# Latent Semantics Local Distribution for CRF-based Image Semantic Segmentation

Giuseppe Passino
giuseppe.passino@elec.qmul.ac.uk
Ioannis Patras
ioannis.patras@elec.qmul.ac.uk
Ebroul Izquierdo
ebroul.izquierdo@elec.qmul.ac.uk

MMV Group
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, UK

Semantic image segmentation is the task of assigning a label of a semantic category (*e.g.* car, bicycle, tree) to every pixel of an image. This task is posed as a supervised learning problem in which the appearance of areas that correspond to a number of semantic categories are learned from a dataset of manually labelled images. In this context, the major challenge is how to jointly consider visual properties and the context of a pixel at different scales. To this end, our contribution is the proposal of a method that combines a region-based probabilistic graphical model that builds on the recent success of Conditional Random Fields (CRFs) in the problem of semantic segmentation [2, 6], with a salient-points-based bags-of-words paradigm. Visual words located at salient points have proved very powerful for object modelling [1, 5] but have been rarely used for semantic segmentation. Compared to similar methods based on distributed features [6, 7], our approach effectively combines complementary complementary information from patches and salient points. In particular, two different integration strategies are explored. The first one is to build (weighted) histograms of local words distributions, the other one is to consider weighted distributions of latent topics associated to the words by the means of probabilistic Latent Semantic Analysis (pLSA) [5].

The analysis of an image proceeds as follows. In a first stage, the image is oversegmented into patches. These are obtained via the Normalised Cuts (NCuts) spectral clustering algorithm [4]. Patches are then described with visual features that depend on their colour (hue histogram), texture (textons), and position (normalised centre coordinates).

The graph on which the CRF is imposed is then obtained. We use a tree obtained from the connectivity graph given by the oversegmentation. A tree structure allows for fast exact inference in the CRF. This is important in terms of both algorithm performance and for the convergence in the training step. Reducing the patch connectivity leads to limited context consideration: we however maximise the correlation between patches by linking patches coherent in appearance. In the proposed appearance-coherent Minimum spanning Tree (acMST) algorithm, edge weights depend on the similarity between the hue part of the linked patches descriptors. The distance measure is the symmetric Kullback-Leibler divergence, defined for two distributions $P, Q$ as

$$D_{KLs}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \ , \tag{1}$$

where $D_{KL}$ is the (asymmetric) KL divergence.

A labelling $\mathbf{y}$ is a vector $\mathbf{y} = (y_1, \ldots, y_m)$ for $m$ image patches, where $y_i \in \mathscr{L} = \{l_1, \ldots, l_n\}$. The CRF imposes an a posteriori probabilistic distribution for the labelling over the nodes of the patch graph $G = \{\mathscr{V}, \mathscr{E}\}$, given the features $\mathbf{X}$. This is modelled as a Gibbs distribution

$$p(\mathbf{y}|\mathbf{X};\theta) = \exp\left[\Psi(\mathbf{y},\mathbf{X},\theta)\right]/Z(\mathbf{X},\theta) \ , \tag{2}$$

where $\theta$ is the model parameter vector, and $Z$ a normalisation factor. The *local function* $\Psi$ is a summation of terms depending on clique variables, that in our model are unary and pairwise. Therefore

$$\Psi(\mathbf{y},\mathbf{X},\theta) = \sum_{v\in\mathscr{V}} (\theta_{y_v} \cdot \mathbf{x}_v) + \sum_{(i,j)\in\mathscr{E}} \theta_{y_i,y_j} \ . \tag{3}$$

The loose notation $\theta_{y_v}$ in the first summation indicates a linear, category-dependent parameter appearance vector and $\theta_{y_i,y_j}$ a (symmetric) label compatibility coefficient encoding patches dependences. Partial labelling of a subset $\mathscr{V}_a \subseteq \mathscr{V}$ of nodes is considered by averaging over the latent nodes $\mathscr{V}_l = \mathscr{V} \setminus \mathscr{V}_a$, $p(\mathbf{y}_a|\mathbf{X};\theta) = \sum_{\mathbf{y}_l\in\mathscr{V}_l} p(\mathbf{y}|\mathbf{X};\theta)$. Inference in the model is done via the Belief Propagation (BP) algorithm. In the training phase, the Maximum A Posteriori (MAP) training set probability is maximised. A quasi-Newton iterative method is used to find the optimal parameters.

| model | precision | model | precision |
|---|---|---|---|
| Lit. [7] | 84.9 | $\text{WWH}_{6,12}$ | **86.2** |
| Base | 82.9 | $\text{LTD}_{24}$ | 84.0 |

Table 1: Classification precision results (average). "Base" is the baseline without additional descriptors. The subscripts {6,12,24} in WWH and LTD refer to the value of window standard deviations $\{d/6, d/12, d/24\}$.

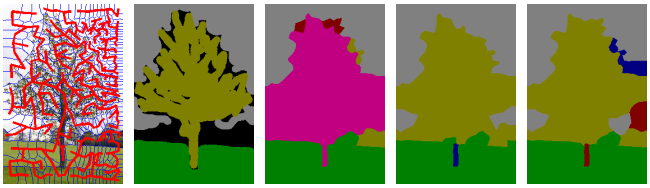

Figure 1: Segmented image (with acMST), ground truth, and segmentation according to baseline model, $\text{WWH}_{6,12}$ and $\text{LTD}_{24}$ respectively.

Descriptors based on visual words taken at salient points are at first detected and represented via the SIFT algorithm [3]. Visual words are obtained by clustering of the descriptors. Descriptors based on local word distributions are calculated for each patch. Both the strategies are based on histograms of word contributions. These are weighted by a Gaussian $w_s(\mathbf{l},\mathbf{l}_p) \propto \mathscr{N}(\|\mathbf{l}-\mathbf{l}_p\|, \sigma_s^2)$, $\mathbf{l},\mathbf{l}_p$ being the word location and patch centre respectively. Descriptor scale changes with the parameter $\sigma_s$, and different scales can be combined by descriptors concatenation. The first proposed descriptor is the Windowed Words Histogram (WWH), that is, the histogram of words in the vicinity of a patch (reduced in dimensionality by PCA). Alternatively, for the Latent Topic Distribution (LTD) descriptor, a posterior distribution over latent topics is at first associated to each word via pLSA. The final patch descriptor is obtained as weighted sum of local posteriors. The advantage of this second approach lays in the separate representation of the semantic content for each visual word. Word contributions are therefore considered as independent in the CRF. The additional descriptors are considered in the CRF framework with additional appearance vectors $\theta^d_{y_v}$ (compare with Eq. (3)).

Results on the publicly available MSRC database of 9 categories show improvements over the baseline and the state of the art, as reported in Table 1. When applying the LTD descriptor, a slight drop of performance can be noticed compared to the WWH. This is likely to be due to the simplification assumption associated to the dimensionality reduction before local word aggregation. In Fig. 1 the segmentation and acMST output, as well as the classification results with different methods, are shown.

[1] L Cao and L Fei-Fei. Spatially coherent latent topic model for concurrent segment. and classific. of objects and scenes. In *ICCV*, 2007.

[2] P Kohli, L Ladicky, and PH Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[3] DG Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[4] DR Martin, CC Fowlkes, and J Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549, 2004.

[5] J Sivic, BC Russell, AA Efros, A Zisserman, and WT Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[6] T Toyoda and O Hasegawa. Random field model for integration of local inform. and global inform. *PAMI*, 30(8):1483–1489, 2008.

[7] J Verbeek and B Triggs. Scene segmentation with CRFs learned from partially labeled images. In *NIPS*, 2007.