Global Structured Models towards Scene Understanding

Ľubor Ladický

Thesis submitted in partial fulfillment of the requirements of the award of

Doctor of Philosophy

Oxford Brookes University

2011

Abstract

Many scene understanding tasks are formulated as a labelling problem that tries to assign a label to each pixel of an image. These discrete labels may vary depending on the task, for example they may correspond to different object classes such as car, grass or sky, or to depths or to intensity after denoising. These labelling problems are typically formulated as a pairwise Markov or Conditional Random Field, modelling the dependencies of labels of pairs of variables in the local neighbourhoods. However, these pairwise models are very restricted in their expressivity. They can not model rich natural statistics and induce desired complex structures in the output labelling. In this thesis we propose global structured formulations beyond pairwise models, showing that they are very useful in computer vision, furthermore that they can still be learnt and optimised efficiently.

First we propose a model, which generalises existing approaches for semantic object class segmentation, formulated in terms of pixels, segments or groups of segments. The proposed method efficiently integrates the strengths of these different approaches, capturing discriminative information across different scales. Next we show how the standard approaches for the semantic object class segmentation problem can be improved by the inclusion of costs based on high level statistics, including object class co-occurrence, which capture knowledge of scene semantics, for example that motorbikes and cows are unlikely to occur together in an image. Then we propose a novel latent random field support vector machine for object detection with a convex MRF regularization and suggest a way to include this information in the object class segmentation formulation. Finally we propose a model that jointly estimates labellings of multiple domains over a product space of labels. We demonstrate the usefulness of this model on the problem of joint object class semantic segmentation and dense $3D$ stereo reconstruction and show that this approach significantly outperforms existing methods. We show that all proposed models can be optimised efficiently using powerful graph cut based move making algorithms.

Acknowledgements

I would like to thank my supervisors Phil Torr and Pushmeet Kohli for showing me the concepts of science. Without their guidance I would hardly know where to start and what to do. I would also like to thank my collaborators, Chris Russell, Karteek Alahari, Sunando Sengupta, Yalin Bastanlar, William Clocksin and Paul Sturgess, and all other (ex-)members of our lab. I am grateful to thank my family for their support. Paul deserves second acknowledgement for his after hours discussions about vision in the Angle & Greyhound, Half Moon and the Duke. These local pubs deserve a credit too; without their awesome ales I would hardly survive the relentless PhD pressure. And special thanks goes to the Duke's bar-maid Iona Caird; only her *truly impressive teasing smile* could make me forget about the NP-hardness of the Markov random field optimisation.

List of Publications

Journals

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

Inference Methods for CRFs with Co-occurrence Statistics

International Journal of Computer Vision, 2011

Invited Paper

Ľubor Ladický, Paul Sturgess, Chris Russell, Sunando Sengupta,
Yalin Bastanlar, William Clocksin, Philip H.S. Torr

Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction

International Journal of Computer Vision, 2011

Invited Paper

Pushmeet Kohli, Ľubor Ladický, Philip H.S. Torr

Robust Higher Order Potentials for Enforcing Label Consistency

International Journal of Computer Vision, 2009

Conferences

Ľubor Ladický, Philip H.S. Torr

Locally Linear Support Vector Machines

International Conference on Machine Learning, 2011

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

Graph Cut based Inference with Co-occurrence Statistics

European Conference on Computer Vision, 2010

Best Paper Award

Ľubor Ladický, Paul Sturgess, Karteek Alahari, Chris Russell, Philip H.S. Torr

What,Where & How Many? Combining Object Detectors and CRFs

European Conference on Computer Vision, 2010

Ľubor Ladický, Paul Sturgess, Chris Russell, Sunando Sengupta,
Yalin Bastanlar, William Clocksin, Philip H.S. Torr
Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction
British Machine Vision Conference, 2010
Best Paper Award

Chris Russell, Ľubor Ladický, Pushmeet Kohli, Philip H.S. Torr
Exact and Approximate Inference in Associative Hierarchical Networks using Graph-Cuts
Conference on Uncertainty in Artificial Intelligence, 2010

Ľubor Ladický, Chris Russell, Pushmeet Kohli, Philip H.S. Torr
Associative Hierarchical CRFs for Object Class Image Segmentation
International Conference on Computer Vision, 2009

Paul Sturgess, Karteek Alahari, Ľubor Ladický, Philip H.S. Torr
Combining Appearance and Structure from Motion Features for Road Scene Understanding
British Machine Vision Conference, 2009

Pushmeet Kohli, Ľubor Ladický, Philip H.S. Torr
Robust Higher Order Potentials for Enforcing Label Consistency
Conference on Computer Vision and Pattern Recognition, 2008

Contents

Chapter 1

Introduction

Scene understanding tasks can be formulated as a labelling problem that tries to assign a label to each unobserved hidden discrete variable. The labels correspond to various estimated properties of an image and may be for example an object class label (road, car, sky, building, ..) in the case of object class image segmentation [88], a depth label in the case of dense stereo reconstruction [54], a real pixel intensity in the case of image denoising [2], or a location of the pictorial structures [57]. The labels for the problems, where multiple labels have to be assigned, are typically conditionally dependent on each other and the output labelling tends to be highly structured. The most natural way to deal with this problem is to incorporate all conditional dependencies in one global probabilistic framework and solve the whole labelling jointly. However, the number of pixels in the image may grow to millions and conditional dependencies may be very complex, and that makes inference in many computer vision problems very hard to solve.

The standard way to deal with this problem is to model the dependencies of labels of pairs of variables in the local neighbourhoods. The most common models are known as pairwise Conditional Random Fields (CRF) (or their special case Markov Random Fields (MRF)). They have become very popular for solving several computer vision problems such as semantic object class segmentation, image denoising or dense stereo reconstruction. However, these pairwise models are very restricted in their expressivity. They cannot model rich natural statistics and induce desired complex structure such as connectivity, label set consistency or planarity of the output labelling. Enforcing these properties would require incorporation of conditional dependencies more general than pairwise and optimisation methods for the general case are infeasible.

In this thesis we show that CRFs with cliques of higher order than pairwise dependencies inducing complex structured properties useful in computer vision can still be solved efficiently using graph cut based methods with relatively low impact on the memory consumption and computational cost. We demonstrate the usefulness of our proposed models on several scene understanding problems such as object class semantic segmentation, dense $3D$ stereo reconstruction, object detection and localization.

In the next sections of this chapter we explain the basics of st-min cut, CRFs and their standard max-flow optimization algorithms. In chapter 2 we show how to formulate CRF problems that contain discriminative features across multiple scales. We demonstrate that our model is a generalization of the most popular models used for object class segmentation. We show that our proposed model can be optimised using efficient graph cut based algorithms. In chapter 3 we explain how models with label set preferences based on co-occurrence statistics can be formulated and optimised. In chapter 4 we propose a novel deformable template model for object detection and localization with MRF priors on the deformation field. We also show how the detector responses can be included in the CRF for object class segmentation. In chapter 5 we propose a model that jointly estimates labels for multiple domains and demonstrate its usefulness on the joint estimation of object class semantic segmentation and dense stereo reconstruction. In the last chapter 6 we conclude and suggest new directions for future research.

1.1 Labelling Problems in Computer Vision

In this section we introduce the principles behind Markov and Conditional Random Field and describe standard inference techniques, that will be used later in this thesis.

1.1.1 Markov and Conditional Random Fields

Discrete Labelling problems involving a large number of hidden variables are typically formulated using probabilistic frameworks called Markov Random Fields (MRFs), which model the conditional dependencies between unobserved hidden variables.

Let us consider the problem of assigning one label from the discrete set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$ for a discrete random variable per image pixel. We use $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ to denote the set of random variables corresponding to the image pixels $i \in \mathcal{V} = \{1, 2, \ldots, N\}$. The neighbourhood system $\mathcal{N}$ of the random field is defined by the sets $\mathcal{N}_i, \forall i \in \mathcal{V}$, where $\mathcal{N}_i$ denotes the set of all

neighbours of the variable X_i. A clique c is a subset of random variables $\mathbf{X}_c \subseteq \mathbf{X}$ which are conditionally dependent on each other. Any possible assignment of labels to the random variables will be called a *labelling* denoted by $\mathbf{x}$, which takes values from $\mathbf{L} = \mathcal{L}^N$.

The posterior distribution $\Pr(\mathbf{x}|\mathbf{D})$ for given data $\mathbf{D}$ over the labellings of the CRF is a *Gibbs* distribution and can be written as:

$$\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z}\exp(-\sum_{c\in\mathcal{C}}\psi_c(\mathbf{x}_c)), \tag{1.1.1}$$

where Z is a normalising constant called the *partition function*, and $\mathcal{C}$ is the set of all cliques [61]. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique $c \subset \mathcal{V}$ where $\mathbf{x}_c = \{x_i : i \in c\}$. The corresponding Gibbs energy is given by:

$$E(\mathbf{x}) = -\log\Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c\in\mathcal{C}}\psi_c(\mathbf{x}_c). \tag{1.1.2}$$

The number of variables in each clique $\mathbf{x}_c$ is called the order of the potential $\psi_c(\mathbf{x}_c)$. The most probable or Maximum a Posteriori (MAP) labelling $\mathbf{x}^*$ of the random field is defined as:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathbf{L}}\Pr(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{x}\in\mathbf{L}}E(\mathbf{x}). \tag{1.1.3}$$

Optimisation methods typically find the most probable solution by finding the labelling with the minimal energy. As the partition function is constant and thus does not affect the solution of the optimisation problem, we shall drop the partition function Z from future equations for compactness.

1.1.2 Pairwise Random Fields

Most labelling problems in vision are formulated as a pairwise MRF, whose energy can be written as the sum of unary and pairwise potentials as:

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}}\psi_i(x_i) + \sum_{i\in\mathcal{V},j\in\mathcal{N}_i}\psi_{ij}(x_i,x_j). \tag{1.1.4}$$

The unary potentials $\psi_i(x_i)$ of the MRF are typically defined as the negative log likelihood of variable X_i taking label x_i, while the pairwise potentials typically encode a smoothness prior which encourages neighbouring pixels in the image to take the same or similar label. The pairwise NRF suffers from a number of problems stemming from its inability to express high-level dependencies between pixels. Despite these limitations, it is widely used and very effective.

In MRFs only the unary potentials depend on the data. MRFs globally conditioned on the data are called Conditional Random Fields (CRFs) [61]. This distinction is rather philosophical, CRFs follow the same principles as MRFs in the early vision [6], and the optimization problems for CRF and MRF are exactly the same. In fact for the most part this thesis is concerned with the minimisation of discrete random field energy functions and the probabilistic interpretation only becomes important if one wants to estimate the CRF parameters.

1.2 Graph Cut based Inference for CRFs

Although the problem of finding the MAP labelling is NP-hard in general, for certain families of energy functions it can be solved exactly in polynomial time. One of those families are CRFs whose graphs form a tree, which can be solved using Belief Propagation [107]. This property does not apply for many computer vision problems, for which the task is to label each pixel in an image and the corresponding graph is a lattice. In this case submodular functions are widely used since they are also exactly solvable in polynomial time.

The binary energy function $E(\mathbf{x})$ is said to be submodular if, for each pair of binary variables $x_i, x_j \in \mathbf{x}$, and each labelling of the remaining variables $\bar{\mathbf{x}}_{ij} = \mathbf{x} \setminus \{x_i, x_j\}$:

$$E(0,0,\bar{\mathbf{x}}_{ij}) + E(1,1,\bar{\mathbf{x}}_{ij}) \leq E(0,1,\bar{\mathbf{x}}_{ij}) + E(1,0,\bar{\mathbf{x}}_{ij}). \tag{1.2.1}$$

Currently the best minimization algorithm [72] can solve general binary submodular problems in $\mathcal{O}(n^6 + n^5 Q)$, where Q is the time taken to evaluate the function and n is the number of variables. This optimisation method is computationally

expensive for computer vision problems with a very large number of nodes. However, it can be shown [37], that all submodular pairwise energies can be optimised by solving the corresponding st-mincut (also called graph cut) problem, which we explain next.

1.2.1 The st-Mincut Problem

In this section we provide a formulation of the st-mincut problem. Consider the directed weighted graph $G(V, E, C)$ with non-negative edge weights, where V is the set of vertices and E the set of edges with corresponding edge costs C. The number of vertices is denoted as $n = |V|$ and the number of edges as $m = |E|$. In the st-mincut problem there are two special terminal vertices called the source s and the sink t.

The st-cut is the partition of the set of vertices into two subsets S and $T = V - S$, such that $s \in S$ and $t \in T$, and the corresponding cost of the cut is defined as:

$$C_{S,T} = \sum_{i \in S, j \in T} c_{ij}. \tag{1.2.2}$$

The st-mincut problem is a problem of finding the partition with the lowest cost of the corresponding mincut:

$$(S^*, T^*) = \arg\min_{S,T} C_{S,T}, \tag{1.2.3}$$

where $T = V - S$. Using the variable $x_i = \delta(i \in T)$, where $\delta(\cdot)$ is the Kronecker δ-function, the st-mincut problem is equivalent to:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \sum_{(s,i) \in E} c_{si} x_i + \sum_{(i,t) \in E} c_{it}(1 - x_i) + \sum_{(i,j) \in E, i,j \notin \{s,t\}} c_{ij}(1 - x_i)x_j. \tag{1.2.4}$$

According to the max-flow min-cut theorem by Ford and Fulkerson [30], in a flow network the minimum cut is equal to the maximum amount of flow passing from the source to the sink. Let us assume all vertices are connected to the source and the sink and $(i, j) \in E \implies (j, i) \in E$ (for all non-existing edges we add an edge

with the weight 0). Then the equivalent max flow problem is defined as:

$$\max \sum_{i \in V} f_{si} \tag{1.2.5}$$

$$\text{s.}t. \qquad 0 \leq f_{ij} \leq c_{ij}, \qquad \forall (i,j) \in E \tag{1.2.6}$$

$$\sum_{j \in N(i)} f_{ji} - f_{ij} = 0, \quad \forall i \in V \setminus \{s,t\}, \tag{1.2.7}$$

where f_{ij} is the flow from node i to node j, c_{ij} is called the capacity of an edge and $N(i)$ is the set of neighbouring vertices connected by an edge to node i. The first set of constraints guarantee that the flow is non-negative and does not exceed the capacity of an edge and the second set of constraints guarantee the conservation of flow for nonterminal vertices. Given a flow f_{ij} the residual capacity r_{ij} of an edge $(i,j) \in E$ is defined as:

$$r_{ij} = c_{ij} - f_{ij} + f_{ji}. \tag{1.2.8}$$

A residual graph for $G_f(V,E,R)$ is a graph with the same set of vertices and edges with corresponding residual capacities as weights. An augmented path is a path from the source to the sink along the nonzero edges in the residual graph. If such a path does not exist, the graph is split into two disjoint sets S connected to the source and T to the sink, which are the solutions of the corresponding min-cut problem.

The most common method for finding the solution of the max flow problem is the augmenting paths algorithm, which iteratively finds the augmenting path of the residual graph and pushes the flow through it, until no such path exists. Various augmenting path algorithms differ only in the strategy of finding the augmenting paths [30, 21, 10].

1.2.2 Exact MAP Estimation for 2-label CRFs

As we already mentioned, all binary pairwise submodular functions can be solved using st-mincut. Let us consider the pairwise energy

$$\begin{aligned} E(\mathbf{x}) &= \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} \psi_{ij}(x_i, x_j) \\ &= \sum_{i\in\mathcal{V}} (g_i^1 x_i + g_i^0(1-x_i)) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} (g_{ij}^{00}(1-x_i)(1-x_j) \\ &+ g_{ij}^{01}(1-x_i)x_j + g_{ij}^{10}x_i(1-x_j) + g_{ij}^{11}x_i x_j), \end{aligned} \tag{1.2.9}$$

where each unary cost g_i^l is taken if $x_i = l \in \{0, 1\}$, and each pairwise cost g_{ij}^{lk} is taken if $x_i = l$ and $x_j = k$. The pairwise cost can be written as:

$$\psi_{ij}(x_i, x_j) = K_{ij} + g_i' x_i + g_j' x_j + c_{ij}(1-x_i)x_j + c_{ij}x_i(1-x_j) \tag{1.2.10}$$

where:

$$K_{ij} = g_{ij}^{00} \tag{1.2.11}$$

$$g_i' = \frac{g_{ij}^{10} + g_{ij}^{11} - g_{ij}^{01} - g_{ij}^{00}}{2}, \tag{1.2.12}$$

$$g_j' = \frac{g_{ij}^{01} + g_{ij}^{11} - g_{ij}^{10} - g_{ij}^{00}}{2}, \tag{1.2.13}$$

$$c_{ij} = \frac{g_{ij}^{01} + g_{ij}^{10} - g_{ij}^{00} - g_{ij}^{11}}{2}. \tag{1.2.14}$$

By the definition of the submodular functions $c_{ij} \geq 0$. By applying this transformation to the energy function and summing up all constant terms to K and all linear terms to $c_i x_i$ for each i:

$$E(\mathbf{x}) = K + \sum_{i\in\mathcal{V}} c_i x_i + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} c_{ij}(1-x_i)x_j + c_{ij}x_i(1-x_j). \tag{1.2.15}$$

Let $c_{it} = c_i$ and $c_{si} = 0$ if $c_i \geq 0$, and $c_{it} = 0$ and $c_{si} = -c_i$ otherwise. Then $c_i x_i = c_{it} x_i$ for $c_i \geq 0$ and $c_i x_i = -c_{si} + c_{si}(1-x_i)$ otherwise. Because the constant term does not have any effect on the argument of the minimum of the

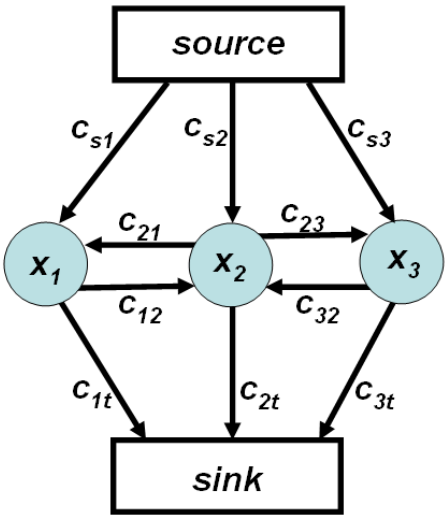


Figure 1.1: *A graph construction for the pairwise* CRF *with pairwise potentials between* x_1 *and* x_2*, and between* x_2 *and* x_3*.*

energy function, the optimisation problem becomes:

$$\begin{aligned} \mathbf{x}^* &= \arg\min_{\mathbf{x}} \sum_{i\in\mathcal{V}} c_{si}x_i + c_{it}(1-x_i) \\ &+ \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} c_{ij}(1-x_i)x_j + c_{ij}x_i(1-x_j). \end{aligned} \tag{1.2.16}$$

This formulation is equivalent to the st-min cut problem (1.2.4) with one vertex per variable x_i with c_{ij} as the set of edge costs. Each $x_i = 0$ if $x_i \in S$, and $x_i = 1$ otherwise. The equivalent graph construction is given in Figure 1.1. This transformation into a pairwise graph is not unique; we have given a transformation with symmetric pairwise edges. Also note that the graph construction can be swapped by swapping the values 0 and 1 of the source and the sink.

The class of functions solvable using max-flow algorithm can be extended to energy functions of orders higher than 2, for which each clique potential $\psi_c(\mathbf{x}_c)$ can be written as:

$$\psi_c(\mathbf{x}_c) = \min_{\mathbf{z}_c} \psi_c^p(\mathbf{x}_c, \mathbf{z}_c), \tag{1.2.17}$$

where $\mathbf{z}_c$ is the set of binary auxiliary variable and $\psi_c^p(\mathbf{x}_c, \mathbf{z}_c)$ is a pairwise submodular function. Potentials satisfying this property are called graph-representable. The most probable labelling is found by solving a submodular pairwise CRF prob-

lem:

$$x^* = \arg\min_{\mathbf{x}} \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) = \arg\min_{\mathbf{x}} \left(\min_{\mathbf{z}} \sum_{c \in \mathcal{C}} \psi_c^p(\mathbf{x}_c, \mathbf{z}_c) \right). \tag{1.2.18}$$

It has been shown that all binary submodular functions of order 3 [55] and several families [55, 31, 50, 81, 49] of submodular functions of higher order than 3 can be transformed into a corresponding pairwise submodular problem. However, the exact characterisation of graph-representable binary functions is not known.

1.2.3 Exact MAP Estimation for the n-label CRFs

The MAP estimation for certain classes of multi-label CRFs [46, 83, 75] can be also solved exactly by solving one st-mincut. This is done by designing an *encoding* [75], in which each state of the multi-label variable corresponds to the state of the multiple binary variables. The edges in the graph are designed in such a way that the cost of each possible cut is equal to the corresponding CRF energy under the chosen encoding scheme. Thus by finding the best cut, the minimal energy is found and the solution can be obtained by inverting the encoding scheme.

One such graph construction was proposed by Ishikawa [46] to solve multi-label problems exactly for pairwise energies with convex priors over ordered sets of labels, where the pairwise energy is called convex [46], if $\psi_{ij}(l_i, l_j) = f(l_i - l_j)$ and $f(.)$ is a discrete function, satisfying $f(i+1) - 2f(i) + f(i-1) \geq 0$. The encoding uses $|\mathcal{L}|$ binary variables $\mathbf{x}_i$ to represent the state of the $|\mathcal{L}|$-label variable y_i as follows:

$$\begin{aligned}
y_i = 1 &\iff \{x_i^1 = 0, x_i^2 = 1, x_i^3 = 1, x_i^4 = 1, .., x_i^{|\mathcal{L}|} = 1\} \\
y_i = 2 &\iff \{x_i^1 = 0, x_i^2 = 0, x_i^3 = 1, x_i^4 = 1, .., x_i^{|\mathcal{L}|} = 1\} \\
y_i = 3 &\iff \{x_i^1 = 0, x_i^2 = 0, x_i^3 = 0, x_i^4 = 1, .., x_i^{|\mathcal{L}|} = 1\} \\
&.. \\
y_i = |\mathcal{L}| &\iff \{x_i^1 = 0, x_i^2 = 0, x_i^3 = 0, x_i^4 = 0, .., x_i^{|\mathcal{L}|} = 0\}.
\end{aligned}$$

This encoding is also called the *battleship*[1] encoding. To disallow all other states

[1] The shape of the graph construction reminds one of a battleship.

of binary variables, sufficiently large pairwise edges $c_{x_i^{j+1}x_i^j} = K$ for each $j \in \{1, 2, ..|\mathcal{L}| - 1\}$ are used, where $K \to \infty$, guaranteeing $x_i^{j+1} = 0 \implies x_i^j = 0$. The unary potential for each variable y_i can be included in the graph under this encoding as a set of edges:

$$c_{sx_i^j} = K \tag{1.2.19}$$

$$c_{x_i^j x_i^{j+1}} = \psi_u(y_i = j) \quad \forall j = 1, ..|\mathcal{L}| - 1 \tag{1.2.20}$$

$$c_{x_i^{|\mathcal{L}|}t} = \psi_u(y_i = |\mathcal{L}|). \tag{1.2.21}$$

Intuitively the cost is taken when there is a transition 0/1 between x_i^j and x_i^{j+1}. To guarantee the positivity of the edges each unary cost can be increased by the same sufficiently large constant. This transformation does not change the optimal labelling.

The pairwise potential $\psi(y_i, y_k)$ is encoded using edges between corresponding binary variables $\mathbf{x}_i$ and $\mathbf{x}_k$. Assuming $y_i = l_i$ and $y_k = l_k$, the directed edge $c_{x_i^j x_k^m}$ is cut if $j \leq l_i$ and $m > l_k$ or $j > l_i$ and $m \leq l_k$. Thus the cost of the cut under the *battleship* encoding is:

$$C(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^{l_i} \sum_{m=l_k+1}^{|\mathcal{L}|} c_{x_i^j x_k^m} + \sum_{j=l_i+1}^{|\mathcal{L}|} \sum_{m=1}^{l_k} c_{x_k^m x_i^j}. \tag{1.2.22}$$

Under the constraint $C(\mathbf{x}_i, \mathbf{x}_k) = f(l_i - l_k)$ the second difference of this function can be shown [46] to be:

$$(f(l_i - l_k + 1) - f(l_i - l_k)) - (f(l_i - l_k) - f(l_i - l_k - 1)) = c_{x_k^{l_k} x_i^{l_i}} + c_{x_i^{l_i} x_k^{l_k}}. \tag{1.2.23}$$

Thus the capacities for the pairwise edges can be set to:

$$c_{x_i^j x_k^m} = c_{x_k^m x_i^j} = \frac{f(l_i - l_k + 1) - 2f(l_i - l_k) + f(l_i - l_k - 1)}{2}. \tag{1.2.24}$$

The resulting cost for any cut is the same as the corresponding CRF energy. All edges are non-negative if $f(\cdot)$ is convex. See [46] for more details. The equivalent graph construction is given in Figure 1.2. The class of pairwise multi-label problems that are exactly solvable in polynomial time using the max-flow

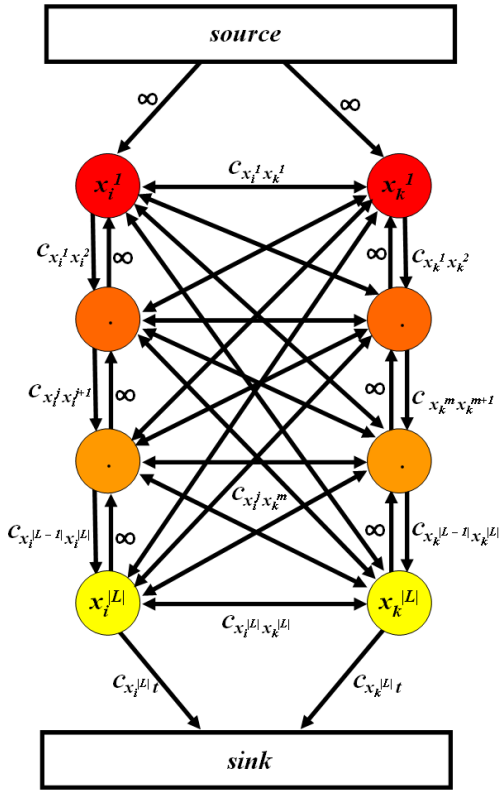


Figure 1.2: *A graph construction for the multi-label problem with convex prior using the battleship encoding [46].*

algorithm can be extended to multi-label submodular functions [83], defined as functions satisfying:

$$\psi(l_1, l_2) + \psi(l_1 + 1, l_2 + 1) \leq \psi(l_1 + 1, l_2) + \psi(l_1, l_2 + 1) \tag{1.2.25}$$

for each pair of variables $x_i, x_j \in \mathbf{x}$ and each pair of labels $l_1, l_2 \in \mathcal{L} \setminus \{l_L\}$.

1.2.4 Approximate MAP Estimation for CRFs

The optimisation problem for finding the MAP labelling for many practical multi-label computer vision problems is NP-hard and approximation algorithms have to be applied. Several methods for general pairwise CRFs have been proposed. These algorithms can be divided into three classes. Relaxation methods [84] formulate the problem as an integer program and relax non-convex constraints. The final labelling is obtained using one of the appropriate rounding schemes such as [48]. Message passing algorithms [107, 52] iteratively update their beliefs in each label

based on messages from their local neighbours. The last class of algorithms for approximately solving of CRFs are move making algorithms.

Move making algorithms iteratively project the problem into a smaller subspace of possible solutions containing the current solution. The solution of each subproblem proposes optimal moves which guarantee that the energy decreases after each move and must eventually converge. The move is optimal in a sense that it leads to the largest decrease in the energy under the move space being considered. The performance of move making algorithms depends dramatically on the size of the move space. The iterated conditional modes [4] (ICM) method allows in each iteration to change a label of one variable to one that reduces the overall energy. The method can be used for arbitrary CRFs, but its move space is very small and, thus tends to get stuck in poor local minima. Graph-Cut based move making algorithms [11] project the problem into a submodular binary one, solvable using max-flow algorithms. Unlike (ICM) their move space is exponential in the number of variables and, if applicable, they have been found to outperform other algorithms in terms of speed and energy [53, 81].

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 $(\alpha, \beta) \in \mathcal{L}^2$ 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:

$$T_{\alpha\beta}(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad x_i \in \{\alpha, \beta\} \text{ and } t_i = 0, \\ \beta & \text{if} \quad x_i \in \{\alpha, \beta\} \text{ and } t_i = 1. \end{cases} \tag{1.2.26}$$

Optimal $\alpha\beta$-swap moves cannot be efficiently found for all general CRF energies. One sufficient condition is the semi-metricity of the pairwise potentials. Pairwise

potentials are called semi-metric [11] if for all pairs of labels $l_a, l_b \in \mathcal{L}$:

$$\psi^p(l_a, l_a) = 0 \tag{1.2.27}$$

$$\psi^p(l_a, l_b) = \psi^p(l_b, l_a) \geq 0. \tag{1.2.28}$$

Trivially,

$$\psi^p(l_a, l_b) + \psi^p(l_b, l_a) - \psi^p(l_a, l_a) - \psi^p(l_b, l_b) = 2\psi^p(l_a, l_b) \geq 0, \tag{1.2.29}$$

thus the $\alpha\beta$-swap projection is submodular and is solvable using graph cut.

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 $\alpha \in \mathcal{L}$ 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:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad t_i = 0 \\ x_i & \text{if} \quad t_i = 1. \end{cases} \tag{1.2.30}$$

A sufficient condition for the submodularity of the projection is the metricity of the pairwise potentials. Pairwise potentials are called metric, if they are semi-metric and for any $l_a, l_b, l_c \in \mathcal{L}$:

$$\psi^p(l_a, l_b) + \psi^p(l_b, l_c) \geq \psi^p(l_a, l_c). \tag{1.2.31}$$

Let the current labels of two nodes be l_b and l_c. The submodular condition for the α-expansion move energy for the label l_a is:

$$\begin{aligned} &\psi^p(l_a, l_b) + \psi^p(l_c, l_a) - \psi^p(l_a, l_a) - \psi^p(l_b, l_c) \\ &\quad = \psi^p(l_a, l_b) + \psi^p(l_c, l_a) - \psi^p(l_b, l_c) \geq 0. \end{aligned} \tag{1.2.32}$$

Thus, under the metricity condition the move energy is submodular and solvable using graph cut.

Typical pairwise potentials for most computer vision problems enforce local smoothness of the labelling. They usually take the form of either a Potts model

$\psi^p(l_a, l_b) = K\delta(l_a \neq l_b)$ for unordered sets of labels, where $\delta(.)$ is Kronecker's δ-function, or a truncated convex prior $\psi^p(l_a, l_b) = K \min(f(|l_a - l_b|), T)$, where $f(.)$ is a convex function and T an optional truncation parameter. Both of these forms satisfy (semi-)metric conditions and thus $\alpha\beta$-swap or α-expansion algorithms can be applied.

Move making algorithms with binary move energies have been generalised to multi-label range swap move energies [104, 58] for pairwise potentials with truncated convex priors, allowing each pixel currently taking a label from a given range to change its label to any other label from that range. The transformation function of $\alpha\beta$-range move is defined as:

$$T_{\alpha\beta}(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad x_i \in [\alpha, \beta] \text{ and } t_i = 1, \\ \alpha + 1 & \text{if} \quad x_i \in [\alpha, \beta] \text{ and } t_i = 2, \\ .. & \\ \beta & \text{if} \quad x_i \in [\alpha, \beta] \text{ and } t_i = \beta - \alpha + 1. \end{cases} \tag{1.2.33}$$

where the pairwise cost is convex over the range $[\alpha - \beta, \beta - \alpha]$. The range move subproblem is solvable using graph cuts [46] as it is explained in the previous section 1.2.3. An expansion version of this range move algorithm allowing each pixel to keep its old label has been proposed in [58]. The move energy in each iteration is over-estimated by a convex function and Ishikawa's standard construction is applied [46]. The authors showed that this inference scheme leads to the same bound on the solution for convex truncated models as the linear programming (LP) relaxation [84]. This result is important because the LP solution is practically not useful for computer vision problems due to its high computational cost.

Chapter 2

Associative Hierarchical CRFs for Object Class Segmentation

Object class image segmentation (see figure 2.1) aims to assign an object label to each pixel of a given image. Over the last few years many different methods have been proposed for this problem. They can be broadly categorised on the basis of their choice of the quantisation (partitioning) of the image space[1]. Some methods are formulated in terms of pixels [88] (representing the finest quantisation), others used segments [3, 32, 108], groups of segments [74], or intersections of multiple segmentations [73], while some have gone to the extreme of looking at the whole image in order to reason about object segmentation [62].

In this chapter we present a model together with an efficient optimisation technique that contains the above mentioned previous methods as special cases, thus allowing for the use of holistic models that integrate the strengths of these different approaches.

2.1 Pixels vs Segments

Each choice of image quantisation comes with its share of advantages and disadvantages. Pixels might be considered the most obvious choice of quantisation. However, pixels by themselves contain a limited amount of information. The colour and intensity of a lone pixel is often not enough to determine its correct object label. Ren and Malik's [76] remark that *'pixels are not natural entities; they are merely a consequence of the discrete representation of images'* captures some of the problems of pixel-based representations.

The last few years have seen a proliferation of unsupervised segmentation methods [15, 24, 86], that perform an initial *a priori* segmentation of the image, applied to object segmentation [3, 32, 108, 40, 80, 108], and elsewhere [43, 91]. These rely upon an initial quantisation over the image space, typically based upon a segmentation of pixels based upon spatial location and colour/texture distribution.

Based upon the assumption that the quantisation is correct a segment based conditional random field (CRF) is defined over the image, and inference is per-

[1] We use the phrase "quantise the image" as opposed to "segment the image" in order to emphasise that a 'quantum' of the image space need not just be a collection of pixels. It could represent a sub-pixel division of the image space.

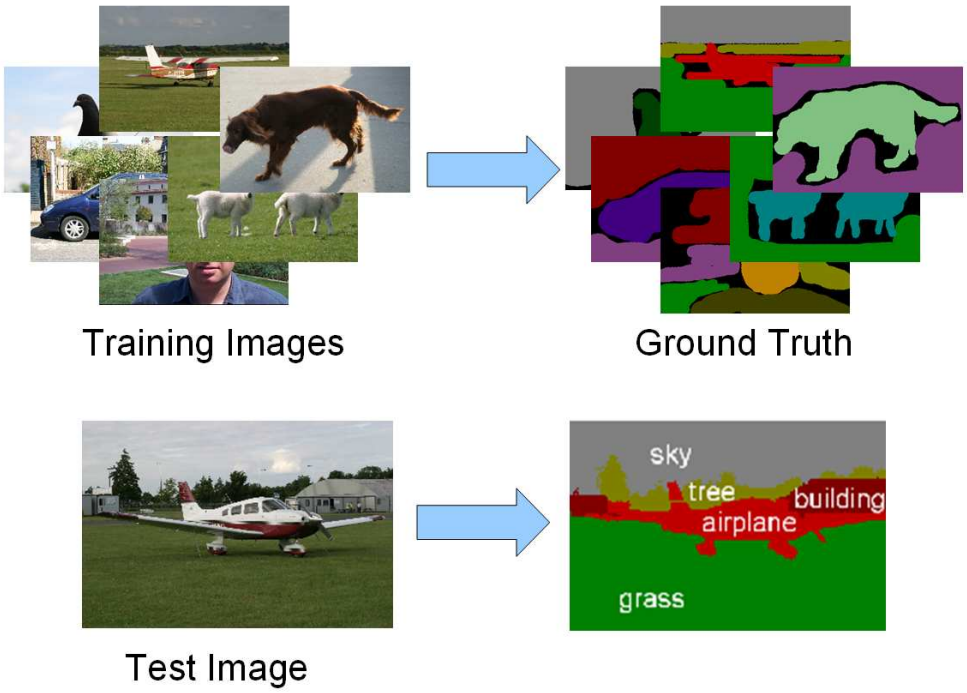


Figure 2.1: *Schematic description of the object class image segmentation problem. Given a set of training images with the corresponding ground truth the task is to build a classifier that will label a test image.*

formed to estimate the dominant label of each segment. This quantisation of the image allows the computation of powerful region-based features which are partially invariant to scale [105].

2.1.1 Use of Multiple Quantisations

Segment based methods work under the assumption that some segments share boundaries with objects in an image. This is not always the case, and this assumption may result in dramatic errors in the labelling (see figure 2.2). A number of techniques have been proposed to overcome errors in the image quantisation. Rabinovich *et al.* [74] suggested finding the most stable segmentation from a large collection of multiple segmentations in the hope that these would be more consistent with object boundaries. Larlus and Juri [62] proposed an approach to the problem driven by object detection. In their algorithm, rectangular regions are detected using a bag-of-words model based upon affine invariant features. These

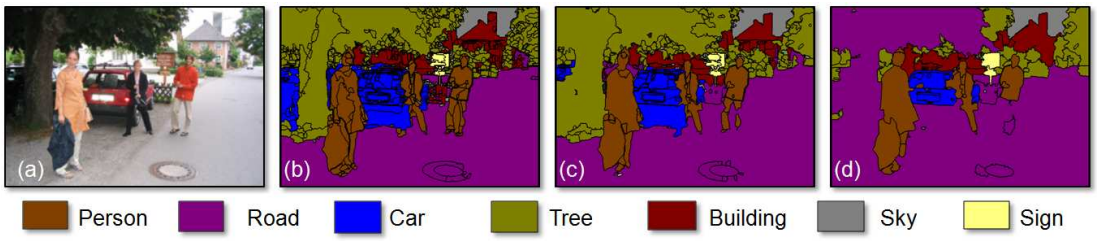


Figure 2.2: *Multiple unsupervised image segmentations. (a) Original image. (b)-(d) Unsupervised image segmentations with different image quantisations. (b), (c) and (d) use three different unsupervised segmentations of the image, in this case mean-shift, with different choices of kernel, to divide the image into segments. Each segment is assigned the label of the dominant object present in it. It can be seen that quantisation (b) is the best for tree, road, and car. However, quantisation (d) is better for the left person and the sign board.*

rectangles are refined using graph cuts to extract boundaries in a grab-cut [78] like approach. Such approaches face difficulties in dealing with cluttered images, in which multiple object classes intersect. Pantofaru *et al.* [73] observed that although segments may not be consistent with object boundaries, the segmentation map formed by taking the intersections of multiple segmentations often is. They proposed finding the most probable labelling of intersections of segments based upon the features of their parent segments. This scheme effectively reduces the image quantisation level. It results in more consistent segments but with a loss in the information content and discriminative power associated with each segment.

Another method to overcome these issues was proposed by Kohli *et al.* [51]. By formulating the labelling problem as a CRF defined over pixels, they were able to recover from misleading segments which spanned multiple object classes. Further, they were able to encourage individual pixels within a single segment to share the same label by defining higher order potentials (functions defined over cliques of size greater than 2) that penalised inconsistent labellings of segments. Their method can be understood as a relaxation of the hard constraint of previous methods, that the image labelling must follow the quantisation of the image space, to a softer constraint in which a penalty is paid for non-conformance.

Given the dependence of previous methods on the image partitioning (quantisation), the key question to be asked is: *What is the correct quantisation of an*

image and how can we find it? This is a difficult question to answer. As we explore the quantisation hierarchy from coarse to fine, we observe that while larger segments are perceptually more meaningful and easier to label correctly, they are less likely to lie inside a single object. Indeed pragmatically, it appears that the finding of an ideal quantisation may not be possible, and that segmentation of different objects in the image may require different quantisations (see figure 2.2).

In this chapter we propose a novel hierarchical CRF formulation of object class segmentation that allows us to unify multiple disparate quantisations of the image space, avoiding the need to make a decision of which is most appropriate. It allows for the integration of features derived from different quantisation levels (pixel, segment, and segment union/intersection). We will demonstrate how many of the state-of-the-art methods based on different fixed image quantisations can be seen as special cases of our model.

Inferring the Maximum a Posteriori solution in this framework involves the minimisation of a higher order function defined over several thousand random variables. We show that the solutions of such difficult function minimisation problems can be efficiently computed using graph-cut [10] based move-making algorithms. However, the contribution is not limited to the application of the novel hierarchical CRF framework to object class segmentation. We also propose new sophisticated potentials defined over the different levels of the quantisation hierarchy, and evaluate the efficacy of our framework on some of the most challenging data sets for object class segmentation, and show that it outperforms state-of-the-art methods based on individual image quantisation levels. We believe this is because: *(i)* Our methods generalise these previous methods allowing them to be represented as particular parameter choices of our hierarchical model. *(ii)* We go beyond these models by being able to use multiple hierarchies of segmentation simultaneously. *(iii)* In contrast to many previous methods that do not define any sort of cost function, or likelihood, we cleanly formulate the CRF energy of our model and show how it can be minimised.

2.1.2 Hierarchical Models and Context

The use of context has been well documented for object recognition and segmentation. It is particularly useful in overcoming ambiguities caused by limited evidence; this often occurs in object recognition where we frequently encounter objects at small scales or low resolution images [44]. Classical Markov and Conditional Random Field models exploit context in a local manner by encouraging adjacent pixels or segments to take the same label. To encode context at different scales Zhu *et al.* [109] introduced the hierarchical image model (HIM) built of rectangular regions with parent-child dependencies. This model captures large-distance dependencies and is solved efficiently using dynamic programming. However, it supports neither multiple hierarchies, nor dependencies between variables at the same level. To encode semantic context and to combine top-down and bottom-up approaches Tu *et al.* [99] proposed a framework in which they showed that the use of object specific knowledge helps to disambiguate low-level segmentation cues.

Our hierarchical CRF model uses a novel formulation that allows context to be incorporated at multiple levels of multiple quantisation, something not previously possible. As we will explain in section 2.5 it leads to improved segmentation results, while keeping the inference tractable.

2.2 CRFs for Object-Class Segmentation

Most pixel labelling problems in vision are formulated as a pairwise CRF whose energy can be written as the sum of unary and pairwise potentials as:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j). \tag{2.2.1}$$

The unary potentials $\psi_i(x_i)$ of the CRF are defined as the negative log likelihood of variable X_i taking label x_i, while the pairwise potential encodes a smoothness prior which encourages neighbouring pixels in the image to take the same label, resulting in a *shrinkage bias* [51].

The pairwise CRF formulation suffers from a number of problems stemming

from its inability to express high-level dependencies between pixels. Despite these limitations, it is widely used and very effective. Shotton *et al.* [88] applied the pairwise CRF to the object class segmentation problem. They defined the unary likelihoods potentials using the result of a boosted classifier over a region about each pixel, that they called *TextonBoost* and were able to obtain good results.

2.2.1 The Robust P^N model

The pairwise CRF formulation of [88] was extended by [51] with the incorporation of robust higher order potentials defined over segments. Their formulation was based upon the observation that pixels lying within the same segment are more likely to take the same label. The energy of the higher order CRF proposed by [51] was of the form:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_l} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{S}} \psi_c^h(\mathbf{x}_c), \tag{2.2.2}$$

where $\mathcal{S}$ is a set of cliques (or segments), and ψ_c^h are higher order potentials defined over them. Their higher order potentials took the form of a Robust P^N model defined as:

$$\psi_c^h(\mathbf{x}_c) = \min_{l \in \mathcal{L}}(\gamma_c^{\max}, \gamma_c^l + k_c^l N_c^l(\mathbf{x}_c)), \tag{2.2.3}$$

satisfying $\gamma_c^l \leq \gamma_c^{\max}, \forall l \in \mathcal{L}$, where $N_c^l(\mathbf{x}_c) = \sum_{i \in c} \delta(x_i \neq l)$ is the number of inconsistent pixels with the label l.

The potential takes cost γ_c^l if all pixels in the segment take the label l. Each inconsistent pixel is penalised with a cost k_c^l. The maximum cost of the potential is truncated to γ_c^{max}. By setting $\gamma_c^l = 0 \; \forall l \in \mathcal{L}$ this potential penalises inconsistent segments and thus encourages label consistency in segments. The weighted version of this potential is:

$$\psi_c^h(\mathbf{x}_c) = \min_{l \in \mathcal{L}}(\gamma_c^{\max}, \gamma_c^l + \sum_{i \in c} w_i k_c^l \delta(x_i \neq l)), \tag{2.2.4}$$

where w_i is the weight of the variable x_i.

This framework enabled the integration of multiple quantisations of the image

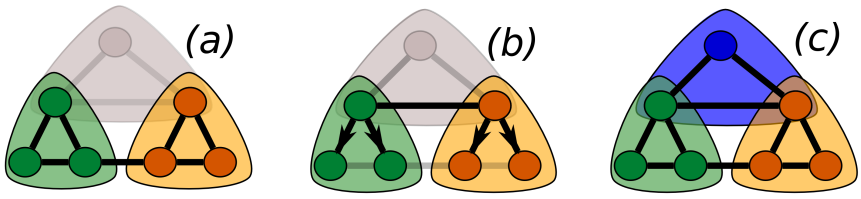


Figure 2.3: Existing models as special cases of our hierarchical model. *The lowest layer of the image represents the pixel layer, the middle layer potentials defined over super-pixels or segments, and the third layer represents our hierarchical terms.* (a) *shows the relationships permitted in a pixel-based* CRF *with Robust* P^N *potentials.* (b) *shows relationships contained within a super-pixel-based* CRF *(the directed edges indicate the one way dependence between the labellings of pixels and super-pixels).* (c) *Our hierarchical* CRF. *See section 2.3.*

space in a principled manner. However unlike our work, their choice of potential was independent of the choice of label and only encouraged pixels within the same segment to take the same label. Similarly, their model is unable to encode the conditional dependencies between segments. These potentials greatly increase the expressiveness of our model, as detailed in section 2.3.

2.2.2 The Robust P^N-Based Hierarchical CRFs

The higher-order P^N potentials of (2.2.4) are equivalent to the minimisation of a pairwise graph defined over the same clique $\mathbf{x}_c$ and a single auxiliary variable $x_c^{(1)}$, that takes values from an extended label set $\mathcal{L}^E = \mathcal{L} \cup \{l_F\}$. The cost function over $\mathbf{x}_c \cup \{x_c^{(1)}\}$ takes the form:

$$\psi_c^p(\mathbf{x}_c, x_c^{(1)}) = \phi_c(x_c^{(1)}) + \sum_{i \in c} \phi_c(x_c^{(1)}, x_i). \tag{2.2.5}$$

where the unary potential over $x_c^{(1)}$, $\phi_c(x_c^{(1)})$ associates the cost γ_c^l with $x_c^{(1)}$ taking a label in $\mathcal{L}$, and $\gamma_c^{\max}$ with $x_c^{(1)}$ taking the *free* label l_F. The pairwise potentials $\phi_c(x_c^{(1)}, x_i)$ are defined as:

$$\phi_c(x_c^{(1)}, x_i) = \begin{cases} 0 & \text{if } y_c = l_F \text{ or } x_c^{(1)} = x_i \\ w_i k_c^{x_c^{(1)}} & \text{otherwise.} \end{cases} \tag{2.2.6}$$

Then

$$\psi_c^h(\mathbf{x}_c) = \min_{x_c^{(1)}} \psi_c^p(\mathbf{x}_c, x_c^{(1)}). \tag{2.2.7}$$

By ensuring that the pairwise edges between the auxiliary variable and its children satisfy the constraint $\sum_i w_i k_c^l \geq 2\phi_c(l), \forall l \in \mathcal{L}$, we can guarantee that the labels of these auxiliary variables carry a clear semantic meaning. If this constraint is satisfied an auxiliary variable may take state $l \in \mathcal{L}$ in a minimal cost labelling, if and only if, the weighted majority of its child nodes take state l. State l_F indicates a heterogeneous labelling of a segment in which no label holds a significant majority. We now extend the model to include pairwise dependencies between auxiliary variables:

$$\begin{aligned} E(\mathbf{x}) &= \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} \psi_{ij}(x_i, x_j) \\ &+ \min_{\mathbf{x}^{(1)}} \Big(\sum_{c\in\mathcal{S}} \psi_c^p(\mathbf{x}_c, x_c^{(1)}) + \sum_{c,d\in\mathcal{S}} \psi_{cd}(x_c^{(1)}, x_d^{(1)}) \Big). \end{aligned} \tag{2.2.8}$$

These pairwise terms can be understood as encouraging consistency between neighbouring cliques. This framework can be further generalised to a hierarchical model where the connection between layers takes the form of (2.2.5) and the weights for each child node in $\phi_c(.)$ are proportional to the sum of the weights in the *"base layer"* belonging to the clique c.

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

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} \psi_{ij}(x_i, x_j) + \min_{\mathbf{x}^{(1)}} E^{(1)}(\mathbf{x}, \mathbf{x}^{(1)}), \tag{2.2.9}$$

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

$$\begin{aligned} E^{(n)}(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}) &= \sum_{c\in\mathcal{S}^{(n)}} \psi_c^p(\mathbf{x}_c^{(n-1)}, x_c^{(n)}) + \sum_{c,d\in\mathcal{S}^{(n)}} \psi_{cd}(x_c^{(n)}, x_d^{(n)}) \\ &+ \min_{\mathbf{x}^{(n+1)}} E^{(n+1)}(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}). \end{aligned} \tag{2.2.10}$$

Where $\mathbf{x}^{(0)} = \mathbf{x}$ refers to the state of the base level, and $\mathbf{x}^{(n)}$ for $n \geq 1$ the state of auxiliary variables.

The inter-layer potential between between two layers of auxiliary variables

takes the form of weighted Robust P^N:

$$\phi_c(\mathbf{x}_c^{(n-1)}, x_c^{(n)}) = \begin{cases} 0 & \text{if } x_c^{(n)} = l_F \text{ or } x_c^{(n)} = \mathbf{x}_c^{(n-1)} \\ w_{\mathbf{x}_c^{(n-1)}} k_c^l & \text{otherwise, where } l = \mathbf{x}_c^{(n-1)}, \end{cases} \tag{2.2.11}$$

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

$$w_{\mathbf{x}_c^{(n-1)}} = \sum_{i \in \mathbf{x}_c^{(n-1)}} w_i. \tag{2.2.12}$$

2.3 Relation to Previous Models

In this section, we draw comparisons with the current state-of-the-art models for object segmentation [32, 73, 74, 108] and show that at certain choices of the parameters of our model, these methods fall out as special cases (illustrated in figure 2.3). Thus, our method not only generalises the standard pairwise CRF formulations over pixels, but also the previous work based on super-pixels and (as we shall see) provides a global optimisation framework allowing us to combine features at different quantisation levels.

We will now show that our model is not only a generalisation of CRFs over pixels, but also of two classes of pre-existing model: *(i)* CRFs based upon disjoint segments [3, 32, 108] (see figure 2.3(b)), and *(ii)* CRFs based upon the intersection of segments [73].

2.3.1 Equivalence to CRFs based on Segments

Let us consider the case with only one segmentation and potentials defined only over this layer. In this case, $c \in \mathcal{S}$ are disjoint (non-overlapping)[2]. To ensure that $x_c^{(1)} \neq l_F, \forall c \in \mathcal{C}$, we assign a high value to $\gamma_c^{\max} \to \infty, \forall c \in \mathcal{C}$. As only the potential $\psi^p(\mathbf{x}_c, x_c^{(1)})$ acts upon $x_i : i \in c$, all pixels in c will take the same label. In this case, the optimal labelling will always be *segment consistent* (i.e. the labelling

[2] This is equivalent to the case where only one particular quantisation of the image space is considered.

of pixels within any segment is homogeneous) and the potential $\psi_c^p(\mathbf{x}_c, x_c^{(1)})$ can now be considered as a unary potential over the auxiliary (segment) variable $x_c^{(1)}$. This allows us to rewrite (2.2.8) as:

$$E(\mathbf{x}^{(1)}) = \sum_{c \in \mathcal{S}^{(1)}} \psi_c(x_c^{(1)}) + \sum_{c,d \in \mathcal{S}^{(1)}} \psi_{cd}(x_c^{(1)}, x_d^{(1)}) \tag{2.3.1}$$

which is exactly the same as the cost associated with the pairwise CRF defined over segments with $\psi_c(x_c^{(1)} = l) = \gamma_c^l$ as the unary cost and $\psi_{cd}(\cdot)$ as the pairwise cost for each segment. In this case, our model becomes equivalent to the pairwise CRF models defined over segments [3, 32, 74, 108].

2.3.2 Equivalence to Models of Segment Intersections

Let us now consider the case with multiple overlapping segmentations and potentials defined only over this layer. If we set $w_i k_c^l = \gamma_c^{\max}$, $\forall i \in \mathcal{V}, l \in \mathcal{L}, c \in \mathcal{S}$, then $x_c^{(1)} \neq l_F$ only if $x_i = x_c^{(1)}, \forall i \in c$. In this case, only the potentials $\sum_{c \ni i} \psi_c^p(\mathbf{x}_c, x_c^{(1)})$ act on x_i.

Consider a pair of pixels i, j that lie in the same intersection of segments *i.e.* $\{c \in S : c \ni i\} = \{c \in S : c \ni j\}$. Then, in a minimal labelling, either $\exists x_c^{(1)} = x_i$, and hence $x_j = x_c^{(1)} = x_i$, or $\forall c \ni i : x_c^{(1)} = l_F$. In the second case there are no constraints acting on x_i or x_j, and a minimal cost labelling can be chosen such that $x_i = x_j$.

Consequentially, there is always a minimal cost labelling consistent with respect to the intersection of segments, in this sense our model is equivalent to that proposed in [73].

2.3.3 Robustness to Misleading Segmentations

As discussed before, the quantisation of image space obtained using unsupervised segmentation algorithms may be misleading since segments may contain multiple object classes. Assigning the same label to all pixels of such segments will

result in an incorrect labelling. This problem can be overcome by using segment quality measures proposed by [74, 76] which can be used to distinguish the *good* segments from *misleading* ones. These measures can be seamlessly integrated in our hierarchical framework by modulating the strength of the potentials defined over segments. Formally, this is achieved by weighting the potentials $\psi_c^h(\mathbf{x}_c, x_c^{(1)})$ according to a quality sensitive measure $Q(c)$ for any segment c.

2.4 Inference for Hierarchical CRFs

It has been experimentally shown [53, 81], that for most computer vision problems graph cut [10] based move making algorithms [11] tend to outperform other approaches in terms of speed and quality.

In this section we show how to find the optimal move if we allow in each α-expansion iteration all variables in the base layer to either keep their old label or change their label to α, and all variables in the auxiliary layers to either keep the old label, change their label to l_F or change the label to α. It can be shown, that if the hierarchy is *well-founded* [81] this kind of move is not optimal only for the hierarchical energy over $|\mathcal{L}| + 1$ labels but also over higher order energy (2.2.9). See [81] for more details.

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

The transformation function for the base layer variables is encoded the same way as standard α-expansion:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad t_i = 0 \\ x_i & \text{if} \quad t_i = 1. \end{cases} \tag{2.4.1}$$

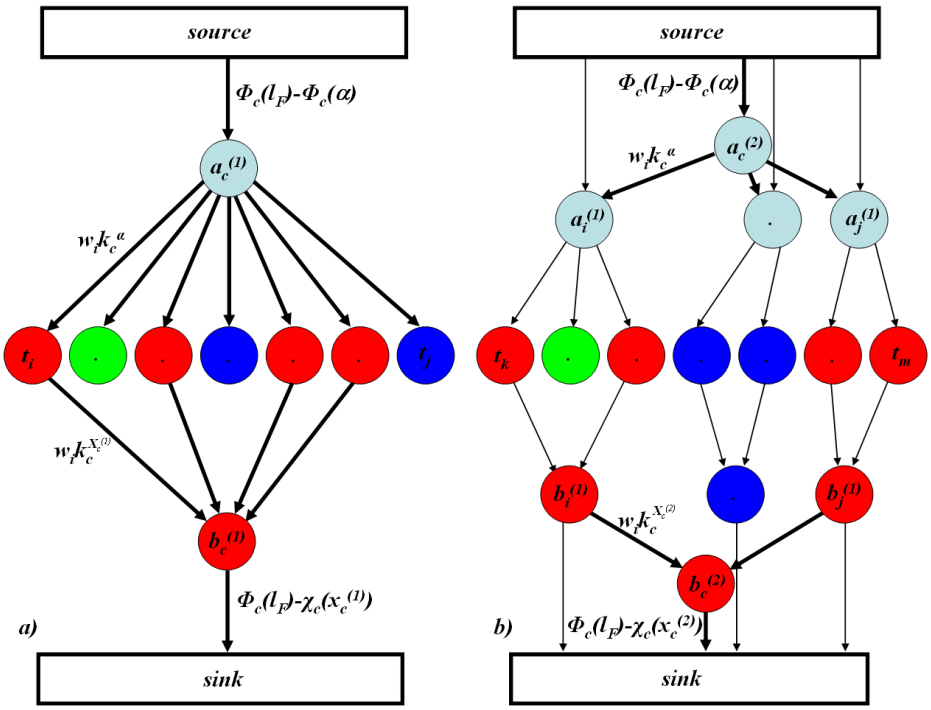


Figure 2.4: *A graph construction for the α-expansion move of the inter-layer connection between a) base layer and the first auxiliary layer, b) between two auxiliary levels. The colour of variables t_i and $b_c^{(n)}$ corresponds to the label before the move. Each variable $a_c^{(n)}$ is connected to each of the variables t_i respectively $a_i^{(n-1)}$ in the clique of the previous level, each variable $b_c^{(n)}$ is connected to each of the variables t_i respectively $b_i^{(n-1)}$ in the clique of the previous level. Edges modelling corresponding inter-layer connection are bold.*

The transformation function for the auxiliary variables is encoded as:

$$T_\alpha(\mathbf{x}_c^{(n)}, a_c^{(n)}, b_c^{(n)}) = \begin{cases} \alpha & \text{if } a_c^{(n)} = 0 \text{ and } b_c^{(n)} = 0 \\ x_c^{(n)} & \text{if } a_c^{(n)} = 1 \text{ and } b_c^{(n)} = 1 \\ l_F & \text{if } a_c^{(n)} = 1 \text{ and } b_c^{(n)} = 0. \end{cases} \tag{2.4.2}$$

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

2.4.1 Graph Construction for the Inter-layer Potential

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

$$\psi_c^p(\mathbf{x}_c, x_c^{(1)}) = \phi_c(x_c^{(1)}) + \sum_{i \in c} \phi_c(x_c^{(1)}, x_i), \tag{2.4.3}$$

where

$$\phi_c(x_c^{(1)}, x_i) = \begin{cases} 0 & \text{if } x_c^{(1)} = l_F \text{ or } x_c^{(1)} = x_i \\ w_i k_c^{x_c^{(1)}} & \text{otherwise.} \end{cases} \tag{2.4.4}$$

The move energy of this potential is:

$$\psi_c^p(\mathbf{t}_c, a_c^{(1)}, b_c^{(1)}) = \begin{cases} \phi_c(\alpha) + \sum_{i \in c} w_i k_c^{\alpha} t_i & \text{if } a_c^{(1)} = 0 \text{ and } b_c^{(1)} = 0 \\ \chi_c(x_c^{(1)}) + \sum_{i \in c} w_i k_c^{x_c^{(1)}} (1 - t_i) \delta(x_i = x_c^{(1)}) & \text{if } a_c^{(1)} = 1 \text{ and } b_c^{(1)} = 1 \\ \phi_c(l_F) & \text{if } a_c^{(1)} = 1 \text{ and } b_c^{(1)} = 0, \end{cases} \tag{2.4.5}$$

where $\chi_c(x_c^{(1)}) = \phi(x_c^{(1)}) + \sum_{i \in c} w_i k_c^{x_c^{(1)}} \delta(x_i \neq x_c^{(1)})$. The move energy can be transformed into:

$$\begin{aligned} \psi_c^p(\mathbf{t}_c, a_c^{(1)}, b_c^{(1)}) &= \phi_c(\alpha) + \chi_c(x_c^{(1)}) - \phi_c(l_F) \\ &+ \sum_{i \in c} w_i k_c^{\alpha} t_i (1 - a_c^{(1)}) + (\phi_c(l_F) - \phi_c(\alpha)) a_c^{(1)} \\ &+ \sum_{i \in c} w_i k_c^{x_c^{(1)}} \delta(x_i = x_c^{(1)}) (1 - t_i) b_c^{(1)} + (\phi_c(l_F) - \chi_c(x_c^{(1)}))(1 - b_c^{(1)}). \end{aligned} \tag{2.4.6}$$

The equivalence can be shown by checking the value of the transformed move energy for each combination of $a_c^{(1)}$ and $b_c^{(1)}$. The move energy is pairwise submodular and thus represents our inter-layer potential. The graph is equivalent to the Robust-P^N graph construction in [51].

For the inter-layer potential between two auxiliary layers $\mathbf{x}^{(n)}$ and $\mathbf{x}^{(n-1)}$ where

$n > 1$, the pairwise cost becomes:

$$\phi_c(x_c^{(n)}, x_d^{(n-1)}) = \begin{cases} 0 & \text{if } x_c^{(n)} = l_F \text{ or } x_c^{(n)} = x_d^{(n-1)} \\ w_d k_c^{x_c^{(n)}} & \text{otherwise.} \end{cases} \tag{2.4.7}$$

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

$$\psi_c^p(\mathbf{a}^{(n-1)}, \mathbf{b}^{(n-1)}, a_c^{(n)}, b_c^{(n)}) = \begin{cases} \phi_c(\alpha) + \sum_{i \in c} w_i k_c^{\alpha} a_i^{(n-1)} & \text{if } a_c^{(n)} = 0 \text{ and } b_c^{(n)} = 0 \\ \chi_c(x_c^{(n)}) + \sum_{i \in c} w_i k_c^{x_c^{(n)}} (1 - b_i^{(n-1)}) \delta(x_i^{(n-1)} = x_c^{(n)}) & \\ & \text{if } a_c^{(n)} = 1 \text{ and } b_c^{(n)} = 1 \\ \phi_c(l_F) & \text{if } a_c^{(n)} = 1 \text{ and } b_c^{(n)} = 0, \end{cases} \tag{2.4.8}$$

where $\chi_c(x_c^{(n)}) = \phi(x_c^{(n)}) + \sum_{i \in c} w_i k_c^{x_c^{(n)}} \delta(x_i^{(n-1)} = x_c^{(n)})$. Similarly to the previous case the move energy can transformed into:

$$\begin{aligned} \psi_c^p(\mathbf{a}^{(n-1)}, \mathbf{b}^{(n-1)}, a_c^{(n)}, b_c^{(n)}) &= \phi_c(\alpha) + \chi_c(x_c^{(n)}) - \phi_c(l_F) \\ &+ \sum_{i \in c} w_i k_c^{\alpha} a_i^{(n-1)} (1 - a_c^{(n)}) + (\phi_c(l_F) - \phi_c(\alpha)) a_c^{(n)} \\ &+ \sum_{i \in c} w_i k_c^{x_c^{(n)}} \delta(x_i^{(n-1)} = x_c^{(n)}) (1 - b_i^{(n-1)}) b_c^{(n)} \\ &+ (\phi_c(l_F) - \chi_c(x_c^{(n)}))(1 - b_c^{(n)}). \end{aligned} \tag{2.4.9}$$

The graph constructions for both cases of inter-layer connection are given in figure 2.4.

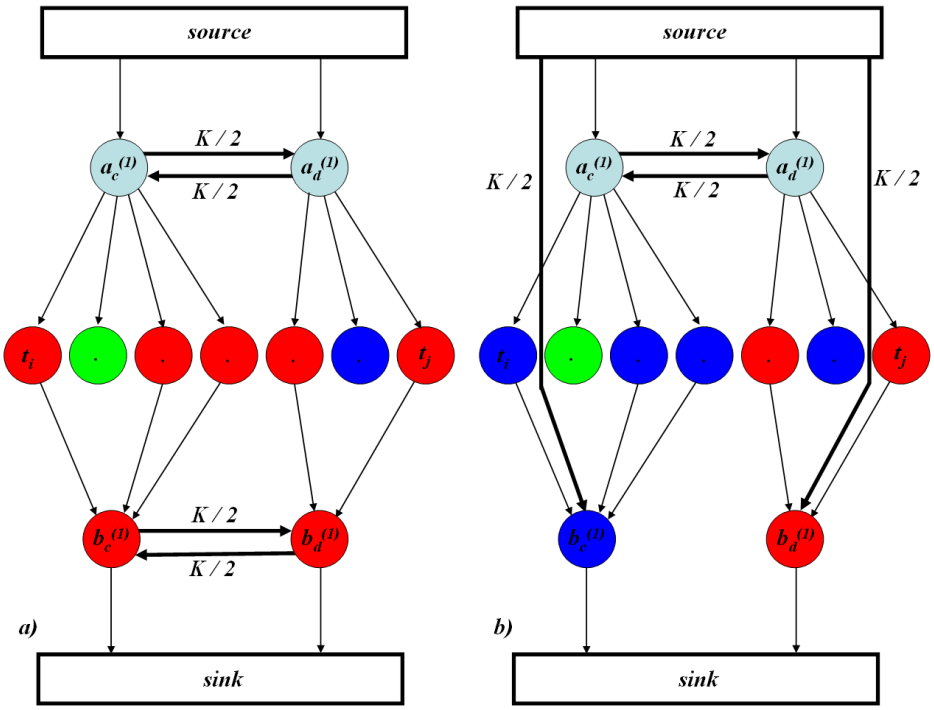


Figure 2.5: *A graph construction for the α-expansion move of the pairwise potential on the auxiliary level if the label before the move was a) the same, b) different. The colours of variables t_i and $b_c^{(n)}$ correspond to the label before the move. Edges modelling corresponding pairwise potentials are bold.*

2.4.2 Graph Construction for the Pairwise Potentials of the Auxiliary Variables

A sufficient condition for the graph-representability of the pairwise potential, that is given in [81] takes the form:

$$\psi_{cd}^{p}(x_c^{(n)}, x_d^{(n)}) = \begin{cases} 0 & \text{if } x_c^{(n)} = x_d^{(n)} \\ \frac{K}{2} & \text{if } (x_c^{(n)} = l_F \text{ and } x_d^{(n)} \neq l_F) \text{ or } (x_c^{(n)} \neq l_F \text{ and } x_d^{(n)} = l_F) \\ K & \text{if } x_c^{(n)} \neq x_d^{(n)} \neq l_F. \end{cases} \tag{2.4.10}$$

In case $x_c^{(n)} = x_d^{(n)}$ the move energy of the pairwise potentials between auxiliary

variables is:

$$\psi^p_{cd}(a^{(n)}_c, b^{(n)}_c, a^{(n)}_d, b^{(n)}_d) = \begin{cases} 0 & \text{if } a^{(n)}_c = a^{(n)}_d \text{ and } b^{(n)}_c = b^{(n)}_d \\ \frac{K}{2} & \text{if } (a^{(n)} \neq a^{(n)}_d \text{ and } b^{(n)}_c = b^{(n)}_d) \\ & \text{or } (a^{(n)} = a^{(n)}_d \text{ and } b^{(n)}_c \neq b^{(n)}_d) \\ K & \text{if } a^{(n)}_c \neq a^{(n)}_d \text{ and } b^{(n)}_c \neq b^{(n)}_d. \end{cases} \tag{2.4.11}$$

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

$$\begin{aligned} \psi^p_{cd}(a^{(n)}_c, b^{(n)}_c, a^{(n)}_d, b^{(n)}_d) &= \frac{K}{2} a^{(n)}_c (1 - a^{(n)}_d) + \frac{K}{2}(1 - a^{(n)}_c) a^{(n)}_d \\ &+ \frac{K}{2} b^{(n)}_c (1 - b^{(n)}_d) + \frac{K}{2}(1 - b^{(n)}_c) b^{(n)}_d. \end{aligned} \tag{2.4.12}$$

The equivalence can be shown by checking all possible combinations of $a^{(n)}_c$, $b^{(n)}_c$, $a^{(n)}_d$ and $b^{(n)}_d$.

In the case that $x^{(n)}_c \neq x^{(n)}_d$ the move energy of the pairwise potential between auxiliary variables becomes:

$$\psi^p_{cd}(a^{(n)}_c, b^{(n)}_c, a^{(n)}_d, b^{(n)}_d) = \begin{cases} 0 & \text{if } a^{(n)}_c = a^{(n)}_c \text{ and } b^{(n)}_c = b^{(n)}_d = 0 \\ K & \text{if } a^{(n)}_c = a^{(n)}_d \text{ and } b^{(n)}_c = b^{(n)}_d = 1 \\ \frac{K}{2} & \text{if } (a^{(n)} \neq a^{(n)}_d \text{ and } b^{(n)}_c = b^{(n)}_d) \\ & \text{or } (a^{(n)} = a^{(n)}_d \text{ and } b^{(n)}_c \neq b^{(n)}_d) \\ K & \text{if } a^{(n)}_c \neq a^{(n)}_d \text{ and } b^{(n)}_c \neq b^{(n)}_d, \end{cases} \tag{2.4.13}$$

and the equivalent pairwise submodular move energy is:

$$\psi^p_{cd}(a^{(n)}_c, b^{(n)}_c, a^{(n)}_d, b^{(n)}_d) = \frac{K}{2} a^{(n)}_c (1 - a^{(n)}_d) + \frac{K}{2}(1 - a^{(n)}_c) a^{(n)}_d + \frac{K}{2} b^{(n)}_c + \frac{K}{2} b^{(n)}_d. \tag{2.4.14}$$

Note that the equivalence holds only for 3×3 allowed configurations of $a^{(n)}_c$, $b^{(n)}_c$, $a^{(n)}_d$ and $b^{(n)}_d$. Graph constructions for both cases $x^{(n)}_c = x^{(n)}_d$ and $x^{(n)}_c \neq x^{(n)}_d$ are given in figure 2.5.

All the previous constructions were made under the assumption that none of the variables already takes the label α or l_F. If a variable in the base layer

already takes the label α, the problem is equivalent to changing each t_i to 0 in all pairwise submodular expressions. If the variable in the auxiliary layer already takes the the label α, both $a_c^{(n)}$ and $b_c^{(n)}$ have to be changed to 0 in all derived expressions. In the case that the auxiliary variable takes the label l_F, the variable can take only label α and label l_F after the move and thus $b_c^{(n)}$ has to be changed to 0. Setting the label of any variable to 0 is equivalent to tying it to the sink or equivalently changing each incoming edge to this variable to the edge going to the sink. Setting the label of any variable to 1 is equivalent to tying it to the source or equivalently changing each outgoing edge of this variable to the edge going to the source. The infinite edge between $a_c^{(n)}$ and $b_c^{(n)}$ is not necessary if the hierarchy is *well-founded*, see [81] for more details.

2.5 Potentials for Hierarchical CRFs

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

In the previous sections we decomposed the energy (2.2.10) into a set of potentials $\psi_c(\mathbf{x}_c)$. In this section we will decompose them further, writing $\psi_c(\mathbf{x}_c) = \lambda_c \xi_c(\mathbf{x}_c)$, where ξ_c is a feature based potential over c and λ_c its weight. Initially we will discuss the learning of potentials $\xi_c(\mathbf{x}_c)$, and later discuss the learning of the weights λ_c.

For our application we used potentials defined over a three-level hierarchy. We refer to elements of each layer as pixels, segments and super-segments respectively. Unsupervised segments are initially found using multiple applications of a fine scale mean-shift algorithm [15]. "Super-segments" are based upon a coarse mean-shift segmentation, performed over the result of the previous segmentations.

2.5.1 Features

Several well-engineered features were experimentally found to be more discriminative then the raw RGB values of pixels. In our application we use textons [69], local binary patterns [71], multi-scale [8] dense SIFT [67] and opponent SIFT [100]. Textons [69] are defined as a clustered 16-dimensional response to 16 different filters - Gaussian, Gaussian derivative and Laplacian filters at different scales. Local binary pattern [71] is a 8-dimensional binary feature consisting of 8 comparisons of the intensity value of the center pixel with its neighbours. The SIFT [67] feature contains the histograms of gradients of 4×4 cells quantised into 8 bins. The resulting 128 dimensional vector is normalised to 1. Opponent SIFT [100] is a variant of coloured SIFT and is built of separate histograms of gradients for 3 channels in the transformed colour space. All features except local binary patterns are quantised to 150 clusters using standard K-means clustering.

2.5.2 Unary Potentials from Pixelwise Features

Unary potentials from pixelwise features are derived from *TextonBoost* [88], and allow us to perform texture based segmentation, at the pixel level, within the same framework. The features used for constructing these potentials are computed on every pixel of the image, and are also called *dense* features. TextonBoost estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. The shape filters are defined by a texton t and rectangular region r. Their response $v_{[t,r]}(i)$ for a given point i is the number of textons t in the region r placed relative to the point i. Corresponding weak classifiers are decision stumps, which split on a shape filter response and one of a set of thresholds. The most discriminative weak classifiers are found using multi-class Gentle Ada-Boost [95].

We observed that textons were unable to discriminate between some classes of similar textures. This motivated us to extend the *TextonBoost* framework by boosting classifiers defined on multiple dense features (such as colour, textons, histograms of oriented gradients (HOG) [18], and pixel location) together. Generalised shape filters are defined by feature type f, feature cluster t and rectangular

region r. Their response $v^f_{[t,r]}(i)$ for given point i is the number of features of type f belonging to cluster t in the region r placed relative to the point i. The pool of weak classifiers contains decision stumps based on the generalised shape filters against a set of thresholds θ. See [95, 88] for further details of the procedure. Our results show that the boosting of multiple features together results in a significant improvement of the performance (note the improvement from the 72% of [88] to 81% of our similar pixel-based CRF in figure 2.10). Further improvements were achieved using exponentially instead of linearly growing thresholds and Gaussian instead of uniform distribution of rectangles around the point. The potential is incorporated into the framework in the standard way as a negative log-likelihood.

2.5.3 Histogram-based Segment Unary Potentials

We now explain the unary potential defined over segments and super-segments. For many classification and recognition problems, the distributions of pixelwise feature responses are more discriminative than any feature alone. For instance, the sky can be either 'black' (night) or 'blue' (day), but is never 'half-black' and 'half-blue'. This consistency in the colour of object instances can be used as a region based feature for improving object segmentation results. The unary potential of an auxiliary variable representing a segment is learnt (using the normalised histograms of multiple clustered pixelwise features) using multi-class Gentle AdaBoost [95], where the pool of weak classifiers is as above, comparing the percentage of features of the cluster t of the feature f with one of the thresholds θ. The selection and learning procedure is identical to [95].

The segment potential is incorporated into the energy as:

$$\phi_c(x^{(1)} = l) = \lambda_s |c| \min(-H_l(c) + K, \alpha^h), \tag{2.5.1}$$

$$\phi_c(x^{(1)} = l_F) = \lambda_s |c| \alpha^h, \tag{2.5.2}$$

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

For our experiments, the cost of pixel labels differing from an associated segment label was set to $k^l_c = (\phi_c(x^{(1)} = l_F) - \phi_c(x^{(1)} = l))/0.1|c|$. This means that

up to 10% of the pixels can take a label different to the segment label without the segment variable changing its state to l_F.

2.5.4 Pairwise Potentials

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

$$\xi^p(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i,j) & \text{otherwise,} \end{cases} \tag{2.5.3}$$

where the function $g(i, j)$ is an edge feature based on the difference in the intensity of colours of neighboring pixels [9]. It is typically defined as:

$$g(i,j) = \theta_p + \theta_v \exp(-\theta_\beta ||I_i - I_j||^2), \tag{2.5.4}$$

where I_i and I_j are the colour vectors of pixel i and j respectively. These encourage neighbouring pixels in the image (having a similar colour) to take the same label. We refer the reader to [9, 78, 88] for details.

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

$$\xi^p_{cd}(x^{(1)}_c, x^{(1)}_d) = \begin{cases} 0 & \text{if } x^{(1)}_c = x^{(1)}_d, \\ g(c,d)/2 & \text{if } (x^{(1)}_c = l_F \text{ and } x^{(1)}_d \neq l_F) \\ & \text{or } (x^{(1)}_c \neq l_F \text{ and } x^{(1)}_d = l_F), \\ g(c,d) & \text{otherwise,} \end{cases} \tag{2.5.5}$$

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

2.6 Learning Weights for Hierarchical CRFs

Having learnt potentials $\xi_c(\mathbf{x}_c)$ as described earlier, the problem remains of how to assign appropriate weights λ_c. This weighting, and the training of CRF parameters in general is not an easy problem and there is a wide body of literature dealing with it [5, 41, 40, 92]. The approach we take to learn these weights uses a coarse to fine, layer-based, local search scheme over a validation set.

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

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

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

$$C(\mathbf{x}^{(i)}) = \sum_{c \in \mathcal{V}^{(i)}} \Delta(x_c^{(i)} \neq l_c \wedge l_c \neq l_F). \tag{2.6.1}$$

We optimise (2.6.1) layer by layer. The full method is given in algorithm 1, where we use $\lambda_1^{(i)}$ to refer to the weighting of unary potentials in the i^{th} layer, $\lambda_2^{(i)}$ the weight of the pairwise terms and $\lambda_h^{(i+1)}$ a scalar modifier of all terms in the $(i+1)^{\text{th}}$ layer or greater. Θ is an arbitrary constant that controls the precision of the final assignment of λ.

An alternative and elegant approach to this is that of [27] which we intend to investigate in future work.

Algorithm 1 *Weight Learning Scheme.*

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

2.7 Experiments

We evaluated the performance of our framework on four data sets: Corel, Sowerby, PASCAL VOC 2008 [22] and MSRC-21 [88].

MSRC-21 The MSRC segmentation data set contains 591 images of resolution 320×213 pixels, accompanied with a hand labelled object segmentation of 21 object classes. Pixels on the boundaries of objects are not labelled in these segmentations. The division into training, validation and test sets occupied 45%, 10% and 45% of the images. Methods are typically compared using global criteria or average-per-class recall criteria (see figure 2.10 for details). For these experiments, the hierarchy was composed of 3 pairs of nested segmentations. The parameters of the mean-shift kernels were chosen as $(6, 5), (12, 10)$; $(6, 7.5), (12, 15)$; and $(6, 9), (12, 18)$. The first value refers to the planar distance between points, and the second refers to the Euclidian distance in the LUV colour space. Quantitative comparison of performance with other methods is given in figure 2.10. Qualitative results are given in figure 2.6.

Corel The Corel segmentation data set contains 100 images of resolution 180×120 pixels of natural sceneries, with a hand labelled object segmentation of 7 object classes. The division into training and test sets occupied 50% and 50% the images. The same parameters as for MSRC data set have been used due to

an insufficient amount of data. Unlike in MSRC dataset, segment-based methods performed better than pixel-based (see figure 2.11 for more details). Qualitative results are given in figure 2.7.

Sowerby The Sowerby segmentation data set contains 106 images of resolution 96×64 pixels of road scenes, with a hand labelled object segmentation of 7 object classes. The division into training and test sets occupied 50% and 50% the images. Similarly to Corel data set, the same parameters as for MSRC dataset have been used due to an insufficient amount of data. Segment-based methods perform better than pixel-based (see figure 2.12 for more details). Small classes performed very badly due to insufficient amount of training and test data. Qualitative results are given in figure 2.8.

PASCAL VOC 2008 This data set was used for the PASCAL Visual Object Category segmentation contest 2008. It is especially challenging given the presence of significant background clutter, illumination effects and occlusions. It contains 511 training, 512 validation and 512 segmented test images of 20 foreground and 1 background classes. The organisers also provided $10,057$ images for which only the bounding boxes of the objects present in the image are marked. We did not use these additional images for training our framework. For this data set we used a two-level hierarchy. The methods are evaluated using intersection vs. union criteria [22] that penalises the performance of classes i and j given a mislabelling of i as j (see figure 2.13). Note that this is not equivalent to the percentage of pixels correctly labelled. Quantitative comparison of performance with other methods is given in 2.13. Qualitative results are given in figure 2.9. The only comparable methods used classification and detection priors trained over a much larger set of images. Note that the reported results are from the actual challenge. For more recent results see chapter 4.

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

2.8 Conclusions

We have presented a generalisation of many previous super-pixel based methods within a principled CRF framework. Our approach enabled the integration of features and contextual priors defined over multiple image quantisations in one optimisation framework that supports efficient MAP estimation using graph cut based move making algorithms. In order to do this, we have examined the use of auxiliary variables in CRFs which have been relatively neglected in computer vision over the past twenty years.

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

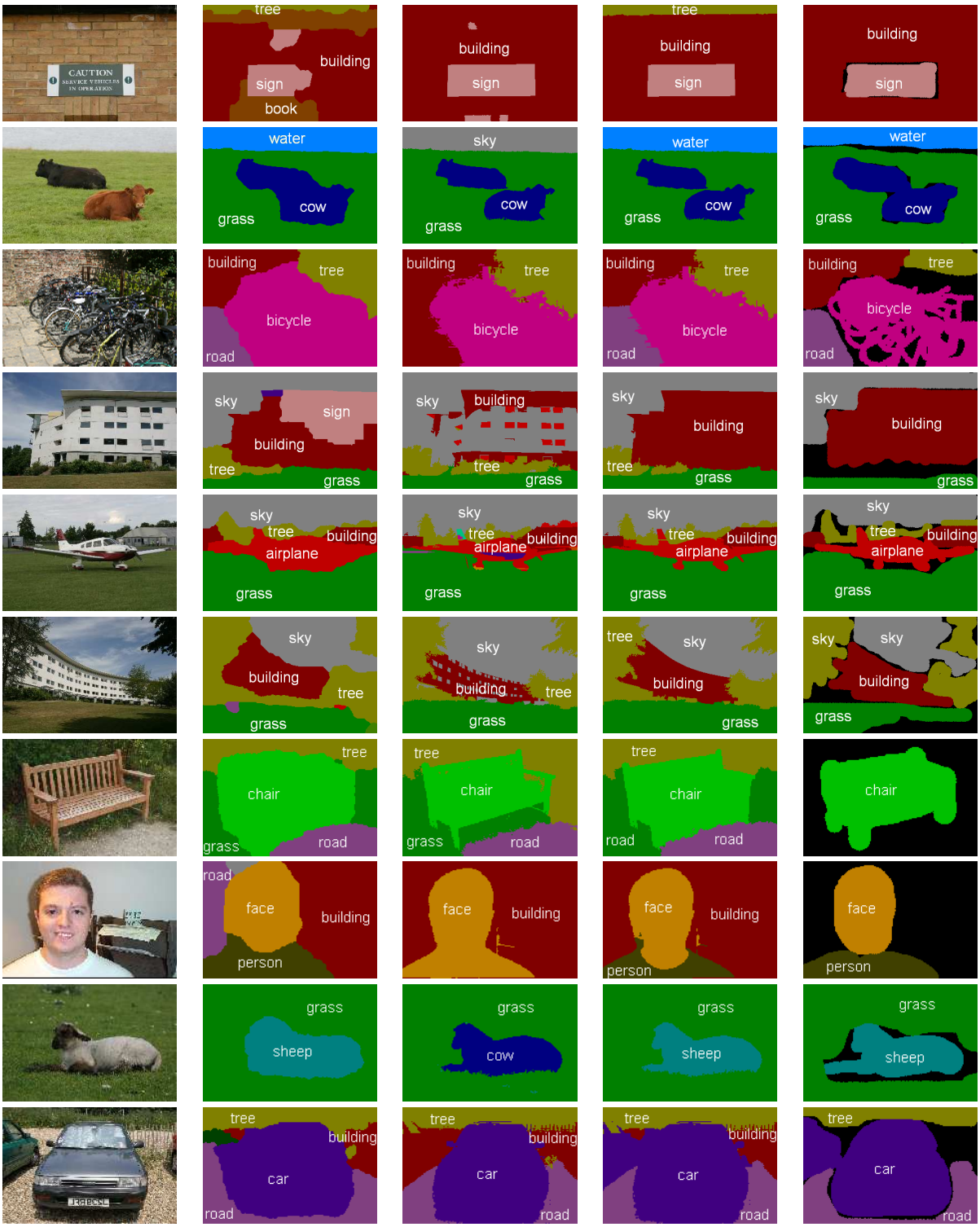
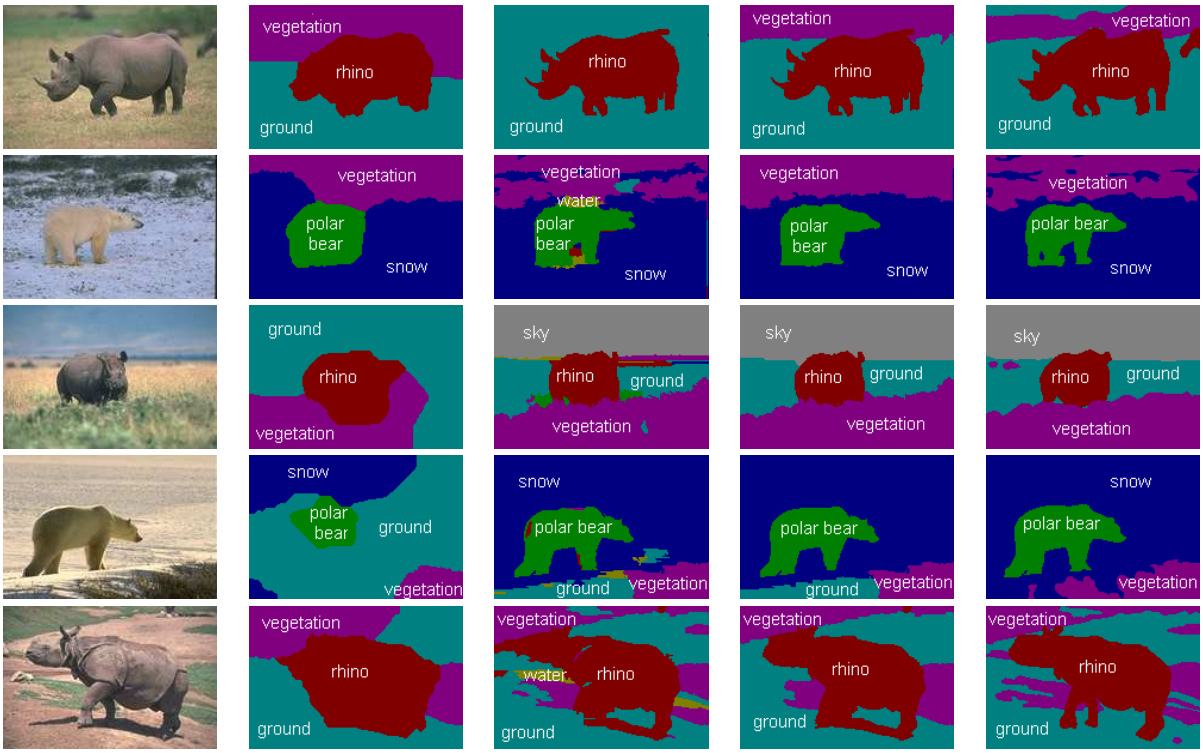
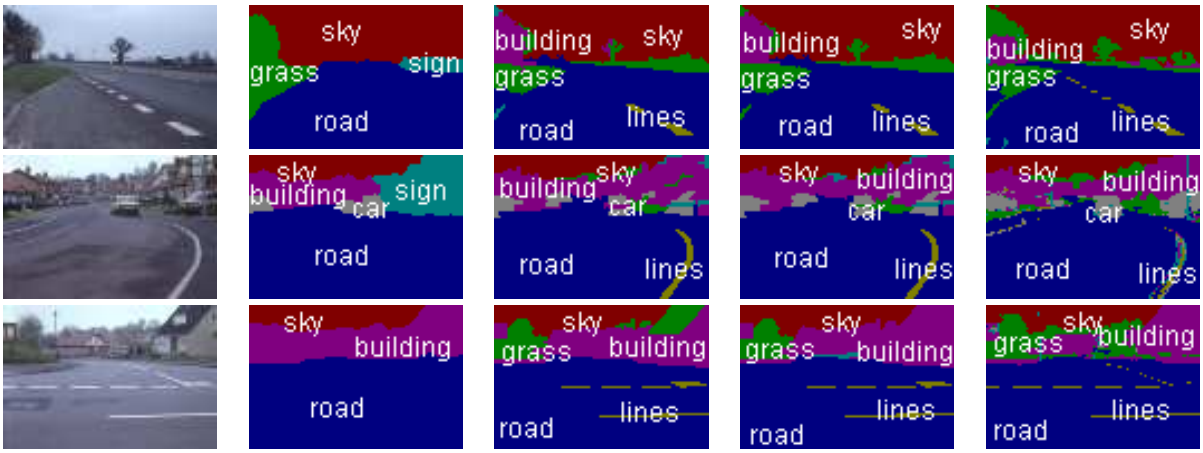


Figure 2.6: *Qualitative results on the MSRC-21 data set comparing non-hierarchical (i.e. pairwise models) approaches defined over pixels (similar to TextonBoost [88]) or segments (similar to [108, 73, 80] described in section 2.3) against our hierarchical model. Regions marked black in the hand-labelled ground truth image are unlabelled.*



Original Image | Pixel-based CRF | Segment-based CRF | Hierarchical CRF | Ground Truth

Figure 2.7: *Qualitative results on the Corel data set comparing approaches defined over pixels or segments against the hierarchical model.*



Original Image | Pixel-based CRF | Segment-based CRF | Hierarchical CRF | Ground Truth

Figure 2.8: *Qualitative results on the Sowerby data set comparing approaches defined over pixels or segments against the hierarchical model.*

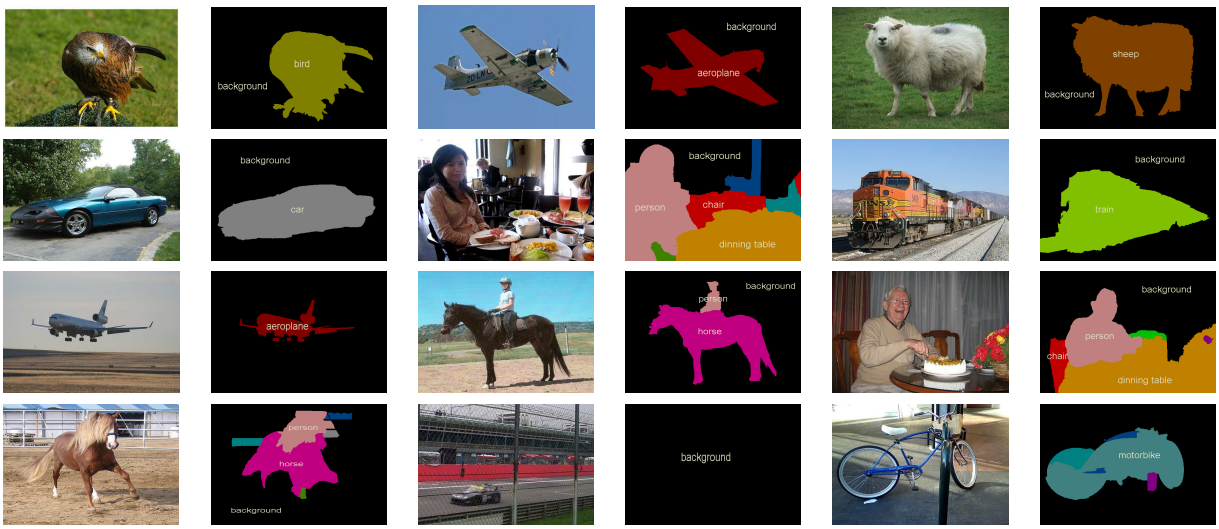


Figure 2.9: *Qualitative results on the VOC-2008 data set.* **Successful segmentations** *(top 3 rows) and* **standard failure cases** *(bottom) - from left to right, context error, detection failure and misclassification.*

	Global	Average	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
[87]	72	67	49	88	79	**97**	**97**	78	82	54	87	74	72	74	36	24	93	51	78	**75**	35	66	18
[88]	72	58	62	**98**	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	07
[3]	70	55	68	94	84	37	55	68	52	71	47	52	85	69	54	05	85	21	66	16	49	44	32
[108]	75	62	63	**98**	**89**	66	54	86	63	71	83	71	79	71	38	23	88	23	88	33	34	43	**32**
Pixel-based CRF	84	76	73	93	84	77	84	96	**85**	91	90	**86**	91	95	**91**	**41**	92	**53**	87	65	**77**	**70**	17
Segment-based CRF	81	66	80	**98**	83	64	81	**99**	59	89	85	68	68	**98**	76	26	85	39	84	30	49	50	07
Hierarchical CRF	**87**	**78**	**81**	96	**89**	74	84	**99**	84	**92**	**90**	**86**	**92**	**98**	**91**	35	**95**	**53**	**90**	62	**77**	**70**	12

Figure 2.10: *Quantitative results on the MSRC data set. The table shows % pixel recall measure $N_{ii}/\sum_j N_{ij}$ for different object classes. 'Global' refers to the overall error $\frac{\sum_{i\in\mathcal{L}} N_{ii}}{\sum_{i,j\in\mathcal{L}} N_{ij}}$, while 'average' is $\sum_{i\in\mathcal{L}} \frac{N_{ii}}{|\mathcal{L}|\sum_{j\in\mathcal{L}} N_{ij}}$. N_{ij} refers to the number of pixels of label i labelled j. The comparison suggests that the incorporation of the classifiers at different scales leads to a significant improvement of the performance.*

	Global	Average	Rhino/Hippo	Polar Bear	Water	Snow	Grass	Ground	Sky
[3]	83	**85**	87	**92**	82	**91**	66	**83**	**94**
Pixel-based CRF	76	72	80	85	88	83	75	57	35
Segment-based CRF	80	78	**92**	65	91	84	81	67	73
Hierarchical CRF	**84**	**85**	**92**	82	**94**	88	**83**	77	76

Figure 2.11: *Quantitative results on the Corel data set. Segment-based method tend to outperform pixel-based ones. Due to the insufficient amount of data the performance largely depends on the random split of the data. The same error measure as for the MSRC dataset has been used. Combining classifiers at different scales led to an improvement of the performance.*

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

Figure 2.12: *Quantitative results on the Sowerby data set. The segment-based method tend to outperform pixel-based ones. Context-based pixel method could not capture small objects due to the insufficient size of the images. The same error measure as for the MSRC dataset has been used. Similarly to other data sets, the hierarchical* CRF *outperformed both approaches over single scale.*

	Average	Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table	Dog	Horse	Motor bike	Person	Potted plant	Sheep	Sofa	Train	TV/monitor
XRCE	**25**	76	26	16	19	**22**	17	**27**	**25**	**24**	08	**25**	10	**18**	**23**	**34**	29	**23**	**32**	**15**	**26**	**37**
UIUC / CMU	19	**79**	32	**21**	08	07	**34**	16	23	10	01	07	08	10	23	25	28	16	04	05	19	32
MPI	13	75	19	08	06	09	04	11	12	06	01	04	16	40	12	16	16	01	20	06	15	13
Hierarchical CRF	20	75	**37**	05	**22**	11	14	14	20	10	**09**	04	**28**	07	17	23	**31**	14	27	12	20	25

Figure 2.13: *Quantitative analysis of VOC2008 results [22] based upon performance the intersection vs. union criteria* ($\frac{\sum_{i \in \mathcal{L}} N_{ii}}{|\mathcal{L}|(-N_{ii}+\sum_{j \in \mathcal{L}} N_{ij}+N_{ji})}$). *Note that all other methods used classification and detection priors trained over a much larger data set that included unsegmented images. The reported results are from the actual challenge, for recent results see chapter 4.*

Chapter 3

Co-occurrence Statistics in CRFs

Standard approaches for the object class segmentation problem 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 chapter 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 objects appearing in an image, co-occurrence potentials can also be used to impose minimum description length (MDL) prior, that encourages a parsimonious description of an image using fewer labels. As discussed eloquently in the recent work [13], the need for a bias towards parsimony becomes increasingly important as the number of classes to be considered increases. Figure 3.1 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.* [74] proposed the integration of such co-occurrence costs that characterise the relationship between two classes. Similarly Torralba *et al.* [96] 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.

3.1 CRFs and Co-occurrence

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

$$E(\mathbf{x}) = \sum \psi_c(\mathbf{x}_c) + K(\mathbf{x}). \tag{3.1.1}$$

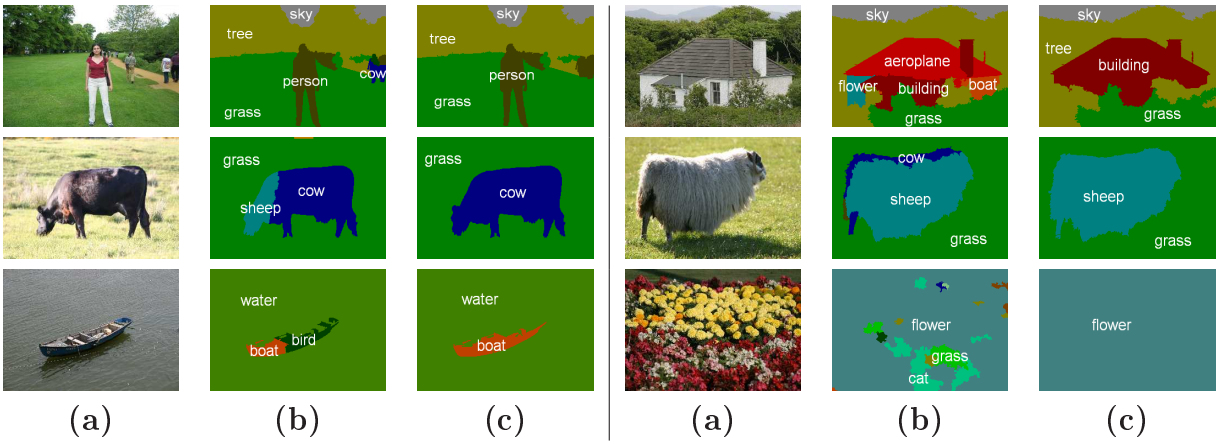


Figure 3.1: **Best viewed in colour:** *Qualitative results of object co-occurrence statistics.* **(a)** *Typical images taken from the MSRC data set [88];* **(b)** *A labelling based upon a pixel based random field model [59] that does not take into account co-occurrence;* **(c)** *A labelling of the 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.*

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 (3.1.1). 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 [96].

(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 [97], the surprise at seeing a polar bear in a street scene should 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 requirement

of any conventional inference algorithm [90] is typically $O(|\mathcal{V}|)$ for vision problems, the memory requirement 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:

$$K(\mathbf{x}) = C(L(\mathbf{x})) \tag{3.1.2}$$

Adding the co-occurrence term to the standard CRF cost function 3.1.1, we have:

$$E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) + C(L(\mathbf{x})). \tag{3.1.3}$$

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.* :

$$L_1 \subset L_2 \implies C(L_1) \leq C(L_2). \tag{3.1.4}$$

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.* $\psi_{\mathcal{V}}(\mathbf{x})$.

Method	Global energy *(i)*	Invariance *(ii)*	Efficiency *(iii)*	Parsimony *(iv)*
Unary ([96])	✓	✗	✓	✗
Pairwise ([74, 32, 97])	✓	✗	✗	✓
Hard decisions ([17])	✗	—	✓	—
Our approach	✓	✓	✓	✓

Figure 3.2: *A comparison of the capabilities of existing image co-occurrence formulations against our new approach. See section 3.1.1 for details.*

3.1.1 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.* [96] proposed the use of additional unary potentials to capture scene based occurrence priors. Their costs took the form:

$$K(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi(x_i). \tag{3.1.5}$$

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 [17], 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)*.

Several researchers ([74, 32], and independently [97]) 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:

$$K(\mathbf{x}) = \sum_{i,j \in \mathcal{V}} \phi(x_i, x_j), \tag{3.1.6}$$

where ϕ was some potential which penalised labels that should not occur together in an image. Unlike our model (3.1.3) 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 [11], 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 3.2.

Two efficient schemes [19, 45] 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 section 3.1.

Hoiem *et al.* [45] 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:

$$C(L(\mathbf{x})) = k||L(\mathbf{x})|| \tag{3.1.7}$$

In a recent work, independently appearing at the same time as ours, Delong *et al.* [19] also proposed the use of a cost over the number of labels present. In general their approach allowed a penalty cost to be taken if any label from a certain subset of labels 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 [93] for discussion of this problem. The general form of their costs is:

$$C(L(\mathbf{x})) = \sum_{L \subseteq \mathcal{L}} k_L \delta(L(\mathbf{x}) \cap L \neq \emptyset), \tag{3.1.8}$$

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

Note that the costs of [19] and [45] both satisfy the inequality:

$$C(L_1 \cup L_2) \leq C(L_1) + C(L_2), \tag{3.1.9}$$

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.

3.1.2 Inference on Global Co-occurrence Potentials

Consider the energy (3.1.3). The inference problem becomes:

$$\begin{aligned} \mathbf{x}^* &= \arg\min_{\mathbf{x} \in \mathcal{L}^{|\mathcal{V}|}} \textstyle\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) + C(L(\mathbf{x})) \\ &\text{s.t. } \mathbf{x} \in \mathcal{L}^{|\mathcal{V}|},\ L(\mathbf{x}) = \{l \in \mathcal{L} : \exists x_i = l\}. \end{aligned} \tag{3.1.10}$$

In this section we show that the problem of minimising this energy can be solved efficiently using move-making $\alpha\beta$-swap and α-expansion moves [11], where the number of additional edges of the graph grows linearly with the number of variables in the graph. In contrast to [74], these algorithms can be applied to large graphs with more than $200,000$ variables.

3.1.3 $\alpha\beta$-Swap Moves

Move making algorithms iteratively project the problem into a smaller subspace of possible solutions containing the 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 a 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 (3.1.3) is additive, we derive graph constructions only for the term $C(L(\mathbf{x}))$. The final graph is the merger of the graph for optimising the standard CRF [11] 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 $(\alpha, \beta) \in \mathcal{L}^2$ 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:

$$T_{\alpha\beta}(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad x_i \in \{\alpha, \beta\} \text{ and } t_i = 0, \\ \beta & \text{if} \quad x_i \in \{\alpha, \beta\} \text{ and } t_i = 1. \end{cases} \tag{3.1.11}$$

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:

$$S = \{L(\mathbf{x}) \cup \{\alpha\} \setminus \{\beta\}, L(\mathbf{x}) \cup \{\beta\} \setminus \{\alpha\}, L(\mathbf{x}) \cup \{\alpha, \beta\}\}\,. \tag{3.1.12}$$

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:

$$E(\mathbf{t}) = \begin{cases} C_\alpha = C(L(\mathbf{x}) \cup \{\alpha\} \setminus \{\beta\}) & \text{if } \forall i \in \mathcal{V}_{\alpha\beta}\,, t_i = 0, \\ C_\beta = C(L(\mathbf{x}) \cup \{\beta\} \setminus \{\alpha\}) & \text{if } \forall i \in \mathcal{V}_{\alpha\beta}\,, t_i = 1, \\ C_{\alpha\beta} = C(L(\mathbf{x}) \cup \{\alpha, \beta\}) & \text{otherwise.} \end{cases} \tag{3.1.13}$$

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 standard pairwise move energy $E'(\mathbf{t})$ without co-occurrence. The optimal move with co-occurence can be found as:

$$\mathbf{t}^* = \arg\min_{\mathbf{t}}(E'(\mathbf{t}') + C_{\alpha\beta}, E'(\mathbf{0}) + C_\alpha, E'(\mathbf{1}) + C_\beta), \tag{3.1.14}$$

where $\mathbf{0}$ and $\mathbf{1}$ are uniform vectors composed entirely of 0 or 1 respectively. If the solution $\mathbf{t}'$ contains both 0s and 1s, it must also be the best mixed solution including co-occurrence term and the optimal move can found by comparing its energy with co-occurrence with energies of homogenous moves. If the solution $\mathbf{t}'$ is composed solely of 0s or 1s, due to the parsimony condition

$$\forall \mathbf{t} : E(\mathbf{t}') \leq E(\mathbf{t}) \implies E'(\mathbf{t}') + C_\alpha \leq E'(\mathbf{t}) + C_{\alpha\beta} \tag{3.1.15}$$

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 exists an efficient solution similar to the one in [19] to find the optimal $\alpha\beta$-swap move for energies with co-occurrence, for illustration we also derive its graph construction efficiently 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*

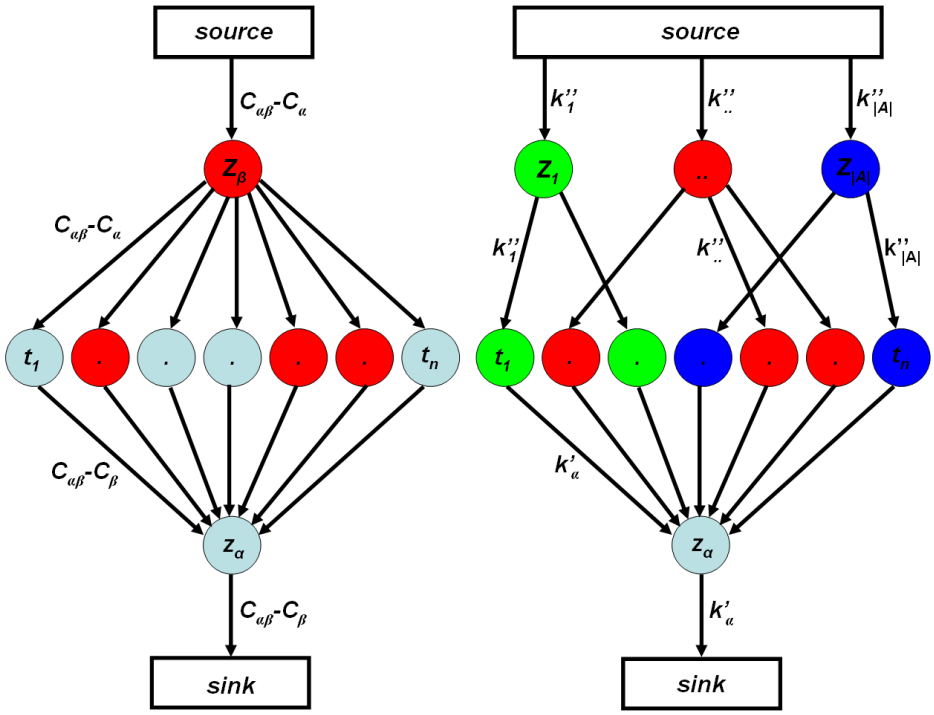


Figure 3.3: *Graph construction for $\alpha\beta$-swap and α-expansion move. In $\alpha\beta$-swap variable x_i will take the label α if the corresponding t_i are tied to the sink after the st-mincut and β otherwise. In α-expansion variable x_i changes the label to α if it is tied to the sink after the st-mincut and remains the same otherwise. Colours represent the labels of the variables before the move.*

auxiliary variables z_α and z_β as:

$$\begin{aligned} E(\mathbf{t}) &= C_\alpha + C_\beta - C_{\alpha\beta} + \min_{z_\alpha, z_\beta} \Big[(C_{\alpha\beta} - C_\alpha) z_\beta \\ &+ (C_{\alpha\beta} - C_\beta)(1 - z_\alpha) + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha,\beta} - C_\alpha) t_i (1 - z_\beta) \\ &+ \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i) z_\alpha \Big]. \end{aligned} \tag{3.1.16}$$

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

3.1.4 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 $\alpha \in \mathcal{L}$ 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:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad t_i = 0 \\ x_i & \text{if} \quad t_i = 1. \end{cases} \tag{3.1.17}$$

To derive a graph-construction that approximates the true cost of an α-expansion move we use the decomposition

$$C(L) = \sum_{B \subseteq L} k_B, \tag{3.1.18}$$

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

$$k_B = C(B) - \sum_{B' \subset B} k_{B'}. \tag{3.1.19}$$

This 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 the 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:

$$\delta_\alpha(\mathbf{t}) = \begin{cases} 1 & \text{if} \quad \exists i \in \mathcal{V} \text{ s.t. } t_i = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1.20}$$

$$\forall l \in A \ , \ \delta_l(\mathbf{t}) = \begin{cases} 1 & \text{if} \quad \exists i \in \mathcal{V}_l \text{ s.t. } t_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1.21}$$

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

$$\begin{aligned} E(\mathbf{t}) &= E_{new}(\mathbf{t}) - E_{old} \\ &= \sum_{B \subseteq A \cup \{\alpha\}} k_B \prod_{l \in B} \delta_l(\mathbf{t}) - C(A). \end{aligned} \tag{3.1.22}$$

Ignoring the constant term and decomposing the sum into parts with and without

terms dependent on α we have:

$$E(\mathbf{t}) = \sum_{B \subseteq A} k_B \prod_{l \in B} \delta_l(\mathbf{t}) + \sum_{B \subseteq A} k_{B \cup \{\alpha\}} \delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t}). \tag{3.1.23}$$

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

$$\delta_\alpha(\mathbf{t}) \prod_{l \in B} \delta_l(\mathbf{t}) = \delta_\alpha(\mathbf{t}) + \prod_{l \in B} \delta_l(\mathbf{t}) - 1. \tag{3.1.24}$$

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 (3.1.22) becomes:

$$\begin{aligned} E(\mathbf{t}) &= \sum_{B \subseteq A} k_{B \cup \{\alpha\}} \delta_\alpha(\mathbf{t}) + \sum_{B \subseteq A} (k_B + k_{B \cup \{\alpha\}}) \prod_{l \in B} \delta_l(\mathbf{t}) \\ &= k'_\alpha \delta_\alpha(\mathbf{t}) + \sum_{B \subseteq A} k'_B \prod_{l \in B} \delta_l(\mathbf{t}), \end{aligned} \tag{3.1.25}$$

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 [70, 79, 58] 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. For any $l' \in A$ we can overestimate $K(A, \mathbf{t})$ by:

$$\begin{aligned} K(A, \mathbf{t}) &\leq K(A \setminus \{l'\}, \mathbf{t}) \\ &+ \delta_{l'}(\mathbf{t}) \min_{S \subseteq A \setminus \{l'\}} \sum_{B \subseteq S} (k'_{B \cup \{l'\}} - k'_B) \\ &= K(A \setminus \{l'\}, \mathbf{t}) + k''_{l'} \delta_{l'}(\mathbf{t}), \end{aligned} \tag{3.1.26}$$

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:

$$K(A, \mathbf{t}) \leq K + \sum_{l \in A} k''_l \delta_l(\mathbf{t}). \tag{3.1.27}$$

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:

$$l' = \arg\min_{l \in A} \min_{S \subseteq A \setminus \{l\}} \sum_{B \subseteq S} (k'_{B \cup \{l\}} - k'_B). \tag{3.1.28}$$

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

$$C(L) = \sum_{B \subseteq L} k_B, \text{ s.t. } k_B \geq 0. \tag{3.1.29}$$

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

$$E_B(\mathbf{t}) = k'_B \prod_{l \in B} \delta_l(\mathbf{t}) \leq k'_B \sum_{l \in B} \rho_l^B \delta_l(\mathbf{t}), \tag{3.1.30}$$

where $\rho_l^B \geq 0$ and $\sum_{l \in B} \rho_l^B = 1$. In practice, to obtain a symmetrical over-estimation of energy, we set $\rho_l^B = 1/|B|$. The moves for the first order occurrence costs [19] 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 section 3.2).

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 $|A|+1$ auxiliary variables $\mathbf{z}$ as:*

$$\begin{aligned} E'(\mathbf{t}) &= \min_{\mathbf{z}} \Bigg[k'_\alpha (1 - z_\alpha) + \sum_{l \in A} k''_l z_l + \sum_{i \in \mathcal{V}} k'_\alpha (1 - t_i) z_\alpha \\ &+ \sum_{l \in A} \sum_{i \in \mathcal{V}_l} k''_l t_i (1 - z_l) \Bigg], \end{aligned} \tag{3.1.31}$$

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 [19] for the optimisation of binary over-estimated move energy (3.1.22). 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 [19].

3.2 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 the previous chapter 2. On the VOC data set, the baseline also makes use of the detector potentials of [60].

The costs $C(L)$ for the MSRC dataset 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

$$z_l^{(m)} = \delta(l \in L(\mathbf{x}^{(m)})) \tag{3.2.1}$$

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

$$C(L) = -w \log \frac{1}{M} \left(1 + \sum_{m=1}^{M} \prod_{l \in L} z_l^{(m)} \right), \tag{3.2.2}$$

where w is the weight of the co-occurrence potential. The form guarantees that

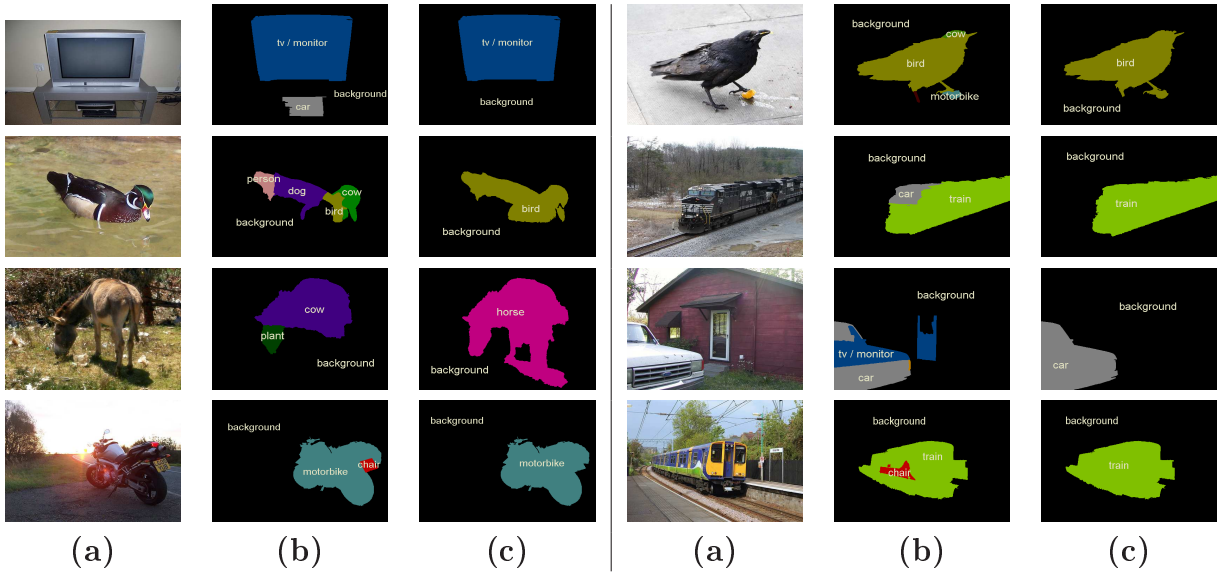


Figure 3.4: **Best viewed in colour: (a)** *Typical images taken from the* VOC-*2009 data set [88];* **(b)** *A labelling based upon a pixel based random field model [59] that does not take into account co-occurrence;* **(c)** *A labelling of the same model using co-occurrence statistics. Note that the co-occurrence potentials perform in a similar way across different data sets, suppressing the smaller classes (see also figure 3.1) 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.*

$C(L)$ is monotonically increasing with respect to L. To avoid over-fitting we approximated the potential $C(L)$ as a second order function:

$$C'(L) = \sum_{l \in L} c_l + \sum_{k,l \in L, k<l} c_{kl}, \tag{3.2.3}$$

where c_l and c_{kl} 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 segment-based CRF and a 1% overall, 2% average per class and 2% in the intersection vs. union measure with the AHCRF.

On the VOC dataset, 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:

$$C(L) = -w \sum_{k<l \in \mathcal{L}} c_{kl}, \tag{3.2.4}$$

where c_{kl} were learnt as:

$$\begin{aligned} c_{kl} &= \min(-\log(\mathbb{P}(k|l) \vee \mathbb{P}(l|k)), T) \\ &= \min(-\log(\mathbb{P}(k|l) + \mathbb{P}(l|k) - \mathbb{P}(k|l)\mathbb{P}(l|k)), T), \end{aligned} \tag{3.2.5}$$

$\mathbb{P}(k|l) = \frac{\mathbb{P}(\{k,l\})}{\mathbb{P}(\{l\})}$, $\mathbb{P}(L) = \frac{\sum_{m=1}^{M} \prod_{l \in L} z_l^{(m)}}{M}$ 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. 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 3.5 and 3.6.

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 [88], the incorporation of co-occurrence potentials leads to a significant improvement in accuracy.

3.3 Conclusion

The importance of co-occurrence statistics is well established [96, 74, 17]. 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

	Global	Average	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
Segment-based CRF	81	66	80	**98**	83	64	81	**99**	59	89	85	68	68	**98**	76	26	85	39	84	30	49	50	07
Segment-based CRF with CO	82	68	81	**98**	83	65	81	**99**	59	91	85	69	68	**98**	76	27	85	39	85	29	49	51	07
Hierarchical CRF	87	78	81	96	**89**	74	**84**	**99**	**84**	92	**90**	86	**92**	**98**	91	**35**	**95**	53	90	62	**77**	**70**	**12**
Hierarchical CRF with CO	**89**	**80**	**83**	96	**89**	**75**	**84**	**99**	**84**	**94**	**90**	**87**	**92**	**98**	**92**	**35**	**95**	**55**	**91**	**64**	**77**	**70**	11

Figure 3.5: *Quantitative results on the MSRC data set. The table shows % pixel accuracy $N_{ii}/\sum_j N_{ij}$ for different object classes. 'Global' refers to the overall error $\frac{\sum_{i\in\mathcal{L}} N_{ii}}{\sum_{i,j\in\mathcal{L}} N_{ij}}$, while 'average' is $\sum_{i\in\mathcal{L}} \frac{N_{ii}}{|\mathcal{L}|\sum_{j\in\mathcal{L}} N_{ij}}$. N_{ij} refers to the number of pixels of label i labelled j.*

	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	78	38	10	**24**	36	**31**	59	37	21	08	**02**	**23**	**14**	17	27	21	15	**16**	15	48	33
Hierarchical CRF with CO	**31**	**82**	**49**	**12**	19	**38**	**31**	**63**	**46**	**24**	**10**	01	**23**	**14**	**22**	**34**	**36**	**18**	12	**23**	**53**	**37**

Figure 3.6: *Quantitative analysis of VOC2009 results on validation set, intersection vs. union measure, defined as $\frac{True\ Positive}{True\ Positive + False\ Negative + False\ Positive}$. 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. For recent result see chapter 4.*

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. 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.

Appendix

Lemma 1 Proof. First we show that:

$$
\begin{aligned}
E_\alpha(\mathbf{t}) &= \min_{z_\alpha}[(C_{\alpha\beta} - C_\beta)(1 - z_\alpha) + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha\beta} - C_\beta)(1 - t_i)z_\alpha] \\
&= \begin{cases} 0 & \text{if } \forall i \in \mathcal{V}_{\alpha\beta} : t_i = 1, \\ C_{\alpha\beta} - C_\beta & \text{otherwise .} \end{cases}
\end{aligned} \tag{3.3.1}
$$

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:

$$
\begin{aligned}
E_\beta(\mathbf{t}) &= \min_{z_\beta}[(C_{\alpha\beta} - C_\alpha)z_\beta + \sum_{i \in \mathcal{V}_{\alpha\beta}} (C_{\alpha,\beta} - C_\alpha)t_i(1 - z_\beta)] \\
&= \begin{cases} 0 & \text{if } \forall i \in \mathcal{V}_{\alpha\beta} : t_i = 0, \\ C_{\alpha\beta} - C_\alpha & \text{otherwise .} \end{cases}
\end{aligned} \tag{3.3.2}
$$

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 [51]. □

See figure 3.3 for graph construction.

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

$$
\begin{aligned}
E_\alpha(\mathbf{t}) &= \min_{z_\alpha} \left[k'_\alpha(1 - z_\alpha) + \sum_{i \in \mathcal{V}} k'_\alpha(1 - t_i)z_\alpha \right] \\
&= \begin{cases} k'_\alpha & \text{if } \exists i \in \mathcal{V} \text{ s.t. } t_i = 0, \\ 0 & \text{otherwise .} \end{cases}
\end{aligned} \tag{3.3.3}
$$

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 $l \in A$:

$$\begin{aligned} E_b(\mathbf{t}) &= \min_{z_l} \left[k_l'' z_l + \sum_{i \in \mathcal{V}_l} k_l'' t_i (1 - z_l) \right] \\ &= \begin{cases} k_l'' & \text{if } \exists i \in \mathcal{V}_l \text{ s.t. } t_i = 1, \\ 0 & \text{otherwise .} \end{cases} \end{aligned} \tag{3.3.4}$$

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 3.3 for graph construction.

Chapter 4

Latent Random Field SVMs for Object Detection

Object detection is typically formulated as a problem where the objective is to find all instances of objects of a given class and enclose each one of them by a tight bounding box. Several methods [14, 102, 39] follow the *bag-of-words* (BOW) approach designed for the scene classification problem and are learnt using support vector machines (SVM) and evaluated on the sliding window across the image. These methods get good results on classification of boxes but struggle to localize objects exactly.

Dalal and Triggs [18] proposed a method to deal with this problem using the classifier learnt directly on raw non-clustered features - histograms of oriented gradients (HOG) over cells composing the bounding box, efficiently matching object shape with the learnt rigid template of edge directions. This method was originally applied to pedestrian detection, but it turned out to be competitive with other methods for a wide range of object classes with distinctive shapes. On the other hand it struggled on data sets containing images with large intra-class variability or images taken from varying view points.

To overcome this problem, Felzenszwalb *et al.* [25] proposed a star-shaped part based model allowing a predetermined number of rigid parts to change their relative location with respect to the centre of the object. Star-shaped models, or pictorial structures, have a long history in vision [29, 23, 106, 26, 57]. However, the contribution of [25] was in the learning. They formulated the problem as a latent SVM [1], which is a subclass of structured SVMs [98, 92], learning both the weights of the classifier and the location of rigid object parts as latent variables. Large intra-class variance was modelled by splitting training samples based on their aspect ratio and training a classifier for different aspect ratios independently.

Motivated by this work, we propose a new latent variable SVM allowing for any deformations of the template, expressed in terms of a deformation field. Rather than restrict ourselves to a star-shaped model, we take inspiration from recent advances in convex and biconvex models [46, 33], that show inference in these models can be very efficient, to make a lattice-connected part-based model. Furthermore we show how to learn several models not expressible using only local deformations for the case where one model is not enough. We propose tractable optimisation for learning parameters of the model and for evaluation.

4.1 Previous Work

First we describe the formulation of Dalal and Triggs [18]. The linear support vector machine (SVM) classifier response for a given image sub-window is based on histograms of oriented gradients (HOG) evaluated on a regular grid of $n = n_x \times n_y$ (in general overlapping) cells, where each cell is a rectangular region of a fixed size $S_x \times S_y$ centred at the point $c_i = [x_i, y_i]$. Let $\mathbf{h}(c_i)$ be the corresponding histograms of gradients with m directions over the cell, centred at the point c_i, and $\mathbf{h}(\mathbf{c})$ the concatenated histograms over all cells. The linear discriminant function takes the form :

$$H(\mathbf{c}) = \mathbf{w}^* \cdot \mathbf{h}(\mathbf{c}) + b^* = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^* h_j(c_i) + b^*, \tag{4.1.1}$$

where $H(\mathbf{c}) > 0$ indicates a positive detection, negative otherwise. The weights $\mathbf{w}^*$ and bias b^* are trained by solving the optimization problem using M training samples with ground truth labels $z^k \in \{-1, 1\}$ as:

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda ||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1..M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \left(\mathbf{w}.\mathbf{h}(\mathbf{c}^k) + b \right), \end{aligned} \tag{4.1.2}$$

where $\mathbf{h}(\mathbf{c}^k)$ are the concatenated histograms of k-th training sample and λ_w is the regularisation strength.

The rigid formulation can be extended [25] to a more flexible one using MI-SVM [1] or Latent SVM [25], two equivalent formulations, that were discovered independently. The classifier takes the form:

$$H(\mathbf{x}) = \max_{\mathbf{z} \in \mathbf{Z}(\mathbf{x})} \mathbf{w}^* \cdot \Phi(\mathbf{x}, \mathbf{z}) + b, \tag{4.1.3}$$

where $\mathbf{z}$ is the set of latent variables, $\mathbf{Z}(\mathbf{x})$ their possible set of states, and $\Phi(\mathbf{x}, \mathbf{z})$

the feature vector. The model is typically trained by alternating between estimating the state of latent variables given the weight vector $\mathbf{w}^*$ and estimating optimal weights $\mathbf{w}^*$ and bias b given the state of latent variables.

4.2 Deformable Template with MRF Priors

In this work we are inspired by recent advances in biconvex models for optical flow and propose a model that allows the deformation of an object using deformation field $\mathbf{d} = [\mathbf{d^x}, \mathbf{d^y}]$ containing an optic flow like deformation parameters $[d_i^x, d_i^y]$ for each cell centre. This can be thought of a set of latent variables (as in the Latent CRF). However, the form of prior we shall choose will be much richer than in [25] and a generalisation in that we will allow for a general pairwise convex CRF to form the latent field.

Formally, the deformation can be defined by a deformation function $D^{\mathbf{d}}(\mathbf{c})$ transforming each cell centre relative to its size as :

$$D^{d_i}(c_i) = D^{d_i}([x_i, y_i]) = [x_i + d_i^x S_x, y_i + d_i^y S_y]. \tag{4.2.1}$$

We restrict deformations d_i^x and d_i^y to the interval $\mathcal{L}^x = (-d_{max}^x, d_{max}^x)$ and $\mathcal{L}^y = (-d_{max}^y, d_{max}^y)$ respectively. The deformation field can be trivially extended to allow any scale or affine deformation. However, this would increase the complexity of the optimisation and evaluation of the classifier.

The deformation field $\mathbf{d}$ is treated as a set of latent variables jointly estimated with the parameters (weights and bias) of the SVM classifier. To penalise improbable deformations the regularisation term is introduced.

The classifier for our deformable template then takes the form :

$$H(\mathbf{c}) = \max_{\mathbf{d}} \left(\mathbf{w}^*.\mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) + b^* - R(\mathbf{d}) \right), \tag{4.2.2}$$

where $R(\mathbf{d})$ is the regularisation term for deformation field $\mathbf{d}$. The regularisation

term takes the form of the pairwise Markov Random Field (MRF) cost :

$$R(\mathbf{d}) = \theta_u \sum_{i=1}^{n} \psi_u(|d_i|) + \theta_p \sum_{i=1}^{n} \sum_{l \in N_i} \psi_p(|d_i - d_l|), \tag{4.2.3}$$

where N_i is the neighbourhood of i-th cell and ψ_u is the unary potential favouring lower deformations and ψ_p pairwise potential enforcing neighbouring patches to take similar deformations and also guaranteeing that a change in cell ordering becomes very improbable. The unary potential ψ_u can be any nondecreasing function with respect to the deformation. The pairwise potential $\psi_p(.)$ is a convex function over an ordered set [46].

The optimisation problem for learning the weights $\mathbf{w}^*$ and the bias b^* becomes:

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1..M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \max_{\mathbf{d}} \left(\mathbf{w}.\mathbf{h}(D^{\mathbf{d}}(\mathbf{c}^k)) + b - R(\mathbf{d})\right). \end{aligned} \tag{4.2.4}$$

So far we have considered $\mathbf{h}(\cdot)$ to be just HOGs. However, the discriminative power of the classifier can be increased by the incorporation of the *bag-of-words* model (BOW) using visual words [14, 102, 39].

We shall use a hierarchical structure with multiple layers of cells at different resolutions forming a spatial pyramid [63], where each cells is connected to its neighbours on the same layer and to its parent and children. We use the same inconsistency cost on deformation between parent and child as the pairwise cost between neighbouring cells on the same layer.

4.3 Learning the Parameters of the Deformable Model

The optimisation problem (4.2.4) for the training stage is non-convex. However, in this section we show, that if the pairwise regularisation cost is convex over an

ordered set of discrete deformations [46], the whole optimisation problem is tri-convex with respect to the weights $\mathbf{w}$ with the bias b, deformation field component $\mathbf{d_x}$ and deformation component $\mathbf{d_y}$. That means, that given two of these three components, estimation of the third one is convex. Thus, we can approximately solve the problem by repeatedly fixing two components and estimating the third.

First, the problem of finding the optimal weight vector $\mathbf{w}$ and bias b, given the deformation fields $\hat{\mathbf{d}}^k$ for each training example becomes:

$$\begin{aligned}(\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda_w ||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1..M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \left(\mathbf{w} \cdot \mathbf{h}(D^{\hat{\mathbf{d}}^k}(\mathbf{c})) + b - R(\hat{\mathbf{d}}^k)\right)\end{aligned} \tag{4.3.1}$$

and can be solved using any standard SVM algorithm [7, 85, 47]. The problem of finding the optimal deformation field $\mathbf{d}^*$ for each training example given current weights $\hat{\mathbf{w}}^k$ becomes:

$$\begin{aligned}\mathbf{d}^* &= \arg\max_{\mathbf{d}} \left(\hat{\mathbf{w}}^k \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) - R(\mathbf{d})\right) \\ &= \arg\min_{\mathbf{d}} \sum_{i=1}^{n} \psi_u(|d_i|) + \sum_{i=1}^{n} \sum_{l \in N_i} \psi_p(|d_i - d_l|) \\ &- \hat{\mathbf{w}}^k.\mathbf{h}(D^{\mathbf{d}}(\mathbf{c})).\end{aligned} \tag{4.3.2}$$

The last term can be decomposed into functions of deformations d_i for each cell which do not depend on each other as:

$$\hat{\mathbf{w}}^k.\mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) = \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{w}_{ij} h_j(D^{d_i}(c_i)). \tag{4.3.3}$$

By defining $\psi_d(d_i) = \sum_{j=1}^{m} \hat{w}_{ij} h_j(D^{d_i}(c_i))$ the optimisation procedure to find the optimal deformation field becomes :

$$\mathbf{d}^* = \arg\min_{\mathbf{d}} \sum_{i=1}^{n} (\psi_u(|d_i|) - \psi_d(d_i)) + \sum_{i=1}^{n} \sum_{l \in N_i} \psi_p(|d_i - d_l|),$$

which is the standard max-a-posteriori (MAP) estimation of the pairwise MRF

problem with $|\mathcal{L}^x||\mathcal{L}^y|$ labels. This problem can be solved by estimating $\mathbf{d}_x$ and $\mathbf{d}_y$ iteratively using graph cut [10] with Ishikawa's graph construction [46].

The whole optimisation procedure starts by initializing the deformation field equal to zero and iteratively estimating the optimal weights and the bias given the deformation field of each training sample and vice versa. Each step is guaranteed not to increase the objective function (4.2.4) and thus the iterative procedure has to converge. The optimisation problem during the testing phase is equivalent to the second part of the training phase and is solved in a similar manner.

4.4 Kernelising the Deformable Template Model

The deformable template model can be kernelised using any positive definite kernel $K(x, y)$. According to the representer's theorem, for every positive definite kernel $K(x, y)$ there exist a function $\Phi(x)$, in general infinite dimensional, such that $K(x, y) = \Phi(x)\Phi(y)$. The classifier then becomes:

$$H(\mathbf{c}) = \max_{\mathbf{d}} \left(\mathbf{w}^*.\Phi(\mathbf{h}(D^{\mathbf{d}}(\mathbf{c}))) + b^* - R(\mathbf{d}) \right). \tag{4.4.1}$$

The optimisation takes the form:

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1..M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \max_{\mathbf{d}} \left(\mathbf{w}.\Phi(\mathbf{h}(D^{\mathbf{d}}(\mathbf{c}^{\mathbf{k}}))) + b - R(\mathbf{d}) \right). \end{aligned} \tag{4.4.2}$$

Even though for many kernels the finite dimensional approximation [103, 68] of the mapping function $\Phi(.)$ can be found, for many other kernels there is no good approximation. However, for the standard kernel SVM it can be shown [36] that the solution can be found in the form $H(\mathbf{c}) = \sum_{k=1}^{M} \alpha_k K(\mathbf{h}(\mathbf{c}), \mathbf{h}(\mathbf{c}^k))$, where $\mathbf{h}(\mathbf{c}^k)$ are the training samples, which for nonzero α_k are called support vectors. As we show next, for our deformation template model these will correspond to deformed training samples.

Given a deformation field $\mathbf{d}$ the optimisation problem becomes a standard kernel SVM. The classifier then becomes a function of deformed training samples:

$$H(\mathbf{c}) = \max_{\mathbf{d}} \left(\sum_{k=1}^{M} \alpha_k K(\mathbf{h}(D^{\mathbf{d}}(\mathbf{c})), \mathbf{h}^S(\mathbf{c}^k)) + b^* - R(\mathbf{d}) \right), \tag{4.4.3}$$

where $\mathbf{h}^S(\mathbf{c}^k) = \mathbf{h}(D^{\hat{\mathbf{d}}^k}(\mathbf{c}^k))$ is k-th training sample deformed by $\hat{\mathbf{d}}^k$.

The estimation of the deformation field given a classifier $H(\mathbf{c})$ becomes:

$$\begin{aligned} \mathbf{d}^* &= \arg\min_{\mathbf{d}} \sum_{i=1}^{n} \psi_u(|d_i|) + \sum_{i=1}^{n} \sum_{l \in N_i} \psi_p(|d_i - d_l|) \\ &- \sum_{k=1}^{M} \hat{\alpha} K(\mathbf{h}(D^{\mathbf{d}}(\mathbf{c})), \mathbf{h}^S(\mathbf{c}^k)), \end{aligned}$$

where $\hat{\alpha} K(\mathbf{h}(D^{\mathbf{d}}(\mathbf{c})), \mathbf{h}^S(\mathbf{c}^k))$ is a higher order potential over all nodes in the graph. This makes the inference problem intractable. If we restrict ourselves to kernel functions decomposable to a weighted sum of kernel functions $K_i(\cdot, \cdot)$ over each cell:

$$K(\mathbf{h}(\mathbf{c}), \mathbf{h}^S(\mathbf{c}^k)) = \sum_{i=1}^{n} \beta_i K_i(\mathbf{h}(c_i), \mathbf{h}^S(c_i^k)), \tag{4.4.4}$$

where $K_i(\cdot, \cdot)$ can be an arbitrary positive definite kernel over bins within the same cell, we can still solve the optimisation problem of finding the optimal $\mathbf{d}^*$ efficiently the same way as for the linear kernel. The property (4.4.4) is trivially satisfied for any additive kernel $K(\cdot, \cdot)$, such as intersection or quasi-linear χ^2-kernel. More general kernel functions $K_i(\cdot, \cdot)$ can be useful for large cells in the spatial pyramid with histograms of visual words [14, 102, 39] and the relative weights β_i can be learnt using Multiple kernel learning [101] as in [102].

4.5 Learning of Different Viewpoints or Poses

The proposed deformation model can not deal with large changes of view point or pose. This problem can be treated by splitting each class into several sub-classes, each representing different view, aspect ratio or pose, and then training classifiers for each subclass (mode) independently [25]. Positive detections are given if any of the trained detector responses are above a certain threshold. This is equivalent

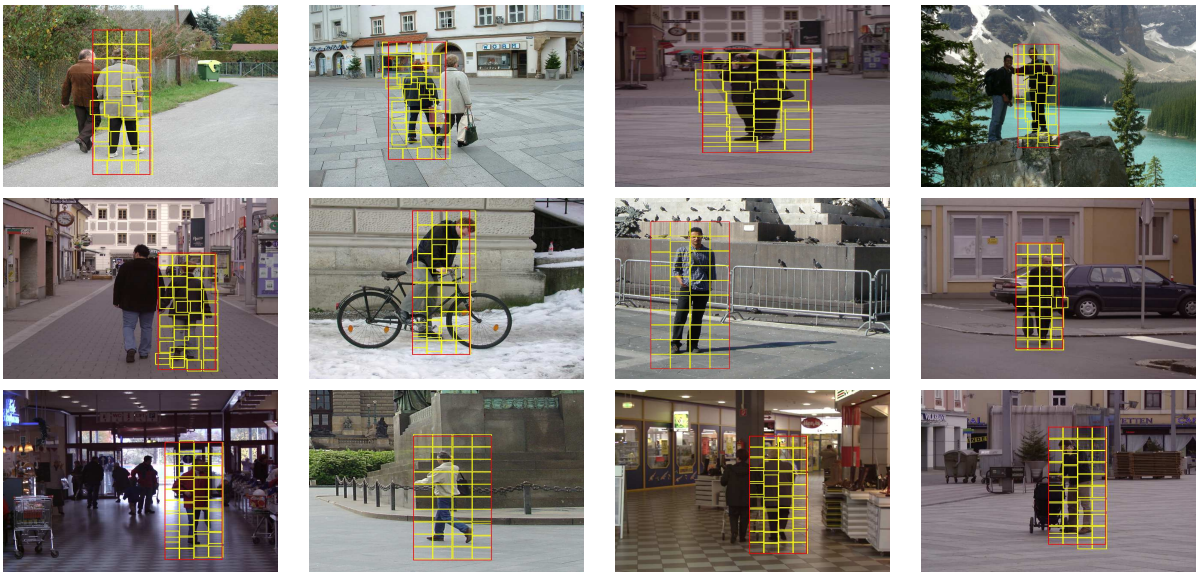


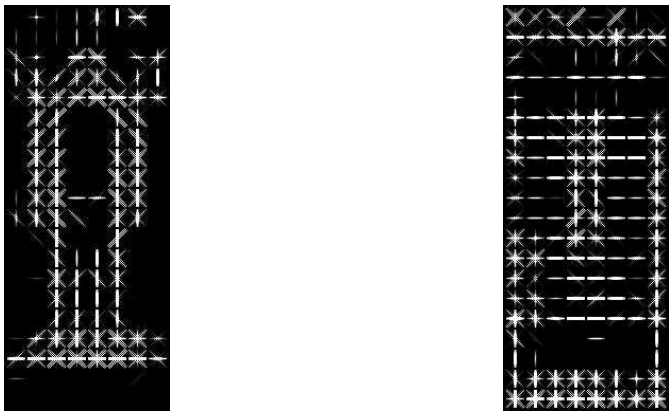
Figure 4.1: *Qualitative results on the INRIA data set. Red boxes are the root bounding boxes and yellow boxes are individual cells on the base layer. The deformation field tries to align the box to both fit the data while keeping locally consistent structure. Good localisation of the person with the original bounding box typically results in lower defomations, while the data-fitting term becomes more important if the root box is not sufficiently good. Only the strongest response for each image is displayed to avoid confusion.*

to the classifier response defined as :

$$H(\mathbf{c}) = \max_t H^t(\mathbf{c}). \tag{4.5.1}$$

where $H^t(\mathbf{c})$ is the t-th classifier response. This approach can not recover from the wrong initial split of the data. Thus instead of fixing the split, we jointly estimate a set of classifiers and the assignment of samples to classifiers.

If the number of models is too high, the amount of data may become insufficient and training may over-fit the data. To deal with this kind of problem we take advantage of feature sharing between models. Intuitively, some of the features, *e.g.* histograms of dense features, tend to be shared between different models of the same object class (whilst some other features *e.g.* HOG feature are not). To induce feature sharing we introduce a regularisation term between the weights of the models of a given class and in the case of a linear kernel the



a) b)

Figure 4.2: *Trained* SVM *weights ((a) positive, b) negative) of the base layer of the deformable template for object detection of a person on INRIA data set using one model. High positive weight implies that the edge is important for successful detection. High negative weight means, that the existence of such edge suggests the object of interest is not present.*

optimisation problem takes the form:

$$
\begin{aligned}
(\mathbf{w}^*, \mathbf{b}^*) &= \arg\min_{(\mathbf{w},\mathbf{b})} \lambda_w \sum_t ||\mathbf{w}^t||^2 + \sum_{k=1}^{M} \xi^k + \lambda_s \sum_{t,t' \neq t} ||P(\mathbf{w}^t) - P(\mathbf{w}^{t'})||_1 \quad (4.5.2) \\
\text{s.t. } \forall k &\in \{1..M\} : \\
\xi^k &\geq 0 \\
\xi^k &\geq 1 - z^k \max_{t,\mathbf{d}^t} \left(\mathbf{w}^t . \mathbf{h}(D^{\mathbf{d}^t}(\mathbf{c}^k)) + b^t - R(\mathbf{d}^t) \right),
\end{aligned}
$$

where $\mathbf{w}$ is the concatenated vector of weights of each model $\mathbf{w}^t$, $\mathbf{b}$ the vector of biases b^t, λ_s the strength of regularisation between models and $P(\mathbf{w}^t)$ projection into the subset of the weights on which we would like to induce the feature sharing. Equivalently we can formulate the optimisation problem for the kernel version. To solve this optimisation problem we can use stochastic gradient decent [7, 85] in a similar fashion as is applied to the multi-class SVM formulation [16] in [85].

4.6 Object Detectors in CRFs for Object Class Segmentation

Object detections contain information, such as shape, which are not included in the traditional CRFs for object class segmentation. Thus, incorporation of these cues should lead to the improvement of the segmentation method. Another problem of CRFs is the impossibility of recovering different instances of an object class. In this section we propose a new formulation that jointly estimates the pixel labelling and all instances of objects of each class.

MAP estimation can be understood as a soft competition among different hypotheses (defined over pixel or segment random variables), in which the final solution maximizes the weighted agreement between them. These weighted hypotheses can be interpreted as potentials in the CRF model. In object class recognition, these hypotheses encourage: (i) variables to take particular labels (unary potentials), and (ii) agreement between variables (pairwise, hierarchical). In this section we introduce an additional set of hypotheses representing object detections for the recognition framework.

Some of the object detection approaches [25, 62] have used their results to perform a segmentation within the detected areas[1]. This approach would include both the true and false positive detections, and segment them assuming they all contain the objects of interest. There is no way of recovering from these erroneous segmentations. Our approach overcomes this issue by using the detection results only as hypotheses that can be rejected in the global CRF energy. In other words, all detections act as soft constraints in our framework, and must agree with other cues from pixels and segments before affecting the object class segmentation result.

Let $\mathcal{D}$ denote the set of object detections, which are represented by bounding boxes enclosing objects, and corresponding scores that indicate the strength of the detections. We define a novel clique potential ψ_d over the set of pixels $\mathbf{x}_d$ belonging to the d-th detection (*e.g.* pixels within the bounding box), with a score H_d and detected label l_d. Figure 4.3 shows the inclusion of this potential

[1] As evident in some of the PASCAL VOC 2009 segmentation challenge entries.

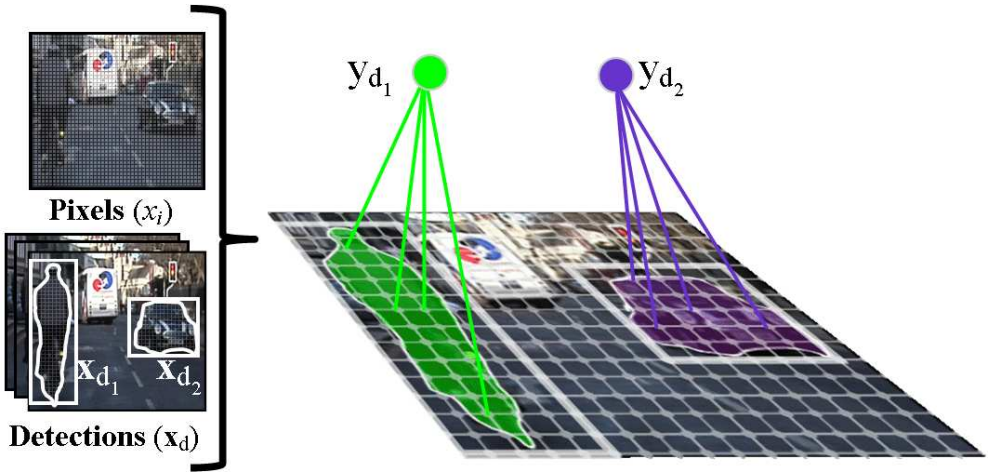


Figure 4.3: *Inclusion of object detector potentials into a* CRF *model. We show a pixel-based* CRF *as an example here. The set of pixels in a detection* d_1 *(corresponding to the bicyclist in the scene) is denoted by* $\mathbf{x}_{d_1}$*. A higher order clique is defined over this detection window by connecting the object pixels* $\mathbf{x}_{d_1}$ *to an auxiliary variable* $y_{d_1} \in \{0, 1\}$*. This variable allows the inclusion of detector responses as soft constraints. (**Best viewed in colour**)*

graphically on a pixel-based CRF. The new energy function is given by:

$$E(\mathbf{x}) = E_{pix}(\mathbf{x}) + \sum_{d \in \mathcal{D}} \psi_d(\mathbf{x}_d, H_d, l_d), \tag{4.6.1}$$

where $E_{pix}(\mathbf{x})$ is any standard pixel-based energy. The minimization procedure should be able to reject false detection hypotheses on the basis of other potentials (pixels and/or segments). We introduce an auxiliary variable $y_d \in \{0, 1\}$, which takes value 1 to indicate the acceptance of d-th detection hypothesis. Let ϕ_d be a function of this variable and the detector response. Thus the detector potential $\psi_d(.)$ is the minimum of the energy values provided by including $(y_d = 1)$ and excluding $(y_d = 0)$ the detector hypothesis, as given below:

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d} \phi_d(y_d, \mathbf{x}_d, H_d, l_d). \tag{4.6.2}$$

We now discuss the form of this function $\phi_d(\cdot)$. If the detector hypothesis is included $(y_d = 1)$, it should: (a) encourage consistency by ensuring that labellings where all the pixels in $\mathbf{x_d}$ take the label l_d should be more probable, *i.e.* the associated energy of such labellings should be lower; (b) be robust to partial inconsistencies, *i.e.* pixels taking a label other than l_d in the detection window.

Such inconsistencies should be assigned a cost rather than completely disregarding the detection hypothesis. The absence of the partial inconsistency cost will lead to a hard constraint where either all or none of the pixels in the window take the label l_d. This allows objects partially occluded to be correctly detected and labelled.

To enable a compact representation, we choose the potential ψ_d such that the associated cost for partial inconsistency depends only on the number of pixels $N_d = \sum_{i \in \mathbf{x_d}} \delta(x_i \neq l_d)$ disagreeing with the detection hypothesis. Let $f(\mathbf{x}_d, H_d)$ define the strength of the hypothesis and $g(N_d, H_d)$ the cost taken for partial inconsistency. The detector potential then takes the form:

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d} \phi_d(y_d, \mathbf{x}_d, H_d, l_d) = \min_{y_d}(-f(\mathbf{x}_d, H_d)y_d + g(N_d, H_d)y_d), \tag{4.6.3}$$

where $\phi_d(y_d, \mathbf{x}_d, H_d, l_d) = -f(\mathbf{x}_d, H_d)y_d + g(N_d, H_d)y_d$.

A stronger classifier response H_d indicates an increased likelihood of the presence of an object at a location. This is reflected in the function $f(\cdot)$, which should be monotonically increasing with respect to the classifier response H_d. As we also wish to penalize inconsistency, the function $g(\cdot)$ should be monotonically increasing with respect to N_d. The number of detections used in the CRF framework is determined by a threshold H_t. The hypothesis function $f(\cdot)$ is chosen to be a linear truncated function using H_t as:

$$f(\mathbf{x}_d, H_d) = w_d|\mathbf{x}_d| \max(0, H_d - H_t), \tag{4.6.4}$$

where w_d is the detector potential weight. This ensures that $f(\cdot) = 0$ for all detections with a response $H_d \leq H_t$. We choose the inconsistency penalizing function $g(\cdot)$ to be a linear function on the number of inconsistent pixels N_d of the form:

$$g(N_d, H_d) = k_d N_d, \tag{4.6.5}$$

where the slope k_d was chosen such that the inconsistency cost equals $f(\cdot)$ when the percentage of inconsistent pixels is p_d, and is given by:

$$k_d = \frac{f(\mathbf{x}_d, H_d)}{p_d|\mathbf{x}_d|}. \tag{4.6.6}$$

Sliding window detectors detect only a bounding box and not the exact set of pixels $\mathbf{x}_d$ belonging to the object. We follow the approach used by submissions in the PASCAL VOC 2009 segmentation challenge and explicitly identify which regions of the box are likely to belong to the object. This provides us a more precise set of object pixels. We can either estimate the foreground and background using local colour model [78] or with pose-based approaches [77] that use a generatively trained likelihood of pixels belonging to the foreground depending on the relative location of pixels within the box. If detectors estimate foreground pixels themselves [65], they may be applied directly. Note that equation (4.6.1) could be defined in a similar fashion over superpixels.

4.7 Inference for Detector Potentials

We now show that our detector potential in equation (4.6.3) can be converted into a form solvable using α-expansion algorithms [11]. In contrast, the related work in [35] suffers from a difficulty to optimize energy. The detector potential $\psi_d(\cdot)$ can be rewritten as follows:

$$\begin{aligned} \psi_d(\mathbf{x}_d, H_d) &= \min_{y_d \in \{0,1\}} (-f(\mathbf{x}_d, H_d) y_d + k_d N_d y_d) \\ &= \min_{y_d \in \{0,1\}} (-f(\mathbf{x}_d, H_d) y_d + k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq d) y_d). \end{aligned} \tag{4.7.1}$$

Now we show that for both cases $\alpha = l_d$ and $\alpha \neq l_d$ the α-expansion move energy can be represented using one auxiliary variable. Similarly to previous chapters, we use the standard transformation function for the α-expansion:

$$T_\alpha(x_i, t_i) = \begin{cases} \alpha & \text{if} \quad t_i = 0 \\ x_i & \text{if} \quad t_i = 1. \end{cases} \tag{4.7.2}$$

Consider the case where $\alpha \neq l_d$. The move energy for the detector potential $\psi_d(\mathbf{x}_d, H_d)$ is:

$$\begin{aligned}\psi_d(\mathbf{t}, y_d) &= -f(\mathbf{x}_d, H_d)y_d + k_d \sum_{i \in \mathbf{x}_d} y_d \delta(x_i \neq l_d) + k_d \sum_{i \in \mathbf{x}_d} \delta(x_i = l_d)(1 - t_i)y_d \\ &= -f'(\mathbf{x}_d, H_d)y_d + k_d \sum_{i \in \mathbf{x}_d} \delta(x_i = l_d)(1 - t_i)y_d, \end{aligned} \tag{4.7.3}$$

where $f'(\mathbf{x}_d, H_d) = f(\mathbf{x}_d, H_d) - k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)$. The move energy is directly in the form of a pairwise submodular function and thus can be solved using graph cut.

For the second case where $\alpha = l_d$ we use the encoding $\bar{y_d} = 1 - y_d$. The move energy for the detector potential $\psi_d(\mathbf{x}_d, H_d)$ is:

$$\psi_d(\mathbf{t}, \bar{y_d}) = -f(\mathbf{x}_d, H_d)(1 - \bar{y_d}) + k_d \sum_{i \in \mathbf{x}_d} t_i(1 - \bar{y_d}). \tag{4.7.4}$$

The move energy is also directly in the form of a pairwise submodular function and thus the optimal α-expansion move can be found using graph cut. In both of these constructions the state of detection y_d can be recovered (in the first case directly, in the second case $y_d = 1 - \bar{y_d}$). If any variable $x_i = \alpha$ before the move, it is equivalent to $t_i = 0$ and dropping the corresponding term from the equation (4.7.4). The equivalent graph constructions for both cases are given in figure 4.4.

4.8 Experiments

We tested the deformable template detector on the INRIA person dataset [18]. This dataset contains 1832 training images and 741 test images of pedestrians and cyclists in an urban environment. We formed a three level spatial pyramid of cells with one cell in the top layer, 2×2 on the second and 4×10 on the base layer. The base layer contained 2×2 subcells containing histograms of oriented gradients, normalised by a sum of gradient responses for each direction in a 3×3 window around the cell. Other two layers contained histograms of visual words of SIFT [67] and Local Binary Patterns [71]. We used the linear kernel on the HOG

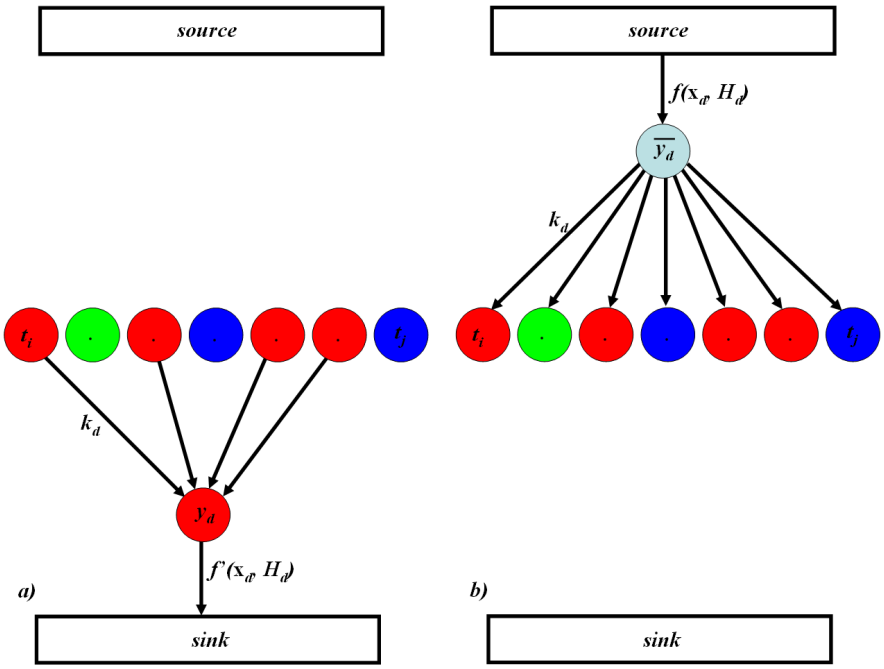


Figure 4.4: *A graph construction for the α-expansion move of the detector potential a) if $l_d \neq \alpha$, b) $l_d = \alpha$. In both of the cases the state of the detection y_d can be recovered. The colour of variables t_i corresponds to the label before the move. The colour of the variable y_d respectively $\bar{y}_d$ corresponds to the label of the detection l_d.*

features and the intersection kernel approximation [102] on the *bag-of-words*. All training boxes were extended by one sixth of their size in height and one third of their size in width to capture edges at the extremities of the object and local context.

In the first round of training, negative samples were picked randomly. In the next two rounds we retrained the model by bootstrapping with hard false positives in the training set [18]. The deformation field was included in the last round of bootstrapping. The positive and negative weights of the trained model are shown in Figure 4.2. During testing the response of the deformable classifier was evaluated on the top 200 windows per image chosen based on the classifier response without deformation. Non-maxima suppression was performed on the resulting set of detections. We tried to train one, and two separate models initialized by clustering feature vectors of the positive samples. Using two models did not lead to significant performance boost. Qualitative results are given in Figure 4.1. We evaluated our performance in recall at a false positive per window

rate (FPPW) of 10^{-4} used also in [18]. We improved our baseline HOG result 87.2% (89% reported in [18]) to 88.1% using two models. However, the deformable template using one model achieved 92.1%, while two models got only 91.5%. These results suggest that the deformable model is partly able to capture small variations between the models. It should be mentioned that the evaluation criteria largely depend on the number of tested windows per image and chosen non-maxima suppression scheme. Thus, the result may not be comparable for different setups and only the relative difference between models matters.

We included the state-of-the-art detectors [102, 25] in the CRF framework for object class segmentation. We tested our framework on the PASCAL VOC 2009 data set. Qualitative results in the intersection vs. union measure on the test set are shown in Figure 4.6. Our approach provides very precise object boundaries and recovers from many failure cases. For example, bird (second row), car (third row), potted plant (fourth row) are not only correctly identified, but also segmented with accurate object boundaries. Quantitative results on this data set are provided in Figure 4.5. We compare our results with the 5 best submissions from the 2009 challenge, and achieve the third best average accuracy. Our method shows the best performance in 3 categories, and a close 2nd/3rd in 10 others.

4.9 Summary

We proposed a new latent SVM for object detection with an MRF prior on the deformation field, that generalises previous work in pictorial structures. We showed how this model can be learnt and optimised efficiently. We showed how to extend this method to multiple models sharing subsets of features. Experimental validation on INRIA data set suggested that the incorporation of the deformation field leads to a quantitative and qualitative improvement of results. We presented a principled way to integrate detectors with the CRF framework, which led to significant improvement of the performance. Unlike many existing methods, our approach supports the robust handling of occluded objects and false detections in an efficient and tractable manner. As a future work we would like to find the

	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
BONN_SVM-SEGM	**36**	**84**	64	22	22	**32**	**40**	**57**	**49**	39	**05**	29	22	20	34	46	34	27	**40**	18	34	**46**
UIUC/CVC_HOCRF	35	80	**67**	**27**	**30**	**32**	30	45	42	25	06	28	11	23	**41**	**53**	32	22	37	**24**	40	30
UOCTTI_LSVM-MDPM	29	79	35	23	19	24	36	41	50	12	09	**29**	01	06	24	35	33	**35**	28	14	34	42
NECUIUC_CLS-DTCT	30	82	42	23	22	22	28	43	52	26	05	19	18	24	27	37	35	09	28	14	36	35
LEAR_SEGDET	26	79	45	16	21	13	29	29	36	25	04	20	01	16	28	30	25	12	31	18	29	32
AH-CRF with Detectors	32	81	46	15	25	21	37	50	44	28	**12**	18	**25**	15	25	38	34	28	30	18	**44**	41

Figure 4.5: *Quantitative analysis of* VOC *2009 test dataset results [22] using the intersection vs union performance measure. Our method is ranked* ***third*** *when compared with the 5 best submissions in the 2009 challenge.*

characterisation of a general class of kernels, for which the corresponding inference problem to estimate the optimal deformation field would still be tractable.

Figure 4.6: *(a) Original test image from* PASCAL VOC *2009 dataset [22], (b) The labelling obtained by [59] without object detectors, (c) The labelling provided by our method which includes detector based potentials. Note that no groundtruth is publicly available for test images in this dataset. Examples shown in the first five rows illustrate how detector potentials not only correctly identify the object, but also provide very precise object boundaries, e.g. bird (second row), car (third row). Some failure cases are shown in the last two rows. This was caused by either incorrect grab-cut solution, by a missed detection or incorrect detections that are very strong and dominate all the other potentials.* ***(Best viewed in colour)***

Chapter 5

Joint Object Class Segmentation and Dense Stereo Reconstruction

The problems of object class segmentation [88, 59], which assigns an object label such as *road* or *building* to every pixel in the image and dense stereo reconstruction, in which every pixel within an image is labeled with a disparity [54], are well suited for being solved jointly. Both approaches formulate the problem of providing a correct labelling of an image as one of Maximum a Posteriori (MAP) estimation over a Conditional Random Field (CRF) [61], which is typically a Potts or truncated linear model. Thus both may use graph cut based move making algorithms, such as α-expansion [11], to solve the labelling problem. These problems *should* be solved jointly, as a correct labelling of object class can help depth labelling, and stereo reconstruction can improve object labelling. Indeed it opens the possibility for the generic stereo priors used previously to be enriched by information about the shape of specific objects. For instance, object class boundaries are more likely to occur at a sudden transition in depth and vice versa, while the height of a point above the ground plane is an extremely informative cue regarding its object class label; e.g. *road* or *sidewalk* lie in the ground plane, and pixels taking labels *pedestrian* or *car* must lie at a constrained height above the ground plane, while pixels taking label *sky* must occur at an infinite depth (zero disparity) from the camera. Figure 5.1 shows our model which explicitly captures these properties.

Object recognition provides substantial information about the 3D location of points in the image. This has been exploited in recent work on single view reconstruction [42, 75, 34, 66], in which a plausible pop-up planar model of a scene is reconstructed from a single monocular image using object recognition and prior information regarding the location of objects in typically photographed scenes. Such approaches only estimate depth from object class, assuming the object class is known. As object recognition is itself a problem full of ambiguity and often requiring knowledge of 3D such a two stage process must, in many cases, be suboptimal.

Other works have taken the converse approach of using 3D information in inferring object class; [44] showed how knowledge of the camera viewpoint and the typical 3D location of objects can be used to improve object detection, while [64] employed Structure-from-Motion (*SfM*) techniques to aid the tracking and

detection of moving objects. However, neither object detection nor the 3D reconstruction obtained gave a dense labelling of every pixel in the image, and the final results in tracking and detection were not used to refine the *SfM* results. The CamVid [12] data set provides sparse *SfM* cues, which have been used by several object class segmentation approaches [12, 89] to generate pixel based image labelling. In these the object class segmentation was not used to refine the 3D structure.

Previous works have attempted to simultaneously solve the problems of object class detection and 3D reconstruction. [45] fitted a 3D model to specific objects, such as *buses* or *cars* within an image by simultaneously estimating 3D location, orientation and object class, while [20] fitted a 3D model of a building to a set of images by simultaneously estimating a wire-frame model and the location of assets such as *window* or *column*. In both of these papers the 3D models are intended to be plausible rather than accurate, and these models are incomplete, they do not provide location or class estimates of every pixel.

None of the discussed works perform joint inference to obtain dense stereo reconstruction and object class segmentation. In this work, we demonstrate that these problems are mutually informative, and benefit from being solved jointly. We consider the problem of scene reconstruction in an urban area [64]. These scenes contain object classes such as *road*, *car* and *sky* that vary in their 3D locations. Compared to typical stereo data sets that are usually produced in controlled environments, stereo reconstruction on this real world data is noticeably more challenging due to large homogeneous regions and problems with photo-consistency. We efficiently solve the problem of joint estimation of object class and depth using modified variants of the α-expansion [11], and range move algorithms [104, 56].

No real world data sets are publicly available that contain both per pixel object class and dense stereo data. In order to evaluate our method, we augmented the data set of [64] by creating hand labeled object class and disparity maps for 70 images. These annotations have been made available for download[1]. Our experimental evaluation demonstrates that joint optimisation of dense stereo re-

[1] `http://cms.brookes.ac.uk/research/visiongroup/files/Leuven.zip`

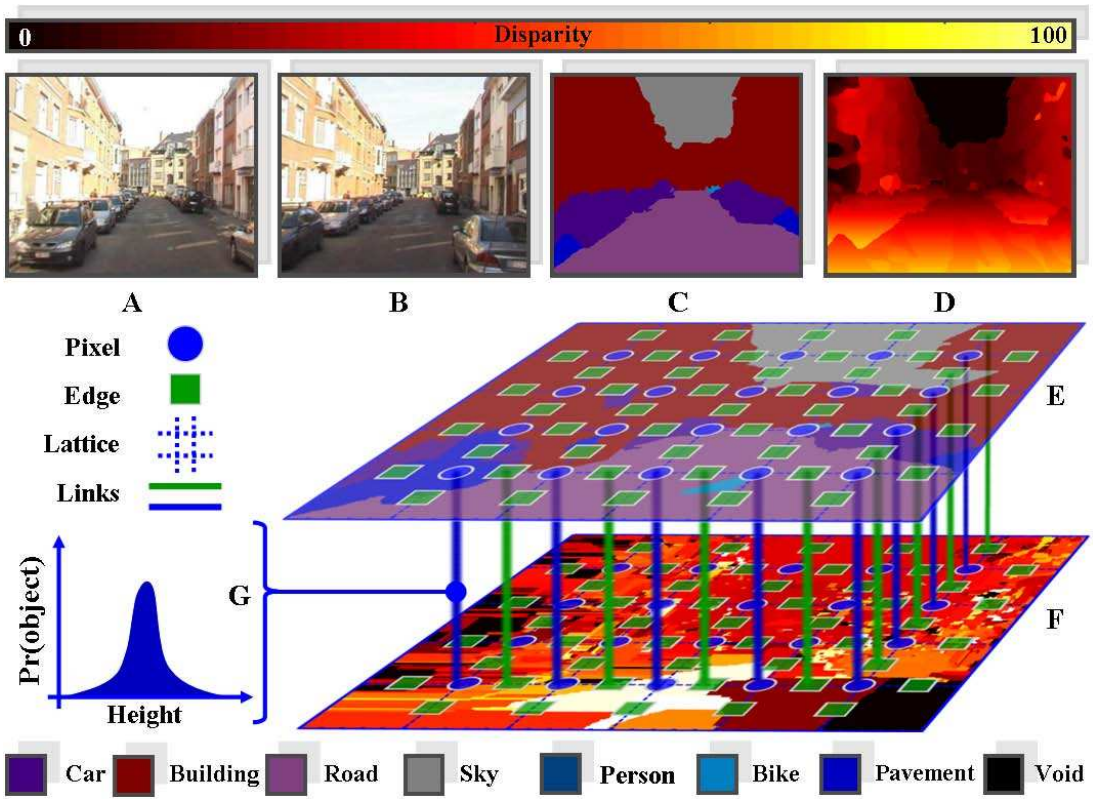


Figure 5.1: *Graphical model of our joint* CRF. *The system takes a left (***A***) and right (***B***) image from a stereo pair that has been rectified. Our formulation captures the dependencies between the object class segmentation problem (***E***, §5.1.1) and the dense stereo reconstruction problem (***F***, §5.1.2) by defining a joint energy on the recognition and disparity labels both on the unary/pixel (blue) and pairwise/edge variables (green) of both problems. The unary potentials of the joint problem encodes the fact that different objects will have different height distributions (***G***,eq. (5.2.1)) learned from our training set containing hand labeled disparities (§5.4). The pairwise potentials encode that object class boundaries, and sudden changes in disparity are likely to occur together, but could also encode different shape smoothness priors for different types of object. The combined optimisation results in an approximate object class segmentation (***C***) and dense stereo reconstruction (***D***). See §5.2 and §5.3 for a full treatment of our model and §5.5 for further results. Best viewed in colour.*

construction and object class segmentation leads to a substantial improvement in the accuracy of the final results.

The structure of the chapter is as follows: In section 5.1 we give the generic formulation of CRFs for dense image labelling, and describe how they can be applied to the problems of object class segmentation and dense stereo reconstruction. Section 5.2 describes the formulation allowing for the joint optimisation of these two problems, while section 5.3 shows how the optimisation can be performed efficiently. The data set is described in section 5.4 and experimental validation follows in 5.5.

5.1 Overview of Dense CRF Formulations

Our joint optimisation consists of two parts, object class segmentation and dense stereo reconstruction. Before we formulate our approach we give an overview of the typically used random field formulation for both problems and introduce the notation used in section 5.2. Both problems have previously been defined as a dense CRF where the set of random variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ corresponds to the set of all image pixels $i \in \mathcal{V} = \{1, 2, \ldots, N\}$. Let $\mathcal{N}$ be the neighbourhood system of the random field defined by the sets $\mathcal{N}_i, \forall i \in \mathcal{V}$, where $\mathcal{N}_i$ denotes the neighbours of the variable Z_i. A clique $c \in \mathcal{C}$ is a set of random variables $\mathbf{Z}_c \subseteq \mathbf{Z}$. Any possible assignment of labels to the random variables will be called a *labelling* and denoted by $\mathbf{z}$, similarly we use $\mathbf{z}_c$ to denote the labelling of a clique. In figure 5.1 E and F depict this lattice structure as a *blue dotted grid*, the variables Z_i are shown as *blue circles*.

5.1.1 Object Class Segmentation using a CRF

The problem of object class segmentation is formulated as in the previous sections as finding a minimal cost labelling of a CRF. In this chapter we use the notation:

$$E^O(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i^O(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^O(x_i, x_j) + \sum_{c \in \mathcal{C}} \psi_c^O(\mathbf{x}_c), \tag{5.1.1}$$

to differentiate between the CRF for object class segmentation and the CRF for the dense stereo reconstruction problem.

5.1.2 Dense Stereo Reconstruction using a CRF

We use the energy formulation of [11, 54] for the dense stereo reconstruction part of our joint formulation. They formulated the problem as one of finding a minimal cost labelling of a CRF defined over a set of random variables $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$, where each variable Y_i takes a state from the label space $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ corresponding to a set of disparities, and can be written as:

$$E^D(\mathbf{y}) = \sum_{i \in \mathcal{V}} \psi_i^D(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^D(y_i, y_j). \tag{5.1.2}$$

The unary potential $\psi_i^D(y_i)$ of the CRF is defined as a measure of colour agreement of a pixel with its corresponding pixel i from the stereo-pair given a choice of disparity y_i. The pairwise terms ψ_{ij}^D encourage neighbouring pixels in the image to have a similar disparity. The cost is a function of the distance between disparity labels:

$$\psi^D(y_i, y_j) = f(|y_i - y_j|), \tag{5.1.3}$$

where $f(.)$ usually takes the form of a linear truncated function $f(y) = \min(k_1 y, k_2)$, where k_1, $k_2 \geq 0$ are the slope and truncation respectively. The unary (*blue circles*) and pairwise (*green squares*) potentials are shown in figure 5.1 *F*. Note that the disparity for a pixel is directly related to the depth of the corresponding 3D point. To partially resolve ambiguities in disparities for low textured objects a Gaussian filter is applied to the unary potentials.

5.1.3 Monocular Video Reconstruction

With minor modification, the formulation of 5.1.2 can also be applied to monocular video sequences, by performing stereo reconstruction over adjacent frames in the video sequence (See figure 5.3). Under the simplifying assumption that the scene remains static, the formulation remains the same. However, without a

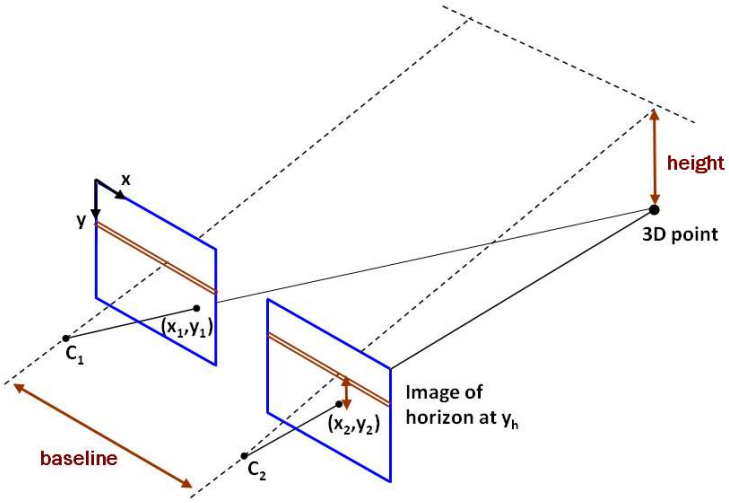


Figure 5.2: An illustration of how 3D information can be reconstructed from a stereo camera rig. Also shown, the relation between disparity (the movement of a point between the pair of images) and height, once ground plane is known.

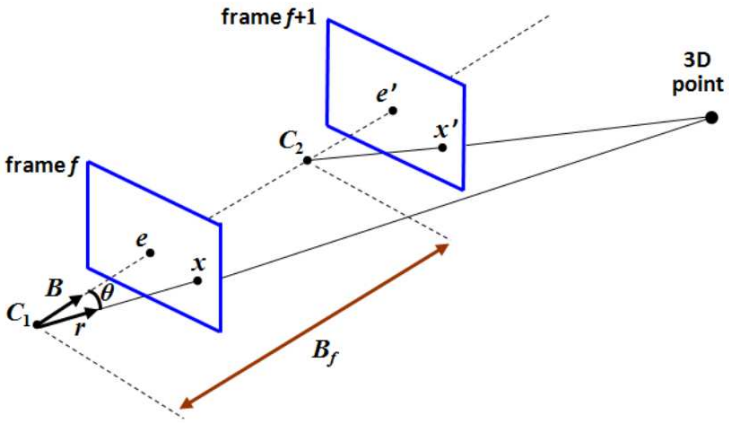


Figure 5.3: An illustration of how 3D information can be reconstructed from the monocular sequence. Details of the conversion of the monocular 3D reconstruction problem into the standard stereo reconstruction are given in the §5.1.3.

fixed baseline between the camera positions in adjacent frames the estimation of disparities, and the mapping of disparities to depths is more complex.

We first pre-process the data, by performing SIFT matching [67] over adjacent frames, before using RANSAC [28, 94] to simultaneously estimate the fundamental matrix, and a corresponding set of inliers from these matches. The fundamental matrix gives us both the epipoles[2] and the epipolar lines, and this allows us to solve the stereo correspondence efficiently by searching along corresponding epipolar lines for a match [38]. Given two images 1, and 2, we write x, x' for a pair of matched points in images 1 and 2 respectively, and use e, e' for the epipoles present in each image. The disparity d is estimated as:

$$d = \left| \, |e - x| - |e' - x'| \, \right| . \tag{5.1.4}$$

Note that we compute the disparity between pixels in a particular frame with those in its previous frame. As the camera moves forward into the image, this guarantees that every unoccluded pixel can be matched. Matching pixels from the current frame against the next would mean that pixels about the edge of the image could not be matched. As with standard stereo reconstruction, the *unary* potential of a particular choice of disparity, or equivalently a match between two pixels, is defined as the pixel difference in RGB space between them.

Converting Monocular Disparity to Stereo Disparity Unlike conventional stereo, disparities in our video sequence are not simply inversely proportional to distances, but also depend on other variables. There are two reasons for this:

- Firstly, the distance traveled between frames by the camera varies with the speed of the vehicle and this implies that the baseline varies from frame to frame.

- Secondly, when the epipole lies in the image the camera can not be approximated as orthographic. The effective baseline, which we define as the component of the baseline normal to the ray, varies substantially within an

[2] The epipoles typically lie within the image as the camera points in the direction of motion.

image from pixel to pixel.

We will describe how disparities in the monocular sequence correspond to distances, and use this to map them into standard form stereo disparities. This allows us to reuse the joint potentials learned for the stereo case, and to directly evaluate both approaches by comparing against the same ground truth.

We define a ray λr, as the set of all values taken by a 3D unit vector r, multiplied by a scalar $\lambda \in \Re$. We define the baseline B_f as the 3D distance traveled by the camera between a pair of frames f and $f+1$ [3]. We let θ be the angle between B and r. Then we define e the epipole, as the intersection point of the baseline and the image plane, and x as the point in the image that the ray λr passes through. Given a disparity d of a point on the ray, the distance s of that point from the camera is:

$$\begin{aligned} s &= K|(B_f - (B_f \cdot r)B_0)|/d \\ &= K|B_f|\sqrt{1-\cos^2\theta}/d \\ &= K|B_f||\sin\theta|/d, \end{aligned} \tag{5.1.5}$$

where K is a constant based on the internal properties of the camera and $B_0 = B_f/|B_f|$ is the unit vector in the direction of B_f.

Noting that $|e-x| \propto \tan\theta$, *i.e.* $\gamma|e-x| = \tan\theta$ for some value γ, and that $|\sin\theta| = \sqrt{\frac{\tan^2\theta}{1+\tan^2\theta}}$, we have

$$s = K|B_f|\sqrt{\frac{\gamma^2(e-x)^2}{1+\gamma^2(e-x)^2}}/d. \tag{5.1.6}$$

Solving s for a conventional stereo pair gives the related equation [54]

$$s = K|B'|/d', \tag{5.1.7}$$

where K is the same constant based on intrinsic camera parameters, $|B'|$ is the distance between the pairs of cameras, assumed to be constant and orthogonal to the field of view of both cameras, and d' is the stereo disparity. Matching the

[3] This value is a part of the standard Leuven data-set, see §5.4, and does not require estimating, in our application, see §5.5.

two equations, and eliminating s, we have

$$d' = \frac{|B'|}{|B_f|} \frac{d}{\sqrt{\frac{\gamma^2(e-x)^2}{1+\gamma^2(e-x)^2}}}. \tag{5.1.8}$$

In case the movement of the camera is very close to translation, orthogonal to the image plane, γ is sufficiently small and the disparity can be approximated by:

$$d' \approx \frac{|B'|\, d}{|B_f|\, \gamma\, |e-x|}. \tag{5.1.9}$$

Given this relationship, unary potentials defined over the monocular disparity d, can be mapped to unary potentials over the conventional stereo disparity d'. This allows standard stereo reconstruction on monocular sequences to be performed as in section 5.1.2, and joint object class and 3D reconstruction from monocular sequences to be performed as described in the following section.

5.2 Joint Formulation of Object Class Labelling and Stereo Reconstruction

We formulate simultaneous object class segmentation and dense stereo reconstruction as an energy minimization of a dense labelling $\mathbf{z}$ over the image. Each random variable $Z_i = [X_i, Y_i]$[4] takes a label $z_i = [x_i, y_i]$, from the product space of object class and disparity labels $\mathcal{L} \times \mathcal{D}$ and correspond to the variable Z_i taking object label x_i and disparity y_i. In general the energy of the CRF for joint estimation can be written as:

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^J(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^J(\mathbf{z}_c), \tag{5.2.1}$$

where the terms ψ_i^J, ψ_{ij}^J and ψ_c^J are a sum of the previously mentioned terms ψ_i^O and ψ_i^D, ψ_{ij}^O and ψ_{ij}^D, and ψ_c^O and ψ_c^D respectively, plus some terms ψ_i^C, ψ_{ij}^C, ψ_c^C, which govern interactions between $\mathbf{X}$ and $\mathbf{Y}$. However, in our case $E^D(\mathbf{y})$ (see §5.1.2) does not contain higher order terms ψ_c^D, and the joint energy is defined

[4] $[X_i, Y_i]$ is the ordered pair of elements X_i and Y_i.

as:

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^J(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^O(\mathbf{x}_c). \tag{5.2.2}$$

If the interaction terms ψ_i^C, ψ_{ij}^C are both zero, then the problems $\mathbf{x}$ and $\mathbf{y}$ are independent of one another and the energy would be decomposable into $E(\mathbf{z}) = E^O(\mathbf{x}) + E^D(\mathbf{y})$ and the two sub-problems could each be solved separately. However, in many real world data sets such as the one we describe in §5.4, this is not the case, and we would like to model the unary and pairwise interaction terms so that a joint estimation may be performed.

5.2.1 Joint Unary Potentials

In order for the unary potentials of both the object class segmentation and dense stereo reconstruction parts of our formulation to interact, we need to define some function that relates $\mathbf{X}$ and $\mathbf{Y}$ in a meaningful way. We could use depth and objects directly, as it may be that certain objects appear more frequently at certain depths in some scenarios. In road scenes we could build statistics relative to an overhead view where the positioning of the objects in the ground plane may be informative, since we expect that *buildings* will lie on the edges of the ground plane, *sidewalk* will tend to lie between *building* and *road* which would occupy the central portion of the ground plane. Building statistics with regard to the real-world positioning of objects gives a stable and meaningful cue that is invariant to the camera position. However models such as this require a substantial amount of data to avoid over-fitting.

We need to model these interactions with limited data. We do this by restricting our unary interaction potential to only modelling the observed fact that certain objects occupy a particular range of real world heights. After calibration we are able to obtain the height above the ground plane via the relation:

$$h(y_i, i) = h_c + \frac{(y_h - y_i)b}{d}, \tag{5.2.3}$$

where h_c is the camera height, y_h is the level of the horizon in the rectified image pair, y_i is the height of the i^{th} pixel in the image, b is the baseline between the

stereo pair of cameras and d is the disparity. This relationship is modeled by estimating the a priori cost of pixel i taking label $z_i = [x_i, y_i]$ by

$$\psi_i^C([x_i, y_i]) = -\log(H(h(y_i, i)|x_i)), \tag{5.2.4}$$

where

$$H(h|l) = \frac{\sum_{i \in \mathcal{T}} \delta(x_i = l)\delta(h(y_i, i) = h)}{\sum_{i \in \mathcal{T}} \delta(x_i = l)} \tag{5.2.5}$$

is a histogram based measure of the naive probability that a pixel taking label l has height h in the training set $\mathcal{T}$. The combined unary potential for the joint CRF is:

$$\psi_i^J([x_i, y_i]) = w_O^u \psi_i^O(x_i) + w_D^u \psi_i^D(y_i) + w_C^u \psi_i^C(x_i, y_i), \tag{5.2.6}$$

where ψ_i^O, and ψ_i^D are the previously discussed costs of pixel i being a member of object class x_i or disparity y_i given the image. w_O^u, w_D^u, and w_C^u are weights. Figure 5.1 G gives a graphical representation of this type of interaction shown as a *blue line* linking the unary potentials (*blue circles*) of $\mathbf{x}$ and $\mathbf{y}$ via a distribution of object heights.

5.2.2 Joint Pairwise Interactions

Pairwise potentials enforce the local consistency of object class and disparity labels between neighbouring pixels. The consistency of object class and disparity are not fully independent, an object classes boundary is more likely to occur here if the disparities of two neighbouring pixels significantly differ. To take this information into account, we chose tractable pairwise potentials of the form:

$$\begin{aligned} \psi_{ij}^J([x_i, y_i], [x_j, y_j]) \quad &= w_O^p \psi_{ij}^O(x_i, x_j) + w_D^p \psi_{ij}^D(y_i, y_j) \\ &+ w_C^p \psi_{ij}^O(x_i, x_j)\psi_{ij}^D(y_i, y_j), \end{aligned} \tag{5.2.7}$$

where $w_O^p, w_D^p > 0$ and w_C^p are weights of the pairwise potential. Figure 5.1 shows this linkage as *green line* between a pairwise potential (*green box*) of each part.

5.3 Inference for the Joint CRF

Optimisation of the energy $E(\mathbf{z})$ is challenging. Each random variable takes a label from the set $\mathcal{L}\times\mathcal{D}$ consequentially, in the experiments we consider (see § 5.4) they have 700 possible states. As each image contains 316×256 random variables, there are $700^{316\times256}$ possible solutions to consider. Rather than attempting to solve this problem exactly, we use graph cut based move making algorithms to find an approximate solution.

Graph cut based move making algorithms start from an initial solution and proceed by making a series of moves or changes, each of which leads to a solution of lower energy. The algorithm is said to converge when no lower energy solution can be found. In the problem of object class labelling, the move making algorithm α-expansion can be applied to pairwise [11] and to higher order potentials [49, 51, 59] and often achieves the best results; while in dense stereo reconstruction, the truncated convex priors (see § 5.1.2) mean that better solutions are found using range moves [56, 104] than with α-expansion.

In object class segmentation, α-expansion moves allow any random variable X_i to either retain its current label x_i or transition to the label α. More formally, given a current solution $\mathbf{x}$ the α-expansion algorithm searches through the space $\mathbf{X}_\alpha$ of size 2^N, where N is the number of random variables, to find the optimal solution, where

$$\mathbf{X}_\alpha = \left\{\mathbf{x}' \in \mathcal{L}^N : x'_i = x_i \text{ or } x'_i = \alpha\right\}. \tag{5.3.1}$$

In dense stereo reconstruction, a range expansion move defined over an ordered space of labels, allows any random variable Y_i to either retain its current label y_i or take any label $l \in [l_a, l_a + r]$. That is to say, given a current solution $\mathbf{y}$ a range move searches through the space $\mathbf{Y}_l$ of size $(r+1)^N$, which we define as:

$$\mathbf{Y}_l = \left\{\mathbf{y}' \in \mathcal{D}^N : y'_i = y_i \text{ or } y'_i \in [l, l + r]\right\}. \tag{5.3.2}$$

A single iteration of α-expansion, is completed when one expansion move for each $l \in \mathcal{L}$ has been performed. Similarly, a single iteration of range moves is completed when $|\mathcal{D}| - r$, moves have been performed.

5.3.1 Projected Moves

Under the assumption that energy $E(\mathbf{z})$ is a metric (as in object class segmentation see §5.1.1) or a semi-metric [11] (as in the costs of §5.1.2 and §5.2) over the label space $\mathcal{L} \times \mathcal{D}$, either α-expansion or $\alpha\beta$ swap respectively can be used to minimize the energy. One single iteration of α-expansion would require $O(|\mathcal{L}||\mathcal{D}|)$ graph cuts to be computed, while $\alpha\beta$-swap requires $O(|\mathcal{L}|^2|\mathcal{D}|^2)$ resulting in slow convergence. In this subsection we show how graph cut based moves can be applied to a simplified, or *projected*, form of the problem that requires only $O(|\mathcal{L}| + |\mathcal{D}|)$ graph cuts per iteration, resulting in faster convergence and better solutions. The new moves we propose are based upon a piecewise optimisation that improves in turn first object class labelling and then depth.

We call a move space *projected* if one of the components of $\mathbf{z}$, i.e. $\mathbf{x}$ or $\mathbf{y}$, remains constant for all considered moves. Alternating between moves in the projected space of $\mathbf{x}$ or of $\mathbf{y}$ can be seen as a form of hill climbing optimisation in which each component is individually optimised. Consequentially, moves applied in the projected space are guaranteed not to increase the joint energy after the move and must converge to a local optima.

We will now show that for energy (5.2.2), projected α-expansion moves in the object class label space and range moves in the disparity label space are of the standard form, and can be optimised by existing graph cut constructs. We note that finding the optimal range move or α-expansion with graph cuts requires that the pairwise and higher order terms are constrained to a particular form. This constraint allows the moves to be represented as a pairwise submodular energy that can be efficiently solved using graph cuts [55]; however, neither the choice of unary potentials nor scaling the pairwise or higher order potentials by a non-negative amount $\lambda \geq 0$ affects if the move is representable as a pairwise sub-modular cost.

5.3.2 Expansion Moves in the Object Class Label Space

For our joint optimisation of disparity and object classes, we propose a new move in the projected object-class label space. We allow each pixel taking label $z_i = [x_i, y_i]$ to either keep its current label or take a new label $[\alpha, y_i]$. Formally, given a current solution $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$ the algorithm searches through the space $\mathbf{Z}_\alpha$ of size 2^N. We define $\mathbf{Z}_\alpha$ as:

$$\mathbf{Z}_\alpha = \left\{ \begin{array}{r} \mathbf{z}' \in (\mathcal{L} \times \mathcal{D})^N : z'_i = [x'_i, y_i] \text{ and} \\ (x'_i = x_i \text{ or } x'_i = \alpha) \end{array} \right\}. \tag{5.3.3}$$

One iteration of the algorithm involves making moves for all α in $\mathcal{L}$ in some order successively. As discussed earlier, the values of the unary potential do not affect the sub-modularity of the move. For joint pairwise potentials (5.2.7) under the assumption that $\mathbf{y}$ is fixed, we have:

$$\begin{aligned} \psi^J_{ij}([x_i, y_i], [x_j, y_j]) &= (w^p_O + w^p_C \psi^D_{ij}(y_i, y_j))\psi^O_{ij}(x_i, x_j) + w^p_D \psi^D_{ij}(y_i, y_j) \\ &= \lambda_{ij}\psi^O_{ij}(x_i, x_j) + k_{ij}. \end{aligned} \tag{5.3.4}$$

The constant k_{ij} does not affect the choice of optimal move and can safely be ignored. If $\forall y_i, y_j \ \lambda_{ij} = w^p_O + w^p_C \psi^D_{ij}(y_i, y_j) \geq 0$, the projection of the pairwise potential is a Potts model and standard α-expansion moves can be applied. For $w^p_O \geq 0$ this property holds if $w^p_O + w^p_C k_2 \geq 0$, where k_2 is defined as in §5.1.2. In practice we use a variant of α-expansion suitable for higher order energies [81].

5.3.3 Range Moves in the Disparity Label Space

For our joint optimisation of disparity and object classes we propose a new move in the project disparity label space. Each pixel taking label $z_i = (x_i, y_i)$ can either keep its current label or take a new label from the range $(x_i, [l_a, l_b])$. To formalize this, given a current solution $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$ the algorithm searches through the space

$\mathbf{Z}_l$ of size $(2+r)^N$, which we define as:

$$\mathbf{Z}_l = \left\{ \begin{array}{r} \mathbf{z}' \in (\mathcal{L} \times \mathcal{D})^N : z'_i = [x_i, y'_i] \text{ and} \\ (y'_i = y_i \text{ or } y'_i \in [l, l+r]) \end{array} \right\}. \tag{5.3.5}$$

As with the moves in the object class label space, the values of the unary potential do not affect the sub-modularity of this move. Under the assumption that $\mathbf{x}$ is fixed, we can write our joint pairwise potentials (5.2.7) as:

$$\begin{aligned} \psi^J_{ij}([x_i, y_i], [x_j, y_j]) &= (w^p_D + w^p_C \psi^O_{ij}(x_i, x_j))\psi^D_{ij}(y_i, y_j) + w^O_d \psi^O_{ij}(x_i, x_j) \\ &= \lambda_{ij} \psi^D_{ij}(y_i, y_j) + k_{ij}. \end{aligned} \tag{5.3.6}$$

Again, the constant k_{ij} can safely be ignored, and if $\forall x_i, x_j \ \lambda_{ij} = w^p_D + w^p_C \psi^O_{ij}(x_i, x_j) \geq 0$ the projection of the pairwise potential is linear truncated and standard range expansion moves can be applied. This property holds if $w^p_D + w^p_C(\theta_p + \theta_v) \geq 0$, where θ_p and θ_v are the weights of the Potts pairwise potential (see §5.1.1).

5.4 Data set

We augment a subset of the Leuven stereo data set[5] of [64] with object class segmentation and disparity annotations. The Leuven data set was chosen as it provides image pairs from two cameras, 150cm apart from each other, mounted on top of a moving vehicle, in a public urban setting. In comparison with other data sets, the larger distance between the two cameras allows better depth resolution, while the real world nature of the data set allows us to confirm our statistical model's validity. However, the data set does not contain the object class or disparity annotations, we require to learn and quantitatively evaluate the effectiveness of our approach.

To augment the data set all image pairs were rectified, and cropped to 316×256, then the subset of 70 non-consecutive frames was selected for human annotation. The annotation procedure consisted of two parts. Firstly we manually labeled each pixel in every image with one of 7 object classes: *Building, Sky, Car,*

[5] `http://www.vision.ee.ethz.ch/~bleibe/cvpr07/datasets.html`

Road, Person, Bike and *Sidewalk*. An 8^{th} label, *Void*, is given to pixels that do not obviously belong to one of these classes. Secondly disparity maps were generated by manually matching by hand the corresponding planar polygons, some examples of which are shown in the figure 5.4 *A, B*, and *D*.

We believe our augmented subset of the Leuven stereo data set to be the first publicly available data set that contains both object class segmentation and dense stereo reconstruction ground truth for real world data. This data differs from commonly used stereo matching sets like the Middlebury [82] data set, as it contains challenging large regions which are homogeneous in colour and texture, such as *sky* and *building*, and suffers from poor photo-consistency due to lens flares in the cameras, specular reflections from windows and inconsistent luminance between the left and right camera. It should also be noted that it differs from the CamVid database [12] in two important ways, CamVid is a monocular sequence, and the 3D information comes in the form of a set of sparse 3D points with outliers[6]. These differences give rise to a challenging new data set that is suitable for training and evaluating models for dense stereo reconstruction, $2D$ and 3D scene understanding, and joint approaches such as ours.

5.5 Experiments

For training and evaluation of our method we split the data set (§5.4) into three sequences: Sequence 1, frames 0-447; Sequence 2, frames 512-800; Sequence 3, frames 875-1174. Augmented frames from sequence 1 and 3 are selected for training and validation, and sequence 2 for testing. All *void* pixels are ignored. We quantitatively evaluate the object class segmentation by measuring the percentage of correctly predicted labels over non-*void* pixels in the test sequence. The dense stereo reconstruction performance is quantified by measuring the number of pixels which satisfy $|d_i - d_i^g| \leq \delta$, where d_i is the label of i-th pixel, d_i^g is corresponding ground truth label and δ is the allowed error. We increment δ from 0 (exact) to 20 (within 20 disparities) giving a clear picture of the performance. The total

[6] The outlier rejection step was not performed on the 3D point cloud in order to exploit large re-projection errors as cues for moving objects. See [12] for more details.

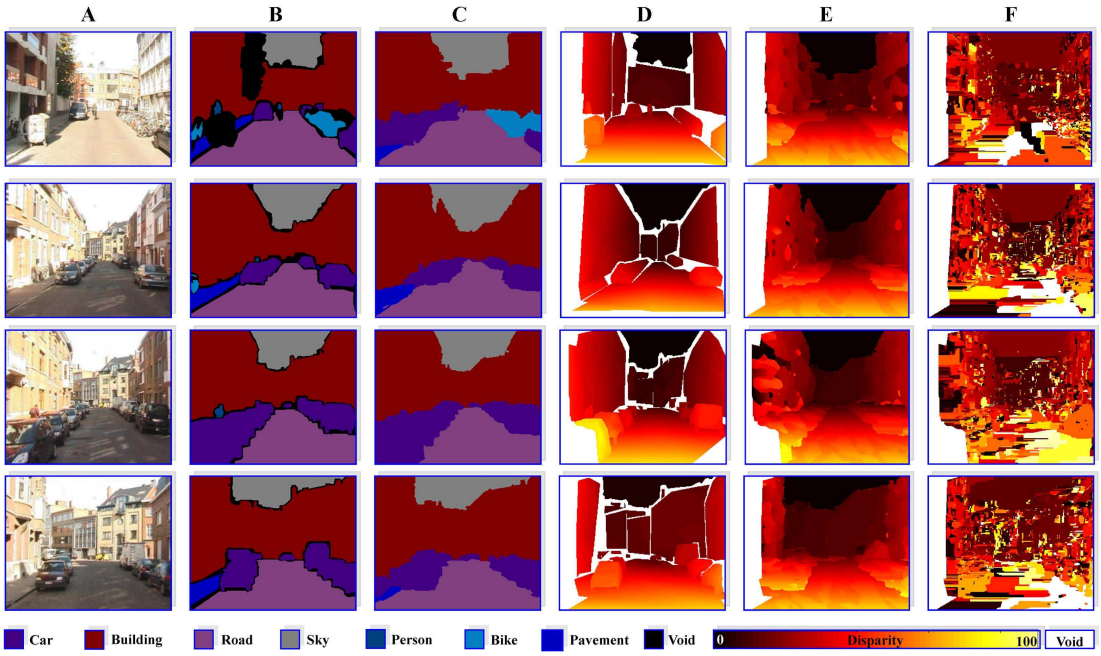


Figure 5.4: Qualitative object class and disparity results for Leuven data set.(**A**) *Original Image.* (**B**) *Object class segmentation ground truth.* (**C**) *Proposed method Object class segmentation result.* (**D**) *Dense stereo reconstruction ground truth.* (**E**) *Proposed method dense stereo reconstruction result.* (**F**) *Stand alone dense stereo reconstruction result (LT). Best viewed in colour.*

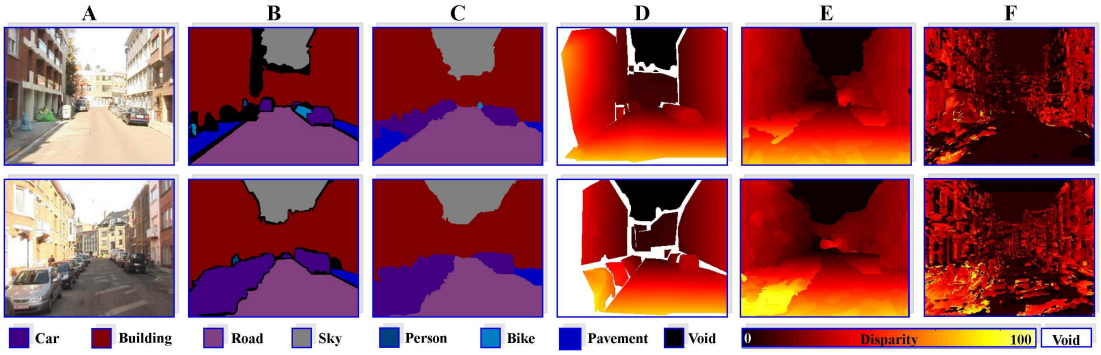


Figure 5.5: Monocular results. *(***A***)* Original Image. *(***B***)* Object class segmentation ground truth. *(***C***)* Proposed method Object class segmentation result. *(***D***)* Dense stereo reconstruction ground truth. *(***E***)* Proposed method dense stereo reconstruction result. *(***F***)* Stand alone dense stereo reconstruction result (LT). The quality of reconstruction improves with the distance from the epipole. Best viewed in colour.

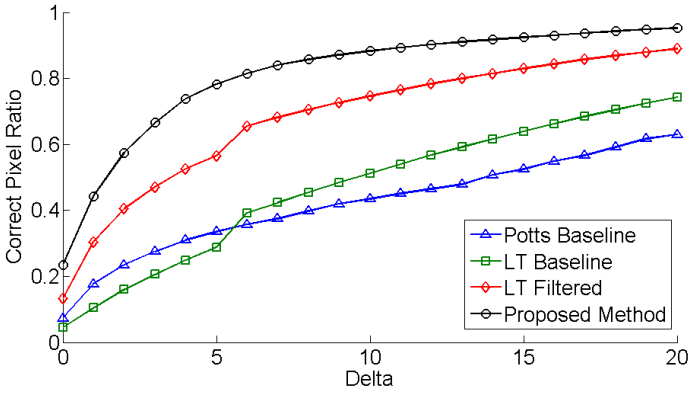


Figure 5.6: Quantitative comparison of the performance of disparity CRFs. *We can clearly see that our joint approach §5.2 (Proposed Method) outperforms standard dense stereo approaches based on the Potts [54] (Potts Baseline), Linear truncated models described in §5.1.2 (LT Baseline) and Linear truncated with Gaussian filtered unary potentials (LT Filtered). The correct pixel ratio is the proportion of pixels which satisfy* $|d_i - d_i^g| \leq \delta$*, where* d_i *is the disparity label of i-th pixel,* d_i^g *is corresponding ground truth label and* δ *is the allowed error. See §5.5 for discussion.*

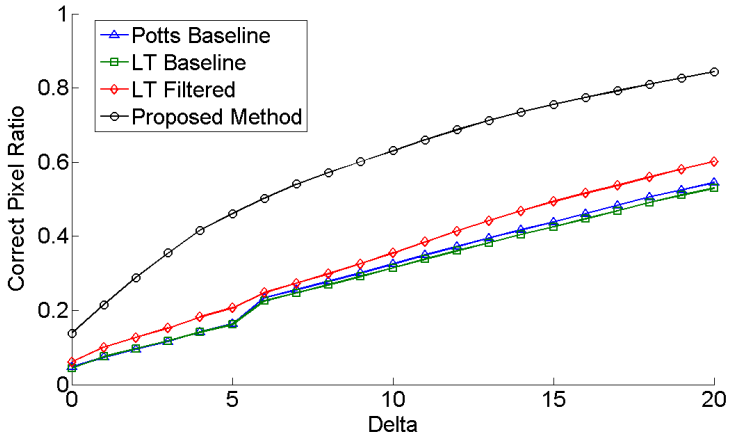


Figure 5.7: Quantitative comparison of the performance of disparity CRFs, on monocular sequences. *As with the stereo pair, we can clearly see that our joint approach §5.2 (Proposed Method) outperforms the stand alone approaches with baseline Potts [54] (Potts Baseline), Linear truncated potentials §5.1.2 (LT Baseline) and Linear truncated with Gaussian filtered unary potentials (LT Filtered). The correct pixel ratio is the proportion of pixels which satisfy* $|d_i - d_i^g| \leq \delta$*, where* d_i *is the disparity label of i-th pixel,* d_i^g *is corresponding ground truth label and* δ *is the allowed error. See §5.5.4 for discussion, and figure 5.4 to compare against conventional stereo.*

	Global	Building	Sky	Car	Road	Sidewalk	Bike
Stand alone	95.7	96.7	99.8	93.5	99.0	60.2	59.3
Joint approach	95.8	96.7	99.8	94.0	98.9	60.6	59.5

Table 5.1: *Quantitative results for object class segmentation of stand alone and joint approach. The pixel accuracy (%) for different object classes. The 'global' measure corresponds to the total proportion of pixels labeled correctly. Per class accuracy corresponds to recall measure commonly used for this task [88, 89, 59]. Minor improvement were achieved for smaller classes that had fewer pixels present in the data set. We assume the difference would be larger for harder data sets. Class* person *was removed from evaluation due to insufficient statistics on the test set.*

number of disparities used for evaluation is 100.

5.5.1 Object Class Segmentation

The object class segmentation CRF as defined in §5.1.1 performed extremely well on the data set, better than we had expected, with 95.7% of predicted pixel labels agreeing with the ground truth. Qualitatively we found that the performance is stable over the entire test sequence, including those images without ground truth. Most of the incorrectly predicted labels are due to the high variability of the object class person, and insufficient training data to learn their appearance. Quantitative comparison of the stand alone and joint method is given in table 5.1.

5.5.2 Dense Stereo Reconstruction

The Potts [54] and linear truncated (LT) baseline dense stereo reconstruction models described in §5.1.2 performed relatively well, with large δ, considering the difficulty of the data, plotted in figure 5.6 as 'Potts baseline' and 'LT baseline'. We found that on our data set a significant improvement was gained by smoothing the unary potentials with a Gaussian blur[7] before incorporating the potential in

[7] This is a form of robust measure, see §3.1 of [82] for further examples.

the CRF framework with linear truncated model, as can be seen in figure 5.6 'LT Filtered'. For qualitative results see figure 5.4 *E*.

5.5.3 Joint Approach

Our joint approach defined in sections §5.2 and §5.3 consistently outperformed the best stand-alone dense stereo reconstruction as can be seen in figure 5.6. Improvement of the object class segmentation was less dramatic, with 95.8% of predicted pixel labels agreeing with the ground truth. We expect to see a more significant improvement on more challenging data sets, and the creation of an improved data set is part of our future work. Qualitative results can be seen in figure 5.4 *C* and *E*.

5.5.4 Monocular Reconstruction

Reconstruction from a monocular sequence is substantially harder than the corresponding stereo problem. Not only does it suffer from the same problems of varying illumination and homogeneous regions, but the effective base-line is substantially shorter making it much harder to recover 3D information with any degree of accuracy, particularly in the region around the epipole (see §5.1.3 and figure 5.5). Despite this, plausible 3D reconstruction is still possible, particularly when performing joint inference over object class and disparity simultaneously, quantitative results can be seen in figure 5.7. Note that the joint optimisation of monocular disparity and object class out performs the pre-existing methods (*LT Baseline* and *Potts Baseline*) over conventional two camera stereo data, and is comparable to the two camera results on *LT filtered*. In figure 5.5 qualitative results can be seen. As expected, these show the quality of reconstruction improves with the distance from the epipole. Consequentially, one of the regions most successfully reconstructed is marked as *void* in the two camera disparity maps, as it is not in the field of view of both cameras. This suggests that the numeric evaluation of figure 5.7 may be overly pessimistic.

5.6 Conclusion

Traditionally the prior in stereo has been fixed to some standard tractable model such as truncated linear on disparities. Within this work we open up the intriguing possibility that the prior on shape should take in account the type of scene and object we are looking at. To do this, we provided a new formulation of the problems, a new inference method for solving this formulation and a new data set for the evaluation of our work. Evaluation of our work shows a dramatic improvement in stereo reconstruction compared to existing approaches. We assume statistically significant gain can be achieved also for object class segmentation, but it would require more challenging data set. The method can be applied to any other scenes where mutual information between 3D location and object label is present. Within this chapter we have focussed on road scenes. Here the object label strongly influences the depth label once we consider a parametrization in terms of height above the road plane. So far we have not considered learning the relation of the smoothness term for depth to object class, for instance the fact walls might be vertical etc. This would be an interesting line for future work. This work puts us one step closer to achieving complete scene understanding, and provides strong experimental evidence that the joint labelling of different problems can bring substantial gains.

Chapter 6

Conclusion and Future Work

6.1 Summary

This thesis is a step towards complete scene understanding, proposing new structured models for labelling problems and efficient algorithms to do inference on them. The work goes beyond standard pairwise conditional random fields, typically used for most labelling tasks. We showed, that for all proposed complex structured formulations efficient graph cut based inference is applicable. The proposed models have been successfully applied to semantic object class segmentation, object detection and dense $3D$ stereo reconstruction yielding *s*tate-of-the-art results for several standard data sets.

The main contributions of this dissertation are:

(i) A new associative hierarchical model that enforces consistency between potentials on different scales. We showed that the proposed model is a generalisation of most of the standard methods used for semantic object class segmentation. We proposed novel potentials for this task and an efficient graph cut based move-making algorithm to deal with the optimisation problem. Published at ICCV'09 and UAI'10.

(ii) A method to include co-occurrence statistics in the CRF framework. Our formulation satisfies all the desired properties of incorporation of such statistics, such us incorporation as a weak constraint, invariance to size and preference of parsimonious solutions. We showed how the model can be efficiently optimised using graph cut based move making algorithms. In practice incorporation of the co-occurrence statistics leads to qualitatively better results. Published at ECCV'10 (Best paper award) and as an invited paper at IJCV'11.

(iii) A novel latent field SVMs for object detection with convex MRF prior on deformation field of the deformable template, that seems to be a natural generalisation of the common methods used for this task. We showed how the latent

variable model can be learnt efficiently. Incorporation of detector responses in the CRF framework for the object class segmentation problem led to a significant improvement of the performance. Published at ECCV'10.

(iv) A novel formulation that models jointly the problem defined over multiple domains with a product space of labels. We proposed efficient projected-move inference to deal with these problems. We demonstrated the usefulness of this model on the joint estimation of dense $3D$ stereo reconstruction and object class segmentation. The method significantly outperforms existing approaches for dense stereo reconstructions of road scenes. Published at BMVC'10 (Best paper award) and as an invited paper at IJCV'11.

6.2 Future work

This thesis dealt with some of the structures in Conditional Random Fields, that are desired to be enforced or induced in the solution of the labelling problem. A natural extension of this work is to analyze other useful properties in computer vision, such as shape, motion or symmetry, and propose new probabilistic formulations and inference methods to deal with them. Based on the experimental results, the representation of several cues in one probabilistic framework seems to be the promising direction to formulate complex computer vision problems and we believe this is just the beginning of an exciting journey into global structured models towards scene understanding.

Bibliography

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 2003.

[2] A. Barbu. Learning real-time mrf inference for image denoising. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] D. Batra, R. Sukthankar, and C. Tsuhan. Learning class-specific affinities for image labelling. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[4] J. Besag. On the statisical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B-48:259–302, 1986.

[5] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *European Conference on Computer Vision*, 2004.

[6] A. Blake and A. Zisserman. *Visual reconstruction.* MIT Press, 1987.

[7] A. Bordes, L. Bottou, and P. Gallinari. SGD-QN: Careful quasi-newton stochastic gradient descent. *Journal of Machine Learning Research*, 10:1737–1754, 2009.

[8] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *International Conference on Image and Video Retrieval*, 2007.

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

[10] Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *Transactions on Pattern Analysis and Machine Intelligence*, 26:1124–1137, 2004.

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

[12] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009.

[13] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[14] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *Conference on Computer Vision and Pattern Recognition*, 2007.

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

[16] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[17] G. Csurka and F. Perronnin. A simple high performance approach to semantic segmentation. In *British Machine Vision Conference*, 2008.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[19] A. Delong, A. Osokin, H. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *Conference on Computer Vision and Pattern Recognition*, 2010.

[20] A. R. Dick, P. H. S. Torr, and R. Cipolla. Modelling and interpretation of architecture from several images. *International Journal of Computer Vision*, 60:111–134, 2004.

[21] Y. Dinitz. Dinitz' algorithm: The original version and even's version. In *Theoretical Computer Science*, 2006.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[23] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Conference on Computer Vision and Pattern Recognition*, 2000.

[24] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004.

[25] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[26] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[27] T. Finley and T. Joachims. Training structural SVMs when exact inference is intractable. In *International Conference on Machine Learning*, 2008.

[28] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, pages 726–740, 1981.

[29] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *Transactions on Computers*, C-22:67–92, 1973.

[30] L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

[31] D. Freedman and P. Drineas. Energy minimization via graph cuts: Settling what is possible. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[32] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[33] B. Goldlücke and D. Cremers. Convex relaxation for multilabel problems with product label spaces. In *European Conference on Computer Vision*, 2010.

[34] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *International Conference on Computer Vision*, 2009.

[35] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *Advances in Neural Information Processing Systems*, 2009.

[36] I. Guyon, B. Boser, and V. Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In *Advances in Neural Information Processing Systems*, 1993.

[37] P. Hammer. Some network flow problems solved with pseudo-boolean programming. *Operations Research*, 1965.

[38] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[39] H. Harzallah, C. Schmid, F. Jurie, and A. Gaidon. Classification aided two stage localization. *Technical report, INRIA, France*, 2008.

[40] X. He, R. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *European Conference on Computer Vision*, 2006.

[41] G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[42] D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Transactions on Graphics*, pages 577–584, 2005.

[43] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *International Conference on Computer Vision*, 2005.

[44] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Conference on Computer Vision and Pattern Recognition*, 2006.

[45] D. Hoiem, C. Rother, and J. M. Winn. 3D layout CRF for multi-view object class recognition and segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[46] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *Transactions on Pattern Analysis and Machine Intelligence*, 25:1333–1336, 2003.

[47] T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[48] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. *Journal of the ACM*, pages 14–23, 2002.

[49] P. Kohli, M. Kumar, and P. H. S. Torr. P^3 and beyond: Solving energies with higher order cliques. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[50] P. Kohli, L. Ladicky, and P. H. S. Torr. Graph cuts for minimizing robust higher order potentials. Technical report, Technical report, Oxford Brookes University, UK, 2008.

[51] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[52] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Transactions on Pattern Analysis and Machine Intelligence*, 28:1568–1583, 2006.

[53] V. Kolmogorov and C. Rother. Comparison of energy minimization algorithms for highly connected graphs. In *European Conference on Computer Vision*, 2006.

[54] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions via graph cuts. In *International Conference on Computer Vision*, 2001.

[55] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts?. *Transactions on Pattern Analysis and Machine Intelligence*, 26:65–81, 2004.

[56] M. Kumar and P. H. S. Torr. Efficiently solving convex relaxations for map estimation. In *International Conference on Machine Learning*, 2008.

[57] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[58] M. P. Kumar, O. Veksler, and P. H. S. Torr. Improved moves for truncated convex models. *Journal of Machine Learning Research*, 12:31–67, 2011.

[59] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *International Conference on Computer Vision*, 2009.

[60] L. Ladicky, C. Russell, P. Sturgess, K. Alahari, and P. H. S. Torr. What, where and how many? Combining object detectors and CRFs. *European Conference on Computer Vision*, 2010.

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

[62] D. Larlus and F. Jurie. Combining appearance models and Markov random fields for category level object segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[63] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, 2006.

[64] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3D scene analysis from a moving vehicle. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[65] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[66] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[67] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[68] S. Maji and A. C. Berg. Max-margin additive classifiers for detection. *International Conference on Computer Vision*, 2009.

[69] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43:7–27, 2001.

[70] M. Narasimhan and J. A. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *Uncertainty in Artificial Intelligence*, 2005.

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

[72] J. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. In *Conference on Integer Programming and Combinatorial Optimization*, 2007.

[73] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *European Conference on Computer Vision*, 2008.

[74] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision*, 2007.

[75] S. Ramalingam, P. Kohli, K. Alahari, and P. H. S. Torr. Exact inference in multi-label CRFs with higher order cliques. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[76] X. Ren and J. Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision*, 2003.

[77] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. S. Torr. Randomized trees for human pose detection. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[78] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[79] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry. In *Conference on Computer Vision and Pattern Recognition*, 2005.

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

[81] C. Russell, L. Ladicky, P. Kohli, and P. H. S. Torr. Exact and approximate inference in associative hierarchical networks using graph cuts. *Uncertainty in Artificial Intelligence*, 2010.

[82] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[83] D. Schlesinger and B. Flach. Transforming an arbitrary minsum problem into a binary one. Technical report, Dresden University of Technology, 2006.

[84] M. Schlesinger. Syntactic analysis of two-dimensional visual signals in noisy conditions. *Kibernetika*, 1976.

[85] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *International Conference on Machine Learning*, 2007.

[86] J. Shi and J. Malik. Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[87] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2008.

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

[89] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *British Machine Vision Conference*, 2009.

[90] 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 *European Conference on Computer Vision*, 2006.

[91] H. Tao, H. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *International Conference on Computer Vision*, 2001.

[92] B. Taskar, V. Chatalbashev, and D. Koller. Learning associative Markov networks. In *International Conference on Machine Learning*, 2004.

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

[94] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24:271–300, 1997.

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

[96] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *International Conference on Computer Vision*, 2003.

[97] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *Transactions on Pattern Analysis and Machine Intelligence*, 30:1483–1489, 2008.

[98] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *The Journal of Machine Learning Research*, 6:1453–1484, 2005.

[99] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Conference on Computer Vision*, 2003.

[100] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[101] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *International Conference on Computer Vision*, 2007.

[102] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *International Conference on Computer Vision*, 2009.

[103] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[104] O. Veksler. Graph cut based optimization for MRFs with truncated convex priors. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[105] J. Wang, P. Bhat, A. Colburn, M. Agrawala, and M. Cohen. Interactive video cutout. *ACM Transactions on Graphics*, 24:585–594, 2005.

[106] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Conference on Computer Vision and Pattern Recognition*, 2000.

[107] Y. Weiss and W. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *Transactions on Information Theory*, 47:736–744, 2001.

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

[109] L. Zhu, Y. Chen, Y. Lin, C. Lin, and A. L. Yuille. Recursive segmentation and recognition templates for 2D parsing. In *Advances in Neural Information Processing Systems*, 2008.