Int J Comput Vis (2013) 103:213–225
DOI 10.1007/s11263-012-0583-y

Inference Methods for CRFs with Co-occurrence Statistics

Ľubor Ladický · Chris Russell · Pushmeet Kohli · Philip H. S. Torr

Received: 21 January 2011 / Accepted: 30 September 2012 / Published online: 9 November 2012

Abstract The Markov and Conditional random fields (CRFs) used in computer vision typically model only local interactions between variables, as this is generally thought to be the only case that is computationally tractable. In this paper we consider a class of global potentials defined over all variables in the CRF. We show how they can be readily optimised using standard graph cut algorithms at little extra expense compared to a standard pairwise field. This result can be directly used for the problem of *class based image segmentation* which has seen increasing recent interest within computer vision. Here the aim is to assign a label to each pixel of a given image from a set of possible object classes. Typically these methods use random fields to model local interactions between pixels or super-pixels. One of the cues that helps recognition is global *object co-occurrence statistics*, a measure of which classes (such as chair or motorbike) are likely to occur in the same image together. There have been several approaches proposed to exploit this property, but all of them suffer from different limitations and typically carry a high computational cost, preventing their application on large images. We find that the new model we propose produces a significant improvement in the labelling compared to just using a pairwise model and that this improvement increases as the number of labels increases.

Keywords Conditional random fields · Object class segmentation · Optimization

1 Introduction

Class based image segmentation is a highly active area of computer vision research as shown by a spate of recent publications (Heitz and Koller 2008; Rabinovich et al. 2007; Shotton et al. 2006; Torralba et al. 2003; Yang et al. 2007). In this problem, every pixel of the image is assigned a choice of object class label, such as grass, person, or dining table. Formulating this problem probabilistically, in order to perform inference, is a difficult problem, as the cost or energy associated with any labelling of the image should take into account a variety of cues at different scales. A good labelling should take account of: low-level cues such as colour or texture (Shotton et al. 2006), that govern the labelling of single pixels; mid-level cues such as region continuity, symmetry (Ren et al. 2005) or shape (Borenstein and Malik 2006) that govern the assignment of regions within the image; and high-level statistics that encode inter-object relationships, such as which objects can occur together in a scene. This combination of cues makes for a multi-scale cost function that is difficult to optimise.

Current state of the art low-level approaches typically follow the methodology proposed in *Texton-boost* (Shotton et al. 2006), in which weakly predictive features such as colour, location, and texton response are used to learn a classifier which provides costs for a single pixel taking a particular

Ľubor Ladický, Chris Russell contributed equally and have joint first authorship.

Ľ. Ladický (✉)
University of Oxford, Oxford, UK
e-mail: lubor@robots.ox.ac.uk

C. Russell
Queen Mary College, University of London, London, UK
e-mail: chrisr@eecs.qmul.ac.uk

P. Kohli
Microsoft Research, Cambridge, UK
e-mail: pkohli@microsoft.com

P. H. S. Torr
Oxford Brookes University, Oxford, UK
e-mail: philiptorr@brookes.ac.uk

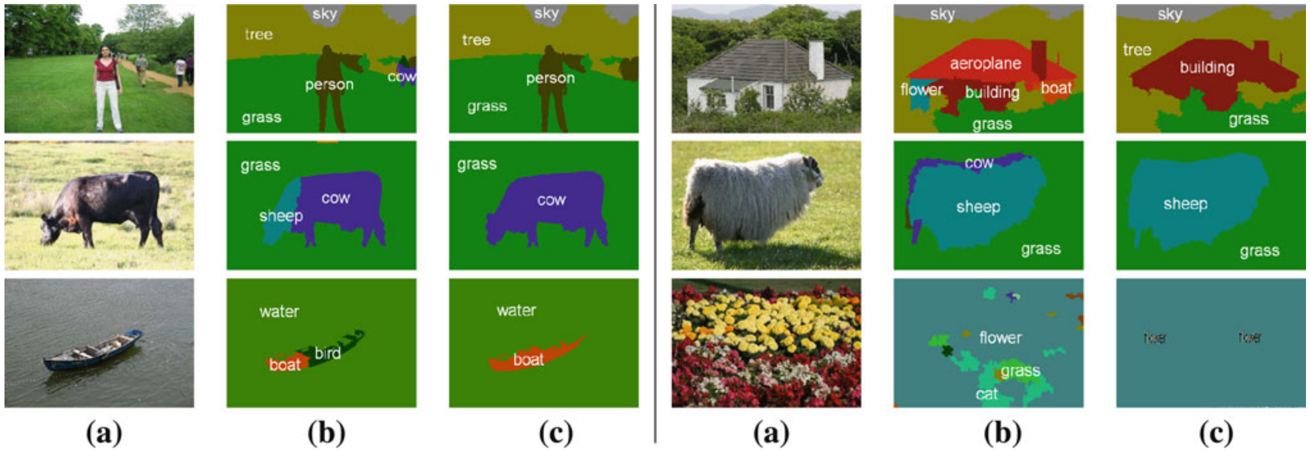

Fig. 1 Best viewed in colour. Qualitative results of object co-occurrence statistics. **a** Typical images taken from the MSRC data set (Shotton et al. 2006). **b** A labelling based upon a pixel based random field model (Ladicky et al. 2009) that does not take into account co-occurrence. **c** A labelling of the same model using co-occurrence statistics. The use of co-occurrence statistics to guide the segmentation results in a labelling that is more parsimonious and more likely to be correct. These co-occurrence statistics suppress the appearance of small unexpected classes in the labelling. *Top left* a mistaken hypothesis of a cow is suppressed *Top right* Many small classes are suppressed in the image of a building. Note that the use of co-occurrence typically changes labels, but does not alter silhouettes

label. These costs are combined in a contrast sensitive conditional random field (CRF) (Lafferty et al. 2001).

The majority of mid-level inference schemes (Russell et al. 2006; Larlus and Jurie 2008) do not consider pixels directly, rather they assume that the image has been segmented into super-pixels (Comaniciu and Meer 2002; Felzenszwalb and Huttenlocher 2004; Shi and Malik 2000). A labelling problem is then defined over the set of regions. A significant disadvantage of such approaches is that mistakes in the initial over-segmentation, in which regions span multiple object classes, cannot be recovered from. To overcome this Gould et al. (2009) proposed a method of reshaping super-pixels to recover from the errors, while works of (Kohli et al. 2008; Ladicky et al. 2009) proposed a novel framework which allowed for the integration of multiple region-based CRFs with a low-level pixel based CRF, and the elimination of inconsistent regions.

These approaches can be improved by the inclusion of costs based on high level statistics, including object class co-occurrence, which capture knowledge of scene semantics that humans often take for granted: for example the knowledge that cows and crocodiles are not kept together and less likely to appear in the same image; or that motorbikes are unlikely to occur near televisions. In this paper we consider object class co-occurrence to be a measure of how likely it is for a given set of object classes to occur together in an image. They can also be used to encode scene specific information such as the facts that computer monitors and stationary are more likely to occur in offices, or that trees and grass occur outside. The use of such costs can help prevent some of the most glaring failures in object class segmentation, such as the labelling of a boat surrounded by water mislabelled as a book.

As well as penalising strange combinations of object class labels appearing in an image, co-occurrence potentials can also be used to impose an minimum description length (MDL) prior, that encourages a parsimonious description of an image using fewer labels. As discussed eloquently in the recent work (Choi et al. 2010), the need for a bias towards parsimony becomes increasingly important as the number of classes to be considered increases. Figure 1 illustrates the importance of co-occurrence statistics in image labelling.

The promise of co-occurrence statistics has not been ignored by the vision community. Rabinovich et al. (2007) proposed the integration of such co-occurrence costs that characterise the relationship between two classes. Similarly Torralba et al. (2003) proposed scene-based costs that penalised the existence of particular classes in a context dependent manner. We shall discuss these approaches, and some problems with them in the next section.

2 CRFs and Co-occurrence

A conventional CRF is defined over a set of random variables $\mathcal{V} = \{1, 2, 3, \ldots, n\}$ where each variable takes a value from the label set $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$ corresponding to the set of object classes. An assignment of labels to the set of random variables will be referred to as a *labelling*, and denoted as $\mathbf{x} \in \mathcal{L}^{|\mathcal{V}|}$. We define a cost function $E(\mathbf{x})$ over the CRF of the form:

$$E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \qquad (1)$$

where the potential ψ_c is a cost function defined over a set of variables (called a clique) c, and $\mathbf{x}_c$ is the state of the set of random variables that lie within c. The set $\mathcal{C}$ of cliques is a subset of the power set of $\mathcal{V}$, *i.e.* $\mathcal{C} \subseteq P(\mathcal{V})$. In the majority of vision problems, the potentials are defined over a clique of size at most 2. *Unary potentials* are defined over a clique of size one, and typically based upon classifier responses [such as ada-boost (Shotton et al. 2006) or kernel SVMs (Schölkopf and Smola 2001)], while *pairwise potentials* are defined over cliques of size two and model the correlation between pairs of random variables.

2.1 Incorporating Co-occurrence Potentials

To model object class co-occurrence statistics a new term $K(\mathbf{x})$ is added to the energy:

$$E(\mathbf{x}) = \sum \psi_c(\mathbf{x}_c) + K(\mathbf{x}). \tag{2}$$

The question naturally arises as to what form an energy involving co-occurrence terms should take. We now list a set of desiderata that we believe are intuitive for any co-occurrence cost.

(i) *Global Energy* We would like a formulation of co-occurrence that allows us to estimate the segmentation using all the data directly, by minimising a *single* cost function of the form (2). Rather than any sort of two stage process in which a hard decision is made of which objects are present in the scene *a priori* as in (Torralba et al. 2003).

(ii) *Invariance* The co-occurrence cost should depend only on the labels present in an image, it should be invariant to the number and location of pixels that object occupies. To reuse an example from (Toyoda and Hasegawa 2008), the surprise at seeing a polar bear in a street scene should not not vary with the number of pixels that represent the bear in the image.

(iii) *Efficiency* Inference should be tractable, *i.e.* the use of co-occurrence should not be the bottle-neck preventing inference. As the memory requirements of any conventional inference algorithm (Szeliski et al. 2006) is typically $O(|\mathcal{V}|)$ for vision problems, the memory requirements of a formulation incorporating co-occurrence potentials should also be $O(|\mathcal{V}|)$.

(iv) *Parsimony* The cost should follow the principle of parsimony in the following way: if several solutions are almost equally likely then the solution that can describe the image using the fewest distinct labels should be chosen. Whilst this might not seem important when classifying pixels into a few classes, as the set of putative labels for an image increases the chance of speckle noise due to misclassification will increase unless a parsimonious solution is encouraged.

While these properties seem uncontroversial, no prior work exhibits property (ii). Similarly, no approaches satisfy properties (i) and (iii) simultaneously. In order to satisfy condition (ii) the co-occurrence cost $K(\mathbf{x})$ defined over $\mathbf{x}$ must be a function defined on the set of labels $L(\mathbf{x}) = \{l \in \mathcal{L} : \exists x_i = l\}$ present in the labelling $\mathbf{x}$; this guarantees invariance to the size of an object:

$$K(\mathbf{x}) = C(L(\mathbf{x})) \tag{3}$$

Adding the co-occurrence term to the CRF cost function (1), we have:

$$E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) + C(L(\mathbf{x})). \tag{4}$$

To satisfy the parsimony condition (iv) potentials must act to penalise the unexpected appearance of combinations of labels in a labelling. This observation can be formalised as the statement that the cost $C(L)$ is monotonically increasing with respect to the label set L *i.e.*:

$$L_1 \subset L_2 \implies C(L_1) \leq C(L_2). \tag{5}$$

The new potential $C(L(\mathbf{x}))$ can be seen as a particular higher order potential defined over a clique which includes the whole of $\mathcal{V}$, i.e. $\psi_{\mathcal{V}}(\mathbf{x})$.

2.2 Prior Work

There are two existing approaches to co-occurrence potentials, neither of which use potentials defined over a clique of size greater than two. The first makes an initial hard estimate of the type of scene, and updates the unary potentials associated with each pixel to encourage or discourage particular choices of label, on the basis of how likely they are to occur in the scene. The second approach models object co-occurrence as a pairwise potential between regions of the image.

Torralba et al. (2003) proposed the use of additional unary potentials to capture scene based occurrence priors. Their costs took the form:

$$K(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi(x_i). \tag{6}$$

While the complexity of inference over such potentials scales linearly with the size of the graph, they are prone to over counting costs, violating (ii), and require an initial hard decision of scene type before inference, which violates (i). As it encourages the appearance of all labels which are common to a scene, it does not necessarily encourage parsimony (iv).

A similar approach was seen in the Pascal VOC2008 object segmentation challenge, where the best performing method, by (Csurka and Perronnin 2008), worked in two stages. Initially the set of object labels present in the image was estimated, and in the second stage, a label from the estimated label set was assigned to each image pixel. As no cost function $K(\cdot)$ was proposed, it is open to debate if it satisfied (ii) or (iv).

Rabinovich et al. (2007); Galleguillos et al. (2008), and independently Toyoda and Hasegawa (2008), proposed co-occurrence as a soft constraint that approximated $C(L(\mathbf{x}))$ as a pairwise cost defined over a *fully connected graph* that took the form:

Table 1 A comparison of the capabilities of existing image co-occurrence formulations against our new approach

Method	Global energy (i)	Invariance (ii)	Efficiency (iii)	Parsimony (iv)
Unary (Torralba et al. 2003)	✓	×	✓	×
Pairwise (Rabinovich et al. 2007; Galleguillos et al. 2008; Toyoda and Hasegawa 2008)	✓	×	×	×
Hard decisions (Csurka and Perronnin 2008)	×	—	✓	—
Our approach	✓	✓	✓	✓

See Sect. 2.2 for details

$$K(\mathbf{x}) = \sum_{i,j \in \mathcal{V}} \phi(x_i, x_j), \tag{7}$$

where ϕ was some potential which penalised labels that should not occur together in an image. Unlike our model (4) the penalty cost for the presence of pairs of labels, that rarely occur together, appearing in the same image grows with the number of random variables taking these labels, violating assumption (ii). While this serves as a functional penalty that prevents the occurrence of many classes in the same labelling, it does not accurately model the co-occurrence costs we described earlier. The memory requirements of inference scales badly with the size of a fully connected graph. It grows with complexity $O(|\mathcal{V}|^2)$ rather than $O(|\mathcal{V}|)$ with the size of the graph, violating constraint (iii). Providing the pairwise potentials are semi-metric (Boykov et al. 2001), it does satisfy the parsimony condition (iv).

To minimise these difficulties, previous approaches defined variables over segments rather than pixels. Such segment based methods work under the assumption that some segments share boundaries with objects in the image. This is not always the case, and this assumption may result in dramatic errors in the labelling. The relationship between previous approaches and the desiderata can be seen in Table 1.

Two efficient schemes (Delong et al. 2010; Hoiem et al. 2007) have been proposed for the minimisation of the number of classes or objects present in a scene. While neither of them directly models class based co-occurrence relationships, their optimisation approaches satisfy the desiderata proposed in Sect. 2.1.

Hoiem et al. (2007), proposed a cost based on the number of objects in the scene, in which the presence of any instance of any object incurs a uniform penalty cost. For example, the presence of both a motorbike and a bus in a single image is penalised as much as the presence of two buses. Minimising the number of objects in a scene is a good method of encouraging consistent labellings, but does not capture any co-occurrence relationship between object classes.

If we view Hoiem's work as assigning a different label to every instance of an object class, their label set costs take the form:

$$C(L(\mathbf{x})) = k||L(\mathbf{x})|| \tag{8}$$

In a recent work, independently appearing at the same time as ours, Delong et al. (2010) also proposed the use of a cost over the number of labels present. In general their approach allowed a penalty cost to be from certain subset is present in an image. They proposed an ingenious use of this cost to combine probabilistic formulations such as Akaike's information criteria, or the Bayesian Information Criteria to efficiently solve a long standing problem in motion segmentation (See also Torr (1998) for discussion of this problem). The general form of their costs is:

$$C(L(\mathbf{x})) = \sum_{L \subseteq \mathcal{L}} k_L \delta(L(\mathbf{x}) \cap L \neq \emptyset), \tag{9}$$

where $\delta(\,)$ is the Kronecker indicator function.

Note that the costs of Delong et al. (2010) and Hoiem et al. (2007) both satisfy the inequality:

$$C(L_1 \cup L_2) \leq C(L_1) + C(L_2), \tag{10}$$

where L_1 and L_2 are any subsets of labels of $\mathcal{L}$. Consequentially, their models are unable to express to co-occurrence potentials which say that certain classes, such as the previously mentioned example of polar bear and street, are less likely to occur together than in separate images.

3 Inference on Global Co-occurrence Potentials

Consider the energy (4) defined in Sect. 2.1. The inference problem becomes:

$$\begin{aligned} \mathbf{x}^* &= \arg\min_{\mathbf{x} \in \mathcal{L}^{|\mathcal{V}|}} \textstyle\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) + C(L(\mathbf{x})) \\ &\quad \text{s.t. } \mathbf{x} \in \mathcal{L}^{|\mathcal{V}|},\ L(\mathbf{x}) = \{l \in \mathcal{L} : \exists x_i = l\}. \end{aligned} \tag{11}$$

In this section we show that the problem of minimising this energy can be reformulated as an integer program and solved using LP-relaxation. We will also show how it can be transformed into pairwise energy by adding one auxiliary variable connected to all pixels in the image and solved using Belief Propagation (Weiss and Freeman 2001) or TRW-S (Kolmogorov 2006). However, reparameterisation methods such as these perform badly on densely connected graphs (Kolmogorov and Rother 2006; Russell et al. 2010). Then we show that the problem can be solved efficiently using move-making $\alpha\beta$-swap and α-expansion moves (Boykov et al. 2001), where the number of additional edges of the graph grows linearly with the number of variables in the graph. In contrast to (Rabinovich et al. 2007), these algorithms can be applied to large graphs with more than 200,000 variables.

3.1 The Integer Programming Formulation, and Its Linear Relaxation

In the following two subsections, we make the simplifying assumption that the cost (1) is currently represented as a pairwise energy. The minimisation of the energy function (4) can be formulated as an Integer Program (IP) (Wainwright et al. 2002; Schlesinger 1976). A vector $\mathbf{z}$ of binary indicator variables is used to represent the assignment of labels. $\mathbf{z}$ is composed of $z_{i;a} \forall a \in \mathcal{L}, \forall i \in \mathcal{V}$, and, $z_{ij;ab} \forall a, b \in \mathcal{L}, (i, j) \in \mathcal{E}$ where $\mathcal{E}$ is the set of edges, to represent the state of variables x_i, x_j such that:

$$z_{i;a} = \begin{cases} 1 & \text{if } x_i = a \\ 0 & \text{otherwise} \end{cases}, \qquad (12)$$

$$z_{ij;ab} = \begin{cases} 1 & \text{if } x_i = a \text{ and } x_j = b \\ 0 & \text{otherwise.} \end{cases}$$

In addition $\mathbf{z}$ is composed of z_L, there are indicator variables that show which subset of labels $L(\mathbf{x})$ is used for the assignment. There are $2^{|\mathcal{L}|}$ such variables in total, one variable z_L for every $L \subseteq \mathcal{L}$. We write:

$$z_L = \begin{cases} 1 & \text{if } L = L(\mathbf{x}) \\ 0 & \text{otherwise.} \end{cases} \qquad (13)$$

Thus, $\mathbf{z}$ is a binary vector of length $|\mathcal{V}| \cdot |\mathcal{L}| + |\mathcal{E}| \cdot |\mathcal{L}|^2 + 2^{|\mathcal{L}|}$. The resulting IP can be written as:

$$\min_{\mathbf{z}} \sum_{i \in \mathcal{V}, a \in \mathcal{L}} \psi_i(a) z_{i;a} + \sum_{\substack{(i,j) \in \mathcal{E}, \\ a,b \in \mathcal{L}}} \psi_{i,j}(a, b) z_{ij;ab}$$
$$+ \sum_{L \subseteq \mathcal{L}} C(L) z_L \qquad (14)$$

such that:

$$\sum_a z_{ij;ab} = z_{j;b}, \qquad \forall (i, j) \in \mathcal{E}, b \in \mathcal{L}, \qquad (15)$$
$$\sum_b z_{ij;ab} = z_{i;a}, \qquad \forall (i, j) \in \mathcal{E}, a \in \mathcal{L}, \qquad (16)$$
$$\sum_a z_{i;a} = 1, \qquad \forall i \in \mathcal{V}, \qquad (17)$$
$$\sum_{L \ni a} z_L \geq z_{i;a}, \qquad \forall i \in \mathcal{V}, a \in \mathcal{L}, L \subseteq \mathcal{L} \qquad (18)$$
$$\sum_{L \subseteq \mathcal{L}} z_L = 1, \qquad (19)$$
$$z_{i;a}, z_{ij;ab}, z_L \in \{0, 1\} \quad \forall i \in \mathcal{V}, \forall (i, j) \in \mathcal{E}, \quad \forall a, b \in \mathcal{L}, \quad \forall L \subseteq \mathcal{L}. \qquad (20)$$

The marginal consistency and uniqueness constraints (15–17) are well-known and used in the standard IP formulation of the labelling problem (Komodakis et al. 2007; Kumar and Torr 2008; Wainwright et al. 2005; Werner 2007). To enforce the consistency between labelling and the label set indicator variables z_L (13), two new properties which we refer to as "inclusion" and "exclusion" properties must be satisfied. The exclusion property which ensures that if $z_L = 1$, no variable takes a label not present in L, is enforced by the exclusion constraints (18). While, the inclusion property guarantees that if $z_L = 1$, then for each label $l \in L$ there exists at least one variable $z_{i;l}$ such that $z_{i;l} = 1$, is enforced by parsimony. To see why this is the case, consider a contrapositive solution where there is a label $l \in L$ not present in the solution. In this case, the solution $\mathbf{z}$ altered by $z_L = 0$ and $z_{L \setminus \{l\}} = 1$ would also satisfy all constraints (15–18) and due to the parsimony property would have the same or lower cost function (14). Thus, there exists a global optima satisfying $z_L = 1$ such that $L(\mathbf{x}) = L$. The constraint (19) guarantees that there is exactly one z_L such that $z_L = 1$. The final constraints (20) ensure that all indicator variables are binary.

The inclusion property can also be explicitly enforced by the set of constraints:

$$\sum_{i \in \mathcal{V}} z_{i;a} \geq z_L, \qquad \forall a \in L \subseteq \mathcal{L}. \qquad (21)$$

In that case the formulation would be applicable also to co-occurrence potentials not satisfying the parsimony property. However, this would encourage degenerate solutions in which only one pixel takes a particular label.

The IP can be converted to a linear program (LP) by relaxing the integral constraints (20) to

$$z_{i;a}, z_{ij;ab}, z_L \in [0, 1] \quad \forall i \in \mathcal{V}, \forall (i, j) \in \mathcal{E}, \quad \forall a, b \in \mathcal{L}, \quad \forall L \subseteq \mathcal{L}. \qquad (22)$$

The resulting linear program can be solved using any general purpose LP solver, and an integer solution, can be induced using rounding schemes such as those of Kleinberg and Tardos (2002). While this approach allows co-occurrence to be computed effectively for small images, over large images the memory and time requirements of standard LP solvers make this approach infeasible.

In many practical cases the co-occurrence cost $C(L)$ is defined as the sum over costs k_L for co-occurrence of subsets of labels, for example all pairs of labels. The cost k_L for each subset is taken if all the labels L are present in an image:

$$C(L) = \sum_{B \subseteq L} k_B, \qquad (23)$$

where $k_B \geq 0$. In general any cost $C(L)$ can be decomposed uniquely into the sum over subsets recursively as:

$$k_B = C(B) - \sum_{B' \subset B} k_{B'}, \qquad (24)$$

however some coefficients k_B may become negative. The positivity constraint $k_B > 0$ did not need to be satisfied for the linear program (15–20). We show, that in the case of the co-occurrence cost defined as a sum over non-negative costs for low-order subset of labels, we can remove the exponential complexity of the linear program (exponential in the number of labels). In this case we will need one variable z_L for each subset, which is either of the cardinality 1 or has a nonzero cost $k_L > 0$. Unlike in (13), $z_L = 1$ if L is the subset of labels present in an image:

$$z_L = \begin{cases} 1 & \text{if } L \subseteq L(\mathbf{x}) \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

In this case, the linear program becomes:

$$\min_{\mathbf{z}} \sum_{i\in\mathcal{V},a\in\mathcal{L}} \psi_i(a)z_{i;a} + \sum_{\substack{(i,j)\in\mathcal{E},\\ a,b\in\mathcal{L}}} \psi_{i,j}(a,b)z_{ij;ab} + \sum_{a\in\mathcal{L}} k(\{a\})z_{\{a\}} + \sum_{\substack{L\subseteq\mathcal{L}\\ |L|\geq 2}} k(L)z_L \tag{26}$$

such that:

$$\sum_a z_{ij;ab} = z_{j;b}, \qquad \forall(i,j)\in\mathcal{E}, b\in\mathcal{L}, \tag{27}$$

$$\sum_b z_{ij;ab} = z_{i;a}, \qquad \forall(i,j)\in\mathcal{E}, a\in\mathcal{L}, \tag{28}$$

$$\sum_a z_{i;a} = 1, \qquad \forall i\in\mathcal{V}, \tag{29}$$

$$z_{\{a\}} \geq z_{i;a}, \qquad \forall i\in\mathcal{V}, a\in\mathcal{L} \tag{30}$$

$$z_L \geq \sum_{a\in L} z_{\{a\}} - |L| + 1, \qquad \forall L\subseteq\mathcal{L}, |L|\geq 2 \tag{31}$$

$$z_{i;a}, z_{ij;ab}, z_L \in \{0,1\} \quad \forall i\in\mathcal{V}, \forall(i,j)\in\mathcal{E}, \quad \forall a,b\in\mathcal{L}, \quad \forall L\subseteq\mathcal{L}. \tag{32}$$

The constraints (30) guarantee that $z_{\{a\}} = 1$ if the label a is present in an image. The constraints (31) enforce that for all L with the cardinality larger than two, $z_L = 1$ if all labels in L are present in an image. In many practical cases, when the cost is defined as a sum over costs for each label as in (Delong et al. 2010), or each pair of labels, this LP program becomes feasible for standard LP solvers.

We next show, that the higher order energy (1) can be transformed into a pairwise energy function with the addition of a single auxiliary variable L that takes $2^{|\mathcal{L}|}$ states.

3.2 Pairwise Representation of Co-occurrence Potentials

The optimization of the energy (4) is equivalent to the pairwise energy function with co-occurrence cost represented using one auxiliary variable z that takes a label from the set of subsets $z \in 2^{\mathcal{L}}$. The unary potential for this auxiliary variable is equal to the corresponding co-occurrence cost:

$$\psi_u(z) = C(z) \qquad \forall z\in 2^{\mathcal{L}}. \tag{33}$$

The exclusion property is enforced by using a sufficiently large pairwise cost $K \rightarrow \infty$ for each pair of inconsistent labelling of pixel $x_i \in \mathbf{x}$ and z as:

$$\psi_p(x_i, z) = K\delta(x_i \notin z) \qquad \forall x_i \in \mathbf{x}. \tag{34}$$

The inclusion property is implicitly encoded in a similar way to the IP formulation as it arises naturally in the usual solutions due to the parsimony. If $z = L$ and there was a label $l \in L$ such that $\forall x_i \in \mathbf{x} : x_i \neq l$, then the solution with $z = L \setminus \{l\}$ would have the same or lower cost $E(\mathbf{x})$).

This formulation allows us to use any approach from the wide body of standard inference techniques (Boykov et al. 2001; Kolmogorov 2006; Szeliski et al. 2006) to minimize this function. However, the complexity grows exponentially with the size of the label set. In the case where the costs can also be decomposed into the sum of positive co-occurrence costs for low-order subsets, the exponential dependency on the size of label set can be removed. The new pairwise formulation contains one variable z_L for each subset with non-zero cost $k_L > 0$. It takes the label $l \in L$, which is currently not present in an image, or label $\emptyset$ if all labels $l \in L$ are present in an image. The unary potential for all auxiliary variables is equal to the corresponding co-occurrence cost, if all labels $l \in L$ are present in an image:

$$\psi_u(z_L) = k(L)\delta(z_L = \emptyset) \qquad \forall L\in 2^{\mathcal{L}}. \tag{35}$$

The consistency of the state of z_L with the labelling on an image is enforced by using a sufficiently large pairwise cost $K \rightarrow \infty$ for each pair of inconsistent labelling of pixel $x_i \in \mathbf{x}$ and z_L as:

$$\psi_p(x_i, z_L) = K\delta(z_L = l)\delta(x_i = l) \quad \forall x_i \in \mathbf{x}, L\in\mathcal{L}. \tag{36}$$

Inference can be performed on this graph using most of the message passing algorithms designed for general pairwise graphs, however in our experiments (explained in Sect. 4) such message passing algorithms were much slower than the graph cut based algorithm, even though the method led to the same solution for every test image.

3.3 $\alpha\beta$-Swap Moves

Move making algorithms iteratively project the problem into a smaller subspace of possible solutions containing current solution. Solving this sub-problem proposes optimal moves which guarantee that the energy decreases after each move and must eventually converge. The performance of move making algorithms depends dramatically on the size of the move space. The expansion and swap move algorithms we consider project the problem into two label sub-problem and under the assumption that the projected energy is pairwise and submodular, it can be solved using graph cuts. We derive graph constructions only for term $C(L(\mathbf{x}))$. The final graph is the merger of the graph for optimising the standard CRF (Boykov et al. 2001) and the derived graph construction for the co-occurrence term.

The swap and expansion move algorithms can be encoded as a vector of binary variables $\mathbf{t} = \{t_i, \forall i \in \mathcal{V}\}$. The transformation function $T(\mathbf{x}^p, \mathbf{t})$ of a move algorithm takes the current labelling $\mathbf{x}^p$ and a move $\mathbf{t}$ and returns the new labelling $\mathbf{x}$ induced by the move.

In an $\alpha\beta$-swap move every random variable x_i whose current label is α or β can transition to a new label of α or β. One iteration of the algorithm involves making moves for all pairs (α, β) in $\mathcal{L}^2$ successively. The transformation function $T_{\alpha\beta}(x_i, t_i)$ for an $\alpha\beta$-swap transforms the label of a random variable x_i as:

$$T_{\alpha\beta}(x_i, t_i) = \begin{cases} \alpha & \text{if } x_i \in \{\alpha, \beta\} \text{ and } t_i = 0, \\ \beta & \text{if } x_i \in \{\alpha, \beta\} \text{ and } t_i = 1. \end{cases} \tag{37}$$

Consider a swap move over the labels α and β, starting from an initial label set $L(\mathbf{x})$. We assume that either α or β is present in the image. Then, after a swap move, the labels present must be an element of S which we define as:

$$S = \{L(\mathbf{x}) \cup \{\alpha\} \setminus \{\beta\}, L(\mathbf{x}) \cup \{\beta\} \setminus \{\alpha\}, L(\mathbf{x}) \cup \{\alpha, \beta\}\}. \tag{38}$$

Let $\mathcal{V}_{\alpha\beta}$ be the set of variables currently taking label α or β. The move energy for $C(L(\mathbf{x}))$ is:

$$E(\mathbf{t}) = \begin{cases} C_\alpha = C(L(\mathbf{x}) \cup \{\alpha\} \setminus \{\beta\}) & \text{if } \forall i \in \mathcal{V}_{\alpha\beta},\ t_i = 0, \\ C_\beta = C(L(\mathbf{x}) \cup \{\beta\} \setminus \{\alpha\}) & \text{if } \forall i \in \mathcal{V}_{\alpha\beta},\ t_i = 1, \\ C_{\alpha\beta} = C(L(\mathbf{x}) \cup \{\alpha, \beta\}) & \text{otherwise.} \end{cases} \tag{39}$$

Note that, if $C(L)$ is monotonically increasing with respect to L then, by definition, $C_\alpha \leq C_{\alpha\beta}$ and $C_\beta \leq C_{\alpha\beta}$.

Let $\mathbf{t}' = \arg\min_{\mathbf{t}} E'(\mathbf{t})$ be the optimal move for the standard pairwise move energy $E'(\mathbf{t})$ without the co-occurrence term. Let us first consider the case where this solution contains both 0 s and 1 s. Because the co-occurrence term is the same for all mixed solutions, this move is better than any other mixed solution also including the co-occurrence term. Thus, the optimal move with co-occurrence term is either homogenous or the same as the optimal move if we did not have a co-occurrence term, and can be found as:

$$\mathbf{t}^* = \arg\min_{\mathbf{t}}(E'(\mathbf{t}') + C_{\alpha\beta}, E'(\mathbf{0}) + C_\alpha, E'(\mathbf{1}) + C_\beta), \tag{40}$$

where $\mathbf{0}$ and $\mathbf{1}$ are uniform vectors composed entirely of 0 or 1 respectively. In case the optimal solution without co-occurrence is homogenous, due to the parsimony condition

$$\forall \mathbf{t} : E(\mathbf{t}') \leq E(\mathbf{t}) \implies E'(\mathbf{t}') + C_\alpha \leq E'(\mathbf{t}) + C_{\alpha\beta} \tag{41}$$

and thus the optimal move is the minimum of the homogenous moves. Note, that this approach can be used only if the parsimony condition is satisfied.

Even though there exist an efficient solution (as described above) similar to the one in (Delong et al. 2010) to find the optimal $\alpha\beta$-swap move for energies with co-occurrence, for illustration we also derive its graph construction directly solvable using graph cuts. It will give us an intuition about the construction of the α-expansion move.

Lemma 1 *For a function $C(L)$, monotonically increasing with respect to L, the move energy can be represented as a binary submodular pairwise cost with two auxiliary variables z_α and z_β as:*

$$\begin{aligned} E(\mathbf{t}) = C_\alpha + C_\beta - C_{\alpha\beta} + \min_{z_\alpha, z_\beta} \Bigg[& (C_{\alpha\beta} - C_\alpha) z_\beta \\ & + (C_{\alpha\beta} - C_\beta)(1 - z_\alpha) + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha,\beta} - C_\alpha) t_i (1 - z_\beta) \\ & + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i) z_\alpha) \Bigg]. \end{aligned} \tag{42}$$

Lemma 1 Proof See appendix. This binary function is pairwise submodular and thus can be solved efficiently using graph cuts.

3.4 α-Expansion Moves

In an α-expansion move every random variable may either retain its current label or transition to label α. One iteration of the algorithm involves making moves for all α in $\mathcal{L}$ successively. The transformation function $T_\alpha(x_i, t_i)$ for an α-expansion move transforms the label of a random variable x_i as:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha & \text{if } t_i = 0 \\ x_i & \text{if } t_i = 1. \end{cases} \tag{43}$$

To derive a graph-construction that approximates the true cost of an α-expansion move we use the decomposition (23), which will allow us to decompose the move energy into the part depending only on the presence of the label α and the part depending only on the presence of all other labels after the move. We do not assume all costs k_B are non-negative.

As a simplifying assumption, let us first assume there is no variable currently taking label α. Let A be set of labels currently present in the image and $\delta_l(\mathbf{t})$ be set to 1 if label l is present in the image after the move and 0 otherwise. Then:

$$\delta_\alpha(\mathbf{t}) = \begin{cases} 1 & \text{if } \exists i \in \mathcal{V} \text{ s.t. } t_i = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

$$\forall l \in A, \delta_l(\mathbf{t}) = \begin{cases} 1 & \text{if } \exists i \in \mathcal{V}_l \text{ s.t.} t_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

The α-expansion move energy of $C(L(\mathbf{x}))$ can be written as:

$$\begin{aligned} E(\mathbf{t}) &= E_{\text{new}}(\mathbf{t}) - E_{\text{old}} \\ &= \sum_{B \subseteq A \cup \{\alpha\}} k_B \prod_{l \in B} \delta_l(\mathbf{t}) - C(A). \end{aligned} \tag{46}$$

Ignoring the constant term and decomposing the sum into parts with and without terms dependent on α we have:

$$E(\mathbf{t}) = \sum_{B \subseteq A} k_B \prod_{l \in B} \delta_l(\mathbf{t}) + \sum_{B \subseteq A} k_{B \cup \{\alpha\}} \delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t}). \tag{47}$$

As either α or all subsets $B \subseteq A$ are present after any move, the following statement holds:

$$\delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t}) = \delta_\alpha(\mathbf{t}) + \prod_{l \in B} \delta_l(\mathbf{t}) - 1. \tag{48}$$

This equality can be checked for all three cases, where either $\delta_\alpha(\mathbf{t})$ or $\prod_{l \in B} \delta_l(\mathbf{t})$ or both are equal to 1. Replacing the term $\delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t})$ and disregarding new constant terms, equation (46) becomes:

$$\begin{aligned} E(\mathbf{t}) &= \sum_{B \subseteq A} k_{B \cup \{\alpha\}} \delta_\alpha(\mathbf{t}) + \sum_{B \subseteq A} (k_B + k_{B \cup \{\alpha\}}) \prod_{l \in B} \delta_l(\mathbf{t}) \\ &= k'_\alpha \delta_\alpha(\mathbf{t}) + \sum_{B \subseteq A} k'_B \prod_{l \in B} \delta_l(\mathbf{t}), \end{aligned} \tag{49}$$

where $k'_\alpha = \sum_{B \subseteq A} k_{B \cup \{\alpha\}} = C(B \cup \{\alpha\}) - C(B)$ and $k'_B = k_B + k_{B \cup \{\alpha\}}$.

$E(\mathbf{t})$ is, in general, a higher-order non-submodular energy, and intractable. However, when proposing moves we can use the procedure described in (Narasimhan and Bilmes 2005; Rother et al. 2005; Kumar et al. 2011) and over-estimate the higher order components $K(A, \mathbf{t}) = \sum_{B \subseteq A} k'_B \prod_{l \in B} \delta_l(\mathbf{t})$ of the cost of moving from the current solution. Let $k'^{(0)}_B = k'_B$ and $K^{(0)}(A, \mathbf{t}) = K(A, \mathbf{t})$. For any $l' \in A$:

$$\begin{aligned} K^{(i)}(A, \mathbf{t}) &= \sum_{B \subseteq A} k'^{(i)}_B \prod_{l \in B} \delta_l(\mathbf{t}) \\ &= \sum_{B \subseteq A \setminus \{l'\}} (k'^{(i)}_B + k'^{(i)}_{B \cup \{l'\}} \delta_{l'}(\mathbf{t})) \prod_{l \in B} \delta_l(\mathbf{t}) \\ &= \sum_{B \subseteq A \setminus \{l'\}} (k'^{(i)}_B + k'^{(i)}_{B \cup \{l'\}}) \prod_{l \in B} \delta_l(\mathbf{t}) \\ &\quad -(1 - \delta_{l'}(\mathbf{t})) \sum_{B \subseteq A \setminus \{l'\}} k'^{(i)}_{B \cup \{l'\}} \prod_{l \in B} \delta_l(\mathbf{t}) \\ &\le \sum_{B \subseteq A \setminus \{l'\}} (k'^{(i)}_B + k'^{(i)}_{B \cup \{l'\}}) \prod_{l \in B} \delta_l(\mathbf{t}) \\ &\quad -(1 - \delta_{l'}(\mathbf{t})) \min_{S \subseteq A \setminus \{l'\}} \sum_{B \subseteq S} k'^{(i)}_{B \cup \{l'\}} \\ &= K^{(i+1)}(A \setminus \{l'\}, \mathbf{t}) - k''_{l'} + k''_{l'} \delta_{l'}(\mathbf{t}), \end{aligned} \tag{50}$$

where $k''_{l'} = \min_{S \subseteq A \setminus \{l'\}} \sum_{B \subseteq S} k'^{(i)}_{B \cup \{l'\}}$ and $k'^{(i+1)}_B = k'^{(i)}_B + k'^{(i)}_{B \cup \{l'\}}$. Coefficients $k''_{l'}$ are always non-negative for all $C(L)$ that are monotonically increasing with respect to L. By applying this decomposition iteratively for any ordering of labels $l' \in A$ we obtain:

$$K(A, \mathbf{t}) \le K + \sum_{l \in A} k''_l \delta_l(\mathbf{t}). \tag{51}$$

The constant term K can be ignored, as it does not affect the location of the optimal move. Heuristically, we pick l' in each iteration as:

$$l' = \arg\min_{l \in A} \min_{S \subseteq A \setminus \{l\}} \sum_{B \subseteq S} k'^{(i)}_{B \cup \{l\}}. \tag{52}$$

The over-estimation is tight for current solution corresponding to $\mathbf{t} = \mathbf{1}$.

In many practical cases the co-occurrence costs is defined as the sum of positive costs of subsets of L, for example all pairs of labels, as:

$$C(L) = \sum_{B \subseteq L} k_B, \text{ s.t. } k_B \ge 0. \tag{53}$$

In the case that k'_B stay non-negative for all $B \in L$, the over-estimation can be done as:

$$E_B(\mathbf{t}) = k'_B \prod_{l \in B} \delta_l(\mathbf{t}) \le k'_B \sum_{l \in B} \rho^B_l \delta_l(\mathbf{t}), \tag{54}$$

where $\rho^B_l \ge 0$ and $\sum_{l \in B} \rho^B_l = 1$. In practice, to obtain a symmetrical over-estimation of energy, we set $\rho^B_l = 1/|B|$. The moves for the first order occurrence costs (Delong et al. 2010) are exact. For second order co-occurrence between labels currently present in the image, the moves removing one of the labels of each pair are over-estimated by a factor of 2. This gives us an intuition why our approximation is appropriate and, in practice, the solution often contains the same label set as in the globally optimal solution (see Sect. 4).

Lemma 2 *For all $C(L)$ monotonically increasing with respect to L the over-estimated move energy can be represented as a binary pairwise graph with $|A|$ auxiliary variables $\mathbf{z}$ as:*

$$\begin{aligned} E'(\mathbf{t}) = \min_{\mathbf{z}} \Bigg[& k'_\alpha (1 - z_\alpha) + \sum_{l \in A} k''_l z_l + \sum_{i \in \mathcal{V}} k'_\alpha (1 - t_i) z_\alpha \\ & + \sum_{l \in A} \sum_{i \in \mathcal{V}_l} k''_l t_i (1 - z_l) \Bigg], \end{aligned} \tag{55}$$

where $\mathcal{V}_l$ is the set of pixels currently taking label l.

Proof See appendix. This binary function is pairwise submodular and thus can be solved efficiently using graph cuts. □

For co-occurrence potentials monotonically increasing with respect to $L(\mathbf{x})$ the problem can be modelled using one binary variable z_l per class indicating the presence of pixels of that class in the labelling, infinite edges for $x_i = l$

and $z_l = 0$ and hyper-graph over all z_l modelling $C(L(\mathbf{x}))$. The derived α-expansion construction can be seen as a graph taking into account costs over all auxiliary variables z_l for each move and over-estimating the hyper-graph energy using unary potentials. Consequentially, the only effect our approximation can have on the final labelling is to over estimate the number of classes present in an image. In practice the solutions found by expansion were generally local optima of the exact swap moves.

Similarly to $\alpha\beta$-swap moves there exists a slightly simpler solution Delong et al. (2010) for the optimisation of binary over-estimated move energy (46). The problem can be solved without the part of move energy $k'_\alpha \delta_\alpha(\mathbf{t})$ corresponding to the cost taken, if label α is introduced to an image after the move, and then the energy after the move is compared the the original energy and the move accepted if the energy has decreased. The proof of equivalence of this approach is similar to the one in Delong et al. (2010).

4 Experiments

We performed a controlled test evaluating the performance of CRF models both with and without co-occurrence potentials. As a base line we used the segment-based CRF and the associative hierarchical random field (AHRF) model proposed in (Ladicky et al. 2009) and the inference method (Russell et al. 2010), which currently offers state of the art performance on the MSRC data set (Shotton et al. 2006). On the VOC data set, the baseline also makes use of the detector potentials of (Ladicky et al. 2010) Figs. 2 and 3.

The costs $C(L)$ for the MSRC data set were created from the training set as follows: let M be the number of images, $\mathbf{x}^{(m)}$ the ground truth labelling of an image m and

$$z_l^{(m)} = \delta(l \in L(\mathbf{x}^{(m)})) \tag{56}$$

an indicator function for label l appearing in an image m. The associated cost was trained as:

$$C(L) = -w \log \frac{1}{M}\left(1 + \sum_{m=1}^{M} \prod_{l \in L} z_l^{(m)}\right), \tag{57}$$

where w is the weight of the co-occurrence potential. The form guarantees, that $C(L)$ is monotonically increasing with respect to L. To avoid over-fitting we approximated the potential $C(L)$ as a second order function:

$$C'(L) = \sum_{l \in L} c_l + \sum_{k,l \in L, k<l} c_{kl}, \tag{58}$$

where c_l and c_{lk} minimise the mean-squared error between $C(L)$ and $C'(L)$ up to the degree $|L| \leq 3$, such that $\forall l, k$: $c_l \geq 0, c_l + c_{lk} \geq 0$.

On the MSRC data set we observed a 3 % overall and 4 % average per class increase in the recall and 6 % in the intersection versus union measure with the of the segment-based CRF and a 1 % overall, 2 % average per class and 2 % in the intersection versus union measure with the AHRF.

On the VOC data set, due to the fact that the data set is unbalanced (all images contain the class background, and 22 % contain the class person, while only 2.8 % contain the class train) and a different performance criterium, the cost $C(L)$ was learnt as a sum of costs for each pair of classes, if they appeared together in the solution as:

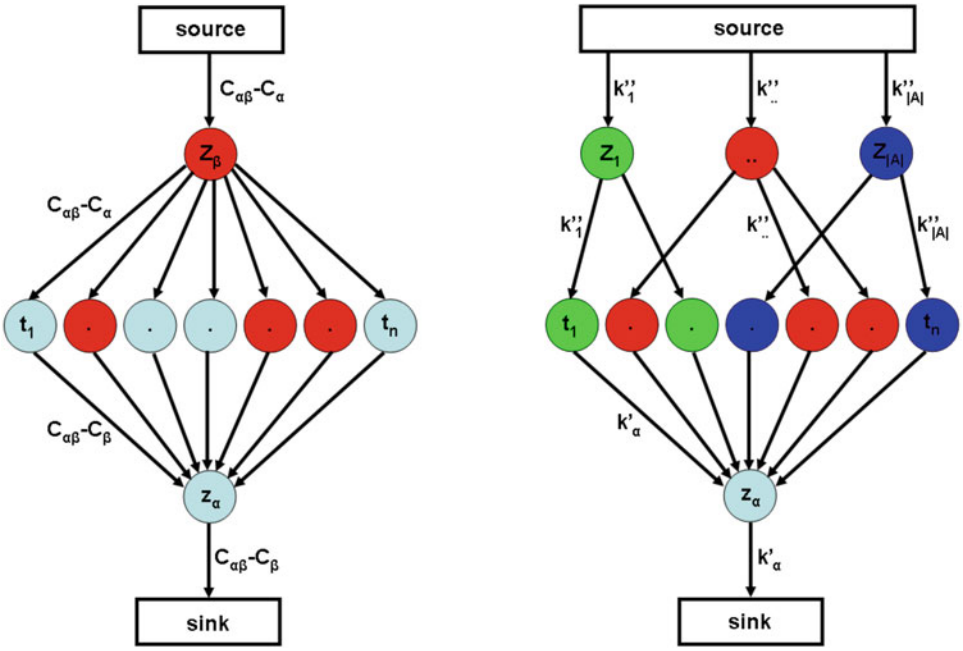

Fig. 2 Graph construction for $\alpha\beta$-swap and α-expansion move. In $\alpha\beta$-swap variable x_i will take the label α if corresponding t_i are tied to the sink after the st-mincut and β otherwise. In α-expansion variable x_i changes the label to α if it is tied to the sink after the st-mincut and remains the same otherwise. *Colours* represent the label of the variables before the move

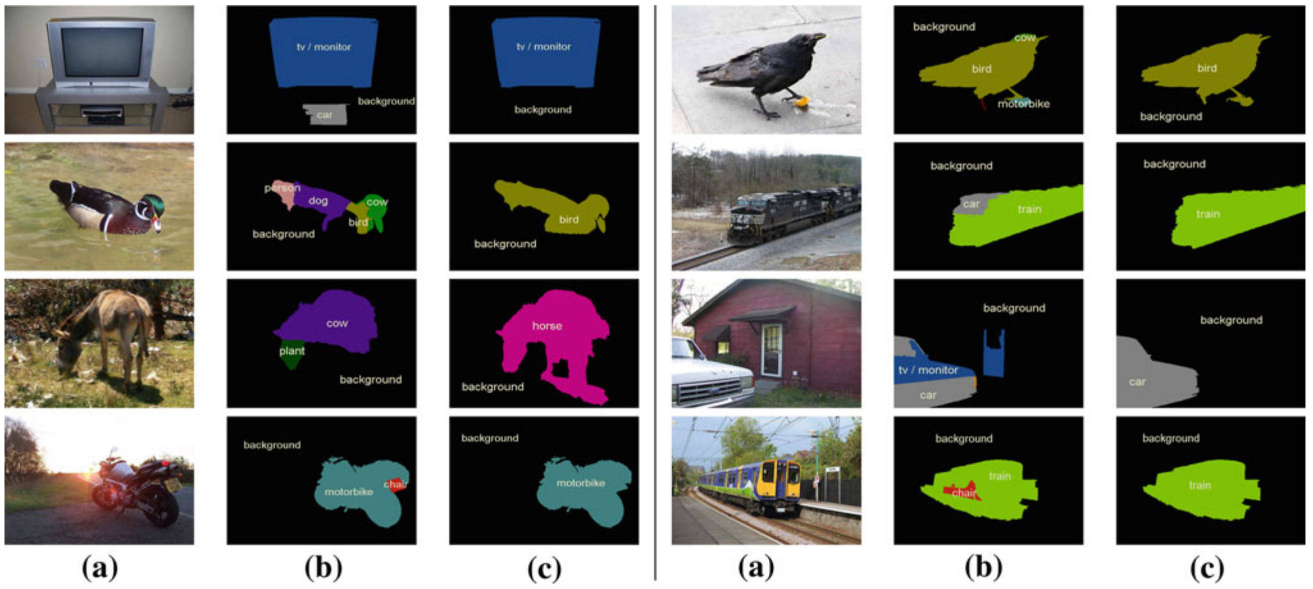

Fig. 3 Best viewed in colour. **a** Typical images taken from the VOC-2009 data set (Shotton et al. 2006). **b** A labelling based upon a pixel based random field model (Ladicky et al. 2009) that does not take into account co-occurrence. **c** A labelling of the same model using co-occurrence statistics. Note that the co-occurrence potentials perform in a similar way across different data sets, suppressing the smaller classes (see also Fig. 1) if they appear together in an uncommon combination with other classes such as a car with a monitor, a train with a chair or a dog with a bird. This results in a qualitative rather than quantitative difference

$$C(L) = -w \sum_{k<l\in\mathcal{L}} c_{kl}, \qquad (59)$$

where c_{kl} were learnt as:

$$\begin{aligned} c_{kl} &= \min(-\log(\mathbb{P}(k|l) \vee \mathbb{P}(l|k)), T) \\ &= \min(-\log(\mathbb{P}(k|l) + \mathbb{P}(l|k) - \mathbb{P}(k|l)\mathbb{P}(l|k)), T), \qquad (60) \end{aligned}$$

$\mathbb{P}(k|l) = \frac{\mathbb{P}(\{k,l\})}{\mathbb{P}(\{l\})}$, $\mathbb{P}(L) = \frac{\sum_{m=1}^{M} \prod_{l\in L} z_l^{(m)}}{M}$ and T is the threshold for the maximum cost.

This heuristically motivated cost ensures that if one class only occurs when another is present, as for example, cow only occurs when grass is present in the image, then the second order co-occurence cost between these classes will be 0. Furthermore it allows simpler LP (26) and pairwise (35) formulations. The comparison on the VOC2009 data set was performed on the validation set, as the test set is not published and the number of permitted submissions is limited. Performance improved by 3.5 % in the intersection versus union measure used in the challenge. We also report the performance on the test set, which is comparable with current state-of-the-art methods. Results for both data sets are given in Tables 2 and 3.

By adding a co-occurrence cost to the CRF we observe constant improvement in pixel classification for almost all classes in all measures. In accordance with desiderata (iv), the co-occurrence potentials tend to suppress uncommon combination of classes and produce more coherent images in the labels space. This results in a qualitative rather than quantitative difference. Although the unary potentials already capture textural context (Shotton et al. 2006), the incorporation of co-occurrence potentials leads to a significant improvement in accuracy.

It is not computationally feasible to perform a direct comparison between the work (Rabinovich et al. 2007) and our potentials, as the AHRF model is defined over individual pixels, and it is not possible to minimise the resulting fully connected graph which would contain approximately 4×10^{10} edges. Similarly, without their scene classification potentials it was not possible to do a like for like comparison with (Torralba et al. 2003).

Average running time on the MSRC data set without co-occurrence was 5.1 s in comparison to 16.1 s with co-occurrence cost. On the VOC2009 data set the average times were 107 s and 388 s for inference without and with co-occurrence costs. We compared the performance of α-expansion with BP and LP relaxation using solver of Benson and Shanno (2007) for general co-occurrence potential on the VOC images sub-sampled to 20 × 20 boxes. All methods produced exactly the same results in terms of energy for every image, however α-expansion took on average 43 ms, BP 0.8 s and LP relaxation 1,200 s. The solution found was the global solution for every image, because the LP found an integer solution. The main reason could be that the sub-sampling simplified the problem making the unary

Table 2 Quantitative results on the MSRC data set, average per class recall measure, defined as $\frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$

	Global	Average	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
Segment CRF	77	64	70	95	78	55	76	95	63	81	76	67	72	73	82	**35**	72	17	88	29	62	45	17
Segment CRF with CO	80	68	77	**96**	80	69	82	98	69	82	79	75	75	81	85	**35**	76	17	89	25	61	50	**22**
Hierarchical CRF	86	75	81	**96**	87	72	84	**100**	77	**92**	86	**87**	87	95	95	27	**85**	33	**93**	43	**80**	62	17
Hierarchical CRF with CO	**87**	**77**	**82**	95	**88**	**73**	**88**	**100**	**83**	**92**	**88**	**87**	**88**	**96**	**96**	27	**85**	**37**	**93**	**49**	**80**	**65**	20

Incorporation of co-occurrence potentials led to a constant improvement for almost every class
Bold values indicate best result for each class

Table 3 Quantitative analysis of VOC2009 results on validation set, intersection vs. union measure, defined as $\frac{\text{True positive}}{\text{True positive} + \text{False negative} + \text{False positive}}$

	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
CRF w/o CO (val)	27.3	77.7	38.3	9.6	**24.0**	35.8	**31.0**	59.2	36.5	21.2	8.3	**1.7**	22.7	**14.3**	17.0	26.7	21.1	15.5	**16.3**	14.6	48.5	33.1
CRF with CO (val)	**30.8**	**82.3**	**49.3**	**11.8**	19.3	**37.7**	30.8	**63.2**	**46.0**	**23.7**	**10.0**	0.5	**23.1**	14.1	**22.4**	**33.9**	**35.7**	**18.4**	12.1	**22.5**	**53.1**	**37.5**
CRF with CO (test)	32.1	81.2	46.1	15.4	24.6	20.9	36.9	50.0	43.9	28.4	11.5	18.2	25.4	14.7	25.1	37.7	34.1	27.7	29.6	18.4	43.8	40.8

Incorporation of co-occurrence potential led to labellings, which visually look more coherent, but are not necessarily correct. Quantitatively the performance improved significantly, on average by 3.5 % per class. For comparison we also report the performance on the test set
Bold values indicate best result for each class

potentials relatively stronger to pairwise potentials. Comparison on larger images was not feasible due to the large memory consumption of LP solver. Message passing BP could have been applied also to the images of the standard size, however it was approximately 20 times slower and converged to the worse solution than α-expansion in terms of energy for every tested image.

5 Conclusion

The importance of co-occurrence statistics is well established (Torralba et al. 2003; Rabinovich et al. 2007; Csurka and Perronnin 2008). In this work we examined the use of co-occurrence statistics and how they can be efficiently incorporated into a global energy or probabilistic model such as a conditional random field. We have shown how they can naturally be encoded by the use of higher order cliques, without a significant computational overhead. Whilst the performance improvements on current data sets are slight, we believe encoding co-occurrence will become increasingly important in the future when, rather than attempting to classify 20 classes in an image we have to classify 20,000 (Sturgess et al. 2012). Even with a false positive rate of 1 % this would still give 200 false positives per image. Co-occurrence information gives a natural way to tackle this problem.

Acknowledgments This study was 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.

Appendix

Lemma 1 Proof First we show that:

$$\begin{aligned} E_\alpha(\mathbf{t}) &= \min_{z_\alpha}[(C_{\alpha\beta} - C_\beta)(1 - z_\alpha) \\ &\quad + \sum_{i\in\mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i)z_\alpha] \\ &= \begin{cases} 0 & \text{if } \forall i \in \mathcal{V}_{\alpha\beta} : t_i = 1, \\ C_{\alpha\beta} - C_\beta & \text{otherwise.} \end{cases} \end{aligned} \tag{61}$$

If $\forall i \in \mathcal{V}_{\alpha\beta} : t_i = 1$ then $\sum_{i\in\mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i)z_\alpha = 0$ and the minimum cost cost 0 occurs when $z_\alpha = 1$. If $\exists i \in \mathcal{V}_{\alpha\beta}$, $t_i = 0$ the minimum cost labelling occurs when $z_\alpha = 0$ and the minimum cost is $C_{\alpha\beta} - C_\beta$. Similarly:

$$\begin{aligned} E_\beta(\mathbf{t}) &= \min_{z_\beta}[(C_{\alpha\beta} - C_\alpha)z_\beta \\ &\quad + \sum_{i\in\mathcal{V}_{\alpha\beta}} (C_{\alpha,\beta} - C_\alpha)t_i(1 - z_\beta)] \\ &= \begin{cases} 0 & \text{if } \forall i \in \mathcal{V}_{\alpha\beta} : t_i = 0, \\ C_{\alpha\beta} - C_\alpha & \text{otherwise.} \end{cases} \end{aligned} \tag{62}$$

By inspection, if $\forall i \in \mathcal{V}_{\alpha\beta} : t_i = 0$ then $\sum_{i\in\mathcal{V}_{\alpha\beta}}(C_{\alpha,\beta} - C_\alpha)t_i(1 - z_\beta) = 0$ and the minimum cost cost 0 occurs when $z_\beta = 0$. If $\exists i \in \mathcal{V}_{\alpha\beta}, t_i = 1$ the minimum cost labelling occurs when $z_\beta = 1$ and the minimum cost is $C_{\alpha\beta} - C_\alpha$.

For all three cases (all pixels take label α, all pixels take label β and mixed labelling) $E(\mathbf{t}) = E_\alpha(\mathbf{t}) + E_\beta(\mathbf{t}) + C_\alpha + C_\beta - C_{\alpha\beta}$. The construction of the $\alpha\beta$-swap move is similar to the Robust P^N model (Kohli et al. 2008). □

See Figs. 2 and 3 for graph construction.

Lemma 2 Proof Similarly to the $\alpha\beta$-swap proof we can show:

$$\begin{aligned} E_\alpha(\mathbf{t}) &= \min_{z_\alpha}\left[k'_\alpha(1 - z_\alpha) + \sum_{i\in\mathcal{V}} k'_\alpha(1 - t_i)z_\alpha\right] \\ &= \begin{cases} k'_\alpha & \text{if } \exists i \in \mathcal{V} \text{ s.t. } t_i = 0, \\ 0 & \text{otherwise .} \end{cases} \end{aligned} \tag{63}$$

If $\exists i \in \mathcal{V} s.t.t_i = 0$, then $\sum_{i\in\mathcal{V}} k'_\alpha(1 - t_i) \geq k'_\alpha$, the minimum is reached when $z_\alpha = 0$ and the cost is k'_α.

If $\forall i \in \mathcal{V} : t_i = 1$ then $k'_\alpha(1 - t_i)z_\alpha = 0$, the minimum is reached when $z_\alpha = 1$ and the cost becomes 0.

For all other $l \in A$:

$$\begin{aligned} E_b(\mathbf{t}) &= \min_{z_l}\left[k''_l z_l + \sum_{i\in\mathcal{V}_l} k''_l t_i(1 - z_l)\right] \\ &= \begin{cases} k''_l & \text{if } \exists i \in \mathcal{V}_l \text{ s.t. } t_i = 1, \\ 0 & \text{otherwise .} \end{cases} \end{aligned} \tag{64}$$

If $\exists i \in \mathcal{V}_l$ s.t. $t_i = 1$, then $\sum_{i\in\mathcal{V}_l} k''_l t_i \geq k''_l$, the minimum is reached when $z_l = 1$ and the cost is k''_l.

If $\forall i \in \mathcal{V}_l : t_i = 0$ then $\sum_{i\in\mathcal{V}_l} k''_l t_i(1 - z_l) = 0$, the minimum is reached when $z_l = 1$ and the cost becomes 0.

By summing up the cost $E_\alpha(\mathbf{t})$ and $|A|$ costs $E_l(\mathbf{t})$ we get $E'(\mathbf{t}) = E_\alpha(\mathbf{t}) + \sum_{l\in A} E_l(\mathbf{t})$. If α is already present in the image $k'_\alpha = 0$ and edges with this weight and variable z_α can be ignored. □

See Figs. 2 and 3 for graph construction.

References

Benson, H. Y., & Shanno, D. F. (2007). An exact primal—dual penalty method approach to warmstarting interior-point methods for linear programming. *Computational Optimization and Applications, 38*(3), 371–399.

Borenstein, E., & Malik, J. (2006). Shape guided object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 969–976) New York.

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

Choi, M. J., Lim, J. J., Torralba, A., & Willsky, A. S. (2010). Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco.

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

Csurka, G., & Perronnin, F. (2008). A simple high performance approach to semantic segmentation. In *British Machine Vision Conference08*, Leeds.

Delong, A., Osokin, A., Isack, H., & Boykov, Y. (2010). Fast approximate energy minimization with label costs. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco.

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

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

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

Heitz, G., & Koller, D. (2008). Learning spatial context: Using stuff to find things. In *European Conference on Computer Vision*, Marseille.

Hoiem, D., Rother, C., & Winn, J. M. (2007). 3d layoutcrf for multi-view object class recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego.

Kleinberg, J., & Tardos, E. (2002). Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM, 49*(5), 616–639.

Kohli, P., Ladicky, L., & Torr, P. H. S. (2008). Robust higher order potentials for enforcing label consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage.

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(10), 1568–1583.

Kolmogorov, V., & Rother, C. (2006). Comparison of energy minimization algorithms for highly connected graphs. In Proceedings of European Conference on Computer Vision (pp. 1–15). Heidelberg: Springer.

Komodakis, N., Tziritas, G., & Paragios, N. (2007). Fast, approximately optimal solutions for single and dynamic mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN.

Kumar, M., & Torr, P. H. S. (2008). Efficiently solving convex relaxations for map estimation. In *International Conference on Machine Learning*. New York: ACM.

Kumar, M. P., Veksler, O., & Torr, P. H. S. (2011). Improved moves for truncated convex models. *Journal of Machine Learning Research,12*, 31–67.

Ladicky, L., Russell, C., Kohli, P., & Torr, P. H. S. (2009). Associative hierarchical crfs for object class image segmentation. In *International Conference on Computer Vision*.

Ladicky, L., Russell, C., Sturgess, P., Alahari, K., & Torr, P. H. S. (2010). What, where and how many? combining object detectors and crfs. *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*.

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

Narasimhan, M., & Bilmes, J. A. (2005). A submodular-supermodular procedure with applications to discriminative structure learning. In *Uncertainty in Artificial Intelligence* (pp. 404–412).

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *International Conference on Computer Vision*, Rio de Janeiro.

Ren, X., Fowlkes, C., & Malik, J. (2005). Mid-level cues improve boundary detection. Tech. Rep. UCB/CSD-05-1382, EECS Department, University of California, Berkeley.

Rother, C., Kumar, S., Kolmogorov, V., & Blake, A. (2005). Digital tapestry. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 589–596).

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

Russell, C., Ladicky, L., Kohli, P., & Torr, P. H. S. (2010). Exact and approximate inference in associative hierarchical networks using graph cuts. *Uncertainty in Artificial Intelligence*, Catalina Island, CA.

Schlesinger, M. (1976). Syntactic analysis of two-dimensional visual signals in noisy conditions. *Kibernetika, 4*, 113–130. (in Russian).

Schölkopf, B., & Smola, A. J. (2001). Learning with kernels: support vector machines, regularization, optimization, and beyond. *Adoptive Computation & Machine Learning*. Cambridge, MA: MIT Press.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell, 22*(8), 888–905.

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* (Vol. 1, pp 1–15).

Sturgess, P., Ladicky, L., Crook, N., & Torr, P. H. S. (2012). Scalable cascade inference for semantic image segmentation. In *British Machine Vision Conference*.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., et al. (2006). A comparative study of energy minimization methods for markov random fields. In *European Conference on Computer Vision*.

Torr, P. H. S. (1998). Geometric motion segmentation and model selection [and discussion]. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences,356*(1740), 1321–1340.

Torralba, A., Murphy, K. P., Freeman, W. T., & Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *Proceedings of the Nineth IEEE International Conference on Computer Vision*.

Toyoda, T., & Hasegawa, O. (2008). Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(8), 1483–1489.

Wainwright, M., Jaakkola, T., & Willsky, A. (2002). *Map estimation via agreement on (hyper)trees: Messagepassing and linear programming approaches*. Cambridge, MA: MIT Press.

Wainwright, M., Jaakkola, T., & Willsky, A. (2005). Map estimation via agreement on trees: Message-passing and linear programming. *IEEE Transactions on Information Theory* (pp. 3697–3717).

Weiss, Y., & Freeman, W. (2001). On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory, 47*(2), 723–735.

Werner, T. (2007). A linear programming approach to max-sum problem: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(7), 1165–1179.

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