Associative Hierarchical Random Fields

Ľubor Ladický[1], Chris Russell[1], Pushmeet Kohli, Philip H. S. Torr

Abstract—This paper makes two contributions: the first is the proposal of a new model – the associative hierarchical random field (AHRF), and a novel algorithm for its optimisation; the second is the application of this model to the problem of semantic segmentation.

Most methods for semantic segmentation are formulated as a labelling problem for variables that might correspond to either pixels or segments such as super-pixels. It is well known that the generation of super pixel segmentations is not unique. This has motivated many researchers to use multiple super pixel segmentations for problems such as semantic segmentation or single view reconstruction. These super-pixels have not yet been combined in a principled manner, this is a difficult problem, as they may overlap, or be nested in such a way that the segmentations form a segmentation tree. Our new hierarchical random field model allows information from all of the multiple segmentations to contribute to a global energy. MAP inference in this model can be performed efficiently using powerful graph cut based move making algorithms.

Our framework generalises much of the previous work based on pixels or segments, and the resulting labellings can be viewed both as a detailed segmentation at the pixel level, or at the other extreme, as a segment selector that pieces together a solution like a jigsaw, selecting the best segments from different segmentations as pieces. We evaluate its performance on some of the most challenging data sets for object class segmentation, and show that this ability to perform inference using multiple overlapping segmentations leads to state-of-the-art results.

Index Terms—Conditional Random Fields, Discrete Energy Minimisation, Object Recognition and Segmentation.

1 Introduction

Semantic segmentation involves the assignment of a 'semantic class' label – such as *person*, or *road*, to every pixel in the image. Until recently, image labelling problems have been formulated using pairwise random fields. However, these pairwise fields are unable to capture the higher-order statistics of natural images which can be used to enforce the coherence of regions in the image or to encourage particular regions to belong to a certain class. Despite these limitations, the use of pairwise models is prevalent in vision. This can largely be attributed to the pragmatism of computer vision researchers; although such models do not fully capture image statistics, they serve as an effective discriminative model that prevents individual pixels from being mislabelled.

Many approaches to multi-scale vision have been proposed where either inference is performed in a 'top-down' approach, i.e. an initial scene based estimates is made, followed by the successive labelling of smaller regions that must be consistent with the initial labelling; or they take a bottom-up approach where starting from the labelling of small regions they progressively assign larger regions a labelling consistent with the small regions.

In this work we propose a novel formulation for multi-scale vision, as captured by a hierarchy of segmentations. This allows for the integration of cues defined at any scale, or over any arbitrary region of the image, and provides a generalisation of many of the segmentation methods prevalent in vision. Our approach provides a unification of the 'top-down' and 'bottom-up' approaches to common to many problems of computer vision. To do this, we propose a new model: the associative hierarchical random field (AHRF) and show how it can be solved efficiently using graph-cut based move-making algorithms.

If two AHRFs are added together to produce a new cost, the resulting cost is also an AHRF, and the sum can also be solved effectively. This allows many different potentials and cues to be incorporated within the model, and for inference to remain practical. This flexibility contributes substantially to the success of our method, and allows us to obtain state of the art results on most existing segmentation data sets. Thus this work contributes both to the general area of random fields, and in its application to computer vision. Next, we shall give some background to the problem of semantic segmentation in computer vision.

1.1 Semantic Segmentation

Over the last few years many different methods have been proposed for semantic segmentation *i.e.* the problem of assigning a set of given object labels such

- *Ľubor Ladický is with the Computer Vision and Geometry lab at the Zürich, Zürich, Switzerland*
- *Chris Russell is with the Department of Computer Science at the University College London, University of London, London, U.K.*
- *Pushmeet Kohli is with the Microsoft Research, Cambridge, U.K.*
- *Philip H. S. Torr is with the Department of Computing, Oxford Brookes University, Oxford, U.K.*

[1] The authors assert equal contribution and joint first authorship

This work is supported by EPSRC research grants, HMGCC, the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. P. H. S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

as *person*, *car*, or *road* to each pixel of a given image, in a manner consistent with human annotations.

Most methods for semantic segmentation are formulated in terms of pixels (Shotton et al, 2006), other methods used segments (Batra et al, 2008; Galleguillos et al, 2008; Yang et al, 2007), groups of segments (Rabinovich et al, 2007), or the intersections of multiple segmentations (Pantofaru et al, 2008), while some have gone to the extreme of looking at the whole image in order to reason about object segmentation (Larlus and Jurie, 2008).

Each choice of image representation comes with its share of advantages and disadvantages. Pixels might be considered the most obvious choice of segmentation. However, pixels by themselves contain a limited amount of information. The colour and intensity of a lone pixel is often not enough to determine its correct object label. Ren and Malik (2003)'s remark that *'pixels are not natural entities; they are merely a consequence of the discrete representation of images'* captures some of the problems of pixel-based representations.

The last few years have seen a proliferation of super-pixel (Comaniciu and Meer, 2002; Felzenszwalb and Huttenlocher, 2004; Shi and Malik, 2000) based methods, that perform an initial *a priori* segmentation of the image, applied to object segmentation (Batra et al, 2008; Galleguillos et al, 2008; He et al, 2006; Russell et al, 2006; Yang et al, 2007), and elsewhere (Hoiem et al, 2005; Tao et al, 2001). These rely upon an initial partitioning of the image, typically based upon a segmentation of pixels based upon spatial location and colour/texture distribution. This clustering of the image allows the computation of powerful region-based features which are partially invariant to scale (Wang et al, 2005).

Super-pixel based methods work under the assumption that some segments share boundaries with objects in an image. This is not always the case, and this assumption may result in dramatic errors in the labelling (see figure 1). A number of techniques have been proposed to overcome errors in super-pixels. Rabinovich et al (2007) suggested finding the most stable segmentation from a large collection of multiple segmentations in the hope that these would be more consistent with object boundaries. Larlus and Jurie (2008) proposed an approach to the problem driven by object detection. In their algorithm, rectangular regions are detected using a bag-of-words model based upon affine invariant features. These rectangles are refined using graph cuts to extract boundaries in a manner similar to (Rother et al, 2004). Such approaches face difficulties in dealing with cluttered images, in which multiple object classes intersect. Pantofaru et al (2008) observed that although segments may not be consistent with object boundaries, the segmentation map formed by taking the intersections of multiple segmentations often is. They proposed finding the most probable labelling of intersections of segments based upon the features of their parent segments. This scheme effectively reduces the size of super-pixels. It results in more consistent segments but with a loss in the information content and discriminative power associated with each segment.

We shall show that each of these models are AHRFs, and that we are free to combine them additively and solve the resulting AHRF.

1.2 Choosing the Correct Segmentation

An earlier approach to dealing with the difficultly of choosing good super-pixels is to delay their choice until much later, and picking super-pixels that are consistent with a 'good' labelling of the image. Gould et al (2009a) proposed an approach in which the choice of super-pixels was integrated with the labelling of the image with object instances. Under their interpretation, super-pixels should physically exist and represent either the entirety of an object or a planar facet if the object class is amorphous and can not be decomposed into individual objects (this includes classes such as *grass*, *building*, or *sky*). Consequently, in their final labelling each pixel belongs to exactly one super-pixel chosen to represent a single instance of an object.

This process of shaping super-pixels to match object outlines is computationally challenging. As discussed in Gould et al (2009b), the optimisation techniques proposed frequently fail to recognise individual instances. Their algorithm is often unable to merge the super-pixels contained within a single instance, even if the super-pixels are correctly labelled by class. The recent work by Kumar and Koller (2010) goes some way to addressing these issues. By using sophisticated LP-relaxations they are able to trade computation time against the quality of the solution found.

Another method to overcome these issues was proposed by (Kohli et al, 2008). By formulating the labelling problem as a higher-order random field defined over pixels, they were able to recover from misleading segments which spanned multiple object classes. Further, they were able to encourage individual pixels within a single segment to share the same label by defining higher-order potentials (functions defined over cliques of size greater than 2) that penalised inhomogeneous labellings of segments. Their method can be understood as a relaxation of the hard constraint of previous methods, that the image labelling must follow super-pixel boundaries, to a softer constraint in which a penalty is paid for nonconformance.

1.3 Overview of our Model

In this paper we propose a novel associative hierarchical random field formulation of semantic segmentation that allows us to combine models defined over different choices of super-pixel, avoiding the need

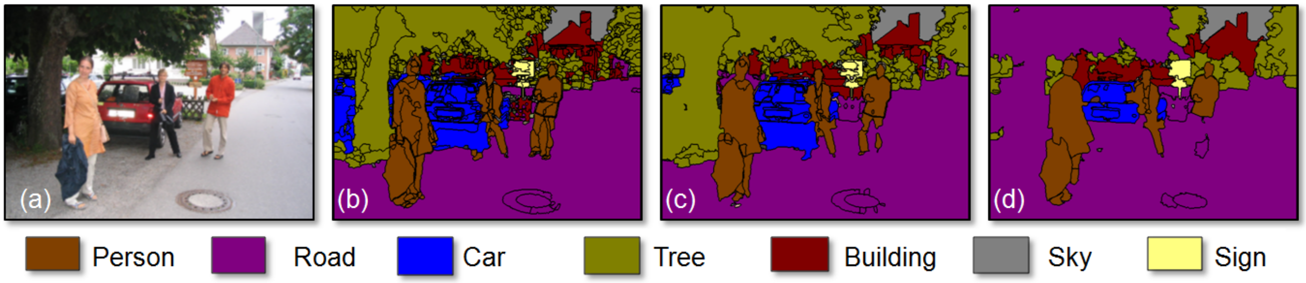


Fig. 1. *Multiple unsupervised image segmentations. (a) Original image. (b)-(d) Unsupervised image segmentations with different size super-pixels. (b), (c) and (d) use three different unsupervised segmentations of the image, in this case mean-shift, with different choices of kernel, to divide the image into segments. Each segment is assigned the label of the dominant object present in it. It can be seen that segmentation (b) is the best for tree, road, and car. However, segmentation (d) is better for the left person and the sign board.*

to make a decision of which is most appropriate. It allows for the integration of features derived from different image scales (pixel, segment, and segment union/intersection). We will demonstrate how many of the state-of-the-art methods based on different fixed super-pixels can be seen as special cases of our model.

Inferring the Maximum a Posteriori solution in this framework involves the minimisation of an energy function that contain higher-order potentials defined over several thousand random variables. We show that the solutions of these difficult problems can be efficiently computed using graph-cut based algorithms similar to the pairwise methods of Boykov et al (2001). The contribution of our work not limited to the problem of inference, and its application of the novel associative hierarchical random field framework to object class segmentation. We propose new sophisticated potentials defined over the different levels of the hierarchy. Evaluating the performance of our framework on some of the most challenging data sets for object class segmentation, we show that it outperforms state-of-the-art methods based on a single choice of scale. We believe this is because:

1) Our methods generalises these previous methods allowing them to be represented as particular parameter choices of our hierarchical model.
2) We go beyond these models by being able to use multiple hierarchies of segmentation simultaneously.
3) The optimisation problem can be minimised effectively.

1.4 Hierarchical Models and Image Context

The use of image context has been well documented for object recognition and segmentation. It is particularly useful in overcoming ambiguities caused by limited evidence; this often occurs in object recognition where we frequently encounter objects at small scales or low resolution images (Hoiem et al, 2006). Classical Random Field models exploit context in a local manner by encouraging adjacent pixels or segments to take the same label. To encode context at different scales Zhu et al (2008) introduced the hierarchical image model (HIM) built of rectangular regions with parent-child dependencies. This model captures large-distance dependencies and can be solved efficiently using dynamic programming. However, it supports neither multiple hierarchies, nor dependencies between variables at the same level. To encode semantic context and to combine top-down and bottom-up approaches Tu et al (2003) proposed a framework in which they showed that the use of object specific knowledge helps to disambiguate low-level segmentation cues.

Our hierarchical random field model uses a novel formulation that allows context to be incorporated at multiple levels of multiple quantization, something not previously possible. As we show in section 6 it leads to state of the art segmentation results.

2 Random Field Formulations for Semantic Segmentation

Consider an ordered set of variables $\mathbf{X} = [X_1, X_2, \ldots, X_n]$, where each variable X_i takes a label from a set $\mathcal{L}$ corresponding to object classes. We write $\mathbf{x} \in \mathcal{L}^n$ for a labelling of $\mathbf{X}$, and use x_i to refer to the labelling of the variable X_i. At times we will refer to the labelling of a subset of variables, corresponding to a *clique* c, for this we use the notation $\mathbf{x}_c$. We use $\mathcal{V} = \{1, 2, \ldots, n\}$ to refer to the set of valid vertices (or indexes) of $\mathbf{X}$, and make use of the common mathematical short hand of $\vee$, $\wedge$ to represent *'or'* and *'and'* respectively. We use Δ as an indicator function, *i.e.*

$$\Delta(\cdot) = \begin{cases} 1 & \text{if } \cdot \text{ is true} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We formulate the general problem of inference, as finding $\mathbf{x}$ the minimiser of an arbitrary cost function

$$\arg\min_{\mathbf{x}} E(\mathbf{x}). \tag{2}$$

Finding $\mathbf{x}$ may correspond to finding the maximum a posteriori (MAP) labelling. Labelling problems in vision are typically formulated as a pairwise random field whose energy can be written as a sum of unary and pairwise potentials:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi^p_{ij}(x_i, x_j), \qquad (3)$$

where N_i is the set of neighbours of vertex i.

The pairwise random field formulation suffers from a number of problems stemming from its inability to express high-level dependencies between pixels, for example, the use of only pairwise smoothing constraints can result in a *shrinkage bias* (Kohli et al, 2008). Despite these limitations, it is widely used and highly effective. Shotton et al (2006) applied the pairwise random field to the object class segmentation problem. They defined unary likelihood potentials using the result of a boosted classifier over a region about each pixel, that they called *TextonBoost*, which provided a substantial increase in performance of existing methods at the time.

This random field (RF) model has several distinct interpretations: As a probabilistic model (Besag, 1986; Lafferty et al, 2001), the unary potentials $\psi_i(x_i)$ of the random field can be interpreted as the negative log likelihood of variable X_i taking label x_i, while the pairwise potential encodes a smoothness prior over neighbouring variables. Under this formulation the *maximum a posteriori* (MAP) estimate corresponds to the minimal cost labelling of cost (3). These cost can also be interpreted as defining a *structured discriminative classifier* (Nowozin et al, 2010; Tsochantaridis et al, 2005), *i.e.* a classifier whose costs do not characterise the log marginal distribution of images, but whose minimum cost labelling is likely to correctly label most of the image. These classifiers can be efficiently learnt in a max-margin framework Alahari et al (2010); Nowozin et al (2010); Szummer et al (2008); Taskar et al (2004a).

The primary contribution of this work is in the proposal of a framework that allows a principled contribution of arbitrary cues from many different cues and models. Therefore, we use the more general discriminative interpretation of our work. The learning method discussed in section 7 is discriminative rather than probabilistic.

2.1 The Robust P^N model

The pairwise random field formulation of (Shotton et al, 2006) was extended by (Kohli et al, 2008) with the incorporation of robust higher-order potentials defined over segments. Their formulation was based upon the observation that pixels lying within the same segment are more likely to take the same label. The energy of the higher-order random field proposed by (Kohli et al, 2008) was of the form:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_l} \psi^p_{ij}(x_i, x_j) + \sum_{c \in \mathcal{S}} \psi^h_c(\mathbf{x}_c), \qquad (4)$$

where $\mathcal{S}$ is a set of cliques (or segments), given by one or more super-pixel algorithms as shown in figure 1, and ψ^h_c are higher-order potentials defined over the cliques. The higher order potentials took the form of a Robust P^N model defined as:

$$\psi^h_c(\mathbf{x}_c) = \min_{l \in \mathcal{L}} \left(\gamma^{\max}_c, \gamma^l_c + \sum_{i \in c} w_i k^l_c \Delta(x_i \neq l) \right), \qquad (5)$$

where w_i is the weight of the variable x_i, and the variables γ satisfy

$$\gamma^l_c \leq \gamma^{\max}_c, \forall l \in \mathcal{L}. \qquad (6)$$

The potential has a cost of γ^l_c if all pixels in the segment take the label l. Each pixel not taking the label l is penalised with an additional cost of $w_i k^l_c$, and the maximum cost of the potential is truncated to γ^{max}_c. This framework enabled the integration of multiple segmentations of the image space in a principled manner.

2.2 Hierarchical Random Fields

To formulate Hierarchical Random Fields, we first show how the higher-order P^N potentials of (5) are equivalent to the cost of a minimal labelling of a set of pairwise potentials defined over the same clique variables $\mathbf{x}_c$ and a single auxiliary variable $x^{(1)}_c$ that takes values from an extended label set[2]

$$\mathcal{L}^E = \mathcal{L} \cup \{L_F\} \qquad (7)$$

and generalise it to a Hierarchical model formed of nested P^N like potentials.

In a minimal cost solution, $x^{(1)}_c$ taking a value l will correspond to the clique $\mathbf{x}_c$ having a dominant label l, *i.e.* the majority of pixels within c must take label l. If it takes the free label L_F, it means that there is no dominant label in the clique, and that segment is unassigned. The cost function over $\mathbf{x}_c \cup \{x^{(1)}_c\}$ takes the form:

$$\psi_c(\mathbf{x}_c, x^{(1)}_c) = \phi_c(x^{(1)}_c) + \sum_{i \in c} \phi_c(x_i, x^{(1)}_c). \qquad (8)$$

where the unary potential over $x^{(1)}_c$, $\phi_c(x^{(1)}_c)$ associates the cost γ^l_c with $x^{(1)}_c$ taking a label in $\mathcal{L}$, and $\gamma^{\max}_c$ with $x^{(1)}_c$ taking the *free* label L_F. The pairwise potentials $\phi_c(x^{(1)}_c, x_i)$ are defined as:

$$\phi_c(x^{(1)}_c, x_i) = \begin{cases} 0 & \text{if } x^{(1)}_c = L_F \vee x^{(1)}_c = x_i \\ w_i k^{x^{(1)}_c}_c & \text{otherwise.} \end{cases} \qquad (9)$$

Then:

$$\psi^h_c(\mathbf{x}_c) = \min_{x^{(1)}_c} \psi_c(\mathbf{x}_c, x^{(1)}_c). \qquad (10)$$

[2] The index (1) refers to the fact that the variable $x^{(1)}_c$ lies in the first layer above the pixel variables x_i (see figure 2).

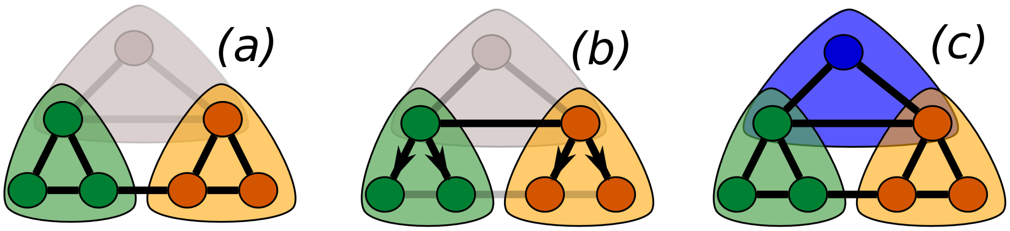


Fig. 2. Existing models as special cases of our hierarchical model. *The lowest layer of the image represents the pixel layer, the middle layer potentials defined over super-pixels or segments, and the third layer represents our hierarchical terms.* (a) *shows the relationships permitted in a pixel-based random field with Robust P^N potentials.* (b) *shows relationships contained within a super-pixel-based random field (the directed edges indicate the one way dependence between the labellings of pixels and super-pixels).* (c) *Our hierarchical random field. See section 3.*

By ensuring that the pairwise edges between the auxiliary variable and individual variables of the clique satisfy the constraint:

$$\sum_{i \in c} w_i k_c^l \geq 2(\phi_c(L_F) - \phi_c(l)), \forall l \in \mathcal{L}, \tag{11}$$

(see section 4.4.1) we can guarantee that the labels of these auxiliary variables carry a clear semantic meaning. If this constraint is satisfied an auxiliary variable may take state $l \in \mathcal{L}$ in a minimal cost labelling, if and only if, the weighted majority of its child variables (*i.e.* the original $\mathbf{x}_c$) take state l. The label L_F indicates a heterogeneous labelling of a segment in which no label holds a significant majority.

We now extend the framework to include pairwise dependencies between auxiliary variables [3]:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi^p_{ij}(x_i, x_j) \tag{12}$$
$$+ \min_{\mathbf{x}^{(1)}} \left(\sum_{c \in \mathcal{S}} \psi_c(\mathbf{x}_c, x_c^{(1)}) + \sum_{c,d \in \mathcal{S}} \psi^p_{cd}(x_c^{(1)}, x_d^{(1)}) \right).$$

These pairwise terms can be understood as encouraging consistency between neighbouring cliques. This framework can be further generalised to a hierarchical model where the connection between layers takes the form of (8) and the weights for each child node in $\phi_c(\cdot)$ are proportional to the sum of the weights in the *"base layer"* belonging to the clique c.

The energy of our new hierarchical model is of the form:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi^p_{ij}(x_i, x_j) \tag{13}$$
$$+ \min_{\mathbf{x}^{(1)}} E^{(1)}(\mathbf{x}, \mathbf{x}^{(1)}),$$

where $E^{(1)}(\mathbf{x}, \mathbf{x}^{(1)})$ is recursively defined as:

$$E^{(\mathrm{n})}(\mathbf{x}^{(n-1)}, \mathbf{x}^{(\mathrm{n})}) = \sum_{c \in \mathcal{S}^{(\mathrm{n})}} \psi^p_c(\mathbf{x}^{(\mathrm{n\text{-}1})}, x_c^{(\mathrm{n})}) \tag{14}$$
$$+ \sum_{c,d \in \mathcal{S}^{(\mathrm{n})}} \psi^p_{cd}(x_c^{(\mathrm{n})}, x_d^{(\mathrm{n})}) + \min_{\mathbf{x}^{(n+1)}} E^{(n+1)}(\mathbf{x}^{(\mathrm{n})}, \mathbf{x}^{(n+1)}).$$

Where $\mathbf{x}^{(0)} = \mathbf{x}$ refers to the state of the base level, and $\mathbf{x}^{(\mathrm{n})}$ for $n \geq 1$ the state of auxiliary variables. The structure of the graph is chosen beforehand and for all layers n beyond the maximal layer in the hierarchy m *i.e.* $n \geq m$,

$$E^{(\mathrm{n})}(\mathbf{x}^{(n-1)}, \mathbf{x}^{(\mathrm{n})}) = 0. \tag{15}$$

The inter-layer potential between between two layers of auxiliary variables takes the form of a weighted Robust P^N potential with the unary term $\phi_c(x_c^{(\mathrm{n})})$ and pairwise term:

$$\phi_c(x_d^{(\mathrm{n\text{-}1})}, x_c^{(\mathrm{n})}) = \begin{cases} 0 & \text{if } x_c^{(\mathrm{n})} = L_F \vee x_c^{(\mathrm{n})} = x_d^{(\mathrm{n\text{-}1})} \\ w_d k_c^{x_c^{(\mathrm{n})}} & \text{otherwise,} \end{cases} \tag{16}$$

where the weights are summed up over the base layer as:

$$w_d = \sum_{j \in d} w_j. \tag{17}$$

Note that (16) encourages $x_c^{(\mathrm{n})}$ to take label L_F if either most of its children take label L_F, or if its children take an inhomogeneous set of labels.

For the remainder of this paper, we assume that the neighbours of auxiliary variables corresponding to super-pixels, will be those variables that correspond to adjacent super-pixels found by the same run of the clustering algorithm. This decision is arbitrary, and the neighbours could equally cross-level of the hierarchy, or connect different clustering algorithms.

3 Relation to Previous Models

In this section, we draw comparisons with the current state-of-the-art models for object segmentation (Galleguillos et al, 2008; Pantofaru et al, 2008; Rabinovich et al, 2007; Yang et al, 2007) and show that at certain choices of the parameters of our model,

[3] When considering multiple cliques, by happenstance, the same clique could occur twice in two different sets of super-pixels. To keep our notation compact, we assume c uniquely specifies both the clique, and the set of super-pixels that gave rise to it. This is purely a notational convenience.

these methods fall out as special cases (illustrated in figure 2). Thus, our method not only generalises the standard pairwise random field formulation over pixels, but also the previous work based on super-pixels and (as we shall see) provides a global optimisation framework allowing us to combine features at different quantization levels.

We will now show that our model is not only a generalisation of random fields over pixels, but also of two classes of preexisting model: *(i)* random fields based upon disjoint segments (Batra et al, 2008; Galleguillos et al, 2008; Yang et al, 2007) (see figure 2(b)), and *(ii)* random fields based upon the intersection of segments (Pantofaru et al, 2008).

3.1 Equivalence to random fields based on Segments

Consider a hierarchy composed of only pixels, and cliques corresponding to one super-pixel based segmentation of the image. In this case, all the segments are disjoint (non-overlapping). In this case, our model becomes equivalent to the pairwise random field models defined over segments (Batra et al, 2008; Galleguillos et al, 2008; Rabinovich et al, 2007; Yang et al, 2007).

To ensure that no segment takes the label L_F, we assign a high value to $\gamma_c^{\max} \to \infty, \forall c \in \mathcal{S}^{(1)}$. In this case, the optimal labelling will always be *segment consistent* (i.e. all pixels within the segment will take the same label) and the potential $\psi_c(\mathbf{x}_c, x_c^{(1)})$ can now be considered as a unary potential over the auxiliary (segment) variable $x_c^{(1)}$. This allows us to rewrite (12) as:

$$E(\mathbf{x}^{(1)}) = \sum_{c \in \mathcal{S}^{(1)}} \psi_c(x_c^{(1)}) + \sum_{c,d \in \mathcal{S}^{(1)}} \psi_{cd}^p(x_c^{(1)}, x_d^{(1)}) \quad (18)$$

which is exactly the same as the cost associated with the pairwise random field defined over segments with $\psi_c(x_c^{(1)} = l) = \gamma_c^l$ as the unary cost and $\psi_{cd}^p(\cdot)$ as the pairwise cost for each segment.

3.2 Equivalence to Models of Segment Intersections

We now consider the case with multiple overlapping segmentations and unary and pairwise potentials defined only upon these segments; this is analogous to the construct of Pantofaru et al (2008). If we set:

$$w_i k_c^l = \gamma_c^{\max}, \forall i \in \mathcal{V}, l \in \mathcal{L}, c \in \mathcal{S}, \quad (19)$$

then:

$$x_c^{(1)} \neq L_F \text{ only if } x_i = x_c^{(1)}, \forall i \in c. \quad (20)$$

In this case, only the potentials $\sum_{c \ni i} \psi_c(\mathbf{x}_c, x_c^{(1)})$ act on x_i.

Consider a pair of pixels i, j that lie in the same intersection of segments *i.e.* :

$$\{c \in S : c \ni i\} = \{c \in S : c \ni j\}. \quad (21)$$

Then, in a minimal labelling, either $\exists x_c^{(1)} = x_i$, and hence $x_j = x_c^{(1)} = x_i$, or $\forall c \ni i : x_c^{(1)} = L_F$. In the second degenerate case there are no constraints acting on x_i or x_j, and a minimal cost labelling can be chosen such that $x_i = x_j$.

Consequently, there is always a minimal cost labelling consistent with respect to the intersection of segments, in this sense our model is equivalent to that proposed in Pantofaru et al (2008).

3.3 Equivalence to tree structured associative models

Tree structured hierarchies such as (Lim et al, 2009; Nowozin et al, 2010; Reynolds and Murphy, 2007; Zhu and Yuille, 2005) have been proposed for semantic segmentation. The structure of these models is clearly a strict subset of ours, as it does not support pairwise connections between variables in the same level, and each variable may only be attached to one variable in the layer above. In the restricted case, in which the label space and edge costs between parent and child are of the same form as those we consider, these models can also be contained in our approach, although this need not hold in general.

3.4 The Relationship with Directed Models

A hierarchical, two-layer, directed model was proposed in (Kumar and Hebert, 2005). This is a hybrid model relatively similar to ours, with unary and pairwise potentials defined over both super-pixels and pixels and pairwise connections between the layers, enforcing consistency. It principally differs from ours in the use of directed edges between layers. These directed edges mean that max-marginals can be computed in a piecewise manner, and propagated from one layer to the other. This makes it more suitable for inference with message passing algorithms than our framework (see section 8.1).

This directed approach does not propagate information through-out the structure. In order to arrive at a consistent hypothesis, that takes account conflicting clues from all levels of the hierarchy, there are two desirable criteria for the propagation of information.

1) We wish for information to be transmitted from the pixel to the segment level and from there back to the pixel level. That is, the labelling of one pixel should affect the label of the segment potential and, from this, the label of other pixels in the same segment.
2) Information should also be transmitted from a segment to the pixel level and back to the segment level. This means that if two segments overlap, the optimal label of one segment should indirectly depend on the labelling of the other.

At most one of the above-mentioned conditions can hold if the connections between layers of the hierarchical model form a directed acyclic graph (DAG). This is the case in the model of Kumar and Hebert (2005).

3.5 Equivalence to the Pylon Model

The recent work by Lempitsky et al (2011) proposed a new approach to segmentation in which, given a tree-structured hierarchy of super-pixels $\mathcal{S}$, they seek a labelling $\mathbf{x}$ where $x_i \in \mathcal{L} \cup \{L_F\}$ that minimise a sum of positive costs over super-pixels and a pairwise regulariser over pairs of pixels:

$$E'(\mathbf{x}) = \sum_{c \in \mathcal{S}} \psi'_c(x_c) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi^p_{ij}(x_i, x_j) \tag{22}$$

here $\mathcal{V}$ denotes the set of pixels in the image. This is a restricted case of our work, in that the pairwise costs ψ' are only defined over pixels, and not super-pixels, while unary potentials are not defined over individual pixels. Unlike our work, they impose the constraint that

$$\psi'_c(L_F) = 0, \ \forall c \in \mathcal{S}, \tag{23}$$

and that for every pixel i in the image $\mathcal{V}$ there exists at least one segment $c \in \mathcal{S}$ such that:

$$x_c \neq L_F \wedge i \in c \tag{24}$$

or equivalently, every pixel belongs to at least one 'active' segment not taking label L_F. In this formulation, assigning the label L_F to a segment c associates no cost with it, and is analogous to discarding it. In a minimum cost labelling, only one super-pixel containing any pixel will take a state other than L_F, and this objective function is similar to those proposed by Gould et al (2009a). Note that every pixel i always has a label $x_i \neq L_F$ associated with it, and this label is shared by any super-pixel that contains it, and does not take label L_F.

We show that the inference of Lempitsky et al (2011), can be performed using the inference techniques we discuss in this paper. We will now show how this is the case by constructing an equivalent cost function of the form described in section 2.2.

We replace the implicit constraint of Lempitsky et al (2011), that if a super-pixel $c \in \mathcal{S}$ takes label $l \in \mathcal{L}$, all its children (pixels or super-pixels it consists of) *i.e.* $j \in \mathcal{S} : j \subset c$ take the label L_F, with the constraint that:

$$x_j = x_c = l, \ \forall j \subset c \tag{25}$$

$$\forall x_c \neq L_F \tag{26}$$

i.e. that a child's labelling must agree with its parent. This can be enforced by setting the weights $w_i k^l_c$ of equation (9) to be sufficiently high. We take the same tree of super-pixels S as our hierarchy, and writing $C(c)$ for the children of c, we set unary potentials to be:

$$\psi_c(x_c) = \psi'_c(x_c) - \sum_{j \in C(c)} \psi'_j(x_c) \ \forall c, x_c \in \mathcal{L}, \tag{27}$$

and

$$\psi_c(L_F) = 0. \tag{28}$$

Taking the pairwise potentials ψ^p_{ij} to be the same as the pylon model, we consider the cost:

$$E(\mathbf{x}) = \sum_{c \in \mathcal{S}} \psi_c(x_c) + \sum_{i \in \mathcal{I}, j \in \mathcal{N}_i} \psi^p_{ij}(x_i, x_j) \tag{29}$$

Now, given any labelling $\mathbf{x}$, we can create a new label $\mathbf{x}'$ by the following transform:

$$x'_{C(i)} = \begin{cases} L_F & \text{if } x_{C(i)} = L_F \wedge x_{C(i)} = x_i \\ x_{C(i)} & \text{otherwise.} \end{cases} \tag{30}$$

Then $E(\mathbf{x}) = E'(\mathbf{x}')^4$, and any minimal $\mathbf{x}$ labelling of $E(\mathbf{x})$, must induce a minimal labelling $\mathbf{x}'$of $E(\mathbf{x}')$, and every pylon model is equivalent to an AHRF.

3.6 Co-occurrence, Harmony and Stack Potentials

Hierarchical models allow the modelling of contextual relationships. The use of context has been well documented by various researchers, for example, the harmony potentials of Boix et al (2011), the contextual potentials of Lim et al (2009) or the stack parameters of Munoz et al (2010). These works proposed the use of potentials that penalise the labelling of a subset of the image if they do not naturally occur together in an image. For example, the combination of *cow, aeroplane* and *sky* would be penalised, while *cow*, *grass*, and *sky* would not. This is in contrast to our potentials which penalises *inhomogeneous* labellings of a region.

Although these works do not formally characterise the potentials they can solve, they all appear to be *monotonic increasing i.e.* their cost never decreases when a new label is added to a subset of an image. As such, these potentials can be understood as *local co-occurrence costs* (or co-occurrence costs defined over a subset of the image) and integrated in our framework using the work (Ladicky et al, 2010).

This use of *local co-occurrence* rather than P^n like potentials, is useful in tree-structured hierarchies that can not make use of pairwise potentials to encourage neighbouring regions to take the same label. In our framework, a region containing a large amount of both *grass* and *cow* would be marked as inhomogeneous and take label L_F, while using *local co-occurrence* the region could be considered as a *cow-field*, and further regions could be merged with it, without modification, only if they take label *cow* or *grass*. This is particularly desirable if no pairwise smoothing exists between adjacent regions in the image. While, in common with Boix et al (2011), we have found *global co-occurrence potentials* useful, we have been unable to find a use for *local co-occurrence*, in our framework, that could not be better handled by pairwise potentials (see Galleguillos et al (2008) for an example of how pairwise potentials can be used to capture local co-occurrence).

[4] Proof follows by induction over the hierarchy from fine to coarse.

3.7 Robustness to Misleading Segmentations

The quantization of image space obtained using unsupervised segmentation algorithms may be misleading since segments may contain multiple object classes. Assigning the same label to all pixels of such segments will result in an incorrect labelling. This problem can be overcome by using the segment quality measures proposed by (Rabinovich et al, 2007; Ren and Malik, 2003) which can be used to distinguish the *good* segments from *misleading* ones. These measures can be seamlessly integrated in our hierarchical framework by modulating the strength of the potentials defined over segments. Formally, this is achieved by weighting the potentials $\psi^h_c(\mathbf{x}_c, x^{(1)}_c)$ according to a quality sensitive measure $Q(c)$ for any segment c, as in Kohli et al (2008).

4 Inference for Hierarchical random fields

The preceding sections introduced the AHRF model and explained how it can be used for semantic segmentation. However, a model is of little use without an efficient and practical method for optimisation. As the associative hierarchical random field can be transformed into a pairwise model over $|\mathcal{L}|+1$ labels, any method designed for solving general pairwise models could be used. In this section we analyse the suitability of various inference methods for AHRF and propose a novel move making algorithm, that outperforms general pairwise methods for this task.

4.1 Inference in Pairwise Random Fields

Although the problem of MAP inference is NP-hard for most associative pairwise functions defined over more than two labels, in real world problems many conventional algorithms provide near optimal solutions over grid connected random fields (Szeliski et al, 2006). However, the dense structure of hierarchical random fields makes traditional message passing algorithms such as loopy belief propagation (Weiss and Freeman, 2001) and tree-reweighted message passing (Kolmogorov, 2006) converge slowly to high cost solutions (Kolmogorov and Rother, 2006). The difficulties faced by these message passing algorithms can be attributed to the presence of frustrated cycles (Sontag et al, 2008; Werner, 2009) that can be eliminated via the use of cycle inequalities, but only by significantly increasing run time.

Graph cut based move making algorithms do not suffer from this problem and have been successfully used for minimising pairwise functions defined over densely connected networks encountered in vision.

Examples of move making algorithms include α-expansion, $\alpha\beta$-swap (Boykov et al, 2001), and range moves (Kumar et al, 2011; Veksler, 2007) for truncated convex potentials. In case of $\alpha\beta$-swap, a sufficient condition for every possible move to be submodular is the semi-metricity of pairwise potentials (Boykov et al, 2001). In the case of α-expansion a sufficient condition is the metricity of pairwise potentials (Boykov et al, 2001). For symmetric potentials this is also the necessary condition; otherwise one can always construct a state from which the expansion move is not submodular.

These moves differ in the size of the space searched for the optimal move. While expansion and swap search a space of size at most 2^n while minimising a function of n variables, range moves explores a much larger space of K^n where $K > 2$ (see Kumar et al (2011); Veksler (2007) for more details). Of these move making approaches, only $\alpha\beta$-swap can be directly applied to associative hierarchical random fields – other methods require that the inter-layer cost either form a metric, or are truncated convex with respect to some ordering of the labels.

Move-making algorithms start from an arbitrary initial solution of the problem and proceed by making a series of changes each of which leads to a solution of the same or lower energy (Boykov et al, 2001). At each step, the algorithms project a set of candidate moves into a Boolean space, along with their energy function. If the resulting projected energy function (also called the *move energy*) is both submodular and pairwise, it can be exactly minimised in polynomial time by solving an equivalent st-mincut problem. These optima can then be mapped back into the original space, returning the optimal move within the move set. The move algorithms run this procedure until convergence, iteratively picking the best candidate as different choices of range are cycled through.

4.1.1 *Minimising Higher-Order Functions*

A number of researchers have worked on the problem of MAP inference in higher-order random fields. Lan et al (2006) proposed approximation methods for BP to make efficient inference possible in higher-order MRFs. This was followed by the recent works of Potetz and Lee (2008); Tarlow et al (2008, 2010) in which they showed how belief propagation can be efficiently performed in random fields containing moderately large cliques. However, as these methods were based on BP, they were quite slow and took minutes or hours to converge.

In the graph-cut based literature, there have been several related works on transforming general higher-order binary functions into pairwise methods Fix et al (2011); Ishikawa (2009, 2011); Ramalingam et al (2011). These methods address more general potentials than we consider here, however, they explicitly target potentials defined over small cliques and do not scale efficiently to the large problems we consider.

To perform inference in the P^n models, Kohli et al (2007, 2008), first showed that certain projection of the higher-order P^n model can be transformed into

$\mathbf{x}$	$\ldots$	β	γ	α	γ	γ	$\ldots$
Proposed Moves	$\ldots$	α	α	α	α	α	$\ldots$
Move choice $\mathbf{t}$	$\ldots$	0	1	1	1	0	$\ldots$
$\mathbf{x}'$	$\ldots$	α	γ	α	γ	α	$\ldots$

Fig. 3. *An illustration of move encoding in α-expansion. Starting from an initial model $\mathbf{x}$ a new move $\mathbf{t}$ is proposed which causes two variables to change their label to α. This results in the new labelling $\mathbf{x}'$.*

submodular pairwise functions containing auxiliary variables. This was used to formulate higher-order expansion and swap move making algorithms.

4.2 Pairwise Inference with α-Expansion

The algorithm α-expansion makes a sequence of progressive decisions to replace the labels of a subset of the random field with label α. At each step in the sequence, for a given choice of α, the optimal subset which results in the lowest possible cost is found via graph-cuts (see figure 3). The steps proposed by α-expansion, can be can be encoded as a *transformation vector* of binary variables $\mathbf{t} = \{t_i, \forall i \in \mathcal{V}\}$. Each component t_i of $\mathbf{t}$ encodes a partial decision, about what the state of the variable x_i should change to. In α-expansion, $t_i = 1$ encodes the decision that x_i should remain constant, while $t_i = 0$ indicates that it should transition to a new label α. See figure 3 for an illustration.

The use of these transformation vectors makes the problem of finding the optimal move equivalent to minimising a pseudo-Boolean cost function from $2^n \to \mathbb{R}$. Consequently, it follows from Kolmogorov and Zabih (2004), that if these cost functions can be shown to be pairwise submodular, the optimal move can be efficiently found using graph-cut.

The inter-layer pairwise potentials associated with our hierarchy (see (8)) are not metric, so it is not clear whether every possible expansion move is submodular. While variants of α-expansion can be applied to arbitrary pairwise potentials Rother et al (2005), the moves are not guaranteed to be optimal, unless the cost of making all such moves is known to be submodular.

Now we show that the form of the inter-layer pairwise potentials allows optimal alpha-expansion move to be computed. A sufficient condition to guarantee this is to find a transformation, that maps all pairwise cost into the metric representation of (Boykov et al, 2001). Note that this transformation is not necessary in the actual implementation of the inference method; it only serves as a proof, that each expansion move energy is submodular.

First, we assume that all variables in the hierarchy take values from the same label set $\mathcal{L}^E$ defined in equation (7). Where this is not true, for example, original variables $\mathbf{x}^{(0)}$ at the base of the hierarchy never take label L_F, we augment the label set with the label L_F and associate a prohibitively large unary cost with it. Secondly, we make the inter-layer pairwise potentials *symmetric*[5] and *metric* by performing a local reparameterisation operation.

Lemma 1: The inter-layer pairwise functions

$$\phi_{ic}^{(n)}(x_i^{(n-1)}, x_c^{(n)}) = \begin{cases} 0 & \text{if } x_c^{(n)} = L_F \\ & \vee\ x_c^{(n)} = x_i^{(n-1)} \\ w_i k_c^l & \text{if } x_c^{(n)} = l \neq L_F \\ & \wedge\ x_i^{(n-1)} \neq x_c^{(n)} \end{cases} \tag{31}$$

of (13) can be written as:

$$\begin{aligned} \phi_{ic}^{(n)}(x_i^{(n-1)}, x_c^{(n)}) &= \psi_i^{(n-1)}(x_i^{(n-1)}) + \psi_c^{(n)}(x_c^{(n)}) \\ &+ \Phi_{ic}^{(n)}(x_i^{(n-1)}, x_c^{(n)}), \end{aligned} \tag{32}$$

where

$$\Phi_{ic}^{(n)}(x_i^{(n-1)}, x_c^{(n)}) =$$

$$\begin{cases} 0 & \text{if } x_i^{(n-1)} = x_c^{(n)} \\ w_i k_c^l/2 & \text{if } x_i^{(n-1)} = L_F \wedge\ x_c^{(n)} = l \\ & \vee\ x_i^{(n-1)} = l \wedge\ x_c^{(n)} = L_F \\ w_i(k_c^{l_1} + k_c^{l_2})/2 & \text{if } x_i^{(n-1)} = l_1 \neq L_F \\ & \wedge\ x_c^{(n)} = l_2 \neq L_F \end{cases} \tag{33}$$

and

$$\psi_c^{(n)}(x_c^{(n)}) = \begin{cases} w_i k_c^l/2 & \text{if } x_c^{(n)} = l \in \mathcal{L} \\ 0 & \text{otherwise,} \end{cases} \tag{34}$$

$$\psi_i^{(n-1)}(x_i^{(n-1)}) = \begin{cases} -w_i k_c^l/2 & \text{if } x_i^{(n-1)} = l \in \mathcal{L} \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Proof: There are five cases for the inconsistency cost $C = \psi_i(x_i^{(n-1)}) + \psi_c(x_c^{(n)}) + \psi_{ic}(x_c^{(n)}, x_i^{(n-1)})$. For each one of them the cost before and after the reparameterisation stays the same.

Labelling	Cost
$x_c^{(n)} = L_F$, $x_i^{(n-1)} = L_F$	$C = 0$
$x_c^{(n)} = L_F$, $x_i^{(n-1)} = l \in \mathcal{L}$	$C = -w_i k_c^l/2 + w_i k_c^l/2 = 0$
$x_c^{(n)} = l \in \mathcal{L}$, $x_i^{(n-1)} = L_F$	$C = w_i k_c^l/2 + w_i k_c^l/2 = w_i k_c^l$
$x_c^{(n)} = x_i^{(n-1)} = l \in \mathcal{L}$	$C = w_i k_c^l/2 - w_i k_c^l/2 = 0$
$x_c^{(n)} = l_1 \in \mathcal{L}$, $x_i^{(n-1)} = l_2 \in \mathcal{L}, l_1 \neq l_2$	$C = w_i k_c^{l_1}/2 - w_i k_c^{l_2}/2 + w_i(k_c^{l_1} + k_c^{l_2})/2 = w_i k_c^{l_1}$.

□

The potential (32) trivially satisfies the metricity condition as $\phi_{ic}^{(n)}(\alpha, \beta) + \phi_{ic}^{(n)}(\beta, \gamma) \geq \phi_{ic}^{(n)}(\alpha, \gamma)$ for all possible cases of $\alpha, \beta, \gamma \in \mathcal{L}^E$. Thus, every possible α-expansion move is submodular.

[5] A pairwise potential $\psi_{ij}^p(x_i, x_j)$ is said to be *symmetric* if $\psi_{ij}^p(x_i, x_j) = \psi_{ij}^p(x_j, x_i)$ for all choices of x_i and x_j.

4.3 Range-move α-Expansion and $\alpha\beta$-Swap

Let us consider a generalisation of the swap and expansion moves proposed in Boykov et al (2001) over our pairwise formulation. In a standard swap move, the set of all moves considered is those in which a subset of the variables currently taking label α or β change labels to either β or α. In our range swap the moves considered allow any variables taking labels α, L_F or β to change their state to any of α, L_F or β. Similarly, while a normal α-expansion move allows any variable to change to some state α, our range expansion allows any variable to change to states α or L_F.

This approach always considers a set of moves that completely contains the moves considered by α-expansion. As such any local optima of range-move expansion is a local optima of α-expansion.

These moves can be seen as a variant on the ordered range moves proposed in Kumar et al (2011). However, Kumar et al (2011) required that there exists an ordering of the labels $\{l_1, l_2, \ldots, l_n\}$ such that the cost function is convex over the range $\{l_i, l_{i+1} \ldots l_{i+j}\}$ for some $j \geq 2$, and our range moves require no such ordering.

4.4 Transformational Optimality

Consider an energy function $E(\mathbf{x})$ defined over the variables $\mathbf{x} = \{\mathbf{x}^{(h)}, h \in \{1, 2, \ldots, H\}\}$ of a hierarchy with H levels. We call a move making algorithm *transformationally optimal* if and only if any move $(\mathbf{x}^*, \mathbf{x}^a)$ proposed by the algorithm satisfies the property:

$$E^a(\mathbf{x}^*, \mathbf{x}^a) = \min_{\mathbf{x}'} E^a(\mathbf{x}^*, \mathbf{x}') \tag{36}$$

i.e. $\mathbf{x}^a$ is a minimiser of $E^a(\mathbf{x}^*, \cdot)$. Inserting this into equation (13) we have:

$$E(\mathbf{x}^*) = E'(\mathbf{x}^*) + E^a(\mathbf{x}^*, \mathbf{x}^a). \tag{37}$$

This implies that the partial move $\mathbf{x}^*$ proposed by a transformationally optimal algorithm over

$$E'(\mathbf{x}^{(1)}) + E^a(\mathbf{x}^{(1)}, \mathbf{x}^a) \tag{38}$$

must function as a move that directly minimise the higher-order cost of equation (13). Experimentally, our transformationally optimal algorithms converge faster, and to better solutions than standard approaches, such as α-expansion. Moreover, unlike standard approaches, our transformationally optimal algorithms are guaranteed to find the exact solution for binary AHRFs.

We now show that when applied to hierarchical random fields, subject to two natural requirements, namely hierarchical consistency, and a restricted choice of pairwise potentials, *range* moves are transformationally optimal.

4.4.1 Hierarchical Consistency

To guarantee transformational optimality, for α-expansion and $\alpha\beta$-swap, we must constrain the set of higher order potentials further. To do this we will introduce the notion of hierarchical consistency, which simply says auxiliary variables should agree with the state of their child variables in a minimal cost labelling. This has two important properties: first it restricts the space of possible labels taken by an auxiliary variable, allowing us to guarantee transformational optimality; more importantly it will enforce agreement between the different layers of the hierarchy and guarantee a consistent labelling across the entire hierarchy.

Consider a clique c with an associated auxiliary variable $x_c^{(i)}$. Let $\mathbf{x}_l$ be a labelling such that $x_c^{(i)} = l \in \mathcal{L}$ and $\mathbf{x}_{L_F}$ be a labelling that differs from it only in that the variable $x_c^{(i)}$ takes label L_F. We say a clique potential is *hierarchically consistent* only if it satisfies the constraint:

$$E(\mathbf{x}_l) \geq E(\mathbf{x}_{L_F}) \implies \frac{\sum_{i\in c} w_i k_c^l \Delta(x_i = l)}{\sum_{i \in c} w_i k_c^l} > 0.5. \tag{39}$$

The property of hierarchical consistency is also required in computer vision for the cost associated with the hierarchy to remain meaningful. The labelling of an auxiliary variable within the hierarchy should be reflected in the state of the clique associated with it. If an energy is not hierarchically consistent, it is possible that the optimal labelling of regions of the hierarchy will not reflect the labelling of the base layer.

To understand why this consistency is important, we consider a case where this is violated. Consider a simple energy function consisting of a base layer of 10 pixels $\mathbf{x}^{(0)}$ and only one clique, with associated auxiliary variable x_c, defined over the base layer. We assume that unary potentials defined over individual pixels have a preference for the class *cow* while the higher-order potential defined over the clique expresses a preference for class *sheep*. More formally we set:

$$\psi_i(x_i) = \begin{cases} 2 & \text{if } x_i = \textit{sheep} \\ 0 & \text{if } x_i = \textit{cow} \end{cases} \quad \forall x_i \in \mathbf{x}^{(0)} \tag{40}$$

$$\phi_c(x_c) = \begin{cases} 0 & \text{if } x_c = \textit{sheep} \\ 20 & \text{if } x_c = \textit{cow} \\ 20 & \text{if } x_c = L_F \end{cases} \tag{41}$$

And we define the pairwise terms between the clique variables as

$$\phi_{c,i}(x_c, x_i) = \begin{cases} 1 & \text{if } x_c \neq L_F \wedge x_c \neq x_i \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

For simplicity, we set all pairwise terms within the base layer to 0, and disregard them. Then a minimal labelling of the solution occurs when, $x_i = \textit{cow}$

$\forall x_i \in \mathbf{x}^{(0)}$ and $x_c =$ *sheep*. This labelling is incoherent, insomuch as we believe at the base scale that a region is *cow*, and at a coarser scale that the same region is *sheep*. Our requirement of *hierarchical consistency* prohibits such solutions by insisting that the minimal cost labelling of auxiliary variables correspond to either the dominant label in base layer, or to the label L_F.

The constraint (39) is enforced by construction, weighting the relative magnitude of $\psi_i(l)$ and $\psi^p_{ij}(b_j, x_c^{(i)})$ to guarantee that $\forall l \in \mathcal{L}$:

$$\psi_i(l) + \sum_{j \in N_i/c} \max_{b_j \in \mathcal{L} \cup \{L_F\}} \psi^p_{ij}(b_j, x_c^{(i)}) < 0.5 \sum_{i \in c} w_i k_c^l. \tag{43}$$

If this holds, in the degenerate case where there are only two levels in the hierarchy, and no pairwise connections between the auxiliary variables, our AHRF is exactly equivalent to the P^n model.

At most one $l \in \mathcal{L}$ at a time can satisfy (39), assuming the hierarchy is consistent. Given a labelling for the base layer of the hierarchy $\mathbf{x}^{(1)}$, an optimal labelling for an auxiliary variable in $\mathbf{x}^{(2)}$ associated with some clique must be one of two labels: L_F and some $l \in \mathcal{L}$. By induction, the choice of labelling of any clique in $\mathbf{x}^{(j)} : j \geq 2$ must also be a decision between at most two labels: L_F and some $l \in \mathcal{L}$.

4.4.2 Transformational Optimality under Swap range moves

Transformational optimality for $\alpha\beta$-swap based range moves for AHRFs further requires that there are no pairwise connections between variables in the same level of the hierarchy, except in the base layer.

From (14), if an auxiliary variable x_c may take label γ or L_F, and one of its children $x_i | i \in c$ take label δ or L_F, the cost associated with assigning label γ or L_F to x_c is independent of the label of x_i with respect to a given move. This means that under a swap move, a clique currently taking label $\delta \notin \{\alpha, \beta\}$ will continue to do so. This follows from (5) as the cost associated with taking label δ is only dependent upon the weighted average of child variables taking state δ, and this remains constant. Hence the only clique variables that may have a new optimal labelling under the swap are those currently taking state α, L_F or β, and these can only transform to one of the states α, L_F or β. As the range moves map exactly this set of transformations, the move proposed must be transformationally optimal, and consequently the best possible $\alpha\beta$-swap over the energy (13).

4.4.3 Transformational Optimality under Expansion Moves

Using range expansion moves, we can maintain transformational optimality while incorporating pairwise connections into the hierarchy — provided condition (39) holds, and the energy can be exactly represented in our submodular moves.

In order for this to be the case, the symmetric pairwise connections must be both multi-label submodular Schlesinger and Flach (2006) over any ordering α, L_F, β and a metric. The only potentials that satisfy these conditions are linear over the orderings $\alpha, L_F, \beta\ \forall \alpha, \beta$, hence must be of the form:

$$\psi^p_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ K_l & \text{if } (x_i = l \wedge x_j = L_F) \\ & \quad \vee (x_i = L_F \wedge x_j = l) \\ K_{l_1} + K_{l_2} & \text{if } x_i = l_1 \wedge x_j = l_2 \end{cases} \tag{44}$$

where $K_l \in \mathbb{R}_0^+$.

A similar argument to that of the optimality of $\alpha\beta$-swap can be made for α-expansion. As the label α is 'pushed' out across the base layer, the optimal labelling of some $x^{(n)}$ where $n \geq 2$ must either remain constant or transition to one of the labels L_F or α. Again, the range moves map exactly this set of transforms and the suggested move is both transformationally optimal, and the best expansion of label α over the higher-order energy of (13).

5 Graph Constructions for Range-move Expansion

In this section we show graph constructs we used in our experiments. Alternatively, we could have used graph construction of (Schlesinger and Flach, 2006) for multi-label submodular functions over the ordering (α, L_F, β) or $(\alpha, L_F, \text{old label})$ respectively. We show the transformation of multi label problem into a binary pairwise one in detail for the sake of completeness.

The move energy will be encoded using one binary variable t_i for each variable x_i in the base layer. This captures the two possible states $\{\alpha, x_i\}$ of the base layer variables after the move. It will use two binary variables $a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}$ for each variable $x_c^{(\mathrm{n})}$ in the auxiliary layer to encoding their three possible states $\{\alpha, L_F, x_c^{(\mathrm{n})}\}$ under a move, where x_i and $x_c^{(\mathrm{n})}$ are the states of the corresponding variables before the move.

The transformation vector (see figure 3) for the base layer variables is encoded the same way as standard α-expansion:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad t_i = 0 \\ x_i & \text{if} \quad t_i = 1. \end{cases} \tag{45}$$

While the transformation vector for the auxiliary variables is encoded as:

$$T_\alpha(\mathbf{x}_c^{(\mathrm{n})}, a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}) = \begin{cases} \alpha & \text{if} \quad a_c^{(\mathrm{n})} = 0 \wedge b_c^{(\mathrm{n})} = 0 \\ x_c^{(\mathrm{n})} & \text{if} \quad a_c^{(\mathrm{n})} = 1 \wedge b_c^{(\mathrm{n})} = 1 \\ L_F & \text{if} \quad a_c^{(\mathrm{n})} = 1 \wedge b_c^{(\mathrm{n})} = 0. \end{cases} \tag{46}$$

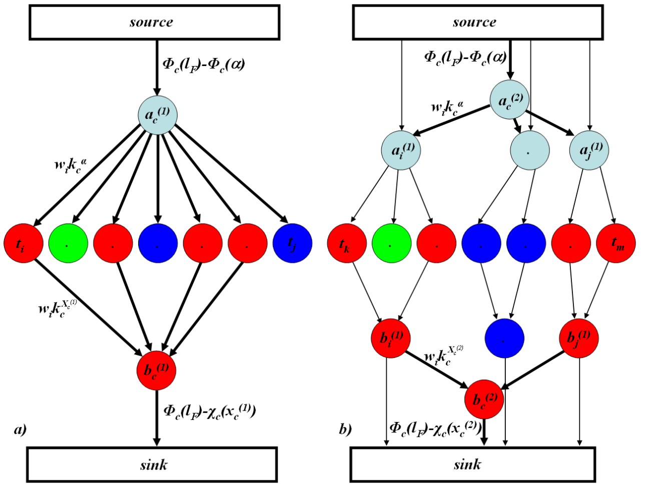


Fig. 4. *A graph construction for the α-expansion move of the inter-layer connection between a) base layer and the first auxiliary layer, b) between two auxiliary levels. The colour of variables t_i and $b_c^{(\mathrm{n})}$ corresponds to the label before the move. Each variable $a_c^{(\mathrm{n})}$ is connected to each of the variables t_i respectively $a_i^{(\mathrm{n-1})}$ in the clique of the previous level, each variable $b_c^{(\mathrm{n})}$ is connected to each of the variables t_i respectively $b_i^{(\mathrm{n-1})}$ in the clique of the previous level. Edges modelling corresponding inter-layer connection are bold.*

To prohibit the combination $a_c^{(\mathrm{n})} = 0$ and $b_c^{(\mathrm{n})} = 1$, we add an edge $K(1-a_c^{(\mathrm{n})})b_c^{(\mathrm{n})}$ with sufficiently large $K \to \infty$. The energy is additive, thus we can find equivalent graph constructions for each term separately.

5.1 Graph Construction for the Inter-layer Potential

Let us first assume none of the variables currently takes a label α or L_F and consider the inter-layer term between the base layer $\mathbf{x}_c$ and the first auxiliary layer:

$$\psi_c^p(\mathbf{x}_c, x_c^{(1)}) = \phi_c(x_c^{(1)}) + \sum_{i \in c} \phi_c(x_c^{(1)}, x_i), \tag{47}$$

where

$$\phi_c(x_c^{(1)}, x_i) = \begin{cases} 0 & \text{if } x_c^{(1)} = L_F \vee x_c^{(1)} = x_i \\ w_i k_c^{x_c^{(1)}} & \text{otherwise.} \end{cases} \tag{48}$$

Writing

$$K_{i,c}^{(\mathrm{n})} = w_i k_c^{x_c^{(\mathrm{n})}} \Delta(x_i = x_c^{(\mathrm{n})}) \tag{49}$$

and

$$\chi_c(x_c^{(\mathrm{n})}) = \phi(x_c^{(\mathrm{n})}) + \sum_{i \in c} w_i k_c^{x_c^{(\mathrm{n})}} \Delta(x_i \neq x_c^{(\mathrm{n})}) \tag{50}$$

The move energy of this potential is:
$\psi_c^p(\mathbf{t}_c, a_c^{(1)}, b_c^{(1)}) =$

$$\begin{cases} \phi_c(\alpha) + \sum_{i \in c} w_i k_c^{\alpha} t_i & \text{if } a_c^{(1)} = 0 \wedge b_c^{(1)} = 0 \\ \chi_c(x_c^{(1)}) + \sum_{i \in c} K_{i,c}^{(1)}(1 - t_i) & \text{if } a_c^{(1)} = 1 \wedge b_c^{(1)} = 1 \\ \phi_c(L_F) & \text{if } a_c^{(1)} = 1 \wedge b_c^{(1)} = 0, \end{cases} \tag{51}$$

The move energy can be transformed into:

$$\begin{aligned} \psi_c^p(\mathbf{t}_c, a_c^{(1)}, b_c^{(1)}) = & \phi_c(\alpha) + \chi_c(x_c^{(1)}) - \phi_c(L_F) \\ & + \sum_{i \in c} w_i k_c^{\alpha} t_i (1 - a_c^{(1)}) \\ & + (\phi_c(L_F) - \phi_c(\alpha))\, a_c^{(1)} \\ & + \sum_{i \in c} w_i k_c^{x_c^{(1)}} \Delta(x_i = x_c^{(1)})(1 - t_i) b_c^{(1)} \\ & + (\phi_c(L_F) - \chi_c(x_c^{(1)}))\, (1 - b_c^{(1)}). \end{aligned} \tag{52}$$

The equivalence can be shown by checking the value of the transformed move energy for each combination of $a_c^{(1)}$ and $b_c^{(1)}$. The move energy is pairwise submodular and thus represents our inter-layer potential. The graph is equivalent to the Robust-P^N graph construction in (Kohli et al, 2008).

For the inter-layer potential between two auxiliary layers $\mathbf{x}^{(\mathrm{n})}$ and $\mathbf{x}^{(\mathrm{n-1})}$ where $n > 1$, the pairwise cost becomes:

$$\phi_c(x_c^{(\mathrm{n})}, x_d^{(n-1)}) = \begin{cases} 0 & \text{if } x_c^{(\mathrm{n})} = L_F \\ & \vee x_c^{(\mathrm{n})} = x_d^{(\mathrm{n-1})} \\ w_d k_c^{x_c^{(\mathrm{n})}} & \text{otherwise.} \end{cases} \tag{53}$$

The condition $x_c^{(\mathrm{n})} = x_d^{(\mathrm{n-1})}$ is satisfied if both auxiliary variables satisfy $a_c^{(\mathrm{n})} = a_d^{(\mathrm{n-1})}$ and $b_c^{(\mathrm{n})} = b_d^{(\mathrm{n-1})}$. A label of a child is not consistent with a label α if $a_i^{(\mathrm{n-1})} = 1$, a label of a child is not consistent with an old label if $b_i^{(\mathrm{n-1})} = 0$. Thus, the move energy of this potential is:
$\psi_c^p(\mathbf{a}^{(\mathrm{n-1})}, \mathbf{b}^{(\mathrm{n-1})}, a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}) =$

$$\begin{cases} \phi_c(\alpha) + \sum_{i \in c} w_i k_c^{\alpha} a_i^{(\mathrm{n-1})} & \text{if } a_c^{(\mathrm{n})} = 0 \wedge b_c^{(\mathrm{n})} = 0 \\ \chi_c(x_c^{(\mathrm{n})}) + \sum_{i \in c} K_{i,c}^{(\mathrm{n})}(1 - b_i^{(\mathrm{n-1})}) & \text{if } a_c^{(\mathrm{n})} = 1 \wedge b_c^{(\mathrm{n})} = 1 \\ \phi_c(L_F) & \text{if } a_c^{(\mathrm{n})} = 1 \wedge b_c^{(\mathrm{n})} = 0, \end{cases}$$

As with the previous case, the move energy can transformed into:
$\psi_c^p(\mathbf{a}^{(\mathrm{n-1})}, \mathbf{b}^{(\mathrm{n-1})}, a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}) =$

$$\begin{aligned} & \phi_c(\alpha) + \chi_c(x_c^{(\mathrm{n})}) - \phi_c(L_F) \\ & + \sum_{i \in c} w_i k_c^{\alpha} a_i^{(\mathrm{n-1})} (1 - a_c^{(\mathrm{n})}) + (\phi_c(L_F) - \phi_c(\alpha)) a_c^{(\mathrm{n})} \\ & + \sum_{i \in c} w_i k_c^{x_c^{(\mathrm{n})}} \Delta(x_i^{(\mathrm{n-1})} = x_c^{(\mathrm{n})})(1 - b_i^{(\mathrm{n-1})}) b_c^{(\mathrm{n})} \\ & + (\phi_c(L_F) - \chi_c(x_c^{(\mathrm{n})}))(1 - b_c^{(\mathrm{n})}). \end{aligned} \tag{54}$$

The graph constructions for both cases of inter-layer connection are given in figure 4.

5.1.1 Graph Construction for the Pairwise Potentials of the Auxiliary Variables

We assume that the pairwise potentials considered are in the same form as eq. (44). There are two case to consider: In the first case, $x_c^{(\mathrm{n})} = x_d^{(\mathrm{n})}$, and the move energy of the pairwise potentials between auxiliary

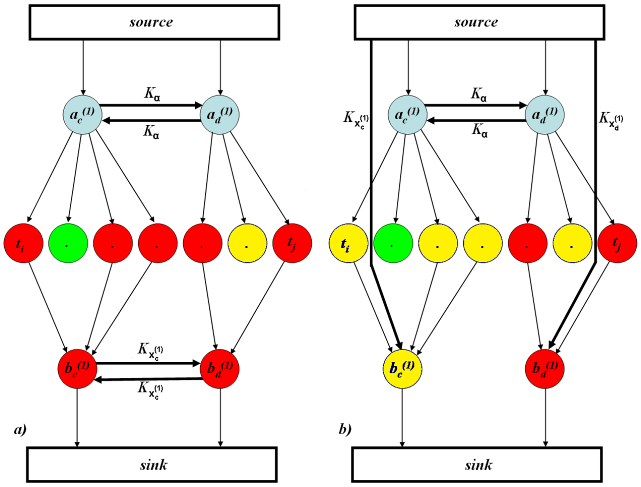


Fig. 5. *A graph construction for the α-expansion move of the pairwise potential on the auxiliary level if the label before the move was a) the same, b) different. The colours of variables t_i and $b_c^{(\mathrm{n})}$ correspond to the label before the move. Edges modelling corresponding pairwise potentials are shown in bold.*

variables is:
$\psi^p_{cd}(a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}, a_d^{(\mathrm{n})}, b_d^{(\mathrm{n})}) =$

$$\begin{cases} 0 & \text{if } a_c^{(\mathrm{n})} = a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = b_d^{(\mathrm{n})} \\ K_\alpha & \text{if } a_c^{(\mathrm{n})} \neq a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = b_d^{(\mathrm{n})} \\ K_{x_c^{(\mathrm{n})}} & \text{if } a_c^{(\mathrm{n})} = a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} \neq b_d^{(\mathrm{n})} \\ K_\alpha + K_{x_c^{(\mathrm{n})}} & \text{if } a_c^{(\mathrm{n})} \neq a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} \neq b_d^{(\mathrm{n})}. \end{cases} \tag{55}$$

This move energy can be transformed into a pairwise submodular one as:

$$\begin{aligned} \psi^p_{cd}(a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}, a_d^{(\mathrm{n})}, b_d^{(\mathrm{n})}) &= K_\alpha a_c^{(\mathrm{n})}(1 - a_d^{(\mathrm{n})}) \\ &+ K_\alpha(1 - a_c^{(\mathrm{n})})a_d^{(\mathrm{n})} + K_{x_c^{(\mathrm{n})}} b_c^{(\mathrm{n})}(1 - b_d^{(\mathrm{n})}) \\ &+ K_{x_c^{(\mathrm{n})}}(1 - b_c^{(\mathrm{n})})b_d^{(\mathrm{n})}. \end{aligned} \tag{56}$$

The equivalence can be seen by checking all possible combinations of $a_c^{(\mathrm{n})}$, $b_c^{(\mathrm{n})}$, $a_d^{(\mathrm{n})}$ and $b_d^{(\mathrm{n})}$.

In the case where $x_c^{(\mathrm{n})} \neq x_d^{(\mathrm{n})}$ the move energy of the pairwise potential between auxiliary variables is:
$\psi^p_{cd}(a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}, a_d^{(\mathrm{n})}, b_d^{(\mathrm{n})}) =$

$$\begin{cases} 0 & \text{if } a_c^{(\mathrm{n})} = a_c^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = b_d^{(\mathrm{n})} = 0 \\ K_{x_c^{(\mathrm{n})}} + K_{x_d^{(\mathrm{n})}} & \text{if } a_c^{(\mathrm{n})} = a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = b_d^{(\mathrm{n})} = 1 \\ K_\alpha & \text{if } a^{(\mathrm{n})} \neq a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = b_d^{(\mathrm{n})} \\ K_{x_c^{(\mathrm{n})}} & \text{if } a_c^{(\mathrm{n})} = a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = 1 \wedge b_d^{(\mathrm{n})} = 0 \\ K_{x_d^{(\mathrm{n})}} & \text{if } a_c^{(\mathrm{n})} = a_d^{(\mathrm{n})} \wedge b_c^{(\mathrm{n})} = 0 \wedge b_d^{(\mathrm{n})} = 1 \\ K_{x_d^{(\mathrm{n})}} + K_\alpha & \text{if } a_c^{(\mathrm{n})} = b_c^{(\mathrm{n})} = 0 \wedge a_d^{(\mathrm{n})} = b_d^{(\mathrm{n})} = 1 \\ K_{x_c^{(\mathrm{n})}} + K_\alpha & \text{if } a_c^{(\mathrm{n})} = b_c^{(\mathrm{n})} = 1 \wedge a_d^{(\mathrm{n})} = b_d^{(\mathrm{n})} = 0, \end{cases} \tag{57}$$

and the equivalent pairwise submodular move energy is:

$$\begin{aligned} \psi^p_{cd}(a_c^{(\mathrm{n})}, b_c^{(\mathrm{n})}, a_d^{(\mathrm{n})}, b_d^{(\mathrm{n})}) &= K_\alpha a_c^{(\mathrm{n})}(1 - a_d^{(\mathrm{n})}) \\ &+ K_\alpha(1 - a_c^{(\mathrm{n})})a_d^{(\mathrm{n})} + K_{x_c^{(\mathrm{n})}} b_c^{(\mathrm{n})} + K_{x_d^{(\mathrm{n})}} b_d^{(\mathrm{n})}. \end{aligned} \tag{58}$$

This equivalence holds for the 3×3 allowed configurations of $a_c^{(\mathrm{n})}$, $b_c^{(\mathrm{n})}$, $a_d^{(\mathrm{n})}$ and $b_d^{(\mathrm{n})}$. Graph constructions for both cases $x_c^{(\mathrm{n})} = x_d^{(\mathrm{n})}$ and $x_c^{(\mathrm{n})} \neq x_d^{(\mathrm{n})}$ are given in figure 5.

All the previous constructions were made under the assumption that none of the variables already takes the label α or L_F. If a variable in the base layer already takes the label α, the problem is equivalent to changing each t_i to 0 in all pairwise submodular expressions. If the variable in the auxiliary layer already takes the the label α, both $a_c^{(\mathrm{n})}$ and $b_c^{(\mathrm{n})}$ have to be changed to 0 in all derived expressions. In the case that the auxiliary variable takes the label L_F, the variable can take only label α and label L_F after the move and thus $b_c^{(\mathrm{n})}$ has to be changed to 0. Setting the label of any variable to 0 is equivalent to tying it to the sink or equivalently changing each incoming edge to this variable to the edge going to the sink. Setting the label of any variable to 1 is equivalent to tying it to the source or equivalently changing each outgoing edge of this variable to the edge going to the source. The infinite edge between $a_c^{(\mathrm{n})}$ and $b_c^{(\mathrm{n})}$ is not necessary if the hierarchy is *hierarchically consistent*, see 4.4.1.

5.2 $\alpha\beta$-swap

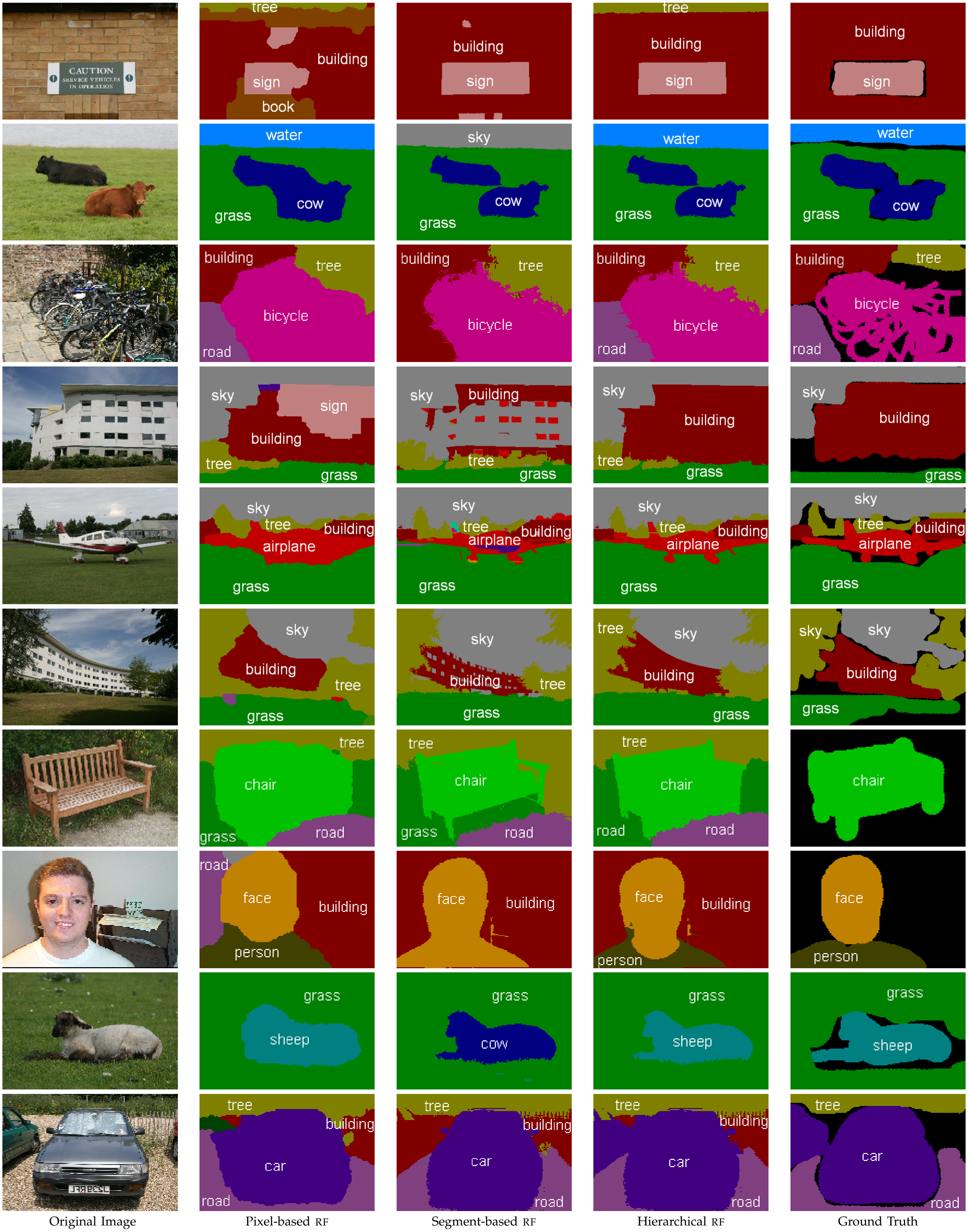
For $\alpha\beta$-swaps the graph-construction can be built by applying these steps: relabelling all variables taking label α to β, and applying α-expansion to only a subset of variables, which took the labels α or β before the move.

6 Hierarchical Random Fields for Semantic Segmentation

Having described the definition and intuition behind the AHRF framework, in this section we describe the set of potentials we use in the object-class segmentation problem. This set includes unary potentials for both pixels and segments, pairwise potentials between pixels and between segments and connective potentials between pixels and their containing segments.

In the previous sections we decomposed the energy (14) into a set of potentials $\psi_c(\mathbf{x}_c)$. In this section we will decompose them further, writing $\psi_c(\mathbf{x}_c) = \lambda_c \xi_c(\mathbf{x}_c)$, where ξ_c is a feature based potential over c and λ_c its weight. Initially we will discuss the learning of potentials $\xi_c(\mathbf{x}_c)$, and later discuss the learning of the weights λ_c.

For our application we used potentials defined over a three-level hierarchy. Although, the hierarchy can be extended indefinitely, we found that performance saturated beyond three-levels on the datasets we considered. We refer to elements of each layer as pixels, segments and super-segments respectively. Unsupervised segments are initially found using multiple applications of a fine scale mean-shift algorithm (Comaniciu and Meer, 2002). "Super-segments" are based upon a coarse mean-shift segmentation, performed over the result of the previous segmentations.



Original Image Pixel-based RF Segment-based RF Hierarchical RF Ground Truth

Fig. 6. *Qualitative results on the MSRC-21 data set comparing non-hierarchical (i.e. pairwise models) approaches defined over pixels (similar to TextonBoost (Shotton et al, 2006)) or segments (similar to (Pantofaru et al, 2008; Russell et al, 2006; Yang et al, 2007) described in section 3) against our hierarchical model. Regions marked black in the hand-labelled ground truth image are unlabelled.*

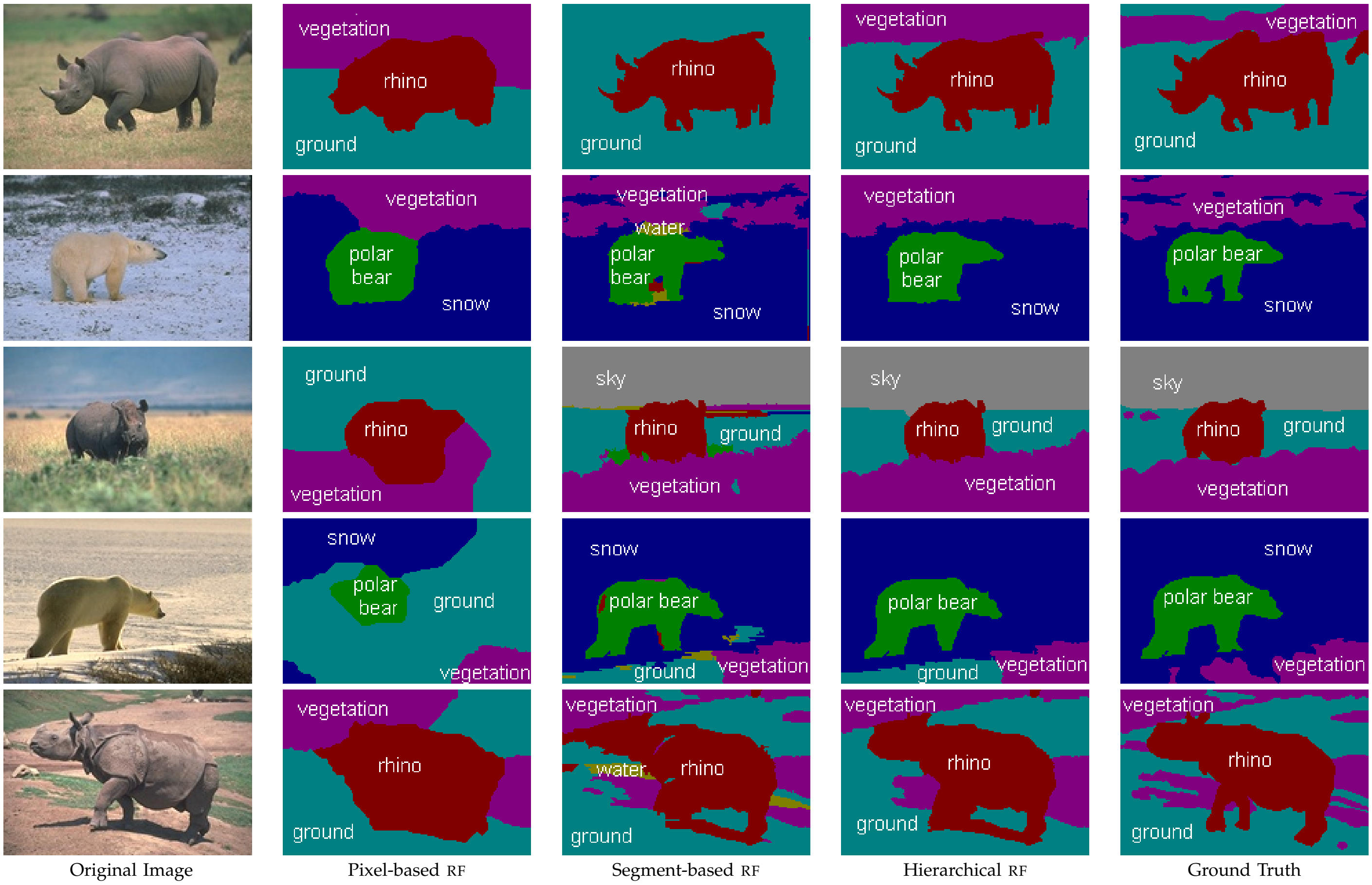


Fig. 7. *Qualitative results on the Corel data set comparing approaches defined over pixels or segments against the hierarchical model.*

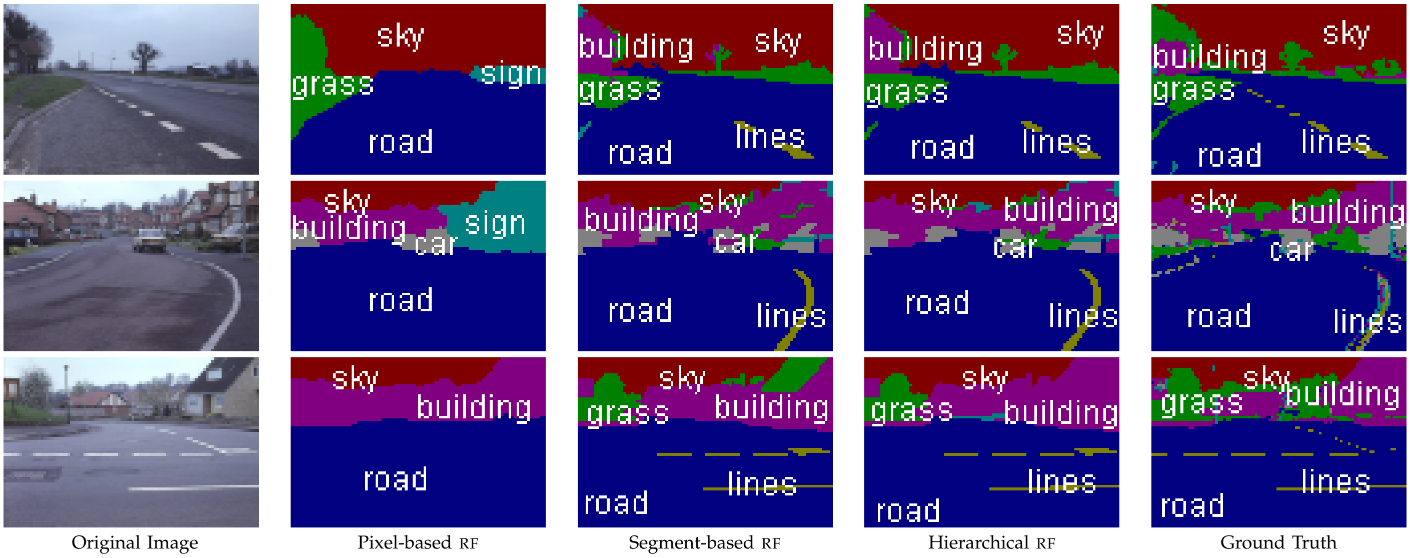


Fig. 8. *Qualitative results on the Sowerby data set comparing approaches defined over pixels or segments against the hierarchical model.*

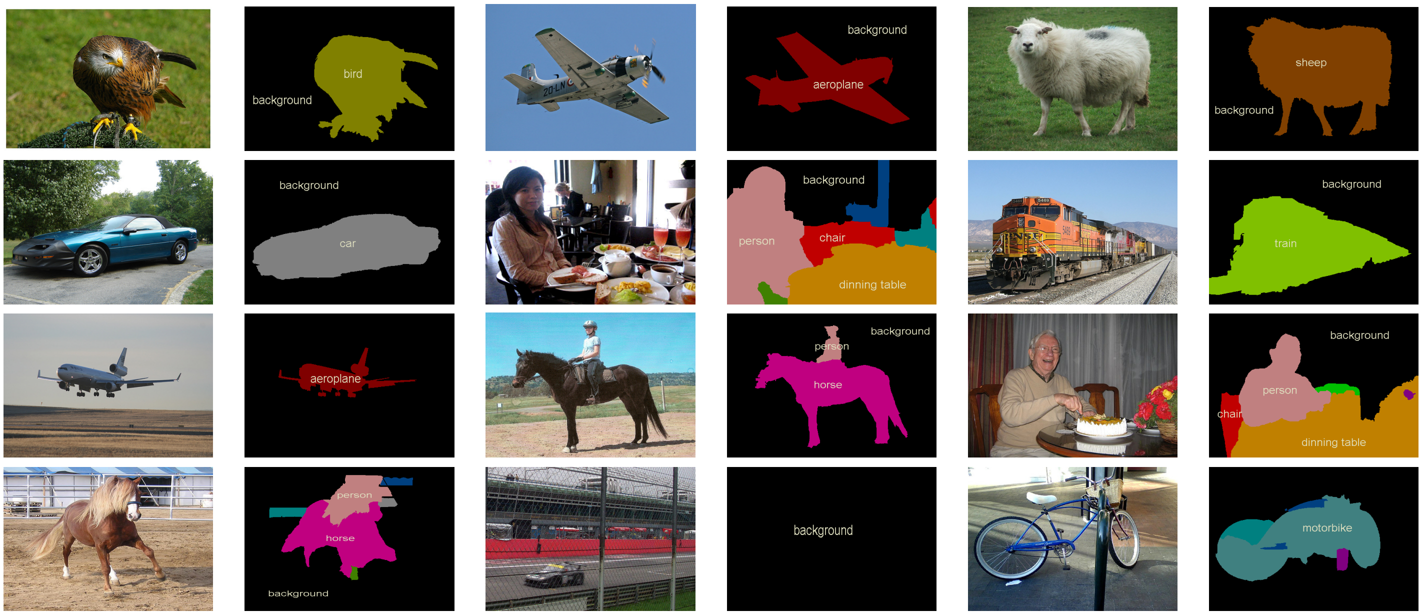


Fig. 9. *Qualitative results on the VOC-2008 data set.* **Successful segmentations** *(top 3 rows) and* **standard failure cases** *(bottom) - from left to right, context error, detection failure and misclassification.*

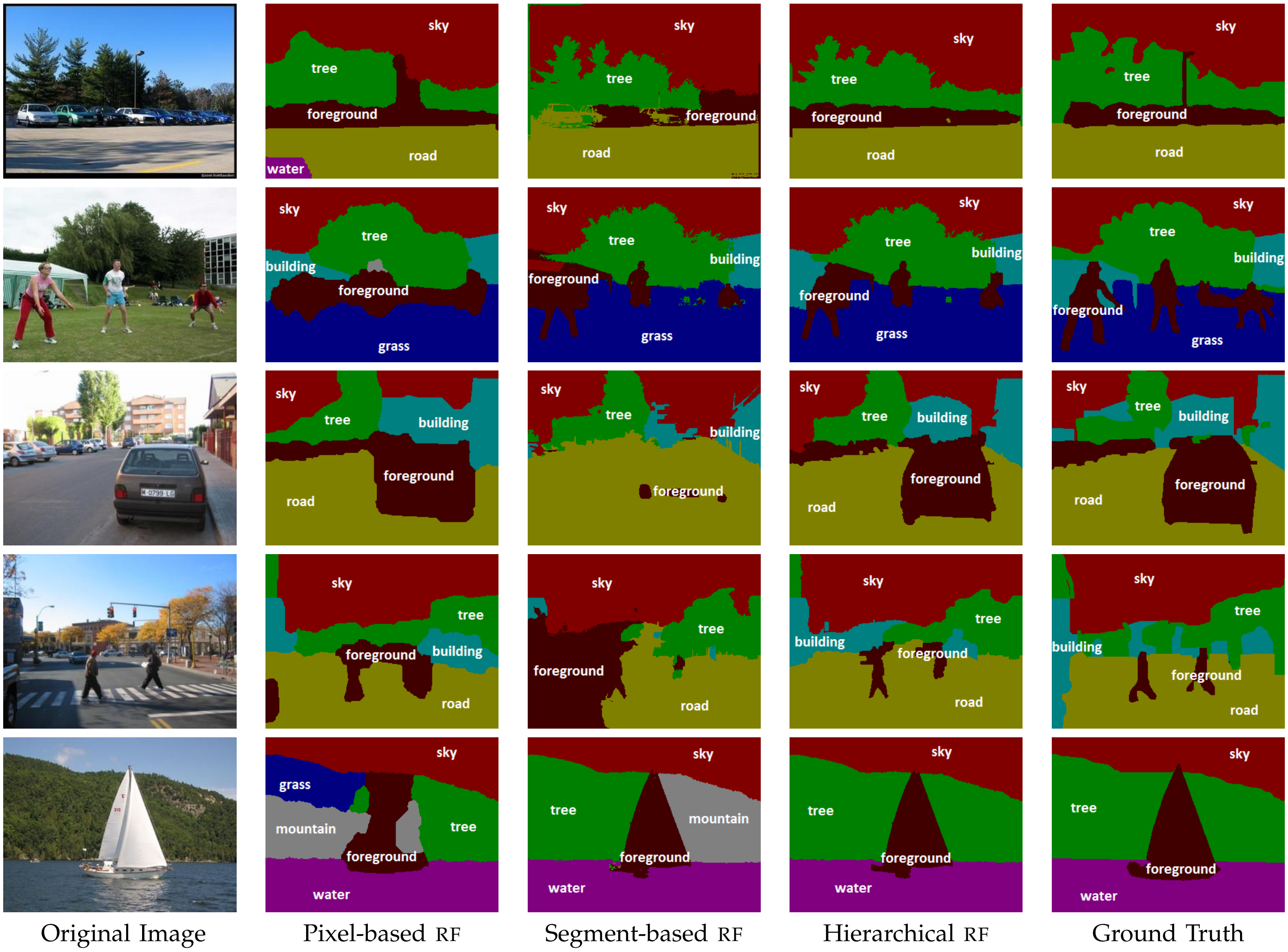


Fig. 10. *Qualitative results on the Stanford data set.*

	Global	Average	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
(Shotton et al, 2008)	72	67	49	88	79	**97**	**97**	78	82	54	87	74	72	74	36	24	93	51	78	**75**	35	66	18
(Shotton et al, 2006)	72	58	62	**98**	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	07
(Batra et al, 2008)	70	55	68	94	84	37	55	68	52	71	47	52	85	69	54	05	85	21	66	16	49	44	32
(Yang et al, 2007)	75	62	63	**98**	**89**	66	54	86	63	71	83	71	79	71	38	23	88	23	88	33	34	43	**32**
Pixel-based RF	84	76	73	93	84	77	84	96	**85**	91	90	**86**	91	95	**91**	**41**	92	**53**	87	65	**77**	**70**	17
Segment-based RF	81	66	80	**98**	83	64	81	**99**	59	89	85	68	68	**98**	76	26	85	39	84	30	49	50	07
Hierarchical RF	**87**	**78**	**81**	96	**89**	74	84	**99**	84	**92**	**90**	**86**	**92**	**98**	**91**	35	**95**	**53**	**90**	62	**77**	**70**	12

TABLE 1

Quantitative results on the MSRC data set. The table shows % pixel recall measure $N_{ii}/\sum_j N_{ij}$ for different object classes. 'Global' refers to the overall error $\frac{\sum_{i\in\mathcal{L}} N_{ii}}{\sum_{i,j\in\mathcal{L}} N_{ij}}$, while 'average' is $\sum_{i\in\mathcal{L}} \frac{N_{ii}}{|\mathcal{L}|\sum_{j\in\mathcal{L}} N_{ij}}$. N_{ij} refers to the number of pixels of label i labelled j. The comparison suggests that the incorporation of the classifiers at different scales leads to a significant improvement of the performance.

	Global	Average	Rhino/Hippo	Polar Bear	Water	Snow	Grass	Ground	Sky
(Batra et al, 2008)	83	**85**	87	**92**	82	**91**	66	**83**	**94**
Pixel-based RF	76	72	80	85	88	83	75	57	35
Segment-based RF	80	78	**92**	65	91	84	81	67	73
Hierarchical RF	**84**	**85**	**92**	82	**94**	88	**83**	77	76

TABLE 2

Quantitative results on the Corel data set. Segment-based method tend to outperform pixel-based ones. Due to the insufficient amount of data the performance largely depends on the random split of the data. The same error measure as for the MSRC data set has been used. Combining classifiers at different scales led to an improvement of the performance.

	Global	Average	Sky	Grass	Road Line	Road	Building	Sign	Car
Pixel-based RF	83	47	92	83	00	89	28	07	33
Segment-based RF	89	**60**	**94**	87	**47**	94	**61**	**10**	35
Hierarchical RF	**91**	**64**	**97**	**96**	45	**98**	59	09	**43**

TABLE 3

Quantitative results on the Sowerby data set. The segment-based method tend to outperform pixel-based ones. Context-based pixel method could not capture small objects due to the insufficient size of the images. The same error measure as for the MSRC data set has been used. As with other data sets, the hierarchical random field outperformed both approaches over single scale.

	Average	Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table	Dog	Horse	Motor bike	Person	Potted plant	Sheep	Sofa	Train	TV/monitor
XRCE	**25.4**	75.9	25.8	15.7	19.2	**21.6**	17.2	**27.3**	**25.5**	**24.2**	7.9	**25.4**	9.9	**17.8**	**23.3**	**34.0**	28.8	**23.2**	**32.1**	**14.9**	**25.9**	**37.3**
UIUC / CMU	19.5	**79.3**	31.9	**21.0**	8.3	6.5	**34.3**	15.8	22.7	10.4	1.2	6.8	8.0	10.2	22.7	24.9	27.7	15.9	4.3	5.5	19.0	32.1
MPI	12.9	75.4	19.1	7.7	6.1	9.4	3.8	11.0	12.1	5.6	0.7	3.7	15.9	3.6	12.2	16.1	15.9	0.6	19.7	5.9	14.7	12.5
Hierarchical RF	20.1	75.0	**36.9**	4.8	**22.2**	11.2	13.7	13.8	20.4	10.0	**8.7**	3.6	**28.3**	6.6	17.1	22.6	**30.6**	13.5	26.8	12.1	20.1	24.8

TABLE 4

Quantitative analysis of VOC2008 results (Everingham et al, 2008) based upon performance the intersection vs. union criteria $\left(\frac{\sum_{i\in\mathcal{L}} N_{ii}}{|\mathcal{L}|(-N_{ii}+\sum_{j\in\mathcal{L}} N_{ij}+N_{ji})}\right)$. All other methods used classification and detection priors trained over a much larger data set that included unsegmented images. The reported results are from the actual challenge.

	Global	Average	Sky	Tree	Road	Grass	Water	Building	Mountain	Foreground
Pixel-based RF	77.9	67.8	90.9	71.3	85.9	82.7	70.0	76.9	8.9	**63.7**
Segment-based RF	77.3	68.1	94.3	65.1	89.9	**88.4**	70.6	77.9	**17.3**	49.0
Hierarchical RF	**80.9**	**70.4**	**94.8**	**71.6**	**90.6**	88.0	**73.5**	**82.2**	10.2	59.9

TABLE 5

Quantitative results on the Stanford data set. The table shows % pixel recall measure $N_{ii}/\sum_j N_{ij}$ for different object classes. 'Global' refers to the overall score $\frac{\sum_{i\in\mathcal{L}} N_{ii}}{\sum_{i,j\in\mathcal{L}} N_{ij}}$, while 'average' is $\sum_{i\in\mathcal{L}} \frac{N_{ii}}{|\mathcal{L}|\sum_{j\in\mathcal{L}} N_{ij}}$. N_{ij} refers to the number of pixels of label i labelled j. The comparison suggests that the incorporation of the classifiers at different scales leads to a significant improvement of the performance. Other papers typically report only the global overall score. For comparison, the pixel CRF approach of (Gould et al, 2009a) gets 74.3%*, their region-based method (Gould et al, 2009a)* 76.4%*, the follow-up paper of (Kumar and Koller, 2010)* 79.4%*, the segment-based method of (Tighe and Lazebnik, 2010)* 77.5% *and the image parsing using recursive neural networks of (Socher et al, 2011)* 78.1%.

	Global	Average	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
Best Mean-Shift seg.	71	57	60	91	76	46	59	88	68	68	64	53	71	67	48	15	83	24	76	39	51	40	**17**
3 Mean-Shift seg.	76	64	59	**98**	81	55	**74**	98	71	84	70	58	70	69	45	**26**	86	29	89	55	62	49	**17**
Best SLIC seg.	70	56	52	95	79	48	57	92	59	74	54	51	60	62	47	21	80	23	78	45	40	35	14
3 SLIC seg.	77	63	58	96	**84**	57	68	95	73	80	69	67	76	67	56	22	87	27	86	58	59	42	09
Best Graph-Cut seg.	69	56	48	94	74	49	60	93	61	76	61	51	58	67	47	23	78	28	76	50	45	35	13
3 Graph-Cut seg.	75	63	51	94	79	55	67	97	74	83	70	60	67	68	58	20	85	**30**	84	61	58	45	13
All 9 segmentations	**83**	**69**	**65**	97	83	**58**	73	**99**	**84**	**87**	**75**	**69**	**77**	**75**	**73**	**26**	**90**	26	**90**	**66**	**66**	**50**	14

TABLE 6

The comparison of performances of three different sources of super-pixels on the MSRC data set, Mean-Shift (Comaniciu and Meer, 2002), SLIC (Achanta et al, 2012) and Graph-cut segmentations (Zhang et al, 2011) and their combination. The results suggest that the combination of multiple segmentations of different kind leads to a significant improvement of the performance. The results were obtained using the same random split of training and test images. These results show the use of 1, 3, or 9 sets of different super-pixel runs in the second layer, and not a three layer hierarchy as in table 1.

6.1 Features

Several well-engineered features were experimentally found to be more discriminative then the raw RGB values of pixels. In our application we use textons (Malik et al, 2001), local binary patterns (Ojala et al, 1994), multi-scale (Bosch et al, 2007) dense SIFT (Lowe, 2004) and opponent SIFT (van de Sande et al, 2008). Textons (Malik et al, 2001) are defined as a clustered 16-dimensional response to 16 different filters - Gaussian, Gaussian derivative and Laplacian filters at different scales. Local binary pattern (Ojala et al, 1994) is a 8-dimensional binary feature consisting of 8 comparisons of the intensity value of the centre pixel with its neighbours. The SIFT (Lowe, 2004) feature contains the histograms of gradients of 4×4 cells quantized into 8 bins. The resulting 128 dimensional vector is normalised to 1. Opponent SIFT (van de Sande et al, 2008) is a variant of coloured SIFT and is built of separate histograms of gradients for 3 channels in the transformed colour space. All features except local binary patterns are quantized to 150 clusters using standard K-means clustering.

6.2 Unary Potentials from Pixel-wise Features

Unary potentials from pixel-wise features are derived from *TextonBoost* (Shotton et al, 2006), and allow us to perform texture based segmentation, at the pixel level, within the same framework. The features used for constructing these potentials are computed on every pixel of the image, and are also called *dense* features. TextonBoost estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. The shape filters are defined by a texton t and rectangular region r. Their response $v_{[t,r]}(i)$ for a given point i is the number of textons t in the region r placed relative to the point i. Corresponding weak classifiers are decision stumps, which split on a shape filter response and one of a set of thresholds. The most discriminative weak classifiers are found using multi-class Gentle Ada-Boost (Torralba et al, 2004).

We observed that textons were unable to discriminate between some classes of similar textures. This motivated us to extend the *TextonBoost* framework by boosting classifiers defined on multiple dense features (such as colour, textons, histograms of oriented gradients (HOG) (Dalal and Triggs, 2005), and pixel location) together. Generalised shape filters are defined by feature type f, feature cluster t and rectangular region r. Their response $v^{f}_{[t,r]}(i)$ for given point i is the number of features of type f belonging to cluster t in the region r placed relative to the point i. The pool of weak classifiers contains decision stumps based on the generalised shape filters against a set of thresholds θ. See (Shotton et al, 2006; Torralba et al, 2004) for further details of the procedure. Our results show that the boosting of multiple features together results in a significant improvement of the performance (note the improvement from the 72% of (Shotton et al, 2006) to 81% of our similar pixel-based random field in figure 1). Further improvements were achieved using exponentially instead of linearly growing thresholds and Gaussian instead of uniform distribution of rectangles around the point. The potential is incorporated into the framework in the standard way as a negative log-likelihood:

$$\phi_i(x_i = l) = -\log \frac{e^{H_l(i)}}{\sum_{l' \in \mathcal{L}} e^{H_{l'}(i)}} = -H_l(i) + K_i, \quad (59)$$

where $H_l(i)$ is the Ada-Boost classifier response for a label l and a pixel i and $K_i = \log \sum_{l' \in \mathcal{L}} e^{H_{l'}(i)}$ a normalising constant.

6.3 Histogram-based Segment Unary Potentials

We now explain the unary potential defined over segments and super-segments. For many classification and recognition problems, the distributions of pixel-wise feature responses are more discriminative than any feature alone. For instance, the sky can be either 'black' (night) or 'blue' (day), but is never 'half-black' and 'half-blue'. This consistency in the colour of object instances can be used as a region based feature for improving object segmentation results. The unary potentials of auxiliary segment variables are learnt using multi-class Gentle Ada-Boost (Torralba et al, 2004) over the normalised histograms of multiple clustered pixel-wise features. The pool of week classifiers are decision stumps that return 1 if more that θ % of a segment belongs to cluster t of feature f, and 0 otherwise. The selection and learning procedure is identical to (Torralba et al, 2004).

The segment potential is incorporated into the energy as:

$$\begin{aligned} \phi_c(x^{(1)} = l) = &\quad \lambda_s |c| \min(-H_l(c) + K_c, \alpha^h), && (60) \\ \phi_c(x^{(1)} = L_F) = &\quad \lambda_s |c| \alpha^h, && (61) \end{aligned}$$

where $H_l(c)$ is the response given by the Ada-boost classifier to clique c taking label l, α^h a truncation threshold and $K = \log \sum_{l' \in \mathcal{L}} e^{H_{l'}(c)}$ a normalising constant.

For our experiments, the cost of pixel labels differing from an associated segment label was set to $k^l_c = (\phi_c(x^{(1)} = L_F) - \phi_c(x^{(1)} = l))/0.1|c|$. This means that up to 10% of the pixels can take a label different to the segment label without the segment variable changing its state to L_F.

6.4 Pairwise Potentials

The pairwise terms on the pixel level $\psi^p_{ij}(\cdot)$ take the form of the classical contrast sensitive potentials.

$$\xi^p(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i,j) & \text{otherwise,} \end{cases} \quad (62)$$

where the function $g(i,j)$ is an edge feature based on the difference in the intensity of colours of neighbouring pixels (Boykov and Jolly, 2001). It is typically defined as:

$$g(i,j) = \theta_p + \theta_v \exp(-\theta_\beta ||I_i - I_j||^2), \quad (63)$$

where I_i and I_j are the colour vectors of pixel i and j respectively. These encourage neighbouring pixels in the image (having a similar colour) to take the same label. We refer the reader to (Boykov and Jolly, 2001; Rother et al, 2004; Shotton et al, 2006) for details.

To encourage neighbouring segments with similar texture to take the same label, we used pairwise potentials based on the squared Euclidean distance of normalised histograms of colour between corresponding auxiliary variables:

$$\xi^p_{cd}(x^{(1)}_c, x^{(1)}_d) = \begin{cases} 0 & \text{if } x^{(1)}_c = x^{(1)}_d, \\ g(c,d)/2 & \text{if } (x^{(1)}_c = L_F \wedge x^{(1)}_d \neq L_F) \\ & \quad \vee (x^{(1)}_c \neq L_F \wedge x^{(1)}_d = L_F), \\ g(c,d) & \text{otherwise,} \end{cases} \quad (64)$$

where $g(c,d) = \min(|c|,|d|) |\mathbf{h}(\mathbf{x}^{(1)}_c) - \mathbf{h}(\mathbf{x}^{(1)}_d)|^2_2$ and $\mathbf{h}(\cdot)$ is the normalised histogram of colours of given segment.

7 Learning Weights for Hierarchical random fields

Having learnt potentials $\xi_c(\mathbf{x}_c)$ as described earlier, the problem remains of how to assign appropriate weights λ_c. This weighting, and the training of random field parameters in general is not an easy problem and there is a wide body of literature dealing with it (Blake et al, 2004; He et al, 2006; Hinton and Osindero, 2006; Taskar et al, 2004b). The approach we take to learn these weights uses a coarse to fine, layer-based, local search scheme over a validation set.

We first introduce additional notation: $\mathcal{V}^{(i)}$ refers to variables contained in the i^{th} layer of the hierarchy, while $\mathbf{x}^{(i)}$ is the labelling of $\mathcal{V}^{(i)}$ associated with a MAP estimate over the truncated hierarchical random field consisting of the random variables $\mathbf{v}' = \{v \in \mathcal{V}^{(k)} : k \geq i\}$. Given the validation data we can determine a dominant label L_c for each segment c, such that $L_F = l$ when $\sum_{i \in l} \Delta(x_i = l) = 0.5|c|$, and if there is no such dominant label, we set $L_c = L_F$.

Algorithm 1 *Weight Learning Scheme.*

```
for $i$ from $m$, the maximal layer, down to $0$ do
    $s_1, s_2, s_h, d_1, d_2, d_h = 1$
    while $s_1, s_2$ or $s_h \geq \Theta$ do
        for $t \in \{1, 2, h\}$ do
            $\lambda_t'^{(i)} \leftarrow \lambda_t^{(i)} + d_t s_t$
            Perform MAP estimate of $\mathbf{x}_i$ using $\lambda_t'$ instead of $\lambda_t$
            if $C(\mathbf{x}_i)$ has decreased then
                $\lambda_t \leftarrow \lambda_t'$
            else
                $s_t \leftarrow s_t/2$, $d_t \leftarrow -d_t$
            end if
        end for
    end while
end for
```

At every level of the hierarchy, the label of a clique $x_c^{(i)}$ must correspond to the dominant label of this clique in the ground truth (or L_F) for its pixels to be correctly labelled. Based on this observation, we propose a simple heuristic which we optimise for each layer.

At each layer, we seek to minimise the discrepancy between the dominant ground truth label of a clique l_c, and the value $x_c^{(i)}$ of the MAP estimate. Formally, we choose parameters λ to minimise

$$C(\mathbf{x}^{(i)}) = \sum_{c \in \mathcal{V}^{(i)}} \Delta(x_c^{(i)} \neq l_c \wedge l_c \neq L_F). \qquad (65)$$

We optimise (65) layer by layer. The full method is given in algorithm 1, where we use $\lambda_1^{(i)}$ to refer to the weighting of unary potentials in the i^{th} layer, $\lambda_2^{(i)}$ the weight of the pairwise terms and $\lambda_h^{(i+1)}$ a scalar modifier of all terms in the $(i+1)^{\text{th}}$ layer or greater. Θ is an arbitrary constant that controls the precision of the final assignment of λ.

An alternative and elegant approach to this is that of (Finley and Joachims, 2008) which we intend to investigate in future work.

8 Experiments

We evaluated the performance of our framework on four data sets: Corel, Sowerby, Stanford (Gould et al, 2009a), PASCAL VOC 2008 (Everingham et al, 2008) and MSRC-21 (Shotton et al, 2006)

MSRC-21 The MSRC segmentation data set contains 591 images of resolution 320×213 pixels, accompanied with a hand labelled object segmentation of 21 object classes. Pixels on the boundaries of objects are not labelled in these segmentations. The division into training, validation and test sets occupied 45%, 10% and 45% of the images. Methods are typically compared using global criteria or average-per-class recall criteria (see table 1 for details). For these experiments, the hierarchy was composed of 3 pairs of nested segmentations. The parameters of the mean-shift kernels were chosen as $(6, 5), (12, 10)$; $(6, 7.5), (12, 15)$; and $(6, 9), (12, 18)$. The first value refers to the planar distance between points, and the second refers to the Euclidean distance in the LUV colour space. A quantitative comparison of performance with other methods is given in table 1. Qualitative results are given in figure 6. A quantitative comparisons of different super-pixel methods is given in table 6.

Corel The Corel segmentation data set contains 100 images of resolution 180×120 pixels of natural scenarios, with a hand labelled object segmentation of 7 object classes. The division into training and test sets occupied 50% and 50% the images. The same parameters as for MSRC data set have been used due to an insufficient amount of data. Unlike in MSRC data set, segment-based methods performed better than pixel-based (see table 2 for more details). Qualitative results are given in figure 7.

Sowerby The Sowerby segmentation data set contains 106 images of resolution 96×64 pixels of road scenes, with a hand labelled object segmentation of 7 object classes. The division into training and test sets occupied 50% and 50% the images. As with the Corel data set, there was insufficient training data to tune parameters on a validation set, and the parameters tuned for the MSRC data set were used. Segment-based methods perform better than pixel-based (see table 3 for more details). Small classes performed very badly due to their highly variable appearance and insufficient training and test data. Qualitative results are given in figure 8.

Stanford The Stanford segmentation data set contains 715 images of resolution 320×240 pixels of road scenes, with a hand labelled object segmentation of 7 background classes one one foreground class. The set of images are split into two equally sized training and testing sets. The same parameters as for the MSRC data set were used. Quantitative comparison of performance is given in table 5. Qualitative results are given in figure 10.

PASCAL VOC 2008 This data set was used for the PASCAL Visual Object Category segmentation contest 2008. It is especially challenging given the presence of significant background clutter, illumination effects and occlusions. It contains 511 training, 512 validation and 512 segmented test images of 20 foreground and 1 background classes. The organisers also provided $10,057$ images for which only the bounding boxes of the objects present in the image are marked. We did not use these additional images for training our framework. For this data set we used a two-level hierarchy. The methods are evaluated using intersection vs. union criteria (Everingham et al, 2008) that penalises the performance of classes i and j given a mislabelling of i as j (see table 4). It should be emphases that this is not the same as the percentage

Method	Best	$E(\text{meth}) - E(\min)$	$\frac{E(\text{meth})}{E(\min)}$	Time
Range exp	265	75	1.000	6.1s
Range swap	137	9034	1.059	20s
α-expansion	109	256	1.002	6.3s
$\alpha\beta$-swap	42	9922	1.060	42s
TRW-S	12	38549	1.239	500s
BP	6	13456	1.081	120s
ICM	5	45955	1.274	25s

Fig. 12. *Comparison of methods on 295 testing images. From left to right the columns show the number of times they achieved the best energy (including ties), the average difference (E(method) $-$ $E(\min)$), the average ratio (E(method)$/E(\min)$) and the average time taken. All three inference methods proposed in this paper: the α-expansion of section 4.2, and the transformationally optimal range expansion and swap (section 5) significantly outperformed existing inference methods both in speed and accuracy. See table 11 for individual examples.*

of pixels correctly labelled. Quantitative comparison of performance with other methods is given in 4. Qualitative results are given in table 9. The only comparable methods used classification and detection priors trained over a much larger set of images. The reported results we show are from the actual challenge, and not the performance of the current implementation.

The hierarchical random field significantly outperformed random field approaches at single scale (pixels, segments) on all data sets. Experimentally, the approach was robust to the choice of the parameters and typically the same parameters performed well on all data sets. This suggests that the improvement of the performance comes from the incorporation of the different discriminative cues across multiple scales.

8.1 Comparison of Inference Methods

We evaluate α-expansion, $\alpha\beta$-swap, TRW-S, Belief Propagation (BP), Iterated Conditional Modes, and both the expansion and swap based variants of our unordered range moves on the problem of object class segmentation over the MSRC data set (Shotton et al, 2006), in which each pixel within an image must be assigned a label representing its class, such as grass, water, boat or cow. For *bp* we followed the same efficient implementation as TRW-S (Kolmogorov, 2006), but without the averaging step. Kappes et al (2013) showed that many of combinatoric approaches work well on pairwise cost functions defined over one set of segments. However, they would not scale to the number of variables or large cliques we consider.

We express the problem as a three layer hierarchy. Each pixel is represented by a random variable of the base layer. The second layer is formed by performing multiple unsupervised segmentations over the image, and associating one auxiliary variable with each segment - note that this use of several hierarchies results in overlapping segments. The children of each of these variables in $x^{(2)}$ are the variables contained within the segment, and pairwise connections are formed between adjacent segments. Further details are given in section 8 (MSRC).

We tested each algorithm on 295 test images, with an average of 70,000 pixels/variables in the base layer and up to 30,000 variables in a clique, and ran them either until convergence, or for a maximum of 500 iterations. In the table in figure 12 we compare the final energies obtained by each algorithm, showing the number of times they achieved an energy lower than or equal to all other methods, the average difference $E(\text{method}) - E(\min)$ and average ratio $E(\text{method})/E(\min)$. Empirically, the message passing algorithms TRW-S and BP appear ill-suited to inference over these dense hierarchical random fields. In comparison to the graph cut based move making algorithms, they had higher resulting energy, higher memory usage, and exhibited slower convergence.

While it may appear unreasonable to test message passing approaches on hierarchical energies when higher-order formulations such as (Komodakis and Paragios, 2009; Potetz and Lee, 2008) exist, we note that for the simplest hierarchy that contains only one additional layer of nodes and no pairwise connections in this second layer, higher-order and hierarchical message-passing approaches will be equivalent, as inference over the trees that represent higher-order potentials is exact. Similar relative performance by message passing schemes was observed in these cases.

In all tested images both α-expansion variants outperformed TRW-S, BP and ICM.

The later methods only obtained minimal cost labellings in images in which the optimal solution found contained only one label i.e. they were entirely labelled as grass or water. The comparison also shows that unordered range move variants usually outperform vanilla move making algorithms. The higher number of minimal labellings found by the range-move variant of $\alpha\beta$-swap in comparison to those of vanilla α-expansion can be explained by the large number of images in which two labels strongly dominate – unlike standard α-expansion both range move algorithms are guaranteed to find a global optimum of such a two label sub-problem. The typical behaviour of all methods alongside the lower bound of TRW-S can be seen in figure 12 and further, alongside qualitative results, in figures 11.

9 Conclusion

This work is a generalisation of many previous super-pixel based methods within a principled random field framework. Our approach enabled the integration of features and contextual priors defined over multiple super-pixels in one optimisation framework that supports efficient MAP estimation using graph cut based move making algorithms. To do this, we have examined the use of auxiliary variables in random fields

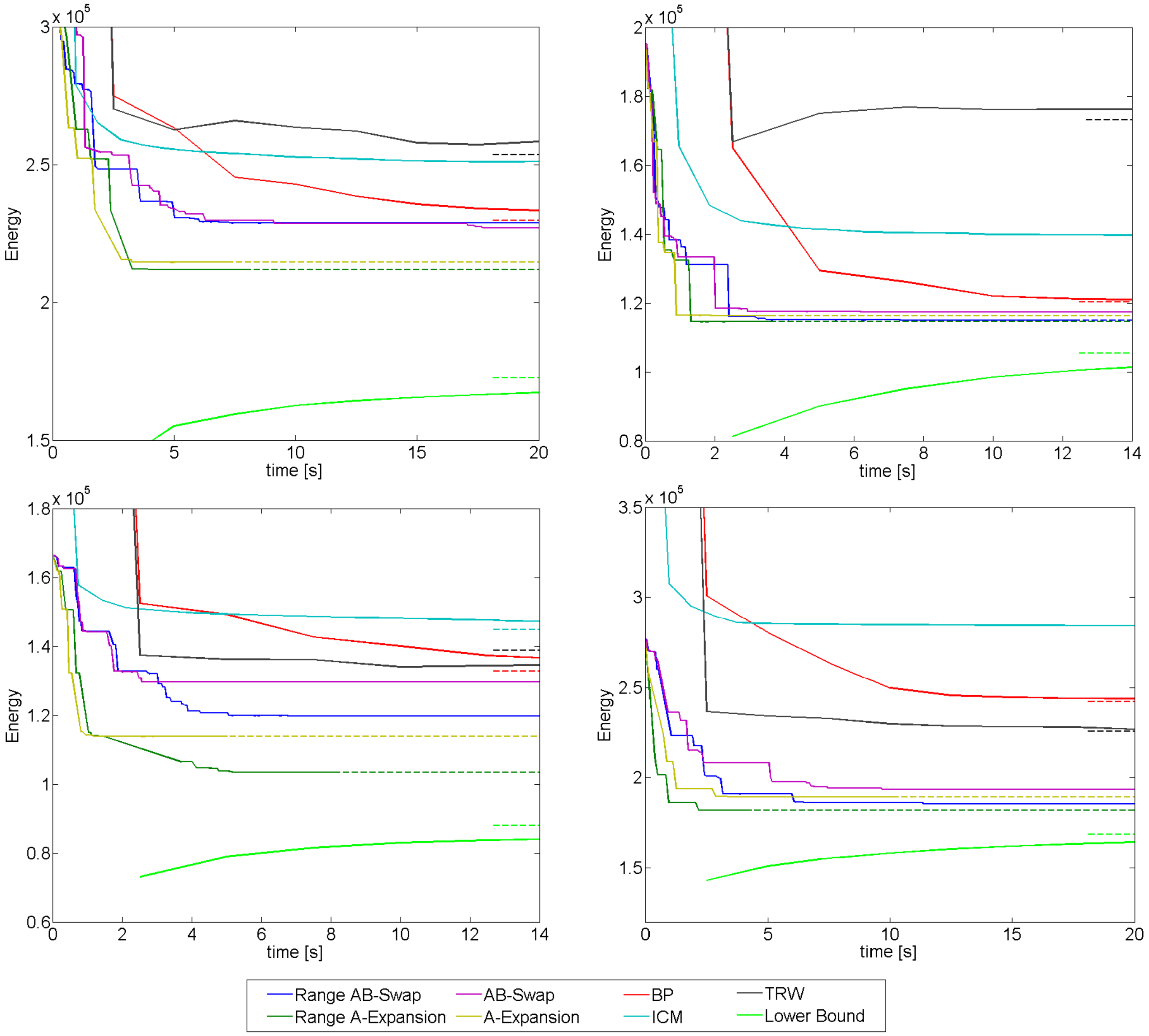


Fig. 11. **Best Viewed in Colour.** *This figure shows additional quantitative results taken from the* MSRC *data set (Shotton et al, 2006). Dashed lines indicate the final converged solution. The slow convergence and poor solutions found by* TRW *and* BP *are to be expected given the large number of cycles present in the graph. Of the remaining move making schemes, the relatively weak performance of $\alpha\beta$-swap and* ICM *is in line with the restricted space of moves available to them. While the three methods derived in this paper significantly outperform all other approaches, range α-expansion reliably dominates.*

which have been relatively neglected in computer vision over the past twenty years.

In doing so, we have shown that higher-order random fields are intimately related to pairwise models. This observation has allowed us to characterise some of the higher-order potentials which can be solved using modified expansion and swap algorithms.

We demonstrated the usefulness of our algorithms on the problem of object class segmentation where they have been shown to outperform state of the art approaches over challenging data sets both in speed and accuracy. We believe that similar improvements can be achieved for many other higher-order labelling problems both in computer vision and machine learning in general.

The flexibility and generality of our framework allowed us to propose and use novel pixel and segment based potential functions and achieve state-of-the-art results on some of the most challenging data sets for object class segmentation. We believe that use of the hierarchical random fields will yield similar improvements for other labelling problems. The source code is publicly available[6].

References

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2012) SLIC superpixels compared to

[6] http://cms.brookes.ac.uk/staff/PhilipTorr/ale.htm

state-of-the-art superpixel methods. Transactions on Pattern Analysis and Machine Intelligence

Alahari K, Russell C, Torr PHS (2010) Efficient piecewise learning for conditional random fields. In: Conference on Computer Vision and Pattern Recognition

Batra D, Sukthankar R, Tsuhan C (2008) Learning class-specific affinities for image labelling. In: Conference on Computer Vision and Pattern Recognition

Besag J (1986) On the statisical analysis of dirty pictures. Journal of the Royal Statistical Society

Blake A, Rother C, Brown M, Perez P, Torr P (2004) Interactive image segmentation using an adaptive GMMRF model. In: European Conference on Computer Vision

Boix X, Cardinal G, van de Weijer J, Bagdanov AD, Serrat J, Gonzalez J (2011) Harmony potentials: Fusing global and local scale for semantic image segmentation. International Journal on Computer Vision

Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In: International Conference on Image and Video Retrieval

Boykov Y, Jolly M (2001) Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: International Conference on Computer Vision

Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. Transactions on Pattern Analysis and Machine Intelligence

Comaniciu D, Meer P (2002) Mean shift: A robust approach toward feature space analysis. Transactions on Pattern Analysis and Machine Intelligence

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2008) The PASCAL Visual Object Classes Challenge (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/index.html

Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. International Journal of Computer Vision

Finley T, Joachims T (2008) Training structural SVMs when exact inference is intractable. In: International Conference on Machine Learning

Fix A, Gruber A, Boros E, Zabih R (2011) A graph cut algorithm for higher-order Markov random fields. In: International Conference on Computer Vision

Galleguillos C, Rabinovich A, Belongie S (2008) Object categorization using co-occurrence, location and appearance. In: Conference on Computer Vision and Pattern Recognition

Gould S, Fulton R, Koller D (2009a) Decomposing a scene into geometric and semantically consistent regions. In: International Conference on Computer Vision

Gould S, Gao T, Koller D (2009b) Region-based segmentation and object detection. In: Advances in Neural Information Processing Systems

He X, Zemel R, Ray D (2006) Learning and incorporating top-down cues in image segmentation. In: European Conference on Computer Vision

Hinton GE, Osindero S (2006) A fast learning algorithm for deep belief nets. Neural Computation

Hoiem D, Efros A, Hebert M (2005) Geometric context from a single image. In: International Conference on Computer Vision

Hoiem D, Efros AA, Hebert M (2006) Putting objects in perspective. In: Conference on Computer Vision and Pattern Recognition

Ishikawa H (2009) Higher-order clique reduction in binary graph cut. In: Conference on Computer Vision and Pattern Recognition

Ishikawa H (2011) Transformation of general binary MRF minimization to the first-order case. Transactions on Pattern Analysis and Machine Intelligence

Kappes JH, Andres B, Hamprecht FA, Schnorr C, Nowozin S, Batra D, Kim S, Kausler BX, Lellmann J, Komodakis N, Rother C (2013) A comparative study of modern inference techniques for discrete energy minimization problems. In: Conference on Computer Vision and Pattern Recognition

Kohli P, Kumar M, Torr PHS (2007) P^3 and beyond: Solving energies with higher order cliques. In: Conference on Computer Vision and Pattern Recognition

Kohli P, Ladicky L, Torr PHS (2008) Robust higher order potentials for enforcing label consistency. In: Conference on Computer Vision and Pattern Recognition

Kolmogorov V (2006) Convergent tree-reweighted message passing for energy minimization. Transactions on Pattern Analysis and Machine Intelligence

Kolmogorov V, Rother C (2006) Comparison of energy minimization algorithms for highly connected graphs. In: European Conference on Computer Vision

Kolmogorov V, Zabih R (2004) What energy functions can be minimized via graph cuts?. Transactions on Pattern Analysis and Machine Intelligence

Komodakis N, Paragios N (2009) Beyond pairwise energies: Efficient optimization for higher-order MRFs. In: Conference on Computer Vision and Pattern Recognition

Kumar MP, Koller D (2010) Efficiently selecting regions for scene understanding. In: Conference on Computer Vision and Pattern Recognition

Kumar MP, Veksler O, Torr PHS (2011) Improved moves for truncated convex models. Journal of Machine Learning Research

Kumar S, Hebert M (2005) A hierarchical field framework for unified context-based classification. In: International Conference on Computer Vision

Ladicky L, Russell C, Kohli P, Torr PHS (2010) Graph cut based inference with co-occurrence statistics. In: European Conference on Computer Vision

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In: International Conference on Machine Learning

Lan X, Roth S, Huttenlocher D, Black M (2006) Efficient belief propagation with learned higher-order markov random fields. In: European Conference on Computer Vision

Larlus D, Jurie F (2008) Combining appearance models and Markov random fields for category level object segmentation. In: Conference on Computer Vision and Pattern Recognition

Lempitsky V, Vedaldi A, Zisserman A (2011) A pylon model for semantic segmentation. In: Advances in Neural Information Processing Systems

Lim JJ, Arbelaez P, Gu C, Malik J (2009) Context by region ancestry. In: International Conference on Computer Vision

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision

Malik J, Belongie S, Leung T, Shi J (2001) Contour and texture analysis for image segmentation. International Journal of Computer Vision

Munoz D, Bagnell JA, Hebert M (2010) Stacked hierarchical labeling. In: European Conference on Computer Vision

Nowozin S, Gehler PV, Lampert CH (2010) On parameter learning in CRF-based approaches to object class image segmentation. In: European Conference on Computer Vision

Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Conference on Computer Vision and Image Processing

Pantofaru C, Schmid C, Hebert M (2008) Object recognition by integrating multiple image segmentations. In: European Conference on Computer Vision

Potetz B, Lee TS (2008) Efficient belief propegation for higher order cliques using linear constraint nodes. Computer Vision and Image Understanding

Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S (2007) Objects in context. In: International Conference on Computer Vision

Ramalingam S, Russell C, Ladicky L, Torr PH (2011) Efficient minimization of higher order submodular functions using monotonic boolean functions. Arxiv preprint arXiv:11092304

Ren X, Malik J (2003) Learning a classification model for segmentation. In: International Conference on Computer Vision

Reynolds J, Murphy K (2007) Figure-ground segmentation using a hierarchical conditional random field. In: Canadian Conference on Computer and Robot Vision

Rother C, Kolmogorov V, Blake A (2004) Grabcut: interactive foreground extraction using iterated graph cuts. In: SIGGRAPH

Rother C, Kumar S, Kolmogorov V, Blake A (2005) Digital tapestry. In: Conference on Computer Vision and Pattern Recognition

Russell B, Freeman W, Efros A, Sivic J, Zisserman A (2006) Using multiple segmentations to discover objects and their extent in image collections. In: Conference on Computer Vision and Pattern Recognition

van de Sande KEA, Gevers T, Snoek CGM (2008) Evaluation of color descriptors for object and scene recognition. In: Conference on Computer Vision and Pattern Recognition

Schlesinger D, Flach B (2006) Transforming an arbitrary minsum problem into a binary one. Tech. rep., Dresden University of Technology

Shi J, Malik J (2000) Normalized cuts and image segmentation. Transactions on Pattern Analysis and Machine Intelligence

Shotton J, Winn J, Rother C, Criminisi A (2006) *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: European Conference on Computer Vision

Shotton J, Johnson M, Cipolla R (2008) Semantic texton forests for image categorization and segmentation. In: Conference on Computer Vision and Pattern Recognition

Socher R, Lin CC, Ng AY, Manning CD (2011) Parsing natural scenes and natural language with recursive neural networks. International Conference on Machine Learning

Sontag D, Meltzer T, Globerson A, Jaakkola T, Weiss Y (2008) Tightening LP relaxations for MAP using message passing. In: Uncertainty in Artificial Intelligence

Szeliski R, Zabih R, Scharstein D, Veksler O, Kolmogorov V, Agarwala A, Tappen M, Rother C (2006) A comparative study of energy minimization methods for Markov random fields. In: European Conference on Computer Vision

Szummer M, Kohli P, Hoiem D (2008) Learning CRFs using graph cuts. In: European Conference on Computer Vision

Tao H, Sawhney H, Kumar R (2001) A global matching framework for stereo computation. In: International Conference on Computer Vision

Tarlow D, Zemel R, Frey B (2008) Flexible priors for exemplar-based clustering. In: Conference on Uncertainty in Artificial Intelligence

Tarlow D, Givoni I, Zemel R (2010) HOP-MAP: Efficient message passing with high order potentials. In: Artificial Intelligence and Statistics

Taskar B, Chatalbashev V, Koller D (2004a) Learning associative markov networks. In: International Conference on Machine Learning

Taskar B, Chatalbashev V, Koller D (2004b) Learning associative Markov networks. In: International Conference on Machine Learning

Tighe J, Lazebnik S (2010) Superparsing: Scalable nonparametric image parsing with superpixels. In: European Conference on Computer Vision

Torralba A, Murphy K, Freeman W (2004) Sharing features: efficient boosting procedures for multiclass object detection. In: Conference on Computer Vision and Pattern Recognition

Tsochantaridis I, Joachims T, Hofmann T, Altun Y (2005) Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research

Tu Z, Chen X, Yuille AL, Zhu SC (2003) Image parsing: Unifying segmentation, detection, and recognition. International Conference on Computer Vision

Veksler O (2007) Graph cut based optimization for MRFs with truncated convex priors. In: Conference on Computer Vision and Pattern Recognition

Wang J, Bhat P, Colburn A, Agrawala M, Cohen M (2005) Interactive video cutout. ACM Transactions on Graphics

Weiss Y, Freeman W (2001) On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. Transactions on Information Theory

Werner T (2009) High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF). In: Conference on Computer Vision and Pattern Recognition

Yang L, Meer P, Foran DJ (2007) Multiple class segmentation using a unified framework over mean-shift patches. In: Conference on Computer Vision and Pattern Recognition

Zhang Y, Hartley RI, Mashford J, Burn S (2011) Superpixels via pseudo-boolean optimization. In: International Conference on Computer Vision

Zhu L, Yuille AL (2005) A hierarchical compositional system for rapid object detection. In: Advances in Neural Information Processing Systems

Zhu L, Chen Y, Lin Y, Lin C, Yuille AL (2008) Recursive segmentation and recognition templates for 2D parsing. In: Advances in Neural Information Processing Systems

Ľubor Ladický received the PhD in computer vision at the Oxford Brookes University in 2011. Then he worked as a post doctoral researcher in the Visual Geometry Group at the University of Oxford. Currently he is a research assistant at the the Computer Vision and Geometry lab at the ETH Zürich. His current research interests are machine learning, computer vision and discrete optimization.

Chris Russell received an MMath degree in mathematics from the University of Oxford followed by a PhD degree from Oxford Brookes University in 2011. He was a postdoctoral researcher at Queen Mary, University of London and currently works at University College London. He works on machine learning and 3d reconstruction from video.

Pushmeet Kohli is a research scientist in the Machine Learning and Perception group at Microsoft Research Cambridge, an associate of the Psychometric centre and Trinity hall, University of Cambridge. Pushmeet's research revolves around Intelligent Systems and Computational Sciences, and he publishes in the fields of Machine Learning, Computer Vision, Information Retrieval, and Game Theory. His current research interests include human behaviour analysis and the prediction of user preferences. Pushmeet is interested in designing autonomous and intelligent computer vision, bargaining and trading systems which learn by observing and interacting with users on social media sites such as Facebook. He is also investigating the use of new sensors such as KINECT for the problems of human pose estimation, scene understanding and robotics.

Philip H.S. Torr did his PhD (DPhil) at the Robotics Research Group of the University of Oxford under Professor David Murray of the Active Vision Group. He worked for another three years at Oxford as a research fellow, and is still maintains close contact as visiting fellow there. He left Oxford to work for six years as a research scientist for Microsoft Research, first in Redmond USA in the Vision Technology Group, then in Cambridge UK founding the vision side of the Machine learning and perception group. He is now a Professor in Computer Vision and Machine Learning at Oxford Brookes University.