# Continuous Factor Graphs for Holistic Scene Understanding

Niko Sünderhauf, Ben Upcroft, Michael Milford
Australian Centre for Robotic Vision, Queensland University of Technology, Brisbane, Australia
niko.suenderhauf@qut.edu.au

## Abstract

*We propose a novel mathematical formulation for the holistic scene understanding problem and transform it from the discrete into the continuous domain. The problem can then be modeled with a nonlinear continuous factor graph, and the MAP solution is found via least squares optimization. We evaluate our method on the realistic NYU2 dataset.*

## 1. Introduction and Motivation

*Holistic* approaches to scene understanding exploit the rich semantic and spatial relations between individual objects in a scene or between objects and the entire scene to boost the performance of individual object and scene classifiers. *Discrete* graphical models such as conditional random fields (CRFs) are commonly applied to model and solve this problem [2,5]. In this paper we explore the application of a *continuous* graphical model (a factor graph) for this task.

We are particularly interested in examining how scene understanding can be addressed jointly with structure from motion or monocular SLAM approaches where continuous factor graphs are extensively used: Object classification implicitly provides size cues about the objects that can be used for depth initialization and 3D reconstruction. Vice versa, the 3D information obtained by SfM contain valuable cues for object recognition. This paper therefore contributes towards expressing both scene understanding and SfM / SLAM in a common mathematical framework.

## 2. Factor Graphs for Scene Understanding

Factor graphs [4] are graphical models that are commonly applied to model probabilistic estimation problems over hidden continuous variables $\mathcal{X}$ and evidence $\mathcal{Z}$. The maximum-a-posteriori estimate of the distribution $P(\mathcal{X}|\mathcal{Z})$ forms an optimization problem $\mathcal{X}^* = \operatorname{argmax}_{\mathcal{X}} P(\mathcal{X}|\mathcal{Z}) = \operatorname{argmax}_{\mathcal{X}} \prod_i P_i(\bar{\mathcal{X}}_i|\bar{\mathcal{Z}}_i)$ where $\bar{\mathcal{X}}_i \subseteq \mathcal{X}$ and $\bar{\mathcal{Z}}_i \subseteq \mathcal{Z}$ are factored subsets of $\mathcal{X}$ and $\mathcal{Z}$. If the single factors $P_i$ are continuous and Gaussian, they are of the general form $P_i(\bar{\mathcal{X}}_i|\bar{\mathcal{Z}}_i) = \eta \exp -\frac{1}{2}\|\mathbf{e}_i(\bar{\mathcal{X}}_i, \bar{\mathcal{Z}}_i)\|^2_{\mathbf{\Sigma}_i}$ where $\mathbf{e}_i(\bar{\mathcal{X}}_i, \bar{\mathcal{Z}}_i)$ is a problem-specific error function. We can solve for the $\mathcal{X}^*$ by least squares optimization.

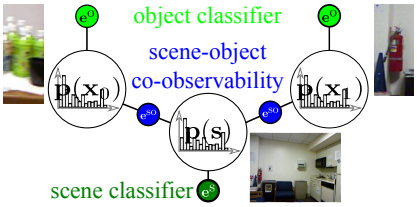


Figure 1. We model the scene understanding problem with a factor graph over *continuous* variables. In contrast to previous work – where the variables are *discrete* – we can perform exact MAP inference using efficient nonlinear least squares optimization.

### 2.1. From Discrete to Continuous Variables

The scene understanding problem addressed here aims at finding the optimal discrete label assignment to observed objects $x_i$ and the scene type $s$, given the observed image and prior semantic knowledge. In order to model and solve the scene understanding problem with continuous factor graphs, we have to transform the problem from the discrete into a continuous domain. Instead of creating the graphical model over discrete variables $x_i$ and $s$, we utilize the *probability distributions* $\mathbf{p}(\mathbf{x}_i)$ and $\mathbf{p}(\mathbf{s})$ as high-dimensional, continuous variables in our formulation: $\mathcal{X} = \{\mathbf{p}(\mathbf{x}_0), \ldots, \mathbf{p}(\mathbf{x}_n), \mathbf{p}(\mathbf{s})\}$. Likewise, we interpret the results of the individual classifiers for the objects and the scene type as *measurements* or *observations*: $\mathcal{Z} = \{\mathbf{z}_0^{\text{object}}, \ldots \mathbf{z}_n^{\text{object}}, \mathbf{z}^{\text{scene}}\}$.

Our formulation corresponds to a probabilistic estimation over probability distributions. The results of the maximum-a-posteriori (MAP) inference therefore are *distributions* over the object and scene classes: $\mathcal{X}^* = \{\mathbf{p}^*(\mathbf{x}_0), \ldots, \mathbf{p}^*(\mathbf{x}_n), \mathbf{p}^*(\mathbf{s})\}$. To retrieve the optimal class label $x_i^*$ for the $i$-th object, another operation $x_i^* = \operatorname{argmax} \mathbf{p}^*(\mathbf{x}_i)$ is performed, and identically executed for $s^*$. This is in contrast to the MAP inference step in a CRF where — due to the discrete formulation of the problem — the MAP results are class labels directly [2,5].

### 2.2. Definition of Factors and Error Functions

In this section we develop the factorization of $P(\mathcal{X}|\mathcal{Z})$ and define the individual error functions for these factors.

**The Object Measurement Factor:** This unary factor captures the relation between an object distribution variable $\mathbf{p}(\mathbf{x}_i)$ and the associated observation $\mathbf{z}_i^{\text{object}}$ which is a distribution created by an object classifier, e.g. a ConvNet. The factor models that in the absence of all other information, the observation $\mathbf{z}_i^{\text{object}}$ is the most likely configuration for $\mathbf{p}(\mathbf{x}_i)$. The error function of the object measurement factor is defined as: $\mathbf{e}_i^{\text{object}} = \mathbf{p}(\mathbf{x}_i) \ominus \mathbf{z}_i^{\text{object}}$. The $\ominus$ operator is the *total variation* measure for discrete distributions that is defined as: $\mathbf{p}(\mathbf{x}) \ominus \mathbf{q}(\mathbf{x}) = \frac{1}{2}\sum_i |p(x_i) - q(x_i)|$.

**The Scene Measurement Factor:** Analogous to the object measurement factor defined above, the scene measurement factor is given as: $\mathbf{e}^{\text{scene}} = \mathbf{p}(\mathbf{s}) \ominus \mathbf{z}^{\text{scene}}$.

**The Scene-Object Factor:** This factor captures the co-occurrence information contained in the conditional probability distribution $\mathbf{p}(\mathbf{x}|\mathbf{s})$ that can be either learned from training data or modelled with the help of a human expert. The error function is defined as:

$$\mathbf{e}_i^{\text{scene-object}} = \begin{pmatrix} p(s_0) \cdot (\mathbf{p}(\mathbf{x}_i) \ominus \mathbf{p}(\mathbf{x}_i|s_0)) \\ \vdots \\ p(s_M) \cdot (\mathbf{p}(\mathbf{x}_i) \ominus \mathbf{p}(\mathbf{x}_i|s_M)) \end{pmatrix} \quad (1)$$

Intuitively, if no object classifications $\mathbf{z}_i^{\text{object}}$ are available, this factor would drive the factor graph's MAP solution towards the conditionals, weighted by the estimated probability for the individual scene labels $s_i$.

**Constructing and Solving the Graph:** Given the three factors defined above, the overall MAP inference problem encoded in the graph (see Fig. 1 for an illustration) is:

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmin}} \|\mathbf{e}^{\text{s}}\|^2_{\mathbf{\Sigma}^{\text{s}}} + \sum_{i=0}^{n} \|\mathbf{e}_i^{\text{o}}\|^2_{\mathbf{\Sigma}_i^{\text{o}}} + \sum_{i=0}^{n} \|\mathbf{e}_i^{\text{so}})\|^2_{\mathbf{\Sigma}_i^{\text{so}}} \quad (2)$$

The superscripts in $\mathbf{e}^{\text{s}}, \mathbf{e}^{\text{o}}$, and $\mathbf{e}^{\text{so}}$ denote the error functions of the scene, object, and scene-object factors. We define the covariance matrices $\mathbf{\Sigma}^{\text{s}} = -H(\mathbf{z}^{\text{scene}})$ and $\mathbf{\Sigma}^{\text{o}} = -H(\mathbf{z}_i^{\text{object}})$. Here $H(\mathbf{p}) = \sum_i p_i \log(p_i)$ is the entropy of the distribution $\mathbf{p}$. This adjusts the influence of a factor according to the certainty of the classifier that produced the associated the observation. Following the same considerations, $\mathbf{\Sigma}_i^{\text{so}} = \text{diag}\left(-\frac{1}{M} H(\mathbf{p}(\mathbf{x}_i|s_j))\right)$.

We use `gtsam` [1] to implement the error functions and maintain the factor graph. The nonlinear optimization problem (2) is solved with the iterative Levenberg-Marquardt algorithm provided by [1]. Our implementation re-normalizes the estimated variables after each iteration so that they remain proper probability distributions.

## 3. Experimental Evaluation

We evaluate our approach on the NYU2 dataset [6]. To create the object and scene classification observations $\mathbf{z}_i^{\text{object}}$ and $\mathbf{z}^{\text{scene}}$ we use the pretrained `AlexNet` [3] and `Places205` [7] ConvNets.

**Dataset Preparation:** We identified 95 object classes from NYU2 that are known to `AlexNet`. Using the pixel-accurate ground truth labels we extracted bounding boxes for all object instances of these common classes. Similarly we identified 10 indoor scene types that are shared between NYU2 and `Places205`. Using the same split as in [6] we divided the NYU2 dataset into a training (2528 objects from 621 scenes) and test subset (1667 objects from 463 scenes).

**Learning the Conditional Probability $\mathbf{p}(\mathbf{x}|\mathbf{s})$:** The conditional probability model $\mathbf{p}(\mathbf{x}|\mathbf{s})$ that is used in the scene-object factors $\mathbf{e}_i^{\text{scene-object}}$ is learned from the training dataset. This is done by determining the relative object class frequencies for each scene type. Non-occurring objects were given a small but non-zero prior probability.

**Results:** The stand-alone `AlexNet` classified 42.8% of all 1667 objects correctly. With our approach – exploiting the scene classification from `Places205` and the scene-object factors – this accuracy increased by 2.8% to 45.6%.

## 4. Conclusions

We proposed a novel formulation of the holistic scene understanding problem using a factor graph over continuous variables and solving it via iterative nonlinear least squares optimization. Our evaluation demonstrated the feasibility of this approach: The object classification accuracy on NYU2 improved when exploiting scene-object co-occurrence information. Although the achieved gain of 2.8% is moderate, it is in the order of magnitude that can be expected from previous work: [5] gained 0.98% when fusing appearance with scene-object information in their CRF framework on a similar subset of the NYU2 dataset. In future work we will model factors that incorporate object-object relations, scene geometry cues, or object sizes, comparable to [5]. Going beyond that, we will demonstrate the full potential of our proposed formulation by jointly modeling and solving the MonoSLAM and scene understanding problem.

## References


[1] GTSAM – The Georgia Tech Smoothing and Mapping Library. https://collab.cc.gatech.edu/borg/gtsam/.

[2] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? Text-to-Image Coreference. In *CVPR*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[4] F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Information Theory*, 2001.

[5] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

[6] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[7] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.