such template and concatenated them to obtain a summary descriptor containing pure object templates independent of context. We obtained qualitatively similar results on using only the relevant template for each experiment.

In contrast to earlier work (Ehinger et al., 2009), we used a much richer appearance model having part statistics for eight object regions in each view. Depending on the depicted view in the HOG appearance template, these regions roughly overlap with head, arms, torso, and legs in the case of persons, and wheels, windows and doors, hood and headlights, trunk and rear windscreen for cars (Felzenszwalb et al., 2010), and this allows us to quantify the part deformations relative to the coarse appearance template. We recorded the average and standard deviation of displacement of each part in a normalized space with unit height and width ($n = 16$). Part deformation was quantified by normalizing each detection into a unit square and finding the relative displacement of each detected part from the respective mean part location in all target matches over 1,300 scenes used in each task. This was done separately for matches to car appearance and for matches to person appearance. We found that summarizing the HOG detector results in this manner was more informative than using average HOG histograms directly (see Section S1).

We avoided using convolutional neural network activations because these often pick up both target and context features, which we subsequently confirmed (see Section S4).

Nontarget features

Nontarget features consisted of binary labels denoting the presence or absence (0 for *absent*, 1 for *present*) for a variety of objects in the scene ($n = 67$). These binary labels were obtained by manually annotating each scene exhaustively with all the visual objects present in it (see Section S1). Cars and people were specifically excluded from these labels since they are potential targets. We also excluded large regions, such as sky and mountain, that occur at scales large enough to influence coarse scene structure. Although we have chosen scenes with a large variety of nontarget objects for this study, it remains possible that the frequency of nontargets are not

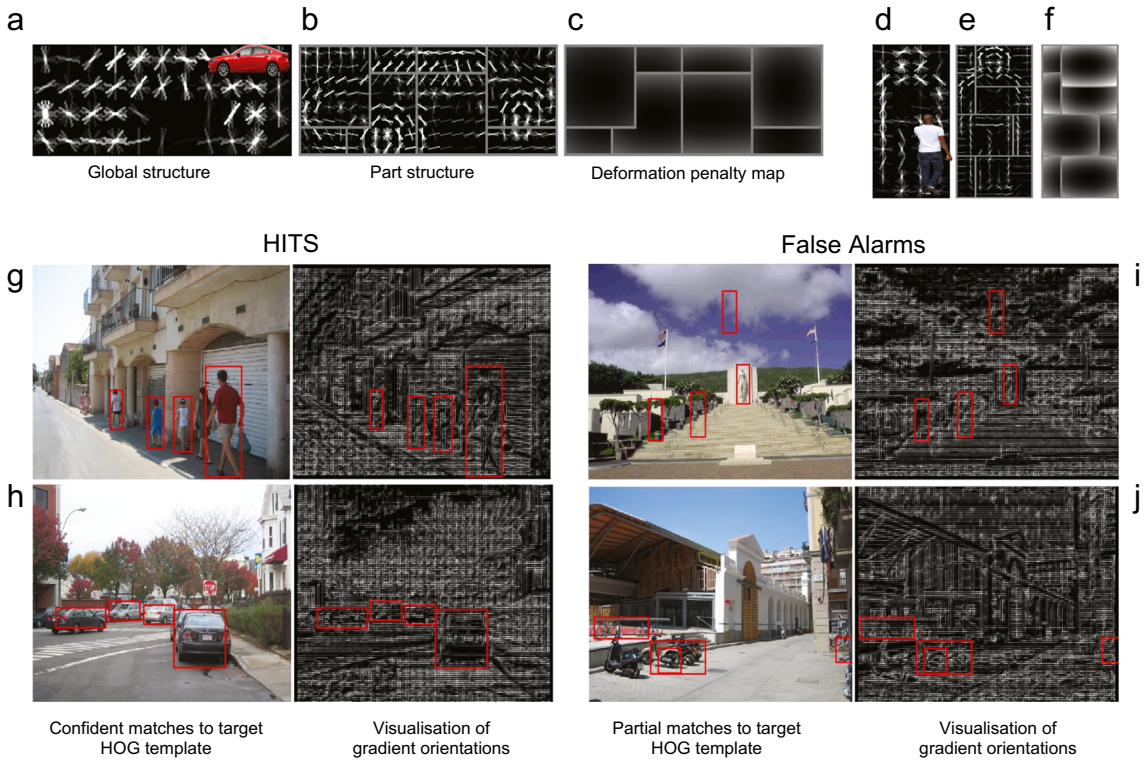

Fig. 2 Target feature models for person and car detection. Each panel shows a visualization of the features extracted by the target feature model for people and cars. The target feature model is a part-based model in which target detection is based on a collection of part and whole-object detectors, each of which match the orientations present in local image patches with a part of whole template. **a** Histogram of oriented gradient structure learnt for one of six canonical views of isolated cars. **b** Histogram of oriented gradient structure learnt for eight parts within each view of a car. **c** Deformation penalties that are imposed on the eight parts defined within each template for a car, where part detections in whiter regions incur more penalty. **d** Histogram of oriented gradient structure learnt for one of six canonical views of isolated people. **e** Histogram of oriented gradient structure learnt for eight parts within each view of a person. **f** Deformation penalties that are imposed on the eight parts defined within each template for a person (part detections in *whiter regions* incur more penalty). **g, h** High confidence matches to person and car HOG templates, respectively. **i, j** Partial matches to person or car HOG templates arising from coarse or fine-grained scene structure. (Color figure online)

representative of natural experience. This remains a poorly studied issue: many common datasets, such as MSCOCO (Lin et al., 2014), contain similar numbers of categories, but it is unclear whether humans are sensitive to nontarget occurrence statistics in general.

Coarse scene features

Coarse spatial envelope GIST features (Oliva & Torralba, 2001) were extracted from blurred versions of scenes (n = 512). The blurred scenes were obtained by convolving images with a low-pass Gaussian filter ($\sigma = 20$ pixels), such that internal details of objects and parts were no longer discernible. To confirm that this level of blurring removed all target-related features, we trained HOG-based object detectors for both cars and people on blurred scenes and confirmed that they yielded poor detection accuracy (<5% correct across 100 representative target-present scenes). We also tried encoding coarse structure using activations of a deep convolutional neural network optimized for scene categorization (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2014), but GIST features yielded better predictions of detection performance (data not shown).

Results

We characterized human performance on two rapid object detection tasks: 30 subjects performed a person detection task and a separate group of 31 subjects performed a car detection task (see Method section, Fig. 1). Subjects were extremely accurate in both tasks (average accuracy: 94% ± 0.03 in the person task; 92% ± 0.03 in the car task, considering only their first response to each unique scene). In general, they were faster to confirm the presence of a target than to confirm its absence (average reaction times for present and absent responses: 0.43 s and 0.50 s in the person task, $p < .001$ for main effect of condition in an ANOVA on response times with subject and condition [present/absent] as factors; 0.44 s and 0.53 s in the car task, $p < .001$ for main-effect of condition; $p < .001$ for interaction between subject and condition in both tasks). To assess whether reaction times are systematic across

scenes, we calculated the correlation between the average response times of one half of the subjects with the other half. This yielded a significant positive correlation for detection and rejection in both tasks (average split-half correlation for detection and rejection times: rs = .53 & .29 in the person task, rs = .68 & .29 in the car task, $p < .001$ in all cases). These numbers can be used to derive the upper limit for model performance (see Method section; rd = .69 & .45 for person task, rd = .81 & .45 in the car task). Thus, subjects were highly consistent in both tasks.

Computational models for person detection

We started out by characterizing the influence of targets, nontarget, and coarse scene features on person detection. To this end, we tested a number of models based on combinations of target, nontarget, and coarse scene features (see Tables 1 and 2). We evaluated models for their ability to predict the average responses to novel scenes that were never used in model fitting (Table 1) or their ability to predict responses of novel subjects that were never used in model fitting (Table 2). Both measures yielded qualitatively similar results (see Tables 1 and 2).

We sought to identify which combination of the three channels would provide the best account of the data. We illustrate our model selection process using the example of person detection using model generalization to new scenes (see Table 1). The best model was one that contained target and coarse scene features for the following reasons: (1) this model had significantly better fits than models informed by any single information type ($p < .001$ in all cases); (2) this model yielded significantly better fits compared to other pairs of channels ($p < .005$ vs. TN, $p < .001$ vs. NC); (3) this model had comparable performance to a full model containing target, nontarget, and coarse scene features ($p > .25$). We obtained similar results with model generalization to new subjects (see Table 2). Note, however, that target features are the dominant influence on detection with only a slight benefit from including coarse scene features.

The performance of the best person detection model on novel scenes is shown in Fig. 3. This model yielded a significant correlation between predicted and observed responses across all scenes (r = .45, $p < .001$; Table 1, Fig. 3a), where the reliability of the data is (rd = .69), suggesting that the model captures the major sources of variation in person detection. Likewise, it yielded a significant correlation between its predictions for each individual subject and the observed responses for that subject (average r across 30 subjects: r = .15; this correlation was significantly different from zero, sign-rank test, $p < .001$; see Table 2), reliability of the data is (rd = .23).

The predictions of the person detection model can be interpreted as being inversely related to the strength of evidence for a person in the image: the stronger the person features in the image, the faster will be the response. Would strong evidence in favor of a person in a person-absent scene result in a slow response? To test this possibility, we used person-absent scenes as input to the person detection model and compared these predictions with the observed person-rejection response times. This yielded a negative correlation (r = -.17, $p < .001$; Fig. 3a). Thus, features that speed up person detection in a scene slow down person rejection.

To gain further insights into model performance, we grouped scenes by their detection time to see whether there were

Table 1 Model generalization to new scenes in person and car detection

Model name	*dof*	*Person detection*	*Car detection*		
		rc	*p*(Model > TC)	*rc*	*p*(Mode l > TC)
Noise ceil		0.69 ± 0.02		0.81 ± 0.02	
TNC	60	0.44 ± 0.01	0.29	0.55 ± 0.01	0.087
T	20	0.41 ± 0.01	0.005	0.50 ± 0.01	0
N	20	0.14 ± 0.02	0	0.17 ± 0.02	0
C	20	0.30 ± 0.01	0	0.41 ± 0.01	0
TN	40	0.40 ± 0.01	0.002	0.48 ± 0.01	0
TC	40	**0.45 ± 0.01**	-	**0.57 ± 0.01**	0
NC	40	0.30 ± 0.01	0	0.42 ± 0.01	0

Note. The best performing model for both detection tasks was one that contained target as well as coarse scene features (values indicated in bold). We characterized the performance of each model using the average cross-validated correlation (*rc*) over 5,000 random 80–20 splits. Because the number of random splits is entirely arbitrary, a direct statistical comparison is meaningless. Instead, we took the statistical significance to be the probability (i.e., fraction of times) each given model yielded a correlation higher than the best model. Thus a probability of 0.05 means that the model X outperformed the best model only 5% of the time, implying that the best model is superior to model X. We chose the widely used criterion of $\alpha = 0.05$ for statistical significance. Note that model performance sometimes reduces with the inclusion of extra features because of overfitting. Model abbreviations: T, N, C represent targets, nontargets, and context. TN = targets & nontargets, TC = targets & context, rc = correlation between model predictions and observed response times, dof = degrees of freedom in the model

Table 2 Model generalization to new subjects in person and car detection

Model name	*dof*	*Person detection*		*Car detection*	
		rc	*p*(Model > TC)	*rc*	*p*(Model > TC)
Noise Ceil		0.23 ± 0.09		0.34 ± 0.1	
TNC	60	0.15 ± 0	0.1	0.23 ± 0	0.18
T	20	0.14 ± 0	0	0.21 ± 0	0
N	20	0.05 ± 0	0	0.07 ± 0	0
C	20	0.10 ± 0	0	0.17 ± 0	0
TN	40	0.13 ± 0	0	0.2 ± 0	0
TC	40	**0.15 ± 0**	–	**0.24 ± 0**	–
NC	40	0.10 ± 0	0	0.18 ± 0	0

New. Each subject is held out and models are trained for car/person detection RT. Correlations of model predictions with the held out subject's RT are averaged over all participants in the task. The best performing model for both detection tasks was one that contained target as well as coarse scene features (values indicated in bold). To compare the best model with each other model, we performed a paired nonparametric statistical comparison (ranksum test) between the individual subject correlations produced by the best model and each individual model. The resulting statistical significance is reported under the *p*(Model > T) column. Abbreviations as in Table 1

common visual elements as illustrated for a few representative scenes that produced fast or slow responses (see Fig. 3b, top row). The observed responses closely follow the predictions from a model trained on target and coarse scene features. As shown in Fig. 3b, relatively strong person evidence (e.g., stronger and greater number of matches to person templates using HOG models) predicted relatively fast responses typically associated with larger scales and numbers of people.

Computational models for car detection

Next, we investigated models based on target, nontarget, and coarse scene features for their ability to predict car detection performance. The results are summarized using model generalization to new scenes (see Table 1). As with person detection, we found that the best model for car detection included target and coarse scene features but did not benefit from including nontarget features. Target features were the predominant influence on detection with only a slight benefit from including coarse scene features. This model (target + coarse scene features) yielded a significant correlation between predicted and observed responses ($r = .57$, $p < .001$; see Table 1 and Fig. 4a). Given the reliability of the data itself ($rd = .81$), this indicates that the model captures the major sources of variation in car detection. We obtained similar results using model generalization to new subjects (see Table 2).

Next, we tested whether partial evidence in favor of target presence, which speeds up the detection of targets that are actually present, would slow down target rejection, as observed in the person detection task. To test this prediction, we used the target feature model (trained on car-present scenes) to predict the response times on car-absent scenes. These predicted responses show a weak but significant negative correlation with the observed car-absent response times ($r = -.03$, $p < .001$; see Fig. 4a). Thus, features that speed up car detection in a scene slow down car rejection and suggests that car rejection is influenced by target features.

To elucidate model performance on car detection, we illustrate model features for representative scenes with fast or slow detection times (Fig. 4B). Scenes with fast detection responses generally contained high confidence matches for cars in the model and scene layouts that are strongly associated with cars and also allow for easy detection. We conclude that target and coarse scene features play a dominant role in car detection.

Computational models for person rejection

We then investigated whether computational models can explain times taken by humans to confirm the absence of people in an image. To this end we again computed features corresponding to target, nontarget, and coarse scene information from each scene, and asked whether the average person-absent response time can be explained as a linear combination of these features. The results are summarized in Table 3 for model generalization to new scenes and in Table 4 for model generalization to new subjects. The best model in this case was one that included both target and nontarget information (see Tables 3 and 4). Including context features did not significantly improve the fit (see Table 3). This model yielded a significant correlation between predicted and observed responses ($r = .33$, $p < .001$; see Table 3, Fig. 3c), close to the reliability of the data ($rd = .45$), suggesting that the model captures the major sources of variation in person rejection. Importantly, the combined target-nontarget model performed significantly better than the two simpler models (vs. target, $p < .0005$; vs. nontarget, $p < .0005$) as well as other models informed by two information channels (vs. TC, $p < .001$; vs. NC, $p < .001$), adding coarse scene information did not

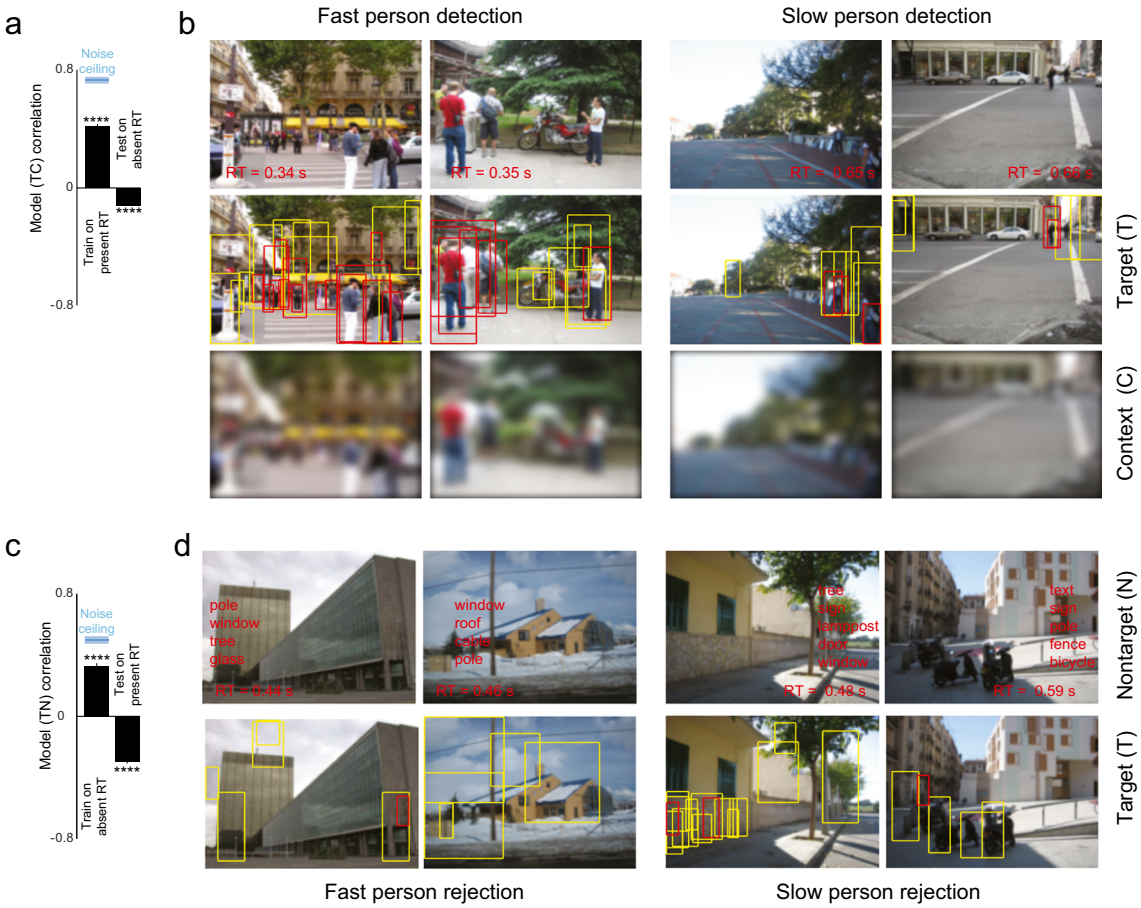

Fig. 3 Computational models predict person detection (Experiment 1). **a** *Left bar:* Correlation between the best model predictions and person detection response times. The noise ceiling, which estimates the upper limit on model performance given the reliability of the data, is also shown for comparison (*blue*). *Right bar:* Correlation between best model predictions on target-absent scenes and person rejection response times, showing a negative correlation. Here and throughout, *error bars* represent standard deviations in cross-validated correlations obtained across 1,000 random 80-20 splits of the data. **b** Visualization of model features in scenes with fast (*left*) and slow (*right*) observed detection response times. The first row depicts the original scene with the average response time shown in red. The second row depicts target features: scenes are overlaid with weak (*yellow boxes*) and strong (*red boxes*) matches to target features. The third row depicts the blurred scenes used to extract coarse scene features. **c, d** Analogous plots for person rejection with inset text in first row showing nontarget labels and second row depicting weak (*yellow boxes*) and strong (*red boxes*) matches to target features. (Color figure online)

improve the performance of this model ($p > .05$). We obtained similar results using model generalization to new subjects (see Table 4).

As before, we reasoned that features that slow down person rejection are presumably those that contribute evidence in favor of person present. To confirm this prediction, we took the model trained on person-absent scenes (i.e., the target-nontarget model) and calculated its responses to person-present scenes. This yielded a negative correlation with target-present responses ($r = -.28$, $p < .001$; Fig. 3c), confirming that features that slow down person rejection indeed speed up person detection.

To illustrate model performance on person rejection, we illustrate model features for representative scenes with fast or slow detection times (Fig. 3c). Person rejection was slowed down by partial target matches as well as by nontarget objects that co-occur with people, such as trees, pillars, and parked vehicles. It is also slowed by nontarget objects such as doors that are similar to people in HOG feature space (see Fig. S2a, Supplementary Section S3). Taken together our results show that the target features and nontarget objects are the dominant influence on person rejection.

Computational models for car rejection

We then investigated whether computational models can explain times taken by humans to confirm the absence of cars in an image. To this end we again computed features corresponding to target, nontarget, and coarse scene information from each scene, and asked whether the average car-absent response time can be explained as a linear combination of these features. The results are summarized in Table 3 for model generalization to new scenes and in Table 4 for model generalization to new subjects.

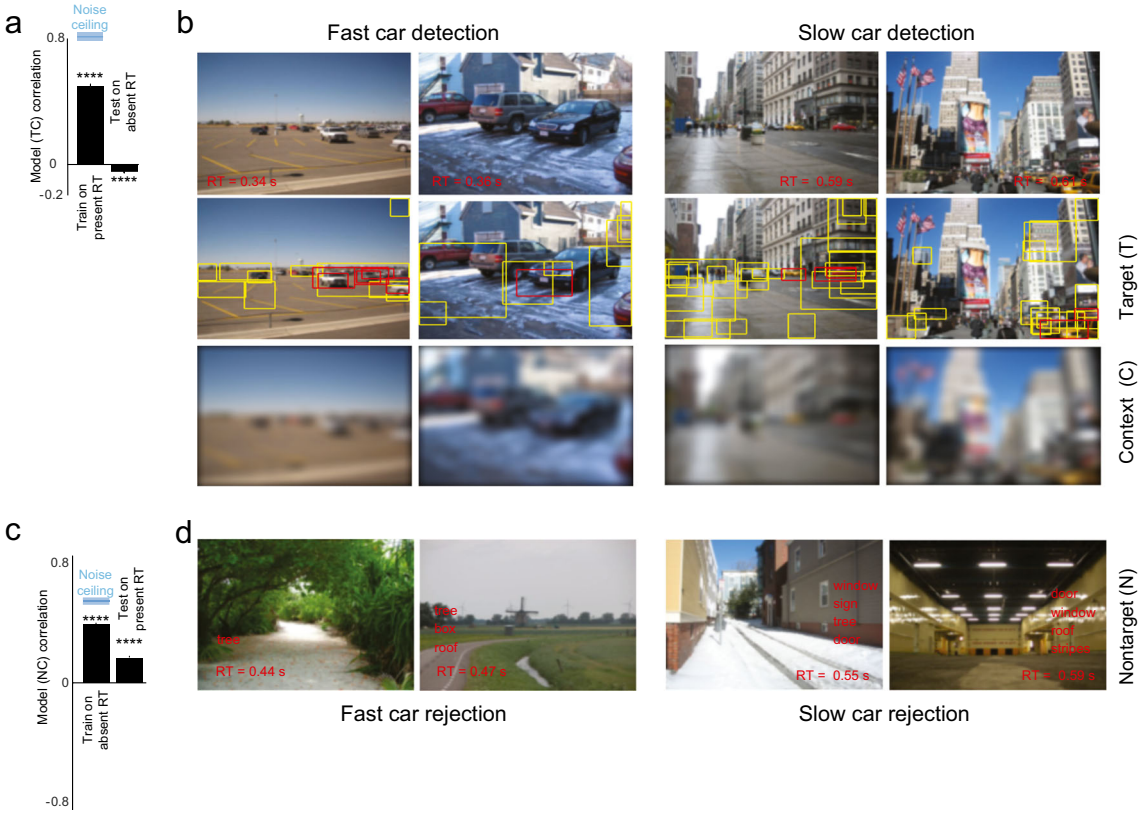

Fig. 4 Computational models predict car detection (Experiment 2). **a–d** Analogous plots as in Fig. 3, but for car detection. **a** Correlation between best model predictions and car detection response as well as best model prediction on car absent scenes and car rejection response times, all conventions as in Fig. 3. **b** Visualizing original scenes with overlaid observed car detection response times in the first row, target matches in second row and coarse scene information in the third row. **c, d** Analogous plots for car rejection with **d** showing nontarget labels visualized on novel scenes sorted according to increasing car rejection response times. (Color figure online)

The best models for car rejection were those containing nontarget features alone, with no contribution of either target or coarse scene features (see Table 3). This model yielded a significant correlation with car-absent response times ($r = .34$, $p < .001$; see Table 3 and Fig. 4c), where the reliability of the data itself is ($rd = .45$). Adding the target or coarse scene information did not improve performance. The nontarget model performed significantly better than models containing only target or only coarse scene information (vs. target, $p < .0005$; vs. coarse scene information, $p < .0005$), and adding other information channels did not significantly improve the performance of the nontarget only model (vs. all other models with two or three information types, $p > .05$). We obtained similar results using model generalization to new subjects (see

Table 3 Model generalization to new scenes in person and car rejection

Model name	*dof*	*Person rejection*		*Car rejection*	
Noise ceil		*rc* 0.45 ± 0.02	*p*(Model > TN)	*rc* *0.45 ± 0.02*	*p*(Model > N)
TNC	60	0.29 ± 0.02	0.09	0.34 ± 0.02	0.42
T	20	0.25 ± 0.01	0	0.06 ± 0.02	0
N	20	0.20 ± 0.02	0	**0.34 ± 0.01**	–
C	20	0.14 ± 0.02	0	0.15 ± 0.02	0
TN	40	**0.33 ± 0.02**	–	0.35 ± 0.01	0.65
TC	40	0.23 ± 0.02	0	0.13 ± 0.02	0
NC	40	0.21 ± 0.02	0	0.35 ± 0.01	0.58

Note. The best model for person rejection was the TN, model and the best model for car rejection was the N model. Conventions are as in Table 1

Table 4 Model generalization to new subjects in person and car rejection

Model name	*dof*	*Person rejection*		*Car rejection*	
		rc	*p*(Model > TN)	*rc*	*p*(Model > N)
Noise ceil		0.12 ± 0.08		0.12 ± 0.05	
TNC	60	0.07 ± 0	0.04	0.085 ± 0	0.71
T	20	0.07 ± 0	0.001	0.02 ± 0	0
N	20	0.05 ± 0	0	**0.09 ± 0**	–
C	20	0.03 ± 0	0	0.04 ± 0	0
TN	40	**0.085 ± 0**	–	0.09 ± 0	0.8
TC	40	0.06 ± 0	0	0.03 ± 0	0
NC	40	0.05 ± 0	0	0.09 ± 0	0.67

Note. Each subject is held out, and models are trained for car/person detection/rejection RT. The best model for person rejection was the TN model, and the best model for car rejection was the N model. All conventions are as in Table 2

Table 4). Both analyses show that nontargets have the dominant influence on car-absent responses.

As before, we reasoned that features that slow down car rejection are presumably those that contribute evidence in favor of car presence. To test this prediction, we took the model trained on car-absent scenes using nontarget features alone and calculated its responses to car-present scenes. These predicted responses were positively correlated with car-present response times ($r = .15$, $p < .001$; see Fig. 4c). Though this is contrary to our prediction, it agrees with participant feedback indicating that many nontargets can add to clutter and slow down detection (see Section S3).

When visually inspecting scenes grouped according to slow or fast rejection (Fig. 4d), we found that increase in the presence of associated nontarget objects, like signs and poles, elements of urban facades like windows, and other coarse scene structures such as urban environments all slow down car-absent responses—and indeed these correspond to scenes likely to contain cars (see Section S3).

Task specificity of computational models

So far, we have shown that target detection and rejection in both car and person tasks can be predicted using computational models. However, these results do not yet provide conclusive evidence that the models are specific in predicting car or person detection performance. It is possible, in principle, that a model that accurately predicts person detection (or rejection) reaction times may similarly predict car detection (or rejection) reaction times, for example because it capitalizes on factors that generally influence response times (e.g., visibility, clutter, scene typicality). To test for task specificity, we analyzed the subset of scenes common to both tasks: these included 225 car-and-person-present scenes and 650 scenes with both targets absent. We first asked whether, across scenes containing both cars and people, scenes that took longer for person detection also took longer for car detection. We observed no significant correlation between the average detection times across the 225 car-and-person-present scenes in the two tasks ($r = -.12$ $p = .08$; see Fig. 5a). In contrast, there was a positive correlation between car rejection and person rejection times across the two tasks ($r = .34$, $p < .001$; see Fig. 5a). This indicates the presence of common features such as clutter that slow down target rejection in both tasks.

The above correlations raise the possibility that computational models may be capturing task-independent scene features rather than task-specific ones. To test this possibility, we took the best model for each condition and recalculated its partial correlation on the detection times after accounting for task-independent factors as well as low-level factors. For example, in the case of the best model for person detection on 225 common scenes containing both cars and people, we regressed out the observed as well as predicted car detection times on the same set of scenes, as well as low-level factors such as size of the largest target, its eccentricity, number of targets and measures of clutter (see Section S2 for detailed description of low-level features). This yielded a positive correlation ($r = .40$, $p < .001$; see Fig. 5b). Likewise, the best model for person rejection remained significant after regressing out observed and predicted car rejection times and low-level factors ($r = .27$, $p < .001$; see Fig. 5b) across 650 common scenes.

We performed similar analyses on the performance of computational models for car detection and rejection. The best model for car detection yielded a positive partial correlation after accounting for person detection times and low-level factors ($r = .43$, $p < .001$; see Fig. 5c). Likewise, the best model for car rejection yielded a positive correlation after factoring out predicted as well as observed person rejection times and low-level factors ($r = .35$, $p < .001$; see Fig. 5c).

The drop in best model performance in the above analyses indicates that the best models contained contributions from

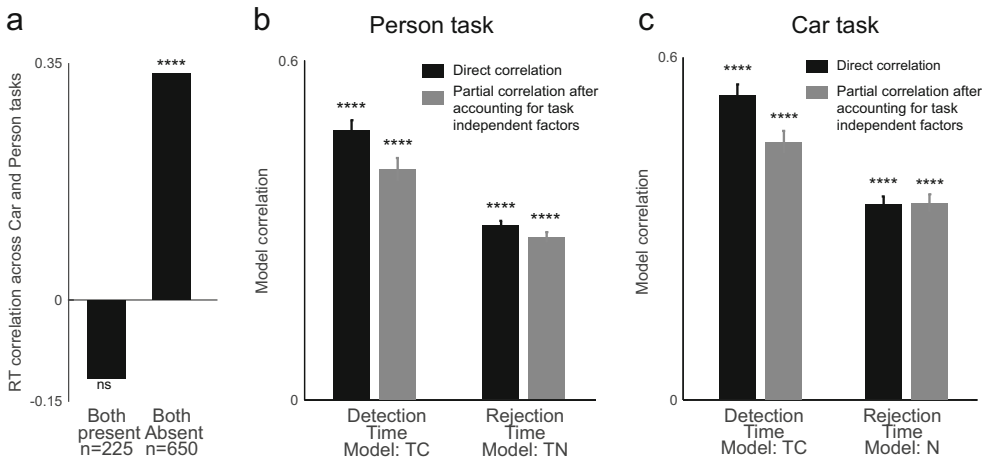

Fig. 5 Task-specificity of computational models. **a** *Left*: Correlation between person detection times (Experiment 1) and car detection times (Experiment 2) across the 225 scenes containing both people and cars. *Right:* Correlation between person rejection times (Experiment 1) and car rejection times (Experiment 2) across the 650 target-absent scenes that contained neither cars nor people. **b** Partial correlation analysis for person task: *Left:* Best model performance on person detection with observed detection times (*black*) is reduced but not abolished after factoring out dependence on car detection times and low-level factors such as size and eccentricity of the largest target, number of targets and clutter (*gray*). *Right:* Best model performance on person rejection (*black*) is reduced but not abolished after factoring out dependence on car rejection times and low-level factors (*gray*). **c, d** Analogous plots for the car task

task-independent features and low-level factors. However, the drop is relatively minor, suggesting that target detection and rejection in both tasks depend largely on task-specific features.

Validity of parceling scene information into three feature channels

The approach described in the preceding sections involve modeling human performance during rapid object detection using features derived from three distinct information channels. But what evidence do we have that there are three distinct channels? We addressed this question computationally by asking whether models trained on the entire scene perform better than models that use features segregating into the three channels. To this end, we selected a popular convolutional neural network (CNN) implementation (Krizhevsky et al., 2012), projected the features of its most informative layer (FC7) along their principal components such that it had the same degrees of freedom as the best model for each of the four conditions (detection/rejection × person/car). This model yielded comparable fits to the data (correlation with detection and rejection times: $r = .38$ & $.37$ for person task, $r = .05$ & $.41$ for car task, $p < .001$). In comparison, our models using best subsets of target, nontarget, and coarse scene information showed similar fits ($r = .45$ & $.33$ for person detection and rejection; $r = .57$ & $.34$ for car detection and rejection; $p < .001$). We conclude that parceling scene information into the three distinct channels results in similar performance on object detection as using the intact scenes.

Incidentally, our general approach can be used to elucidate how targets, nontargets, and coarse scene features explain the performance of deep neural networks as well, and we have included this analysis in the Supplementary Material (see Section S4).

Assessment of data set bias

The relative contributions of target, nontarget, and coarse scene features obtained in our results could potentially be specific to the scenes used in the study. For instance, our finding that nontargets do not contribute to car or person detection might be a consequence of the fact that they were not predictive of target presence in the scenes used here. To assess this possibility we estimated the degree to which each channel could predict the presence or absence of a target across scenes. As before, we calculated the five-fold cross-validated accuracy of linear classifiers trained on the features obtained from each channel. Target features were the most informative as expected (average accuracy: 93% for detecting cars; 92% for detecting people). Importantly, adding nontarget or context features did not improve performance (average accuracy: targets + nontargets: 93% for car, 92% for person; targets + coarse scene: 93% for car, 91% for people). Classifiers trained only on nontargets and coarse scene features also performed well above chance (average accuracy: 72% & 71%, respectively, for detecting cars; 73% & 73%, respectively, for detecting people), and combining them yielded a slight improvement (accuracy for nontargets + context: 77% for car, 81% for person). Classifiers trained on all three channels together were only marginally better than classifiers trained on only target features (average accuracy: 93% for car, 94% for person). Thus, at least for the scenes in our study, target features alone yielded the best possible performance. By contrast, our

analyses reveal that humans are using distinct additional feature channels for target detection and rejection.

Performance of state-of-the-art CNNs

Even though deep convolutional networks have obtained 90% to 95% categorization accuracies on top-five results on natural scenes (Russakovsky et al., 2015), the actual performance on the best result (or top-one) classification is much lower (74% on detecting people, 87% for cars) using a popular convolutional network (Ren, He, Girshick, & Sun, 2016). To investigate this specifically on our image set, we evaluated the performance of an AlexNet architecture fine-tuned for 20-way object classification (Lapuschkin, Binder, Montavon, Muller, & Samek, 2016) on the PASCAL object dataset (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010). This deep network had an accuracy of 73% for car classification and 83% for person classification. By contrast, using the most confident car or person detections using HOG models resulted in much higher accuracies (93% accuracy on car, 92% on person). We therefore used HOG-based features for evaluating target features in this study.

General discussion

Here we have shown that human performance on person and car detection in real-world scenes contain systematic scene-by-scene variations that can be explained using distinct contributions from target, nontarget, and coarse scene features. While these three types of information are known to influence object detection, their relative contributions on a scene-by-scene basis have not been previously established.

Specifically, our main findings are as follows: (1) Target-present responses in both car and person tasks are driven mainly by target features and to a smaller degree by coarse scene features but were not influenced by nontarget objects; (2) target-absent responses in the person task were predicted by target and nontarget features but not by coarse scene features; (3) target-absent responses in the car task were predicted primarily by the presence of nontarget objects; (4) features that speed up target detection slow down rejection, and vice-versa; and (5) human performance in both tasks is influenced largely by task-specific features and to a smaller degree by task-independent features. Below we review our findings in the context of the existing literature.

Our main finding is that human performance during rapid object detection can be understood by separating scene features into distinct channels comprising target, nontarget, and coarse scene features. Importantly, separating features into distinct channels did not adversely affect the ability of models to predict human performance, thus validating this separation. While specific types of target features (Harel & Bentin, 2009: Ullman et al., 2002), similar nontargets (Zelinsky, Peng, & Samaras, 2013b), co-occurring nontargets (Auckland et al., 2007) and coarse scene features (Ehinger et al., 2009; Oliva & Torralba, 2008) can affect object detection, in this study we have quantified their relative contributions. Our results also represent a general approach to determine the influence of targets, nontargets, and coarse scene features in any task.

Our approach resembles that of Ehinger et al. (2009), who separately assessed the contributions of targets and coarse scene features towards the distribution of eye fixations. Although our approach also includes target and coarse scene features, our goal and experimental methodology are different in several ways. First, we sought to explain object detection in briefly flashed scenes that precluded shifts in eye position or attention, in contrast to the Ehinger et al. (2009) task, which allowed free viewing until the target was detected. The difference in viewing duration alone and/or the response modality (hand vs. eye) could potentially drive different feature channels. Secondly, we used more sophisticated target features, rigorously separated the contributions of targets, nontargets, and coarse scene features. Finally, we used two detection tasks (people, cars) to establish both task-specific and task-general contributions of each feature channel.

Our finding that different channels contribute towards target detection and rejection offers possible insight into how humans perform this task. Our analysis of data set bias shows that in principle, object detection could be performed entirely using target features without recourse to nontargets or coarse scene features. Yet humans show evidence of using nontarget and coarse context features, suggesting that their feature representations may be optimized for real-world scene statistics that are potentially different from the scenes used in our study. However, the fact that target features alone suffice for object detection in computational models raises the important question of whether different features contribute to target detection and rejection. According to widely accepted models of decision making, evidence in favor of a response accumulates towards a threshold and upon reaching it, triggers a ballistic motor response (Kiani, Corthell, & Shadlen, 2014; Schall, Purcell, Heitz, Logan, & Palmeri, 2011). In principle, a single accumulator could produce two responses by triggering one response when evidence reaches the threshold and an opposite response when it does not reach threshold within a prescribed time limit. This is unlikely for several reasons: First, rejection times are systematically modulated by scenes, which rules out a fixed time limit predicted by a single accumulator. Second, the contribution of different channels towards detection and rejection suggests two separate accumulators for rejection and detection. Third, at least in person detection, target features contribute to both detection and rejection but have opposite effects, again suggesting two accumulators driven by the same evidence in opposite directions.

We found that the target detection in both tasks is driven mainly by target features with a relatively small contribution

of context but not nontarget features. The fact that the specific target features used in our modeling—namely, HOG features with deformable parts—predict human performance indicates that the underlying features are similar or at least partially correlated (see Fig. 2, Figs. S1–S2). However the model consisting of HOG features and coarse scene features did not fully explain the observed data, which could indicate features not captured by either channel (Delorme et al., 2010; Evans & Treisman, 2005; Reeder & Peelen, 2013; Ullman et al., 2002). We have also found that target features influence rejection times in the person task, suggesting that partial matches to target features slow down rejection. This is consistent with visual search studies where search becomes difficult even when a few distractors are similar to the target (Duncan & Humphreys, 1989; Jacob & Hochstein 2010; Motter & Holsapple, 2007).

Our finding that coarse scene features contribute only weakly to object detection is at odds with the many demonstrations of contextual influences in the literature. For instance, context strongly facilitates search by narrowing down possible target locations (Castelhano & Heaven, 2010; Malcolm et al., 2014; Torralba et al., 2006) and is a major determinant of eye fixations made during searching (Ehinger et al., 2009). Likewise, search in natural scenes is more efficient than expected given the number of distractors (Wolfe et al., 2011) and is facilitated by familiar spatial configurations of objects (Kaiser, Stein, & Peelen, 2014). These studies differ from ours in that scenes remained visible for a relatively long time, allowing scene context (or nontargets) to guide eye movements. Thus it is possible that context only weakly facilitates object detection at first (as in our study) but plays an increasingly important role as it builds up in time across multiple fixations (as reported by others). Whether contextual influences build up with viewing time or across multiple fixations is an important issue that is beyond the scope of this study. Nonetheless the fact remains that subjects are highly accurate in finding objects even in briefly flashed scenes and we have shown that this behaviour is driven by specific types of features.

We have shown that nontargets can have task-independent as well as task-specific influences (see Fig. S3). When nontargets slow down target-absent responses in both tasks, this can be interpreted as salient objects that divert attention and slow down responses. However the finding that a nontarget object can affect one task but not the other could be because of semantic associations with the target (Auckland et al., 2007) or due to biasing of target-related features throughout the visual field. Our results place important constraints on the degree to which nontargets guide object detection; for instance, we have shown that they are the dominant influence in car rejection but contribute equally to person rejection as target features.

Finally, the predictions of individual channels offer additional insights into the features underlying object detection in real-world scenes. First, for people but not cars, target-absent responses are longer when there are partial matches to targets in the scenes (i.e., false alarms in the HOG model; see Fig. 3c). This may be because people are easily confused with other upright vertical objects, yielding many more false alarms. False alarms for people in our computational models occurred for objects such as tree trunks, poles, and other upright objects matching in scale and shape with images of people. On the other hand, cars are very reliably detected (Fig. 2) and yield very few false alarms. Thus, the lack of modulation of car-absent responses by target features could be because cars, being rigid objects with distinctive parts (e.g., wheels), are relatively easy to detect. Car false alarms largely seem to arise from box-like structures such as entrances and building facades that could easily be discounted by our visual system (see Fig. S2).

Conclusions

To summarize, our results show that humans have systematic variation in object detection performance across real-world scenes that can be predicted using computational modeling. Our results provide a quantitative framework to study the contribution of target, nontarget, and coarse scene features in real-world object detection.

Acknowledgments This work was funded through the ITPAR collaborative grant (to S.P.A. & M.V.P.) from the Department of Science and Technology, Government of India and the Province of Trento. H.K. was supported through a postdoctoral fellowship from the DST Cognitive Science Research Initiative.

References

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review, 14*(2), 332–337.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5,* 617–629.

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception, 25*(3), 343–352.

Barenholtz, E. (2013). Quantifying the role of context in visual object recognition. *Visual Cognition, 22*(1), 30–56. doi:10.1080/13506285.2013.865694

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception & Psychophysics, 72*(5), 1283–1297.

Choi, M. J., Torralba, A., & Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognition Letters, 33*(7), 853–862. doi:10.1016/j.patrec.2011.12.004

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the 2005 I.E. Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005*, I. (pp.886–893). Available at: http://ieeexplore.ieee.org/document/1467232/?reload=true

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2010). Key visual features for rapid categorization of animals in natural scenes. *Frontiers in Psychology*, *1*(JUN), 21.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433–458.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*(6/7), 945–978.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology. Human Perception and Performance, 31*(6), 1476–1492.

Everingham, M., Ali Eslami, S. M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2014). The Pascal Visual Object Classes challenge: A retrospective. *International Journal of Computer Vision, 111*(1), 98–136.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology*, 2(OCT), 243.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminative trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1627–1645.

Harel, A., & Bentin, S. (2009). Stimulus type, level of categorization, and spatial-frequencies utilization: Implications for perceptual categorization hierarchies. *Journal of Experimental Psychology: Human Perception and Performance, 35*(4), 1264–1273. www.ncbi.nlm.nih.gov/pubmed/19653764

Jacob, M., & Hochstein, S. (2010). Graded recognition as a function of the number of target fixations. *Vision Research, 1,* 107–117.

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision, 8*(13), 1–18. Retrieved from http://eprints.gla.ac.uk/32899/

Joubert, O. R., Rousselet, G. A., Fabre-Thorpe, M., & Fize, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision, 9*(1), 2.1–16.

Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences, 111*(30), 1217–11222.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329–1342.

Krizhevsky, A., Sulskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Proceedings of the Advances in Neural Information and Processing Systems (NIPS)* (Vol. 25, pp. 1–9). Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012

Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., & Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. *2016 I.E. Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 17). Retrieved from http://iphome.hhi.de/samek/pdf/LapCVPR16.pdf

Lewis, M. B., & Edmonds, A. J. (2003). Face detection: Mapping human performance. *Perception, 32*(8), 903–920.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596–601. Retrieved from http://www.pnas.org/content/99/14/9596.full

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8693* LNCS(PART 5), 740–755. doi:10.1007/978-3-319-10602-1_48.

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: Strategic and incremental information accumulation for scene categorization. *Psychological Science, 25*(5), 1087–1097.

Mohan, K., & Arun, S. P. (2012). Similarity relations in visual search predict rapid visual categorization. *Journal of Vision, 12,* 19–19.

Morrison, D. J., & Schyns, P. G. (2001). Usage of spatial scales for the categorization of faces, objects, and scenes. *Psychonomic Bulletin & Review, 8*(3), 454–469.

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

Neider, M. B., & Zelinsky, G. J. (2011). Cutting through the clutter: Searching for targets in evolving complex scenes. *Journal of Vision, 11*(14), 1–16.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Oliva, A., & Torralba, A. (2008). Building the gist of a scene: The role of global image features in recognition. In *Progress in Brain Research* (pp. 23–39). doi: 10.1016/S0079-6123(06)55002-2

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature, 460*(7251), 94–97.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Reeder, R. R., & Peelen, M. V. (2013). The contents of the search template for category-level search in natural scenes. *Journal of Vision, 13*(3), 1–13.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *Poster presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence, PP*(99). Retrieved from https://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S.,…Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. doi:10.1007/s11263-015-0816-y

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1/3), 157–173.

Schall, J. D., Purcell, B. A., Heitz, R. P., Logan, G. D., & Palmeri, T. J. (2011). Neural mechanisms of saccade target selection: Gated accumulator model of the visual-motor cascade. *European Journal of Neuroscience, 33*(11), 1991–2002.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104*(5), 6424–6429. doi:10.1073/pnas.0700622104

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision, 53*(2), 169–191.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). 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.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.

Vighneshvel, T., & Arun, S. P. (2013). Does linear separability really matter? Complex visual search is explained by simple search. *Journal of Vision, 13,* 10. Retrieved from http://www.journalofvision.org/content/13/11/10.short

Walther, D. B., & Fei-Fei, L. (2007). Task-set switching with natural scenes: Measuring the cost of deploying top-down attention. *Journal of Vision, 7*(11), 9.1–12.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception & Psychophysics, 73*(6), 1650–1671. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3153571&tool=pmcentrez&rendertype=abstract

Zelinsky, G. J., Peng, Y., & Samaras, D. (2013a). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision, 13*(14), 1–13.

Zelinsky, G. J., Peng, Y., & Samaras, D. (2013b). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision, 13*(14). Retrieved from http://www.journalofvision.org/content/13/14/10.abstract?ct

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene CNNs. *Arxiv, 12*. Retrieved from http://arxiv.org/abs/1412.6856

Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research, 50*(20), 2062–2068.