# Hierarchical Image-Region Labeling via Structured Learning

Julian McAuley[1,2]
julian.mcauley@nicta.com.au
Teofilo de Campos[1,3]
t.decampos@st-annes.oxon.org
Gabriela Csurka[1]
gabriela.csurka@xrce.xerox.com
Florent Perronnin[1]
florent.perronnin@xrce.xerox.com

[1] Xerox Research Centre Europe
Meylan, France
[2] Australian National University/NICTA
Canberra, Australia
[3] University of Surrey
Guildford, United Kingdom

(all authors were at Xerox at the time of this work)

When classifying and segmenting images, some categories may be possible to identify based on global properties of the image (i.e., features extracted at a large scale), whereas others will depend on highly local information (i.e., features extracted at a fine scale); in some cases, these features may be very different, and may even provide conflicting information. In this paper, we aim to address this issue by using a graphical model which combines features extracted at multiple scales, and *learns* the extent to which the features provide consistent information across scales.

The nodes of our graphical model are shown in Figure 1. In this image, features are extracted at four scales, with the scale halving at each level. Edges in our graphical model are formed by connecting nodes at different scales: we connect two nodes precisely when the corresponding image regions overlap at two adjacent scales, so that our graphical model is tree-structured (in this case, a quad-tree).

Our image features are based on those from [1], in which image-level, region-level, and patch-level classifiers are proposed. We enforce that the labeling given to the image regions is consistent across scales, i.e., a region labeled as 'cat' at one scale should not be labeled as 'dog' at another scale. We enforce these constraints using the hierarchy shown in Figure 2, though in principle more complex hierarchies could be used.

The tree-structured nature of our model allows us to perform efficient and exact inference using *max-sum belief propagation*. Our method has running time and memory requirements of $O(|\mathscr{M}||\mathscr{H}|^2)$, where $\mathscr{M}$ is the set of image regions, and $\mathscr{H}$ is the set of classes. This gives our model an advantage over grid-structured models (for example), in which inference is typically approximate. A consequence of using a tree-structured model is that we are no longer directly enforcing neighbourhood constraints; instead, these are indirectly enforced through our hierarchy.

We train our method using *structured learning*, in the framework of [4]. This requires only that we are able to solve the inference problem, and that our *loss function* decomposes over the cliques in our model (i.e., the edges); this is certainly true of the Hamming loss, i.e., the proportion of incorrectly labelled regions.

Our method is evaluated on the PASCAL VOC2007 and VOC2008 datasets [2, 3] (the 2008 data was used for training, the 2007 data for testing). In Table 1 we show the performance of our method compared to a baseline (which is based on the work of [1]), and a non-learning approach (which assigns equal weights to all features). In Table 2, we see the contribution to the performance made at each image level. Our method seems to give the most substantial benefit at higher levels (i.e., smaller scales); in contrast, first-order methods become unreliable when using highly local features. We see that if structured learning is applied, we benefit substantially from the use of hierarchical constraints.

[1] G. Csurka and F. Perronnin. A simple high performance approach to semantic segmentation. In *BMVC*, 2008.

[2] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[3] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, 2008.

[4] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Predicting Structured Data*, pages 823–830, 2004.

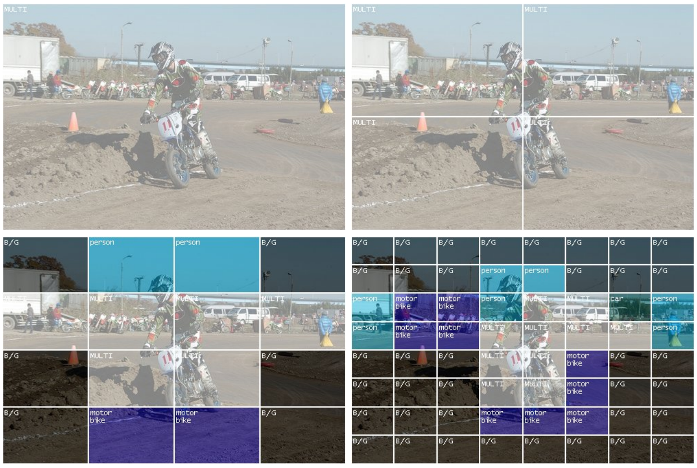
MULTI
MULTI MULTI
MULTI MULTI
B/G person person B/G
MULTI MULTI MULTI MULTI
B/G MULTI MULTI B/G
B/G motor bike motor bike B/G
B/G B/G B/G B/G B/G B/G B/G B/G
B/G B/G B/G person person B/G B/G B/G
person motor bike motor bike person MULTI MULTI car person
person motor bike motor bike MULTI MULTI MULTI MULTI person
B/G B/G B/G MULTI MULTI motor bike B/G B/G
B/G B/G B/G MULTI MULTI motor bike B/G B/G
B/G B/G B/G motor bike motor bike motor bike B/G B/G
B/G B/G B/G B/G B/G B/G B/G B/G

Figure 1: Nodes on the first four levels of our model. The correct labels (used for training our model) are shown on the image using different colours.

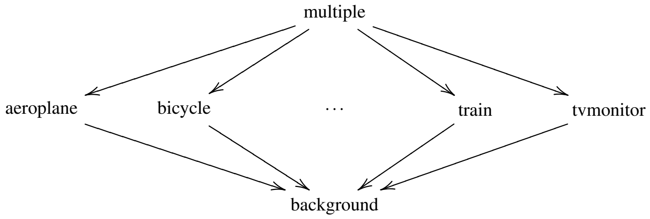

Figure 2: The hierarchy used in our paper. At the top of the hierarchy, we simply state that a region may contain multiple classes; at the bottom we state that a region does not contain any class. Class labels are those from [2, 3].

| | Training | Validation | Testing |
|---|---|---|---|
| Baseline (see [1]) | 0.272 (0.004) | 0.273 (0.004) | 0.233 (0.003) |
| Non-learning | 0.235 (0.006) | 0.224 (0.005) | 0.233 (0.004) |
| Learning | **0.460** (**0.006**) | **0.456** (**0.006**) | **0.374** (**0.004**) |

Table 1: The performance of our method on the training and validation datasets (VOC2008), and on the testing dataset (VOC2007).

| | Level 0 | Level 1 | Level 2 | Level 3 |
|---|---|---|---|---|
| Baseline (see [1]) | 0.342 | 0.214 | 0.232 | 0.163 |
| Non-learning | **0.426** | 0.272 | 0.137 | 0.112 |
| Learning | 0.413 | **0.307** | **0.349** | **0.444** |

Table 2: The contribution to the performance made at each image level (on the test set).