Harmony Potentials for Joint Classification and Segmentation

Josep M. Gonfaus[1,2,*] Xavier Boix[1,*]
Joost van de Weijer[1,2] Andrew D. Bagdanov[1] Joan Serrat[1,2] Jordi Gonzàlez[1,2]
[1]Centre de Visió per Computador [2]Dept. of Computer Science, Universitat Autònoma de Barcelona.

Abstract

Hierarchical conditional random fields have been successfully applied to object segmentation. One reason is their ability to incorporate contextual information at different scales. However, these models do not allow multiple labels to be assigned to a single node. At higher scales in the image, this yields an oversimplified model, since multiple classes can be reasonable expected to appear within one region. This simplified model especially limits the impact that observations at larger scales may have on the CRF model. Neglecting the information at larger scales is undesirable since class-label estimates based on these scales are more reliable than at smaller, noisier scales.

To address this problem, we propose a new potential, called harmony potential, which can encode any possible combination of class labels. We propose an effective sampling strategy that renders tractable the underlying optimization problem. Results show that our approach obtains state-of-the-art results on two challenging datasets: Pascal VOC 2009 and MSRC-21.

1. Introduction

Object class image segmentation aims to assign predefined class labels to every pixel in an image. It is a highly unconstrained problem and state-of-the-art approaches focus on exploiting contextual information available around each pixel. Similarly to [24], we distinguish three scales of context: the local, mid-level and global scales. The local scale, defined at the pixel or super-pixel level, is typically represented by local image features such as color and texture. Mid-level scales also consider labels and features of neighboring regions, and the global scale considers the entire image. One of the most successful trends in object class image segmentation poses this labeling problem as one of energy minimization of a conditional random field (CRF) [20, 6, 22]. In this paper we also adopt this framework but focus on the crucial point of how to efficiently represent and combine context at various scales.

Representing the image at the global scale has been intensively studied in the field of image classification [23, 13, 21, 3, 18]. The image is generally represented by histograms over visual words, and these representations are further enriched to incorporate, for example, spatial relationships [13]. Though local regions may also be described by a bag-of-words defined over local features such as color, texture or shape, the complex representations that have considerably improved image classification performance cannot be expected to improve local region classification. The reason is that these regions lack the complexity encountered at larger scales. Therefore, in contrast to existing CRF-based methods [17, 8, 22] 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. An additional advantage is that for training at the global scale we only require data labelled with a whole-image label. This type of label information is more abundantly available than the more expensive, hand-segmented ground-truth required for learning at the local and mid-scale scales.

Verbeek and Triggs [22] proposed to use global scale information to improve estimation at local scales. However, they use the same image representation regardless of scale. The use of image classification results was also used by Csurka *et al.* [3] to reduce the number of classes to consider in an image. In addition to image classification, shape priors have been investigated in order to guide the segmentation process [10]. Bounding boxes obtained from detection classifiers have been used as well as a prior for the segmentation [14]. Finally, Li *et al.* [15] employ the user tags provided by Flickr as an additional cue to infer the presence of an object in the image.

As mentioned before, CRFs are theoretically sound models for combining information at local and mid-level scales [20, 11]. A smoothness potential between neighboring nodes models the dependencies between 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. Generally the final effect such CRFs is

[*] Both authors contributed equally to this work.

merely a smoothing of local predictions. To overcome this problem, hierarchical CRFs have been proposed in which higher level nodes describe the class-label configuration of the smaller regions [17, 8, 24]. One of the main advantages of this approach is that mid-level context is based on larger regions, and hence can lead to more certain estimations.

A drawback of existing hierarchical models is that to make them tractable they must be oversimplified by allowing regions to have just a single label [17], or in a more recent paper, an additional free label which basically cancels the information obtained at the larger scales [8]. Even though these models might be valid for the lower mid-level scales close to the pixel level, they do not model very well the higher mid-level scales. At the highest scales, far away from pixels, they impose a rather unrealistic model since multiple classes appear together. The free label 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 present a new CRF for object class image segmentation that addresses the problems mentioned above. Our model is a two-level CRF that uses labels, features and classifiers appropriate to each level. The lowest level of nodes represents superpixels labeled with single labels, while a single global node on top of them permits any combination of primitive local node labels. A new consistency potential, which we term the *harmony potential*, is also introduced which 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 data sets, namely Pascal VOC2009 and MSCR-21.

2. Consistency Potentials for Labeling Problems

In this section we present a CRF that jointly uses global and local information for labeling problems. Although several labeling approaches have used this idea in the past [17, 12, 10], they differ in the definition of the relationship between the local and global parts. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the graph that represents our probabilistic model, where the set $\mathcal{V}$ is used for indexing random variables, and $\mathcal{E}$ is the set of undirected edges representing compatibility relationships between random variables. We use $\mathbf{X} = \{X_i\}$ to denote the set of random variables or nodes, where $i \in \mathcal{V}$. Let $\mathcal{C}$ represent the set of maximal cliques, *i.e.* the set of maximal complete subgraphs. Then, according to the Hammersley-Clifford theorem, the probability of a certain configuration can be written as the normalized negative exponential of an energy function $E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \varphi_c(\mathbf{x}_c)$, where φ_c is the potential function of clique $c \in \mathcal{C}$. The optimal labeling $\mathbf{x}^*$ is obtained by inferring the Maximum a Posteriori (MAP) probability, or, equivalently, by minimizing the global energy:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x}). \qquad (1)$$

In our approach, in order to relate global and local information, we designate one random variable for 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 global edges $\mathcal{E}_G$ connects 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 connection between the local nodes.

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

$$\sum_{i \in \mathcal{V}} \phi(x_i) + \sum_{(i,j) \in \mathcal{E}_L} \psi_L(x_i, x_j) + \sum_{(i,g) \in \mathcal{E}_G} \psi_G(x_i, x_g). \quad (2)$$

The unary term $\phi(x_i)$ depends on a single probability $P(X_i = x_i | \mathbf{O}_i)$, where $\mathbf{O}_i$ is the observation that affects X_i in the model. The smoothness potential $\psi_L(x_i, x_j)$ 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 $\psi_G(x_i, x_g)$ expresses the dependency between the local nodes and the global node.

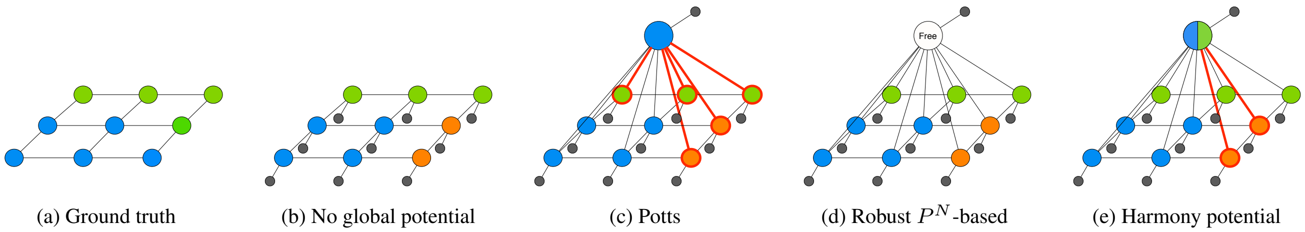
Some authors used the graphical model $\mathcal{G}$ as the basic structure to be repeated recursively in a hierarchical graph [17, 12]. In this paper we review the Potts and the robust P^N-based consistency potentials, which were used in a hierarchical CRF for segmentation. Then, we define a new one that we call harmony potential. Figure 1 briefly illustrates the characteristics of the different CRF models compared in this paper.

Potts Potential. In the basic graph used to build the tree structured model of Plath *et al.* [17][1], the consistency potential is defined as a Potts model:

$$\psi_G(x_i, x_g) = \gamma_i^l \mathrm{T}[x_i \neq x_g], \qquad (3)$$

where $\mathrm{T}[\cdot]$ is the indicator function and γ_i^l is the cost of labeling x_i as l. Since this potential encourages assigning

[1] Although the model of Plath *et al.* [17] does not have a smoothness term, we can understand its basic graph as a specific instance of $\mathcal{G}$.



(a) Ground truth (b) No global potential (c) Potts (d) Robust P^N-based (e) Harmony potential

Figure 1. Example of the penalization behavior of four different CRF models for a labeling problem with labels {blue, green, orange}, where (a) is the ground-truth. (b) Without consistency potentials only the smoothness potentials penalize 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}.

the same label as the global node to all the local nodes, this potential is unable to support any kind of heterogeneity.

Robust P^N-Based Potential. Ladicky *et al.* [12] show that the robust P^N potential defined in [8] can be understood as the sum of the pairwise connections between global and local nodes. In this case, the global node has an extended label set $\mathcal{L}^E = \mathcal{L} \cup \{l_F\}$, where l_F, which stands for "free label", means that any possible label in $\mathcal{L}$ can be assigned to local nodes with cost:

$$\psi_G(x_i, x_g) = \begin{cases} 0 & \text{if } x_g = l_F \text{ or } x_g = x_i \\ \gamma_i^l & \text{otherwise, where } l = x_i \end{cases} . \quad (4)$$

The model is recursively used to build up a hierarchical graph for object segmentation.

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.

Harmony potential. In this work we introduce a new consistency potential, which we call the *harmony potential*. 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.

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, will draw labels from $\mathcal{P}(\mathcal{L})$, the *power set* of $\mathcal{L}$, defined as the set of all subsets of $\mathcal{L}$. In this context it is intended to represent 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 hopelessly intractable. In the sequel, we propose a ranked subsampling 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 defined as:

$$\psi_G(x_i, x_g) = \gamma_i^l \mathrm{T}[x_i \notin x_g]. \quad (5)$$

Notice that we apply a penalization 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 Eq. (4), 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 Eq. (5) 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.

However, the use of the power set $\mathcal{P}(\mathcal{L})$ as the global node label set is also the main drawback of the harmony potential. In most interesting cases optimizing a problem with $2^{|\mathcal{L}|}$ possible labels is intractable. 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.

The incorporation of global information through the harmony potential is novel with respect to existing techniques exploiting image-level priors such as [19]. While such techniques rely on the image classification scores, our probabilistic framework incorporates the uncertainty of classification with the selected labels of local nodes in a joint-probabilistic manner.

3. Ranked subsampling of $\mathcal{P}(\mathcal{L})$

We have shown 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. This in turn allows us to establish a ranking of subsets that prioritizes the optimization over the $2^{|\mathcal{L}|}$ possible labels for the global node.

Optimizing for the best assignment of global label x_g^* implies maximizing $P(\ell = x_g^*|\mathbf{O})$. This is very difficult in practice due to the $2^{|\mathcal{L}|}$ possible labels and the lack of an analytic expression for $P(\ell = x_g^*|\mathbf{O})$. Instead of working directly with this posterior, we base the selection of which labels are better to use for inferring a solution on the probability that a certain label $\ell \in \mathcal{P}(\mathcal{L})$ appears in $\mathbf{x}^*$, given all the observations $\mathbf{O}$ required by the model. This is done by the following approximation of the posterior:

$$P(\ell \subseteq x_g^*|\mathbf{O}) \propto P(\ell \subseteq x_g^*)P(\mathbf{O}|\ell \subseteq x_g^*). \tag{6}$$

This approximation allows us to effectively rank possible global node labels, and thus to prioritize candidates in the search for the optimal label x_g^*. This is done by picking the best $M' \leq 2^{|\mathcal{L}|}$ subsets of $\mathcal{L}$ that maximize the posterior in Eq. (6). The posterior $P(\ell \subseteq x_g^*|\mathbf{O})$ 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 exhaustively consider all labels in $\mathcal{P}(\mathcal{L})$, but at least we consider the most likely candidates for x_g^*.

The two terms of Eq. (6) are:

- *Prior:* $P(\ell \subseteq x_g^*)$. We can approximate this probability 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.*

  $$P(\ell \subseteq x_g^*) \approx \frac{1}{K} \sum_{I_i \in \mathcal{I}} \mathrm{T}[\ell \subseteq t_g^i], \tag{7}$$

  where K is the normalization factor, and 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.

- *Likelihood:* $P(\mathbf{O}|\ell \subseteq x_g^*)$. Due to the high dimensionality of $\mathbf{O}$, the estimation of the likelihood is indeed challenging. To overcome this problem, we propose the following approximation:

  $$P(\mathbf{O}|\ell \subseteq x_g^*) \approx P(\mathbf{O}_g|\ell \subseteq x_g^*), \tag{8}$$

  where $\mathbf{O}_g$ are the only observations that influence the global node in the model. Notice that Eq. (8) only takes 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 can *design* a function that computes $P(\mathbf{O}_g|\ell \subseteq x_g^*)$ using probabilities that only involve labels in $\mathcal{L}$. In our case, we propose to choose the most pessimistic likelihood given that a single label $l_k \in \ell$ is encoded in x_g^*:

  $$P(\mathbf{O}_g|\ell \subseteq x_g^*) \approx \min_{k|l_k \in \ell} \left\{P(\mathbf{O}_g|l_k \in x_g^*)\right\} \tag{9}$$

  $$\propto \min_{k|l_k \in \ell} \left\{P(l_k \in x_g^*|\mathbf{O}_g)\right\}. \tag{10}$$

  Observe that Eq. (10) follows from the assumption that labels in $\mathcal{L}$ are equiprobable. Since x_g^* tends to correspond to the ground-truth, $P(l_k \in x_g^*|\mathbf{O}_g)$ can be estimated as the probability that label l_k appears in the ground-truth knowing the global observation, *i.e.* $P(l_k \in X_g|\mathbf{O}_g)$.

It can be easily shown that if $\ell_1, \ell_2 \in \mathcal{P}(\mathcal{L})$, with $\ell_1 \subseteq \ell_2$, and assuming the approximations in Eq. (7) and Eq. (10), then

$$P(\ell_1 \subseteq x_g^*|\mathbf{O}) \geq P(\ell_2 \subseteq x_g^*|\mathbf{O}). \tag{11}$$

This property can be exploited to select the best M' labels with the maximum computed $P(\ell \subseteq x_g^*|\mathbf{O})$ by any branch-and-bound-like algorithm. If the branching is done by incrementing the number of encoded labels of each candidate label, then Eq. (11) can be used to massively prune away large sets of candidates.

4. Joint Classification and Segmentation

Now that we have described the general structure of the graphical model $\mathcal{G}$, in this section we address how we specialize it to solve the problem of joint segmentation and classification of images. For more implementation specificities we refer the reader to Section 5.1. The local nodes X_i represent the semantic labeling of superpixels (*i.e.* groups of pixels) obtained with an unsupervised segmentation method. Since all pixels inside a superpixel take the same label, an oversegmentation of the image is required in order to avoid having two different objects in the same superpixel. We establish a smoothness potential between local nodes that share a boundary in the unsupervised segmentation. The global node X_g represent the semantic classification of the whole image. It is connected by the harmony potential to each local node X_i.

The unary potentials $\phi(x_k)$ are derived from classification scores obtained using a bag-of-words representation over the region represented by the node. We differentiate between the computation of the unary potentials of the local nodes $\phi_L(x_k)$, where $i \in \mathcal{V}_L$, and of the global node $\phi_G(x_g)$.

Local Unary Potential. The unary potential of the local nodes is:

$$\phi_L(x_i) = -\mu_L K_i \omega_L(x_i) \log P(X_i = x_i|\mathbf{O}_i), \qquad (12)$$

where μ_L is the weighting factor of the local unary potential, K_i normalizes over the number of pixels inside superpixel i, and $\omega_L(x_i)$ is a learned per-class normalization. $P(X_i = x_i|\mathbf{O}_i)$ is the classification score given an observed representation $\mathbf{O}_i$ of the region, which is based on a bag-of-words built from features of superpixel i and those superpixels adjacent to it. These classifiers are trained for each label independently of the others.

The aim of $\omega_L(x_i)$ is to calibrate the confidence in the output of each classifier. One of the main drawbacks of learning independent classifiers for multi-class problems is that at some point we have to merge the classification scores. Since each classifier is trained independently, it is not taken into account the bias between classes. This effect is more noticeable when the number of training samples of each class is unbalanced. In the learning stage, we find the $\omega_L(x_i)$ that properly normalize each class. We found this to significantly improve results.

Often, object classifiers are trained to differentiate objects from one class from *any* other class. However, the harmony potential will already take care of penalizing the coexistence of objects from classes which are not likely to be in that image. Hence, the superpixel classifiers do not need to be so general, and can be specialized on discriminating between a certain object class and *just* those classes of objects which appear simultaneously in the same image. This means that the negative examples of the training set are just the superpixels of another class in the same image. In that way, the training data does not need to contain the variability encountered in the whole dataset.

In our implementation we have no nodes representing mid-level scales in the image. To also benefit from the *context at the mid-level* we investigated extending the representation at the local scale with mid-level context information. Fulkerson *et al.* [5] have shown that a single bag-of-words extracted not only inside of the superpixel, but also in the area adjacent to it, is able to better describe the superpixel. In this paper, we propose to consider two different bag-of-words: one for the superpixel and another for the region adjacent to it, in order to finally concatenate both descriptors. In that way, despite doubling the dimensionality of the descriptor, our method is able to be more discriminative than [5], specially at boundaries. In table 1 results of these two strategies are summarized, showing that concatenating mid-level information yields a 3% gain on the Pascal VOC 2009 dataset.

Distance (in pixels)	0	10	50	100
Summation [5]	20.0	24.7	25.5	24.1
Concatenation	20.0	26.5	27.6	27.7

Table 1. Results on the validation set of Pascal VOC 2009 for different mid-level context sizes (in pixels). See [4] for evaluation criteria details. Effect of the size of the area adjacent to the superpixel to build the bag-of-words, before per class normalization is applied.

Global Unary Potential. The global unary potential is defined as:

$$\phi_G(x_g) = -\mu_G \omega_G(x_g) \log P(X_g = x_g|\mathbf{O}_g), \qquad (13)$$

where μ_G is the weighting factor of the global unary potential, and $\omega_G(x_g)$ is again a per-class normalization like the one used in the local unary potential. The main difference comes from the computation of $P(X_g = x_g|\mathbf{O}_g)$, which is the posterior:

$$P(X_g = x_g|\mathbf{O}_g) \propto P(\mathbf{O}_g|X_g = x_g)P(X_g = x_g). \quad (14)$$

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.* $\sum_{I_i \in \mathcal{I}} \mathrm{T}[x_g = t_g^i]$. Since learning $P(\mathbf{O}_g|X_g = x_g)$ for each combination of labels is unfeasible, we employ the same approximation here as in Eq. (9) and Eq. (10), where we use the most pessimistic likelihood knowing which combinations of labels are present and which not:

$$P(\mathbf{O}_g|X_g = x_g) \propto \min \left\{ \min_{k|l_k \notin x_g} \{P(l_k \notin X_g|\mathbf{O}_g)\}, \right.$$
$$\left. \min_{k|l_k \in x_g} \{P(l_k \in X_g|\mathbf{O}_g)\} \right\}, \quad (15)$$

where $P(l_k \notin X_g|\mathbf{O}_g) = 1 - P(l_k \in X_g|\mathbf{O}_g)$. $P(l_k \in X_g|\mathbf{O}_g)$ is the classification score given the representation $\mathbf{O}_g$ of the whole image, which is based again on a bag-of-words. The per class normalization factor $\omega_G(x_g)$ is determined by the most pessimistic l_k, *i.e.* the l_k which gives the minimum of Eq. (15).

Smoothness Potential. The smoothness term is given by

$$\psi_L(x_i, x_j) = \lambda_L K_{ij} \theta(c_{ij}) \mathrm{T}[x_i \neq x_j] \qquad (16)$$

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 the superpixels indexed by 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 use a learned set of parameters θ as modulation costs. Thus, unlike other approaches, we discretize in several bins the possible values of c_{ij} and the parameter $\theta(c_{ij})$ is learned in order to establish a more accurate potential.

Harmony Potential. We take into account the classification score of the whole image $P(X_g = x_g|\mathbf{O}_g)$ to define the penalization applied by the consistency potential in Eq. (5):

$$\gamma_i^l = -\lambda_G K_i(\gamma_{min} + \omega_G(x_g)\log P(X_g = x_g|\mathbf{O}_g)) \quad (17)$$

where λ_G is the weighting factor of the consistency term, K_i normalizes over the number of pixels contained in the superpixel i, and γ_{min} is the minimum penalization applied which is set to 1.

Learning CRF Parameters. Learning the several parameters of the CRF potentials is a key step to attain state-of-the-art results on the labeling problem. We learn the parameters of the different potentials by iterating a two-step procedure until convergence. In the first step, we train the weighting factors of the potentials λ_G, λ_L, μ_L, μ_G, while in the second step we learn the local and global per class normalization $\omega_L(l)$ and $\omega_G(l)$. These two sets of parameters are quite decoupled, and this division reduces the size of the parameter space at each step. New samples are obtained with a simple Gibbs-like sampling algorithm, where we vary a single parameter at a time.

5. Experiments

We evaluate our method on two of the most challenging datasets for object class segmentation: the Pascal VOC 2009 Segmentation Challenge [4] and the MSCR-21 Dataset [20]. VOC 2009 contains 20 object classes plus the background class, MSCR-21 contains 21 classes. The Pascal dataset focusses on object recognition, and normally only one or few objects are present in the image, surrounded by background. On the other hand, the MSCR-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.

5.1. Implementation

Unsupervised Segmentation. Regions are created by oversegmenting the image with the quick-shift algorithm using the same parameters as Fulkerson *et al.* [5]. As it is stated in that paper, the results of this segmentation preserve nearly all the object boundaries. By working directly on the superpixels level instead of the pixel level, the number of nodes in the CRF is significantly reduced, typically from 10^5 to 10^2 per image. Therefore the inference algorithm converges drastically faster.

Local Classification Scores $P(X_i = x_i|\mathbf{O}_i)$. We extract patches over a grid with 50% of overlapping at several scales (12, 24, 36 and 48 pixels of diameter). These patches are described by shape features (SIFT) and by color features (RGB histogram). In the case of MSCR-21 the location feature is also added, using a 5×5 grid. In order to build a bag-of-words representation, we quantize with K-means the shape features to 1000 words and the color features to 400 words.

We use a different SVM classifier with intersection kernel [16] for each label to obtain the classification scores. Each classifier is learnt with a similar number of positive and negative examples: around a total of 8.000 superpixel samples for MSCR-21, and 20.000 for VOC 2009 for each class.

Global Classification Scores $P(X_g = x_g|\mathbf{O}_g)$. In the case of VOC 2009, the global classification score is based on a comprehensive image classification method. We use a bag-of-words representation [23], based on shape SIFT, color SIFT [21], together with spatial pyramids [13] and color attention [18]. Furthermore, the training of the global node only requires weakly labeled image data, and can therefore be done on a large set of 7054 images. In the case of MSCR-21, we use a simpler bag-of-word representation based on SIFT, RGB histograms and spatial pyramids [13]. In both methods, we use an SVM with χ^2 kernel as a classifier.

Inference. The optimal MAP configuration $\mathbf{x}^*$ can be inferred using any popular message passing or graph cut based algorithm. In all the experiments we use α-expansion graph cuts[2] [9]. The global node uses the M' first most probable labels obtained in the ranked subsampling. We set M' to a value such that no significant improvements are observed beyond it. We set $M' = 100$ for all experiments. The average time to do MAP inference for an image in MSCR-21 is 0.24 seconds and in VOC 2009 is 0.32 seconds.

CRF Parameters. Regarding the smoothness potential, we use seven bins to discretize c_{ij}, where for each c_{ij} bin we respectively set $\theta(c_{ij}) = \{200, 100, 80, 20, 10, 5, 0\}$.

In MSCR-21 we do the learning of the CRF parameters over a 5-fold cross-validation of the union of training and validation sets. The parameters obtained are λ_G = 10, λ_L = 0.7, μ_L = 7 , μ_G = 50. In the case of VOC 2009, we use the available validation set to train the CRF parameters. The parameters are $\lambda_G = 10$, $\lambda_L = 0.2$, $\mu_L = 5$, $\mu_G = 11$. Since the background class always appears in combination with other classes, we do not allow the harmony potential to apply any penalization to the background class.

5.2. Results

MSCR-21. In Table 2, our results are compared with other state-of-the-art methods. Furthermore, we show the results without consistency potentials and those obtained with robust P^N-based potentials.

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

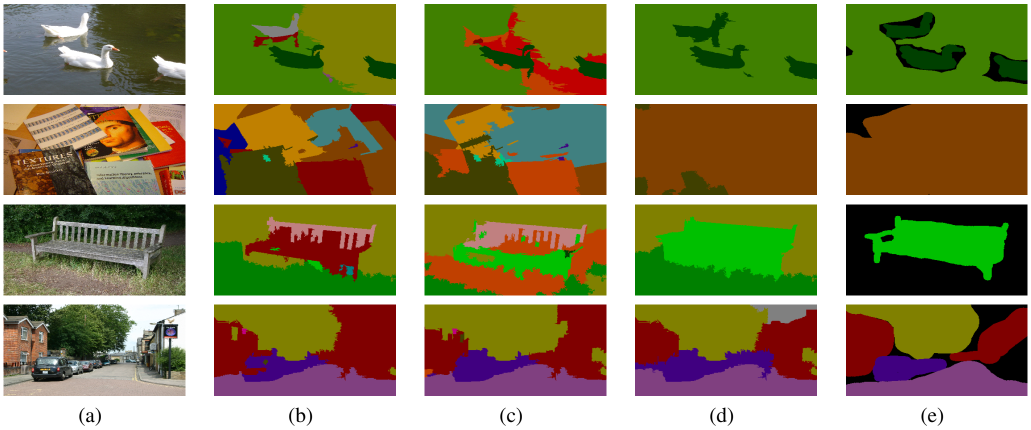


Figure 2. **Qualitative results for the MSCR-21 Dataset.** Comparison between (b) no consistency potentials, (c) robust P^N-based potentials, and(d) harmony potentials. (e) ground-truth images.

	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.* [19]	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 [7]	53	**97**	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13	78	68
Pixel-based CRF [12]	73	92	85	75	78	92	75	76	**86**	79	87	96	**95**	31	81	34	84	53	61	60	15	81	72
Hierarchical CRF *et al.* [12]	**80**	96	**86**	74	87	**99**	74	**87**	**86**	**87**	82	**97**	**95**	30	86	31	**95**	51	**69**	**66**	09	**86**	**75**
w/o Consistency Potential	70	92	83	67	54	93	66	71	63	64	82	66	39	28	75	42	83	60	26	52	12	74	61
Robust P^N Based Potentials	52	79	74	80	86	80	**88**	66	65	75	**97**	88	81	36	86	50	65	**85**	23	61	35	72	69
Harmony Potential	60	78	77	91	68	88	87	76	73	77	93	**97**	73	**57**	**95**	**81**	76	81	46	56	**46**	77	**75**

Table 2. **MSRC-21 segmentation results.** The average score provides the per-class average. The global scores gives the percentage of correctly classified pixels.

The results show that without consistency potentials we obtain a baseline of only 61% due to the simple features used, while [12] obtains a considerably better 72%. From our baseline, the harmony potentials are able to improve by 15%, whereas the robust P^N-based potentials by 8%. We believe this improvement is due to the fact that harmony potentials better model the heterogeneity of images. Our parameters are optimized on the per-class average score, on which we obtain state-of-the-art results. The results are especially remarkable for some of the difficult object classes such as birds and boats.

In Figure 2 we provide segmentation results for different potentials. In the first three rows the robust P^N-based potentials are unable to deal with noisy multiclass problems. By assigning the free label to the global node, no penalization is applied and the behavior is similar to using no consistency potential at all. The harmony potentials successfully exploit the more reliable estimates based on the whole image to improve classification at the lower scale.

Pascal VOC 2009. In Table 3 the results on the Pascal VOC 2009 datasets are summarized. We only provide the results of the winners BONN SVM-SEGM [2] and the method most similar to ours BROOKESMSRC [12], which obtained state-of-the-art results on MSRC-21. Our method obtained best results on 6 out of the 20 classes in the Pascal VOC 2009 challenge.

Our method outperforms the hierarchical CRF method on all but three classes. Comparing the results of the two methods on the two datasets, it can be observed that we outperform the hierarchical CRFs especially for the object classes in the MSRC-21 data set. Since the PASCAL challenge only contains object classes, we significantly outperform them on this set.

In Figure 3 segmentation results are depicted. The results show that harmony potentials are able to deal with multiclass images, partial occlusion, and to correctly classify the background.

6. Conclusions

We have 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.

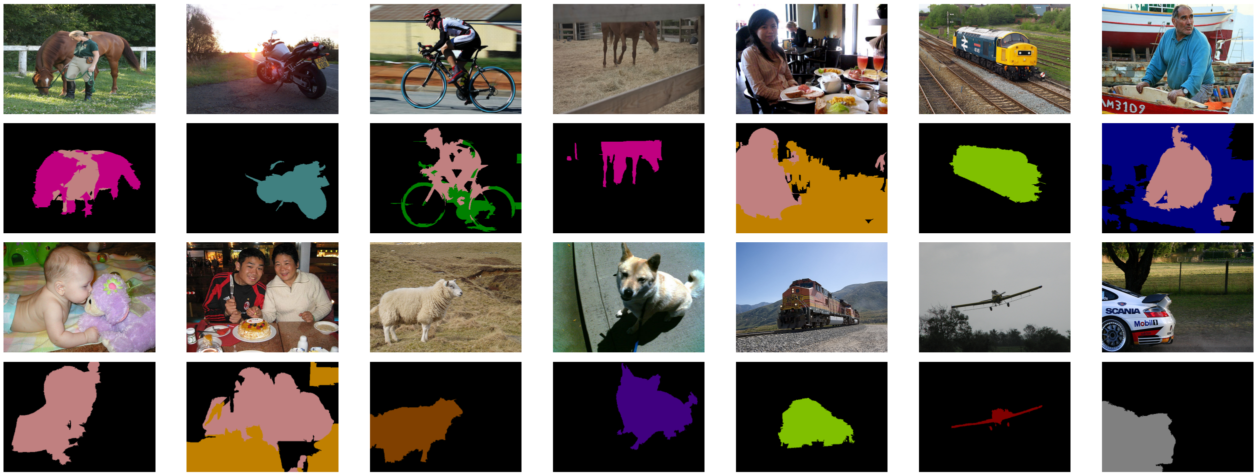


Figure 3. **Qualitative results of Pascal VOC 2009.** The original image (top) and our successful segmentation result (bottom).

	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**
BONN SVM-SEGM [2]	**83.9**	**64.3**	21.8	21.7	**32.0**	**40.2**	**57.3**	**49.4**	**38.8**	5.2	28.5	22.0	19.6	33.6	45.5	**33.6**	**27.3**	**40.4**	18.1	33.6	**46.1**	**36.3**
BROOKESMSRC AHCRF [12]	79.6	48.3	6.7	19.1	10.0	16.6	32.7	38.1	25.3	5.5	9.4	25.1	13.3	12.3	35.5	20.7	13.4	17.1	18.4	37.5	36.4	24.8
Harmony potential	80.5	62.3	**24.1**	**28.3**	30.5	32.7	42.2	48.1	22.8	**9.1**	**30.1**	7.9	**21.5**	**41.9**	**49.6**	31.5	26.1	37.0	**20.1**	**39.4**	31.1	34.1

Table 3. **Pascal VOC 2009 segmentation results.** Comparison with state-of-the-art methods. See [4] for evaluation criteria details. Note that these results are slightly different than those submitted for Pascal VOC Challenge 2009.

Acknowledgements

This work has been supported by the EU projects ERGTS-VICI-224737 and VIDI-VIDEO IST-045547; by the Spanish Research Program Consolider-Ingenio 2010: MIPRCV (CSD200700018); and by the Spanish projects TIN2009-14501-C02-02, TIN2009-14173 and TRA2007-62526/AUT. Joost van de Weijer acknowledges the support of a Ramon y Cajal fellowship, and Xavier Boix the support of the FPU fellowship AP2008-03378.

References

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9), 2004.

[2] J. Carreira and C. Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR*, 2010.

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

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[5] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[6] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 2008.

[7] J. Jiang and Z. Tu. Efficient scale space auto-context for image segmentation and labeling. In *CVPR*, 2009.

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

[9] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(9), 2004.

[10] M. P. Kumar, P. Torr, and A. Zisserman. Obj cut. In *CVPR*, 2005.

[11] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.

[12] L. Ladicky, C. Russell, P. Kohli, , and P. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[14] V. Lempitsky, P. Kohli, C. Rother, , and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[15] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[16] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[17] N. Plath, M. Toussaint, and S. Nakajima. Multi-class image segmentation using conditional random fields and global classification. In *ICML*, 2009.

[18] F. Shahbaz Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *ICCV*, 2009.

[19] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[20] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[21] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE PAMI*, 2010.

[22] J. Verbeek and B. Triggs. Scene segmentation with crfs learned from partially labeled images. In *NIPS*, 2008.

[23] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2), 2007.

[24] L. Zhu, Y. Chen, Y. Lin, C. Lin, and A. L. Yuille. Recursive segmentation and recognition templates for 2d parsing. In *NIPS*, 2008.