Int J Comput Vis (2012) 96:83–102
DOI 10.1007/s11263-011-0449-8

Harmony Potentials

Fusing Global and Local Scale for Semantic Image Segmentation

Xavier Boix · Josep M. Gonfaus · Joost van de Weijer · Andrew D. Bagdanov · Joan Serrat · Jordi Gonzàlez

Received: 11 November 2010 / Accepted: 4 April 2011 / Published online: 23 April 2011

Abstract The Hierarchical Conditional Random Field (HCRF) model have been successfully applied to a number of image labeling problems, including image segmentation. However, existing HCRF models of image segmentation do not allow multiple classes to be assigned to a single region, which limits their ability to incorporate contextual information across multiple scales. At higher scales in the image, this representation yields an oversimplified model since multiple classes can be reasonably expected to appear within large regions. This simplified model particularly limits the impact of information at higher scales. Since class-label information at these scales is usually more reliable than at lower, noisier scales, neglecting this information is undesirable. To address these issues, we propose a new consistency potential for image labeling problems, which we call the *harmony potential*. It can encode any possible combination of labels, penalizing only unlikely combinations of classes. We also propose an effective sampling strategy over this expanded label set that renders tractable the underlying optimization problem. Our approach obtains state-of-the-art results on two challenging, standard benchmark datasets for semantic image segmentation: PASCAL VOC 2010, and MSRC-21.

Keywords Semantic object segmentation · Hierarchical conditional random fields

Both authors contributed equally to this work.

X. Boix (✉) · J.M. Gonfaus · J. van de Weijer · A.D. Bagdanov · J. Serrat · J. Gonzàlez
Centre de Visió per Computador, Barcelona, Spain
e-mail: boixbosch@vision.ee.ethz.ch

J.M. Gonfaus
e-mail: gonfaus@cvc.uab.cat

J.M. Gonfaus · J. van de Weijer · J. Serrat · J. Gonzàlez
Department of Computer Science, Universitat Autònoma de Barcelona, Barcelona, Spain

X. Boix
Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

1 Introduction

Semantic image segmentation aims to assign predefined class labels to every pixel in an image, and is a crucial step for automatic understanding of an image. Image segmentation belongs to the general class of labeling problems, some of which, like image classification and stereo vision, date back to the early days of computer vision. Image segmentation is highly under-constrained, and state-of-the-art approaches focus on exploiting contextual information available around each pixel and at different scales of the image. One of the recent trends in semantic image segmentation is the use of Conditional Random Field (CRF) models with consistency potentials, which are able to cast the semantic segmentation task as an energy minimization problem over pixel or superpixel labelings. Continuing along these lines, we show in this article that the CRF model, when equipped with a new consistency potential which we call the *harmony potential*, can be used to efficiently fuse contextual information at the global and local context scales.

It is well known that context plays an important role for the recognition of objects in human vision (Oliva and Torralba 2007). The classification of an image region ignoring its context, and focusing only on the information within the object boundaries, is often an impossible task. The global context provides an important cue in the recognition of the objects, probably even more important than the objects

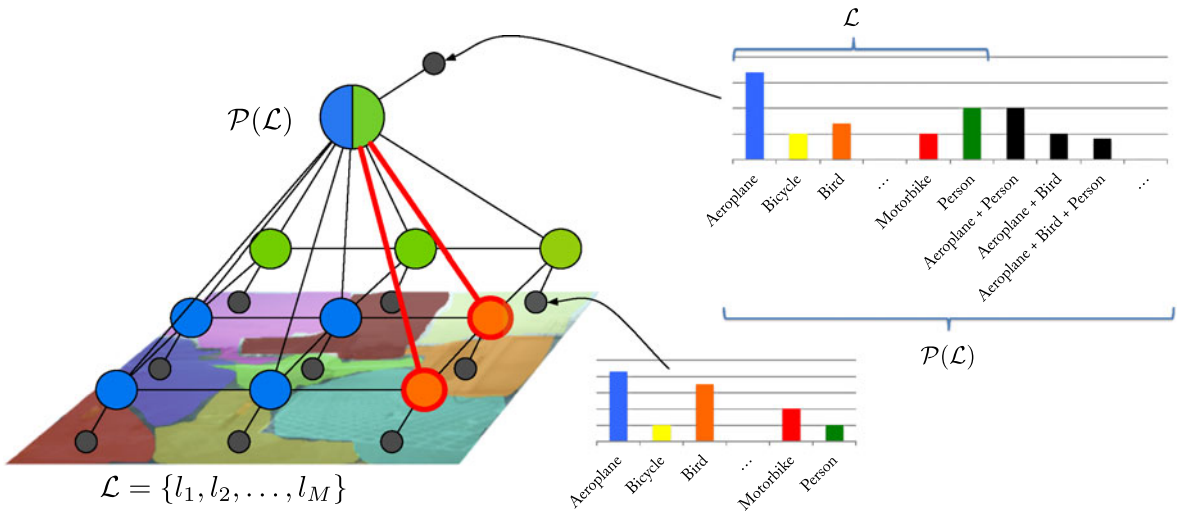


Fig. 1 Overview of our method. Illustration of the HCRF applied to image segmentation. Local nodes represent the random variables over superpixel labels, which take values from the set of class labels $\mathcal{L}$. Local nodes are connected when their superpixels share a boundary. The global node is a random variable over $\mathcal{P}(\mathcal{L})$, the power set of $\mathcal{L}$, which allows it to take any possible combination of the class labels as its label. The global node represents the classification of the whole image into semantic categories. Harmony potentials connect the global node to all local nodes

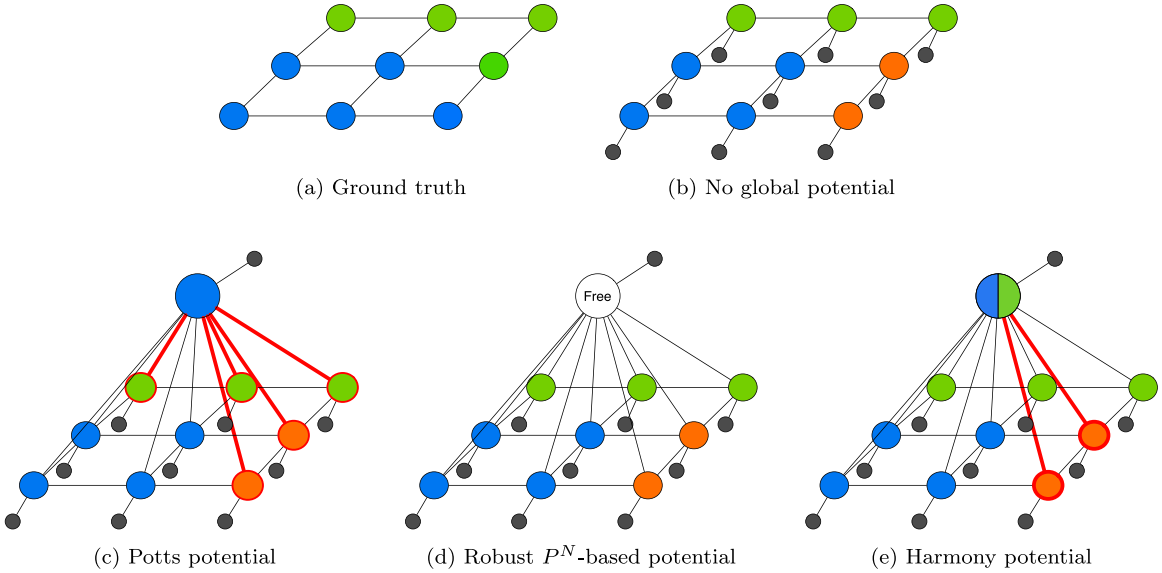
themselves. In a living room one expects sofas, lamps, tables, chairs, but not airplanes or trains.

Predicting the presence of a certain kind of objects based on the global image scale has been intensively studied in the field of image classification (Zhang et al. 2007; Lazebnik et al. 2006; van de Sande et al. 2010; Csurka and Perronnin 2010; Shahbaz et al. 2009). The image is generally represented by histograms over visual words, which are further enriched to incorporate, for example, spatial relationships (Lazebnik et al. 2006). These works use features of both objects and context to infer the presence of objects. Though local regions may also be described by a bag-of-words over local features such as color, texture or shape, the more complex representations that have considerably improved image classification performance cannot be expected to improve local region classification. The reason is that these regions lack of the complexity encountered at larger scales. Therefore, in contrast to existing CRF-based methods (Plath et al. 2009; Verbeek and Triggs 2008), we propose to adapt the classification method to the scale of the region. In particular, we use methods investigated by the image classification community to improve classification at the global scale in order to improve classification at the local scale of superpixels.

CRFs are theoretically sound models for combining information at multiple scales (Shotton et al. 2009; Kumar and Hebert 2005). A smoothness potential between neighboring nodes models the dependencies between the class labels of regions. However, since nodes at the lowest scale often represent small regions in the image, labels based only on their observations can be very noisy. Often, the final effect of such CRFs is merely a smoothing of local predictions. To overcome this problem, hierarchical CRFs have been proposed in which lower level nodes describe the class label configuration of the smaller regions (Plath et al. 2009; Kohli et al. 2009b; Zhu et al. 2008). One of the main advantages of this approach is that the higher-level context is based on larger regions, and hence can lead to more accurate estimations.

A drawback of existing hierarchical models is that to make them tractable they are often oversimplified by limiting regions to take just a single label (Plath et al. 2009), or in a more recent paper, an additional "free label" which basically cancels the information obtained at larger scales (Kohli et al. 2009b; Ladicky et al. 2009). Even though these models might be valid for scales close to the pixel level, they do not model very well the higher scales, much less the global scale. At the highest scales, far away from pixels, they impose a rather unrealistic model since multiple classes often appear together. The "free label" approach does not overcome this drawback because it does not constrain the combinations of classes which are not likely to appear simultaneously in one image. To summarize: the requirement to obtain tractable CRF models has led to oversimplified models of images, models which do not properly represent real-world images.

In this paper, we also adopt the hierarchical CRF framework but improve it by focusing on the crucial issue of how to efficiently represent and combine information at various scales. Our model is a two-level CRF that uses labels, features and classifiers appropriate to each scale. Figure 1 gives an overview of our approach to semantic image segmen-



(a) Ground truth (b) No global potential

(c) Potts potential (d) Robust P^N-based potential (e) Harmony potential

Fig. 2 (Color online) Example of the penalization behavior of different models for a labeling problem with labels {*blue*, *green*, *orange*}, where (**a**) is the ground-truth. (**b**) Without consistency potentials only the smoothness potential penalizes discontinuities in the labeling. (**c**) The Potts consistency potential adds an extra penalization (indicated in *red*) for each label different from the global node. (**d**) The Robust P^N-based potential, when the global node takes the "free label", does not penalize any combination of labels. (**e**) The harmony potential, which allows combinations of labels in the global node, correctly penalizes the orange labeling if the global node takes label {*blue*, *green*}

tation. It shows how consistency potentials can be defined to effectively relate semantic context in an image with local observations. The lowest level nodes represent superpixels labeled with single labels, while a global node on top of them constrains possible combinations of primitive local node labels below (Fig. 2e). A new consistency potential, which we term the harmony potential, is introduced and enforces consistency of local label assignment with the label of the global node. We propose an effective sampling strategy for global node labels that renders tractable the underlying optimization problem. Experiments yield state-of-the-art results for object class image segmentation on two challenging datasets: PASCAL VOC 2010 and MSRC-21.

In the next section we review the existing literature on semantic image segmentation. Section 3 describes the common framework for context-based probabilistic labeling. Then, in Sects. 4 and 5 we introduce a new type of a consistency potential: the harmony potential. Section 6 then specializes this framework for the problem of object segmentation and image classification by defining the concrete unary, smoothness and consistency potentials we use. In Sect. 7 we present results, and finally we draw some conclusions in Sect. 8.

2 Related Work

Image segmentation enjoys a long history as one of the mainstream topics of research in the computer vision community. It has long been approached as a bottom-up process based on low-level image features such as color, texture, and edge-detection (Marr 1982; Tu and Zhu 2002; Martin et al. 2004). In evaluation against human segmentation of images, acceptable results can be obtained (Martin et al. 2001), but common consensus is that for further improvement top-down semantic information is needed.

Advances in object recognition (Schmid and Mohr 1997; Lowe 2004; Sivic and Zisserman 2003) allowed for the recognition of semantic classes in images to aid image segmentation. Early works incorporating top-down information include (Mori et al. 2004) which combine segmentation and recognition, and the work on image parsing pioneered by the early work of Tu and Zhu (2002) and continuing with (Chen et al. 2005; Zhu et al. 2008). The image parsing approach, in general, uses a generative model of image formation and segments an image by decomposing it into its constituent patterns represented as a hierarchical parse tree. The tree of constituent patterns that maximizes a posterior is selected as the final image segmentation. These developments gave birth to the field of semantic segmentation where the goal is to both segment the image and classify pixels into a set of predefined semantic categories.

In this section, we discuss the most relevant recent approaches and classify them according to the scale of the context on which the segmentation is based. We distinguish three levels of scale. Firstly, the local scale is defined by a local patch or superpixel, usually obtained from an oversegmentation of the image. Secondly, the mid-level scale con-

sists of a neighborhood of patches or superpixels. We also consider as mid-level scale the outputs of sliding-window approaches as used in object detection, since they typically consist of multiple superpixels. Finally, the global scale is the entire image, which enables us to incorporate more sophisticated context. Approaches like our method, which are based on graphical models that enforce global consistency, will not be discussed here, but rather will be discussed in relation to our work in Sect. 3.

2.1 Local Scale

Bottom-up image segmentation methods try to label each pixel with the most likely class relying only on local information (Shotton et al. 2009; Yang et al. 2007; Pantofaru et al. 2008; Jiang and Tu 2009; Fulkerson et al. 2009). These methods tend to yield rough and noisy object segmentations, since many ambiguities are still present in the local observations. However, these methods are well suited for classes for which shape is not informative, which are better described by the local textures. These classes are referred to as *stuff* classes (Adelson 2001).

Since pixels alone are often not informative enough, one needs to consider a patch around them, which is described by multiple features. Typically, shape features such as SIFT (Lowe 2004), color features like local color histograms, and texture features like LBPs (Ojala et al. 2002) are used as local descriptors. Due to redundancy at the pixel level and for computational efficiency, a common approach is to sample randomly or in a regular grid from all possible locations, rather than representing features at the pixel level (Nowak et al. 2006). The main drawback of such approaches is that the image is partitioned in a uniform way, whereas natural images usually are not.

A solution to this problem is to use an initial unsupervised segmentation algorithm like (Felzenszwalb and Huttenlocher 2004; Comaniciu and Meer 2002; Vedaldi and Soatto 2008; Vazquez et al. 2011). This enables us to construct the low-level partitions of an image using a superpixel-based approach, which minimizes the risk of containing more than one object in a single superpixel (Fulkerson et al. 2009; Jiang and Tu 2009). Since unsupervised image segmentation is known to be unstable, Pantofaru et al. (2008) proposed combining several bottom-up segmentations. Fulkerson et al. (2009) investigated the benefits of using superpixels and conclude that they have lower computational requirements, provide coherent regions on which to obtain feature statistics, and preserve object boundaries.

2.2 Mid-level Scale

Mid-level scale is usually exploited in the form of object detection, hierarchical segmentation and enlarged local regions. It is usually used by top-down object segmentation approaches, which use the mid-level context scale to disambiguate local predictions and, in contrast to bottom-up approaches, they use *a priori* knowledge about the whole object such as its structure (Levin and Weiss 2009). They incorporate global object properties, like shape masks or histograms of oriented gradients (Yang et al. 2010; Leibe et al. 2008; Winn and Jojic 2005; Kumar et al. 2005; Lempitsky et al. 2009; Carreira and Sminchisescu 2010). However, since they rely on the global appearance of the object, occluded and less salient objects become more difficult to segment.

Several approaches are built upon the bounding boxes obtained from a detection method (Lempitsky et al. 2009; Gould et al. 2009; Ladicky et al. 2010b). For instance, Yang et al. (2010) merge several object detections by layering the scene, and infers which object is in front of the other. Since it can be understood as a refinement of detection methods, its performance remains bounded by the detection accuracy.

Other approaches incorporate the structure of object parts. In Leibe et al. (2008), the relative part location is determined by using a codebook and the generalized Hough transform, and Kumar et al. (2005) cast the problem as an energy minimization over a set of predefined parts and their relative locations. In Winn and Jojic (2005), an unsupervised procedure is able to segment an object class using a learned class mask and a deformation field. Also using an unsupervised procedure, Carreira and Sminchisescu (2010) select the most plausible figure-ground hypotheses and combine them in a later stage (Li et al. 2010).

Other works apply a coarse-to-fine approach based on a hierarchical representation (Zhu et al. 2008; Lim et al. 2009; Ladicky et al. 2009). The main strength of these methods is their ability to encode the context of a region, but they usually fail when background classes are not labeled in the training data since the semantic context can not be retrieved.

In our method, we apply mid-level scale information to improve the classification of superpixels. This is done by enriching the superpixel description with information about its neighbors. We use the object detection of Felzenszwalb et al. (2010) as an additional mid-level cue to improve superpixel classification.

2.3 Global Scale and Context

Global-scale information as used in image classification is often sufficient to determine the presence or absence of an object in a scene. Often, these methods rely more on contextual features rather than the object itself. The composition of the image can reveal the plausibility that an object does or does not appear in the image. Some segmentation algorithms use this information without taking into account its reliability, and only consider in the image the detected objects (Csurka and Perronnin 2010; Plath et al. 2009), or reweight the local predictions like in Shotton et al. (2008).

Several authors have noted the importance of context to obtain good classification (Oliva and Torralba 2007; Galleguillos and Belongie 2010). Context can be any information that is not directly produced by the appearance of an object. As stated in Oliva and Torralba (2007), in many cases the local appearance of an image is not enough to correctly classify the object class, and context plays an important role in disambiguating it. For example, the notion of semantic co-occurrence is shown to be helpful in the CRF formulation of Rabinovich et al. (2007). Closely related to our previous approach (Gonfaus et al. 2010) is the recent work of Ladicky et al. (2010a), where the co-occurrence statistics are incorporated directly into the graph cut inference procedure. To do so, it uses the principle of parsimony, which for similar likely solutions chooses the solution with fewer labels. Similarly, the model by Delong et al. (2010) penalizes over the quantity of different labels present in the image but without taking into account any co-occurence statistics. In contrast to these works, we adapt the representation to the context scale and use more sophisticated global classifiers rather than semantic co-ocurrence. We show that this greatly improves the results (see Sect. 7).

Another way of exploiting global image information is by inferring 3D scene geometry to discover where objects are likely to appear and how big they can be (Hoiem et al. 2007; Hoiem et al. 2008; Munoz et al. 2009). Splitting the image into regions allows the design of more sophisticated relations within the classes in an image. For example, based on confident familiar detections, other objects can be discovered (Lee and Grauman 2010), or inter-class relations can be learned in Jain et al. (2010), or hierarchical models can be approximated by sequentially fitting simple two-level models in a coarse-to-fine manner (Munoz et al. 2010).

As discussed in the introduction, we use image classification to provide global-scale information. We also learn the co-occurrence of classes from the training data and incorporate all of these cues into a hierarchical CRF model. In the next section we introduce the labeling problem as MAP estimation in preparation for the definition of the harmony potential in Sect. 4.

3 Labeling as MAP Estimation in Graphical Models

We present a model for labeling problems that jointly uses global and local scales and introduce the existing labeling approaches that use this same idea (Plath et al. 2009; Ladicky et al. 2009; Kumar et al. 2005). We show the different ways they define the relationship between the local and global context scales.

3.1 Hierarchical CRFs for Labeling

Graphical models are sound representations of joint probability distributions (Lauritzen 1996; Wainwright and Jordan 2008). A graphical model uses a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent a probabilistic model composed of a set $\mathbf{X} = \{X_i\}_{i\in\mathcal{V}}$ of random variables, each of which corresponds to a node in the graph. Each node is indexed with an element of the set $\mathcal{V} = \{1, 2, \ldots, N\}$. We use $\mathbf{x} = \{x_i\}_{i\in\mathcal{V}}$ to denote a possible state or instantiation of $\mathbf{X}$. That is, $\mathbf{x} = \{x_i\}_{i\in\mathcal{V}}$ represents hypothetical assignment of value x_i to random variable X_i in $\mathbf{X}$. In this paper, we only consider undirected graphical models, and represent the edges of the graph with the set $\mathcal{E}$ of tuples (i, j), where $i, j \in \mathcal{V}$. The edges define a set of conditional independence assumptions, where each edge represents the compatibility between the nodes it connects, and for which the Markov property holds:

$$P(X_i = x_i|\mathbf{X}_{\{j\neq i\}}) = P(X_i = x_i|\mathbf{X}_{\{j|(i,j)\in\mathcal{E}\}}). \tag{1}$$

These models are called Markov Random Fields (MRF), or Conditional Random Fields (CRF) when compatibility between nodes is conditioned on some measurement.

A clique is a subgraph in which every node is connected to all other nodes in the subgraph. Let $\mathcal{C}$ represent the set of cliques that are not a subset of any other clique. These are known as *maximal cliques*, and according to the Hammersley-Clifford theorem (Hammersley and Clifford 1971) the probability that $\mathbf{X}$ takes value $\mathbf{x}$ in a CRF, conditioned on $\mathbf{O}$, follows a Gibbs distribution:

$$P(\mathbf{X} = \mathbf{x}|\mathbf{O}) = \frac{1}{Z}\prod_{c\in\mathcal{C}} e^{-\varphi_c(\mathbf{x}_c)}, \tag{2}$$

where φ_c is the compatibility function or potential of a clique $c \in \mathcal{C}$, and $\mathbf{x}_c = \{x_i\}_{i\in c}$ is the state $\mathbf{x}$ restricted to the nodes in clique $c \in \mathcal{C}$. For the sake of simplicity, we do not explicitly indicate the dependence of φ_c on $\mathbf{O}$. The potential functions $\varphi_c(\mathbf{x}_c)$ do not have a probabilistic interpretation, but encode *a priori* knowledge about random variables in a clique. $Z = \sum_{\mathbf{x}}\prod_{c\in\mathcal{C}} e^{-\varphi_c(\mathbf{x}_c)}$, called the partition function, is a normalization constant whose exact computation is usually intractable. We define the energy of state $\mathbf{x}$ as

$$E(\mathbf{x}) = -\log P(\mathbf{X} = \mathbf{x}|\mathbf{O}) - \log Z = \sum_{c\in\mathcal{C}} \varphi_c(\mathbf{x}_c). \tag{3}$$

CRFs have been broadly used to model dependencies in labeling problems (Shotton et al. 2009; Kumar and Hebert 2005). The simplest and most common only involves the local context scale. Since nodes at the lowest scales often represent small regions in the image, labels based only on their observations can be very noisy. To reduce such noisy labeling, a smoothness potential between neighboring local nodes is defined to model the dependencies between regions. However, the final effect of such CRFs is merely a smoothing of local predictions. Li and Huttenlocher (2008) attempted to overcome this problem using a connectivity pattern with long range dependencies. Other authors use high-order cliques in the original connectivity pattern, and then

convert them into order two cliques by the introduction of new variables (Ramalingam et al. 2008; Rother et al. 2009; Ishikawa 2009; Kohli and Kumar 2010).

In addition to local scale, Hierarchical CRFs (HCRFs) are used for combining different scales of context (Plath et al. 2009; Ladicky et al. 2009; Zhu et al. 2008). This approach consists on building a hierarchy of variables on top of the graph. Higher level nodes describe the class-label configuration of larger image regions, while those lower in the hierarchy still describe local scale at the pixel or super-pixel level. One of the main advantages of these approaches is that higher level context is based on larger regions, and hence can lead to better estimations.

Our treatment of the HCRF formulation is limited to an instantiation of a graphical model $\mathcal{G}$ relating a global context scale with the local one. We designate a random variable as the global node and one for each local node. Thus, $\mathcal{V} = \mathcal{V}_G \cup \mathcal{V}_L$, where $\mathcal{V}_G = \{g\}$ is the index associated with the global node, and $\mathcal{V}_L = \{1, 2, \ldots, N\}$ are the indexes associated with each local node. All of these random variables take a discrete value from a set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$. Analogously, we define two subsets of edges: $\mathcal{E} = \mathcal{E}_G \cup \mathcal{E}_L$. The set of edges $\mathcal{E}_G$ contains edges connecting the global node X_g with each of the local nodes X_i, for $i \in \mathcal{V}_L$. The set of local edges $\mathcal{E}_L$ is the pairwise connections between local nodes.

The energy function of the graph $\mathcal{G}$ can be written as the sum of the unary, smoothness and consistency potentials, respectively:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi_i(x_i) + \sum_{(i,j) \in \mathcal{E}_L} \psi_{ij}^L(x_i, x_j) + \sum_{(i,g) \in \mathcal{E}_G} \psi_{ig}^G(x_i, x_g). \tag{4}$$

The unary term ϕ_i depends on a single probability $P(\mathbf{O}_i | X_i = x_i)$, where $\mathbf{O}_i$ is the observation that affects X_i in the model. The smoothness potential ψ_{ij}^L determines the pairwise relationship between two local nodes. It represents a penalization for two connected nodes having different labels, and usually depends also on an observation. The consistency potential ψ_{ig}^G expresses the dependency relationship between the labels of a local node and the global node.

Some authors used this graphical model $\mathcal{G}$ as a basic structure that is repeated recursively to form a larger, hierarchical graph (Plath et al. 2009; Ladicky et al. 2009). Doing so, mid-level context scale can be easily added to the model. However, the definition of the relationships between these context scales, i.e. the consistency potential, is an important issue that has to be clarified. Before introducing our framework, we first review existing consistency potentials applied to image labeling problems.

3.2 Existing Consistency Potentials

In the following we review the Potts and the robust P^N-based consistency potentials, which have been used in a HCRF for labeling problems. In Sect. 4, we extend these potentials to a new one that we call harmony potential. Figure 2 briefly illustrates the characteristics of the different models compared in this paper.

3.2.1 Potts Potential

In the basic graph used to build the tree structured model by Plath et al. (2009) the consistency potential is defined as a Potts model:

$$\psi_{ig}^G(x_i, x_g) = \gamma_i(x_i) \mathrm{T}[x_i \neq x_g], \tag{5}$$

where $\mathrm{T}[\cdot]$ is the indicator function and $\gamma_i(x_i)$ is the cost of labeling $x_i \in \mathcal{L}$. Since this potential encourages assigning the same label as the global node to all the local nodes, this potential is unable to support any kind of heterogeneity in the region below the global node.

3.2.2 Robust P^N-Based Potential

In this case, the global node has an extended label set $\mathcal{L}^E = \mathcal{L} \cup \{l_F\}$, where l_F stands for a "free label". This special label means that any possible label in $\mathcal{L}$ can be assigned to local nodes without any cost. Thus, the potential becomes

$$\psi_{ig}^G(x_i, x_g) = \begin{cases} 0 & \text{if } x_g = l_F \text{ or } x_g = x_i, \\ \gamma_i(x_i) & \text{otherwise.} \end{cases} \tag{6}$$

The model is recursively used to build up a hierarchical graph for object segmentation (Ladicky et al. 2009), and inference can be achieved using graph cuts (Russell et al. 2010).

This potential enforces labeling consistency when the vast majority of local nodes have the same label and, unlike the Potts model, does not force a certain labeling when the solution is heterogeneous. However, in the heterogeneous case, not applying any penalization is not always the best decision. When a particular subset of labels $\ell \subset \mathcal{L}$ appears in the ground-truth and $x_g = l_F$, the robust P^N-based potential will not penalize any assigned label not in the subset ℓ.

This potential is equivalent to the high-order robust P^N potential previously introduced by Kohli et al. (2009b) and is an extension of the P^N Potts potential (Kohli et al. 2009a). The P^N Potts potential is a high order potential that, rather than adding a penalization for each mislabeling as in (6), penalizes a constant value when all nodes do not take the same label.

4 The Harmony Potential

The main drawback of existing consistency potentials is that to make inference tractable they usually must be oversimplified by allowing regions to have just a single class label (Potts), or adding a "free label" which basically cancels the information obtained at the larger scales (Robust P^N-based). At the highest scales, far away from pixels, they impose a rather unrealistic model since multiple classes appear together. The requirement to obtain tractable inference has led to oversimplified HCRF models, that do not properly represent larger context scales.

The harmony potential generalizes the robust P^N-based potential, which is itself a generalization of the Potts potential. As in music harmony describes pleasant combinations of tones when played simultaneously, here we employ this term to describe likely combinations of labels. In this section we formally define the harmony potential, show how it is a natural generalization of the robust P^N-based potential, and its equivalence to a high order graphical model.

4.1 Definition

Let $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$ denote the set of class labels from which local nodes X_i take their labels. The global node X_g, instead of taking labels from this same set, will draw labels from $\mathcal{P}(\mathcal{L})$, the *power set* of $\mathcal{L}$. In this context, the power set represents all possible combinations of primitive labels from $\mathcal{L}$. This expanded representation capability is what gives the harmony potential its power, although its cardinality $2^{|\mathcal{L}|}$ renders most optimization problems over the entire label set for the global node. In the sequel, we propose a ranked sub-sampling strategy that effectively reduces the size of the label set that must be considered.

$\mathcal{P}(\mathcal{L})$ is able to encode any combination of local node labels, and the harmony potential subsequently establishes a penalty for local node labels not encoded in the label of the global node. The harmony potential is simply defined as:

$$\psi_{ig}^{G}(x_i, x_g) = \gamma_i(x_i)\mathrm{T}[x_i \notin x_g]. \tag{7}$$

Note that $\psi_{ig}^{G}(x_i, x_g)$ penalizes when x_i is not encoded in x_g, but not when a particular label in x_g does not appear in the x_i.

Analyzing the definition of the robust P^N-based potential in (6), we see that l_F is essentially a "wildcard" label that represents *any possible label* from $\mathcal{L}$. Setting $x_g = \mathcal{L} \in \mathcal{P}(\mathcal{L})$ in the harmony potential in (7) similarly applies no penalty to any combination of local node labels, since $l \in x_g = \mathcal{L}$ for *any* local label l. In this way the harmony potential generalizes the robust P^N-based potential by admitting wildcard labels at the global node, while also allowing concrete and heterogeneous label combinations to be enforced by the global node.

The incorporation of global information through the harmony potential is novel with respect to existing techniques exploiting image-level priors such as Shotton et al. (2008). While such techniques rely on global information, our probabilistic framework incorporates the uncertainty of X_g with the selected labels of local nodes in a joint-probabilistic manner. The harmony potential intrinsically handles the heterogeneity of the labeling problem, mainly because the label set of the global node is the power set of local node labels. We can observe in (7) how, unlike the P^N-based potential, the harmony potential is able to distinguish between combinations of labels and to apply a different penalization according to the compatibility of these combinations.

4.2 Equivalence to a High Order Model

High order graphical models are able to encode complex dependencies between sets of random variables. Models with high-order potentials have been successfully applied in applications ranging from image denoising (Roth and Black 2009) and stereo vision (Woodford et al. 2009) to labeling problems (Kohli et al. 2009a). However, it is not always possible to infer a satisfactory MAP configuration because of the complexity of the model. More expressive potentials are needed but without sacrificing the reliability of MAP inference.

Recently, several authors pointed out that some high-order potentials can be transformed into pairwise models by extending them with extra random variables (Ramalingam et al. 2008; Rother et al. 2009; Ishikawa 2009; Kohli and Kumar 2010). Following this idea, it can be shown that the harmony potential is in fact equivalent to a high-order model.

Let $\psi^H(\mathbf{x}_L)$ be a high-order potential that encodes a dependency between all local nodes and the global scale observation $\mathbf{O}_g$. $\mathbf{x}_L$ is the set of local nodes labels $\{x_i\}_{i\in\mathcal{V}_L}$. We define a new graphical model $\mathcal{G}_H$ from $\mathcal{G}$, where we substitute all harmony potentials and the global random variable X_g by the high-order potential ψ^H. This gives rise to a model which has the following energy function

$$E_H(\mathbf{x}_L) = \sum_{i\in\mathcal{V}_L} \phi_i(x_i) + \sum_{(i,j)\in\mathcal{E}_L} \psi_{ij}^{L}(x_i, x_j) + \psi^H(\mathbf{x}_L). \tag{8}$$

Note that the model does not have a global random variable X_g, but takes into account the global scale observation $\mathbf{O}_g$ inside ψ^H.

According to the transformation proposed by Rother et al. (2009), the graphical models $\mathcal{G}_H$ and $\mathcal{G}$ are equivalent if the high-order potential ψ^H is defined as

$$\psi^H(\mathbf{x}_L) = \min_{\ell \in \mathcal{P}(\mathcal{L})} \left\{ \gamma_g(\ell) + \sum_{i \in \mathcal{V}_L} \gamma_i(x_i) \mathrm{T}[x_i \notin \ell] \right\}, \quad (9)$$

where $\gamma_g(\ell)$ is a constant that depends on the global scale observation $\mathbf{O}_g$. Note that what makes ψ^H a high-order potential is the minimum operation: it takes into account all random variables in order to choose which $\ell \in \mathcal{P}(\mathcal{L})$ minimizes the summation. The main idea behind this transformation is that the global node X_g is now encoded in ψ^H through the auxiliary variable ℓ. A proof of this is provided in Appendix A.

In the same way the harmony potential is expressed as a high-order clique, Ladicky et al. (2009) show that the pairwise robust P^N-based potential in (6) is equivalent to the high-order robust P^N potential defined by Kohli et al. (2009b), which is

$$\psi^H(\mathbf{x}_L) = \min_{l \in \mathcal{L}} \left\{ \gamma_g(l_F), \gamma_g(l) + \sum_{i \in \mathcal{V}_L} \gamma_i(x_i) \mathrm{T}[x_i \neq l], \right\}. \quad (10)$$

Here we can also observe that the high-order version of the harmony potential is a generalization of the high-order robust P^N potential. The harmony potential, as shown in (9), is the minimum value taken over the power set $\mathcal{P}(\mathcal{L})$, while in the robust P^N potential the minimum is only taken over $\gamma_g(l_F)$, that represents the wildcard label, and the values given by $\mathcal{L}$. This wildcard label is included in $\mathcal{P}(\mathcal{L})$, and hence in the minimization in (9) since $\mathcal{L} \in \mathcal{P}(\mathcal{L})$.

We have shown that the use of the power set $\mathcal{P}(\mathcal{L})$ as the label set for the global node is what gives more expressive power to the harmony potential. However, since in most interesting cases optimizing a problem with $2^{|\mathcal{L}|}$ possible labels is intractable, the harmony potential also makes inference into a challenging problem. In the next section we describe how to select the labels of the power set that are the most likely to appear in the optimal configuration.

5 Ranked Sampling of $\mathcal{P}(\mathcal{L})$

In the previous section we showed that the harmony potential can be used to specify which labels are likely to appear in the local nodes, and it also gives rise to a model with which we can infer the most probable combinations of local node labels. However, because the harmony potential is built using all combinations of labels, the excessive cardinality $2^{|\mathcal{L}|}$ of the label set renders exact inference infeasible. For models with variables on very large domains, inference is usually made possible by discarding labels (Freeman et al. 2000; Coughlan and Ferreira 2002) or sampling the label space (Ihler and McAllester 2009; Koller et al. 1999; Sudderth et al. 2002). Along these lines, we establish a ranking of subsets that prioritizes the optimization over the $2^{|\mathcal{L}|}$ possible labels for the global node, and then apply any suitable inference algorithm such as Loopy Belief Propagation (LBP) (Frey and MacKay 1998) or Graph Cuts (Boykov and Kolmogorov 2004). In this section, we focus on the selection of labels for the global node.

Optimizing for the best assignment of global label x_g^* implies maximizing $P(X_g = \ell|\mathbf{O})$, where $\ell \in \mathcal{P}(\mathcal{L})$. This is very difficult in practice due to the $2^{|\mathcal{L}|}$ possible labels and the lack of an analytic expression for $P(X_g = \ell|\mathbf{O})$. An approximation of this probability allows us to effectively rank possible global node labels, and thus to prioritize candidates in the search for the optimal label x_g^*. We pick the best $M' \leq 2^{|\mathcal{L}|}$ subsets of $\mathcal{L}$ that maximize an approximation of the posterior $P(X_g = \ell|\mathbf{O})$. This approximation establishes an order on subsets of the (unknown) optimal labeling of the global node x_g^* that guides the consideration of global labels. We may not be able to consider all labels in $\mathcal{P}(\mathcal{L})$ during inference, but at least we can consider the most likely candidates for the global nodes.

In the following subsections, we introduce a branch-and-bound algorithm that is used to sample $\mathcal{P}(\mathcal{L})$, and then the approximation of the posterior $P(X_g = \ell|\mathbf{O})$.

5.1 Branch-and-Bound Sampling

A branch-and-bound algorithm allows us to find an approximately optimal solution to the labeling problem without having to exhaustively search the whole space of image labellings. We require at this point a bounding strategy that discards large sets of candidate labels without pruning away any potentially optimal solutions. In Algorithm 1 we summarize a recursive branch-and-bound algorithm to do just that. It establishes a search tree where a label is built incrementally by increasing the number of considered semantic classes. At each level of the tree, an extra class is considered and a decision is made whether to encode it in the label or not. For instance, let $\ell'' \in \mathcal{P}(\mathcal{L}'')$ be a partially built label at the k-th level of the search tree, where $\mathcal{L}'' \subset \mathcal{L}$. After a branching to the $(k+1)$-th level, we take into consideration one extra class label l_{branch} to build $\ell' \in \mathcal{P}(\mathcal{L}')$, and consider the probability that this class is encoded in ℓ' or not. At the leaves of the search tree we obtain the labels in $\mathcal{P}(\mathcal{L})$ and all classes have been taken into account.

During the exploration of the tree, the algorithm maintains a set $\mathcal{S}$ of the $M' \leq 2^{|\mathcal{L}|}$ labels with the highest posterior $P(X_g = \ell|\mathbf{O})$. An upper bound $\gamma_{\ell'}$ of this posterior is evaluated for each partially built label $\ell' \in \mathcal{P}(\mathcal{L}')$. If the upper bound $\gamma_{\ell'}$ is lower than all the posteriors of the labels in the set $\mathcal{S}$, we can discard all labels below ℓ' in the tree. Since these pruned labels have a posterior lower or equal to the

```
function S=Branch&Bound(ℓ'', S, k)
    for ℓ' = {ℓ'', {ℓ'', l_branch}} do
        if ∃ℓ ∈ S : γ_ℓ' ≥ q(ℓ) then
            if k = |L| then
                ℓ' ↣ S
            else
                S = Branch&Bound(ℓ', S, k + 1);
            end
        end
    end
end
```

Algorithm 1: Branch-and-bound algorithm for selecting the M' labels with highest posterior $q(\ell) \propto P(X_g = \ell|\mathbf{O})$. The set $\mathcal{S}$ stores the best found labels

upper bound, we are sure that none of them has a posterior high enough to be selected. This pruning is what maintains tractable computational costs.

5.2 Approximating $P(X_g = \ell|\mathbf{O})$

We first decompose the posterior using Bayes rule,

$$P(X_g = \ell|\mathbf{O}) \propto P(X_g = \ell)P(\mathbf{O}|X_g = \ell). \tag{11}$$

This breaks the posterior into the prior and the likelihood, each of which are approximated separately.

We can approximate the prior $P(X_g = \ell)$ from the ground-truth of the training set $\mathcal{I}$: it is approximated by a histogram of the number of models where the set ℓ appears encoded in the ground-truth, i.e.

$$P(X_g = \ell) \propto \sum_{I_i \in \mathcal{I}} \mathrm{T}[\ell \subseteq t_g^i], \tag{12}$$

where t_g^i is the ground-truth label of the global node for the training image $I_i \in \mathcal{I}$. Note that this prior has the advantage that it incorporates semantic co-occurrence of classes: buses do not occur with televisions, though they do occur quite often with cars.

The high dimensionality of $\mathbf{O}$ makes the estimation of the likelihood $P(\mathbf{O}|X_g = \ell)$ very challenging. To overcome this problem, let $\mathbf{O}_g^{l_k}$ be $\mathbf{O}$ restricted to only those observations that influence the global node in the model and are specific for each encoded object class $l_k \in \mathcal{L}$. Thus, the likelihood can be approximated as

$$P(\mathbf{O}|X_g = \ell) \approx P(\{\mathbf{O}_g^{l_k}\}_{l_k \in \mathcal{L}}|X_g = \ell). \tag{13}$$

Note that it only takes only into account the observations of the global node individually, and discards any relationship between it and the other random variables. In order to facilitate the computation of this probability, we assume conditional independence among the global observations $\{\mathbf{O}_g^{l_k}\}_{l_k \in \mathcal{L}}$,

$$P(\{\mathbf{O}_g^{l_k}\}_{l_k \in \mathcal{L}}|X_g = \ell) = \prod_{k|l_k \notin \ell} P(\mathbf{O}_g^{l_k}|l_k \notin X_g) \prod_{k|l_k \in \ell} P(\mathbf{O}_g^{l_k}|l_k \in X_g) \tag{14}$$

$$\propto \prod_{k|l_k \notin \ell} P(l_k \notin X_g|\mathbf{O}_g^{l_k}) \prod_{k|l_k \in \ell} P(l_k \in X_g|\mathbf{O}_g^{l_k}), \tag{15}$$

where $P(l_k \notin X_g|\mathbf{O}_g^{l_k}) = 1 - P(l_k \in X_g|\mathbf{O}_g^{l_k})$. Note that (15) follows from the assumption that labels in $\mathcal{L}$ are equiprobable.

Because we are interested in ranking the labels, we approximate a quantity proportional to $P(X_g = \ell|\mathbf{O})$ rather than the probability itself. Denoting this quantity as $q(\ell)$ and using (12) and (15), $q(\ell)$ is defined as:

$$\sum_{I_i \in \mathcal{I}} \mathrm{T}[\ell \subseteq t_g^i] \prod_{k|l_k \notin \ell} P(l_k \notin X_g|\mathbf{O}_g^{l_k}) \prod_{k|l_k \in \ell} P(l_k \in X_g|\mathbf{O}_g^{l_k}). \tag{16}$$

For each partially built label $\ell' \in \mathcal{P}(\mathcal{L}')$ in the branch-and-bound search exploration, we need an upper bound $\gamma_{\ell'}$ of $q(\ell)$ for all possible labels ℓ built by branching from ℓ'. As mentioned before, this serves to prune all labels ℓ for which $\gamma_{\ell'}$ is smaller than the worst label in the list of solutions $\mathcal{S}$. It is easy to show that the quantity $q(\ell')$ is an upper bound of the labels build from itself (the proof is given in Appendix B), i.e.

$$\gamma_{\ell'} = q(\ell') \geq q(\ell). \tag{17}$$

This is because after branching from ℓ' and considering whether the label $l_k \in \mathcal{L}$ is present or not, neither decision can lead to an increase of the quantity $q(\ell')$. Note that this does not mean that the posterior $P(X_g = \ell|\mathbf{O})$ is necessarily lower when more single labels are present. $q(\ell')$ is computed using a partially built label ℓ', and only the subset of labels $\mathcal{L}' \subset \mathcal{L}$ are taken into account.

5.3 Effects of Sampling $\mathcal{P}(\mathcal{L})$

In order to validate our hypothesis about the impact of such sampling, we performed a simple experiment (see Sect. 7 for a detailed description of the datasets and implementation details used in all our experiments). We analyze the performance of the system for different numbers of sampled label combinations. Results are shown in Fig. 3 for the MSRC-21 and PASCAL datasets. The gain of adding label combinations is more significant for MSRC-21 since it is inherently more multiclass than the PASCAL dataset. Despite the fact that we cannot compare with the use of all possible combination of labels because it is computationally unfeasible, we

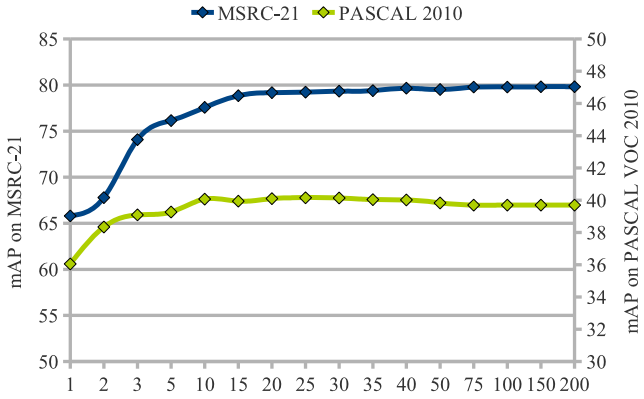


Fig. 3 Ranked sampling of $\mathcal{P}(\mathcal{L})$. Mean Average Precision (mAP) achieved by allowing more combinations of labels at the global node

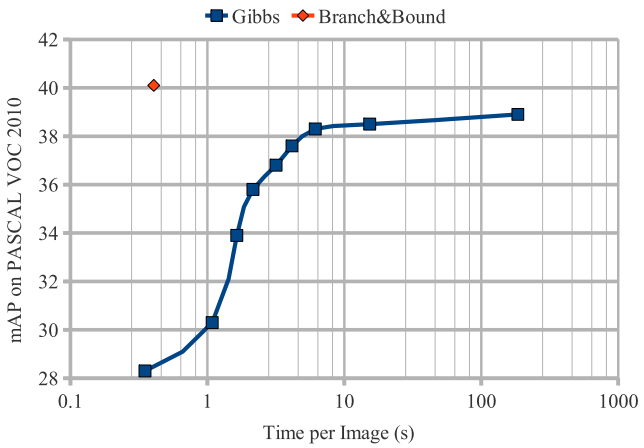


Fig. 4 Comparison to Gibbs sampling. Mean Average Precision (mAP) achieved by sampling $\mathcal{P}(\mathcal{L})$ with 50 labels against time required by Gibbs sampler to converge. Note that with our sampling, inference is only done once, while with Gibbs sampling inference is done at every iteration

observe that the performance quickly stabilizes after considering only a few combinations.

It is also important to note the poor performance of using just the best combination of labels. The reason for this is that a global classifier cannot always decisively identify the exact combination of true labels as the best combination over all of them. This shows that we cannot blindly rely on the best combination according to the global classifier, since we obtain far superior performance by considering more. Although these combinations are less likely from the global classifier point of view, they are more suitable from the point of view of our HCRF which jointly uses the global and local context scales.

As another experiment, Fig. 4 shows a comparison to the use of Gibbs sampling to select labels for the global node. By iteratively flipping one of the M labels on or off in the global label, one can infer a solution without the approximation used in our branch-and-bound algorithm. The results using Gibbs sampling eventually reach the performance achieved by our branch-and-bound method, but it is important to note that the number of Gibbs sampling iterations required to achieve this performance is, on average, more than 50 seconds per image. Our ranked sampling approach achieves state-of-the-art performance using only 50 labels for the global node and requires less then half a second to segment an image.

6 Fusing Local and Global Scales

In the previous section we described the structure of our HCRF. Now we address how to apply it to fuse information at local and global scales for semantic image segmentation. To illustrate the choices made in this section we will show results on the two datasets on which we will evaluate our method in Sect. 7: the PASCAL VOC 2010 Segmentation Challenge (Everingham et al. 2010) and the MSRC-21 dataset (Shotton et al. 2009).

In Fig. 1 we show an overview of the HCRF for image segmentation. The local nodes $\{X_i\}_{i\in\mathcal{V}_L}$ represent random variables over the semantic labelings of superpixels. We obtain the set of superpixels using an unsupervised segmentation method. Since all pixels inside a superpixel can take only a unique label, an oversegmentation of the image is required so that superpixels do not cross object boundaries. Regions are created by over-segmenting the image with the quick-shift algorithm (Vedaldi and Soatto 2008) using the same parameters as Fulkerson et al. (2009). By working directly on the superpixel level instead of the pixel level, the number of nodes in the CRF is significantly reduced, typically with an image of 500×300 pixels, the reduction goes from 150.000 to an average of 500 nodes per image. Therefore, the inference algorithm converges drastically faster.

The local nodes that share a boundary are connected with a smoothness potential, and the global node X_g represents the semantic classification of the whole image. That is, it expresses whether the image contains or not each of the semantic categories over which the segmentation problem is defined. It is connected by the harmony potential to each local node.

We differentiate between the unary potentials of the local nodes $\phi_i^L(x_i)$, where $i \in \mathcal{V}_L$, and the unary potential of the global node $\phi_g^G(x_g)$. This is because we adapt each potential to its scale. The larger scale of the global node allows us to use more sophisticated representations, such as spatial pyramids (Lazebnik et al. 2006), which are unsuitable at smaller scales. To improve classification accuracy at the local nodes we further extend their observations with mid-level scale information.

6.1 Local Unary Potential

The unary potential associated with local nodes is based only on information at the superpixel scale. At this level, the ambiguity that exists between classes leads to unreliable classification scores. To improve superpixel classification accuracy, we combine both local and mid-level information in the unary potential.

The superpixel descriptors are based on a bag-of-words over both appearance and color features. To benefit from context at the mid-level scale, we extend the representation at the local scale with mid-level context information. Fulkerson et al. (2009) showed that a combination of features extracted not only inside superpixels, but also in the area adjacent to them, better describes superpixels. We use two different bags-of-words: one for the superpixel and another for the regions adjacent to it. These are then concatenated to form the final feature representation of the superpixel. We found that this combination better describes and distinguishes object boundaries.

We use a variety of cues to represent superpixels, and we train one classifier for each of them. We denote by $s_i(k, x_i)$ the classification score for class label $x_i \in \mathcal{L}$ at node $i \in \mathcal{V}_L$ obtained using the cue indexed by $k \in \mathcal{F}$, where $\mathcal{F}$ is the set that indexes the cues. Thus, for each superpixel we have several classification scores, one for each cue and semantic class.

We compute the unary potential by weighting the classification scores $\{s_i(k, x_i)\}_{k\in\mathcal{F}}$ through a sigmoid function. The unary potential becomes:

$$\phi_i^L(x_i) = -\mu_L K_i \log \prod_{k\in\mathcal{F}} \frac{1}{1+\exp(f_i(k, x_i))}, \tag{18}$$

$$f_i(k, x_i) = a(k, x_i)s_i(k, x_i) + b(k, x_i), \tag{19}$$

where μ_L is the weighting factor of the local unary potential, K_i normalizes over the number of pixels inside the superpixel. We have two sigmoid parameters for each class/cue pair: $a(k, x_i)$ and $b(k, x_i)$. The usage of a sigmoid to convert classification scores into probabilities is common practice (Platt 1999). Here, we simultaneously learn all the sigmoids on a validation set.

We use four different cues, each describing different aspects of mid and low-level context scale. The different cues also exploit different training sets in order to discriminate between certain subsets of classes. An earlier version of our work (Gonfaus et al. 2010) was based only on the first of these cues. Our four cues are:

1. *Foreground-background classifier* (*FG-BG*): Object classifiers are generally trained to differentiate between objects from one class and objects from *any* other class. However, the harmony potential already takes care of penalizing the coexistence of objects from classes which are not likely to be in the image. Hence, the superpixel classifiers need not be so general, and can instead be specialized to discriminate between a specific object class and *only* those classes of objects which appear simultaneously in the same image. The FG-BG classifier is designed to discriminate objects from their own background, and thus, the negative examples of the training set are those superpixels in the same image not intersecting any instance of the object class.
2. *Object class against other objects* (*CLASS*): When several classes share similar backgrounds, such as cows and horses, or cats and dogs, the FG-BG classifier might lead to high probabilities for several foreground classes, and thus, it does not discriminate between classes. In this case, both classes are highly probable, but usually only one of them appears in the same image. In order to disambiguate these cases, the CLASS classifier is trained to discriminate between each class and all other object classes.
3. *Location* (*LOC*): We use the position of the superpixel as an additional cue. For instance, this cue allows us to learn that many objects tend to be in the center of the image, dining tables are often at the bottom, or sky is most likely to be at the top.
4. *Object detection* (*OBJ*): We incorporate object detection into the unary potentials to exploit another source of mid-level information. We use the part-based object detector of Felzenszwalb et al. (2010) to obtain a score for each bounding box in the image. We convert these detection scores to superpixel scores by selecting the highest scoring detection intersecting each pixel of the superpixel. We then compute the mean of pixel-level scores over the superpixel.

In Fig. 5 we show per-cue maps of the probability of superpixels belonging to four PASCAL classes. In this example, the bottle class is very poorly segmented by FG-BG, especially compared to the segmentation using CLASS and OBJ. Note also the LOC cue reduces the noisy segmentation of the dining table in the top-right of the image.

In Fig. 6 we show the individual performance of the four cues described above on the PASCAL VOC 2010 validation dataset. Of the individual cues, FG-BG is significantly better than all others. However, from this table we see that the CLASS cue is complementary to FG-BG since their combination increases performance by more than three percent. Combining all four cues obtains the best results.

6.2 Global Unary Potential

The global unary potential is defined as:

$$\phi_g^G(x_g) = -\mu_G \log(P(X_g = x_g)P(\mathbf{O}_g|X_g = x_g)), \tag{20}$$

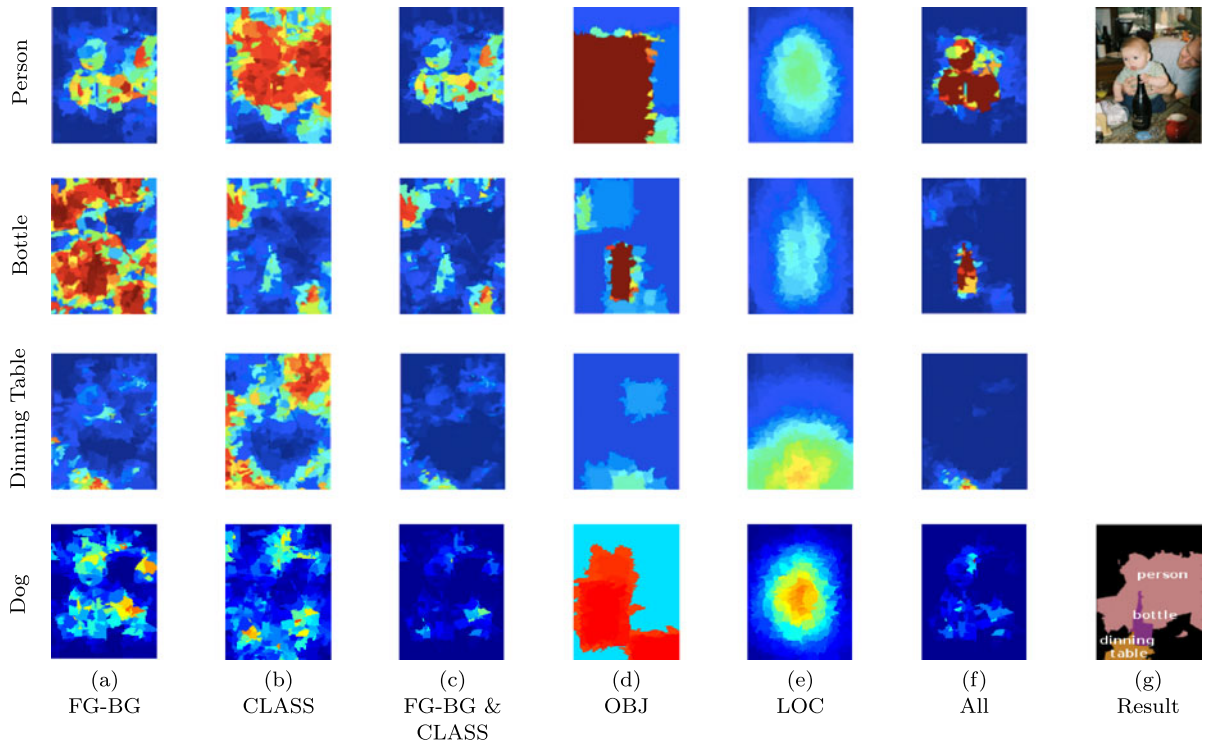


Fig. 5 Example of the local unary potentials. Examples of responses for the different cues for Person, Bottle, Dinning Table and Dog classes. (**a**) FG-BG, (**b**) CLASS, (**c**) combination of FG-BG and CLASS, (**d**) OBJ, (**e**) LOC, (**f**) all, (**g**) input and results image. See the text for explanation

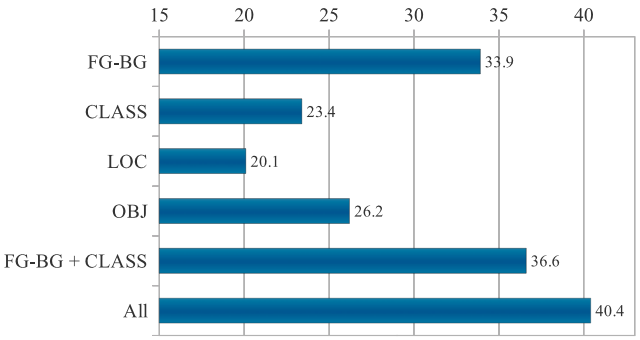


Fig. 6 Combining cues. Segmentation results on PASCAL 2010 dataset. Results are shown for the four cues used in our method: foreground-background (FG-BG), object class against other objects (CLASS), location (LOC) and object detection (OBJ)

where μ_G is the weighting factor of the global unary potential. The prior $P(X_g = x_g)$ can be approximated by the frequency that label x_g appears in the ground-truth image of the training-set, i.e. $\sum_{I_i \in \mathcal{I}} \mathrm{T}[x_g \subseteq t_g^i]$. Since learning $P(\mathbf{O}_g | X_g = x_g)$ for all combinations of labels is unfeasible, we employ the same approximation here as in (14) and (15),

$$P(\mathbf{O}_g | X_g = x_g) = P(\{\mathbf{O}_g^{l_k}\}_{l_k \in \mathcal{L}} | X_g = x_g) \tag{21}$$

$$\propto \prod_{k | l_k \notin x_g} P(l_k \notin X_g | \mathbf{O}_g^{l_k}) \prod_{k | l_k \in x_g} P(l_k \in X_g | \mathbf{O}_g^{l_k}), \tag{22}$$

where

$$P(l_k \notin X_g | \mathbf{O}_g^{l_k}) = 1 - P(l_k \in X_g | \mathbf{O}_g^{l_k}).$$

$P(l_k \in X_g | \mathbf{O}_g^{l_k})$ is obtained transforming through a sigmoid the classification score given the representation $\mathbf{O}_g^{l_k}$ of the whole image, which is based again on a bag-of-words.

6.3 Smoothness Potential

The smoothness term is given by

$$\psi_{ij}^L(x_i, x_j) = \lambda_L K_{ij} \theta(c_{ij}) \mathrm{T}[x_i \neq x_j] \tag{23}$$

where λ_L is the weighting factor of the smoothness term, K_{ij} normalizes over the length of the shared boundary between superpixels, and $c_{ij} = \|c_i - c_j\|$ is the norm of the difference of the mean RGB colors of superpixels i and j. In our case, instead of relying on a predefined function to relate the smoothness cost with the color difference between superpixels, we empirically define a set of parameters θ as modulation costs.

6.4 Consistency Potential

In our approach we use the harmony potential as the consistency potential. Recall from (7) that the harmony potential is defined as:

$$\psi_{ig}^{G}(x_i, x_g) = \gamma_i(x_i)\mathrm{T}[x_i \notin x_g]. \tag{24}$$

We define the penalization factor as $\gamma_i(x_i) = \lambda_G K_i$, where λ_G is the weighting factor of the consistency term, and K_i normalizes over the number of pixels contained in the superpixel.

6.5 Learning HCRF Parameters

Learning the parameters of the CRF potentials is a key step in attaining state-of-the-art results on the labeling problem. In our case, we have two groups of parameters that must be learned.

First, it is necessary to calibrate the classification scores because the classifiers are learned independently for each class and are trained without taking into account the others classes. In this case, the classification scores are unbalanced, and their relative strength is unknown. The outputs scores of individually trained classifiers are effectively incomparable. In order to overcome this problem, the usage of the sigmoid functions for the local and global unary potential enables us to weight the importance of each cue for each class, and also weight the strength of each classifier with respect to the others. We found this to significantly improve results.

In addition to these per-class, per-cue sigmoid parameters, we must also learn the weighting parameters of the different potentials: λ_G, λ_L, μ_L and μ_G. We learn both groups of parameters by iterating a two-step procedure until convergence. In the first step, we train the weighting factors of the potentials, while in the second step we learn the per-class, per-cue sigmoid parameters $a(k, x_i)$ and $b(k, x_i)$ of the local unary potential and the per-class sigmoid parameters of the global unary potential. These two sets of parameters are quite decoupled, and this division reduces the size of the parameter space at each step. We use $\boldsymbol{\pi}$ to denote the set of parameters to be learned.

In each step we randomly generate new instances of parameters $\boldsymbol{\pi}$ and select the one that maximizes the performance of the segmentation on a validation set. We obtain new parameter instances with a simple Gibbs sampling-like algorithm in which each time we vary one, randomly chosen parameter $\pi \in \boldsymbol{\pi}$. Only if the segmentation performance increases on the validation set do we keep the new parameter value. We vary the parameter using a normal distribution with 0 mean and deviation $\sigma(t)$ which depends on the iteration number t. At each new iteration, if some improvement has been achieved, we multiply $\sigma(t)$ by a factor in order to reduce the variability of the parameters when we are near convergence. This factor is a compromise between computational cost and the possibility of getting stuck in local extrema.

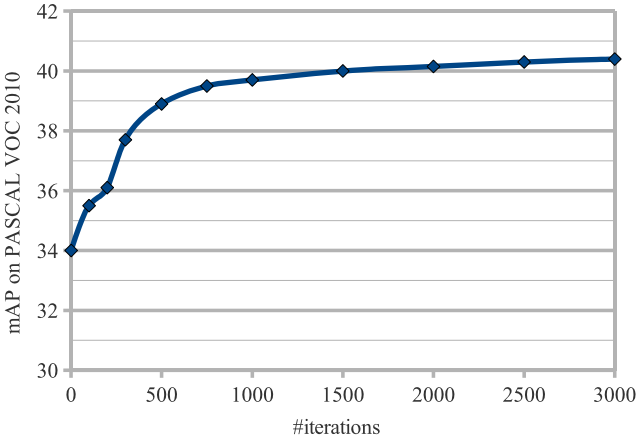


Fig. 7 Parameter optimization. Improvement of performance on PASCAL VOC 2010 validation set as a function of number of iterations, showing the importance of per-class normalization

In Fig. 7 the improvement from learning the parameters described in this section is shown for the PASCAL VOC 2010. An absolute performance gain of over 5% is obtained.

7 Experiments

We evaluate our method on two challenging datasets for object class segmentation: the PASCAL VOC 2010 Segmentation Challenge (Everingham et al. 2010) and the MSRC-21 dataset (Shotton et al. 2009). VOC 2010 contains 20 object classes plus the background class, MSRC-21 contains 21 classes. The PASCAL dataset focuses on object recognition, and normally only one or few objects are present in the image, surrounded by background. In contrast, the MSRC-21 contains fully labeled images, where the background is divided in different regions, such as grass, sky or water. After giving the most relevant implementation details, we discuss the results obtained on both datasets.

7.1 Implementation Details

We extract patches over a grid with 50% overlap at several scales (12, 24, 36 and 48 pixels of diameter). These patches are described by shape (SIFT), color (RGB histogram) and the SSIM self-similarity descriptor (Shechtman and Irani 2007). In order to build a bag-of-words representation, we quantize with K-means the shape features to 1.000 words, the color features to 400 words and the SSIM descriptor to 300 words.

We use a different SVM classifier with intersection kernel (Maji et al. 2008) for each label to obtain classification scores. Each classifier is learned using a similar number of

positive and negative examples: around a total of 8.000 superpixel samples for MSRC-21, and 20.000 for VOC 2010 for each class.

The feature assignment to build the bag-of-words is done using nearest neighbor, and as mentioned we concatenate the bag-of-words of the inside of the superpixel with that of region around it. Thus, the description of a single superpixel has a dimension of $2 \times (1.000 + 400 + 300)$ bins. The contextual area of a superpixel is extended up to 4 times the size of the feature.

In the case of VOC 2010, the global classification score is based on a comprehensive image classification method. We use a bag-of-words representation (Zhang et al. 2007), based on shape SIFT, color SIFT (van de Sande et al. 2010), together with spatial pyramids (Lazebnik et al. 2006) and color attention (Shahbaz et al. 2009) based on the Color Name feature (van de Weijer et al. 2009). Furthermore, the training of the global node only requires weakly labeled image data, and can therefore be done on the larger set of 10.103 images labeled for image classification. In the case of MSRC-21, we use a simpler bag-of-words representation based on SIFT, RGB histograms, SSIM and spatial pyramids (Lazebnik et al. 2006) with max-pooling (Yang et al. 2009). In both methods, we use an SVM with intersection kernel as a classifier.

The global node uses the M' most probable labels obtained by ranked sampling. We set M' to a value such that no significant improvements are observed beyond it, which was found to be $M' = 50$ for all experiments. An approximate MAP configuration $\mathbf{x}^*$ can be inferred using a message passing or graph cut based algorithm. In all the experiments we use α-expansion graph cuts[1] (Boykov et al. 2001), where α can be any label present in the CRF, which is the union between the M' labels of the global node and the set $\mathcal{L}$ of labels of the local nodes. The average time to segment an image in MSRC-21 is just 0.24 seconds and in VOC 2010 it is 0.32 seconds.

7.2 Results for MSRC-21

In Table 1, our results are compared with other state-of-the-art methods. We also show the results without consistency potentials and results obtained with Potts and robust P^N-based potentials. It should be noted that we optimized our system on the average per-class recall.

The results show that without consistency potentials we obtain a baseline of 71% average recall. From this baseline, Potts potentials improve by 5%, robust P^N-based potentials by 6%, and harmony potentials by 9%, obtaining state-of-the-art results of 80% average recall. In Fig. 8 we provide segmentation results for different potentials. Overall, adding consistency potentials smooths segmentation results and removes small segments. In the first row the global classifier punishes the presence of cow, allowing it to correctly label the region as dog. The third row provides an example where semantic co-occurrence helps to correctly label the water region. Since in the training set the combination of dog and human is unlikely, the results of the harmony potential deteriorate in the fourth row. In the last row, the incorrect recognition of the water region as road results in an incorrect classification of the boat as bicycle.

Looking at the global score, the best scores are obtained by Ladicky et al. (2010b). Their hierarchical CRF model achieves excellent performance on the stuff classes such as building, grass, sky, water. On the other hand, on some of the difficult and less frequent object classes we obtain significantly better results: on boat, bird, chair and boat we more than double the performance of Ladicky et al. (2010b).

7.3 Results for PASCAL VOC 2010

In Table 2 the results on the PASCAL VOC 2010 dataset for both the validation and the test sets are summarized. Performance is evaluated for each class using average precision (see the PASCAL VOC evaluation criteria defined in Everingham et al. (2010)).

To analyze the influence of both the co-occurrence (CO) used to compute the prior and the introduction of image classification results at the global node, we performed several experiments on the validation set. Not using either of them, hence without global consistency (see Fig. 2b), gives an overall score of 31.2%. Introducing consistency in the form of CO without global observation improves results to 33.4%, which is consistent with the gain reported in Ladicky et al. (2010a). Only using the information from image classification at the global node (without CO) yields a performance increase to 35.3%. Including both CO and global observation leads to an overall average precision of 40.4% (referenced as *All cues* in Table 2).

Figure 9 shows the results of our method compared to the method without consistency potentials (obtaining a mAP of 31.2% on the validation set). This allows us to illustrate the influence of the global node and the global classifier on the segmentation results. In most cases the harmony potential removes unlikely classes and significantly improved results are obtained. It is worth noting that labels in the local nodes that are not encoded in the global node label combination are penalized by the harmony potential, but may still appear in the final segmentation (always at a cost). We have found that about 15% of the image segmentations contain labels that are not encoded in the global label. This happens mainly for two reasons: a failure in the global image classifier, or due to a combination of labels that has never been

[1] Our implementation uses the min-cut/max-flow libraries provided by Boykov and Kolmogorov (2004).

Table 1 MSRC-21 segmentation results. The average score provides the average per-class recall. The global scores gives the percentage of correctly classified pixels

		building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**Global**	**Average**
(Shotton et al, 2008)		49	88	79	**97**	**97**	78	82	54	**87**	74	72	74	36	24	**93**	51	78	75	35	66	18	72	67
(Jiang and Tu, 2009)		53	**97**	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13	78	68
Pixel CRF (Ladicky et al, 2009)		73	92	85	75	78	92	75	76	86	79	87	96	95	31	81	34	84	53	61	60	15	81	72
Hier. CRF (Ladicky et al, 2009)		80	96	86	74	87	99	74	87	86	**87**	82	**97**	95	30	86	31	**95**	51	69	66	09	86	75
Hier. CRF with CO (Ladicky et al, 2010a)		**82**	95	88	73	88	**100**	83	**92**	88	**87**	88	96	**96**	27	85	37	93	49	**80**	65	20	**87**	77
Our method	w/o Consistency	66	93	82	59	66	95	**88**	77	81	83	87	77	82	42	84	33	79	65	44	57	54	79	71
	Potts	63	92	**90**	81	71	97	81	71	72	69	**94**	86	83	43	82	73	84	79	64	62	52	81	76
	Robust P^N	60	92	85	76	75	96	76	75	72	75	**94**	96	86	**57**	82	75	84	79	60	63	**59**	81	77
	Harmony	66	87	84	81	83	93	81	82	78	86	**94**	96	87	48	90	**81**	82	**82**	75	**70**	52	83	**80**
	Harmony w/ Im. tags	68	93	92	86	88	97	91	85	73	86	94	100	89	77	100	96	89	95	94	60	74	89	87

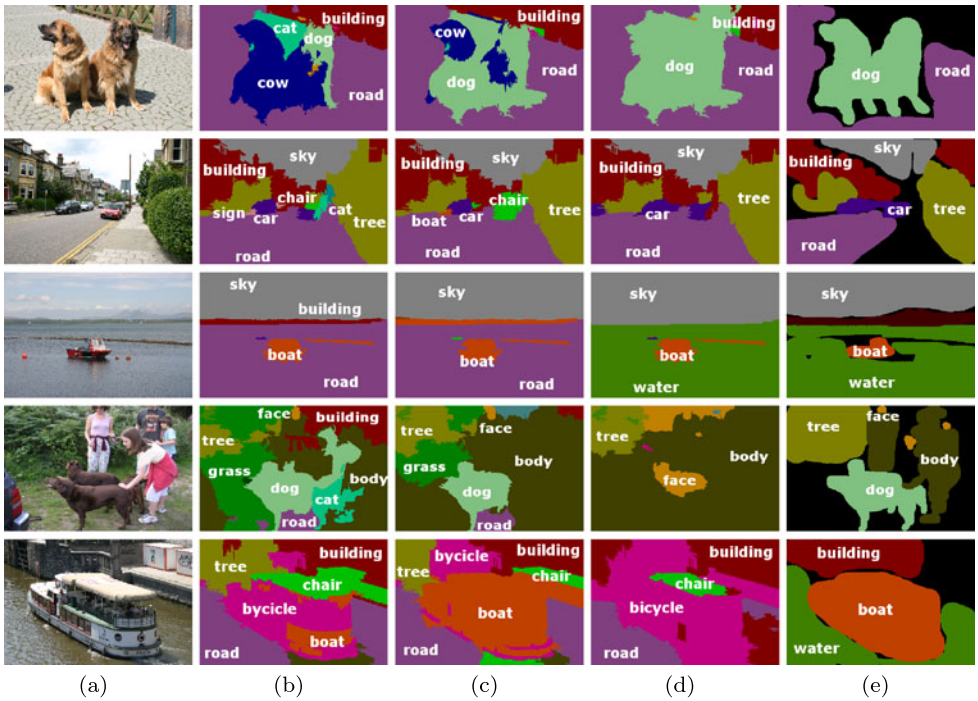


Fig. 8 Qualitative results for the MSRC-21 dataset. Comparison between (**b**) no consistency potentials, (**c**) robust P^N-based potentials, and (**d**) harmony potentials. (**e**) Ground-truth images. In the first three rows the harmony potential successfully improves segmentation results. The last two rows show failure cases of harmony potentials

Table 2 PASCAL VOC 2010 segmentation results. Comparison of the harmony potential with state-of-the-art methods

		Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dinning Table	Dog	Horse	Motorbike	Person	Potted Plant	Sheep	Sofa	Train	TV/Monitor	**Average**
VALIDATION SET																							
no global, no CO		76.7	47.5	29.3	20.2	26.2	30.3	54.7	54.4	33.1	7.3	23.9	9.5	22.6	26.3	42.4	34.3	10.9	23.2	12.7	43.5	26.7	31.2
no global, with CO		73.0	49.4	32.3	22.0	31.7	31.8	51.7	54.8	35.5	11.7	21.8	8.3	23.9	29.5	46.8	38.4	9.9	24.6	17.7	50.0	36.1	33.4
global, no CO		80.3	53.0	31.4	21.9	27.8	33.1	57.9	54.7	33.6	13.0	29.6	18.8	20.5	27.9	50.3	38.1	11.9	30.3	18.3	47.5	42.0	35.3
All cues		82.6	61.2	26.0	32.4	41.2	38.2	60.9	57.2	38.2	13.7	45.4	27.4	31.6	26.7	48.2	41.1	20.5	39.6	23.3	54.7	38.0	40.4
Im. tags		85.0	71.3	38.1	46.2	59.2	50.5	70.3	65.2	65.4	20.7	72.0	51.3	63.8	57.6	65.9	50.0	42.3	69.7	39.4	67.9	50.6	57.2
TEST SET																							
BONN SVR		**84.2**	52.5	27.4	32.3	34.5	**47.4**	60.6	54.8	**42.6**	9.0	32.9	**25.2**	27.1	32.4	47.1	38.3	**36.8**	**50.3**	**21.9**	35.2	40.9	39.7
BERKELEY		82.0	49.7	23.3	20.6	19.0	47.1	58.1	53.6	32.5	0.0	31.1	0.0	29.5	**42.9**	41.9	43.8	16.6	39.0	18.4	38.0	41.5	34.7
BROOKES		70.1	31.0	18.8	19.5	23.9	31.3	53.5	45.3	24.4	8.2	31.0	16.4	15.8	27.3	48.1	31.1	31.0	27.5	19.8	34.8	26.4	30.3
STANFORD		80.0	38.8	21.5	13.6	9.2	31.1	51.8	44.4	25.7	6.7	26.0	12.5	12.8	31.0	41.9	44.4	5.7	37.5	10.0	33.2	32.3	29.1
UC3M		73.4	45.9	12.3	14.5	22.3	9.3	46.8	38.3	41.7	0.0	35.9	20.7	**34.1**	34.8	33.5	24.6	4.7	25.6	13.0	26.8	26.1	27.8
UOCTTI		80.0	36.7	23.9	20.9	18.8	41.0	62.7	49.0	21.5	8.3	21.1	7.0	16.4	28.2	42.5	40.5	19.6	33.6	13.3	34.1	**48.5**	31.8
Our method	FG-BG	80.2	**57.0**	**28.7**	29.3	31.7	27.0	57.6	48.5	35.2	8.3	29.9	22.6	25.2	33.0	52.6	35.9	25.2	39.7	16.9	43.4	24.7	35.8
	All cues	82.2	52.6	26.8	**37.7**	**35.4**	34.4	**63.3**	**61.0**	32.1	**11.9**	**36.6**	23.9	33.7	36.8	**61.6**	**45.0**	26.6	40.5	20.4	**43.8**	36.4	**40.1**

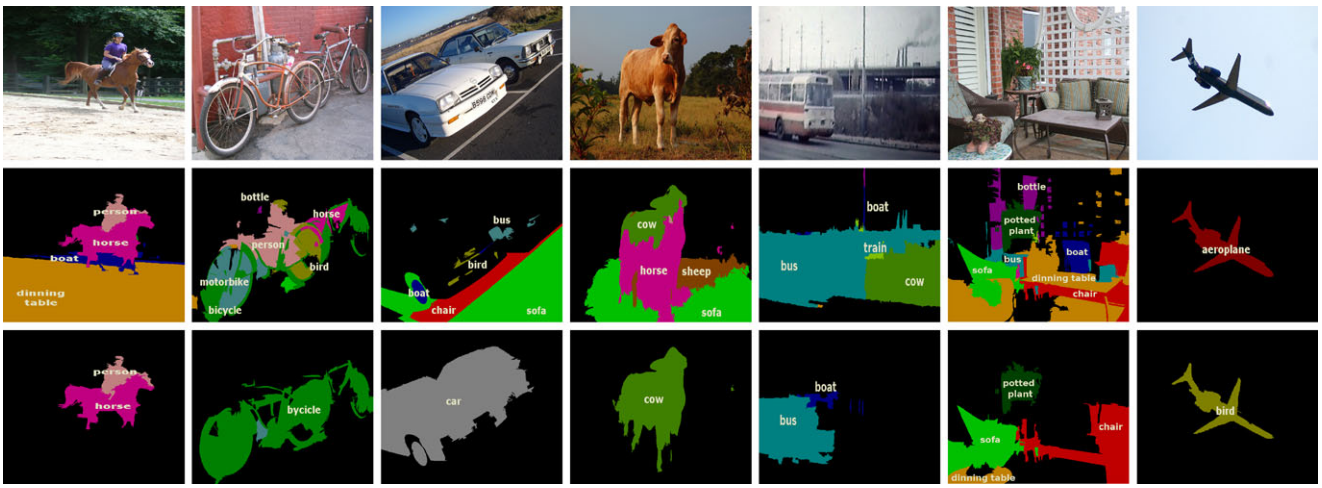


Fig. 9 Qualitative results for the PASCAL VOC 2010 dataset. Comparison between not using the harmony potential (*middle row*) and using it with an image categorization method (*bottom row*). The first four columns show examples of successful segmentation using the harmony potential. Columns five and six show results with label combinations never seen in the training images. Finally, the last column show a failure case, caused by a higher probability of birds at the global scale

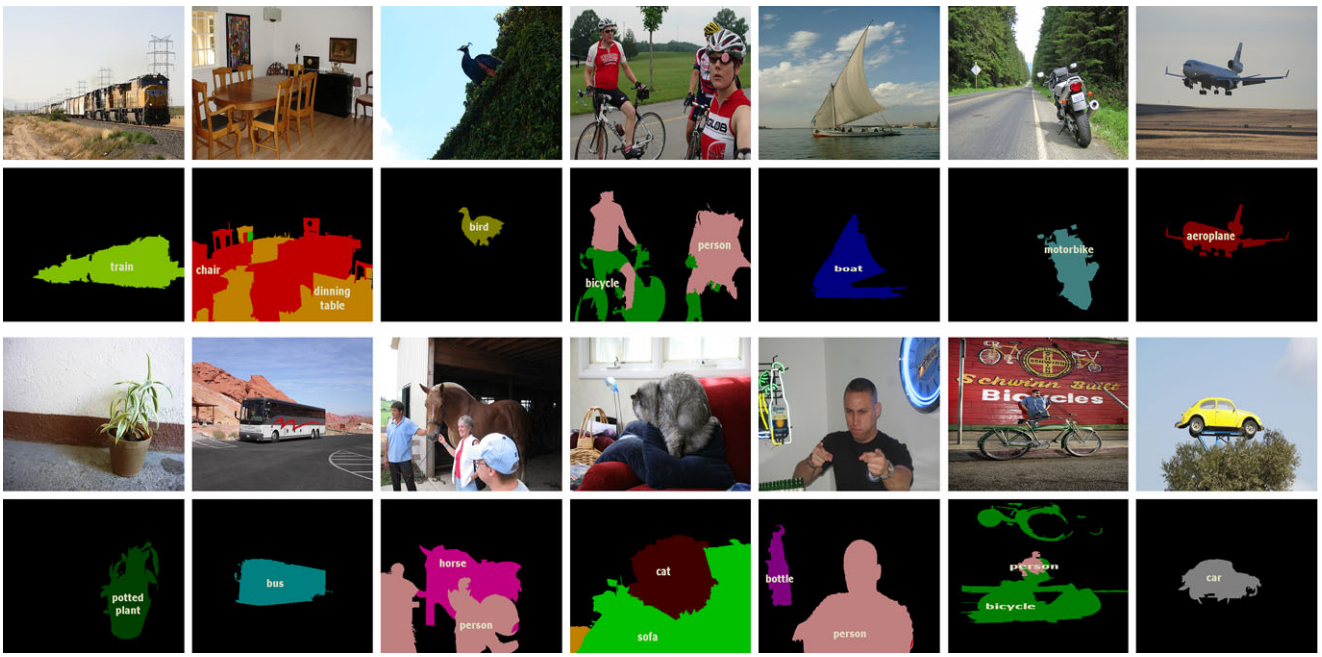


Fig. 10 Qualitative results of PASCAL VOC 2010. The original image (*top*) and our successful segmentation result (*bottom*)

seen during training. As an example, columns five and six in Fig. 9 show two examples of the latter case. The last column shows an error caused by the global classifier, which converts the aeroplane into a bird. It should also be noted that there are weights balancing the importance of global evidence versus local evidence (see μ_L and μ_G in (18) and (20), respectively).

Compared to our early work (Gonfaus et al. 2010) which was only based on the FG-BG cue instead of the four cues we use now, we obtain an absolute performance gain of almost 5% in average precision. We also compare our results to the best submission to the PASCAL VOC 2010 challenge.

Most related to our work is the submission of BROOKES (Ladicky et al. 2010a) which is also a hierarchical CRF method. Because of the lack of stuff classes in the PASCAL dataset, the performance gain of the harmony potentials is especially pronounced. Overall we obtain the best results on eleven out of the twenty classes, and obtain slightly better mean average precision than the BONN SVR (Li et al. 2010) submission. For several classes the results of our

method and those of BONN diverge significantly, which indicates that both methodologies could be combined to obtain better results.

A variety of segmentation results are shown in Fig. 10. The results show that harmony potential is able to deal with multiclass images, partial occlusion, and to correctly classify the background. Notice the difficulties on the chair class in the second column, which are also reflected in an average precision of only 11.9% on chairs.

7.4 Influence of Image Classification

The success of our image segmentation algorithm is partially dependent on the quality of image classification. To have a better understanding of how improved image classification can influence results we performed an additional experiment using perfect image classification information, meaning that $P(X_g = x_g|\mathbf{O}_g) = 1$ for the actual label combination and zero for the other label combinations. This situation could arise, for example, when image tags are available.[2] Results are given for MSRC-21 in Table 1, and for the PASCAL VOC 2010 validation set in Table 2. Results on PASCAL are shown only for the validation set because this experiment requires groundtruth labels which are not available for the test set.

The results show that for both datasets a significant gain can be obtained by improving global classification scores. The MSRC-21 dataset mean average precision goes up by 7% to 87%, and for PASCAL by 17% to 57%. For PASCAL the performance gain is especially significant for the easily confusable animal classes such as cat, dog, horse, cow and sheep. For these classes perfect classification scores help to choose the correct class and relative performance gains are around 100%. Other classes such as chair, bicycle, and sofa even with image tags remain very difficult to localize and mean average precision remains below 50%.

8 Conclusions

We presented a new CRF model for object class image segmentation. Existing CRF models only allow a single label to be assigned to the nodes representing the image at different scales. In contrast, we allow the global node, which represents the whole image, to take any combination of class labels. This allows us to better exploit class-label estimates based on observations at the global scale. This is especially important because for inference of the global node label we can use the full power of state-of-the-art image classification techniques. Experiments show that our new CRF model obtains state-of-the-art results on two challenging datasets.

For future work, we are especially interested in combining the various potentials into hierarchical CRFs. The Potts potential is appropriate as a smoothness potential at the lowest scales, for mid-level scales the robust P^N-based potential is more appropriate, whereas at the highest scales harmony potentials better model the heterogeneity of real-world images. Given the fact that for our model inference for a single image takes less than one second, it seems feasible to investigate hierarchical CRF models with heterogeneous potentials.

Acknowledgements We gratefully acknowledge Fahad Shahbaz Khan for providing image classification results for the PASCAL dataset. This work has been supported by the EU projects ERGTS-VICI-224737, VIDI-VIDEO IST-045547, FP7-ICT-24314 and FP7-ICT-248873; by the Spanish Research Program Consolider-Ingenio 2010: MIPRCV (CSD2007-00018); and by the Spanish projects TIN2009-14501-C02-02, TIN2009-14173 and TRA2010-21371-C03-01. Joost van de Weijer acknowledges the support of a Ramon y Cajal fellowship, and Xavier Boix the support of the FPU fellowship AP2008-03378.

Appendix A

Let $E(\mathbf{x})$ and $E(\mathbf{x}_L)$ be the energies of the models $\mathcal{G}$ and $\mathcal{G}_H$, which are:

$$E(\mathbf{x}) = K(\mathbf{x}_L) + \phi_g^G(x_g) + \sum_{(i,g)\in\mathcal{E}_G} \psi_{ig}^G(x_i, x_g), \tag{25}$$

where $\phi_g^G(x_g)$ is the global unary potential, and

$$E_H(\mathbf{x}_L) = K(\mathbf{x}_L) + \psi^H(\mathbf{x}_L). \tag{26}$$

For the sake of simplicity we have abbreviated the smoothness and local potentials with the term $K(\mathbf{x}_L)$. Recall that $\mathbf{x} = \{\mathbf{x}_L, x_g\}$.

Let the consistency potential ψ_{ig}^G of $E(\mathbf{x})$ be the harmony potential in (7). We want to prove that if the high-order potential ψ^H of $E(\mathbf{x}_L)$ is defined as in (9), both models give the same configuration $\mathbf{x}^\star$ when doing inference, in other words: are equivalent.

Rewriting the high-order energy of $\mathbf{x}_L^\star$ it becomes

$$\begin{aligned} E_H(\mathbf{x}_L^\star) &= \min_{\mathbf{x}_L} E_H(\mathbf{x}_L) \qquad (27)\\ &= \min_{\mathbf{x}_L} \Big\{ K(\mathbf{x}_L) \\ &\quad + \min_{\ell\in\mathcal{P}(\mathcal{L})} \Big\{ \gamma_g(\ell) + \sum_{i\in\mathcal{V}_L} \gamma_i(x_i)\mathrm{T}[x_i \notin \ell] \Big\} \Big\} \qquad (28) \end{aligned}$$

[2] It should be noted that in case of perfect classifier the global node is not necessary and simply restricting the label set of the local nodes would obtain similar scores.

$$= \min_{\mathbf{x}_L, \ell \in \mathcal{P}(\mathcal{L})} \left\{ K(\mathbf{x}_L) + \gamma_g(\ell) + \sum_{i \in \mathcal{V}_L} \gamma_i(x_i) \mathrm{T}[x_i \notin \ell] \right\}. \tag{29}$$

Then, substituting the auxiliary variable ℓ by the random variable X_g (Rother et al. 2009):

$$E_H(\mathbf{x}_L^\star, x_g^\star) = \min_{\mathbf{x}_L, x_g \in \mathcal{P}(\mathcal{L})} E_H(\mathbf{x}_L, x_g) \tag{30}$$

$$= \min_{\mathbf{x}_L, x_g \in \mathcal{P}(\mathcal{L})} \left\{ K(\mathbf{x}_L) + \gamma_g(x_g) + \sum_{(i,g) \in \mathcal{E}_G} \gamma_i(x_i) \mathrm{T}[x_i \notin x_g] \right\}, \tag{31}$$

which it turns to be $E(\mathbf{x}^\star)$ if we set $\gamma_g(x_g) = \phi_g^G(x_g)$, because the summation term is by definition the harmony potential, i.e. $\psi_{ig}^G(x_i, x_g) = \gamma_i(x_i)\mathrm{T}[x_i \notin x_g]$.

Appendix B

Let $\ell_1 \in \mathcal{P}(\mathcal{L}'')$ and $\ell_2 \in \mathcal{P}(\mathcal{L}')$ be two partially built labels in the branch-and-bound procedure. ℓ_2 is obtained after branching, considering one extra label in ℓ_1: i.e. either $\ell_2 = \{\ell_1, l_{branch}\}$ (adding ℓ_{branch}) or $\ell_2 = \ell_1$ (adding nothing). We must prove that in both cases $q(\ell_1) \geq q(\ell_2)$. Assuming (12) and (15), we can decompose the $q(\ell_1)$ into its constituent factors: the likelihood $q_{lhood}(\ell_1)$ and the prior $q_{prior}(\ell_1)$. It is then sufficient to show that these constituent components bound $q_{lhood}(\ell_2)$ and $q_{prior}(\ell_2)$, respectively.

When l_{branch} is added, for the likelihood we have

$$q_{lhood}(\ell_2)$$

$$= \prod_{k | l_k \notin \ell_2} P(l_k \notin X_g | \mathbf{O}_g^{l_k}) \prod_{k | l_k \in \ell_2} P(l_k \in X_g | \mathbf{O}_g^{l_k}) \tag{32}$$

$$= P(l_{branch} \in X_g | \mathbf{O}_g^{l_{branch}}) \cdot q_{lhood}(\ell_1) \tag{33}$$

$$\leq q_{lhood}(\ell_1), \tag{34}$$

Equality (32) is obtained from (15), and (33) is due to the fact that $\ell_2 = \{\ell_1, l_{branch}\}$. The final inequality follows from the fact that $P(l_{branch} \in X_g | \mathbf{O}_g^{l_{branch}}) \leq 1$. When l_{branch} is not added, in (33) instead of $P(l_{branch} \in X_g | \mathbf{O}_g^{l_{branch}})$ we have $P(l_{branch} \notin X_g | \mathbf{O}_g^{l_{branch}})$ and it follows in the same way.

For the prior we have

$$q_{prior}(\ell_2) = \sum_{I_i \in \mathcal{I}} \mathrm{T}[\ell_2 \subseteq t_g^i] \tag{35}$$

$$\leq \sum_{I_i \in \mathcal{I}} \mathrm{T}[\ell_1 \subseteq t_g^i] \tag{36}$$

$$= q_{prior}(\ell_1).$$

Equality (35) comes from (12), and the inequality (36) since $\ell_1 \subseteq \ell_2$ and hence $\ell_1 \subseteq t_g^i \implies \ell_2 \subseteq t_g^i$.

References

Adelson, E. H. (2001). On seeing stuff: the perception of materials by humans and machines. In *Proceedings of the SPIE: human vision and electronic imaging VI*.

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1124–1137.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1222–1239.

Carreira, J., & Sminchisescu, C. (2010). Constrained parametric min-cuts for automatic object segmentation. In *Proc. computer vision and pattern recognition*.

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

Coughlan, J. M., & Ferreira, S. J. (2002). Finding deformable shapes using loopy belief propagation. In *Proc. European conf. on computer vision*.

Csurka, G., & Perronnin, F. (2010). An efficient approach to semantic segmentation. *International Journal of Computer Vision* doi:10.1007/s11263-010-0344-8.

Delong, A., Osokin, A., Isack, H. N., & Boykov, Y. (2010). Fast approximate energy minimization with label costs. In *Proc. computer vision and pattern recognition*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Freeman, W. T., Pasztor, E. C., & Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision*, *40*(1), 25–47.

Frey, B., & MacKay, D. (1998). A revolution: belief propagation in graphs with cycles. In *Advances in neural information processing systems*.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In *Proc. IEEE int. conf. on computer vision*.

Galleguillos, C., & Belongie, S. (2010). Context based object categorization: a critical survey. *Computer Vision and Image Understanding*, *114*, 712–722.

Gonfaus, J., Boix, X., van de Weijer, J., Bagdanov, A., Serrat, J., & Gonzàlez, J. (2010). Harmony potentials for joint classification and segmentation. In *Proc. computer vision and pattern recognition*.

Gould, S., Gao, T., & Koller, D. (2009). Region-based segmentation and object detection. In *Advances in neural information processing systems*.

Hammersley, J. M., & Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished.

Hoiem, D., Efros, A. A., & Hebert, M. (2007). Recovering surface layout from an image. *International Journal of Computer Vision*, *75*(1), 151–172.

Hoiem, D., Efros, A. A., & Hebert, M. (2008). Putting objects in perspective. *International Journal of Computer Vision 80*(1), 3–15.

Ihler, A., & McAllester, D. (2009). Particle belief propagation. In *Proc. int. conf. on artificial intelligence and statistics*.

Ishikawa, H. (2009). Higher-order clique reduction in binary graph cut. In *Proc. computer vision and pattern recognition*.

Jain, A., Gupta, A., & Davis, L. (2010). Learning what and how of contextual models for scene labeling. In *Proc. European conf. on computer vision*.

Jiang, J., & Tu, Z. (2009). Efficient scale space auto-context for image segmentation and labeling. In *Proc. computer vision and pattern recognition*.

Kohli, P., & Kumar, M. P. (2010). Energy minimization for linear envelope MRFs. In *Proc. computer vision and pattern recognition*.

Kohli, P., Kumar, M. P., & Torr, P. H. (2009a). P^3 and beyond: move making algorithms for solving higher order functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(9), 1645–1656.

Kohli, P., Ladický, L., & Torr, P. H. (2009b). Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, *82*(3), 302–324.

Koller, D., Lerner, U., & Angelov, D. (1999). A general algorithm for approximate inference and its application to hybrid Bayes nets. In *Proc. annual conference on uncertainty in artificial intelligence*.

Kumar, M. P., Torr, P., & Zisserman, A. (2005). Obj cut. In *Proc. computer vision and pattern recognition*.

Kumar, S., & Hebert, M. (2005). A hierarchical field framework for unified context-based classification. In *Proc. IEEE int. conf. on computer vision*.

Ladicky, L., Russell, C., Kohli, P., & Torr, P. (2009). Associative hierarchical crfs for object class image segmentation. In *Proc. IEEE int. conf. on computer vision*.

Ladicky, L., Russell, C., Kohli, P., & Torr, P. H. S. (2010a). Graph cut based inference with co-occurrence statistics. In *Proc. European conf. on computer vision*.

Ladicky, L., Sturgess, P., Alahari, K., Russell, C., & Torr, P. H. S. (2010b). What, where & how many? combining object detectors and crfs. In *Proc. European conf. on computer vision*.

Lauritzen, S. L. (1996). *Graphical models. Oxford statistical science series*. London: Oxford University Press.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proc. computer vision and pattern recognition*.

Lee, Y., & Grauman, K. (2010). Object-graphs for context-aware category discovery. In *Proc. computer vision and pattern recognition*.

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, *77*(1–3), 259–289.

Lempitsky, V., Kohli, P., Rother, C., & Sharp, T. (2009). Image segmentation with a bounding box prior. In *Proc. IEEE int. conf. on computer vision*.

Levin, A., & Weiss, Y. (2009). Learning to combine bottom-up and top-down segmentation. *International Journal of Computer Vision*, *81*(1), 1645–1656.

Li, F., Carreira, J., & Sminchisescu, C. (2010). Object recognition as ranking holistic figure-ground hypotheses. In *Proc. computer vision and pattern recognition*.

Li, Y., & Huttenlocher, D. P. (2008). Sparse long-range random field and its application to image denoising. In *Proc. European conf. on computer vision*.

Lim, J. J., Arbelaez, P., Gu, C., & Malik, J. (2009). Context by region ancestry. In *Proc. IEEE int. conf. on computer vision*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *Proc. computer vision and pattern recognition*.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE int. conf. on computer vision*.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). 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.

Mori, G., Ren, X., Efros, A. A., & Malik, J. (2004). Recovering human body configurations: combining segmentation and recognition. In *Proc. computer vision and pattern recognition*.

Munoz, D., Bagnell, J. A., Vandapel, N., & Hebert, M. (2009). Contextual classification with functional max-margin Markov networks. In *Proc. computer vision and pattern recognition*.

Munoz, D., Bagnell, J. A., & Hebert, M. (2010). Stacked hierarchical labeling. In *Proc. European conf. on computer vision*.

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Proc. European conf. on computer vision*.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

Pantofaru, C., Schmid, C., & Hebert, M. (2008). Object recognition by integrating multiple image segmentations. In *Proc. European conf. on computer vision*.

Plath, N., Toussaint, M., & Nakajima, S. (2009). Multi-class image segmentation using conditional random fields and global classification. In *Proc. international conference on machine learning*.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *Proc. IEEE int. conf. on computer vision*.

Ramalingam, S., Kohli, P., Alahari, K., & Torr, P. H. S. (2008). Exact inference in multi-label crfs with higher order cliques. In *Proc. computer vision and pattern recognition*.

Roth, S., & Black, M. J. (2009). Fields of experts. *International Journal of Computer Vision*, *82*(2), 205–229.

Rother, C., Kohli, P., Feng, W., & Jia, J. (2009). Minimizing sparse higher order energy functions of discrete variables. In *Proc. computer vision and pattern recognition*.

Russell, C., Ladicky, L., Kohli, P., & Torr, P. H. (2010). Exact and approximate inference in associative hierarchical random fields using graph-cuts. In *Proc. annual conference on uncertainty in artificial intelligence*.

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(10), 1582–1596.

Schmid, C., & Mohr, R. (1997). Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 530–535.

Shahbaz Khan, F., van de Weijer, J., & Vanrell, M. (2009). Top-down color attention for object recognition. In *Proc. IEEE int. conf. on computer vision*.

Shechtman, E., & Irani, M. (2007). Matching local self-similarities across images and videos. In *Proc. computer vision and pattern recognition*.

Shotton, J., Johnson, M., & Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *Proc. computer vision and pattern recognition*.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, *81*(1), 2–23.

Sivic, J., & Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. In *Proc. IEEE int. conf. on computer vision*.

Sudderth, E. B., Ihler, A. T., Ihler, E. T., Freeman, W. T., & Willsky, A. S. (2002). Nonparametric belief propagation. In *Proc. computer vision and pattern recognition*.

Tu, Z., & Zhu, S. C. (2002). Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 657–673.

Tu, Z., Chen, X., Yuille, AL, & Zhu, S. C. (2005). Image parsing: unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, *63*(2), 18–25.

Vazquez, E., Baldrich, R., van de Weijer, J., & Vanrell, M. (2011). Describing reflectances for colour segmentation robust to shadows, highlights and textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 917–930.

Vedaldi, A., & Soatto, S. (2008). Quick shift and kernel methods for mode seeking. In *Proc. European conf. on computer vision*.

Verbeek, J., & Triggs, B. (2008). Scene segmentation with crfs learned from partially labeled images. In *Advances in neural information processing systems*.

Wainwright, M. J., & Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Hanover: Now Publishers Inc.

van de Weijer, J., Schmid, C., Verbeek, J., & Larlus, D. (2009). Learning color names for real-world applications. *IEEE Transactions on Image Processing*, *18*(7), 1512–1523.

Winn, J., & Jojic, N. (2005). Locus: learning object classes with unsupervised segmentation. In *Proc. IEEE int. conf. on computer vision*.

Woodford, O., Torr, P. H., Reid, I., & Fitzgibbon, A. (2009). Global stereo reconstruction under second-order smoothness priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2115–2128.

Yang, J., Yuz, K., Gongz, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Proc. computer vision and pattern recognition*.

Yang, L., Meer, P., & Foran, D. J. (2007). Multiple class segmentation using a unified framework over mean-shift patches. In *Proc. computer vision and pattern recognition*.

Yang, Y., Hallman, S., Ramanan, D., & Fowlkes, C. (2010). Layered object detection for multi-class segmentation. In *Proc. computer vision and pattern recognition*.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, *73*(2), 213–238.

Zhu, L., Chen, Y., Lin, Y., Lin, C., & Yuille, A. L. (2008). Recursive segmentation and recognition templates for 2D parsing. In *Advances in neural information processing systems*.