

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

Giuseppe Passino  
giuseppe.passino@elec.qmul.ac.uk  
Ioannis Patras  
ioannis.pstras@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

---

## Abstract

Semantic image segmentation is the task of assigning a semantic label 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. This paper proposes 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, with a salient-points-based bags-of-words paradigm. In a first stage, the image is oversegmented into patches. Then, in a CRF-based formulation we learn both the appearance for each semantic category and the neighbouring relations between patches. In addition to patch features, we also consider information extracted on salient points that are detected in the patch's vicinity. A visual word is associated to each salient point. Two different types of information are used. First, we consider the local weighted distribution of visual words. Using local (i.e. centred at each patch) word histograms enriches the classical global bags-of-word representation with positional information on word distributions. Second, we consider the un-normalised local distribution of a set of latent topics that are obtained by probabilistic Latent Semantic Analysis (pLSA). This distribution is obtained by the weighted accumulation of the latent topic distributions that are associated to the visual words in the area. The advantage of this second approach lays in the separate representation of the semantic content for each visual word. This allows us to consider the word contributions as independent in the CRF formulation without introducing too strong simplification assumptions. Tests on a publicly available dataset demonstrate the validity of the proposed salient point integration strategies. The results obtained with different configurations show an advance compared to other leading works in the area.

## 1 Introduction

The paper proposes a method for semantic segmentation of natural images. This task entails the association of each image pixel with a label of a semantic category (*e.g.* car, bicycle, tree). An example of semantic segmentation is given in Fig. 1 for an image from the Microsoft Research Cambridge (MSRC) database<sup>1</sup>. The latter is used throughout the paper and described in Section 4.



Figure 1: Semantic segmentation example for an image from the MSRC database. The legend specifies the different categories used in the paper.

Semantic segmentation is conceptually different from classical segmentation, the ultimate goal being the correct category association to image areas rather than simple object boundary detection. Furthermore, semantic segmentation methods are based on a bottom-up analysis aimed at the labelling of each pixel. This is in contrast with object detection methods based on top-down object models, in which object instances are searched within the image by finding the best model match.

The major challenge in the task of semantic segmentation is how to jointly consider visual properties and the context of a pixel at different scales. The short-range relationships have been successfully modelled via probabilistic graphical models. Of particular interest are methods based on discriminative models such as Conditional Random Fields (CRF) [8], the discriminative version of a Markov Random Field (MRF). With CRFs, shared (non local to single nodes) features can be naturally integrated into the model without the introduction of oversimplifying independence hypotheses. Additionally, since appearance probability distributions are not explicitly modelled, fewer training examples are required, resulting in reduced computational complexity.

Long-range connection cliques in the CRF make exact inference intractable. For this reason CRFs are used to enforce locally smooth labelling [6, 19]. The introduction of higher order potentials defined over regions [6] results in improvements in the accuracy on the object boundaries rather than in more consistent overall labelling configurations. Since close pixels are usually semantically strongly correlated, a reasonable strategy is to cluster groups of pixels into patches and apply the CRF at patch-level [21, 24]. This reduces the graph size and the computational complexity. In this framework, the main research question is how to include visual features and long-range information in the model to describe long-range interactions between graph nodes.

## 1.1 Related Work

With generative models, part dependences at different scales are made explicit by modelling the presence and layout of objects in the scene [9]. This often has a great cost in terms of training and inference complexity. In contrast, with discriminative models, one way towards greater context awareness is the use of global clues. In particular, distributed [19] or global features [21, 24] can be integrated in the patch feature vector. These features do not directly relate to semantic categories, and have to be carefully chosen in order to capture context: descriptors calculated in large areas typically suffer from background noise and fail to focus on objects of interest. Verbeek and Triggs [24] consider aggregate versions of local descriptors, with the disadvantage of accepting a degree of redundancy in the feature set. Other works rely on simple, appearance-based image-level descriptors [21] that are too generic and not tailored to the description of object instances, giving therefore weak context information. In

*Textonboost* [19], distributed features focus on appearance coherence between near pixels rather than longer range semantic coherence.

Another approach is to integrate independently learnt long-range constraints in the probabilistic model [4, 21]. This greatly increases the complexity of both inference and learning because these constraints are based on categories (e.g., through specification of allowed label patterns) rather than on features, that introduce long-range label dependencies. Learning constraints on a large number of elements needs large training sets. Such constraints have limited generalisation properties, tending to be bound to the specific training data from which they are obtained.

## 1.2 Paper contributions and overview

In the light of what said so far, we propose a method to integrate distributed information to local patch analysis in a CRF-based framework for semantic segmentation. The graphical model is patch-based, where the patches are obtained through oversegmentation. In this way, several advantages are achieved: first, we split the complexity of handling the correlation among neighbouring pixels from the short-scale statistical dependencies between patches. Secondly, object boundary detection accuracy is improved in comparison with methods based on rectangular patches [2]. Furthermore, descriptors extracted from coherent patches are less affected to noise deriving from co-presence of different categories.

The additional information interacts with the probabilistic model as an additional feature, thus retaining structural simplicity and efficient inference properties. A key point of our contribution is that these additional features are derived from distribution of *visual words* capturing local appearance at salient points in the vicinity of a patch. Such information has proved very powerful for object modelling [1, 16, 20] but has been rarely used for semantic segmentation. The reason is that salient points are sparse and therefore fail to cover the full area of the image homogeneously; dense descriptors are used instead [2, 11]. Nonetheless, considering both features at salient points and densely extracted features for content description can be beneficial since they contain largely complementary information. Salient points are localised at stable extrema of the scale-space, and the associated descriptors are designed to optimise the representation of such areas. Descriptors extracted at patches obtained by low-level segmentation provide dense coverage of the image, and describe areas homogeneous in the feature space. We consider two alternative methods to integrate visual words. 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) [8].

Both the proposed descriptors are calculated in the vicinity of the patch. In this way, significant information related to the visual words location is integrated in the bag-of-words paradigm, which in despite of the recently obtained success relies only on the relatively basic information of word co-presence in the image. Moreover, changing the size of the descriptors support allows to consider word contributions at different scales.

The remainder of the paper is organised as follows. In Section 2 our base model is presented. In Section 3 the two different proposed strategies for integrating contextual information extracted at salient points are presented. The experimental results are discussed in Section 4. Finally, Section 5 terminates the discussion by presenting our conclusions.

## 2 Labelling Framework

### 2.1 Probabilistic Model

The patch-based CRF is defined over a graph  $G = \{\mathcal{V}, \mathcal{E}\}$  in which each node  $v_i \in \mathcal{V}$  is associated to a random variable  $y_i$  over the sample space  $\mathcal{L} = \{l_1, \dots, l_n\}$  describing the label of the  $i$ -th image patch. A labelling  $\mathbf{y}$  is the vector  $\mathbf{y} = (y_1, \dots, y_m)$  for the  $m$  image patches. The set of variables is Markovian, that is, each variable is independent on the entire graph when conditioned on its neighbours. This property allows the factorisation of the conditioned probability function for  $\mathbf{y}$ , that is a Gibbs distribution

$$p(\mathbf{y}|\mathbf{X}; \theta) = \frac{\exp \Psi(\mathbf{y}, \mathbf{X}, \theta)}{Z(\mathbf{X}, \theta)}, \quad (1)$$

where  $\theta$  is the model parameter vector,  $\mathbf{X}$  the observation (features) and  $Z(\mathbf{X}, \theta)$  a normalisation factor. The *local function*  $\Psi$  is a summation of terms depending on clique variables, so that  $\Psi(\mathbf{y}, \mathbf{X}, \theta) = \sum_{c \in \mathcal{C}} \phi_c(\mathbf{y}_c, \mathbf{X}, \theta)$ , where  $\mathcal{C}$  is the set of cliques and  $\mathbf{y}_c$  is the projection of  $\mathbf{y}$  on the clique  $c$ , that is,  $\mathbf{y}_c = (y_i : i \in c)$ . In our model, cliques are unary and pairwise [24, 19, 24], so that

$$\Psi(\mathbf{y}, \mathbf{X}, \theta) = \sum_{v \in \mathcal{V}} \sum_{k \in \mathcal{K}_1} \theta_k \phi_k^1(y_v, \mathbf{X}) + \sum_{(i,j) \in \mathcal{E}} \sum_{k \in \mathcal{K}_2} \theta_k \phi_k^2(y_i, y_j, \mathbf{X}), \quad (2)$$

where  $\mathcal{K}_1$  and  $\mathcal{K}_2$  are the sets of indices  $k$  of the parameter vector  $\theta$  referring to different unary and pairwise potentials, respectively. The unary potentials encode the appearance model. They have the form  $\phi_k^1(y, \mathbf{x}) = x_{i_k} \delta(y, l_{j_k})$ , where  $\delta$  is the Kronecker delta and the indices  $i_k, j_k$  are indexed in  $k$  to span the feature vector and the label set respectively. The pairwise potential functions encode a symmetric category look-up table. Their form is  $\phi_k^2(y, y') = \delta(y, l_{i_k}) \delta(y', l_{j_k})$ , where the indices  $i_k, j_k$  span the label set. The pairwise potentials are independent of patch features, since in previous works these have been shown to be unhelpful [24]. The model handles the presence of unlabelled patches in the training set. Whenever only a subset  $\mathcal{V}_a \subseteq \mathcal{V}$  of nodes is assigned, leaving the nodes  $\mathcal{V}_l = \mathcal{V} \setminus \mathcal{V}_a$  latent, the likelihood of the *assigned* nodes,

$$p(\mathbf{y}_a|\mathbf{X}; \theta) = \frac{\sum_{\mathbf{y}_l} \exp \Psi(\mathbf{y}, \mathbf{X}, \theta)}{Z(\mathbf{X}, \theta)}, \quad (3)$$

is considered in spite of Eq. (1). An additional “void” category in the label space accounts for observations that are not associated to any of the given semantic classes.

**Training and Inference.** Inference in the model is done via the Belief Propagation (BP) algorithm. In the training phase, the Maximum A Posteriori (MAP) criterion is used as fitness function to be maximised. A quasi-Newton iterative method is used to find the maximum of the training set likelihood

$$\log(L) = \sum_i \log(L_i) - \frac{\|\theta\|^2}{2\sigma_\theta^2}, \quad L_i = p(\mathbf{y}_i|\mathbf{X}_i; \theta), \quad (4)$$

where  $i$  is the image index in the training set and the term  $\frac{\|\theta\|^2}{2\sigma_\theta^2}$  is a Gaussian prior imposed to the parameters to control overfitting. By introducing a *conditioned* graph (indicated by

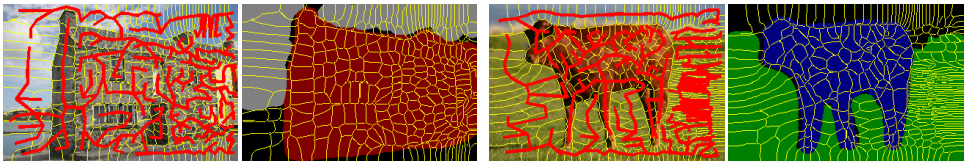


Figure 2: Oversegmentation (in yellow) with overlaid acMST (in red) for two database images, and oversegmentation overlaid to their hand-made labelling.

the subscript  $c$ ) where all the labelled nodes are assigned, the log-likelihood of the  $i$ -th image assumes the form

$$\log(L_i) = Z_i(\mathbf{X}_i, \theta) - Z_{ci}(\mathbf{y}_{ai}, \mathbf{X}_i, \theta) . \quad (5)$$

The gradient of Eq. (5) can be written in terms of conditioned or unconditioned clique variables marginal probabilities, that are provided by the BP algorithm.

When labelling a new image, the optimal labelling

$$\mathbf{y}_{\text{opt}} = \arg \max_{\mathbf{y}} \{p(\mathbf{y}|\mathbf{X}; \theta)\} , \quad (6)$$

is iteratively found via the max-sum algorithm.

## 2.2 Patch Extraction

The advantages of using oversegmentation in the context of part-based scene analysis with random fields have been proved [9, 15]. In our work patches are obtained via the Normalised Cuts (NCuts) spectral clustering algorithm [18], accounting for texture and edge information in the pixel similarity measure [13].

Patch features are based on colour, texture, and positional information. For colour we use an invariant hue descriptor [22], in the form of a 30-bins hue histogram. For texture we use textons [21], histograms of texture words obtained by clustering non-normalised, multi-scale, multi-orientation Difference-of-Gaussians (DoG) filterbank responses. We have chosen a 300 words dictionary, built across the entire database. The descriptor dimension is then reduced to 40 by the application of a Principal Component Analysis (PCA) algorithm to ease the learning of the appearance by the CRF. The normalised centre coordinate vector of the patch is used as position feature.

## 2.3 Graph Connectivity

The oversegmentation provides the connectivity for the patches. We aim at building a tree structure from the original graph for two main reasons. First, an open structure allows for fast exact inference with BP: the number of BP iterations is linear with respect to the diameter of the tree (that is, the greatest distance between any two nodes), while in presence of a closed graph the Loopy BP (LBP) algorithm, a generalisation of BP, does not offer guarantees of convergence [9]. Other than for the algorithm performance, exact inference is crucial for the convergence of the optimisation of a cost function based on differential terms as in Eq. (5). Reducing the patch connectivity leads to limited context consideration: we however maximise the correlation between patches by linking patches coherent in appearance. We propose the appearance-coherent Minimum spanning Tree (acMST) algorithm to build the

tree. In this MST algorithm, the weight of an edge is 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

$$D_{KLs}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) , \quad (7)$$

where  $D_{KL}$  is the (asymmetric) KL divergence. In Fig. 2 an example of oversegmentation and acMST is presented, as well as a visual idea of how the oversegmentation matches the hand-labelling. Quantitatively, the use of acMST instead of another connectivity criterion as weighting the edges on the distance between the patch centres reported an increase in performance of over 1% on the MSRC dataset.

### 3 Distributed Descriptors

The basic method described in Section 2 is enriched with distributed descriptors to account for context on a scale larger than the one considered with the CRF connections. These descriptors, being based on visual words taken at salient points, carry information that is complementary to the one associated to image patches. We propose two methods to integrate visual words into the CRF framework, the first one based on words histograms, the second one on latent topics histograms. The two strategies however differ only on the means by which the distributed feature is computed, not on the integration in the CRF.

In both the methods, salient points are at first detected and descriptors extracted with the SIFT algorithm [14]. A dictionary of 1000 visual words is then obtained via  $k$ -means clustering over all the dataset. Then, for each patch an additional feature vector that accounts for local word distributions is calculated, according to the two different strategies detailed in the remainder of this section. In both the cases, the contribution of the single word is weighted on its distance from the patch centre, thus achieving the advantages related to bag-of-words models [20] while not discarding positional information. The weight  $w_s$  is a Gaussian  $w_s(\mathbf{l}, \mathbf{l}_p) \propto \mathcal{N}(\|\mathbf{l} - \mathbf{l}_p\|, \sigma_s^2)$ , where  $\mathbf{l}, \mathbf{l}_p$  are the word location and patch centre respectively. By changing the variance  $\sigma_s^2$  of the window we consider salient points in narrower or broader neighbourhoods. Finally, the additional feature vector  $\mathbf{x}_d$  is integrated in the CRF via singleton potential factors  $\phi_k^3(y, \mathbf{x}_d) = x_{d,i_k} \delta(y, l_{j_k})$ . The additional factors do not compromise the complexity of the inference in the probabilistic model since distributed information is accounted for at feature-level rather than in terms of label patterns.

**Windowed Word Histograms.** The first of the proposed descriptors is the Windowed Words Histogram (WWH), that is, the histogram of words in the vicinity of a patch, weighted as detailed earlier. The descriptor length is then reduced to avoid an unbalance in the number of parameters associated to different features in the CRF, that would worsen the learning. PCA is used to reduce the dimensionality of the descriptor to 20. The word histograms enclose full word co-presence information, and the PCA acts over all the descriptors in the dataset, thus optimising the descriptivity of the reduced feature vectors.

**Latent Topic Distributions.** An alternative descriptor is the Latent Topic Distribution (LTD). This descriptor is obtained by associating a compact representation to the single words and simply accumulating them in the vicinity of each patch. Visual words are at first considered globally when the pLSA algorithm [1] is used to associate a distribution of latent

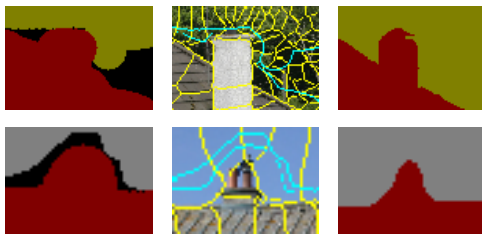


Figure 3: Detail of two examples in which the ground truth provided with the database is not consistent with the actual object category near the object boundaries. On the left, the ground truth image; on the centre, the original image with segmentation and ground truth categories boundaries; on the right, our labelling.

topic posteriors to each word. We decided to use a total of 20 latent topics, that are expected to represent different traits of some of the categories (visual words in general cover different categories unequally). To obtain the final patch descriptor, these posteriors are weighted as described earlier and summed up to form the final descriptor.

As for the WWH descriptor, the pLSA entails a dimensionality reduction, but in this case that happens before the integration of contributions from different words in the local window. This is a simplifying assumption, since some information related to words co-presence is mainly lost before the windowing step. On the other hand, the use of pLSA ensures that distributions of words in the entire image is considered when associating the descriptor to the words. Additionally, this representation allows for more flexibility, as for example for a dynamic choice of the window centres and scales [17], that is however not performed at the moment. Finally, adding topic distributions associated to different words is not coherent with the assumption of independent words, but allows to consider local word densities in an effective way.

## 4 Experimental Results

For the assessment of the proposed methods, we used the publicly available MSRC database of nine categories. This challenging set of images presents complex outdoor scenes, and indoor scenes with faces in cluttered background. Instances of object categories in general appear at different scales and exhibit large variability in terms of appearance. Multiple instances of the same category can appear in one image. The ground truth labelling of the images is provided at pixel level. In Fig. 1 an example with associated ground truth is shown. The database contains 9 semantic categories, and ambiguous pixels are left unlabelled (“void”). The use of this dataset allows us to compare the results with previous research works [23, 24]. To increase the reliability of the results, they are averaged over runs on three different splits of training/test images, in a ratio of 75%/25%.

**Performance estimation.** In the remainder of the paper, unless explicitly stated otherwise, we present the classification accuracy at patch-level. We consider that a patch is classified correctly if its label is the same as that of the majority of the pixels in the ground truth label map. Results on the estimation of the labelling accuracy at pixel level, which are obtained by labelling each pixel of a patch with the corresponding patch label and then comparing

	Building (14.5%)	Grass (30.1%)	Tree (14.1%)	Cow (7.2%)	Sky (13.4%)	Airplane (2.8%)	Face (3.2%)	Car (7.5%)	Bicycle (7.3%)	Average
Lit. [23]	74.0	88.7	64.4	77.4	95.7	<b>92.2</b>	88.8	81.1	78.7	82.3
Lit. [24]	73.6	91.1	<b>82.1</b>	73.6	<b>95.7</b>	78.3	89.5	<b>84.5</b>	81.4	84.9
Base	63.0	94.2	68.9	84.4	93.7	75.8	92.9	76.4	86.5	82.9
WWH <sub>G</sub>	50.3	87.4	70.3	73.7	81.2	65.7	85.2	66.2	83.1	75.3
WWH <sub>3</sub>	71.2	94.2	71.7	85.4	94.0	73.1	96.9	70.0	95.2	84.8
WWH <sub>6</sub>	74.9	94.6	72.1	87.9	94.9	73.0	99.3	74.6	94.2	86.2
WWH <sub>12</sub>	68.1	<b>95.1</b>	75.4	<b>87.4</b>	94.3	73.2	96.8	79.7	90.1	85.6
WWH <sub>6,12</sub>	<b>76.7</b>	94.6	71.4	86.3	95.0	73.1	<b>99.3</b>	73.2	<b>93.7</b>	<b>86.2</b>
LTD <sub>6</sub>	57.5	94.2	76.9	84.3	92.5	73.9	88.2	80.6	88.5	83.2
LTD <sub>12</sub>	60.2	94.3	76.5	84.2	93.2	73.5	91.9	80.3	89.6	83.9
LTD <sub>24</sub>	61.5	94.1	76.6	83.9	93.8	74.3	91.6	80.4	88.3	84.0

Table 1: Classification precision results for different categories and in weighted average. Categories relative occurrences are shown under the name. “Base” is the base model without additional descriptors. The subscripts  $\{G,3,6,12,24\}$  in WWH and LTD refer to the value of window standard deviations  $\{+\infty, d/3, d/6, d/12, d/24\}$ , respectively.

with the ground truth label map, highlighted inaccuracies in the localisation of the object borders in the ground truth maps. In some occasions this penalises the accurate segmentation provided by our segmentation method. This is illustrated in Fig. 3 in which details of both the ground truth and the estimated label field are depicted for two example images.

**Windowed Words Histogram.** Labelling results are presented in Table 1, reporting average precision and the category breakdown. For the WWH descriptor we chose  $\sigma_s \in \{+\infty, d/3, d/6, d/12\}$ , where  $d$  is the image diagonal, for the standard deviation of the Gaussian that determines the aperture of the histogram window. When the WWH descriptor is applied to the model, results are sensibly affected. However, if the scale of the window is not consistent with coherent words, the descriptors fail in conveying relevant information and this results to noise in the features and therefore in degraded performance. In particular, we experience an improvement for all the scales except the global one. This is in contrast with previous findings [24], that can be however explained with the fact that the location of the words on salient point ties them to objects increasing their positional significance. We also tested a combination of descriptors at multiple scales (the two best-performing ones), obtaining further improvements. The combination has been obtained simply appending the related feature vectors. Our strategy clearly outperforms similar methods in the literature tested over the same dataset. The variability on the single categories precisions when changing the window size parameter is due to the fact that in the training the global accuracy is maximised rather than the single category ones, and the categories better described at each scale are favoured in different cases.

**Local Topic Distribution.** When applying the LTD descriptor, a slight drop of performance can be noticed in comparison with the WWH, even though the performance still reports improvements over the base model, placing the method roughly at the same level of similar works in the literature. As hinted when introducing the LTD descriptor, this is likely to be due to the simplification assumption associated to the dimensionality reduction before local word aggregation. The independence between contributions is highlighted in the very reduced performance difference with different configurations, with  $\sigma_s = \{d/6, d/12, d/24\}$ .



Figure 4: Two examples of segmented images from the MSRC database. In the first column the original images, on the second one the ground truth, while the last three columns contain the segmentation according to the base model, the  $WWH_{6,12}$  and the  $LTD_{12}$  respectively.

To have a visual insight of the improvements associated to our method, labelled examples are shown in Fig. 4. In the selected images is possible to see how visual words help in enhancing the precision in the classification of border patches or entire objects.

## 5 Conclusions

We proposed two different strategies to improve context awareness for image semantic segmentation through the application of distributed descriptors to a CRF graphical model. The graphical model is based on patches that densely cover the image. Distributed descriptors built on visual words taken at salient points complement the patches. Visual words link to robust visual traits of object instances, and their reliability in describing object categories is proved. In contrast, dense patches are homogeneous in colour and texture and isolate clusters of pixels that are likely to belong to the same object.

Two descriptors are proposed and analysed. The WWH descriptor is based on histograms of visual words in the vicinity of each patch, where the single word contributions to the histogram are weighted on the distance from the patch centre. In this way, only local words are accounted when labelling a patch. Word position information is considered as a result, addressing one of the main shortcomings of the bag-of-words model.

In the WWH descriptor the feature vector is reduced in dimensionality via PCA to make the integration to the CRF framework possible. In contrast, the proposed LTD descriptor associates a compact representation, *i.e.* a distribution of latent topics, to the single word, and topics distributions are then simply added up for each patch, weighting the contributions on the word distances from the patch centre. The simplification introduced in reducing the dimensionality of the feature before the integration step partially affects the results, but it introduces greater flexibility in considering words distributions. In particular, one of the future directions of the work is towards dynamic windows when considering local word distributions, to reflect the fact that the words associated to each object instance in the image are not always close to the patch centre. Experiments on the MSRC public dataset show

clear improvements of our method when compared to other recent proposals in the field.

## 6 Acknowledgements

The research leading to this paper was supported by the European Commission under contract FP7-216444, Peer-to-peer Tagged Media, PetaMedia.

## References

- [1] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *IEEE 11th International Conference on Computer Vision*, 2007. doi: <http://dx.doi.org/10.1109/ICCV.2007.4408965>. URL <http://dx.doi.org/10.1109/ICCV.2007.4408965>.
- [2] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2005. doi: <http://dx.doi.org/10.1109/CVPR.2005.16>. URL <http://dx.doi.org/10.1109/CVPR.2005.16>.
- [3] Brendan J. Frey and David J. Mackay. A revolution: Belief propagation in graphs with cycles. In *Advances in Neural Information Processing Systems*, volume 10, 1997. URL <http://books.nips.cc/papers/files/nips10/0479.pdf>.
- [4] Xuming He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *2004 IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2004. doi: <http://dx.doi.org/10.1109/CVPR.2004.1315232>. URL <http://dx.doi.org/10.1109/CVPR.2004.1315232>.
- [5] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001. URL <http://portal.acm.org/citation.cfm?id=599631>.
- [6] Pushmeet Kohli, L'ubor Ladicky, and Philip H. Torr. Robust higher order potentials for enforcing label consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. doi: <http://dx.doi.org/10.1109/CVPR.2008.4587417>. URL <http://dx.doi.org/10.1109/CVPR.2008.4587417>.
- [7] Sanjiv Kumar and Martial Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *Advances in Neural Information Processing Systems*, 2003. URL <http://citeseer.ist.psu.edu/679063.html>.
- [8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning*, 2001.
- [9] Diane Larlus, Jakob Verbeek, and Frédéric Jurie. Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields. *International Journal of Computer Vision*, 2009. URL <http://lear.inrialpes.fr/pubs/2009/LVJ09/LVJ09.pdf>.

- [10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Conference on Computer Vision and Pattern Recognition*, 2006. doi: <http://dx.doi.org/10.1109/CVPR.2006.68>. URL <http://dx.doi.org/10.1109/CVPR.2006.68>.
- [11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, 2004.
- [12] Jitendra Malik, Serge Belongie, Thomas K. Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001. URL <http://citeseer.ist.psu.edu/malik01contour.html>.
- [13] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004. doi: 10.1109/TPAMI.2004.1273918. URL <http://dx.doi.org/10.1109/TPAMI.2004.1273918>.
- [14] Giuseppe Passino, Ioannis Patras, and Ebroul Izquierdo. Aspect coherence for graph-based image labelling. In *International Conference on Visual Information Engineering*, July 2008. URL [http://www.citeulike.org/pdf/user/zeppe/article/3171472/passino\\_08\\_aspect.pdf](http://www.citeulike.org/pdf/user/zeppe/article/3171472/passino_08_aspect.pdf).
- [15] I. Patras, E. A. Hendriks, and R. L. Lagendijk. Video segmentation by map labeling of watershed segments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):326–332, 2001. doi: 10.1109/34.910886. URL <http://dx.doi.org/10.1109/34.910886>.
- [16] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *2006 IEEE Conference on Computer Vision and Pattern Recognition*, 2006. doi: <http://dx.doi.org/10.1109/CVPR.2006.326>. URL <http://dx.doi.org/10.1109/CVPR.2006.326>.
- [17] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object class segmentation using random forests. In *British Machine Vision Conference*, 2008. URL [http://research.microsoft.com/pubs/72423/Criminisi\\_bmvc2008.pdf](http://research.microsoft.com/pubs/72423/Criminisi_bmvc2008.pdf).
- [18] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. URL <http://citeseer.ist.psu.edu/shi97normalized.html>.
- [19] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, 2006. URL [http://research.microsoft.com/~carrot/publications\\_files/paper\\_eccv06-Rec.pdf](http://research.microsoft.com/~carrot/publications_files/paper_eccv06-Rec.pdf).
- [20] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *2005 IEEE International Conference on Computer Vision*, volume 1, 2005. doi: <http://dx.doi.org/10.1109/ICCV.2005.77>. URL <http://dx.doi.org/10.1109/ICCV.2005.77>.

- [21] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1483–1489, 2008. doi: <http://dx.doi.org/10.1109/TPAMI.2008.105>. URL <http://dx.doi.org/10.1109/TPAMI.2008.105>.
- [22] Joost van de Weijer and Cordelia Schmid. Coloring local feature extraction. In *European Conference on Computer Vision*, volume Part II, 2006. URL <http://lear.inrialpes.fr/pubs/2006/VS06>.
- [23] Jakob Verbeek and Bill Triggs. Region classification with markov field aspect models. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007. doi: 10.1109/CVPR.2007.383098. URL <http://dx.doi.org/10.1109/CVPR.2007.383098>.
- [24] Jakob Verbeek and Bill Triggs. Scene segmentation with CRFs learned from partially labeled images. In *Advances in Neural Information Processing Systems*, 2007. URL <http://lear.inrialpes.fr/pubs/2007/VT07a>.