Interactive Segmentation with Super-Labels

Andrew Delong* Lena Gorelick* Frank R. Schmidt Olga Veksler Yuri Boykov

University of Western Ontario, Canada *authors contributed equally

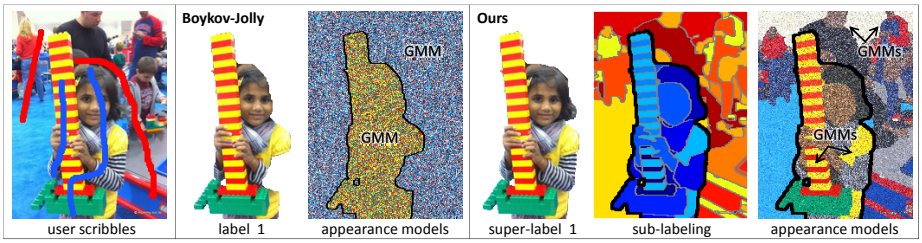


Fig. 1. Given user scribbles, typical MRF segmentation (Boykov-Jolly) uses a GMM to model the appearance of each object label. This makes the strong assumption that pixels inside each object are i.i.d. In contrast, we define a two-level MRF to encourage inter-object coherence among *super-labels* and intra-object coherence among *sub-labels*.

Abstract. In interactive segmentation, the most common way to model object appearance is by GMM or histogram, while MRFs are used to encourage spatial coherence among the object labels. This makes the strong assumption that pixels *within* each object are i.i.d. when in fact most objects have multiple distinct appearances and exhibit strong spatial correlation among their pixels. At the very least, this calls for an MRF-based appearance model within each object itself and yet, to the best of our knowledge, such a "two-level MRF" has never been proposed.

We propose a novel segmentation energy that can model complex appearance. We represent the appearance of each object by a set of distinct spatially coherent models. This results in a two-level MRF with "super-labels" at the top level that are partitioned into "sub-labels" at the bottom. We introduce the *hierarchical Potts* (hPotts) prior to govern spatial coherence within each level. Finally, we introduce a novel algorithm with EM-style alternation of proposal, α-expansion and re-estimation steps.

Our experiments demonstrate the conceptual and qualitative improvement that a two-level MRF can provide. We show applications in binary segmentation, multi-class segmentation, and interactive co-segmentation. Finally, our energy and algorithm have interesting interpretations in terms of semi-supervised learning.

1 Introduction

The vast majority of segmentation methods model object appearance by GMM or histogram and rely on some form of spatial regularization of the object labels. This includes interactive [1–3], unsupervised [4–9], binary [1–3, 8, 9] and multi-class [4–7, 10] techniques. The interactive methods make the strong assumption that all pixels within an entire object are i.i.d. when in fact many objects are composed of multiple regions with distinct appearances. Unsupervised methods try to break the image into small regions that actually *are* i.i.d., but these formulations do not involve any high-level segmentation of objects.

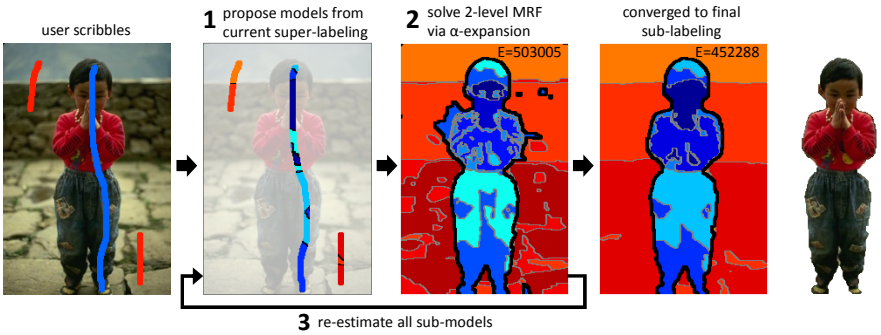


Fig. 2. We iteratively propose new models by randomly sampling pixels from super-labels, optimize the resulting two-level MRF, and re-estimate model parameters.

We propose a novel energy that unifies these two approaches by incorporating unsupervised learning into interactive segmentation. We show that this more descriptive object model leads to better high-level segmentations. In our formulation, each object (*super-label*) is automatically decomposed into spatially coherent regions where each region is described by a distinct appearance model (*sub-label*). This results in a two-level MRF with super-labels at the top level that are partitioned into sub-labels at the bottom. Figure 1 illustrates the main idea. We introduce the *hierarchical Potts* (hPotts) prior to govern spatial coherence at both levels of our MRF. The hierarchical Potts prior regularizes boundaries between objects (super-label transitions) differently from boundaries within each object (sub-label transitions). The unsupervised aspect of our MRF allows appearance models of arbitrary complexity and would severely over-fit the image data if left unregularized. We address this by incorporating global sparsity prior into our MRF via the energetic concept of "label costs" [7].

Since our framework is based on multi-label MRFs, a natural choice of optimization machinery is α-expansion [11, 7]. Furthermore, the number, class, and parameters of each object's appearance models are not known *a priori* — in order to use powerful combinatorial techniques we must propose a finite set of possibilities for α-expansion to select from. We therefore resort to an iterative graph-cut process that involves random sampling to propose new models, α-expansion to update the segmentation, and re-estimation to improve the current appearance models. Figure 2 illustrates our algorithm.

The remainder of the paper is structured as follows. Section 2 discusses other methods for modeling complex appearance, MDL-based segmentation, and related iterative graph-cut algorithms. Section 3 describes our energy-based formulation and algorithm in formal detail. Section 4 shows applications in interactive binary/multi-class segmentation and interactive co-segmentation; furthermore it describes how our framework easily allows appearance models to come from a mixture of classes (GMM, plane, *etc.*). Section 5 draws an interesting parallel between our formulation and multi-class semi-supervised learning in general.

2 Related Work

Complex appearance models. The DDMCMC method [6] was the first to emphasize the importance of representing object appearance with complex mod-

els (*e.g.* splines and texture based models in addition to GMMs) in the context of unsupervised segmentation. However, being unsupervised, DDMCMC does not delineate objects but rather provides low-level segments along with their appearance models. Ours is the first multi-label graph-cut based framework that can learn a mixture of such models for segmentation.

There is an interactive method [10] that decomposes objects into spatially coherent sub-regions with distinct appearance models. However, the number of sub-regions, their geometric interactions, and their corresponding appearance models must be carefully designed for each object of interest. In contrast, we automatically learn the number of sub-regions and their model parameters.

MDL-based segmentation. A number of works have shown that *minimum description length* (MDL) is a useful regularizer for unsupervised segmentation, *e.g.* [5–7]. Our work stands out here in two main respects: our formulation is designed for semi-supervised settings and explicitly weighs the benefit of each appearance model against the 'cost' of its inherent complexity (*e.g.* number of parameters). To the best of our knowledge, only the unsupervised DDMCMC [6] method allows arbitrary complexity while explicitly penalizing it in a meaningful way. However, they use a completely different optimization framework and, being unsupervised, they do not delineate object boundaries.

Iterative graph-cuts. Several energy-based methods have employed EM-style alternation between a graph-cut/α-expansion phase and a model re-estimation phase, *e.g.* [12, 2, 13, 14, 7]. Like our work, Grab-Cut [2] is about interactive segmentation, though their focus is binary segmentation with a bounding-box interaction rather than scribbles. The bounding box is intuitive and effective for many kinds of objects but often requires subsequent scribble-based interaction for more precise control. Throughout this paper, we compare our method to an iterative multi-label variant of Boykov-Jolly [1] that we call iBJ. Given user scribbles, this baseline method maintains one GMM per object label and iterates between α-expansion and re-estimating each model.

On an algorithmic level, the approach most closely related to ours is the unsupervised method [14, 7] because it also involves random sampling, α-expansion, and label costs. Our framework is designed to learn complex appearance models from partially-labeled data and differs from [14, 7] in the following respects: (1) we make use of hard constraints and the current super-labeling to guide random sampling, (2) our hierarchical Potts potentials regularize sub- and super-labels differently, and (3), again, our label costs penalize models based on their individual complexity rather than using uniform label costs.

3 Modeling Complex Appearance via Super-Labels

We begin by describing a novel multi-label energy that corresponds to our two-level MRF. Unlike typical MRF-based segmentation methods, our actual set of discrete labels (appearance models) is not precisely known beforehand and we need to estimate both the number of unique models and their parameters. Section 3.1 explains this energy formulation in detail, and Section 3.2 describes our iterative algorithm for minimizing this energy.

3.1 Problem Formulation

Let $\mathcal{S}$ denote the set of super-labels (scribble colors) available to the user and let $\mathcal{P}$ denote the indexes of pixels in the input image I. By "scribbling" on the image, the user interactively defines a partial labeling $g : \mathcal{P} \to \mathcal{S} \cup \{\text{none}\}$ that assigns to each pixel p a super-label index $g_p \in \mathcal{S}$ or leaves p unlabeled ($g_p = \text{none}$). Our objective in terms of optimization is to find the following:

1. an unknown set of $\mathcal{L}$ of distinct appearance models (sub-labels) generated from the image, along with model parameters θ_ℓ for each $\ell \in \mathcal{L}$
2. a complete sub-labeling $f : \mathcal{P} \to \mathcal{L}$ that assigns one model to each pixel, and
3. a map $\pi : \mathcal{L} \to \mathcal{S}$ where $\pi(\ell) = i$ associates sub-label ℓ with super-label i, *i.e.* the sub-labels are grouped into disjoint subsets, one for each super-label; any π defines a parent-child relation in what we call a two-level MRF.

Our output is therefore a tuple $(\mathcal{L}, \theta, \pi, f)$ with set of sub-labels $\mathcal{L}$, model parameters $\theta = \{\theta_\ell\}$, super-label association π, and complete pixel labeling f. The final segmentation presented to the user is simply $(\pi \circ f) : \mathcal{P} \to \mathcal{S}$ which assigns a scribble color (super-label index) to each pixel in $\mathcal{P}$.

In a good segmentation we expect the tuple $(\mathcal{L}, \theta, \pi, f)$ to satisfy the following three properties. First, the super-labeling $\pi \circ f$ must respect the constraints imposed by user scribbles, *i.e.* if pixel p was scribbled then we require $\pi(f_p) = g_p$. Second, the labeling f should exhibit spatial coherence both among sub-labels and between super-labels. Finally, the set of sub-labels $\mathcal{L}$ should contain as many appearance models as is justified by the image data, *but no more*.

We propose an energy for our two-level MRFs[1] that satisfies these three criteria and can be expressed in the following form[2]:

$$E(\mathcal{L}, \theta, \pi, f) \;=\; \sum_{p \in \mathcal{P}} D_p(f_p) \;+\; \sum_{pq \in \mathcal{N}} w_{pq} V(f_p, f_q) \;+\; \sum_{\ell \in \mathcal{L}} h_\ell \delta_\ell(f) \qquad (1)$$

The unary terms D of our energy express negative log-likelihoods of appearance models and enforce the hard constraints imposed by the user. A pixel p that has been scribbled ($g_p \in \mathcal{S}$) is only allowed to be assigned a sub-label ℓ such that $\pi(\ell) = g_p$. Un-scribbled pixels are permitted to take any sub-label.

$$D_p(\ell) \;=\; \begin{cases} -\ln \Pr(I_p | \theta_\ell) & \text{if } g_p = \text{none} \;\vee\; g_p = \pi(\ell) \\ \infty & \text{otherwise} \end{cases} \qquad (2)$$

The pairwise terms V are defined with respect to the current super-label map π as follows:

$$V(\ell, \ell') \;=\; \begin{cases} 0 & \text{if } \ell = \ell' \\ c_1 & \text{if } \ell \neq \ell' \text{ and } \pi(\ell) = \pi(\ell') \\ c_2 & \text{if } \pi(\ell) \neq \pi(\ell') \end{cases} \qquad (3)$$

We call (3) a *two-level Potts* potential because it governs coherence on two levels: c_1 encourages sub-labels within each super-label to be spatially coherent, and

[1] In practice we use non-uniform w_{pq} and so, strictly speaking, (1) is a *conditional random field* (CRF) [15] rather than an MRF.

[2] The dependence of D on (π, θ) and of V on π is omitted in (1) for clarity.

c_2 encourages smoothness among super-labels. This potential is a special case of our more general class of *hierarchical Potts* potentials introduced in Appendix 7, but two-level Potts is sufficient for our interactive segmentation applications. For image segmentation we assume $c_1 \leq c_2$, though in general any V with $c_1 \leq 2c_2$ is still a metric [11] and can be optimized by α-expansion. Appendix 7 gives general conditions for hPotts to be metric. It is commonly known that smoothness costs directly affect the expected length of the boundary and should be scaled proportionally to the size of the image. Intuitively c_2 should be larger as it operates on the entire image as opposed to smaller regions that correspond to objects. The weight $w_{pq} \geq 0$ of each pairwise term in (1) is computed from local image gradients in the standard way (*e.g.* [1, 2]).

Finally, we incorporate a model-dependent "label costs" [7] to regularize the number of unique models in $\mathcal{L}$ and their individual complexity. A label cost h_ℓ is a global potential that penalizes the use of ℓ in labeling f through indicator function $\delta_\ell(f) = 1 \Leftrightarrow \exists f_p = \ell$. There are many possible ways to define the weight h_ℓ of a label cost, such as Akaike information criterion (AIC) [16] or Bayesian information critierion (BIC) [17]. We use a heuristic described in Section 4.2.

3.2 Our SUPERLABELSEG Algorithm

We propose a novel segmentation algorithm based on the iterative PEARL framework [7]. Each iteration of PEARL has three main steps: propose candidate models by random sampling, segment via α-expansion with label costs, and re-estimate the model parameters for the current segmentation. Our algorithm differs from [7] as follows: (1) we make use of hard constraints g and the current super-labeling $\pi \circ f$ to guide random sampling, (2) our two-level Potts potentials regularize sub- and super-labels differently, and (3) our label costs penalize models based on their individual complexity rather than uniform penalty.

The proposal step repeatedly generates a new candidate model ℓ with parameters θ_ℓ fitted to a random subsample of pixels. Each model is proposed in the context of a particular super-label $i \in \mathcal{S}$, and so the random sample is selected from the set of pixels $\mathcal{P}_i = \{\, p \,|\, \pi(f_p) = i \,\}$ currently associated with i. Each candidate ℓ is then added to the current label set $\mathcal{L}$ with super-label assignment set to $\pi(\ell) = i$. A heuristic is used to determine a sufficient number of proposals to cover the set of pixels $\mathcal{P}_i$ at each iteration.

Once we have candidate sub-labels for every object a naïve approach would be to directly optimize our two-level MRF. However, being random, not all of an object's proposals are equally good for representing its appearance. For example, a proposal from a small sample of pixels is likely to over-fit or mix statistics (Figure 3, proposal 2). Such models are not characteristic of the object's overall appearance but are problematic because they may incidentally match some portion of another object and lead to an erroneous super-label segmentation. Before allowing sub-labels to compete over the entire image, we should do our best to ensure that all appearance models within each object are relevant and accurate.

Given the complete set of proposals, we first re-learn the appearance of each object $i \in \mathcal{S}$. This is achieved by restricting our energy to pixels that are currently labeled with $\pi(f_p) = i$ and optimizing via α-expansion with label costs [7];

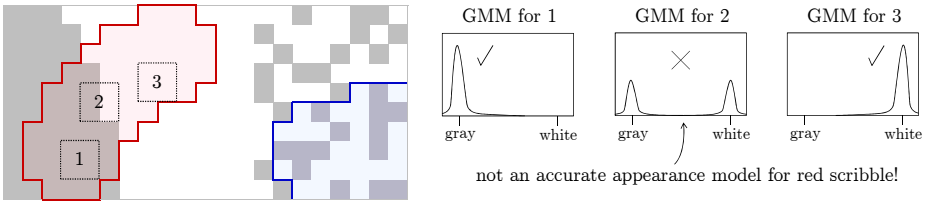


Fig. 3. The object marked with red has two spatially-coherent appearances: pure gray, and pure white. We can generate proposals for the red object from random patches 1–3. However, if we allow proposal 2 to remain associated with the red object, it may incorrectly claim pixels from the blue object which actually *does* look like proposal 2.

this ensures that each object is represented by an accurate set of appearance models. Once each object's appearance has been re-learned, we allow the objects to simultaneously compete for all image pixels while continuing to re-estimate their parameters. Segmentation is performed on a two-level MRF defined by the current $(\mathcal{L}, \theta, \pi)$. Again, we use α-expansion with label costs to select a good subset of appearance models and to partition the image. The pseudo-code below describes our SuperLabelSeg algorithm.

SuperLabelSeg(g) where $g : \mathcal{P} \to \mathcal{S} \cup \{\text{none}\}$ is a partial labeling

$\mathcal{L} = \{\}$ // empty label set with global f, π, θ undefined
Propose(g) // initialize $\mathcal{L}, \pi, \theta$ from user scribbles
repeat
    Segment($\mathcal{P}, \mathcal{L}$) // segment entire image using all available labels $\mathcal{L}$
    Propose($\pi \circ f$) // update $\mathcal{L}, \pi, \theta$ from current super-labeling
until converged

Propose(z) where $z : \mathcal{P} \to \mathcal{S} \cup \{\text{none}\}$

for each $i \in \mathcal{S}$
    $\mathcal{P}_i = \{\, p \,|\, z_p = i \,\}$ // set of pixels currently labeled with super-label i
    repeat sufficiently
        generate model ℓ with parameters θ_ℓ fitted to random sample from $\mathcal{P}_i$
        $\pi(\ell) = i$
        $\mathcal{L} = \mathcal{L} \cup \{\ell\}$
    end
    $\mathcal{L}_i = \{\, \ell \,|\, \pi(\ell) = i \,\}$
    Segment($\mathcal{P}_i, \mathcal{L}_i$) // optimize models and segmentation within super-label i

Segment($\hat{\mathcal{P}}, \hat{\mathcal{L}}$) where $\hat{\mathcal{P}} \subseteq \mathcal{P}$ and $\hat{\mathcal{L}} \subseteq \mathcal{L}$

let $f|_{\hat{\mathcal{P}}}$ denote current global labeling f restricted to $\hat{\mathcal{P}}$
repeat
    $f|_{\hat{\mathcal{P}}} = \text{argmin}_{\hat{f}} E(\hat{\mathcal{L}}, \theta, \pi, \hat{f})$ // segment by α-expansion with label costs [7]
                                                // where we optimize only on $\hat{f} : \hat{\mathcal{P}} \to \hat{\mathcal{L}}$
    $\mathcal{L} = \mathcal{L} \setminus \{\, \ell \in \hat{\mathcal{L}} \,|\, \delta_\ell(\hat{f}) = 0 \}$ // discard unused models
    $\theta = \text{argmin}_\theta E(\hat{\mathcal{L}}, \theta, \pi, \hat{f})$ // re-estimate each sub-model params
until converged

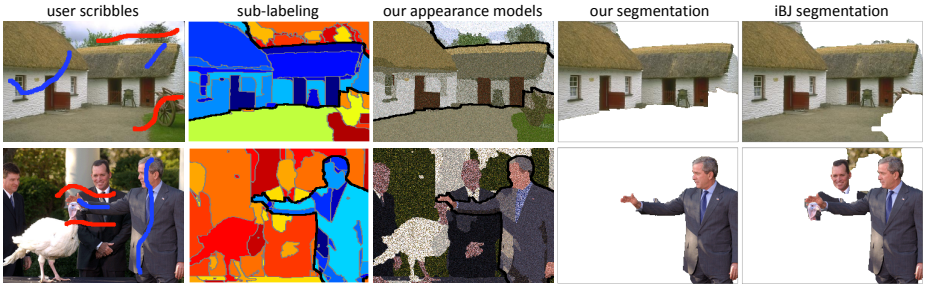


Fig. 4. Binary segmentation examples. The second column shows our final sub-label segmentation f where blues indicate foreground sub-labels and reds indicate background sub-labels. The third column is generated by sampling each I_p from model θ_{f_p}. The last two columns compare our super-label segmentation $\pi \circ f$ and iBJ.

4 Applications and Experiments

Our experiments demonstrate the conceptual and qualitative improvement that a two-level MRF can provide. We are only concerned with scribble-based MRF segmentation, an important class of interactive methods. We use an iterative variant of Boykov-Jolly [1] (iBJ) as a representative baseline because it is simple, popular, and exhibits a problem characteristic to a wide class of standard methods. By using one appearance model per object, such methods implicitly assume that pixels within each object are i.i.d. with respect to its model. However, this is rarely the case, as objects often have multiple distinct appearances and exhibit strong spatial correlation among their pixels. The main message of all the experiments is to show that by using multiple distinctive appearance models per object we are able to reduce uncertainty near the boundaries of objects and thereby improve segmentation in difficult cases. We show applications in interactive binary/multi-class segmentation and interactive co-segmentation.

Implementation details: In all our experiments we used publicly available α-expansion code [11, 7, 4, 18]. Our non-optimized matlab implementation takes on the order of one to three minutes depending on the size of the image, with the majority of time spent on re-estimating the sub-model parameters. We used the same within- and between- smoothness costs ($c_1 = 5, c_2 = 10$) in all binary, multi-class and co-segmentation experiments. Our proposal step uses distance-based sampling within each super-label whereby patches of diameter 3 to 5 are randomly selected. For re-estimating model parameters we use the Matlab implementation of EM algorithm for GMMs and we use PCA for planes. We regularize GMM covariance matrices to avoid overfitting in (L, a, b) color space by adding constant value of 2.0 to the diagonal.

4.1 Binary Segmentation

For binary segmentation we assume that the user wishes to segment an object from the background where the set of super-labels (scribble indexes) is defined by $\mathcal{S} = \{F, B\}$. In this specific case we found that most of the user interaction is spent on removing disconnected false-positive object regions by scribbling over them with background super-label. We therefore employ a simple heuristic: after

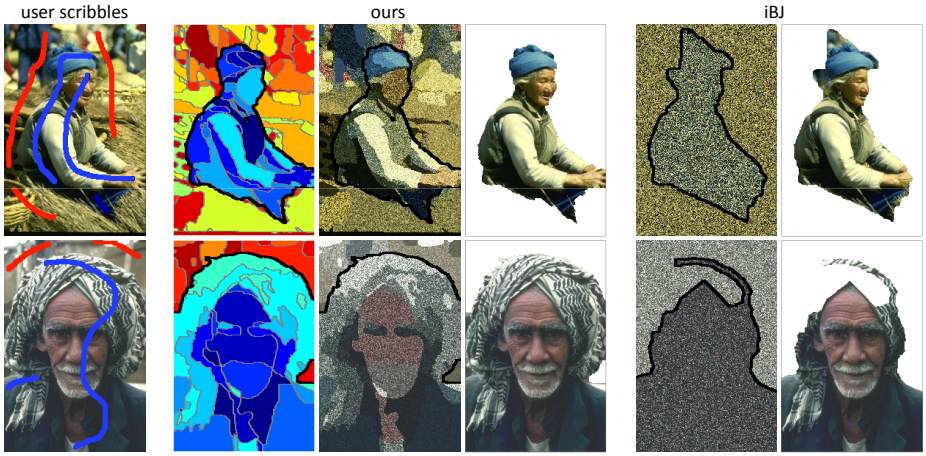


Fig. 5. More binary segmentation results showing scribbles, sub-labelings, synthesized images, and final cut-outs.

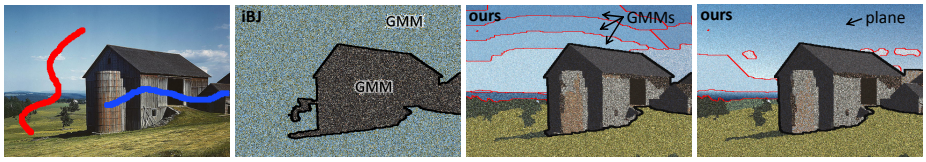


Fig. 6. An image exhibiting gradual changes in color. Columns 2–4 show colors sampled from the learned appearance models for iBJ, our two-level MRF restricted to GMMs only, and ours with both GMMs and planes. Our framework can detect a mix of GMMs (grass,clouds) and planes (sky) for the background super-label (top-right).

convergence we find foreground connected components that are not supported by a scribble and modify their data-terms to prohibit those pixels from taking the super-label F. We then perform one extra segmentation step to account for the new constraints. We apply this heuristic in all our binary segmentation results for both SuperLabelSeg and iBJ (Figures 4, 5, 6). Other heuristics could be easily incorporated in our energy to encourage connectivity, *e.g.* star-convexity [19, 20].

In Figure 4, top-right, notice that iBJ does not incorporate the floor as part of the background. This is because there is only a small proportion of floor pixels in the red scribbles, but a large proportion of a similar color (roof) in the blue scribbles. By relying directly on the color proportions in the scribbles, the learned GMMs do not represent the actual appearance of the objects in the full image. Therefore the ground is considered *a priori* more likely to be explained by the (wrong) roof color than the precise floor color, giving an erroneous segmentation despite the hard constraints. Our methods relies on spatial coherence of the distinct appearances within each object and therefore has a sub-label that fits the floor color tightly. This same phenomenon is even more evident in the bottom row of Figure 5. In the iBJ case, the appearance model for the foreground mixes the statistics from all scribbled pixels and is biased towards the most dominant color. Our decomposition allows each appearance with spatial support (textured fabric, face, hair) to have good representation in the composite foreground model.

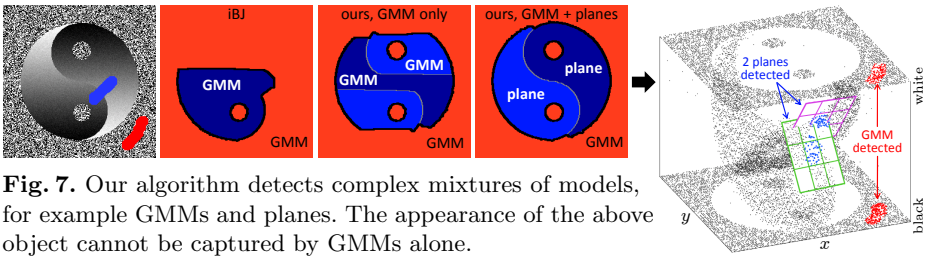


Fig. 7. Our algorithm detects complex mixtures of models, for example GMMs and planes. The appearance of the above object cannot be captured by GMMs alone.

4.2 Complex Appearance Models

In natural images objects often exhibit gradual change in hue, tone or shades. Modeling an object with a single GMM in color space [1, 2] makes the implicit assumption that appearance is piece-wise constant. In contrast, our framework allows us to decompose an object into regions with distinct appearance models, each from an arbitrary class (*e.g.* GMM, plane, quadratic, spline). Our algorithm will choose automatically the most suitable class for each sub-label within an object. Figure 6 (right) shows such an example where the background is decomposed into several grass regions, each modeled by a GMM in (L, a, b) color space, and a sky region that is modeled by a plane[3] in a 5-dimensional (x, y, L, a, b) space. Note the gradual change in the sky color and how the clouds are segmented as a separate 'white' GMM.

Figure 7 show a synthetic image, in which the foreground object breaks into two sub-regions, each exhibiting a different type of gradual change in color. This kind of object appearance cannot be captured by a mixture of GMM models. In general our framework can incorporate a wide range of appearance models as long as there exists a black-box algorithm for estimating the parameters θ_l, which can be used at the line **6** of SEGMENT. The importance of more complex appearance models was proposed by DDMCMC [6] for unsupervised segmentation in a completely different algorithmic framework. Ours is the first multi-label graph-cut based framework that can incorporate such models.

Because our appearance models may be arbitrarily complex, we must incorporate individual model complexity in our energy. Each label cost h_ℓ is computed based on the number of parameters θ_ℓ and the number ν of pixels that are to be labeled, namely $h_\ell = \frac{1}{2}\sqrt{\nu}|\theta_\ell|$. We set $\nu = \#\{\, p \,|\, g_p \neq \text{none} \,\}$ for line **2** of the SUPERLABELSEG algorithm and $\nu = |\mathcal{P}|$ for lines **4**,**5**. Penalizing complexity is a crucial part of our framework because it helps our MRF-based models to avoid over-fitting. Label costs balance the number of parameters required to describe the models against the number of data points to which the models are fit. When re-estimating the parameters of a GMM we allow the number of components to increase or decrease if favored by the overall energy.

4.3 Multi-Class Segmentation

Interactive multi-class segmentation is a straight-forward application of our energy (1) where the set of super-labels $\mathcal{S}$ contains an index for each scribble color.

[3] In color images a 'plane' is a 2-D linear subspace (a *2-flat*) of a 5-D image space.

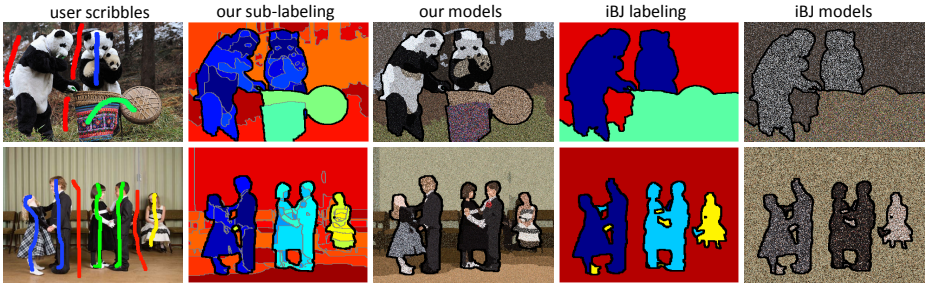


Fig. 8. Multi-class segmentation examples. Again, we show the color-coded sub-labelings and the learned appearance models. Our super-labels are decomposed into spatially coherent regions with distinct appearance.

Figure 8 shows examples of images with multiple scribbles corresponding to multiple objects. The resulting sub-labelings show how objects are decomposed into regions with distinct appearances. For example, in the top row, the basket is decomposed into a highly-textured colorful region (4-component GMM) and a more homogeneous region adjacent to it (2-component GMM). In the bottom row, notice that the hair of children marked with blue was so weak in the iBJ appearance model that it was absorbed into the background. The synthesized images suggest the quality of the learned appearance models. Unlike the binary case, here we do not apply the post-processing step enforcing connectivity.

4.4 Interactive Co-segmentation

Our two-level MRFs can be directly used for interactive co-segmentation [3, 21]. Specifically, we apply our method to co-segmentation of a collection of similar images as in [3] because it is a natural scenario for many users. This differs from 'unsupervised' binary co-segmentation [8, 9] that assumes dissimilar backgrounds and similar-sized foreground objects. Figure 9 shows a collection of four images with similar content. Just by scribbling on one of the images our method is able to correctly segment the objects. Note that the unmarked images contain background colors not present in the scribbled image, yet our method was able to detect these novel appearances and correctly segment the background into sub-labels.

5 Discussion: Super-Labels as Semi-Supervised Learning

There are evident parallels between interactive segmentation and semi-supervised learning, particularly among graph cut methods ([1] versus [22]) and random walk methods ([23] versus [24]). An insightful paper by Duchenne *et al.* [25] explicitly discusses this observation. Looking back at our energy and algorithm from this perspective, it is clear that we actually do semi-supervised learning applied to image segmentation. For example, the grayscale image in Figure 7 can be visualized as points in a 3D feature space where small subsets of points have been labeled either blue or red. In addition to making 'transductive' inferences, our algorithm automatically learned that the blue label is best decomposed into two

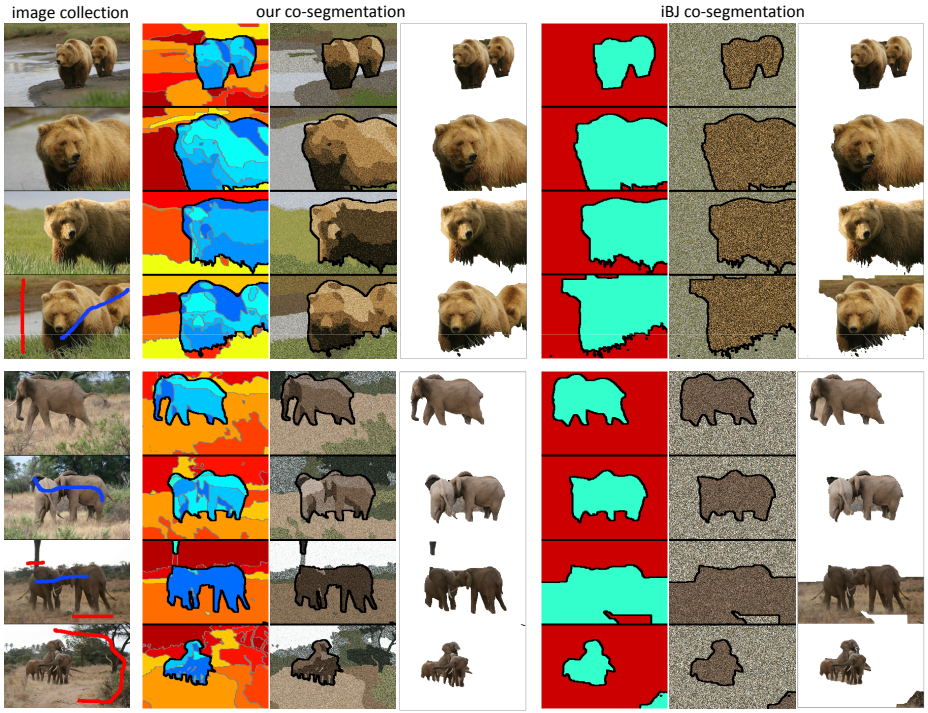


Fig. 9. Interactive co-segmentation examples. Note that our method detected sub-models for grass, water, and sand in the 1st and 3rd bear images; these appearances were not present in the scribbled image.

linear subspaces (green & purple planes in Figure 7, right) whereas the red label is best described by a single bi-modal GMM. The number, class, and parameters of these models was not known *a priori* but was discovered by SuperLabelSeg.

Our two-level framework allows each object to be modeled with arbitrary complexity but, crucially, we use spatial coherence (smooth costs) and label costs to regularize the energy and thereby avoid over-fitting. Setting $c_1 < c_2$ in our smooth costs V corresponds to a "two-level clustering assumption," *i.e.* that class clusters are better separated than the sub-clusters within each class. To the best of our knowledge, we are first to suggest iterated random sampling and α-expansion with label costs (SuperLabelSeg) as an algorithm for multi-class semi-supervised learning. These observations are interesting and potentially useful in the context of more general semi-supervised learning.

6 Conclusion

In this paper we raised the question of whether GMM/histograms are an appropriate choice for modeling object appearance. If GMMs and histograms are not satisfying generative models for a natural image, they are equally unsatisfying for modeling appearance of complex objects within the image.

To address this question we introduced a novel energy that models complex appearance as a two-level MRF. Our energy incorporates both elements of interactive segmentation and unsupervised learning. Interactions are used to

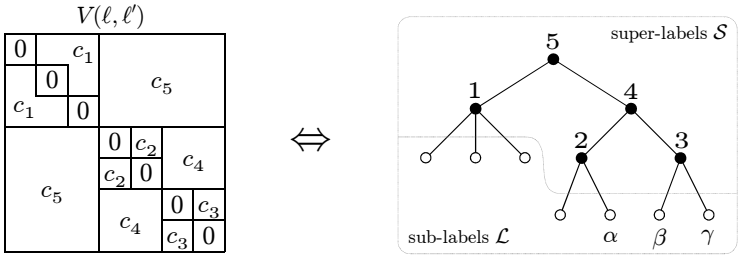
provide high-level knowledge about objects in the image, whereas the unsupervised component tries to learn the number, class and parameters of appearance models within each object. We introduced the hierarchical Potts prior to regularize smoothness within and between the objects in our two-level MRF, and we use label costs to account for the individual complexity of appearance models. Our experiments demonstrate the conceptual and qualitative improvement that a two-level MRF can provide.

Finally, our energy and algorithm have interesting interpretations in terms of semi-supervised learning. In particular, our energy-based framework can be extended in a straight-forward manner to handle general semi-supervised learning with ambiguously-labeled data [26]. We leave this as future work.

7 Appendix — Hierarchical Potts

In this paper we use two-level Potts potentials where the smoothness is governed by two coefficients, c_1 and c_2. This concept can be generalized to a *hierarchical Potts* (hPotts) potential that is useful whenever there is a natural hierarchical grouping of labels. For example, the recent work on hierarchical context [27] learns a tree-structured grouping of the class labels for object detection; with hPotts potentials it is also possible to learn pairwise interactions for segmentation with hierarchical context. We leave this as future work.

We now characterize our class of hPotts potentials and prove necessary and sufficient conditions for them to be optimized by the α-expansion algorithm [11]. Let $\mathcal{N} = \mathcal{L} \cup \mathcal{S}$ denote combined set of sub-labels and super-labels. A hierarchical Potts prior V is defined with respect to an irreducible[4] tree over node set $\mathcal{N}$. The parent-child relationship in the tree is determined by $\pi : \mathcal{N} \to \mathcal{S}$ where $\pi(\ell)$ gives the parent[5] of ℓ. The leaves of the tree are the sub-labels $\mathcal{L}$ and the interior nodes are the super-labels $\mathcal{S}$. Each node $i \in \mathcal{S}$ has an associated Potts coefficient c_i for penalizing sub-label transitions that cross from one sub-tree of i to another. An hPotts potential is a special case of general pairwise potentials over $\mathcal{L}$ and can be written as an $|\mathcal{L}| \times |\mathcal{L}|$ "smooth cost matrix" with entries $V(\ell, \ell')$. The coefficients of this matrix are block-structured in a way that corresponds to some irreducible tree. The example below shows an hPotts potential V and its corresponding tree.



Let $\pi^n(\ell)$ denote n applications of the parent function as in $\pi(\cdots\pi(\ell))$. Let $\text{lca}(\ell, \ell')$ denote the lowest common ancestor of ℓ and ℓ', *i.e.* $\text{lca}(\ell, \ell') = i$ where

[4] A tree is irreducible if all its internal nodes have at least two children.

[5] The root of the tree $r \in S$ is assigned $\pi(r) = r$.

$i = \pi^n(\ell) = \pi^m(\ell')$ for minimal n, m. We can now define an hPotts potential as

$$V(\ell, \ell') \;=\; c_{\mathrm{lca}(\ell,\ell')} \tag{4}$$

where we assume $V(\ell, \ell) = c_\ell = 0$ for each leaf $\ell \in \mathcal{L}$. For example, in the tree illustrated above $\mathrm{lca}(\alpha, \beta)$ is super-label 4 and so the smooth cost $V(\alpha, \beta) = c_4$.

Theorem 1. *Let V be an hPotts potential with corresponding irreducible tree π.*

$$V \textit{ is metric on } \mathcal{L} \quad \Longleftrightarrow \quad c_i \leq 2c_j \;\; \textit{for all} \;\; j = \pi^n(i). \tag{5}$$

Proof. The metric constraint $V(\beta, \gamma) \leq V(\alpha, \gamma) + V(\beta, \alpha)$ is equivalent to

$$c_{\mathrm{lca}(\beta,\gamma)} \leq c_{\mathrm{lca}(\alpha,\gamma)} + c_{\mathrm{lca}(\beta,\alpha)} \tag{6}$$

for all $\alpha, \beta, \gamma \in \mathcal{L}$. Because π defines a tree structure, for every α, β, γ there exists $i, j \in \mathcal{S}$ such that, without loss of generality,

$$\begin{aligned} &j = \mathrm{lca}(\alpha, \gamma) = \mathrm{lca}(\beta, \alpha), \text{ and} \\ &i = \mathrm{lca}(\beta, \gamma) \;\text{ such that }\; j = \pi^k(i) \text{ for some } k \geq 0. \end{aligned} \tag{7}$$

In other words there can be up to two unique lowest common ancestors among (α, β, γ) and we assume ancestor i is in the sub-tree rooted at ancestor j, possibly equal to j. For any particular (α, β, γ) and corresponding (i, j) inequality (6) is equivalent to $c_i \leq 2c_j$. Since π defines an irreducible tree, for each (i, j) there must exist corresponding sub-labels (α, β, γ) for which (6) holds. It follows that $c_i \leq 2c_j$ holds for all pairs $j = \pi^k(i)$ and completes the proof of (5). ■

References

1. Boykov, Y., Jolly, M.P.: Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. Int'l Conf. on Computer Vision (ICCV) **1** (2001) 105–112
2. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. In: ACM SIGGRAPH. (2004)
3. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: iCoseg: Interactive Co-segmentation with Intelligent Scribble Guidance. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2010)
4. Kolmogorov, V., Zabih, R.: What Energy Functions Can Be Optimized via Graph Cuts. IEEE Trans. on Patt. Analysis and Machine Intelligence **26** (2004) 147–159
5. Zhu, S.C., Yuille, A.L.: Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence **18** (1996) 884–900
6. Tu, Z., Zhu, S.C.: Image Segmentation by Data-Driven Markov Chain Monte Carlo. IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002) 657–673
7. Delong, A., Osokin, A., Isack, H., Boykov, Y.: Fast Approximate Energy Minimization with Label Costs. Int'l Journal of Computer Vision (2011) (in press).

8. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of Image Pairs by Histogram Matching - Incorporating a Global Constraint into MRFs. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2006)
9. Vicente, S., Kolmogorov, V., Rother, C.: Cosegmentation Revisited: Models and Optimization. In: European Conf. on Computer Vision (ECCV). (2010)
10. Delong, A., Boykov, Y.: Globally Optimal Segmentation of Multi-Region Objects. In: Int'l Conf. on Computer Vision (ICCV). (2009)
11. Boykov, Y., Veksler, O., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. IEEE Trans. on Pattern Analysis and Machine Intelligence (2001)
12. Birchfield, S., Tomasi, C.: Multiway cut for stereo and motion with slanted surfaces. In: Int'l Conf. on Computer Vision. (1999)
13. Zabih, R., Kolmogorov, V.: Spatially Coherent Clustering with Graph Cuts. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2004)
14. Isack, H.N., Boykov, Y.: Energy-based Geometric Multi-Model Fitting. Int'l Journal of Computer Vision (2011) (accepted).
15. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Int'l Conf. on Machine Learning (ICML). (2001)
16. Akaike, H.: A new look at statistical model identification. IEEE Trans. on Automatic Control **19** (1974) 716–723
17. Schwarz, G.: Estimating the Dimension of a Model. Annals of Statistics **6** (1978) 461–646
18. Boykov, Y., Kolmogorov, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. IEEE Trans. on Pattern Analysis and Machine Intelligence **29** (2004) 1124–1137
19. Veksler, O.: Star Shape Prior for Graph-Cut Image Segmentation. In: European Conf. on Computer Vision (ECCV). (2008)
20. Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A.: Geodesic Star Convexity for Interactive Image Segmentation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2010)
21. Schnitman, Y., Caspi, Y., Cohen-Or, D., Lischinski, D.: Inducing Semantic Segmentation from an Example. In: Asian Conf. on Computer Vision (ACCV). (2006)
22. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: Int'l Conf. on Machine Learning. (2001)
23. Grady, L.: Random Walks for Image Segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence **28** (2006) 1768–1783
24. Szummer, M., Jaakkola, T.: Partially labeled classication with markov random walks. In: Advances in Neural Information Processing Systems (NIPS). (2001)
25. Duchenne, O., Audibert, J.Y., Keriven, R., Ponce, J., Segonne, F.: Segmentation by transduction. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2008)
26. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from Ambiguously Labeled Images. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2009)
27. Choi, M.J., Lim, J., Torralba, A., Willsky, A.S.: Exploiting Hierarchical Context on a Large Database of Object Categories. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2010)