# A Bayesian Model of the Effect of Object Context on Visual Attention

**Ben Allison (ballison@inf.ed.ac.uk)**
**Frank Keller (keller@inf.ed.ac.uk)**
**Moreno I. Coco (mcoco@inf.ed.ac.uk)**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

## Abstract

Research in visual cognition has demonstrated that scene understanding is influenced by the contextual properties of objects, and a number of computational models have been proposed that capture specific context effects. However, a general model that predicts the fit of an arbitrary object with the context established by the rest of the scene is until now lacking. In this paper, we explain the contextual fit of objects in visual scenes using Bayesian topic models, which we induce from a database of annotated images. We evaluate our models firstly on synthetic object intrusion data, and then on eye-tracking data from a spot-the-difference task and from an object naming experiment. For the synthetic data, we find that our models are able to detect object intrusions accurately. For the eye-tracking data, we show that context scores derived from our models are associated with fixation latencies on target objects.

**Keywords:** visual attention; object context; Bayesian modeling; eye-tracking data.

## Introduction

Real-world objects are often related to each other and typically form a coherent scene. For example, a toothbrush is likely to occur with a tube of toothpaste, a mirror, a sink; it is unlikely to occur with a sauce pan, a salt shaker, a cooker. For a given object, it is therefore possible to determine whether it is in context in a scene (toothbrush in bathroom), or out of context (toothbrush in kitchen). Experimental evidence shows that context information facilitates human object recognition (Bar, 2004). In visual search tasks, eye fixations are targeted towards contextually appropriate regions (Torralba et al., 2006), and out-of-context objects attract fixations earlier than in-context objects (Underwood et al., 2008).

In computer vision, being able to detect out of context objects is useful for object labeling. The local detectors standardly used for this task only consider the visual features of the pixels within the bounding box of the object of interest (Felzenszwalb et al., 2010). Local detectors are therefore prone to confusing objects that are visually similar (e.g., fork and toothbrush). This problem can be addressed by combing a local detectors with a model of object context, i.e., a model that determines which objects occur together. While this approach has been shown to increase object labeling performance (Choi et al., 2010; Galleguillos et al., 2008), the context models used are simple, typically relying on co-occurrence statistics over object labels. Furthermore, the context models used in computer vision are not designed to capture human performance (e.g., in visual search). Therefore, these models have not been evaluated on tasks such as detecting out-of-context objects.

In this paper, we present a new model of object context based on a more complex notion of object label co-occurrence that makes use of latent (i.e., unlabeled and unobserved) scene types: the Latent Scene Type model. This model allows us to exploit the common structure of scenes in order to estimate reliable parameters even for infrequently occurring objects. We investigate two model variants: the first is Latent Dirichlet Allocation (LDA, Blei et al. 2003), a standard model of word-topic co-occurrence, which we use to capture object-scene type co-occurrence. The second model variant is formulated as a Bayesian mixture of multinomials, which assumes one latent scene type per scene (rather than one per object, as in LDA).

We test both model variants on the task of producing context judgments for objects in scenes. We first use a synthetic data set for evaluation (in this data, context objects have been artificially inserted). In the second evaluation study, we use our model to mimic the data from an eye tracking experiment in which human participants had to spot out-of-context objects. Finally, we demonstrate that our model can predict fixation latencies in an object naming experiment which included out-of-context objects.

## Related Work

To our knowledge, ours is the first model to attempt to quantify the degree of fit between arbitrary objects in a scene, and to correlate the predictions of such a model with human behavior in scene viewing tasks. However, a number of models have been proposed to capture context effects on visual attention; a prominent example is the Contextual Guidance Model (CGM, Torralba et al. 2006), which combines bottom-up saliency with global scene information (scene gist, Oliva & Torralba 2006). The model is trained on a set of images in which the target objects are labeled; from this data a probability distribution of typical positions of objects is learned. This distribution is conditioned on the scene gist, essentially a coarse-grained representation of global image features. Gist is a latent variable in the model, comparable to scene type in our approach. The CGM has been evaluated on eye-tracking data from visual search experiments, and can successfully predict the scene-type-specific search behavior that participants exhibit. However, the model is not specifically designed to detect out-of-context objects, and has not been evaluated on tasks that require an estimate of the contextual fit of an object.

In the computer vision literature, the work closest to ours in spirit, if not in ultimate task, is that of Choi et al. (2010). The authors use a generative model of images features, scene gist, the set of objects in the image, and their locations, to re-rank the output of a local object detector to respect contextual interactions between objects, and show an improvement over the baseline detector. The co-occurrence model used by Choi et al. (2010) is a fairly simple binary tree of presence features whose principal purpose is to facilitate inference on other aspects of the image.

The model of object context proposed in this paper is formulated as a topic model. While topic models have been studied extensively in both the language and vision literature, they originate from applications to text, beginning with Latent Dirichlet Allocation. LDA assumes a document is sampled from a mixture of multinomials, where the multinomial from which words are drawn is sampled once per word, and the mixture co-efficients are sampled once per document. A corpus is then a distribution over mixture co-efficients. This approach can be adapted fairly straightforwardly for modeling objects instead of words, and scenes instead of documents. However, we note that instead of being used as descriptive tools to provide insight into collections, in this paper we are interested in the predictive aspects of a topic model and want to test how well they correlate with human scene viewing data. In this respect, while the models we used are standard, the purpose for which we use them is novel and we derive new metrics to correlate with human behavior.

Topic models have been applied to images in the computer vision literature (Wang et al., 2009; Li et al., 2009), but rather than describing the sampling of object labels, these models specify how discrete-valued image patches are sampled (by quantizing continuous image features), and the relation between these patches and the labels applied to the image.

## Models of Latent Scene Type

This paper presents a model of context, and by extension contextual fit, which rests solely on the set of object labels in the image. The method we employ has two components: a distribution over the set of labels in the scene, and the application of such a model to a continuous measure of how well any object fits with that scene. The observation of sets of objects is explained through latent scene types, which can be thought of as simple clusters of objects which are likely to co-occur. We then use the predictive distribution over new sets of objects, as derived from our latent scene type models, to determine the fit of target objects to the scenes.

### A Model Of the Probability of a Set of Object Labels

We experiment with two models of the probability of a set of object labels, both of which are topic models. Topic models comprise mixtures over multinomial distributions, where in this case the multinomial outcomes correspond to object labels. The first topic model, Latent Dirichlet Allocation, is fairly standard, while the second one, the mixture of multinomials model, is less commonly used. Each of the models describes a distribution over a vector of counts, which we call $o$, such that $o_i$ is the count of the $i$-th object in the current image (note that for most images, most of these elements will be zero).

**Latent Dirichlet Allocation** There has been much interest in the use of topic models as descriptive tools, able to infer structure in collections of documents, or other collections of discrete entities (Blei et al., 2003; Wang et al., 2009). For the LDA model, the predictive distribution over new sets of object labels is given by:

$$p_{LDA}(o|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_n \sum_{z_n} p(z_n|\theta)\, p(l_n|z_n,\beta)\right) d\theta \tag{1}$$

where $l_n$ is the label for the $n$-th object in the image, and $z_n$ is the (latent) topic assignment for this label—counting the $l_n$ gives $o$ as defined above. The $z$s are indicators explaining which latent scene type was used to generate the current label. As in the original paper, $\alpha$ is the Dirichlet prior on $\theta$, and $\beta$ is the topic–word probability matrix that gives the probability of each object label in each topic. The above is evaluated using a particle-filter-inspired Monte Carlo method described by Wallach et al. (2009).

**Mixture of Multinomials** The mixture of multinomials model is defined over the same count vector $o$ as above, but for this model the scene type $z$ is sampled only once per scene. The parameters to the model are $\phi$ (the mixture coefficients) and $\theta$ (the parameters for the component multinomials). The distribution over the observable variables, $o$, is:

$$p(o) = \sum_z \phi_z p(o|\theta_z) \tag{2}$$

where the distribution $p(o|\theta_z)$ is a multinomial parametrized by the vector $\theta_z$ (its components giving the probabilities of each possible label occurring within that component).

We explore two variants of this model—the first uses maximum a posteriori (MAP) estimation to fix the parameters to the (approximate) posterior mode—the single best estimate of model parameters. This can be done using EM, and we employ uniform priors on both sets of parameters. Conditioned on some observations (training data, which we label $D$), the maximum likelihood method stipulates:

$$\hat{\phi},\hat{\theta} = \arg\max_{\theta,\phi} p(D|\phi,\theta) \tag{3}$$

$$p_{ml}(o|D) = p\left(o|\hat{\phi},\hat{\theta}\right) \tag{4}$$

This predictive distribution (4) is what we are interested in: exploring the probability of new scenes given our training data.

However, from a computational as well as cognitive perspective, given only limited samples from the process we should feel uneasy about saying with any certainty what the values of the parameters are. Instead, we suggest that given our experience we have beliefs about what is likely to happen, but we retain uncertainty and factor this in to our predictions. In light of this, we also employ a Bayesian version of this model, which integrates over the full parameter space given our training data $D$:

$$p_{bayes}(o|D) = \int p(\phi,\theta|D)\, p(o|\phi,\theta)\, d\phi, d\theta \tag{5}$$

For the mixture of multinomials model we cannot evaluate this integral in closed form, so we sample mixture models from their posterior $p(\phi,\theta|D)$—i.e., we retain uncertainty about which model best explains our data, and average over this uncertainty in deriving our predictions. Assuming Dirichlet priors (in our case, uniform) on $\phi$ and $\theta$ leads to Dirichlet posteriors over these same parameters, conditioned on assignments of training observations to latent mixture components; it is these assignments that we sample. We then evaluate the $p(o|\phi,\theta)$ at each of these sampled points; in practice we do not sample different mixture models for each new $o$ we wish to evaluate, but run the sampler once in training and store all mixture models sampled. This allows us to simply average over the sampled components for the predictive distribution, leading to a deterministic evaluation of (5) as simply the mean of the probability $p(o|\phi,\theta)$ under the sampled models.

As a final note for all these models, while inference techniques are approximate in all cases, and different between the MoM and LDA models, we are confident that the particular approximations do not overly sway the models' ultimate performance. While using heldout probability as the metric of concern show disparities between different approximations for the LDA model in Wallach et al. (2009), our uses of the models are different. Most of the problems we consider are decision problems where the exact probability is less of a concern than the relative probabilities of the scenes under two models, and in the final section we are interested in the correlation between the probabilities and some other continuous measure which is unlikely to be affected by (relatively) small changes due to approximation error.

## Detecting Out-of-context Objects with Scene Probability

The previous section presented models which have been used previously in other fields for describing the co–occurrence of entities. We turn in this section to the manipulation of these models to derive quantities which we will correlate with human performance.

There are two distinct tasks we explore in this paper: which of two objects is more probable given a scene, and whether a given object belongs to a scene or not. Here, we briefly describe the use of the models we defined in the previous section to achieve these tasks.

Firstly, the conditional probability of some object (label) in question $o'$ given a set of object labels $o$ (where $o$ is the count vector as above) is:

$$p(o'|o) = \frac{p(o \cup o')}{\sum_{o^{new}} p(o^{new} \cup o)} \tag{6}$$

That is, the probability of the count vector which includes the new label $o'$, normalized by the probability of the context for all possible objects which could be added to the scene.

To determine which of $o^1$ and $o^2$ better fits some context $o$ we can compare $p(o^1|o)$ and $p(o^2|o)$ computed as in (6)—we may simply interested in which of these is the larger, or perhaps in the ratio between these two quantities. Note that in either case, the normalizing constant can be dropped since it is common to both (this speeds up computation considerably).

Secondly, to determine whether $o'$ is in context or not, we can compare $p(o'|o)$ with the quantity obtained by marginalizing out the extra object, namely:

$$p(o^{\text{new}}|o) = \sum_{o^n} p(o^n)\, p(o^n|o) \tag{7}$$

where $p(o^n)$ is the probability of $o^n$ occurring in any scene, for which we use simply the fraction of all objects across all scenes which are $o^n$. For this paper, we explore both the decision problem (is $p(o'|o) > p(o^{\text{new}}|o)$, i.e., is the object in context or not) and the continuous scores derived as above.

## Evaluation on Synthetic Out-of-context Objects

We construct our first test set based on the Spatial Envelope data set (Oliva & Torralba, 2001). Here, the models will be used to determine whether an object is in context with respect to the rest of a scene, or not (Equation (7)). The images in the data set contain full object annotations, but also scene type labels. These allow us to construct test data for the scenario we are interested in. (Note, however, that this is the only use of overt scene type labels in this paper; the scene types in our model are latent.) The data set is annotated using LabelMe conventions, but does not overlap with the LabelMe data from which our models are estimated. In terms of objects per scene, there are on the order of ten objects in each image, and the number of images is reasonably balanced between scene types. We extract scenes which are either rural or urban (the two top level scene types). We produce frequency counts of objects within these two categories, and compute a $\chi^2$ statistic for each to measure the distinctiveness of that object in that class. We then select the 25 most distinctive objects for each class which occur in at least ten scenes, and extract all scenes containing each of these objects. These distinctive objects are treated as the targets, and the other objects in the image form the contexts.

The original scenes form examples of in-context objects—to produce out-of-context ones, for each scene we replace the in-context target with a randomly selected member of the distinctive list for the other category. This produces a set of just over 26,000 scenes, equally balanced between in- and out-of-context objects, to use for further experimentation. We divide this into 6,000 scenes for development (model selection and parametrization), with the remainder being used for held-out testing. In all cases, the held-out data are unobserved until all model parameters are fixed. Table 1 shows examples of the data we produce.

| Target | In/Out Context | Context |
|---|---|---|
| **stone** | *in* | stick:1 stone:1 tree_trunk_fallen:2 trees:1 ground:2 brushes:1 |
| **buildings** | *in* | skyscraper:1 building_occluded:2 buildings:1 sky:1 skyscraper_occluded:1 |
| **road** | *out* | tree:1 stone:3 river_water:1 trees:1 field:1 sky:1 stones:2 rocky_mountain:1 |
| **sea_water** | *out* | window:11 car_occluded:2 pot_plant_occluded:1 sidewalk:1 person_occluded:1 arcade:1 palm_tree:1 car:1 window_occluded:1 person_walking:3 person_woman_walking:1 traffic_light:1 hall:1 building:1 road:1 |

Table 1: Some examples of the synthetic data—the context is depicted as a sparse vector over the outcomes in the form [*label*:*count*], which is then reduced as appropriate for a trimmed vocabulary

|  | LDA |  | ML-MoM |  | B-MoM |  |
|---|---|---|---|---|---|---|
| $\|T\|$ | 500V | 1000V | 500V | 1000V | 500V | 1000V |
| 50 | 0.737 | 0.747 | 0.674 | 0.679 | 0.896 | 0.895 |
| 100 | 0.759 | 0.801 | 0.660 | 0.662 | 0.897 | 0.899 |

Table 2: Accuracy (proportion of decisions where the correct determination is made) on the synthetic data

**Results** Table 2 shows results on the synthetic dataset. The Bayesian Mixture of Multinomials is clearly superior to the other two models, and the larger vocabulary size and greater dimensionality improves this slightly. The LDA model shows greater sensitivity to parametrization than the other two, and the maximum likelihood model is considerably worse than the others across all parameter settings. Of particular note is the maximum likelihood model getting *worse* as the dimensionality increases; this is a classic result for non-Bayesian models, where as the parameter space expands it is less and less well summarized by a single point (the mode) and that mode becomes harder to find.

## Modeling Human Experimental Data

The evaluation study presented in the previous section used artificially generated data. It showed that the Latent Scene Type Model is highly accurate at detecting out-of-context objects which have been inserted into a scene. In the present study, we validate this result using a data set from an eye tracking experiment by Underwood et al. (2008). In this experiment, participants had to perform a search task (determine whether two scenes are the same or different); in the different-scene condition, the target object was either out of context or in context, with saliency being controlled. An example pair of scenes can be found in Figure 1. The results show that scenes with in-context objects are inspected for longer and received more fixations than scenes with out-of-context objects. Also the in-context objects themselves were detected later and required more fixations prior to detection than out-of-context objects.

We expect our Latent Scene Type Model to capture the behavioral effect of out-of-contextness demonstrated by Underwood et al.'s study: out-of-context objects should receive lower probabilities than in-context objects in their data set.

Underwood et al.'s study contains 80 pairs of scenes. In one scene in the pair, the target object is in context (congruent, in the language of that paper) and in the other it is out of context. (Saliency was also manipulated in the study, but this is not of interest here.) We manually listed the objects in each scene (the contexts are identical between pairs, and two pairs are identical save for their targets). Checking the labels against our LabelMe training data revealed that 25% of target objects were observed in LabelMe, and just over 30% of all objects. LabelMe contains mainly outdoor scenes, while the experimental data set are all indoor scenes, predominantly kitchen, utility room or bathroom scenes, in which the objects have been carefully arranged.

We therefore iteratively relabeled the target objects to establish a closer match with the LabelMe database, choosing in some cases synonyms and in others (direct) hypernyms. This was a manual process which relied on linguistic resources such as WordNet. This produced a target coverage rate of just over 70%, making it possible to use 45 of the 80 scene pairs, with each scene having on average approximately 60% of its context object appear in the training data (note that this increase was incidental, as we optimized the coverage of the target objects and simply propagated corrections through to contexts as well so as to reduce the amount of manual engineering). The selected scenes contain an average of ten objects in total.

Given the small size of the test set, we were not able to split off a separate development set, and therefore retained the parameters as set in the previous section on the synthetic data set.

**Results** Table 3 shows results on the Underwood et al. data for the task of detecting which of two possible objects is out of context. As in previous sections, we note that the Bayesian version of the mixture of multinomials model performs better than the maximum likelihood version of the model, but given the small dataset it is not possible to compare the B-MoM and LDA models except to say that both are significantly different from a random (50%) baseline as established by a binomial test. Note that while seeming disappointing initially, the performance of the models here is limited because the Un-

| Method | LDA | ML-MoM | B-MoM |
|---|---|---|---|
| Accuracy | 31/45 | 24/45 | 29/45 |

Table 3: Proportion of scenes in the Underwood et al. (2008) data where the correct determination was made. The LDA and B-MoM models are significantly different from a random (50%) baseline, but not one another

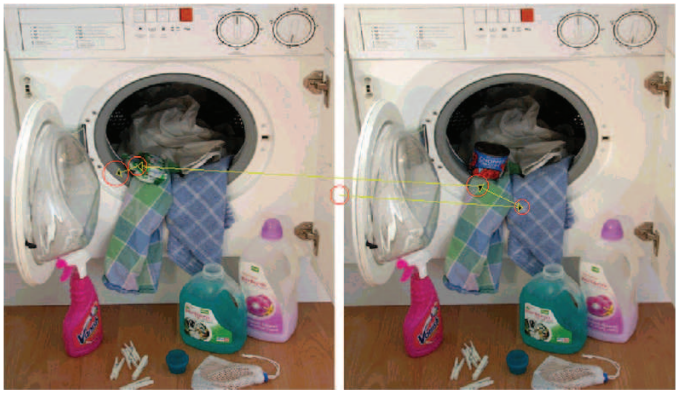
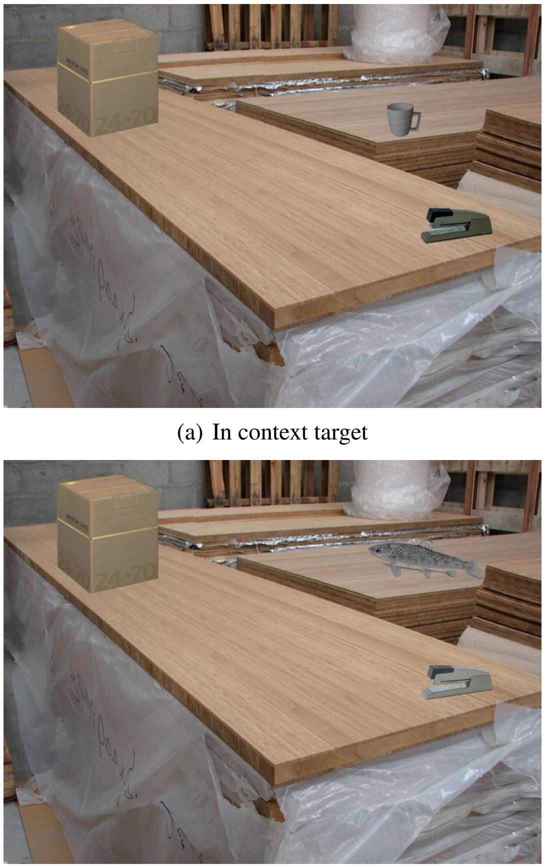


Figure 1: A pair of example scenes from the eye tracking experiment of Underwood et al. (2008). The target object in the left hand image is the sock (in context), while in the right hand image it is the can of soup (out of context)



(a) In context target

(b) Out of context

Figure 2: An example of a scene with an in-context target (cup) and the same scene with an out-of-context target (fish)

derwood et al. scenes are staged indoor shots featuring many objects that occur infrequently, if at all, in our training data (only 60% of context objects appeared at all). The next section presents an evaluation where objects are more frequently observed.

## Modeling an Object Naming Dataset

The third evaluation of our models used eye-tracking data from an object naming experiment by Coco et al. (2012). In this study, 24 participants were presented with 28 photo-realistic scenes and asked to name the five most important objects in the scene. In each scene, an object of interest and two competitors were inserted using Photoshop. The Saliency (Salient, Non-Salient) and Contextual Fit (In-Context, Out-of-Context) of the object of interest was manipulated. In contrast to Underwood et al. (2008), this study shows that out-of-context objects are less likely to be named than in-context objects. Moreover, first fixation latency, i.e., the time to land on a target object for the first time from scene onset, is longer for out-of-context than for in-context objects. A naming task demands a joint evaluation of both linguistic and visual information, thus even if an out-of-context object might be visually more informative, it is linguistically less relevant.

We first evaluate our models on the task of determining which of two objects is in context, identical to that presented in the previous section. Then, we investigate whether the contextual scores calculated by the models are correlated with the visual responses observed on the associated objects. We employ linear mixed effects model (LME, Baayen et al. 2008) analysis to investigate how first fixation latency (the dependent measure) correlates with model score (our predictor). LME is more appropriate than simple correlation because there were many other factors considered in the experimental data which affect the dependent measure, including frequency and saliency of objects in the scene and size of the objects. Employing LME means we are able to control for these factors by including them as covariates in the analysis.

On the basis of the experimental data, we expect the model score to be negatively associated with first fixation latency, i.e., the more out-of-context an object is, the longer it takes to fixate it. Together with the Score, we include as predictors the Saliency of the object, and the type of Model. As a random effect, we include Scene. We residualize first fixation latency by the area of the object (in pixel square) to reduce the effect of area on the dependent measure. We select the final LME model by following a forward step-wise procedure, where nested models are compared on the basis of log-likelihood improvement. In the following, we report the coefficients of the predictors found significant after model selection.

| Method | LDA | ML-MoM | B-MoM |
|---|---|---|---|
| Accuracy | 46/56 | 33/56 | 50/56 |

Table 4: Proportion of scenes in the Underwood data where the correct determination was made. The LDA and B-MoM models are significantly different from a random (50%) baseline, but not one another

| | LDA | | ML-MoM | | B-MoM | |
|---|---|---|---|---|---|---|
| | I | O | I | O | I | O |
| S | 0.0106 | 0.0001 | 0.0010 | 0.0010 | 0.0071 | 0.0006 |
| NS | 0.0064 | 0.0002 | 0.0010 | 0.0010 | 0.0054 | 0.0012 |

Table 5: Average context scores across the conditions—**I** is in context, **O** out of context, **S** is the salient condition and **NS** is the non-salient condition. ML-MoM scores are in fact slightly different to one another, but both contexts are highly improbable under the model

**Results** We first present the results for the decision problem. There are fifty-six decisions to be made (28 pairs of scenes, each in the salient and non-salient condition), with the goal being to determine which of the pair is out of context. Table 4 shows the results on this task, where we once again see that the LDA and Bayesian models are significantly above chance, but given the limited sample size not significantly different from one another. Table 5 shows the mean context scores across the conditions for each of the three models, where we see that the effects of the models in the decision problem (Table 4) are equally visible on the continuous scale.

When using the LME to check whether model score is a predictor of first fixation latency, we find a significant effect with $\beta_{Score} = -0.1309; p < 0.0001$: the more in-context an object is, the shorter the latency. We do not find an effect of Saliency and Model. This result also echoes the experimental finding obtained by Coco et al. (2012), and shows that the scores generated by our models can capture the patterns in the eye-movement responses.

## General Discussion

This paper introduced the Latent Scene Type models for describing the fit of objects to scenes. Our models quantify how well a target object fits an observed context (the other objects in the image). Sets of objects are generated by latent scene types, with scene types representing objects which tend to co-occur. We choose a Bayesian formulation for our models, as this is attractive from a cognitive point of view: a cognitive process operates with finite experience, which means that it has to estimate a model of the world based on a limited sample (in our case of context and objects). Committing to a single parameter setting based on a limited sample is difficult; it therefore seems more plausible to integrate over the full parameter space, which is the hallmark of Bayesian models. The Bayesian approach therefore captures the uncertainty faced by a cognitive process with access to limited data.

We showed that the Latent Scene Type models perform well on the task of detecting out-of-context objects in a synthetic dataset. Furthermore, we successfully applied the models to two eye-tracking datasets, one involving a spot-the-difference task, the other involving object-naming. In both cases, the models were able to successfully detect out-of-context objects, and in the case of the naming data, we also showed that model scores are associated with first fixation latencies on a target object (either in or out of context).

## Acknowledgments

The work reported here was funded by the European Research Council under award number 203427 "Synchronous Linguistic and Visual Processing".

## References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

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

Blei, D., Ng, A., & Jordan, M. (2003, March). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Choi, M., Lim, J., Torralba, A., & Willsky, A. (2010). Exploiting hierarchical context on a large database of object categories. In *Proceedings of cvpr'10* (p. 129-136).

Coco, M. I., Malcolm, G. L., & Keller, F. (2012). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Journal of Vision.* (under review)

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010, September). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence*, *32*(9), 1627 -1645.

Galleguillos, C., Rabinovich, A., & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *Ieee conference on computer vision and pattern recognition (cvpr).* Anchorage, AK.

Li, L.-J., Socher, R., & Fei-Fei, L. (2009). Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *Proceedings CVPR'09.*

McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit.*

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

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. In *Progress in brain research* (p. 2006).

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, *113*(4), 766–786.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, *17*, 159–170.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings ICML'09* (pp. 1105–1112). New York, NY, USA: ACM.

Wang, C., Blei, D., & Fei-Fei, L. (2009). Simultaneous image classification and annotation. In *In proceedings CVPR'09.*