Gates for Handling Occlusion in Bayesian Models of Images: An Initial Study

Conference Paper · October 2011

DOI: 10.1007/978-3-642-24455-1_21 · Source: DBLP

Dominik Endres
Philipps University of Marburg

Martin A. Giese
University of Tuebingen

Marina Kolesnik
Fraunhofer Institute for Applied Information Technology FIT

Gates for Handling Occlusion in Bayesian Models of Images: an Initial Study

Daniel Oberhoff[1], Dominik Endres[2], Martin A. Giese[2], and Marina Kolesnik[1]

[1] Fraunhofer FIT-LIFE, Schloss Birlinghoven, St. Augustin, Germany
daniel.oberhoff@fit.fraunhofer.de, marina.kolesnik@fit.fraunhofer.de
[2] Theoretical Sensomotorics, Cognitive Neurology, University Clinic Tübingen,
CIN, HIH and University of Tübingen, Germany
dominik.endres@klinikum.uni-tuebingen.de, martin.giese@uni-tuebingen.de

Abstract. Probabilistic systems for image analysis have enjoyed increasing popularity within the last few decades, yet principled approaches to incorporating occlusion *as a feature* into such systems are still few [11,7,8]. We present an approach which is strongly influenced by the work on *noisy-or* generative factor models (see e.g. [3]). We show how the intractability of the hidden variable posterior of *noisy-or* models can be (conditionally) lifted by introducing gates on the input combined with a sparsifying prior, allowing for the application of standard inference procedures. We demonstrate the feasibility of our approach on a computer vision toy problem.

Cite as:
Oberhoff D., Endres D., Giese M.A., Kolesnik M. (2011). Gates for Handling Occlusion in Bayesian Models of Images: an Initial Study In *KI2011: Advances in Artificial Intelligence, LNAI 7006, Springer*, 228-232.
The original publication can be found at www.springerlink.com,
DOI: 10.1007/978-3-642-24455-1.

1 Introduction

Both computer vision systems and models of (mammalian) biological vision have been researched extensively in the past few decades [4,5,10]. A key decision to make in the design of both types of system is which aspects of the visual environment are represented explicitly, and what details are to be disregarded by the system. The former determines the *specificity* of (the parts of) the system, while the latter are referred to as *invariances*. For example, invariance against shift, rotation, scaling and deformative transformations have been extensively modeled, which is due to these transformations being an ubiquitous part of the processes that generate natural images. We are concerned with another aspect of the image-generating process which has received less attention: *occlusion*. We argue, like [11,7], that occlusion is an important (and frequent) enough aspect to warrant explicit modeling, rather than treating it as noise. In other words, we propose to treat occluded parts of objects in an image as unobserved data, rather than formulating a generative model which expects these parts to be visible (this would e.g. be the assumption underlying linear generative models, or standard *noisy-or* models).

2 The Gated Convolutional Model

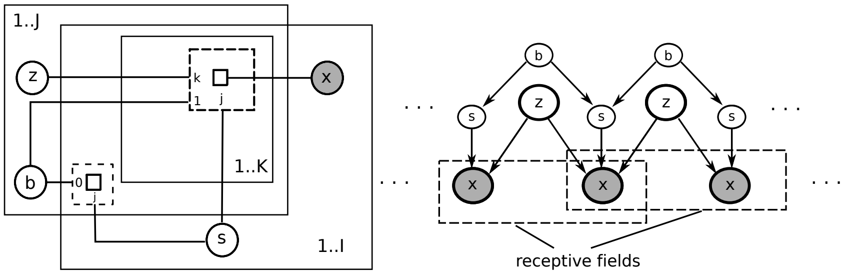

Fig. 1. Factor (*left*) and directed (*right*) graphs of a model layer with categorical factors and sparsity-promoting bits b. The factor graph uses plate notation and gates [9]. The observable data x are explained by the emission model (here: Gaussian clusters) associated with a latent variable z, if the gating variable s which connects x and z at a given point in the image is 'on'. A gating variable s can only be 'on' for a given x/z combination if the corresponding b is 'on'. We use a sparsity-promoting prior on b, i.e. most b are 'off' most of the time. Thus, the model will try to explain the image with few z only.

Our model is a convolutional directed Bayesian model in which the image is generated from a set of discrete latent variables, each of which has connections with a local receptive field in the image, where neighboring receptive fields are shifted by one pixel with respect to each other[3]. This effectively makes it a mixture model over (overlapping) image patches of fixed size. The novelty of our approach lies in the fact that we do not combine the predictions of multiple mixture models converging on one input, nor do we infer about a global ordering over the generated image patches [7,8]. Instead we introduce auxiliary variables which assign each input to a single mixture model only. This effectively implements an occlusion model because the image generated by one mixture model can 'occlude' parts of the image generated by a neighboring mixture model. In the following we will call these auxiliary variables 'gates'. Interestingly, this

[3] Our receptive fields are quadratic regions of the image and pixels contain all available channels of the image. For example, in an RGB image a pixel comprises a three-tuple with one real number per channel.

approach corresponds to the common variational approximation to the posterior of the QMR-DT multiple-causes model [6], where similar selector variables are introduced, effectively turning the noisy-or into an exclusive-or. We employ mixture models with categorical latent variables mainly for computational simplicity; factor models [7,8] could in principle also be used. To promote the forming of proper object hypotheses and to avoid simplistic solutions where each input pixel is modeled by a separate latent variable, we place sparsity promoting priors on the gates, encouraging them to use only few latent variables to explain the input.

The joint distribution for a layer with gated mixture models is:

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{z}, \mathbf{s}, \mathbf{b}, \boldsymbol{\Theta}) = \prod_{ijk} &\left[p(x_i \mid \Theta_{ijk})^{s_{ij} z_{jk}} p(z_j) \right]^{b_j} \\
&\times \prod_{ij} \mathbb{1} \left[b_j \vee (\neg s_{ij} \wedge \neg b_j) \right] p(\mathbf{b}) p(\mathbf{s}) p(\boldsymbol{\Theta}).
\end{aligned} \tag{1}
$$

where i is an input variable index, j is a latent variable index, k is the mixture component index (components are shared among latent variables, i.e. the network is convolutional). Θ_{ijk} are the parameters of the emission model (here: Gaussian clusters with diagonal covariance) which connect the latent variables z_j to the observable data x_i. The latent variables z_j are vectors in 1-out-of-K encoding. Similarly, the gate variables $\mathbf{s}$ are comprised of I vectors in 1-out-of-J encoding. $p(\mathbf{z})$ and $p(\mathbf{s})$ are the priors over the latent and gate variables, respectively. We furthermore introduce one binomial variable b_j per latent variable, indicating the availability of latent variables for explaining the data. When the b_j associated with a latent variable is 'off', gate settings which assign an input to the latent variable become *forbidden*. By assigning a low prior probability to the 'on' state of each b_j, sparsity of the b_j is encouraged, therefore the inputs will be assigned to a small subset of latent variables. The second product with the indicator function ensures that no gate switches to a latent variable whose bit is off. $p(\mathbf{b})$ is a product of identical binomial distributions for each bit. The above graphical model is sketched in plate and gate notation (see [9] for an explanation of the gate notation) in figure 1. Note that due to the convolutional nature of the network together with the possibility to 'pick' a subset of receptive fields for explaining the data, we can get shift invariance for free, since the gates can always choose a receptive field with the 'right' shift to explain some subset of input pixels (i.e. an 'object').

Inference in our model can be performed efficiently by blocked Gibbs sampling, since the latent variables are independent given the gates, and the gates are independent given the bits and the latent variables. During sampling gate proposals within one receptive field are generated together with proposals for the corresponding latent variable. Note that a latent variable whose bit is 'off' does not need to be updated, since it has no observable effects.

Learning the parameters: We put conjugate exponential priors on all parameters and use the inferred latent and gate variable distributions to compute an approximate posterior. In effect, this is variational Bayesian expectation maximization (VBEM) [1]. For the mixture prior we use the truncated stick breaking approximation to the Dirichlet process, thus avoiding the need for random initialization and reducing the effect of the (arbitrary) choice of the number of available mixture components [2].

3 Results and Conclusion

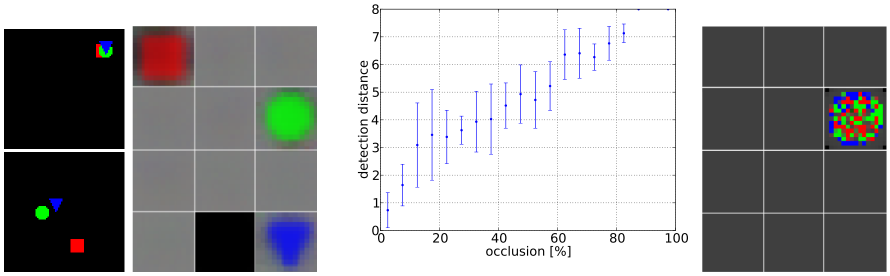

Fig. 2. *Left*: two example frames from the toy example sequence. The full sequence is one hundred frames long. Each frame has a size of 80x80 pixels. *Middle Left*: Expected receptive fields of the mixture components learned from the toy data. Due to the sparsity promoting prior the model is forced to explain the data with few latent variables, leading to a proper representation of the encountered objects and the background by separate mixture components. *Middle Right*: object localization accuracy on the toy data set. Localization degrades gracefully with increasing occlusion. *Right*: Expected receptive field of the mixture components learned without the sparsity prior (i.e. all latents are always available for explaining the input).

We demonstrate the ability of the model to separate objects that have a constant shape but occlude each other in various configurations on a toy data set consisting of an artificially generated RGB image sequence in which three geometric shapes of red greed and blue color move randomly over a black background. The shapes occasionally occlude each other, examples are shown in figure 2, left[4]. For this experiment receptive field sizes are chosen to match the size of the 'objects'.

To promote sparse solutions, we assign a very low 'on' probability to each b_j[5]. We use 5 Gibbs iterations per latent variable during inference. The model learns a separate component for each of the three objects and the background (see figure 2, middle left). For comparison we have also trained a model without the

[4] Here the input pixels thus consist of RGB-triples, and thus a separate gate variable exists for each such triple.

[5] we usually used a sparsity prior of 0.0001, but the exact value, as long as it is below the actual expected sparsity, did not make much of a difference

sparsity promoting prior, meaning that any latent can be used to explain the input at any time, and the resulting expected receptive fields are shown in figure 2, right. In this case a single high entropy component is learned, and the image is explained purely by the gates, i.e. no object concept is formed.

The model with sparsity promoting prior can now be used as an object detector: A detection is indicated by a latent variable, with enabled associated b_j, selecting one of the object representing components. The object location of this detector is defined as the most probable location of the original object image within the expected receptive field of the corresponding mixture component. In figure 2, middle right, we plotted the accuracy of this localization as a function of object occlusion. Localization performance degrades gracefully with increasing occlusion.

To conclude, we have shown a feasible approach for modeling occlusion in a Bayesian image model. Its main features are a gated 'competition' between possible foreground explanations and strong sparsity promotion. We have demonstrated the viability of this approach by applying it to a toy dataset, were we could also demonstrate the key role of the sparsity prior for forming stable object hypothesises.

Acknowledgments: this work was supported by EU projects FP7-ICT-215866 SEARISE, FP7-249858-TP3 TANGO and FP7-ICT-248311 AMARSi.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer-Verlag (2006)
2. Blei, D.M., Jordan, M.I.: Variational methods for the dirichlet process. In: In Proceedings of the 21st International Conference on Machine Learning (2004)
3. Courville, A., Eck, D., Bengio, Y.: An infinite factor model hierarchy via a noisy-or mechanism. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 405–413 (2009)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (1). pp. 886–893 (2005)
5. Földíak, P.: Learning invariance from transformation sequences. Neural Computation 3, 194–200 (1991)
6. Jaakkola, T.S., Jordan, M.I.: Variational probabilistic inference and the qmr-dt network. J. Artif. Int. Res. 10, 291–322 (May 1999), http://portal.acm.org/citation.cfm?id=1622859.1622869
7. Le Roux, N., Heess, N., Shotton, J., , Winn, J.: Learning a generative model of images by factoring appearance and shape. Tech. rep., Microsoft Research (2010)
8. Lücke, J., Turner, R., Sahani, M., Henniges, M.: Occlusive components analysis. In: Advances in NIPS (2009)
9. Minka, T., Winn, J.: Gates: A graphical notation for mixture models. In: Advances in NIPS (2008)
10. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
11. Tamminen, T., Lampinen, J.: A bayesian occlusion model for sequential object matching. In: Proc. British Machine Vision Conference 2004. pp. 547–556 (2004)