Graph Cut based Inference with Co-occurrence Statistics

Lubor Ladicky[1,3], Chris Russell[1,3], Pushmeet Kohli[2], and Philip H.S. Torr[1]

[1] Oxford Brookes
[2] Microsoft Research

Abstract. Markov and Conditional random fields (CRFs) used in computer vision typically model only local interactions between variables, as this 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 an improvement in the labelling compared to just using a pairwise model.

1 Introduction

Class based image segmentation is a highly active area of computer vision research as is shown by a spate of recent publications [11,22,29,31,34]. 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 as a likelihood, 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 [29], that govern the labelling of single pixels; mid-level cues such as region continuity, symmetry [23] or shape [2] 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.

[3] The authors assert equal contribution and joint first authorship
This work was supported by EPSRC, HMGCC and the PASCAL2 Network of Excellence. Professor Torr is in receipt of a Royal Society Wolfson Research Merit Award.

Current state of the art low-level approaches typically follow the methodology proposed in *Texton-boost* [29], 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 label. These costs are combined in a contrast sensitive Conditional Random Field CRF [19].

The majority of mid-level inference schemes [25,20] do not consider pixels directly, rather they assume that the image has been segmented into super-pixels [5,8,28]. 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 [10] proposed a method of reshaping super-pixels to recover from the errors, while the work [17] proposed a novel hierarchical 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 sheep 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 cow as half cow and half sheep, or the mistaken labelling of a boat surrounded by water as a book.

As well as penalising strange combinations of objects appearing in an image, co-occurrence potentials can also be used to impose an MDL[1] prior that encourages a parsimonious description of an image using fewer labels. As discussed eloquently in the recent work [4], 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. In [22] Rabinovich *et al.* proposed the integration of such co-occurrence costs that characterise the relationship between two classes. Similarly Torralba *et al.* [31] 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

[1] Minimum description length

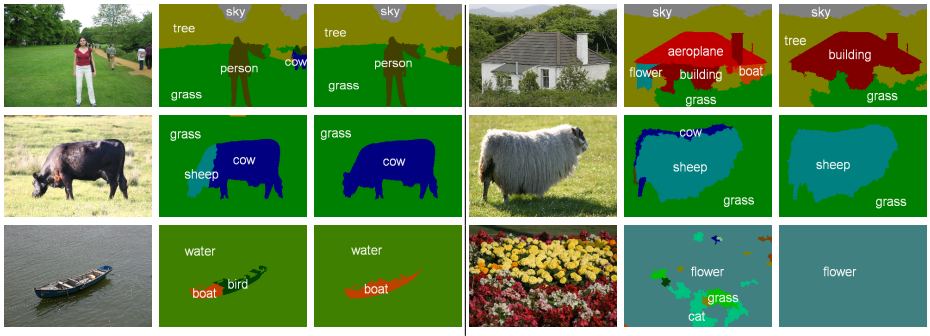

Fig. 1. **Best viewed in colour:** *Qualitative results of object co-occurrence statistics.* **(a)** *Typical images taken from the MSRC data set [29];* **(b)** *A labelling based upon a pixel based random field model [17] that does not take into account co-occurrence;* **(c)** *A labelling of the same 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.*

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:

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

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 [29] or kernel SVMs [27]), 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 [31].

(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 [32], 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 [30] 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:

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

Embedding the co-occurrence term in 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 $C(L)$ 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 uses 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.* [31] 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 [6], 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)*.

Method	Global energy *(i)*	Invariance *(ii)*	Efficiency *(iii)*	Parsimony *(iv)*
Unary [31]	✗	✗	✓	✗
Pairwise [22,9,32]	✓	✗	✗	✓
Csurka [6]	✗	—	✓	—
Our approach	✓	✓	✓	✓

Fig. 2. *A comparison of the capabilities of existing image co-occurrence formulations against our new approach. See section 2.2 for details.*

Rabinovich *et al.* [9,22], and independently [32], 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:

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

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 [3], 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 figure 2.

Two efficient schemes [7,12] 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 do satisfy our desiderata.

One such approach was proposed by Hoiem *et al.* [12], who used 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.

In a recent work, appearing at the same time as ours, Delong *et al.* [7] proposed the use of a soft cost over the number of labels present in an image for clustering. While the mathematical formulation they propose is more flexible than this, they do not suggest any applications of this increased flexibility. Moreover, their formulation is less general than ours as it does not support the full range of monotonically increasing label set costs.

3 Inference on global co-occurrence potentials

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

$$\begin{aligned} \mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{L}^{|\mathcal{V}|}} \textstyle\sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c) + C(L(\mathbf{x})) \\ \text{s.t. } \mathbf{x} \in \mathcal{L}^{|\mathcal{V}|},\ L(\mathbf{x}) = \{l \in \mathcal{L} : \exists x_i = l\}. \end{aligned} \tag{8}$$

In the general case the problem of minimising this energy can be reformulated as an integer program and approximately solved as an LP-relaxation [16]. This LP-formulation can be transformed using a Lagrangian relaxation into a pairwise energy, allowing algorithms, such as Belief Propagation [33] or TRW-S [14], that can minimise arbitrary pairwise energies to be applied [16]. However, reparameterisation methods such as these perform badly on densely connected graphs [15,26].

In this section we show that under assumption, that $C(L)$ is monotonically increasing with respect to L, the problem can be solved efficiently using $\alpha\beta$-swap and α-expansion moves [3], where the number of additional edges of the graph grows linearly with the number of variables in the graph. In contrast to [22], these algorithms can be applied to large graphs with more than $200,000$ variables.

Move making algorithms project the problem into a smaller subspace in which a sub-problem is efficiently solvable. 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. Because the energy (4) is additive, we derive graph constructions only for term $C(L(\mathbf{x}))$. Both the application of swap and expansion moves to minimise the energy, and the graph construction for the other terms proceed as described in [3].

3.1 $\alpha\beta$-Swap Moves

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 $\mathbf{x}$ which has been 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:

$$T_{\alpha\beta}(x_i, t_i) = \begin{cases} \alpha \text{ if } x_i = \alpha \text{ or } \beta \text{ and } t_i = 0, \\ \beta \text{ if } x_i = \alpha \text{ or } \beta \text{ and } t_i = 1. \end{cases} \tag{9}$$

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:

$$S = \{L(\mathbf{x}) \cup \{\alpha\} \setminus \{\beta\}, L(\mathbf{x}) \cup \{\beta\} \setminus \{\alpha\}, L(\mathbf{x}) \cup \{\alpha, \beta\}\}. \tag{10}$$

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:

$$E(\mathbf{t}) = \begin{cases} C_\alpha = C(L(\mathbf{x}) \cup \{\alpha\} \setminus \{\beta\}) \text{ if } \forall i \in \mathcal{V}_{\alpha\beta}\,, t_i = 0, \\ C_\beta = C(L(\mathbf{x}) \cup \{\beta\} \setminus \{\alpha\}) \text{ if } \forall i \in \mathcal{V}_{\alpha\beta}\,, t_i = 1, \\ C_{\alpha\beta} = C(L(\mathbf{x}) \cup \{\alpha, \beta\}) \qquad \text{otherwise.} \end{cases} \tag{11}$$

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

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:*

$$\begin{aligned} E(\mathbf{t}) = C_\alpha + C_\beta - C_{\alpha\beta} + \min_{z_\alpha, z_\beta} \Big[& (C_{\alpha\beta} - C_\alpha) z_\beta + (C_{\alpha\beta} - C_\beta)(1 - z_\alpha) \\ & + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha,\beta} - C_\alpha) t_i (1 - z_\beta) + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i) z_\alpha) \Big]. \end{aligned} \tag{12}$$

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

3.2 α-Expansion Moves

In an α-expansion move every random variable can 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:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha \text{ if } t_i = 0 \\ x_i \text{ if } t_i = 1. \end{cases} \tag{13}$$

To derive a graph-construction that approximates the true cost of an α-expansion move we rewrite $C(L)$ as:

$$C(L) = \sum_{B \subseteq L} k_B, \tag{14}$$

where the coefficients k_B are calculated recursively as:

$$k_B = C(B) - \sum_{B' \subset B} k_{B'}. \tag{15}$$

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:

$$\delta_\alpha(\mathbf{t}) = \begin{cases} 1 & \text{if} \quad \exists i \in \mathcal{V} \text{ s.t. } t_i = 0, \\ 0 \text{ otherwise.} \end{cases} \tag{16}$$

$$\forall l \in A \,,\, \delta_l(\mathbf{t}) = \begin{cases} 1 & \text{if} \quad \exists i \in \mathcal{V}_l \text{ s.t. } t_i = 1, \\ 0 \text{ otherwise.} \end{cases} \tag{17}$$

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

$$E(\mathbf{t}) = E_{new}(\mathbf{t}) - E_{old} = \sum_{B \subseteq A \cup \{\alpha\}} k_B \prod_{l \in B} \delta_l(\mathbf{t}) - C(A).$$

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

$$E(\mathbf{t}) = \sum_{B \subseteq A} k_B \prod_{l \in B} \delta_l(\mathbf{t}) + \sum_{B \subseteq A} k_{B \cup \{\alpha\}} \delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t}). \tag{18}$$

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

$$\delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t}) = \delta_\alpha(\mathbf{t}) + \prod_{l \in B} \delta_l(\mathbf{t}) - 1. \tag{19}$$

Replacing the term $\delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t})$ and disregarding new constant terms, equation (18) becomes:

$$E(\mathbf{t}) = \sum_{B \subseteq A} k_{B \cup \{\alpha\}} \delta_\alpha(\mathbf{t}) + \sum_{B \subseteq A} (k_B + k_{B \cup \{\alpha\}}) \prod_{l \in B} \delta_l(\mathbf{t}) = k'_\alpha \delta_\alpha(\mathbf{t}) + \sum_{B \subseteq A} k'_B \prod_{l \in B} \delta_l(\mathbf{t}), \tag{20}$$

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 [21,24] and overestimate second term $K(A,\mathbf{t}) = \sum_{B\subseteq A} k'_B \prod_{l\in B} \delta_l(\mathbf{t})$ of the cost of moving from the current solution.

For any $l' \in A$ we can overestimate $K(A,\mathbf{t})$ by

$$\begin{aligned} K(A,\mathbf{t}) &\leq K(A\setminus\{l'\},\mathbf{t}) + \delta_{l'}(\mathbf{t}) \min_{S\subseteq A\setminus\{l'\}} \sum_{B\subseteq S} (k'_{B\cup\{l'\}} - k'_B) \\ &= K(A\setminus\{l'\},\mathbf{t}) + k''_{l'}\delta_{l'}(\mathbf{t}), \end{aligned} \tag{21}$$

where $k''(l')$ is 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 :

$$K(A,\mathbf{t}) \leq K + \sum_{l\in A} k''_l \delta_l(\mathbf{t}). \tag{22}$$

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

$$l' = \arg\min_{l\in A} \min_{S\subseteq A\setminus\{l\}} \sum_{B\subseteq S} (k'_{B\cup\{l\}} - k'_B). \tag{23}$$

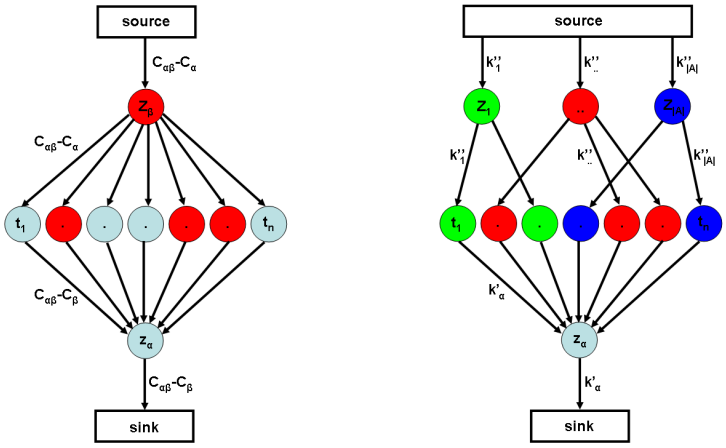

Fig. 3. *Graph construction for $\alpha\beta$-swap and α-expansion move. In $\alpha\beta$-swap variable x_i will take the label α if corresponding t_i are tied to the sink after the st-mincut and β otherwise. In α-expansion variable x_i changes the label to α if it is tied to the sink after the st-mincut and remains the same otherwise. Colours represent the label of the variables before the move.*

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

$$E'(\mathbf{t}) = \min_{\mathbf{z}} \left[k'_\alpha(1-z_\alpha) + \sum_{l\in A} k''_l z_l + \sum_{i\in\mathcal{V}} k'_\alpha(1-t_i)z_\alpha + \sum_{l\in A}\sum_{i\in\mathcal{V}_l} k''_l t_i(1-z_l) \right], \tag{24}$$

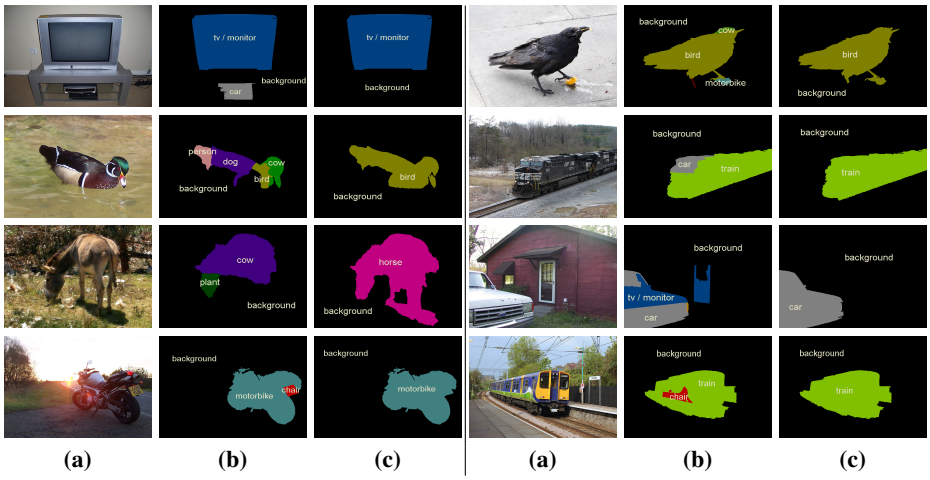

Fig. 4. **Best viewed in colour:** **(a)** *Typical images taken from the* VOC*-2009 data set [29];* **(b)** *A labelling based upon a pixel based random field model [17] that does not take into account co-occurrence;* **(c)** *A labelling of the same model using co-occurrence statistics. Note that the co-occurrence potentials perform in a similar way across different data sets, suppressing the smaller classes (see also figure 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.*

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. Note that the energy approximation is exact, unless existing classes are removed from the labelling. 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.

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 [17] and the inference method [26], which currently offers state of the art performance on the MSRC data set [29]. On the VOC data set, the baseline also makes use of the detector potentials of [18]. The costs $C(L)$ 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

$$z_l^{(m)} = \delta(l \in L(\mathbf{x}^{(m)})) \tag{25}$$

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

$$C(L) = -w \log \frac{1}{M} \left(1 + \sum_{m=1}^{M} \prod_{l \in L} z_l^{(m)} \right), \tag{26}$$

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:

$$C'(L) = \sum_{l \in L} c_l + \sum_{k,l \in L, k<l} c_{kl}, \tag{27}$$

where c_l and c_{lk} minimise the mean-squared error between $C(L)$ and $C'(L)$.

On the MSRC data set we observed a 3% overall and 4% average per class increase in the recall and 6% in the intersection vs. union measure with the of the segment-based CRF and a 1% overall, 2% average per class and 2% in the intersection vs. union measure with the AHRF. 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 vs. union measure used in the challenge. The performance on the test set was 32.11% which is comparable with current state-of-the-art methods. Results for both data sets are given in tables 5 and 6.

By adding a co-occurrence cost into 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 [29], 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 [22] 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 [31].

Average running time on the MSRC data set without co-occurrence was $5.1s$ in comparison to $16.1s$ with co-occurrence cost. On the VOC2009 data set the average times were $107s$ and $388s$ for inference without respectively with co-occurrence costs. We compared the performance of α-expansion with LP relaxation using solver given in [1] for general co-occurrence potential on the sub-sampled images [16]. Both methods produced similar results in terms of energy, however α-expansion was approximately $42,000$ times faster.

5 Conclusion

The importance of co-occurrence statistics has been well established [31,22,6]. In this work we have examined the use of co-occurrence statistics and how they might be incorporated into a global energy or likelihood model such as a conditional random field. We

Fig. 5. *Quantitative results on the MSRC data set, average per class recall measure, defined as* $\frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}}$. *Incorporation of co-occurrence potentials led to a constant improvement for almost every class.*

	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

Fig. 6. *Quantitative analysis of VOC2009 results on validation set, intersection vs. union measure, defined as* $\frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative} + \textit{False Positive}}$. *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.*

	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
Hierarchical CRF	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
Hierarchical CRF with CO	**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**

have discovered that they can naturally be encoded by the use of higher order cliques, without a significant computational overhead. Our new framework provides significant advantages over state of the art approaches including efficient scalable inference. We performed a controlled test evaluating the performance of CRF models both with and without co-occurrence potentials and the incorporation of these potentials results in quantitatively better and visually more coherent labellings.

References

1. H. Y. Benson and D. F. Shanno. An exact primal—dual penalty method approach to warm-starting interior-point methods for linear programming. *Comput. Optim. Appl.*, 2007. 11
2. E. Borenstein and J. Malik. Shape guided object segmentation. In *CVPR*, 2006. 1
3. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001. 5, 6
4. M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2
5. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002. 2
6. G. Csurka and F. Perronnin. A simple high performance approach to semantic segmentation. In *BMVC08*, 2008. 5, 11
7. A. Delong, A. Osokin, H. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *CVPR*, 2010. 6
8. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 2
9. C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 5
10. S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 2
11. D. K. G. Heitz. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 1

12. D. Hoiem, C. Rother, and J. M. Winn. 3d layoutcrf for multi-view object class recognition and segmentation. In *CVPR*, 2007. 6
13. P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 14
14. V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 2006. 6
15. V. Kolmogorov and C. Rother. C.: Comparison of energy minimization algorithms for highly connected graphs. In *ECCV*, 2006. 6
16. L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph Cut based Inference with Co-occurrence Statistics — Technical report, 2010. 6, 11
17. L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 2, 3, 10
18. L. Ladicky, C. Russell, P. Sturgess, K. Alahari, and P. Torr. What, where and how many? Combining object detectors and CRFs. *ECCV*, 2010. 10
19. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001. 2
20. D. Larlus and F. Jurie. Combining appearance models and markov random fields for category level object segmentation. In *CVPR*, 2008. 2
21. M. Narasimhan and J. A. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *UAI*, 2005. 9
22. A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1, 2, 5, 6, 11
23. X. Ren, C. Fowlkes, and J. Malik. Mid-level cues improve boundary detection. Technical Report UCB/CSD-05-1382,Berkeley, Mar 2005. 1
24. C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry. In *CVPR*,2005. 9
25. B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2
26. C. Russell, L. Ladicky, P. Kohli, and P. Torr. Exact and approximate inference in associative hierarchical networks using graph cuts. *UAI*, 2010. 6, 10
27. B. Schlkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, 2001. 3
28. J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*,, 2000. 2
29. J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV (1)*, 2006. 1, 2, 3, 10, 11
30. R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields. In *ECCV*, 2006. 4
31. A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Computer Vision, Proceedings.*, 2003. 1, 2, 3, 4, 5, 11
32. T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *PAMI*, 2008. 4, 5
33. Y. Weiss and W. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *Transactions on Information Theory*, 2001. 6
34. L. Yang, P. Meer, and D. J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*, 2007. 1

Appendix

Lemma 1 Proof. First we show that:

$$E_\alpha(\mathbf{t}) = \min_{z_\alpha} \left[(C_{\alpha\beta} - C_\beta)(1 - z_\alpha) + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i) z_\alpha \right]$$
$$= \begin{cases} 0 & \text{if } \forall i \in \mathcal{V}_{\alpha\beta} : t_i = 1, \\ C_{\alpha\beta} - C_\beta & \text{otherwise.} \end{cases} \quad (28)$$

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:

$$E_\beta(\mathbf{t}) = \min_{z_\beta} \left[(C_{\alpha\beta} - C_\alpha) z_\beta + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha,\beta} - C_\alpha) t_i (1 - z_\beta) \right]$$
$$= \begin{cases} 0 & \text{if } \forall i \in \mathcal{V}_{\alpha\beta} : t_i = 0, \\ C_{\alpha\beta} - C_\alpha & \text{otherwise.} \end{cases} \quad (29)$$

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 [13]. □

See figure 3 for graph construction.

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

$$E_\alpha(\mathbf{t}) = \min_{z_\alpha} \left[k'_\alpha (1 - z_\alpha) + \sum_{i \in \mathcal{V}} k'_\alpha (1 - t_i) z_\alpha \right] = \begin{cases} k'_\alpha \text{ if } \exists i \in \mathcal{V} \text{ s.t. } t_i = 0, \\ 0 \qquad \text{otherwise.} \end{cases} \quad (30)$$

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 $l \in A$:

$$E_b(\mathbf{t}) = \min_{z_l} \left[k''_l z_l + \sum_{i \in \mathcal{V}_l} k''_l t_i (1 - z_l) \right] = \begin{cases} k''_l \text{ if } \exists i \in \mathcal{V}_l \text{ s.t. } t_i = 1, \\ 0 \qquad \text{otherwise.} \end{cases} \quad (31)$$

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 figure 3 for graph construction.