VISUAL COGNITION, 2012, iFirst, 1–23

Psychology Press
Taylor & Francis Group

Influence of the amount of context learned for improving object classification when simultaneously learning object and contextual cues

Sophie Marat and Laurent Itti

Department of Computer Science, University of Southern California, Los Angeles, CA, USA

Humans use visual context to improve object recognition. Yet, many machine vision algorithms still focus on local object features, discarding surrounding features as unwanted clutter. Here we study the impact of learning contextual cues while training an object classifier. In a new image database with 10 object categories and 28,800 images, objects were presented in contextual or uniform backgrounds. Both the fraction of contextual backgrounds during training and the spatial extent of context were analysed. Local object features and broader context features were extracted by two biologically inspired algorithms, previously used for object and scene classification, respectively: HMAX, applied to a tight window around every object, and a "Gist" algorithm, applied to a larger yet still localized window. The descriptors from both algorithms were combined and processed by a Support Vector Machine. The recognition rate increased from 29%, without contextual cues, to 43% for objects presented in their context.

Keywords: Amount and size of context; Biologically inspired algorithms; Learning; Local context; Object classification.

Every day we use contextual cues to more efficiently find objects like our keys next to our wallet on our table. A growing number of studies of human perception (visual cognition and cognitive neuroscience) demonstrate the

Please address all correspondence to Sophie Marat, Department of Computer Science, University of Southern California, Hedco Neuroscience Building, Room 10, 3641 Watt Way, Los Angeles, CA 90089, USA. E-mail: sophie.marat.ilab@gmail.com

Supported by the National Science Foundation (grant number BCS-0827764), and the Army Research Office (W911NF-08-1-0360 and W911NF-11-1-0046), the Defense Advanced Research Projects Agency (HR0011-10-C-0034), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

http://www.psypress.com/viscog http://dx.doi.org/10.1080/13506285.2012.667007

useful relationships between the objects present in a visual scene and the context in which they are present. It is easier to find and recognize objects that are embedded in a consistent scene rather than in an inconsistent one (Biederman, 1981; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008). Conversely, scene recognition can also be facilitated by the presence of consistent objects (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Mack & Palermi, 2010). Even without scene context, presenting an isolated object among a set of other isolated but contextually related objects improves object recognition (Auckland, Cave, & Donnelly, 2007). It was also shown that context can improve reading efficiency (Bicknell & Levy, 2012) and that without context (meaningless sentences) readers have different viewing strategy compared to normal text (Schad & Engbert, 2012). Neurologically, fMRI studies have shown the use of contextual cues by humans (Bar, 2004).

Contextual information can be provided by different sources. Biederman (1981) suggested that no more than five classes of relations are needed to characterize the organization between an object and its setting in real-world scenes: Support (objects do not float in air), interposition (objects occlude their background), probability (objects tend to be found in some particular context, for example, a car on a road and not in the sky), position (objects tend to occupy specific positions in a congruent scene, e.g., cars lie mostly in the lower part of images and birds in the upper part), and finally size (objects have a restricted range of sizes compared to others, e.g., a mug is smaller than a chair). From these relations we can derive different kinds of contextual information that can be used and combined in different ways, and which raised the main questions to be answered: How to combine context with object classification? What kind of context to consider? What spatial extent of context to use?

Historically, in computer vision, objects to be classified were presented as isolated from their context, which was considered as clutter. Recently, more and more studies have investigated context as an interaction, and how context can improve object recognition. Context can be used in pre-processing to restrict the region to analyse in an image, reducing false alarms and processing time. When context refers to global information of the scene, e.g., "Gist", it can provide an oracle for the type of objects to expect in the scene and guide the search to their most probable locations and scales before the actual classification step (Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Torralba, 2003; Torralba, Oliva, Castelhano, & Henderson, 2006; Torralba & Sinha, 2001). Gist can also be used to improve a local feature classifier by providing a prior on the most probable location of the object that is combined with the object features in a probabilistic framework (location priming; Chikkerur, Serre, Tan, & Poggio, 2010; Torralba, Murphy, & Freeman, 2004, 2010). In Heitz and Koller (2008), contextual cues (such as road, tree, etc.) are used to modulate the output of an object detector in a probabilistic framework using a set of hand-labelled relation candidates to

judge the congruence of the object and the context. A car identified on a road would be more probable than one in the sky. Context can also be processed in terms of interactions between the different objects in a scene. In 2007, Rabinovich, Vedaldi, Galleguillos, Wiewiora, and Belongie leveraged semantic context among objects in a postprocessing step that influences the raw score of an object detector. Ambiguous outputs of the detector that are not congruent with the context given by other object detection will be changed to a more suitable object. For example, a "lemon" recognized next to a tennis racket and a person will be corrected as a "tennis ball". In all these studies contextual information is a high-level semantic knowledge (classification of the scene category, semantic background segmentation and a set of hard-coded possible relations between object and context, object label, and their set of relations) that may require complex processing (texture and object recognition).

Several studies have investigated which contextual cues to use and their possible combination. Perko and Leonardis (2010) compared the contribution of object cooccurrence, geometric cues (expressed as ground, vertical structure and sky presence maps), texture cues (anisotropy, polarity, and texture contrast maps) and their combination. They found that texture was better than geometry and object cooccurrence and that combining all the cues outperformed each single one. In 2009, Divvala, Hoeim, Hays, Efros, and Herbert used scene Gist, three dimensional (3-D) geometric context, semantic context, photogrammetric context to compute a scene context cue (object presence, location, and size information) and a spatial support cue (better bounding box and shape estimation given its presence, location, and size in the image). They compared a given object detector to that detector enhanced with scene context or with scene context and spatial support and also quantified the impact of each context source independently. They concluded that context reduced the overall detection errors and that the remaining errors were more reasonable (confusion between similar classes, e.g., between bicycles and motorcycles). A study from Wolf and Bileschi (2006) compared contextual cues from low-level features (colour and texture descriptors) and high semantic level (segmentation of buildings, trees, roads, and skies). They showed that accurate context could be determined from low level features and that high level semantic context was superfluous. They also concluded that even if the context was useful for predicting the location of the objects, context only marginally impacted object detection when objects were clearly visible, which is in contradiction with all the previous studies. Even if most studies agreed that contextual cues help object recognition, the choice of the cues to consider is still a highly debated topic.

In addition to selecting the right contextual cues, the spatial extent of context to integrate has also been explored. In 2009, Blaschko and Lampert, both a local context considering a tight neighbourhood around the object

and a global context considering the whole image are combined with an object classifier in a bag-of-words framework. Their work showed that integration of local and global context improved object detection. Uijlings, Smeulders, and Scha (2009) proposed to investigate more finely the spatial extent of an object by varying the window size used for object and context analysis. The context was defined as a compact box randomly selected in the image and with no overlap with the object's bounding box. Their recommendations on the size of the bounding box to consider are that rigid objects were best recognized if the bounding box was tight around the object, nonrigid objects showed no significant influence of contextual informations, and finally objects that were defined by their function (e.g., chair, horse, boat, etc.) benefited from more spatial context. They were also interested in the influence of localization accuracy of the bounding box around the detected object, showing that precise location improved recognition. The question of the spatial extent of context has instigated few studies and would benefit from more investigation.

To build further on previous work, we propose a new database where objects were photographed while presented on top of a contextual or a uniform background. This database allows us to investigate parametrically the influence of the spatial extent of the contextual window on the proposed model for different testing conditions. This database also enables us to raise the question of the influence of the fraction of images with contextual background in the training data. Previous studies have compared object detection with or without context, but, to the best of our knowledge, none has investigated the fraction of contextual cues that should be learned. What happens when learning a mix of isolated objects and objects presented in their congruent context? This situation may occur as some available database contains objects isolated and another contains objects embedded in congruent context. Finally, the database allows us to evaluate the influence of context in the classification error patterns for the different object classes. This allows us to derive new conclusions and guidelines for how congruent context can be exploited while training and testing object recognition algorithms.

The next section of this paper presents the algorithms used for extracting the local object features, the context features, and how to combine them into a context-enhanced object classifier. Then the database used for training and testing the algorithms is described. The role of context is then analysed in the evaluation section. Finally, the findings are discussed before the conclusion.

INCORPORATING CONTEXT IN OBJECT CLASSIFIER

Our proposed model extracts features both specifically for object classification and for contextual cues; these features are then combined and learned

together (see Figure 1). For this purpose, two biologically inspired algorithms were chosen: HMAX for object detection, and Gist for contextual cues.

Features for object classification

The features used for object detection are extracted using HMAX, a state-of-the-art biologically inspired object detector available online (Riesenhuber & Poggio, 1999; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). HMAX is based on an architecture inspired by current theory of the feed-forward path of object classification in the ventral stream of the primate visual cortex. It is a hierarchical system that builds an increasingly complex feature representation via alternating template matching (using Gaussian-like tuning) and maximum pooling operations, achieving a trade-off between selectivity and invariance. The result of HMAX is a feature vector that represents the strongest features of an object and can be used to train a classifier, e.g., a Support Vector Machine (SVM), to recognize the object's category.

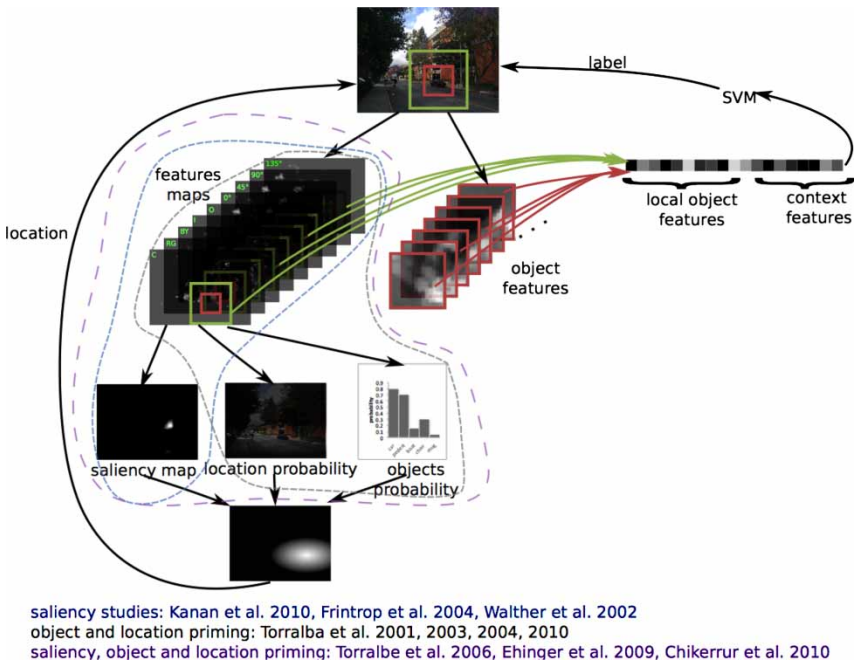

Figure 1. Schema of proposed model for object classification, and previous studies. Saliency studies: Frintop, Nutcher, Surmann, and Hetzberg (2004), Kanan and Cottrell (2010), Walther, Itti, Riesenhuber, Poggio, and Koch, C. (2002); object and location priming: Torralba (2003), Torralba et al. (2001, 2004, 2010); saliency, object, and location priming: Chikkerur et al. (2010), Ehringer et al. (2009), Torralba et al. (2006). To view this figure in colour, please see the online issue of the Journal.

HMAX was chosen for its ability to generalize over different instances of objects within a class. It is more robust to intraclass variation than other approaches such as SIFT (Lowe, 2004) which are limited to recognizing specific object instances. HMAX was also shown to outperform several state-of-the-art algorithms (e.g., SIFT, histogram of gradient, etc.) on standard databases (Serre et al., 2007). In this study we use the HMAX as our object detector, but the proposed approach is general and could be applied just as well to other object detection algorithms such as a histogram of gradient, deformable part model, bag-of-word, etc.

Features for contextual cues

The contextual cues are provided by a Gist algorithm. The Gist is computed using a biologically plausible algorithm available online (Siagian & Itti, 2007). The approach is to exploit statistical summaries of colour and texture measurements in predetermined image subdivisions, as different types of scenes may exhibit large differences in these distributions (e.g., the distribution of colours in forest vs. city scenes). The algorithm uses a multiscale set of early visual features (intensity, orientation, and colour) usually used to compute saliency maps (Itti, Koch, & Niebur, 1998). For each of these feature maps, a Gist vector is extracted by averaging feature responses over fixed subregions of that feature map.

Gist algorithms have been previously used as a source of contextual information (Dalal & Triggs, 2005; Torralba, 2003; Torralba et al., 2004, 2010). In these papers the Gist is used to categorize to whole scene from which the object "probability" and "position" as named by Biederman (1981) are inferred.

Our purpose here is to extract more local contextual information, spatially confined to the neighbourhood of an object (see Figure 1). This enables us to gather "probability" information from the near context of an object, e.g., a car is usually found on a road independently of the road being close to a lake, a desert, or a forest. Furthermore, the database we used contains shots from an aerial viewpoint; no "position" information could be extracted from the knowledge of the scene category.

Combining object and context

Here we consider images where objects are centred within the frame and fit within a bounding box. We note that, for more general use, the object classifier proposed could be coupled with an attentional processing step that will be able to select pertinent regions in an image for object detection (using the saliency map generated in the computation of the contextual cues). Indeed, it has been shown (Elazary & Itti, 2010) that interesting objects are

salient; therefore, attentional processing based on saliency can be used to highlight candidate regions of interest for objects, removing the need for a brute force approach such as sliding-window (Frintrop et al., 2004; Kanan & Cottrell, 2010; Walther et al., 2002).

HMAX extracts local object features within its bounding box (256×256 pixels, this choice is discussed later in the Results), represented in red in Figure 2 at the centre of the frame. Meanwhile, the Gist algorithm extracts contextual cues in the object's surrounding in a larger window. Two different sizes of windows are considered to evaluate the contextual influence: A small one (512×512 pixels, in green in Figure 2) and a big one (720×720 pixels, in blue in Figure 2).

The two feature vectors generated by the two algorithms are then concatenated to form a single feature vector gathering the object and contextual information (see Figure 1). An SVM classifier then processes this vector. Thus, information about the object and the context are learned together while training the classifier. Therefore, object classification is done in a common step without requiring pre- or postprocessing of the object features by the context features. The SVM classifier used in this paper is a multiclass classifier with radial basis functions, which gives, for each feature vector, the probability that the corresponding object belongs to each category of objects (Chang & Lin, 2001).

IMAGE DATABASE

Many image databases have been proposed to evaluate object detectors and are available online (Caltech 101, Caltech 256, LabelMe, ALOI, PASCAL VOC, etc.). These databases present different numbers of object categories,

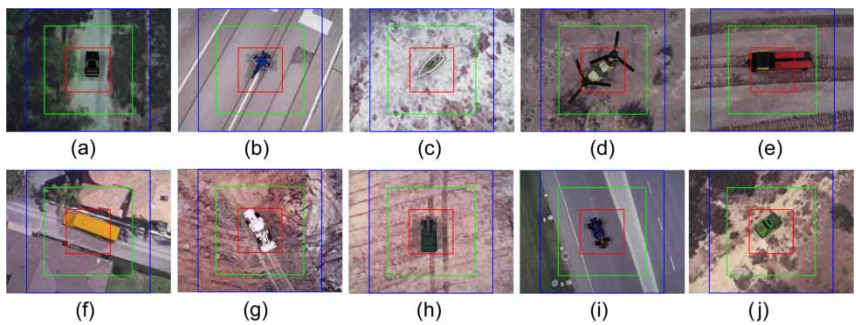

Figure 2. Example of bounding boxes for objects (red) and the context (green and blue), on samples object of the database: (a) Car, (b) plane, (c) boat, (d) helicopter, (e) train, (f) bus, (g) equipment, (h) tank, (i) Formula 1 car, (j) military vehicle. To view this figure in colour, please see the online issue of the Journal.

ranges of variation in appearance inside the same category, object scale, and also different background conditions (isolated objects or in clutter). However, the size of the area surrounding of the objects can be highly variable and sometimes too narrow, especially when the image contains a close view of the object. Moreover, in each of these databases, objects are presented in isolation or in context, but none of them presents the same object categories in different context conditions (isolated and present in congruent, neutral, or incongruent contextual scenes).

We thus introduce a new database (available online[1]) of 28,800 images where objects are presented with both a consistent context and a uniform background. This enables us to investigate learning with different proportions of images with contextual information or with uniform background. The database is composed of 10 different object classes: Car, plane, boat, helicopter, train, bus, equipment, tank, Formula 1 car, and military vehicle (see Figure 2). Each class contains 10 different items. To stay as close as possible to real-world conditions while retaining control of context information, the database is composed of top and near top views of Micro-Machines vehicle models on printed aerial maps that correspond to the scale of the models. Each item is present at the centre of the frame, which correspond to the centre of the printed map, with 12 different backgrounds: Six uniform backgrounds of different colours and six backgrounds with contextually congruent information. These six uniform backgrounds are the same for all the different classes, and so cannot give any information regarding the class of object present. The contextual backgrounds are the same within a class (same printed maps), but different between different classes. For each of these conditions, 24 views (960×720 pixels) are taken, comprised of three different camera positions (one top and two near-top views) and eight rotations of the scene. The contexts for the nonuniform backgrounds are defined as follows:

1. Car: On road (urban and countryside)
2. Plane: On runway, tarmac
3. Boat: On the sea
4. Helicopter: Different terrain and runway
5. Train: On rail track
6. Bus: On road
7. Equipment: Construction equipment on trail in various terrain
8. Tank: On different trail and terrain
9. Formula 1 car: On circuit road
10. Military vehicles: On road in various terrain.

[1] www.ilab.usc.edu

Different classes could share a common context, e.g., car and bus on roads. In those cases, the pictures of the road backgrounds used for the different classes were different. There is no background picture overlap between the different classes.

For each class, the 10 different models are split into two groups: Those used for training the classifier (“learned items”) and those for which the classifier is not trained on (“new items”). Similarly, the 12 backgrounds are separated into the “learned backgrounds” group, which is used for training, and the “new backgrounds” group. Each set contains three uniform backgrounds and three contextually congruent backgrounds. This enables us to perform three different evaluations:

- The “all new set”: Both the items and the backgrounds are different than the ones learned for the training (“new items” and “new backgrounds”)
- The “learned item set”: Only the backgrounds are different from the training data, while the items are identical to ones used for the training (“learned items” and “new backgrounds”)
- The “learned background set”: Only the items are different; they are placed on the same backgrounds as used in the training (“new items” and “learned backgrounds”).

This setting could help us evaluate independently the contribution of learning the background and the items.

RESULTS

First we present some general remarks and the influence of the analysis window size on the different algorithms and their combination when training was done only on images presenting objects within their congruent context, as is usually the case in studies on context. Then, the influence of the amount of contextual exemplars learned as an aid to object classification is studied. Finally the influence of context on the classification error pattern is discussed.

Influence of the analysis window size for objects presented in their context

Tables 1 and 2 present the average, over the 10 classes, of the true positive detection rate across the different subsets of the database (“all new set”, “learned item set”, and “learned background set”). Training was done with objects presented with contextual background and testing was done with objects presented in contextual background (see Table 1) or in uniform

TABLE 1
Mean classification on images with contextual background

	HMAX 256	*HMAX 512*	*HMAX 720*	*Gist 512*	*Gist 720*	*HMAX 256 Gist 512*
All new set	39.36	33.97	31.75	42.72	31.86	42.58
Learned item set	87.16	84.27	80.50	64.63	51.77	79.83
Learned background set	46.50	44.22	46.30	80.61	87.25	68.11

Percentage of true positive with training images that contain only objects in contextual background for classification using different features on different portion of the dataset: "All new set" (new item on new background), "learned item set" (learned item on new background), "learned background set" (new item on learned background).

background (see Table 2). These tables compare the results of SVM classifiers trained on features provided by HMAX alone (HMAX k, where k is the size of the window used for HMAX features extraction), the Gist algorithm alone (Gist l, where l is the size of the window used for the Gist features extraction), or the combination of these two features vectors (HMAX k Gist l).

Evaluation when using only HMAX features. First, we note that the results on our testing data with contextual background using the "all new set" (see Table 1) are lower than reported in the literature (Serre et al., 2007). This could be explained by the challenging nature of our image database, specifically that it presents objects at eight different orientations. The classifier learned is therefore required to generalize to both the different items of each class and also for their diverse poses. In the testing data with contextual backgrounds in the "learned item set" the results are closer to the ones reported in Serre et al. (2007).

TABLE 2
Mean classification on images with uniform background

	HMAX 256	*HMAX 512*	*HMAX 720*	*Gist 512*	*Gist 720*	*HMAX 256 Gist 512*
All new set	34.00	31.97	29.91	37.94	27.19	36.11
Learned item set	75.50	75.50	72.75	47.63	35.61	62.08
Learned background set	34.14	29.55	27.86	34.94	24.88	39.38

Percentage of true positive with training images that contain only objects in contextual background for classification using different features on different portion of the dataset: "All new set" (new item on new background), "learned item set" (learned item on new background), "learned background set" (new item on learned background).

HMAX is used to extract object features within a narrow window around the object (256×256 pixels). Because HMAX extracts features in the whole window surrounding the object, it may also extract some contextual features from the surrounding background. To study this behaviour, HMAX features were computed on the same windows as the ones used for the Gist context (512×512 and 720×720 pixels) to evaluate the impact of the size of the neighbourhood on classification using only HMAX features. Increasing the window size for HMAX features extraction lowers classification performance when evaluation is done on objects atop a contextual background (see Table 1), as more differing contextual information is processed in the HMAX window. The "learned background set" shows closer performance independently of the window size. In this case, the backgrounds are the same for training and testing. Thus, learning the background is no more penalizing, but it is not improving either. The evaluation of objects in uniform backgrounds (see Table 2) shows less impact of the size of the analysis window than the evaluation with contextual backgrounds as there is no more context in testing. HMAX cannot simultaneously and efficiently retain information of the object and information of the context. When the object covers a smaller fraction of the analysis window, object features become less prevalent and classification accuracy may suffer.

Therefore, in accordance with its purpose, HMAX principally relies on object features and considers the context to a lesser extent. Therefore, the narrower window (256×256 pixels) is the best suited for HMAX.

Evaluation when using only Gist features. Although we aim to use Gist features as a contextual complement to local HMAX features, it is possible that these features alone may already be able to classify objects. Thus, here we explore classification accuracy when only using Gist features. Increasing the window size of the Gist analysis gives mixed results on classification accuracy. Table 1 shows that a larger analysis window lowers classification accuracy for the "all new set" and the "learned item set", but improves it for the "learned background set". Testing objects in a uniform background (see Table 2) shows that widening the analysis window penalizes the result for every testing condition. Extracting the Gist on an overly large window takes into account a wider context which improves performance when the context is similar to the one learned, but lowers performance when the context is different from the learned one, even for semantically congruent contexts. For example, a small analysis window will capture the road as context for a car, but a bigger window will also capture the divergent context such as a house or a forest near the road.

While testing on a uniform background (see Table 2), the best results were obtained for the "learned item set". Just as HMAX was unintentionally sensitive to context features within its window, the Gist algorithm is also

sensitive to object features since the object is present in the window used for the context. This also explains the fact that, in Table 1, the "learned item set" yields better results than the "all new set".

Although the Gist algorithm was not designed for object classification, it can still outperform HMAX, not only in the case of the "learned background set" but also for the "all new set" (see Table 1). Surprisingly, Gist can also outperform HMAX even when testing on uniform backgrounds (see Table 2). The coarse information given by the Gist features, extracted on a spatially limited neighbourhood of the object, appears to be sufficient to classify objects even when the background of the tested object is uniform. Absent any context information, the Gist features will be extracted only from the objects. The coarse signature of the object, extracted by the Gist algorithm, may be more robust to intraclass variation than HMAX. Moreover, unlike HMAX, the Gist considers colour information, which can explain the Gist performance. Some objects tend to have specific colours (boat, tank, helicopter, etc.), similar between training and testing, which may improve the Gist results. However, the Gist is less efficient in recognizing particular items than HMAX (as shown by the "learned item set" in Tables 1 and 2).

In conclusion, Gist features are helpful for integrating contextual information about the neighbourhood of the object. However, analysis windows that are too wide may include context information outside the proximal surrounding of the object, which may penalize classification when the context is not consistent at larger distances.

Evaluation when using HMAX and Gist features combined. HMAX is more suited to extracting object information and Gist is more adapted to extracting contextual information in a spatially limited neighbourhood. The combination of HMAX and Gist (HMAX 256 Gist 512) improves the performances by being more robust to the different testing condition and giving overall the best results (see Tables 1 and 2).

As proposed by Uijlings et al. (2009), we tested the robustness of the classifier to the localization of the object (see Table 3). As before, the patches to be classified are the same size as previously (256×256 pixels for HMAX and 512×512 pixels for the Gist), but instead of being centred on the object, the patches are shifted with an offset of 20 pixels on the right (7.8% of the HMAX window). The training data is the same as previously and not shifted. The offset is arbitrarily chosen to the right but would have lead to the same results if considering the left side as the different images contain different orientations of the scene, so a 20 pixels offset to the right for orientation 0° corresponds to a 20 pixels offset to the left for orientation 180°.

The performance of the combination of HMAX and Gist is slightly lower on the shifted test set compared to the nonshifted test set. The results of

TABLE 3
Mean classification on shifted images

	HMAX 256 Gist 512	
Object background	*Context shifted*	*Uniform shifted*
All new set	40.87	34.84
Learned item set	78.68	61.99
Learned background set	66.25	37.82

Percentage of true positive with training images that contain only objects not shifted in contextual background for classification of images with shifted objects on contextual or uniform backgrounds on different portion of the dataset: "All new set" (new item on new background), "learned item set" (learned item on new background), "learned background set" (new item on learned background).

HMAX 256 Gist 512 on the shifted test set are still overall better than for HMAX and Gist considered independently on the nonshifted test set. The combination of HMAX and Gist is robust to such a shift of the bounding box around the object.

Influence of the percentage of contextual exemplars considered

The amount of contextual cues is manipulated by varying, in the training process, the percentage of images that contain objects in their consistent context (called $p_{context}$ in the following) and in a uniform background. $p_{context}$ varies from 0% (all training images contain only objects in uniform background) to 100% (all training images contain only objects in congruent contexts). From the pool of training images, which by definition contain "learned items" on "learned backgrounds", a training set is randomly chosen which have the desired $p_{context}$. The results of classification presented in the following are averaged over 10 evaluations[2] with different randomizations of the training set with the desired $p_{context}$. As previously, testing is done on different subsets of the testing part of the database and with contextually congruent backgrounds (see Figure 3) and uniform backgrounds (see Figure 4).

Figures 3 and 4 present the results as the average, over the 10 classes, of the true positive detection rate (p_{hit}). This is calculated both as a function of the percentage of training images that contain objects in congruent context ($p_{context}$) and across different testing conditions.

[2] Note that for $p_{context} = 0\%$ and $p_{context} = 100\%$ the whole corresponding pool of training image is used resulting in 10 identical selections.

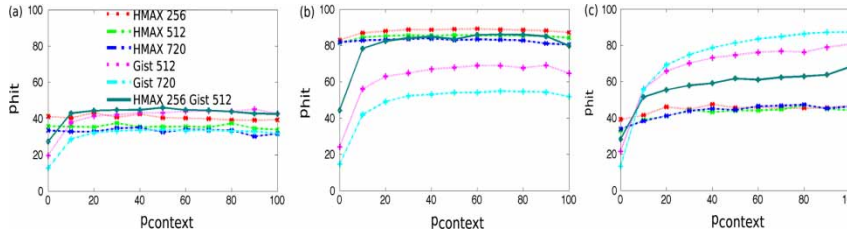

Figure 3. Mean classification on images with contextual background. Percentage of true positive (p_{hit}) as a function of the percentage of training images that contain object in contextual background ($p_{context}$) for classification using different features on different portion of the dataset: (a) "All new set" (new item on new background), (b) "learned item set" (learned item on new background), (c) "learned background set" (new item on learned background). To view this figure in colour, please see the online issue of the Journal.

Evaluation when using only HMAX features. The evaluation of the "all new set" of objects in contextual backgrounds (see Figure 3a) shows that HMAX performance varies with $p_{context}$. The difference in performance for HMAX (independently of the window size) for $p_{context} = 0\%$ and 100% is weakly significant, $F(1, 59) = 4.45$, $p = .039$. However, Figure 3(b) and (c) shows significant improvement with increased $p_{context}$, $F(1, 59) = 12.24$, $p < .001$, and $F(1, 59) = 366.11$, $p < .001$, respectively. In contrast, evaluating the classifier with objects in uniform context (see Figure 4) shows lower performance with increased $p_{context}$, $F(1, 59) = 507.06$, $p < .001$ for the "all new set", $F(1, 59) = 1719.9$, $p < .001$ for the "learned item set", and $F(1, 59) = 576.14$, $p < .001$ for the "learned background set". As said previously, HMAX considers also contextual features inside the analysis window, which makes it sensible to a small extent to $p_{context}$ variation in the training data.

Figure 3 and Figure 4 also show that the conclusion on the analysis window size presented in the previous section for $p_{context} = 100\%$ can be

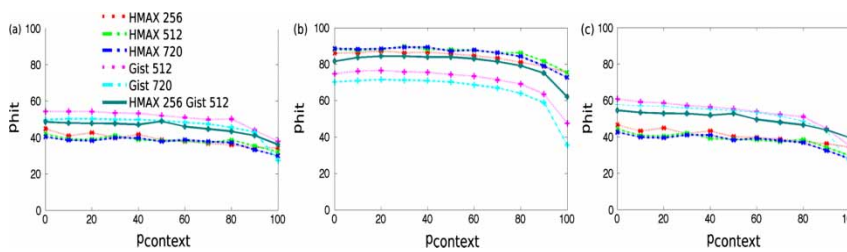

Figure 4. Mean classification on images with uniform background. Percentage of true positive (p_{hit}) as a function of the percentage of training images that contain object in contextual background ($p_{context}$) for classification using different features on different portion of the dataset: (a) "All new set" (new item on new background), (b) "learned item set" (learned item on new background), (c) "learned background set" (new item on learned background). To view this figure in colour, please see the online issue of the Journal.

extended to all the $p_{context}$. For the evaluation with objects presented in context Figure 3, HMAX 256 outperforms HMAX 512 and HMAX 720, $F(2, 329) = 881.43$, $p < .001$ for the "all new set", $F(2, 329) = 398.14$, $p < .001$ for the "learned item set", and $F(2, 329) = 11.11$, $p < .001$ for the "learned background set". HMAX 512 is significantly better than HMAX 720 when testing on images with contextual background for the "all new set" and the "learned item set", $F(1, 219) = 198.19$, $p < .001$, and $F(1, 219) = 222.48$, $p < .001$, respectively, but not for the "learned background set" or when testing with uniform background (see Figure 4), $F(1, 219) = 2.57$, $p = .11$, $F(1, 219) = 3.65$, $p = .057$), $F(1, 219) = 0.72$, $p = .39$, and $F(1, 219) = 1.32$, $p = .25$, respectively.

HMAX can be influenced by both the size of the window analysis and by $p_{context}$. However, considering the smallest window analysis (HMAX 256) enables it to be more robust to the variation of $p_{context}$.

Evaluation when using only Gist features. By design, Gist features are expected to be more sensitive to different $p_{context}$ than HMAX. When learning with an increasing $p_{context}$, classification accuracy on the testing set with contextual backgrounds is invariably improved (see Figure 3). Gist classifiers tested on the "all new set" with objects in contextual backgrounds benefit from higher $p_{context}$ (see Figure 3a). The difference in performance for Gist (independently of the window size) for $p_{context} = 0\%$ and 100% is significant, $F(1, 39) = 205.03$, $p < .001$. Notably, learning only 10% of images with objects in context is sufficient to boost classification compared to learning only isolated objects, $F(1, 39) = 163.55$, $p < 0.001$. The same conclusion can be reached from Figure 3(b) with the "learned item set", $F(1, 39) = 446.73$, $p < .001$ for the difference between $_{pcontext} = 0\%$ and 100%, and $F(1, 39) = 224.08$, $p < 0.001$ for the difference between $p_{context} = 0\%$ and 10%. The "learned background set" (see Figure 3c) shows a larger influence of $p_{context}$, $F(1, 39) = 2964.4$, $p < .001$ for the difference between $_{pcontext} = 0\%$ and 100%, and $F(1, 39) = 1503$, $p < .001$ for the difference between $p_{context} = 0\%$ and 10%. For the "learned background set", most of the increase in performance is still acquired with the first 10% of images with objects in context, $F(1, 39) = 1503$, $p < .001$ for the difference between $p_{context} = 0\%$ and 10%. However, increasing $p_{context}$ from 10% to 100% improves more object classification than for the "all new set", $F(1, 39) = 778.65$, $p < .001$ or the "learned item set", $F(1, 39) = 476.43$, $p < .001$. Since the contextual backgrounds considered are the same for training and testing, the more contextual backgrounds are included in the training, the better the performance.

Turning to testing objects in a uniform background (see Figure 4) shows that learning too much unused contextual information penalizes classification performance. The decrease in performance for Gist (independently of

the window size) for $p_{context} = 0\%$ and 100% is significant, $F(1, 39) = 209.52$, $p < .001$ for the "all new set", $F(1, 39) = 438.47$, $p < .001$ for the "learned item set", and $F(1, 39) = 587.99$, $p < .001$ for the "learned background set". Learning unused context lowers the performance, but the scores are still better than when the context is present but not learned.

The conclusions about the analysis window size can be extended to all $p_{context}$. Gist 512 outperforms Gist 720 for the "all new set" and "learned item set" (see Figure 3a and b), $F(1, 219) = 127.52$, $p < .001$, and $F(1, 219) = 68.70$, $p < .001$, respectively, and Gist 720 outperforms Gist 512 for the "learned background set" for $p_{context} > 10\%$ (see Figure 3c), $F(1, 199) = 26.36$, $p < .001$, even if they are not significantly different, when considering all the pc_{ontext}, $F(1, 219) = 3.54$, $p = .06$. For the evaluation on uniform background, Gist 512 still outperforms Gist 720 for the "all new set", $F(1, 219) = 26.26$, $p < .001$, and the "learned item set", $F(1, 219) = 18.42$, $p < .001$ (see Figure 4a and b), even if the difference between Gist 512 and Gist 720 is smaller than when testing on contextual background (see Figure 3). The difference is not significant for the "learned background", $F(1, 219) = 4.26$, $p = .04$.

It can also be noted that contextual cues influence more the results than the size of the analysis window. For the "all new set" with contextual background (see Figure 3a), considering $p_{context} = 10\%$ improves the results from 19.61 for Gist 512 and $p_{context} = 0\%$ to 37.93 and from 12.72 from Gist 720 and $p_{context} = 0\%$ to 28.67. But increasing the window size from 512 to 720 drop the mean performance over $p_{context}$ from 40.70 to 30.83.

Evaluation when using HMAX and Gist features combined. HMAX has been shown to be more robust to the variations of the amount of context learned (see Figure 3). Gist takes the maximum advantage of increasing $p_{context}$ when contextual backgrounds are also present in testing, but increasing $p_{context}$ can also penalize isolated object classification. Both the evaluation of HMAX features alone and Gist features alone have shown that using an appropriate size for the analysis window, e.g., not too big, is also important. The proposed concatenation of HMAX and Gist features is a good compromise of the two algorithms taking advantage of the context learned when it is helpful and not too much when the object is not in its usual context. The most important testing condition is the "all new set" with contextual background (see Figure 3), which represents a testing condition with contextual background and with testing and training that are not overlapping. In this condition the combination of HMAX and Gist (HMAX 256 Gist 512) is significantly better than HMAX alone (HMAX 256), $F(1, 219) = 17.18$, $p < .001$, and significantly better than Gist alone (Gist 512), $F(1, 219) = 5.57$, $p = .019$.

Influence of the context onto classification error patterns

To study the classification errors made by the proposed model and to see their evolution with $p_{context}$, we plot the confusion matrices that show the repartition of predictions for each class of objects for the "all new set". A confusion matrix represents the predicted classes as a function of the ground truth classes. A perfect classification algorithm corresponds to a diagonal matrix.

Figure 5 presents the confusion matrices for the classification using HMAX and Gist, HMAX alone and Gist alone with a $p_{context} = 0\%$. The first thing to be pointed out is that when the learning is done with all the objects on uniform backgrounds ($p_{context} = 0\%$), and the testing is done with objects placed on their natural context (see Figure 6a) classification is seriously impaired. This could be expected as Figure 3(a) shows a very low percentage of true positives for the average over all the classes. When $p_{context} = 0\%$, only the object features are learned, but on the testing data the classifier tries to identify both the object category and the context features, and so tries to classify context features using object ones. This could be seen in Figure 5(b) where using only HMAX (so mostly object features) shows a better classification for every class. The use of only the Gist information (considering more the context) gives more impaired results, as we can see in

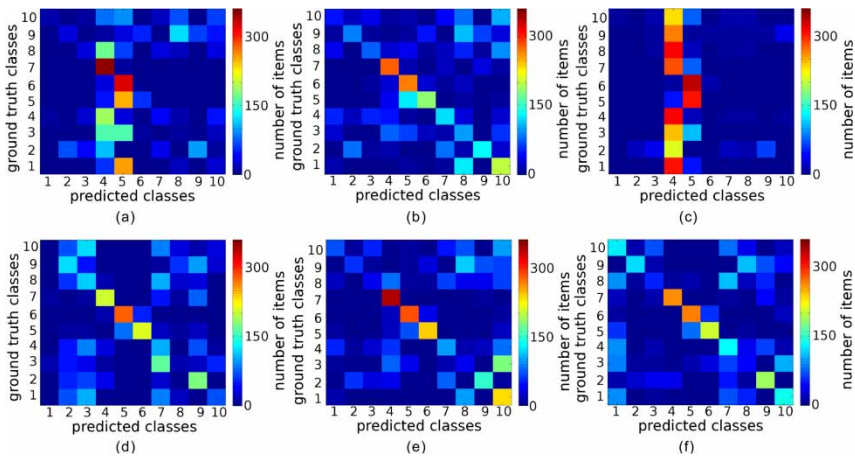

Figure 5. Confusion matrices for object classification in the "all new set" using different features and considering different $p_{context}$, all except (d) are evaluated on contextual background. (a) Using HMAX and Gist with $p_{context} = 0\%$, (b) using only HMAX with $p_{context} = 0\%$, (c) using only Gist with $p_{context} = 0\%$, (d) using HMAX and Gist with $p_{context} = 100\%$ evaluated on uniform background, (e) using HMAX and Gist with $p_{context} = 10\%$, (f) using HMAX and Gist with $p_{context} = 100\%$. The class numbers are given in the description of the different classes section. To view this figure in colour, please see the online issue of the Journal.

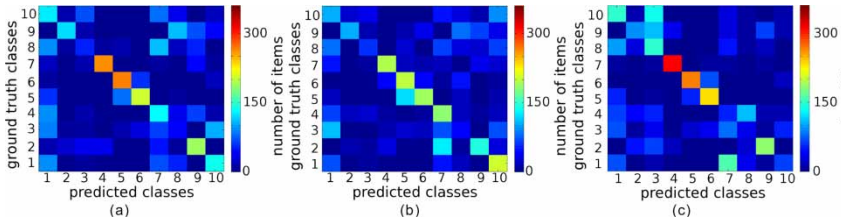

Figure 6. Confusion matrices with $p_{context}$ = 100% evaluated on context background with (a) HMAX and Gist, (b) HMAX alone, and (c) Gist alone. To view this figure in colour, please see the online issue of the Journal.

Figure 5(c) where the classifier predicts almost only helicopter (4) and train (5). Thus, it is not appropriate to try to classify objects that are presented in natural context while learning them only on uniform backgrounds.

At the opposite end from context-free training, Figure 5(d) presents the confusion matrix for the classification of objects on uniform backgrounds while learning objects in their natural context ($p_{context} = 100\%$). The results are less impaired than the ones for the classification of object in uniform background while training with $p_{context} = 0\%$ (see Figure 5a), $F(1, 199) = 4.44$, $p = .03$. More objects are well identified and there is no particular object bias as in the context-free case. Thus, learning objects in context is a better strategy, even while testing on uniform background, rather than training on uniform background and testing on contextual objects. We will now focus on the impact of $p_{context}$ learned while the testing objects are presented in their natural background for the "all new set". Figure 5(e) and (f) show the results for classification of objects presented in contextual background for $p_{context}$ of 10% and 100%, respectively. Classification performance is improved compared to Figure 5(a) with $p_{context} = 0\%$, $F(1, 199) = 12.09$, $p < .001$, and $F(1, 199) = 14.56$, $p < .001$, respectively. As seen on the average curve (see Figure 4a), learning 10% of object in congruent background is sufficient to get almost all the improvement due to contextual cues. The difference between $p_{context} = 10\%$ and $p_{context} = 100\%$ is not significant, $F(1, 199) = 0.01$, $p = .90$. When considering more context during learning, the results are stronger on the diagonal, which is a sign of better classification, the errors also seems to be less spread but higher for some particular object classes.

To understand better the confusion matrix for our model with $p_{context} = 100\%$, we can also look at the confusion matrices meeting the same testing condition (contextual background, "all new set", $p_{context} = 100\%$) for the HMAX features only (see Figure 6b) and the Gist features only (see Figure 6c). First we can see that, for our combined model, learning objects always embedded in context leads the classifier to misclassify objects as cars (1) more often (see Figure 6a). The fact that this bias is also present in both the

HMAX and the Gist results suggests that the bias arises from the objects and/or the conjunction of the object and its close context. While comparing HMAX and Gist results, it can be noticed that the HMAX confusion matrix has errors more spread over the different classes, whereas Gist usually has less sources of error which tend to be stronger than the HMAX errors. The Gist disambiguates classification of ambiguous objects with different context such as: Train vs. bus, tank vs. car, and Formula 1 vs. equipment. Sometimes, using only the Gist can also increase the confusion if the natural contexts share similar features such as: Equipment (7) and tank (8), equipment (7), and military vehicle (10), or, less intuitively, car (1) and boat (3). Using only HMAX, cars (1) can be confused as military vehicles (10) (and the opposite) as the objects share similar features.

The confusion matrices enable us to see that integrating contextual information can decrease the confusion of the classifier between different object classes that share similar object features. We also see that considering the context leads to fewer sources of error, but these errors tend to be stronger. Using context information, these errors are thus more coherent with the similarities of the different contexts.

DISCUSSION

In this study we have shown that learning context while learning to recognize objects improves object classification when the objects are presented in semantically congruent context. Like Uijlings et al. (2009), we found that the conclusion of Wolf and Bileschi (2006) did not generalize on our data. In their study, Wolf and Bileschi found only a marginal help from the context when the target object was unambiguously visible. They concluded that context might be a useful cue when the object appearance was "weak", e.g., low resolution or very noisy images. In our study, we found that contextual information helps in recognizing the different classes of objects presented in their context, even if they are unambiguously visible, and penalizes the classification of isolated objects. This is in accordance with the literature on human perception, where it has been shown that objects are classified more reliably and rapidly when presented in congruent context than incongruent context (Davenport & Potter, 2004; Joubert et al., 2008) or even when presented isolated or in meaningless context (Sun, Simon-Dack, Gordon, & Teder, 2011). This is also consistent with Bicknell and Levy (2012) findings. They concluded that context facilitates reading by allowing readers to reach a given confidence about the word presented more rapidly than for random words.

Using the "learned item set" and "learned background set" show that learning the same instances of image or context as the ones presented in

testing improved classification compared to learning different instances of the same semantic category. In the cognitive literature, similar conclusions have been drawn; presenting a brief preview of the same instance of a scene improved the performance of object detection compared to no scene or a contextually similar scene (Castelhano & Heaven, 2010). The improvement was conjectured to come from the preview of the same scene as the search scene that was guiding attention toward the most likely position of the object. When the identical instance of an object was presented prior to search, search to that particular object was facilitated (Castelhano & Heaven, 2010; Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004). Wolfe et al. (2004) concluded that the human visual system could be biased toward the features of the object for more rapid and accurate recognition. Our findings are compatible with this conclusion. HMAX recognized the same items as the ones present in the training more reliably, as it would be biased to recognize more easily the same features as the ones extracted during the training.

In this study we were also interested in the influence of the size of the analysis window for the context. Results showed that a tighter window around the object was better than a wider one. This seems to be counter-intuitive as the Gist is usually used for processing the whole image. In prior studies, Gist was used to determine the basic category of the scene as a whole, from which the probability of different object classes and their position would be inferred (Dalal & Triggs, 2005; Torralba, 2003; Torralba et al., 2004; Torralba & Sinha, 2001). Thus, a car would have a higher probability to occur in a street scene than in a countryside scene, and the car would be more likely to appear at the bottom of the image. In our experiment the context was defined more locally, such as a car is present on a road in urban or countryside scene. This can explain the fact that the classification is thus better for a smaller context focusing on the road than integrating too divergent information. This is in accordance with the fact that the Gist has been shown to be valuable as spatial layout cues for target search even without semantic congruence with the object (Castelhano & Heaven, 2011). Therefore, it is easier to find keys on a countertop even if it is in a bathroom scene. It would also be interesting to investigate the impact of the variation of the spatial extent of context in human perception.

We have shown that learning a small amount of context ($p_{context} = 10\%$) is sufficient to grasp enough contextual information for efficient object classification in contextual background. A small amount of contextual exemplars is sufficient for our model to generalize to other congruent context. Learning more contextual exemplars has a bigger influence for the "learned background set" where the exemplar where the same as the training and testing. Therefore, learning more context was not improving the

generalization to other context, but the specialization to the particular exemplars learned.

The impact of the object presence in context analysis window could also be questioned. As when the Gist is applied on whole scenes, our use of Gist features also considered not only the objects' surroundings but also the objects present in the scene. Thus, it included the local features of the object at the same time that it captured the summary features of the context. This could seem redundant as HMAX is already providing object features for object detection. However, as it has been shown in the literature, the presence of consistent and inconsistent objects affects scene perception. Our results reinforced the idea that the objects themselves influence the perception of a scene in a categorization task. Scenes that contained congruent objects were categorized more accurately (Davenport & Potter, 2004; Joubert et al., 2007; Mack & Palermi, 2010). Interestingly, Mack and Palermi (2010) also reached the same conclusion as human studies using a model of scene classification based on Gist information (Oliva & Torralba, 2001), which is compatible with our findings. Even the global scene statistics alone were able to reflect the presence of consistent or inconsistent objects. As object–scene consistency is apprehendable by both a Gist model based on global image statistics and by human observers, we chose to include the object in the extracted patch as part of the contextual information in the Gist model.

CONCLUSION

In this paper we studied the influence of contextual feature integration on object classification in two ways: (1) By varying the percentage of the training images that contained objects in contextually congruent contexts rather than uniform backgrounds, and (2) by varying the size of the window used for context features extraction. Both the local features object detector (HMAX) and the context feature extractor (Gist) were evaluated independently with different test data. HMAX showed weak ability to integrate contextual information, whereas Gist was, as expected, more sensitive to context information and more helpful when considering a tighter bounding box around the object. A concatenated feature vector of HMAX and Gist feature descriptors, fed into an object classifier using a Support Vector Machine, enhanced classification results when the context was consistent between the training and testing. Classification results were lowered when training and testing contexts were inconsistent (for example, learning object in uniform background but classifying them in contextual background). Learning objects in contextual background helped classification for object in their natural context; furthermore, only a small amount of object in context was sufficient to improve the results. Considering the context reduced the

sources of error and tended to give more strength to the few remaining, producing more understandable errors.

REFERENCES

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

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

Bicknell, K., & Levy, R. (2012). The utility of modeling word identification from input within models of eye movements in reading. *Visual Cognition*.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–263). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Blaschko, M., & Lampert, C. H. (2009). Object localization with global and local context kernel. In *Proceedings of British Machine Vision conference* (pp. 1–11). London, UK: BMVA Press.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention Perception and Psychophysics*, *75*, 1283–1297.

Castelhnao, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial association in visual search. *Psychonomic Bulletin and Review*, *18*, 890–896.

Chang, C. C., & Lin, C. J. (2001). Libsvm: A library for support vector machines (Software). Retrieved from http://www.csie.ntu.edu.tw/cjlin/libsvm

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research*, *28*, 2233–2247.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of Computer Vision and Pattern Recognition* (pp. 886–893). San Diego, CA: IEEE Computer Society.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*, 559–564.

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., & Herbert, M. (2009). An empirical study of context in object detection. *Proceedings of Computer Vision Pattern Recognition*, 1271–1278.

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*, 945–978.

Elazary, L., & Itti, L. (2010). A Bayesian model for efficient visual search and recognition. *Vision Research*, *50*, 1338–1352.

Frintrop, S., Nutcher, A., Surmann, H., & Hetzberg, J. (2004). Saliency-based object recognition in 3D data. In *Proceedings of international conference on Intelligent Robots and Systems* (Vol. 3, pp. 2167–2172). Sendai, Japan: EEE/RSJ.

Heitz, G., & Koller, D. (2008). Learning spatial context: Using stuff to find things. In *Proceedings of the 10th European conference on Computer Vision: Part I* (pp. 30–43). Marseille, France: Springer.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transaction on PAMI*, *20*(11), 1254–1259.

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*, 1–18.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*, 3286–3297.

Kanan, C., & Cottrell, G. (2010). Robust classification of objects, faces, and flowers using natural image. In *Proceedings of Computer Vision Pattern Recognition* (pp. 2472–2479). San Francisco, CA: IEEE.

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

Mack, M. L., & Palermi, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision*, *10*, 1–11.

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

Perko, R., & Leonardis, A. (2010). A framework for visual-context-aware object detection in still images. *Computer Vision and Image Understanding*, *114*, 700–711.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *Proceedings of the International Conference on Computer Vision* (pp. 1–8). Rio de Janeiro, Brazil: IEEE.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of objects recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Schad, D., & Engbert, R. (2012). The zoom lens of attention: Simulating shuffled versus normal text reading using the SWIFT model. *Visual Cognition*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transaction on PAMI*, *29*, 411–426.

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transaction on PAMI*, *29*, 300–312.

Sun, H.-M., Simon-Dack, S. L., Gordon, R. D., & Teder, W. A. (2011). Contextual influences on rapid object categorization in natural scenes. *Brain Research*, *1398*, 40–54.

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

Torralba, A., Murphy, L. P., & Freeman, W. T. (2004). Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Proceedings of Advances in Neural Information Processing Systems*, *Vol. 16*. Vancouver, Canada: MIT Press.

Torralba, A., Murphy, L. P., & Freeman, W. T. (2010). Using the forest to see the trees: Exploiting context for visual object detection and localization. *Communication of the ACM*, *53*, 107–114.

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 objects search. *Psychological Review*, *113*, 766–786.

Torralba, A., & Sinha, P. (2001). Statistical context priming for object detection. In *Proceedings of International Conference on Computer Vision* (Vol. 1, pp. 763–770). Vancouver, Canada: IEEE.

Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2009). What is the spatial extent of an object? In *Proceedings of Computer Vision Pattern Recognition* (pp. 1–8). Miami, FL: IEEE.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2002). Attentional selection for object recognition—a gentle way. *Lecture Notes in Computer Science*, *2525*, 472–479.

Wolf, L., & Bileschi, S. (2006). A critical view of context. *International Journal of Computer Vision*, *69*(2), 251–261.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, *44*, 1411–1426.