Supplementary Materials: On Parameter Learning in CRF-based Approaches to Object Class Image Segmentation

Sebastian Nowozin[1], Peter V. Gehler[2], and Christoph H. Lampert[3]

[1] Microsoft Research Cambridge, UK,
[2] ETH Zurich, Switzerland,
[3] Institute of Science and Technology, Austria.

1 Notation

We summarize the notation used in the main paper in Table 1.

Symbol	Description
$\mathcal{X}$	Observation domain of a single image region.
$\mathcal{Y} = \{1, \ldots, C\}$	Set of region labels.
$X = \{X_1, \ldots, X_W\}$	Set of observation variables.
$Y = \{Y_1, \ldots, Y_V\}$	Set of label variables.
$\mathcal{F} \subseteq 2^V \times 2^W$	Set of all factors.
$T = \{1, \ldots, \lvert T \rvert\}$	Set of factor types.
$t(F) \in T$	Factor type of factor F.
$\boldsymbol{\mu}_F \in \{0,1\}^{\mathcal{Y}^F}$	Overcomplete representation of $\boldsymbol{y}_F$.
$\boldsymbol{\theta}_F^{t(F)}(\boldsymbol{x}_F, \boldsymbol{w}_{t(F)}) \in \mathbb{R}^{\mathcal{Y}^F}$	Feature function defining the energies for all possible labelings $\boldsymbol{y}_F \in \mathcal{Y}^F$ within the factor F of type $t(F)$.
$H_F^a : \mathcal{X}^F \to \mathbb{R}^{D_a}$	Image feature a extracted from the set of image regions belonging to F.

Table 1. Notation used to describe factor graphs.

2 Image Features

SIFT. We extract color image descriptors using the implementation of van de Sande [1]. In particular, we use the WSIFT descriptors extracted on a regular grid of sizes 6, 10, 20, and 40 pixels. From the set of descriptors of the training images, we subsample 250,000 descriptors and k-means cluster them into 512 codewords. Each region is then assigned a 512-bin L_1-normalized bag-of-words histogram of its feature points.

QPHOG Features. For each image region in the segmentation tree we find the tightest bounding box around the region. We create a binary segmentation mask within the bounding box by assigning each pixel within the region a value of one and each pixel outside the region a value of zero. We rescale the resulting black-and-white image to a fixed size of 96-by-96 pixels and extract pyramid histogram of oriented gradients (PHOG) features using the VGG implementation[4] in eight 360-degree-orientation bins and three pyramid levels producing a 680-dimensional feature f_i for each region i. This feature faithfully encodes the shape of the region. In order to match this shape to some suitable template shapes, the features of all regions in all segmentation trainval images are clustered using k-means with $k = 512$ to obtain cluster centers $u_1, \ldots, u_{512}$. A new "QHOG" feature $q_i \in \mathbb{R}^{512}$ is defined for each region i by $q_{i,j} = K(u_j, f_i; \sigma_{\text{PHOG}})$, where K is a standard Gaussian RBF kernel and we choose σ_{PHOG} such that the mean response over all regions in all images is approximately 0.1.

QHOG Features. The QHOG features are constructed by taking the smallest bounding box containing an image region and extracting a grid of 7-by-7 HOG blocks from the image content using the HOG implementation of Felzenszwalb.[5] All $\approx 230k$ descriptors of the training set are clustered using k-means into 512 codewords and processed as for the QPHOG features.

3 Details regarding Max-Margin Learning

The structured SVM algorithm finds a solution to the following convex[6] non-differentiable optimization problem.

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|^2 + \frac{C_{\text{svm}}}{N} \sum_{n=1}^{N} \max_{\boldsymbol{y} \in \mathcal{Y}^V} (\Delta(\boldsymbol{y}_n, \boldsymbol{y}) + E(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{w}) - E(\boldsymbol{y}; \boldsymbol{x}_n, \boldsymbol{w})), \tag{1}$$

where $\Delta(\boldsymbol{y}_n, \boldsymbol{y})$ is usually assumed to be a semi-metric on $\mathcal{Y}^V$, the set of possible labelings. Intuitively, the objective (1) can be understood as minimizing the energy $E(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{w})$ assigned to true labeling $\boldsymbol{y}_n$ and at the same time maximizing the energy assigned to all other labelings, thus maximizing the margin of the predictor. A larger difference of a labeling to the ground truth, as determined by $\Delta(\boldsymbol{y}_n, \boldsymbol{y})$ leads to a larger enforced margin.

The use of (1) entails the choice of the function $\Delta(\boldsymbol{y}_n, \boldsymbol{y})$, the parameter C_{svm}, and a method to solve (1). For $\Delta : \mathcal{Y}^V \times \mathcal{Y}^V \to \mathbb{R}_+$ we choose the same function as [2], with $\Delta(\boldsymbol{y}^1, \boldsymbol{y}^2) = \sum_{i \in V} s_i \sum_{c \in \mathcal{Y}} (\mu_i^1(c) + \mu_i^2(c) - 2\mu_i^1(c)\mu_i^2(c))$, where $\boldsymbol{\mu}^1$ and $\boldsymbol{\mu}^2$ correspond to the overcomplete representation of $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$,

[4] `http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.html`

[5] `http://people.cs.uchicago.edu/~pff/latent/`

[6] Convexity is easily proven as $E(\boldsymbol{y}; \boldsymbol{x}_n, \boldsymbol{w})$ is a *linear* function in $\boldsymbol{w}$ and thus $\max_{\boldsymbol{y} \in \mathcal{Y}^V} E(\boldsymbol{y}; \boldsymbol{x}_n, \boldsymbol{w})$ is convex by construction.

respectively. The scalar weighting constant $s_i \geq 0$ is the relative image-plane size ($s_i = r_i/(\sum_{j \in V} r_j)$, where r_j is the number of pixels in the j'th region). Thus we give a larger influence to larger regions in the image, something that is not possible in the standard CMLE training method.

Unlike in the intractable general case [3] we can always perform exact MAP labeling for our model for all settings of $\boldsymbol{w}$. Hence, in principle we can also optimize (1) exactly. However, in our case where $\boldsymbol{w}$ contains many parameters and C_{svm} is large the usual cutting plane training procedure [4] leads to very large quadratic programming problems, eventually exhausting available memory; this is true despite pruning inactive constraints in each iteration. Hence we could only evaluate small values of C_{svm}. The performance was always improving as we increased C_{svm}, suggesting that current training methods need to be improved to benefit from the formulation (1).

4 VOC2009 Challenge Evaluation

We have trained one version of our model using the SIFT, QHOG, QPHOG, and STF features and the data-independent pairwise potential on the entire segmentation training and validation set (1499 images) and submitted it to the VOC2009 evaluation server.

The official results from the recent VOC2009 challenge and the results from our model are shown in Table 2. Note that a key distinction must be made between models trained on the segmentation set only (seg, 1499 images) and models trained on both the detection and segmentation sets (cls+seg, 7054+1499 images), as the latter have an almost six times larger set of training images.

The BONN_SVM-SEGM method, which has been trained on the segmentation images only achieves a high performance of 36.3% on the official evaluation measure. It has recently be described in a pair of papers [5, 6] and works by first producing a set of segments as object hypotheses that are then ranked using a learned regression function. The fact that BONN_SVM-SEGM method is not a multi-class random field model, but instead classifies individual image segments from figure ground segmentations indicates that the general random field model could possibly be improved by hierarchical, higher-order factors, a direction that was recently considered by Ladický et al. [7].

References

1. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
2. Nowozin, S., Lampert, C.H.: Global connectivity potentials for random field models. In: CVPR. (2009)
3. Finley, T., Joachims, T.: Training structural SVMs when exact inference is intractable. In: ICML. (2008)
4. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. JMLR **6** (2005) 1453–1484

Submission name	Accuracy	Trained on
CVC_HOCRF	34.5%	cls+seg
NECUIUC_CLS-DTCT	29.7%	cls+seg
UoCTTI_LSVM-MDPM	29.0%	cls+seg
NECUIUC_SEG	28.3%	cls+seg
LEAR_SEGDET	25.7%	cls+seg
UCI_LAYEREDSHAPE	24.7%	cls+seg
BROOKESMSRC_AHCRF	24.8%	?
UC3M_GEN-DIS	14.5%	?
BONN_SVM-SEGM	36.3%	seg
UCLA_SUPERPIXELCRF	13.8%	seg
our model	15.5%	seg

Table 2. VOC 2009 segmentation results on the test set.

5. Carreira, J., Sminchisescu, C.: Constrained parametric min-cuts for automatic object segmentation. In: CVPR. (2010)
6. Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: CVPR. (2010)
7. Ladický, L., Russell, C., Kohli, P.: Associative hierarchical crfs for object class image segmentation. In: ICCV. (2009)