

**Going Beyond Semantic Image Segmentation, Towards
Holistic Scene Understanding, with Associative Hierarchical
Random Fields**

by

Paul Adrian Sturges

Submitted to the Department of Computing and Communication Technologies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Vision

at the

OXFORD BROOKES UNIVERSITY

August 2016

Going Beyond Semantic Image Segmentation, Towards Holistic Scene Understanding, with Associative Hierarchical Random Fields

by

Paul Adrian Sturgess

Submitted to the Department of Computing and Communication Technologies
on August , 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Vision

Abstract

In this thesis we exploit the generality and expressive power of the Associative Hierarchical Random Field (AHRF) graphical model to take its use beyond that of semantic image segmentation, into object-classes, towards a framework for holistic scene understanding. We provide a working definition for the holistic approach to scene understanding, which allows for the integration of existing, disparate, applications into an unifying ensemble. We believe that modelling such an ensemble as an AHRF is both a principled and pragmatic solution. We present a hierarchy that shows several methods for fusing applications together with the AHRF graphical model. Each of the three; feature, potential and energy, layers subsumes its predecessor in generality and together give rise to many options for integration. With applications on street scenes we demonstrate an implementation of each layer. The first layer application joins appearance and geometric features. For our second layer we implement a *things and stuff* co-junction using higher order AHRF potentials for object detectors, with the goal of answering the classic questions: What? Where? and How many? A holistic approach to recognition-and-reconstruction is realised within our third layer by linking two energy based formulations of both applications. Each application is evaluated qualitatively and quantitatively. In all cases our holistic approach shows improvement over baseline methods.

Contents

1	Introduction	13
1.1	Semantic Image Segmentation with AHRF	17
1.2	Holistic Scene Understanding With AHRF	19
1.2.1	Feature level: Geometry and Appearance	19
1.2.2	Potential level: Things and Stuff	20
1.2.3	Energy level: Recognition and Reconstruction	20
1.3	Contributions	21
1.4	Document Map	21
1.5	Publications	22
2	Semantic Image Segmentation with Associative Hierarchical Random Fields	23
2.1	The Labelling Problem	23
2.1.1	Input	24
2.1.2	Output	24
2.1.3	Model	25
2.1.4	Representation	26
2.2	Associative Higher Order Random Field (AHRF)	28
2.2.1	Unary Potentials	29
2.2.2	Associative Pairwise Potentials	30
2.2.3	Associative Higher Order Potentials	31
2.3	Learning	33
2.4	Implementation	35

3	Holistic Scene Understanding with Associative Hierarchical Random Fields	37
3.1	Level-1: The Feature Level	39
3.2	Level-2: The Potential Level	40
3.3	Level-3: The Energy Level	41
4	Holistic Applications: Street Scene Understanding	43
4.1	Feature Layer: Appearance and Geometry	44
4.1.1	Introduction	44
4.1.2	Global Configuration	47
4.1.3	Appearance	47
4.1.4	Geometry	47
4.1.5	Joint Appearance and Geometry	48
4.2	Potential Layer: Things and Stuff	52
4.2.1	Introduction	52
4.2.2	Global Configuration	55
4.2.3	Things	55
4.2.4	Stuff	56
4.2.5	Things and Stuff	57
4.2.6	Inference	60
4.3	Energy Layer: Recognition and Reconstruction	62
4.3.1	Introduction	62
4.3.2	Global Configuration	65
4.3.3	Recognition	65
4.3.4	Reconstruction	66
4.3.5	Recognition and Reconstruction	67
4.3.6	Inference	69
5	Evaluation	73
5.1	CamVid Dataset [7,8]	73
5.2	Leuven Dataset [10,42]	76
5.3	Evaluation Protocol and Metrics	78

5.4	Feature Layer: Appearance and Geometry	81
5.5	Potential Layer: Things and Stuff	84
5.6	Energy Layer: Recognition and Reconstruction	88
6	Conclusion	91
6.1	Feature Layer: Geometry and Appearance	91
6.2	Potential Layer: Things and Stuff	92
6.3	Energy Layer: Recognition and Reconstruction	92
	Appendices	100
A	Original Contributions	101

List of Figures

1-1	Scene Understanding Schematic	14
1-2	Holistic Scene Understanding Schematic	16
1-3	Semantic Image Segmentation Goal	17
1-4	Hierarchical Grouping with AHRF	18
2-1	Factor Graph	28
2-2	AHRF Plate Diagram	29
2-3	Associative Random Field Segment Costs	33
2-4	Piecewise Learning Plate Diagram	34
3-1	Fusion Hierarchy	38
3-2	Level-1 Plate Diagram	39
3-3	Level-2 Plate Diagram	40
3-4	Level-3 Plate Diagram	41
4-1	Examples of Similar Scene Geometry Over Time and Distance	45
4-2	A Conceptual View of Things and Stuff Fusion	53
4-3	Things and Stuff AHRF Graphical Model	59
4-4	Things and Stuff Counting Instances	61
4-5	Recognition and Reconstruction Graphical Model	63
5-1	CamVid Database	74
5-2	CamVid Splits	75
5-3	Leuven Dataset	77
5-4	Evaluation Metrics	80

5-5	Appearance and Geometry Qualitative Results 1	82
5-6	Appearance and Geometry Qualitative Results 2	84
5-7	Things and Stuff Qualitative Results 1	85
5-8	Things and Stuff Qualitative Results 2	86
5-9	Recognition and reconstruction Quantitative Results	89
5-10	Recognition and Reconstruction Qualitative Results	90

List of Tables

5.1	Appearance and Geometry Quantitative Results	83
5.2	Things and Stuff Quantitative Results	87
5.3	Recognition and Reconstruction Quantitative Results	88

Chapter 1

Introduction

The thesis of this dissertation is that a holistic approach to scene understanding is both pragmatic and effective, and can be realised in a principled and efficient manner when represented as a graphical model.

Within the field of computer vision— *The ultimate goal of scene understanding is to build machines that have the same level of visual understanding as humans do: Build machines that see like we do.* This high level definition is schematically depicted in the top part of Fig. 1-1, where a digital input stream is transformed to a representation of human level semantics. However, the exact reasons for why a human, even a young child, can easily understand a scene remains largely elusive, and mimicking this ease of cognition in a machine has proved to be a bewildering task. We could go as far as to say that there is an *elephant in the room* within this goal: Nobody knows what we see; nobody knows how we see it; and nobody knows why we are seeing it. Due to this bewilderment, and the driving need for practical solutions to more specific problems, scene understanding is no longer a stand-alone thesis; it has diverged away from its general roots, and been broken down into a diverse set of specialised applications. These are abstracted in the bottom part of Fig. 1-1 as different coloured boxes. Each of these applications may take an input and transform it to the level of semantics required to perform the specific task being addressed, such as those that form the focus of the VOC challenge [12]

(image classification, object detection, semantic image segmentation) or those found in open source computer vision libraries (some modules of OpenCV [6] could be considered as some organisation of specific scene understanding tasks).

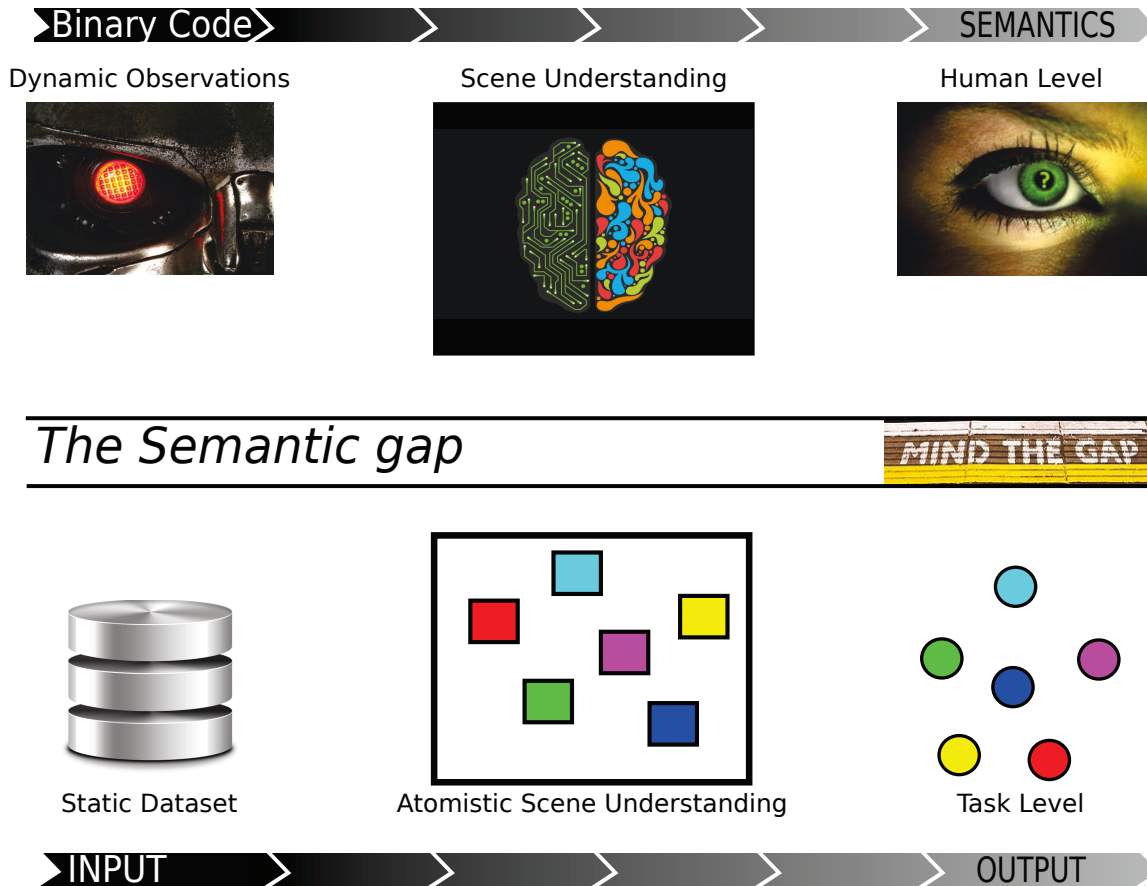


Figure 1-1: **Scene Understanding:** The current state of research into scene understand tends to be fragmented into different tasks, performed on static datasets of consumer photographs (bottom). This is in stark contrast to the original goal of scene understanding that acts upon dynamic, real world, streaming visual data and performs at a human level of semantics (top). This creates a, so called, *semantic gap* between idealised and realised scene understanding.

This trend away from modelling the whole together, towards modelling the parts of the whole, separately, is apparently moving in the wrong direction for our thesis, opening a so called *semantic gap* [65] (Fig. 1-1). However, this does not necessarily have to be the case—better performing parts can lead to a better performing whole—and the decoupling of tasks allows for more focused research that accelerates their individual performances [2, 12, 18, 51, 53, 68]. This

is the core idea behind our thesis. In order to achieve a whole scene understanding system, we first need to provide the glue that holds the parts together, and the architecture that allows for an informed choice of where, and how, to place them. Architecting these systems in such a way that all the parts are intimately inter-connected is termed *Holistic Scene Understanding*, popularised by the pioneering work of Zhu *et al.* [61], Hoiem *et al.* [27, 29] and Gould *et al.* [20, 21, 24] and sustaining much interest within the machine learning and computer vision communities [19, 43, 44, 46, 69, 71].

In providing a general goal for such a holistic approach, we desire one that will impact the practicality of realising a seeing machine, eliminating the proverbial elephant in the room— *The ultimate goal of Holistic Scene Understanding is to ensemble machines that, as a whole, have a high enough level of visual understanding to mimic seeing as humans do: Build an ensemble of machines that, in unison, appear to see like we do.* An interpretation of this is depicted at the top of Fig. 1-2, where we envision a *plug-n-play* system that can *ape* human behaviour for some general tasks by automatically configuring the inter-play between a set of task specific modules, or aspects. We take a step towards this goal by designing a, penultimate, *configurable* system, where we manually model the inter-dependencies between the aspects, as depicted in the lower part of Fig. 1-2.

We assert that taking a holistic approach is: Pragmatic, because it decouples the concern of modelling (possibly) complex interdependencies between modules, from their (independent) development; effective, because the performance of each part can be boosted by the modelling of their interconnections with other parts; efficient, when the interdependencies are represented by graphical models, that themselves have been proven to have efficient inference. We demonstrate these assertions by implementing and evaluating a suite of holistic scene understanding applications on street scenes, a popular playground for joint modelling [19, 69, 71]. We show how to configure the existing Associative Hierarchical Random Field (AHRF) graphical model [37] for fusing different sources of information, thus going beyond its original purpose as a multi-resolution semantic image segmentation system.

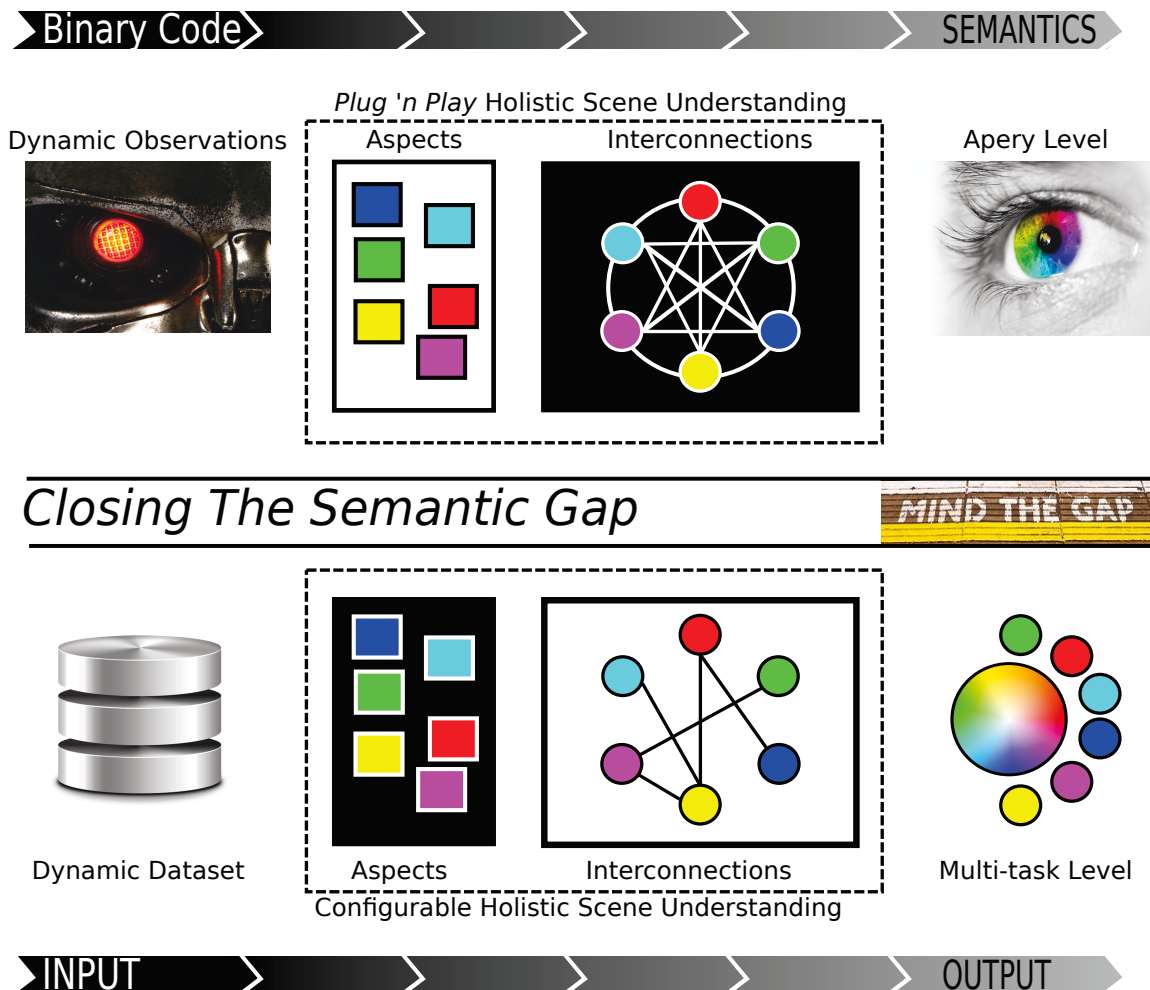


Figure 1-2: **Holistic Scene Understanding:** (Top row) Ultimately we would like to realise an ensemble system that can take dynamic, real-world, streaming data (top left) and interpret it with a level of semantics that can mimic, or ape, humans for general vision based tasks (top right). We envision a plug-n-play system (top middle), where disparate aspects, that can continue to be developed as white box systems, are automatically fused together when plugged in to a black box holistic engine. We take a step towards this goal with a configurable system. Our approach (bottom) takes inputs from datasets acquired from a dynamic source (bottom left), the data is processed by a set of independently developed aspects, here we treat these as black boxes that perform a certain known task, the outputs of these modules are then fused with a hand crafted, or configurable, graphical model (bottom middle). The system then outputs the multiple, enhanced, results from the interconnected tasks. (bottom right)

1.1 Semantic Image Segmentation with AHRF

The problem of Semantic Image Segmentation (Fig.1-3), a fundamental scene understanding task, is to give meaning to an image by way of relating semantic labels with it. Semantic labellings can convey different information and can, roughly, be divided into: low level, such as edges; mid-level, such as groupings; and high level, such as object-classes. We are particularly interested in the case of object-classes, since the AHRF framework was originally conceived in order to tackle this problem, and we extend this framework to address holistic scene understanding. AHRF can be considered as higher order Conditional Random Field (CRF) framework in

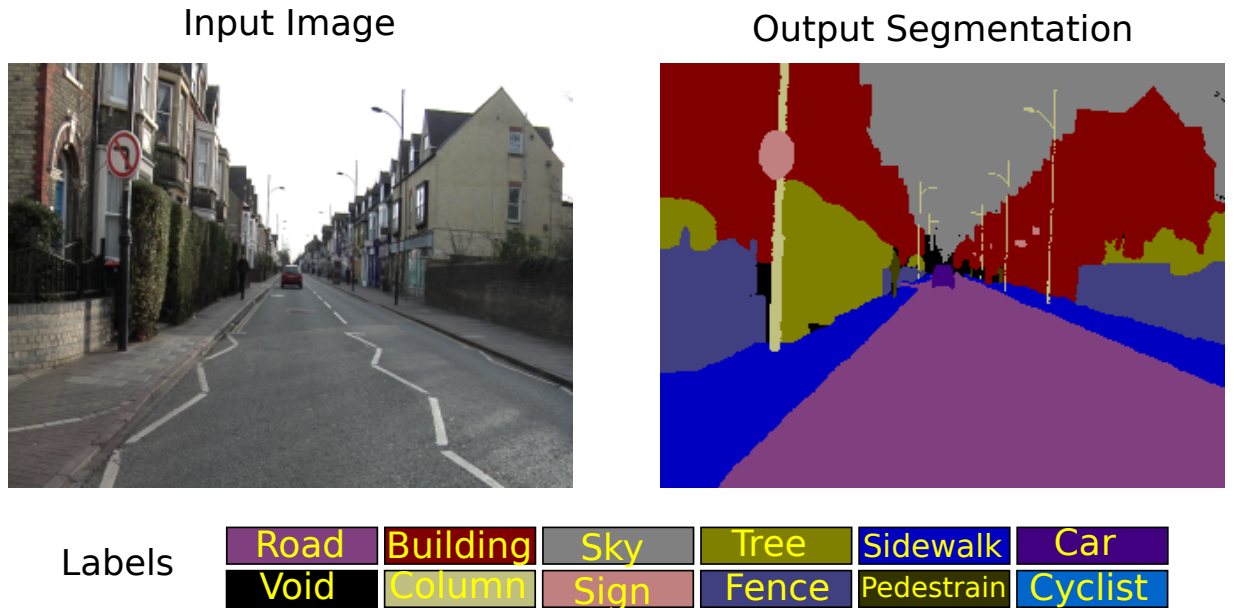


Figure 1-3: **Semantic Image Segmentation:** The goal of semantic image segmentation is to label every pixel, or region, of an input image with an object-class label, as shown in the labels palette.

which a semantic image segmentation is inferred via Maximum A-Priori (MAP) approximate inference. In this context, a first order approximation would only consider elementary units of an image, such as pixels, a second order approximation would extend this to pairs of pixels, and a higher order approximation can exploit any subset of pixels, in order to infer the object-class assignments required to semantically segment the image.

There has been much interest in higher order CRFs. They are successfully used to improve

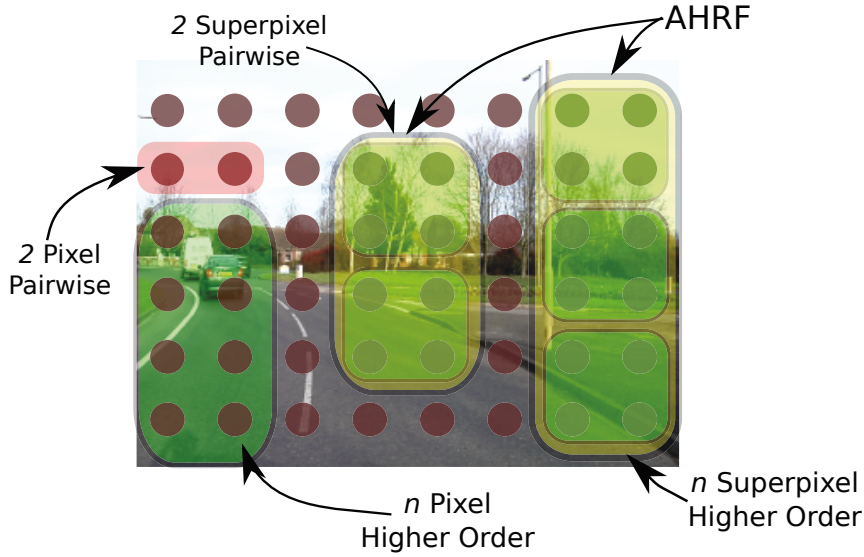


Figure 1-4: **Hierarchical Grouping with AHRF:** Standard approaches to semantic image segmentation model pairwise relations between pixels with the P^2 Potts cost (red). Robust P^n generalises this to higher order costs for superpixels (green). AHRF generalises this further by also modelling a hierarchy of pairwise, and higher order relations between superpixels (yellow).

the results of tasks such as image denoising, restoration [39, 49], texture segmentation [31], object category segmentation [32]. The improvements can be attributed to the fact that higher order relations capture the fine details, including texture and contours, better than pairwise relations.

The AHRF approach to multi-resolution semantic image segmentation builds a hierarchical representation of the image from multiple unsupervised image segmentations, see Fig.1-4 for a sketch of these types of groupings. Whilst there are many ways in which such segments, or *super-pixels*, could be modelled as a CRF, AHRF follows an intuitive progression of generalisations that, at each step, maintain important properties of approximate MAP inference [5]. The base model is the contrast sensitive second order Potts model [3]: A commonly used CRF model that both encourages consistency in the labelling of neighbouring pixels, and attempts to preserve object boundaries in the presence of a contrast change in the image.

The first generalisation moves from pairwise, \mathcal{T}^2 Potts, to n^{th} , or higher order, \mathcal{T}^n Potts [31]. The idea being that each of the regions generated by the pre-segmentation will belong to a single object in the scene and should therefore be constrained to take the same label. However, this

hard region consistency constraint is too strong in practice. As often, a single segment may cross multiple object-class boundaries. The next generalisation adds a segment quality measure and allows for partial inconsistency, in the segment labelling, making for a more robust model—Robust \mathcal{T}^n Potts [32]—which paves the way for, the yet more general, model of AHRF. AHRF builds upon the weighted version of Robust \mathcal{T}^n by allowing for a discriminatively trained segment classifier, which gives a per-class score, rather than a uniform quality score over all labels for a segment. Further, the modelling of pairwise relations between neighbouring regions gives smoothness constraints on the segments, as-well-as the pixels. Finally a hierarchy of segmentations is modelled in such a way that consistency in super-segments, groups of segments, can be encouraged, as-well-as super-pixels. This generality subsumes many of the common CRF models defined over segments [37], and has enough expressive power to go beyond multi-resolution semantic image segmentation, towards holistic scene understanding.

1.2 Holistic Scene Understanding With AHRF

The AHRF graphical model framework has proven to be useful for multi-resolution image parsing. In this dissertation we extend the framework beyond semantic image segmentation towards holistic scene understanding. We demonstrate the efficacy of our thesis with applications to street scene understanding. We define three levels of architecture for fusing information from different sources directly into the graphical model: The Feature Level, that combines features from different modalities; The Potential Level, that fuses potential functions of the graphical model that have different semantic interpretations, over that of object classes; The Energy level, that co-joins graphical models for differing semantic tasks.

1.2.1 Feature level: Geometry and Appearance

In this application we present a framework for semantic image segmentation of road scenes that combines motion and appearance features. It is designed to handle street-level imagery

such as that on Google Street View and Microsoft Bing Maps. We formulate the problem in the AHRF framework. An extended set of appearance-based features is used, which consists of textons, colour, location and histogram of oriented gradient (HOG) descriptors. A boosting approach is then applied to combine the motion and appearance-based features. We evaluate our method both quantitatively and qualitatively on the challenging Cambridge-driving Labelled Video dataset [7]. Our approach shows an overall recognition accuracy of 84% compared a previous state-of-the-art accuracy of 69%.

1.2.2 Potential level: Things and Stuff

Computer vision algorithms for individual tasks such as object recognition, detection and segmentation have shown impressive results. The next challenge is to integrate all these algorithms and address the problem of scene understanding. This application takes a step towards this goal. In our work, we follow the definition of *things* and *stuff* by Forsyth *et al.* [15], where *stuff* is a homogeneous or reoccurring pattern of fine-scale properties, but has no specific spatial extent or shape, and a *thing* has a distinct size and shape. By relating the notion of *things* with object detectors, and *stuff* with segmentation [60], we can jointly reason about regions, objects, and their attributes such as object class, location, and spatial extent. Our model is a AHRF defined on pixels, segments and objects. We define a global energy function for the model, which combines results from sliding window detectors, and low-level pixel-based unary and pairwise relations. Experimental results show that our model achieves significant improvement over the baseline methods.

1.2.3 Energy level: Recognition and Reconstruction

The problems of *dense stereo reconstruction* and *object class segmentation* can both be formulated as CRF based labelling problems, in which every pixel in the image is assigned a label corresponding to either its disparity, or an object class such as road or building. While these two problems are mutually informative, no attempt has been made to jointly optimise their labellings.

In this work we provide a principled energy minimisation framework that unifies the two problems and demonstrate that, by resolving ambiguities in real world data, joint optimisation of the two problems substantially improves performance. To evaluate our method, we augment the street view Leuven data set, producing 70 hand labelled object class and disparity maps. We hope that the release of these annotations will stimulate further work in the challenging domain of street-view analysis.

1.3 Contributions

The key contributions of this dissertation are as follows:-

- ≡ We specify a hierarchy of modelling levels for holistic scene understanding with AHRF. (Chapter 3)
- ≡ We implement and demonstrate an application for each level of the modelling hierarchy. (Chapters 4 & 5)
 - Geometry and Appearance at the Feature level
 - *Things* and *Stuff* (Object Detection and Segmentation) at the Potential level
 - Recognition and Reconstruction at the Energy level
- ≡ We augment existing datasets in order for them to be better suited for the evaluation of Holistic Scene Understanding. (Chapter 5)

1.4 Document Map

In Chapter 2 the background of the AHRF model for semantic image segmentation is outlined, laying down the foundations for our fusion hierarchy for holistic scene understanding, presented in Chapter 3. Our original contributions, in Appendix A, are re-presented in Chapter 4, where we outline them w.r.t our modelling levels: The first application (§4.1) shows a method for

fusing together geometric and appearance features within the first level; the second application (§4.2) specifies how to combine sliding window object detectors with a per-pixel CRF within the second level—giving a holistic *things* and *stuff* system; In the third application (§4.3) we design a method of co-joining two, previously independently treated, CRFs within our third level. One of the CRFs is for object-class segmentation and the other for disparity estimation—giving a holistic recognition and reconstruction system. In Chapter 5 we present qualitative and quantitative results for each of the applications on several datasets. We then conclude in Chapter 6.

1.5 Publications

Original contributions (full text in Appendix A) that form our suite of holistic scene understanding applications:

Geometry and Appearance: Paul Sturges, Karteek Alahari, Lubor Ladicky, Philip H.S. Torr, *Combining Appearance and Structure from Motion Features for Road Scene Understanding*, Proceedings British Machine Vision Conference, 2009.

Things and Stuff: Lubor Ladicky, Paul Sturges, Karteek Alahari, Chris Russell, Philip H.S. Torr, *What, Where & How Many? Combining Object Detectors and CRFs*, Proceedings of the Eleventh European Conference on Computer Vision (ECCV), 2010

Recognition and Reconstruction: Lubor Ladicky, Paul Sturges, Chris Russell, Sunando Sen-gupta, Yalin Bastanlar, William Clocksin, Philip H.S. Torr, *Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction*, Proceedings British Machine Vision Conference (BMVC), 2010 (BMVA Best Science Paper Prize).

Chapter 2

Semantic Image Segmentation with Associative Hierarchical Random Fields

Semantic image segmentation describes the task of partitioning an image into regions that delineate meaningful objects and labelling those regions with an object category label [55]. AHRF [37] is a discriminative structured prediction framework that is designed to perform this task. This chapter explains the framework in order to provide the foundations of our holistic scene understanding hierarchy.

2.1 The Labelling Problem

Set Notation: We use upper-case letters, from the Greek and Roman alphabet, to denote a set: D is a set. We use the same letter in lower case for an element of that set: d is an element of D , $d \in D$. We reserve I , J and K for index sets, $I = \{i \in \mathbb{N} \mid 1 \leq i \leq n\}$. An ordered set is related to its index set by the shorthand D_I , D is indexed by I . If J is a subset of I we denote D_J the subset of D indexed by J . The *power set* $\mathbb{P}(D)$, is the set of all subsets of D . We use boldface to indicate that a variable $v \in V$ has taken on a value, $\mathbf{v} = 100$, and that a set of variables have all taken on their values $\mathbf{V}_{\{1,2,3\}} = \{100, 200, 10\}$.

2.1.1 Input

An input image \mathbf{I} is represented by an ordered set of indexes $I = \{i \mid \mathbb{N} \ 1 \leq i \leq w \bullet h\}$, and corresponding set of variables $D = \{d_i \mid i \in I\}$:

$$\mathbf{I} = \{I, \mathbf{D}\}. \quad \boxed{\text{Input Image}} \quad (2.1)$$

With or shorthand notation we may refer to the image variables directly as D_I . The variables take on colour values, e.g. $\mathbf{I}_{\{1,2,3\}} = \left\{ \begin{array}{|c|c|c|} \hline 255 & 200 & 175 \\ \hline \end{array} \right\}$. The set of all images \mathcal{L} forms the input space, e.g. the set of all $w \bullet h$ grey-scale images is $\mathcal{L} = [0, 255]^{w \times h}$ (input images need not be grey-scale, this is just an example).

2.1.2 Output

An output labelling \mathbf{L} is represented by an ordered set of indexes $I = \{i \mid \mathbb{N} \ 1 \leq i \leq w \bullet h\}$ and a corresponding set of label variables $V = \{v_i \mid i \in I\}$:

$$\mathbf{L} = \{I, \mathbf{V}\}. \quad \boxed{\text{Output Segmentation}} \quad (2.2)$$

With or shorthand notation we may refer to the label variables directly as V_I . The variables take on values from a set of m labels $\mathcal{C} = \{a_1, a_2, a_3, \dots, a_m\}$, e.g. in semantic image segmentation each label represents an object-class, $\mathcal{C} = \left\{ \begin{array}{|c|c|c|} \hline a_1 = car & a_2 = bus & a_3 = van \\ \hline \end{array} \right\}$. The output space is the product of output spaces of single variables $\mathbf{A} = \mathcal{C}_1 \bullet \mathcal{C}_2 \bullet \dots \bullet \mathcal{C}_{w \times h}$. A labelling is a particular assignment of a value to each of the variables $f_I : V_I \in \mathcal{C}$, or equivalently an element of the output space $f_I : V_I \in \mathbf{A}$, e.g. $f_{\{1,2,3\}} = \left\{ \begin{array}{|c|c|c|} \hline f_1(v_1) = a_1 & f_1(v_2) = a_2 & f_3(v_3) = a_2 \\ \hline \end{array} \right\} \leq \left\{ \begin{array}{|c|c|c|} \hline a_1 & a_2 & a_2 \\ \hline \end{array} \right\} / \mathbf{A}$.

2.1.3 Model

Let our variables for the input image D_I and output segmentation V_I be random variables. Let f^D and f be their respective assignment functions. A model of the joint probability over the input and output is a Gibbs distribution $P(f, f^D) = \frac{1}{Z(f, f^D)} e^{-E(f, f^D)}$, where $Z(f, f^D) = \sum_{\{f, f^D\}} e^{-E(f, f^D)}$ normalizes to a valid probability. To obtain a discriminative model, Bayes rule is applied giving $P(f | \mathbf{D}) = \frac{1}{Z(\mathbf{D})} e^{-E(f, \mathbf{D})}$, where $Z(\mathbf{D}) = \sum_f e^{-E(f, \mathbf{D})}$ normalizes. A Gibbs distribution factorises the global energy E into local potential functions ψ :

$$E(f; \mathbf{D}) = \sum_{c \in \mathcal{C}} \psi_c(f_c; \mathbf{D}) \quad \boxed{\text{Global Energy}} \quad (2.3)$$

where $\psi_{c \in \mathcal{C}} \subset 0$, and \mathcal{D} is a set of maximal cliques (A clique is a collection of variables which are all dependant on each other, and such a clique is maximal if it is not properly contained in any other clique). Due to Hammersley—Clifford this is equivalent to a CRF [38] graphical model with graph $\mathcal{H}(\cup_I, \mathcal{G}_{J \subset \{I \times I\}})$, where the vertices represent the variables and the edges the dependencies between them.

Prediction

Given the probabilistic model the prediction of an output segmentation is given by Maximum A Posteriori (MAP) estimation:

$$\begin{aligned} f^* &= \arg \max_f \frac{1}{Z(\mathbf{D})} e^{-E(f; \mathbf{D})} && \boxed{\text{MAP Estimate}} && (2.4) \\ &= \arg \max_f e^{-E(f; \mathbf{D})} \\ &= \arg \max_f E(f; \mathbf{D}) \\ &= \arg \min_f E(f; \mathbf{D}), \end{aligned}$$

i.e. minimising the global energy (2.3) is equivalent to a MAP estimation of the Gibbs distribution.

Inference

For multiple label problems the α -expansion move making algorithm [5] is an efficient method for approximate MAP inference. Given an arbitrary initial labelling, the algorithm minimizes the labelling cost function by making a series of changes (expansion moves) that iteratively decrease it. The algorithm terminates when no more moves can be made that will reduce the cost any further. At each step the move decreases the cost as much as possible (an *optimal* move). The optimal move can be computed quickly (in polynomial time) if the cost function satisfies metric constraints. It is proved, for energies with a maximum clique sizes of 2 (pairwise energies) [5] and higher order cliques [52], that if each expansion move satisfies metric constraints then it is as an optimal move, and any sequence of optimal moves converges to a bounded local optimum. The AHRF framework extensively exploits the efficiency and bounds of α -expansion algorithm for effective semantic image segmentation. Unless otherwise stated, all inference is performed with α -expansion through the entirety of this dissertation.

2.1.4 Representation

The global energy function (2.3) is represented by a factor graph. A factor graph is a bipartite graph that expresses how a global function of several variables factors into a product of local functions [16]. This makes it a perfect representation for a Gibbs factorisation. The energy term of our Gibbs distribution is defined in logarithmic space, thus the products of the factor graph correspond to additions in our case. Fig.2-1 depicts an example of a global energy $E(v_1, v_2, v_3, v_4, v_5) = \psi_{u1}(v_1) + \psi_{u2}(v_2) + \psi_{\rho1}(v_1, v_2) + \psi_{\rho2}(v_4, v_5) + \psi_{\eta1}(v_2, v_3, v_4)$, as a factor graph. It consists of two types of vertices: those associated with variables (the circles in Fig.2-1, called variable nodes) and those associated with local functions (the filled squares in Fig.2-1, called subset nodes). The edges of the factor graph are precisely those that join the variable

node for v_i to the subset node for ψ if and only if v_i is an argument of ψ [16].

Factor Graph: Let V_I be the output variables of a segmentation $L(I, V)$, indexed by I . Let C be a subset of the power set of I , $C \rightarrow \mathbb{P}(I)$ (not including the empty set). Suppose E can be written as a sum (log product) of local functions with arguments indexed by the elements of C . Then a factor graph representation of E (2.5) is a bipartite graph

$$\mathcal{H}(\cup_I, \Psi_C, \mathcal{G}_{I \times C}) \quad \boxed{\text{Factor Graph}} \quad (2.5)$$

with vertex set $I \cup C$ edge set $\{i, c\} : i \in I, c \in C, i \in c$. As stated earlier, we refer to those vertices that are elements of I as variable nodes and those vertices that are elements of C as subset nodes. An edge joins a variable node i to a subset node C if and only if $i \in C$, hence the factor graph is a graphical representation of the relation "element of" in $I \bullet C$. In the example, we have $I = \{1, 2, 3, 4, 5\}$, and $C = \{\{2\}, \{4\}, \{1, 2\}, \{4, 5\}, \{2, 3, 4\}\}$ [16].

Note that the global energy E is necessarily factorised into local factors (*potentials*) for tractable prediction. Consider $E(v_1, v_2, v_3, \dots, v_{10 \times 10}) = \psi_I(v_1, v_2, v_3, \dots, v_{10 \times 10})$ as our global energy for a tiny $10 \bullet 10$ input image, where ψ_I is the potential cost function defined over the whole image. With 10 labels $\mathcal{C} = \{a_1, a_2, \dots, a_{10}\}$, the size of output space \mathcal{M} is a googol—10 000—an absurdly large number. Since the potential is defined over all the variables we need to specify a cost for every one of these 10^{100} possible segmentations, which is impossible. Even if it were possible, we would then need to find the minimum! In contrast consider the same problem, but with a different factorisation $E(v_1, v_2, v_3, \dots, v_{10 \times 10}) = \psi_{u1}(v_1) + \psi_{u2}(v_2) + \psi_{u3}(v_3) + \dots + \psi_{u10 \times 10}(v_{10 \times 10})$. The output space is unchanged, yet now we only need to specify 100 local potential costs, one for each variable. The minimum cost segmentation is now trivially found by $f^* = (\arg \min_{a \in \mathcal{A}} \psi_{u1}(\mathbf{v}_1 = a), \arg \min_{a \in \mathcal{A}} \psi_{u2}(\mathbf{v}_2 = a), \arg \min_{a \in \mathcal{A}} \psi_{u3}(\mathbf{v}_3 = a), \dots, \arg \min_{a \in \mathcal{A}} \psi_{u10 \times 10}(\mathbf{v}_{10 \times 10} = a))$. These examples represent two extremes of complexity, from the impossible, to the trivial (from

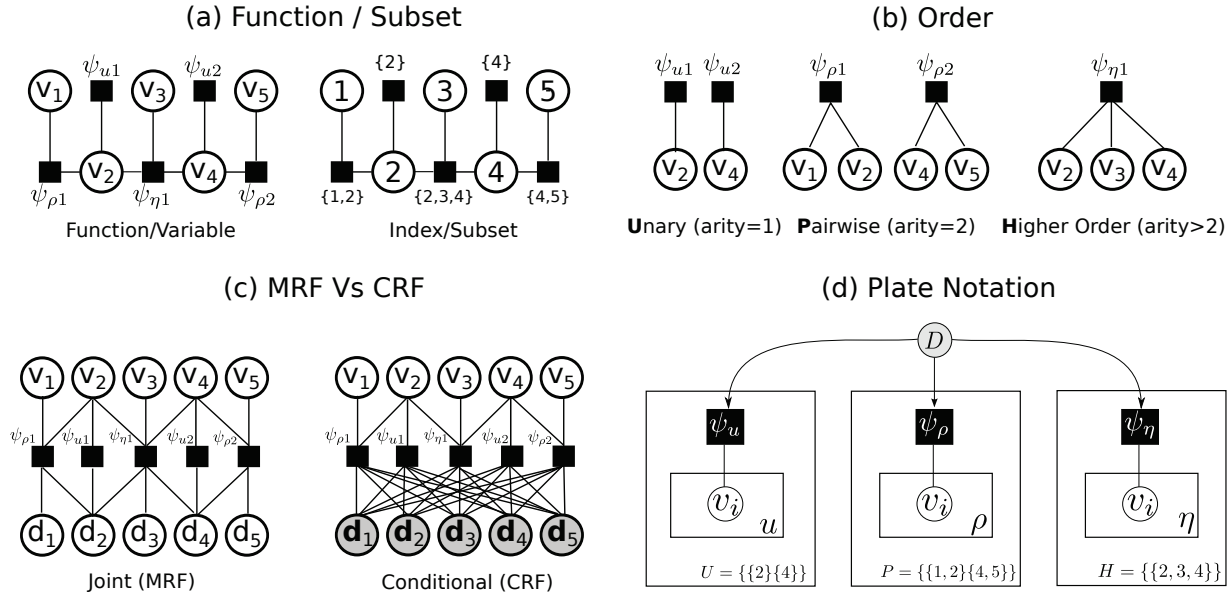


Figure 2-1: **Factor Graph:** An example of a factor graph. Factor nodes are shown as filled squares. Variables nodes are shown as white circles. Variables nodes with known label assignments are filled grey. (a) Shows that the factors can be seen as both functions over variables and subset relations. (b) The factors of the example grouped into common orders (of magnitude). (c) CRFs and MRFs as factor graphs (d) Plate notation for the example CRF factor graph split into orders.

exponential to linear in the number of variables), exemplifying the critical role that factorisation plays in semantic image segmentation.

2.2 Associative Higher Order Random Field (AHRF)

The orders of the factors in the AHRF factor graph are broken down into three special cases defined by their cardinality, unary (U), pairwise (P), and higher order (H), i.e.

$$E(f; \mathbf{D}, \Theta, B) = \Psi^U(f; \mathbf{D}, \Theta^U, \beta^U) + \Psi^P(f; \mathbf{D}, \Theta^P, \beta^P) + \Psi^H(f; \mathbf{D}, \Theta^H, \beta^H), \quad (2.6)$$

where $\beta^U, \beta^P, \beta^H > 0$ are hyper-parameters that serve as a practical method for tuning the bias's of each of the factor cardinalities. These bias may be introduced when employing a piecewise learning strategy of the model parameters $\Theta = \{\Theta^U, \Theta^P, \Theta^H\}$ (Fig.2-4, see *Texton-Boost* [56] for further details. We now specify the form of each order.

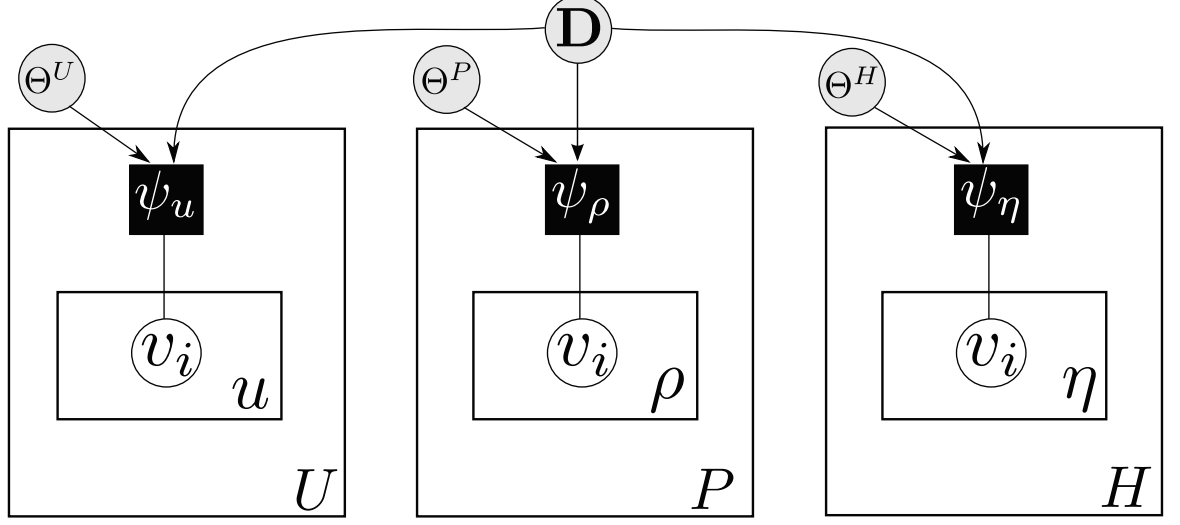


Figure 2-2: **AHRF Plate diagram:** The unary (U), pairwise (P) and higher order (H) factors of AHRF. Note that each order has its own set of parameters. Also note that factors are not independent sets.)

2.2.1 Unary Potentials

The unary sub-set of factors, U , for which each member, $u \in U$, has cardinality $\mathcal{G}_i = 1$, shares the model parameters Θ^U , is associated with a total cost:

$$\Psi^U(f; D, \Theta^U, \beta^U) = \beta^U \times \left\{ \psi_u(f_u; D, \Theta^U) \right\}_{u \in U}. \quad \boxed{\text{Unary Potentials}} \quad (2.7)$$

For semantic image segmentation the AHRF unary potentials assign a per-variable-per-label cost of:

$$\psi_u(f_u = \ell; D, \Theta^U) = \beta_\ell^u \times \log(\Pr(f_u = \ell | D)), \quad \boxed{\text{Unary Cost}} \quad (2.8)$$

where;

$$\Pr(f_u = \ell | D) = \exp(\mathcal{I}_\ell^U(D_u, \Theta_U)), \quad (2.9)$$

where β_ℓ^u is a parameter for compensating for per-label bias, for instance that may be caused by unbalanced data. The notation D_u is used (abused) to represent the data associated with factor $u \in U$ that can be any subset, or subsets, of the data [35, 38]—not just a single pixel value. \mathcal{I} is

a discriminatively trained classifier that outputs label confidences as positive real values (§ 2.3).

A graphical model for the labelling problem that is restricted to only have unary factors is referred to as a *first order approximation*. This has a trivial solution by setting all variables to their minimum cost label independently. Often this serves as a good initialisation for MAP inference, as is always the case in all experiments throughout this dissertation.

2.2.2 Associative Pairwise Potentials

The sub-set of factors, P , for which each member, $\rho \in P$, has cardinality $|\mathcal{G}_\rho| = 2$, shares the model parameters Θ^P , and is associated with a total cost of:

$$\Psi^P(f; D, \Theta^P, \beta^P) = \beta^P \times \prod_{\rho \in P} \psi_\rho(f_\rho; D, \Theta^P). \quad \boxed{\text{Pairwise Potentials}} \quad (2.10)$$

In semantic image segmentation the AHRF pairwise terms encourage smoothness in the labelling and take the form of a contrast sensitive Potts model [3, 56]:

$$\psi_\rho(f_\rho; D, \Theta^P) = w_\rho(D_\rho, \Theta^P) \times \mathcal{T}^2(f_\rho), \quad (2.11)$$

with

$$w_\rho(D_\rho, \Theta^P) = \theta_1^P + \theta_2^P \times \exp(-\theta_3^P \times G_\rho(D_\rho)), \quad \boxed{\text{Contrast}} \quad (2.12)$$

and

$$\mathcal{T}^2(f_\rho) = \begin{cases} 0 & \text{if } f_\rho[1] = f_\rho[2], \\ \sum 1 & \text{otherwise.} \end{cases} \quad \boxed{\mathcal{T}^2 \text{ Potts}} \quad (2.13)$$

The shared pairwise model parameters, $\Theta^P = \{\theta_1^P, \theta_2^P, \theta_3^P\} \subset 0$, are learned using training/validation data (§ 2.3), which compose the penalty for violating label smoothness. This encourages boundaries in the labelling to be consistent with edges in the image. The square brace notation is used to explicitly index the pair of variables in the pairwise factor. D_ρ is the

data associated with factor variables, that can be any subset, or subsets, of the data D [35]. In our case, $G(D_\rho) = \text{Dist}(D_\rho[1], D_\rho[2])$, conditions the smoothness on the distance between the colour vectors of only the two pixels in direct correlation with the pairwise factor, see [3, 50, 56] for more details.

A graphical model for the labelling problem that is restricted to only have unary and pairwise factors is referred to as a *second order approximation*. This is the most common approximation found in the literature. We encode a smooth world prior—the world does not change rapidly from point-to-point, but rather gradually, or smoothly—on these second order factors. We do this by imposing an Ising Lattice structure on the pairwise edge sets: Each variable has a pairwise factor with each of its 4 or 8, spatial nearest neighbours.

2.2.3 Associative Higher Order Potentials

The sub-set of factors, H , for which each member, $\eta \in H$, has cardinality $|\mathcal{G}_\eta| > 2$, shares the model parameters Θ^H , is associated with a total cost of:

$$\Psi^H(f; D, \Theta, \beta^H) = \beta^H \times \prod_{\eta \in H} \psi_\eta(f_\eta; D, \Theta^H). \quad \boxed{\text{Higher Order Potentials}} \quad (2.14)$$

Higher order factors have the capability to model complex interactions between more than two variables, loosening the restrictions to the representational power of the second order approximations, making them better suited for capturing the rich statistics of natural scenes [31]. In semantic image segmentation with AHRF object contiguity is captured through a pre-segmentation of the image and an associativity prior on the segments.

Object contiguity prior For the higher order factors we consider an object contiguity prior—objects in the world tend to form spatially contiguous regions, rather than being inter-dispersed across space—by restricting the set of factors to the corresponding regions/segments/super-pixels that are generated by pre-segmenting the input image, several times, using the mean-shift algorithm [9], with varying values of the parameters.

Associativity prior Given the set of contiguous higher order factors, an associativity prior—for any pair of variables the cost of the prior is lower (or the same) if they take the same label—is employed for each segment, η / H , in order to encourage a consistent labelling of the factors variables. This prior takes the form a generalised ($\eta > 2$ variables) Potts model [32] that measures the *inconsistency* of a segments labelling. Not all segments obtained using unsupervised segmentation (e.g. MeanShift [9]) are equally good, for instance, some segments may contain multiple object classes [32]. Therefore, some measure of segment *quality* is required. Also, the appearance of a image region can be used to discriminate between the object-class/s it contains [36,37]. Taking these properties into account, the total cost is composed of three parts:

inconsistency: A more/less inconsistent labelling of a segment will lead to a higher/lower cost.

Quality: A lower/higher quality segment will lead to a lesser/higher cost for an inconsistent labelling of those segment variables.

Confidence: A lower/higher confidence for a segment to take the label ℓ / \mathcal{M} will lead to a lesser/higher cost for a labelling inconsistent with ℓ .

The inconsistency, quality, and confidence measures are incorporated into the higher order potential for each segment as:

$$\psi_{\eta}(f_{\eta}; D, \Theta^H) = \overbrace{w_{\eta}(D_{\eta}, \Theta^H)}^{\text{Quality}} \times \overbrace{\tilde{\mathcal{T}}^n(f_{\eta})}^{\text{Inconsistency}} + \overbrace{w_{\eta}^{\ell}(D_{\eta}, \Theta^H)}^{\text{Confidence}}, \quad (2.15)$$

with

$$w_{\eta}(D_{\eta}, \Theta^H) = \theta_1^H + \theta_2^H \exp \quad \theta_3^H G_{\eta}(D_{\eta}) \{, \quad \boxed{\text{Seg. Variance}} \quad (2.16)$$

and

$$\mathcal{T}^n(f_{\eta}) = \begin{cases} \sum \frac{w_{\eta} - w_{\eta}^{\ell}}{Q} N_{\ell}(f_{\eta}) \eta^{\theta_4^H} & \text{if } N_{\ell}(f_{\eta}) \geq Q, \\ \eta^{\theta_4^H} & \text{otherwise.} \end{cases} \quad \boxed{\text{Potts } \mathcal{T}^n} \quad (2.17)$$

and

$$w_{\eta}^{\ell}(D_{\eta}, \Theta^H) = \beta_{\ell}^H \times \min(\quad \eta^{\theta_5^H} \log(\exp(\quad \mathcal{I}_{\eta}^{\ell}(D_{\eta}, \Theta^H)), \theta_{\alpha}^H) \quad \boxed{\text{Seg. Classifier}} \quad (2.18)$$

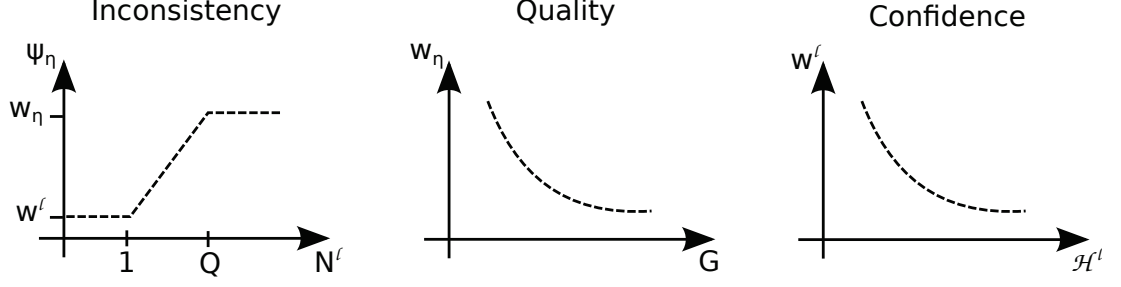


Figure 2-3: **Higher Order Segment Costs:** The segment cost (2.15) is broken down into a truncated, linearly increasing, inconsistency cost (2.17), and decaying exponential quality (2.16), and confidence costs (2.18).

The model parameters, $\{\theta_1^H, \theta_2^H, \theta_3^H, \theta_4^H, \theta_5^H\} \subset 0$, are learned using training/validation data. The robust label inconsistency penalty is a truncated linear function with truncation parameter, Q (satisfying $2Q < \eta$), that controls the the rigidity of the higher order clique potential, and $N_\ell(f_\eta) = \arg \min_{\ell \in \mathcal{L}} (\eta - n_\ell)$ is the number of variables in the clique η not taking the dominant label. β_l^H is a tuning parameter used to compensate for any label bias introduced by the discriminatively trained classifier \mathcal{I}_η^ℓ (§ 2.3), and θ_α^H is a truncation on the classifiers score. The potential takes the cost w_η^ℓ if all pixels in the segment take the label $\ell \in \mathcal{M}$. The potential costs are depicted in 2-3.

2.3 Learning

Piecewise learning is depicted in Fig. 2-4. Each order of factors, unary, pairwise, and higher-order, are treated independently of each other. For each order the model parameters, $\Theta = \{\Theta^U, \Theta^P, \Theta^H\}$, are trained with discriminative classifiers and/or cross-validation. This requires a dataset of factors with known member variables. Each factor in the dataset is treated as an independent sample. Once the discriminative classifiers are trained the model can be fine tuned with the bias', β , parameters.

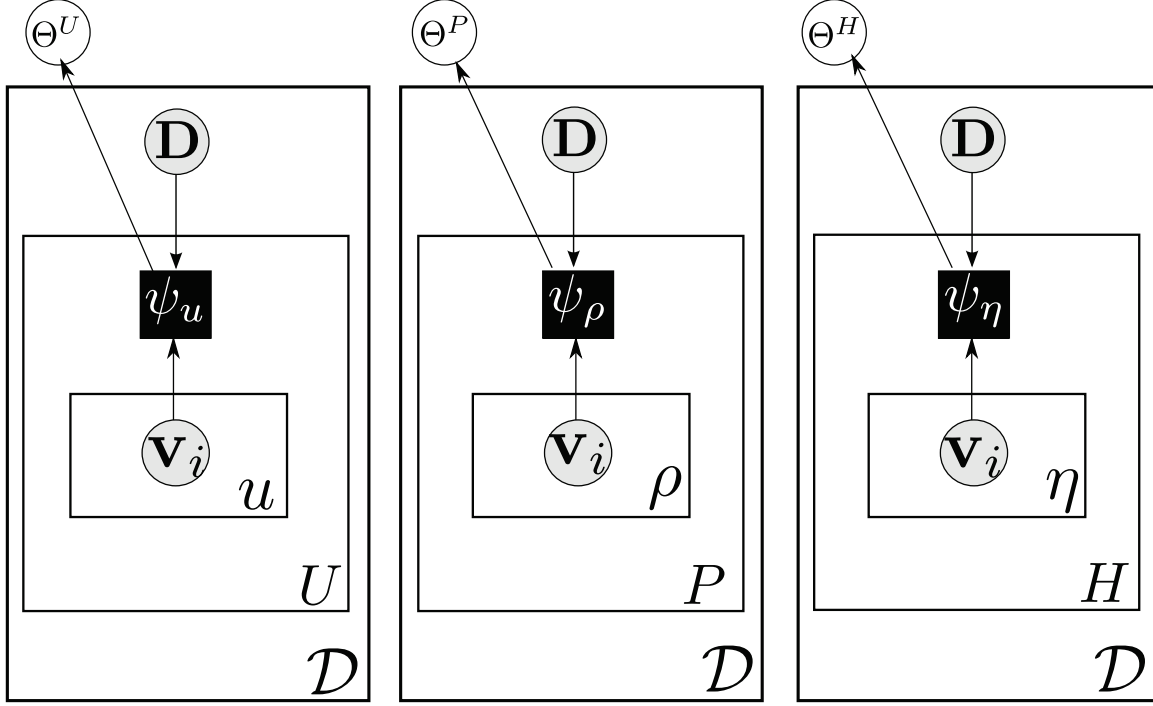


Figure 2-4: **Piecewise Learning Plate diagram** : During training all member variables are visible, and for each factor they are gathered into a dataset \mathcal{E} , the datasets \mathcal{E}^U , \mathcal{E}^P and \mathcal{E}^H are then used to train the models for their respective order of factors.

Classifiers \mathcal{I} : To learn the models responsible for assigning a label-wise confidence value, H , for the unary (2.8) and segment potentials (2.18), a boosting approach is employed. Boosting is an additive model that sums the classification confidence of M weak learners, h . The confidence value outputted by the strong classifier is then reinterpreted as our unary potential using the softmax transformation. A simple form of a boosted classifier could be:

$$H_\ell^U(D_u, \Theta^U) = \left\{ \begin{array}{l} h_m^\ell(D_u, \Theta_m^U), \end{array} \right. \quad \boxed{\text{Strong classifier}} \quad (2.19)$$

where

$$h_m^\ell(D_u, \Theta_m^U) = \theta_a^m \times \delta(D_u^i > \theta_t^m) + \theta_b^m, \quad \boxed{\text{Weak classifier}} \quad (2.20)$$

placed in a *one Vs rest* leaning schema, finally exposing the shared unary parameters $\Theta^U = \{\theta_a^m, \theta_b^m, \theta_t^m\}_{m=1}^M$ as parameters of the weak learner decision stumps. In practice a more sophis-

ticated multi-class boosting approach, with feature sharing, is employed, see *TextonBoost* [56] and references therein for details of the form, and of the learning procedure.

Fine tuning: Any model parameters Θ that are not learnt using the boosting approach, and the bias parameters β are learnt using a cross validation procedure, see Ladicky [36] for details.

2.4 Implementation

The complete AHRF framework is implemented by Ladicky [36] in object oriented c++ code. The library is available from the authors website. The *API* is flexible and allows for the addition of boosted unary potentials, and weighted robust \mathcal{T}^n pairwise and higher order potentials. The source includes α -expansion inference along with the necessary transformation of the energy function.

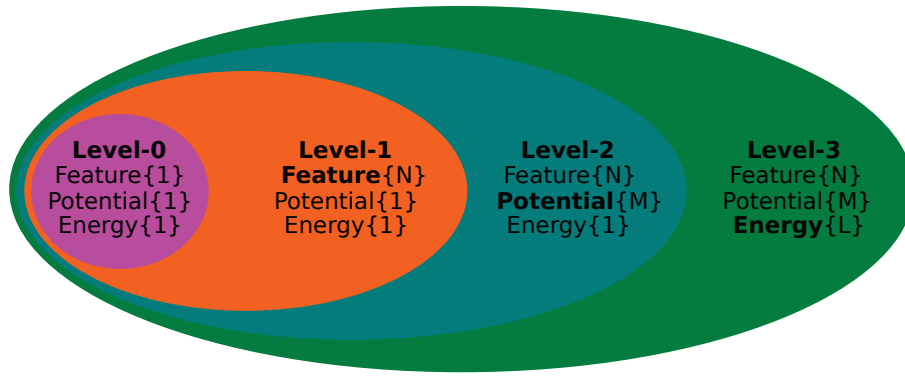
Chapter 3

Holistic Scene Understanding with Associative Hierarchical Random Fields

In this chapter we take AHRF beyond a multi-resolution semantic image segmentation framework, towards a configurable holistic scene understanding one by generalising the labelling problem factor graph to take multiple inputs, and give multiple outputs. We define a fusion hierarchy that with each level becomes more general than its predecessor. We assign a regular random field with a single input to our level-0, i.e.

$$E(f; \mathbf{D}) = \Psi^U(f; \mathbf{D}) + \Psi^P(f; \mathbf{D}) + \Psi^H(f; \mathbf{D}) \quad \boxed{\text{Level-0}} \quad (3.1)$$

as described in Chapter 2, and depicted with plate notation in Fig.2-2 (with parameters not shown). This is extended in level-1—The Feature Level—where we allow for multiple inputs to be fused. Level-2—The Potential Level—generalises this allowing for any-or-all combinations of the inputs to be fused via numerous factors. Ultimately, our level-3—The Energy Level—further generalises the labelling problem, of levels 0-to-2, to a multiple-labelling problem by allowing for many outputs. The specification of each of these levels, as depicted in Fig. 3-1, follows.



Level-3

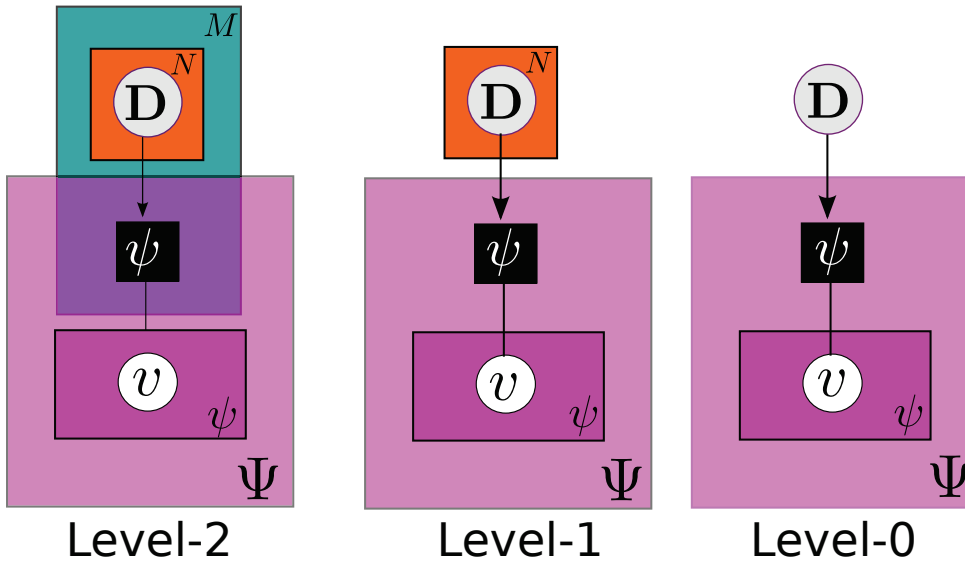
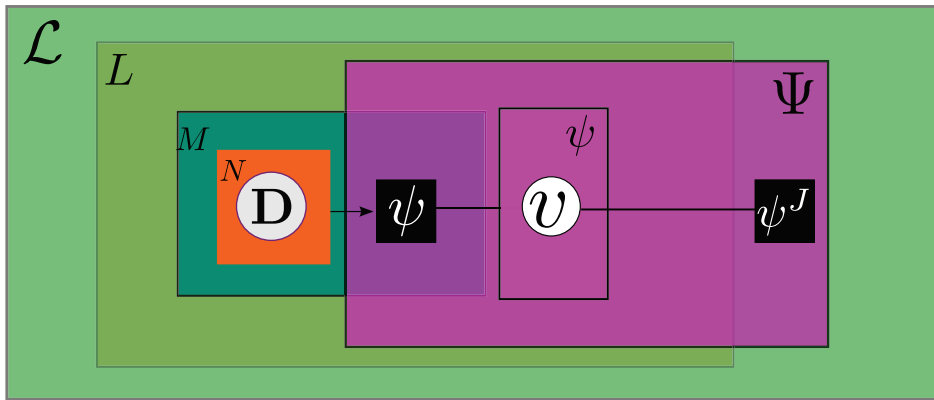


Figure 3-1: **Fusion Hierarchy:** Each level of our fusion hierarchy, depicted as ellipse, subsumes the levels that are shown to be contained within them. Level-0 is the most restrictive, contained within all the other levels, and only allows for one input, indicated by notation $\{1\}$. The feature level fuses multiple inputs, indicated by $\{N\}$. The potential level allows for a set of $\{M\}$ potentials to handle different sets of these fused inputs. The energy level is yet more expressive and subsumes both the potential and feature levels by allowing for L sets of output labels.

3.1 Level-1: The Feature Level

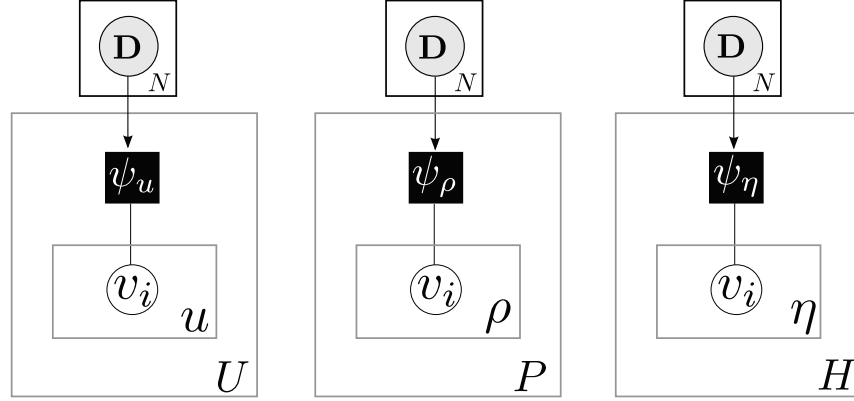


Figure 3-2: Level-1 Plate Diagram

In the feature level, fusion is performed on differing image features or modalities, that highlight some special properties, attributes, or features, of the input image, e.g. appearance and geometry features as in our application §4.1. We use the terminology *Feature Image*, rather than the more standard term *feature vector/descriptor*, to emphasise that we have a feature for every pixel location of the input image. Let $\mathcal{E}_A = \{\mathbf{D}_0, \mathbf{D}_1, \mathbf{D}_2, \dots\}$ be some finite set of feature images indexed by A , e.g. edge gradients, blobs, texture etc. Let \mathcal{P} index $\mathbb{P}(A)$, i.e. \mathcal{P} indexes the space of all possible combinations of the feature images. We denote a joint feature image as $\mathbf{D}|_N$, where $N \subseteq \mathcal{P}$ are the indexes for one particular subset of $\mathbb{P}(\mathcal{E})$, and $\mathbf{D}|_N$ is an arbitrary fusion operator for that set. Now the generalised form of the AHRF cost (2.6) with feature fusion is:

$$E^{\mathcal{N}}(f; \mathbf{D}) = \Psi^{UN}(f; \mathbf{D}|_{N^U}) + \Psi^{PN}(f; \mathbf{D}|_{N^P}) + \Psi^{HN}(f; \mathbf{D}|_{N^H}), \quad \boxed{\text{Level-1}} \quad (3.2)$$

where $N \subseteq \mathcal{P}$ are the index sets of the features to be fused for the unary, pairwise and higher order potentials respectively, and they need not be the same subset for each order (U,P,H) of potentials. Giving the raw image data the 0^{th} index, then $\{N^U, N^P, N^H\} = \{0\} \subseteq \mathcal{P}$ recovers level-0. The plate diagram is shown in Fig.3-2.

3.2 Level-2: The Potential Level

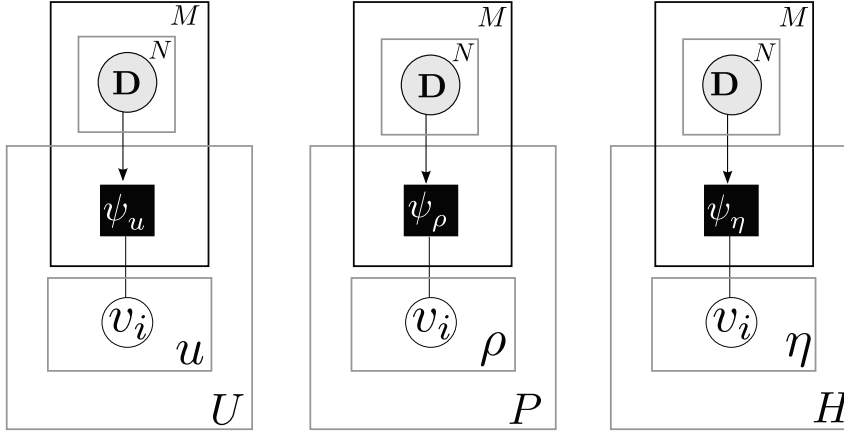


Figure 3-3: Level-2 Plate Diagram

A common approach to integrating different sources of information into a CRF is by defining a potential function for each one. For instance in TextonBoost [56] a unary potential is defined for texture, another for location, and a special type of higher order potential (different to the AHRF ones) is defined for local colour models. Taking the idea of having many potentials to the limit, we can model for all, or any combination of them. Recall that \mathcal{E}_A is a finite set of feature images indexed by A , and that $\mathcal{P} = \mathbb{P}(A)$. This allows us to have a potential defined on a subset of D_A . To generalise this to multiple potentials, each that are dependant on a subset of D_A , we index all those subsets by $\mathcal{N} = \mathbb{P}(\mathcal{P})$. Now $N / M / \mathcal{N}$ is an index set for one particular subset of $\mathbb{P}(\mathcal{E})$ as before, only now we can specify M subsets, rather than just 1 subset, i.e.

$$E^{\mathcal{M}}(f; \mathbf{D}) = \left\{ \right\} \Psi^{UN} \left(_{N \in M^U} + \left\{ \right\} \Psi^{PN} \left(_{N \in M^P} + \left\{ \right\} \Psi^{HN} \left(_{N \in M^H} \quad \boxed{\text{Level-2}} \right) \quad (3.3)$$

where $\Psi^{U/P/HN}$ are the layer-1 potentials (3.2). e.g. given a pair of feature images we could have a possible $2^{|M|}$ joint unary potentials $\Psi^U = \Psi_{\{1\}}^U(f; D_1) + \Psi_{\{2\}}^U(f; D_2) + \Psi_{\{1,2\}}^U(f; \rangle D_1, D_2 | + \Psi_{\{2,1\}}^U(f; \rangle D_2, D_1 |)$. The subsets $M \rightarrow \mathcal{P}$ need not be the same for each order of potentials. It can be easily seen that this subsumes (3.2) with $M = \rangle N \langle$ for all orders (U,P H) of potentials. The plate diagram is shown in Fig.3-3.

3.3 Level-3: The Energy Level

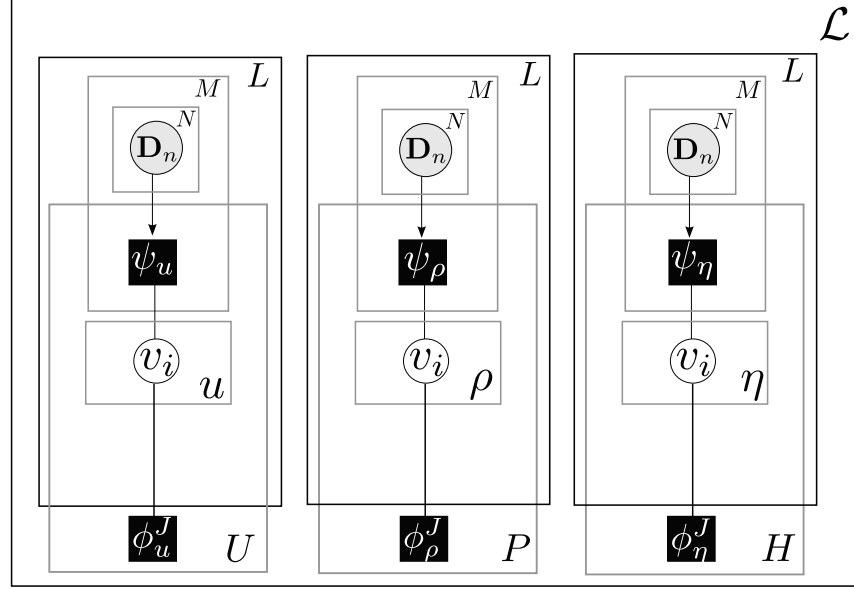


Figure 3-4: **Level-3 Plate Diagram:**

For fusion in the energy level, we provide a framework for fusing energies that share the same factorisation, but have different output spaces, e.g. such as object class and disparity in our recognition and reconstruction application §4.3. Let \mathcal{C}_L be a finite set of label sets indexed by L . Defining an energy for the Cartesian product of their label sets $\mathcal{C}_1 \bullet \mathcal{C}_2 \bullet \mathcal{C}_3, \dots$ can result in intractable MAP inference, because the size of the combined label set $\mathcal{C} \bullet \mathcal{C}' = \mathcal{C} \times \mathcal{C}'$ is too large. An alternative option is the Factorial CRFs [58]. Let $\mathcal{M} = L \bullet L$, giving an index set to all ordered pairs $([,])$ of \mathcal{C}_L . A (pairwise) factorial energy is then defined as:

$$E^{\mathcal{L}}(f^{\mathcal{L}}; \mathbf{D}) = \left\{ \right\} E^{\mathcal{M}^{\ell}} f^{\ell}; \mathbf{D} \left\{ \left(\left\{ \right\}_{\ell \in L} + \left\{ \right\} \epsilon^J \right) [f^{\ell}, f^{\ell'}]; \mathbf{D} \left(\left\{ \right\}_{[\ell, \ell'] \in \mathcal{L}} \right. \boxed{\text{Energy Level}} \quad (3.4)$$

where ℓ / L indexes the ℓ^{th} label set in L , and $E^{\mathcal{M}^{\ell}}$ is a level-2 energy (3.3) with label set \mathcal{C}_{ℓ} . Taking these alone would result in a set of independent labelling problems, one for each label set. However, our holistic approach requires that these interact with each other. The term ϵ^J , assigns a cost over a pair of label sets $\} \mathcal{C}_{\ell}, \mathcal{C}_{\ell'} \langle$ such that we can model these interactions, e.g. Take one energy for predicting foreground Vs background, and another for inferring near Vs

far. For the pairs, foreground and far, background and near, assign a higher cost. For the pairs, foreground and near, background and far, give a lower cost. This configuration would encourage a joint assignment in which foreground objects are nearer than background ones.

Recall that a factor ψ represents both a function/variable and index/subset relation (Fig. 2-1). Given two identically factored energies E^ℓ and $E^{\ell'}$, let the member variables of the joint factor ϕ be exactly those member variables of $\psi_i^\ell \{ \psi_i^{\ell'} \}$, i.e. ϕ joins the factors of E^ℓ with those of $E^{\ell'}$. The joint term of (3.4) is now defined as:

$$\epsilon) [f^\ell, f^{\ell'}]; \mathbf{D} \left(= \Phi^U \right) [f^\ell, f^{\ell'}]; \mathbf{D} \left(+ \Phi^P \right) [f^\ell, f^{\ell'}]; \mathbf{D} \left(+ \Phi^H \right) [f^\ell, f^{\ell'}]; \mathbf{D} \left(\quad (3.5)$$

where $\Phi^{U/P/H}$ are the sets of joint unary/pairwise/higher-order factors that are defined as follows. The plate diagram is shown in Fig.3-4.

Joint Unary Potentials: The set of joint factors $\} \phi_u = \psi_u^\ell \{ \psi_u^{\ell'} \}_{u \in U}$, where $\psi_u = 1$. The total joint cost is given by

$$\Phi^U([f^\ell, f^{\ell'}]; \mathbf{D}) = \left\{ \phi_u([f_u^\ell, f_u^{\ell'}]; \mathbf{D}) \right. \quad \boxed{\text{Joint Unary Potentials}} \quad (3.6)$$

Joint Pairwise Potentials: The set of joint factors $\} \phi_\rho = \psi_\rho^\ell \{ \psi_\rho^{\ell'} \}_{\rho \in P}$, where $\psi_\rho = 2$. The total joint cost is given by

$$\Phi^P([f^\ell, f^{\ell'}]; \mathbf{D}) = \left\{ \phi_\rho([f_\rho^\ell, f_\rho^{\ell'}]; \mathbf{D}) \right. \quad \boxed{\text{Joint Pairwise Potentials}} \quad (3.7)$$

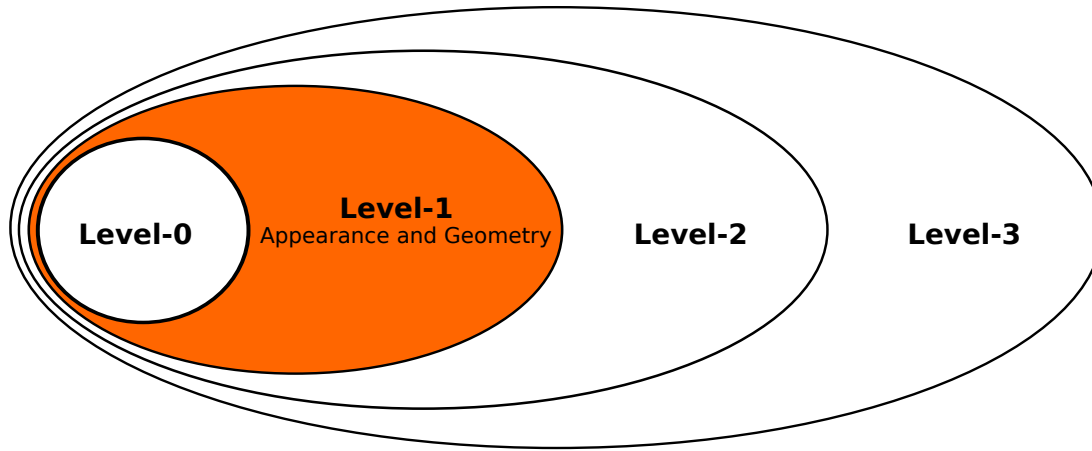
Joint Higher Order Potentials: The set of joint factors $\} \phi_\eta = \psi_\eta^\ell \{ \psi_\eta^{\ell'} \}_{\eta \in H}$, where $\psi_\eta > 2$. The total joint unary cost is given by

$$\Phi^H([f^\ell, f^{\ell'}]; \mathbf{D}) = \left\{ \phi_\eta([f_\eta^\ell, f_\eta^{\ell'}]; \mathbf{D}) \right. \quad \boxed{\text{Joint Higher Order Potentials}} \quad (3.8)$$

Chapter 4

Holistic Applications: Street Scene Understanding

With applications such as Google Street View, Microsoft Bing maps, the problem of street scene understanding has gained more importance than ever. For instance, in mapping applications, there is a need to identify *objects* in the scene in order to anonymise the data, remove transient objects, and localise important street furniture. Identifying these types of objects, such as people, cars, and signs, in street view imagery is challenging because the scenes consist of complex scenarios involving them. Yet, they are highly structured making them an interesting case study for structured prediction with AHRF. We experiment with our 3 level hierarchy for holistic scene understanding with a demonstration application for each level. For the feature level we fuse Appearance and Geometry cues §4.1. For the potential level we fuse things and stuff §4.2. For the energy level we fuse recognition and reconstruction §4.3.



4.1 Feature Layer: Appearance and Geometry

In this application we aim to exploit 3D geometric information to aid in the semantic image segmentation of monocular image sequences filmed from within a driven car. Our work is directly inspired by the contextual modelling of appearance in [56], and the demonstration of the power of 3D geometry for Semantic Image Segmentation of street scenes in [8]. In essence, we combine these two works by fusing appearance and geometry based features directly in the 1st layer of our hierarchy (§3.1).

4.1.1 Introduction

Image sequences from a moving car consist of complex scenarios involving multiple *objects*, such as people, buildings, and cars, making them challenging for purely appearance based Semantic Image Segmentation. For instance pedestrians may wear a large variety of clothing, and cars can be varying in colour, and the road may be made of many differently textured surface materials. Moreover, street scenes (even void of transient objects) that are constructed with the similar materials, will still vary in appearance under differing weather conditions, or even times of day. However the core geometry of the scene remains unscathed by sun, rain, wind or the daily passage of time. In fact the static elements of these scenes, such as the road, buildings, and sky, tend to each share a common 3D geometry, as well as geometric relations between them, across a large variety of geographic locations and wide spans of time; see fig.4-1 for demonstrative

examples. This is argued for and clearly demonstrated in [8], where they are able to produce an accurate semantic segmentation of a street scene purely from 3D geometric information, alone.



Figure 4-1: **Scene Geometry:** On the left hand side we see two street scenes from different geographic areas that strongly share geometric structure, but are not so similar in appearance. On the right side a photograph of the same location of Oxford, snapped in different decades, show strong structural similarity, despite the long gap in time and differing weather conditions.

Modelling complex variations in appearance is difficult. In TextonBoost [56] they combine features that encode the 2d location, texture and local colour of the object classes to tackle the problem. Furthermore, and importantly, they employ the *boosting trick* to learn a layout filter for the texture cues that encode spatial context. This proves to be vital for the Semantic Image Segmentation task. However their approach is developed for generic use, e.g. a cow photographed in a field, or a tourist attraction snapped by a holiday maker. Whereas here we work towards a specific task, i.e. a street scene viewed from within a driven car. These types of highly structured scenes have a more consistent 3D geometry, and geometric relations between the objects contained within them. Pedestrians and cars will be afforded by the ground plane, and the pavement upon which the pedestrians walk are parallel regions along the sides of the road (hence their alternative name *sidewalks*), of which the cars drive along. So along with appearance, we would also like to model the complex geometric properties and their contexts

that are clearly present in street scenes.

Extracting the underlying geometry of a scene is difficult, as it requires us to first obtain 3D information from 2d images. An interesting approach to this problem is presented in [28] with applications to automatic photo pop-ups [26], and holistic scene understanding [29]. In [28] they assume, inter alia, that a large percentage of a depicted scene can be represented by a small number of geometric labels: the ground plane, surfaces roughly perpendicular to the ground, and sky. Under this assumption they aim to statistically learn a mapping from a large set of customised image features to one of the 3 geometric labels, i.e. a labelling problem. There are a few technical issues in using their work for our application. Firstly, if we were to adopt their features, they are largely appearance based, thus including them would not achieve our goal w.r.t geometric information. Secondly, if we were to adopt their geometric labelling, the results are very coarse and thus not practical to further derive features from them. In fact, these labels would be more suitable for fusion within our 3rd layer (§ 3.3). But, even setting these difficulties aside, we note that [28] are actually tackling a harder problem than we require, as they only have a single 2d image, whereas we have a temporal sequence of them. Thus here, we rather follow [8], where the authors exploited Structure-From-Motion (SfM) to help ease this task, and provide more detailed geometric information.

We implement the five motion and structure features proposed by [8], namely: height above the camera, distance to the camera path, projected surface orientation, feature track density, and residual reconstruction error. They are computed from a sparse 3D point cloud obtained from a SfM procedure. As noted by [8], the five cues are tailored for the driving application and are invariant to camera pitch, yaw and perspective distortions.

In summary, our application is inspired by the works of [8,56], but differs in our contribution—we treat the problem as one of holistic scene understanding, which fuses multiple geometric- and multiple appearance-based features within the 1st layer of our proposed architecture (§ 3.1).

4.1.2 Global Configuration

We combine multiple appearance and multiple geometry features within layer-1 of our holistic scene understanding hierarchy (§ 3.1). Let us index the set of data, \mathbf{D} , with special cases: D_0 , the input image; $sfm = \{i \mid 1 \leq i \leq n\}$ —the geometric features; and $app = \{j \mid n+1 \leq j \leq m\}$ —the appearance features. Then, we can write this applications configuration as:

$$\begin{aligned}
 E(f; \mathbf{D}) = & \Psi^U(f; \mathbf{D})|_{\{sfm \cup app\}} \left\{ \begin{array}{l} \boxed{\text{Level-1}} \end{array} \right. & (4.1) \\
 & + \Psi^P(f; \mathbf{D}_0, \cdot) \left\{ \begin{array}{l} \boxed{\text{Level-0}} \end{array} \right. \\
 & + \Psi^H(f; \mathbf{D})|_{\{app\}} \left\{ \begin{array}{l} \boxed{\text{Level-1}} \end{array} \right.
 \end{aligned}$$

where our contribution is focused on the combination of appearance and geometric features within our level-1 fusion, which we have chosen to fuse for the unary factors only. (The pairwise factors depend on the raw image data. The higher-order factors depend on the set of appearance features.).

4.1.3 Appearance

We now describe the appearance-based features employed in our framework. In contrast to [8], which uses only texon histograms and localized bag of semantic texon (BOST) features, our approach uses colour, location, texon, and Histogram of Oriented Gradients (HOG) [11] features.

4.1.4 Geometry

We use the five motion and structure features proposed by [8], namely: height above the camera (m^H); distance to the camera path (m^C); projected surface orientation (m^O); feature track density (m^D); and residual reconstruction error (m^R). For a detailed description of the motion-based features, structure-from-motion pipeline, and the projection of features from 3D to 2D see [8], here we present a summary of their raw features along with their intuitions.

Height Height above the camera is measured as the difference of the y coordinates of a world point and the camera centre, after aligning the car’s *up* vector as the camera’s y axis.

Camera path Distance to the camera path is computed using the entire sequence of camera centres. Let $C(t)$ denote the camera centre in frame t , and W denote a world point. This feature is defined as $\min_t \|W - C(t)\|$.

Surface orientation The surface orientation at any given 3D scene point is estimated from the 2D Delaunay triangles [54] formed using the projected world points in a frame. The intuition behind the orientation features is that although individual 3D coordinates may have inaccurate depths, the relative depths of the points gives an approximate local surface orientation.

Track density The track density feature exploits the well-known fact that objects yield sparse or dense feature tracks based on how fast they are moving, and their texture. For instance, trees, buildings, and other forms of vegetation yield dense feature tracks, while sky and roads give rise to sparse feature tracks. This cue is measured as the 2D map of the feature density.

Backprojection error The residual reconstruction error measures the backprojection error (2D variance) of the estimated 3D world points. This residual error separates moving objects such as people and cars, from stationary ones such as buildings, vegetation, and roads.

4.1.5 Joint Appearance and Geometry

Texton Coding

The original Textons are special filter banks designed to be combined in order to recognize textures (see. *What are Textons?* [73]). The textonisation process (Texton Coding) is also applied to other types of raw features. K -means clustering is performed to quantise each of feature types independently (All features types are whitened to zero mean and unit variance prior to cluster-

ing). We say that¹ the outcome of the clustering process, the cluster centres, is a *dictionary*, T_j , with $\{t_k\}_{k \in K_j}$ *words* for the j^{th} feature type. A texton encoded pixel, t_i / I , is then a nearest neighbour assignment from the feature/cue value to the nearest word in the dictionary:

$$t_i = NN(Q_i, T) = \min_{j \in T} (Dist(Q_i, T_j)) , \quad \boxed{\text{Texton Encoding}} \quad (4.2)$$

where, $Dist$, is the multidimensional Euclidean distance, and Q_i is of the same dimensionality as T_j .

Contextual Texton Pooling

Contextual Texton Pooling for semantic image segmentation is proposed in TextonBoost [56] under the name texture-layout filters. Here we motivate, and re-define them, such that they fit within the pooling stage of a the more widely adapted image classification pipelines, and our layer-1 fusion.

Pooling textons with histogram aggregation over a pooling region, $r(w, h)$, that is the same size as a texton image (4.2), is, in essence, a Bag-of-words representation, as is commonly used in image classification/retrieval pipelines [72]. However, here we are interested in per-pixel, not per-image, classification. A possible work-around is to define a pooling region for each pixel, relative to its co-ordinates, $r(x, y, offset(x, y))$ —giving a bag-of-words-per-pixel. An eloquent adaptation, but it is inadequate to represent the (possibly complex) contextual information that we require for accurate semantic image segmentation. This is because it is a *bag*—an or order-less collection—and thus invariant (by design) to the spatial organisation of the textons within the pooling window. Nevertheless, the size of the region r , defined by its offsets, can be considered as a contextual support: the larger the region, the more context is considered. This is a reasonable counter argument for capturing wider-context with a per-pixel-bag-of-words repre-

¹following conventions in bag of words (BoW) image classification, which in tern follow on from natural language document classification using BoW.

sensation, without the need for spatial layout, but it is also flawed. The flaw is a subtle one, and is best described with a simple example.

Take an object of interest, such as a car, we would like to identify all the pixels that together depict it. Now, it is not far fetched to consider that the parts of the car and other nearby objects, such as the road, could help to classify the pixel correctly. This would necessitate a large region of support, such that given any car pixel it would capture the car wheel and the road. However, such a large support window placed at pixel near to the car, but not belonging to the car, would have a very similar bag-of-words, making it difficult to distinguish the two objects, especially around the objects boundary locations. To overcome this *contextual confusion*, the context support region would be necessarily small, and forced to disregard the wider-context of the vehicle altogether, flawing the original argument for its use to capture wider-context.

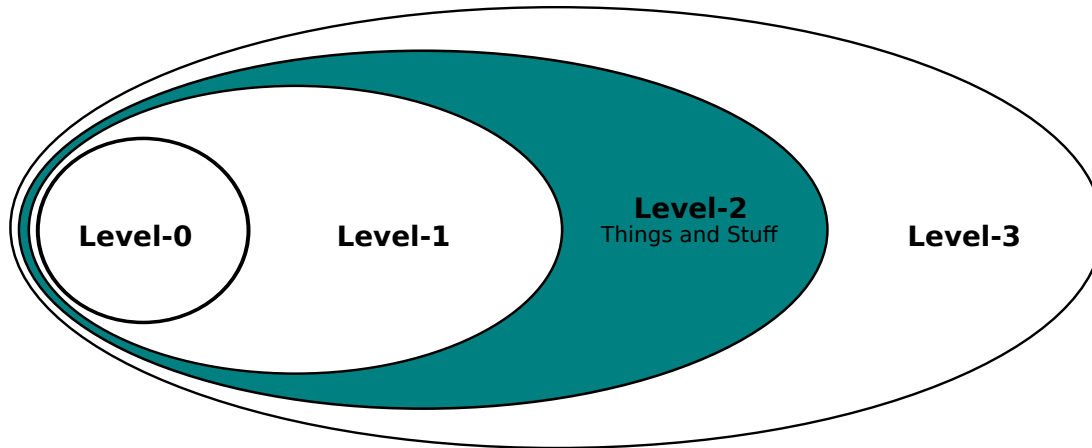
One way to overcome these issues is to allow for multiple contextual support regions, $R = \{r_1, r_2, \dots, r_n\}$, each, on its own, being the same as before, but when combined can represent the spatial layout via the relative positions (the offsets) of the regions. Given a fixed ordering over R , simple concatenation of the BoW histograms, for each region, encodes the layout of the multiple regions. Furthermore, the problem of contextual confusion can be disambiguated via the power of committee, since we can allow for many sized regions, from small to large.

Incorporating spacial layout into the pooling process is termed *contextual pooling*, a well know example in the image classification domain is the spatial pyramid representation [41], that models the coarse layout of scene types; in the face detection domain, the Haar wavelet like contextual pooling of the Viola-Jones face detector proved to be effective at modelling the layout of facial features [64] such as eyes, mouth and nose. In Semantic Image Segmentation TextonBoosts [56] texture-layout filters, have proven to be a powerful representation [12]. These are extended to include multiple appearance features, in [37, 57]. Here we further extend the approach to include multiple appearance and multiple geometric cues. Simply concatenating a large number of histograms for each window would lead to very high dimensional context features. To overcome this the boosting trick is employed for feature selection.

Boosting Trick

We use an adapted version [37] of the boosting approach described in TextonBoost [56] for feature selection and learning for the unary potentials of our model, unlike [8] which uses a randomized decision forest. The shape filters are defined by a rectangular region r and texton t pair. The feature response $v_i(r, t)$ of the shape filter for a given point i is the number of textons of type t in the region r placed relative to the point i . These filters capture the contextual relationships between objects. Each weak classifier compares the (shape filter) response to a threshold. The most discriminative filters are found using the Joint Boosting algorithm [59].

The classifiers defined on geometric and appearance-based features are combined in an adapted boosting approach. The shape filters are now defined by triplets of feature type f , feature cluster t , and rectangular region r . The feature response $v_i(r, f, t)$ for a given point i is the number of features of type f belonging to cluster t in the region r . The weak classifiers compare the responses of shape filters with a set of thresholds. The feature selection and learning procedure is identical to that in [56]. The negative log likelihood given by the classifier is incorporated as the unary potential in the CRF framework as defined in §3.1.



4.2 Potential Layer: Things and Stuff

In this application we aim to address the problems of *what*, *where*, and *how many*: we recognize objects, find their location and spatial extent, segment them, and also provide the number of instances of objects. This problem is particularly challenging in scenes composed of a variety of classes. For instance, road scene datasets [8] contain classes with specific shapes such as people and cars, and background classes such as the sky and grass lawns, which lack a distinctive shape [60] (Figure 4-2). Our holistic solution to this classical recognition problem involves firstly adopting the notion of *things* and *stuff* from Adelson [1], and then relating these notions to the applications of detection and segmentation [60]. In essence we fuse together bounding box detectors [14] and Semantic Image Segmentation [37] within the 2nd layer of our hierarchy (§3.2).

4.2.1 Introduction

The distinction between the two special sets of object classes—*things* and *stuff*—is well known [1, 15, 25]. Adelson [1] emphasized the importance of studying the properties of *stuff* in early vision tasks. Recently, these ideas are being revisited in the context of the new vision challenges, and have been implemented in many forms [25, 47, 60, 61]. In our work, we follow the definition by Forsyth *et al.* [15], where *stuff* is a homogeneous or reoccurring pattern of fine-scale properties, but has no specific spatial extent or shape, and a *thing* has a distinct size and shape. The dis-

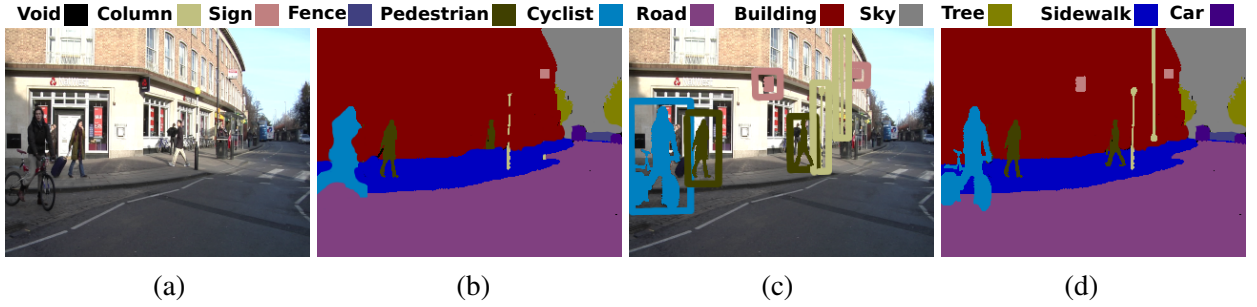


Figure 4-2: **A conceptual view of our method:** (a) An example input image. (b) Object class segmentation result of a typical CRF approach. (c) Object detection result with foreground/background estimate within each bounding box. (d) Result of our proposed method, which jointly infers about objects and pixels. Standard CRF methods applied to complex scenes as in (a) underperform on the “things” classes, e.g. produce inaccurate segmentation of the bicyclist and persons, and misses a pole and a sign, as seen in (b). However, object detectors tend to perform well on such classes. By incorporating these detection hypotheses, shown in (c), into our framework, we aim to achieve an accurate overall segmentation result as in (d). **(Best viewed in colour)**

inction between these classes can also be interpreted in terms of localization. *Things*, such as cars, pedestrians, bicycles, can be easily localized by bounding boxes unlike *stuff*, such as road, sky [60]².

Complete scene understanding requires not only the pixel-wise segmentation of an image, but also an identification of object instances of a particular class. Consider an image of a road scene taken from one side of the street. It typically contains many cars parked in a row. Object class segmentation methods such as [8, 37, 56] would label all the cars adjacent to each other as belonging to a large car segment or blob, as illustrated in Figure 4-4. Thus, we would not have information about the number of instances of a particular object—car in this case. On the other hand, object detection methods can identify the number of objects [14, 62], but cannot be used for background (*stuff*) classes.

A few object detection methods have attempted to combine object detection and segmentation sub-tasks, however they suffer from certain drawbacks. Larlus and Jurie [40] obtained an initial object detection result in the form of a bounding box, and then refined this rectangular

²Naturally what is classified as things or stuff might depend on either the application or viewing scale, e.g. flowers or trees might be things or stuff.

region using a CRF. A similar approach has been followed by entries based on object detection algorithms [14] in the PASCAL VOC 2009 [12] segmentation challenge. This approach is not formulated as one energy cost function and cannot be applied to either cluttered scenes or *stuff* classes. Furthermore, there is no principled way of handling multiple overlapping bounding boxes. Tu *et al.* [61] also presented an effective approach for identifying text and faces, but leave much of the image unlabelled. Gu *et al.* [22] used regions for object detection instead of bounding boxes, but were restricted to using a single over-segmentation of the image. Thus, their approach cannot recover from any errors in this initial segmentation step. In comparison, our method does not make such *a priori* decisions, and jointly reasons about segments and objects.

The work of layout CRF [66] also provides a principled way to integrate things and stuff. However, their approach requires that things must conform to a predefined structured layout of parts, and does not allow for the integration of arbitrary detector responses. Other existing approaches that attempt to jointly estimate segmentation and detection in one optimization framework are the works of [21, 67]. However, the minimization of their cost functions is intractable and their inference methods can get easily stuck in local optima. Thus, their incorporation of detector potentials does not result in a significant improvement of performance. Also, [21] focussed only on two classes (cars and pedestrians), while we handle many types of objects. Joint learning of things and stuff, and the relations between them is presented in [60] within a boosted CRF framework. They propose one representation, for both things and stuff, which is a parametrised convolution model that can represent texture (for stuff) at a fine scale, and templates (for things) at a courser scale. Whilst here, we adopt the current state-of-the-art representations of [14, 37], and then combine them—getting the best from both worlds.

We define a global energy function, modelling within the 2^{nd} layer of our hierarchy for holistic scene understanding (§ 3.2), which combines results from detectors (Figure 4-2(c)), mid-level cues such as superpixels, and low-level pixel-based unary and pairwise relations (Figure 4-2(b)). We also show that, unlike [21, 67], our formulation can be solved efficiently using graph cut based move making algorithms.

4.2.2 Global Configuration

We build upon the layer-1 application, where we defined index sets sfm and app for the geometric and appearance features that we combined, see § 4.1 for details. Here, we combine segmentation and detection within layer-2 of our holistic scene understanding hierarchy (§ 3.2). Under the weak assumption that a box that bounds an object of interest has a larger area than 2-pixels, we model the bounding box detections as higher-order factors. Let us index the set of higher-order factors, with special cases: $S = \{s \mid \mathbb{N} \ 0 \leq s \leq n - 1\}$ —the super-pixels from the base model (Chapter 2); and $B = \{b \mid \mathbb{N} \ n \leq b \leq m - 1\}$ —the object detection bounding boxes. Then, we can write this applications configuration as:

$$\begin{aligned}
 E(f; \mathbf{D}) = & \Psi^U(f; \mathbf{D})|_{\{sfm \cup app\}} \left\{ \begin{array}{l} \text{Level-1} \end{array} \right. & (4.3) \\
 & + \Psi^P(f; \mathbf{D}_0) \left\{ \begin{array}{l} \text{Level-0} \end{array} \right. \\
 & + \left\{ \right\} \Psi^H(f; \mathbf{D}) \left(\begin{array}{l} \text{Level-2} \end{array} \right)_{\{things \cup stuff\}}
 \end{aligned}$$

where our contribution is focused on the inclusion of detectors as higher order factors within our level-2 fusion.

4.2.3 Things

A *thing* has a distinct size and shape. Things are represented by a set bounding box object detections. They are included in the form of a higher order potential over pixels based on detector responses. In order to jointly estimate the class category, location, and segmentation of objects, we augment the standard CRF using responses of one of the most successful detectors on the PASCAL VOC 2009 dataset [14]. We retrain the models on the CamVid dataset for our application. Other detector methods could similarly be incorporated into our framework. In [14] each object is composed of a set of deformable parts and a global template. Both the global template and the parts are represented by HOG descriptors [11], but computed at a coarse and fine level respectively. The task of learning the parts and the global template is posed as a latent

SVM problem, which is solved by an iterative method.

This method produces results as bounding boxes around the detected objects along with a score, which represents the likelihood of a box containing an object. Let B 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 propose a novel potential ψ_b over the set of pixels \mathbf{v}_b belonging to the b^{th} detection (*e.g.* pixels within the bounding box), such that Ψ^{things} (4.3) is defined as:

$$\epsilon^{things}(f, \mathbf{D}_{\{hog\}}) = \left\{ \psi_b(\mathbf{v}_b, H_b, l_b) \right\}_{b \in B} \quad (4.4)$$

with a score H_b and detected label l_b . The full space of possible detections can be very large, however in a SVM-based classifier most of the responses are *ve*, hence the parameter H_t (which defines the detector threshold w.r.t H_b , equation (4.9)) can be set to 0 eliminating a large set of potentials from the problem. In practice a more accurate set of pixels belonging to the detected object is obtained using local foreground and background colour models [50].

4.2.4 Stuff

stuff is a homogeneous or reoccurring pattern of fine-scale properties, but has no specific spatial extent or shape. We model the recognition of *stuff* with a layer-1 energy over appearance and geometric cues. It is identical to the previous application §4.1, and is summarised here as:

$$\epsilon^{stuff} = \Psi^U + \Psi^P + \Psi^{stuff} \quad (4.5)$$

where Ψ^U and Ψ^P are the pixel based potentials, and Ψ^{stuff} are the segment based potentials.

pixel-based potentials.

The pixel-based unary potential estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. Shape filters are defined by

triplets of feature type, feature cluster, and rectangular region. Their response for a given pixel is the number of features belonging to the given cluster in the region placed relative to the given pixel. The most discriminative filters are found using the Joint Boosting algorithm [59]. Details of the learning procedure are given in [37, 56]. To enforce local consistency between neighbouring pixels we use the standard contrast sensitive Potts model [3] as the pairwise potential on the pixel level.

Segment-based potentials.

We also learn unary potentials for the higher order factors which that represent segments. The segment unary potential is also learnt using the Joint Boosting algorithm [59]. The pairwise potentials in higher layers (*e.g.* pairwise potentials between segments) are defined using a contrast sensitive (based on distance between colour histogram features) Potts model. We refer the reader to [37] for more details on these potentials and the learning procedure.

4.2.5 Things and Stuff

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). Existing methods [23, 37, 70] are limited to such hypotheses provided by pixels and/or segments only. We introduce an additional set of hypotheses representing object detections for the recognition framework³.

Some object detection approaches [14, 40] have used their results to perform a segmentation within the detected areas⁴. These approaches include both the true and false positive detections,

³Note that our model chooses from a set of given detection hypotheses, and does not propose any new detections.

⁴As evident in some of the PASCAL VOC 2009 segmentation challenge entries.

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 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. We illustrate this with one of our results shown in Figure 5-7. Here, the false positive detection for “person” class (shown as the large green box on the right) does not affect the segmentation result in (c). Although, the true positive detection for “car” class (shown as the purple box) refines the segmentation because it agrees with other hypotheses. This is achieved by using the object detector responses⁵ to define a clique potential over the pixels, as described below. Figure 4-3 shows the inclusion of this potential graphically on a pixel-based CRF. The new energy function is given by:

$$E(f) = E_{stuff}(f) + E_{things}(f) \quad (4.6)$$

where E_{stuff} (4.5) is a standard energy for semantic image segmentation (see §2) and E_{things} (4.4) is our novel detector based cost. 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 $y_b \in \{0, 1\}$, which takes value 1 to indicate the acceptance of b -th detection hypothesis. Let ϕ_b be a function of this variable and the detector response. Thus the detector potential $\psi_b(\cdot)$ is the minimum of the energy values provided by including ($y_b = 1$) and excluding ($y_b = 0$) the detector hypothesis, as given below:

$$\psi_b(\mathbf{x}_b, H_b, l_b) = \min_{y_b \in \{0, 1\}} \phi_b(y_b, \mathbf{v}_b, H_b, l_b). \quad (4.7)$$

We now discuss the form of this function $\phi_b(\cdot)$. If the detector hypothesis is included ($y_b = 1$), it should: (a) Encourage consistency by ensuring that labellings where all the pixels in \mathbf{v}_b take the label l_b should be more probable, *i.e.* the associated energy of such labellings

⁵This includes sliding window detectors as a special case.

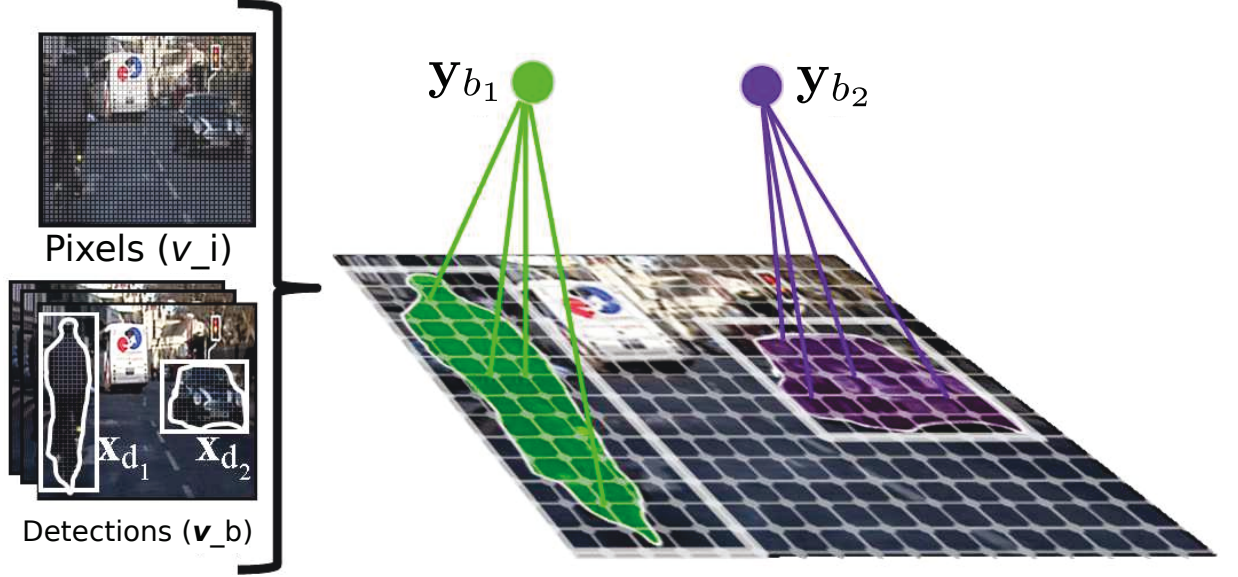


Figure 4-3: **Inclusion of object detector potentials into an AHRF:** We show a pixel-based CRF as an example here. The set of pixels in a detection b_1 (corresponding to the bicyclist in the scene) is denoted by \mathbf{v}_{b_1} . A higher order clique is defined over this detection window by connecting the object pixels \mathbf{v}_{b_1} to an auxiliary variable $y_{b_1} \in \{0, 1\}$. This variable allows the inclusion of detector responses as soft constraints. **(Best viewed in colour)**

should be lower; (b) Be robust to partial inconsistencies, *i.e.* pixels taking a label other than l_b 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_b . This allows objects partially occluded to be correctly detected and labelled.

To enable a compact representation, we choose the potential ψ_b such that the associated cost for partial inconsistency depends only on the number of pixels $N_b = \sum_{i \in \mathbf{v}_b} \delta(v_i \neq l_b)$ disagreeing with the detection hypothesis. Let $f(\mathbf{v}_b, H_b)$ define the strength of the hypothesis and $g(N_b, H_b)$ the cost taken for partial inconsistency. The detector potential then takes the form:

$$\psi_b(\mathbf{v}_b, H_b, l_b) = \min_{y_b \in \{0, 1\}} (f(\mathbf{v}_b, H_b)y_b + g(N_b, H_b)y_b). \quad (4.8)$$

A stronger classifier response H_b 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_b . As we also wish to penalize inconsistency, the function $g(\cdot)$ should be monotonically increasing with respect to N_b . 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:

$$f(\mathbf{v}_b, H_b) = w_b \mathbf{v}_b \max(0, H_b - H_t), \quad (4.9)$$

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

$$g(N_b, H_b) = k_b N_b, \quad k_b = \frac{f(\mathbf{v}_b, H_b)}{p_b \mathbf{v}_d}, \quad (4.10)$$

where the slope k_b was chosen such that the inconsistency cost equals $f(\cdot)$ when the percentage of inconsistent pixels is p_b .

Detectors may be applied directly, especially if they estimate foreground pixels themselves. However, we use sliding window detectors that provide a bounding box around objects. To obtain a more accurate set of pixels \mathbf{v}_b that belong to the object, we use a local colour model [50] to estimate foreground and background within the box. This is similar to the approach used by submissions in the PASCAL VOC 2009 segmentation challenge. Any other foreground estimation techniques may be used.

4.2.6 Inference

One of the main advantages of our framework is that the associated energy function can be solved efficiently using graph cut [4] based move making algorithms (which outperform message passing algorithms [13, 33] for many vision problems). We now show that our detector potential in equation (4.8) can be converted into a form solvable using $\alpha\beta$ -swap and α -expansion



Figure 4-4: **Counting Things:** (a) An Object class segmentation labels all the cars adjacent to each other as belonging to one large blob. (b) Detection methods localize objects and provide information about the number of objects, but do not give a segmentation. (c) Our method jointly infers the number of object instances and the object class segmentation. See §4.2.6 for details. **(Best viewed in colour)**

algorithms [5]. In contrast, the related work in [21] suffers from a difficult to optimize energy. Using equations (4.8), (4.9), (4.10), and $N_b = \sum_{i \in \mathbf{v}_d} \delta(v_i \neq l_b)$, the detector potential $\psi_b(\mathbf{x})$ can be rewritten as follows:

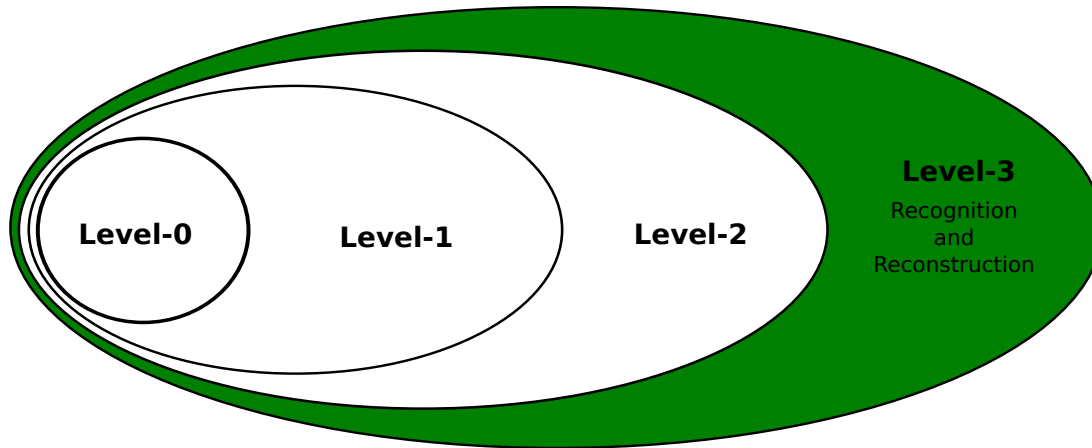
$$\begin{aligned} \psi_b(\mathbf{v}_b, H_b, l_b) &= \min \left(0, f(\mathbf{v}_b, H_b) + k_b \left\{ \delta(v_i \neq l_b) \right\}_{i \in \mathbf{v}_d} \right) \\ &= f(\mathbf{v}_b, H_b) + \min \left(f(\mathbf{v}_b, H_b), k_b \left\{ \delta(v_i \neq l_b) \right\}_{i \in \mathbf{v}_b} \right). \end{aligned} \quad (4.11)$$

This potential takes the form of a Robust P^N potential [32], which is defined as:

$$\psi_h(\mathbf{v}) = \min \left[\gamma_{max}, \min_l \right) \gamma_l + k_l \left\{ \delta(v_i \neq l) \right\}_{i \in \mathbf{v}} \left[\begin{array}{l} \text{ } \end{array} \right], \quad (4.12)$$

where $\gamma_{max} = f(\mathbf{x})$, $\gamma_l = f(\mathbf{x}, \mathcal{A} \neq b)$, and $\gamma_b = 0$. Thus it can be solved efficiently using $\alpha\beta$ -swap and α -expansion algorithms as shown in [32]. The detection instance variables y_b can be recovered from the final labelling by computing y_b as:

$$y_b = \arg \min_{y'_b \in \{0,1\}} f(\mathbf{v}_b, H_b)y'_b + g(N_b, H_b)y'_b. \quad (4.13)$$



4.3 Energy Layer: Recognition and Reconstruction

The aim of this application is to combine two definitive computer vision problems—recognition and reconstruction—in order to improve the accuracy of both. Our solution to this problem involves taking the CRF approach to recognition with the CRF approach to reconstruction, and then co-joining them in a holistic fashion: allowing them to communicate, and update one-another. In essence we fuse together semantic image segmentation [37] and disparity estimation [5, 34] within the 3rd layer of our architecture (§3.3).

4.3.1 Introduction

The problems of object class segmentation [37, 56], 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 labelled with a disparity [34], 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) [38], which is typically a generalised Potts truncated linear model. Thus both may use graph cut based move making algorithms, such as α -expansion [5], to solve the labelling problem. These problems should be solved jointly, as a correct labelling of object class can inform depth labelling and stereo reconstruction can also improve object labelling. To provide some intuition behind this

statement, note that the object class boundaries are more likely to occur at a sudden transition in depth and vice-versa. Moreover, the height of a point above the ground plane is an extremely informative cue regarding its class label, and can be computed from the depth. For example, *road* or *sidewalk* lie in the ground plane, and pixels taking labels *pedestrian* or *car* must lie above the ground plane, while pixels taking label *sky* must occur at an infinite depth from the camera. Figure 4-5 shows our model which explicitly captures these properties. Object class

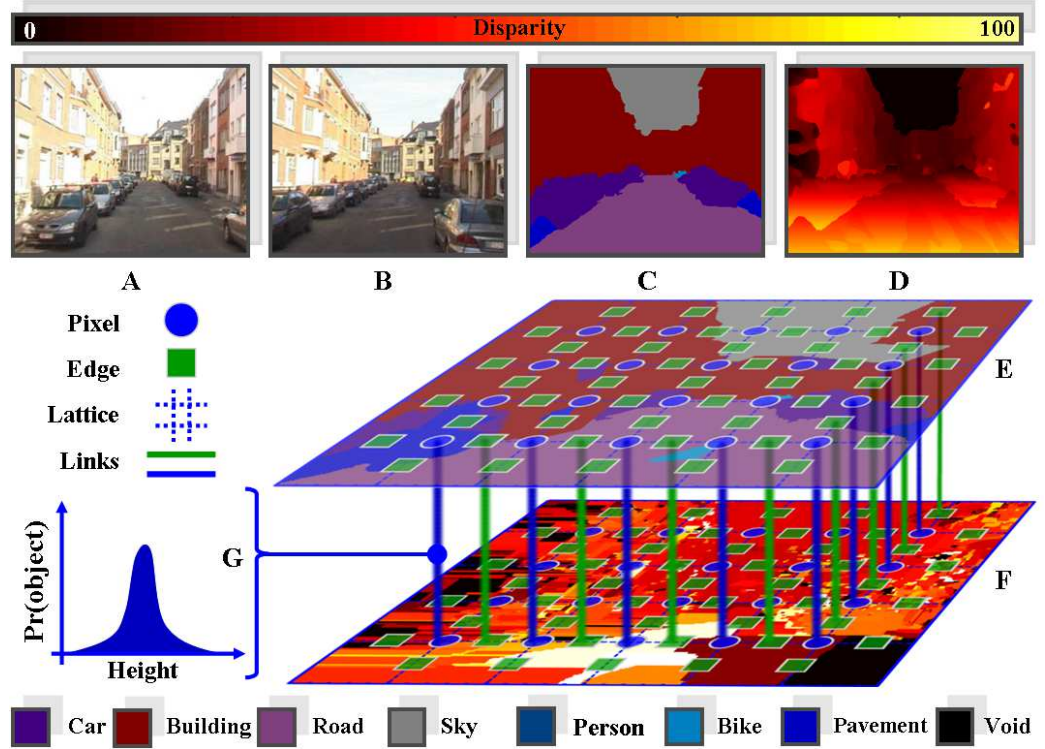


Figure 4-5: **Graphical model.** The system takes a left (A) and right (B) image from a stereo pair that has been rectified. Our formulation captures the co-dependencies between the object class segmentation problem (E, §4.3.3) and the dense stereo reconstruction problem (F, §4.3.4) by allowing interactions between them. These interactions are defined to act between the unary/pixel (blue) and pairwise/edge variables (green) of both problems. The unary potentials are linked via a height distribution (G, eq. (4.22)) learnt from our training set containing hand labelled disparities. The pairwise potentials encode that object class boundaries, and sudden changes in disparity are likely to occur together. The combined optimisation results in an approximate object class segmentation (C) and dense stereo reconstruction (D). View in colour.

recognition yields strong information about 3D structure as shown by the work on photo pop-

up [20, 26, 45, 48]. Here a plausible pop-up or planar model of a scene was reconstructed from a single monocular image using only prior information regarding the geometry of typically photographed scenes, and knowledge of where object boundaries are likely to occur.

Beyond this, many tasks require both object class and depth labelling. For an agent to interact with the world, it must be capable of recognising both objects and their physical location. For example, camera based driverless cars must be capable of differentiating between *road* and other classes, and also of recognising where the road ends. Similarly, several companies wish to provide an automatic annotation of assets (such as *street light*, *drain* or *road sign*) to local authorities. In order to provide this service, assets must be identified, localised in 3D space and an estimation of the quality of the assets made. The use of object labellings to inform scene reconstruction is not new. The aforementioned pop-up method of [20] explicitly used object labels to aid the construction of a scene model, while 3D Layout CRF [30] matched 3D models to object instances. However, in [20] they build a plausible model from the results of object class segmentation, neither jointly solving the two problems, nor attempting to build an accurate 3D reconstruction of the scene, whereas in this application we jointly estimate both. Hoiem *et al.* [30] fit a 3D model not to the entire scene but only to specific objects, and similarly, these 3D models are intended to be plausible rather than accurate.

Leibe *et al.* [42] 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 [8] data set provides sparse *SfM* cues, which were used by several object class segmentation approaches [8, 57] to provide pixel wise labelling. In these works, no dense depth labelling was performed and the object class segmentation was not used to refine the 3D structure.

None of the discussed works perform joint inference to obtain dense stereo reconstruction and object class segmentation. In this application, we demonstrate that the problems are mutually informative, and benefit from being modelled jointly within the 3rd level of our holistic

scene understanding hierarchy (§ 3.3).

4.3.2 Global Configuration

We combine recognition and reconstruction within layer-3 of our holistic scene understanding hierarchy (§ 3.3). Reconstruction is modelled as a disparity (inverse depth) labelling problem. It takes a rectified pair of input images, $\mathcal{L}^{right}(I, \mathbf{D}^{right})$ and $\mathcal{L}^{left}(I, \mathbf{D}^{left})$, and outputs a labelling with $\mathbf{S}(I, \mathbf{Y})$, where each variable y / Y takes on a label from the label set \mathcal{C}^D . Recognition is modelled as an object-class labelling problem that takes the left image as input and outputs a segmentation $\mathbf{S}(I, \mathbf{X})$, where each variable x / X takes on a label from the label set \mathcal{C}^O . The joint labelling problem is a layer-3 fusion using factorial CRF (3.4), giving our joint energy for recognition and reconstruction as:

$$E(f; \{\mathbf{D}^{left}, \mathbf{D}^{right}\}) \quad \boxed{\text{Level-3}}$$

$$= E^O(X; \mathbf{D}^{left}) \quad \boxed{\text{Recognition}} \quad (4.14)$$

$$+ E^D(Y; \{\mathbf{D}^{left}, \mathbf{D}^{right}\}) \quad \boxed{\text{Reconstruction}} \quad (4.15)$$

$$+ \epsilon[X, Y; \mathbf{D}^{left}], \quad \boxed{\text{Joint factors}} \quad (4.16)$$

Our contribution here is on configuring the joint factors (4.16). For completeness we shall also summarise the configurations of the Recognition and Reconstruction parts.

4.3.3 Recognition

For the recognition part of our joint energy (4.14) we follow [32, 37, 56] in formulating the problem of object class segmentation as finding a minimal cost labelling of a CRF defined over a set of random variables $X = \{x_1, \dots, x_N\}$ each taking a state from the label space $\mathcal{C}^O = \{o_1, o_2, \dots, o_k\}$. Each label o_j indicates a different object class such as *car*, *road*, *building* or *sky*. These energies take the form:

$$\begin{aligned}
E^O(f^O; \mathbf{D}^{left}) = & \Psi^{UO} \left(X; \mathbf{D}^{left} \right) \Big|_{\{app\}} \left\{ \begin{array}{l} \text{Level-1} \end{array} \right. & (4.17) \\
& + \Psi^{PO} \left(X; \mathbf{D}_0^{left} \right) \left(\begin{array}{l} \text{Level-0} \end{array} \right) \\
& + \Psi^{HO} \left(X; \mathbf{D}^{left} \right) \Big|_{\{app\}} \left\{ \begin{array}{l} \text{Level-1} \end{array} \right.
\end{aligned}$$

where Ψ^{UO} , Ψ^{PO} and Ψ^{HO} are unary, pairwise and higher order potentials.

Unary Potentials The unary potentials, $\Psi^{UO} = \sum \{w_O^u \psi_u^O \langle_{u \in UO}$, of the CRF describes the cost of a single pixel taking a particular label. The terms are typically computed from colour, texture and location features of the individual pixels and corresponding prelearned models for each object class [56].

Pairwise Potentials The pairwise terms, $\Psi^{PO} = \sum \{w_O^p \psi_p^O \langle_{p \in PO}$, encourage similar neighbouring pixels in the image to take the same label and takes the form of a contrast sensitive Potts model [3, 50, 56]. These potentials are shown in fig. 4-5 *E* as blue circles and green squares respectively.

Higher Order Potentials The higher order terms, $\Psi^{HO} = \sum \{w_O^\eta \psi_\eta^O \langle_{\eta \in HO}$, describe potentials defined over cliques containing more than two pixels. In our work we follow [37] and use their hierarchical potentials based upon region based features, which significantly improve the results of object class segmentation.

4.3.4 Reconstruction

We use the energy formulation of [5, 34] 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 $\mathbf{Y} = \{Y_1, \dots, Y_N\}$, where each variable Y_i takes a state from the label space $\mathcal{E} = \{d_1, d_2, \dots, d_m\}$ corresponding to a set of disparities, and can be

written as:

$$E^D(Y; \{\mathbf{D}^{left}, \mathbf{D}^{right}\}) = \Psi^{UD}(Y; \mathbf{D}^{left}, \mathbf{D}^{right}) + \Psi^{PD}(Y; \mathbf{D}^{left}) \quad (4.18)$$

The unary (*blue circles*) and pairwise (*green squares*) potentials are shown in fig. 4-5 *F*. Note that the disparity for a pixel is directly related to the depth of the corresponding 3D point.

Unary Factors The unary potentials, $\Psi^{UD} = \sum_u w_D^u \psi_u^D$, of the disparity CRF are defined as a measure of colour agreement of a pixel \mathbf{D}_i^{left} with its corresponding pixel \mathbf{D}_i^{right} from the stereo-pair given a choice of disparity.

Pairwise Factors The pairwise terms, $\Psi^{PD} = \sum_{\rho \in PD} w_D^\rho \psi_\rho^D$, encourage neighbouring pixels in the image to have a similar disparity. The cost is a function of the distance between disparity labels:

$$\psi_\rho(y_i, y_j) = g(|y_i - y_j|), \quad (4.19)$$

where $g(\cdot)$ usually takes the form of linear truncated function $g(y) = \min(k_1 y, k_2)$, where $k_1, k_2 \in \mathbb{R}$ are the slope and truncation respectively.

4.3.5 Recognition and Reconstruction

The co-joining part of our factorised joint labelling problem for object-class and disparity (4.16) is defined as:

$$\epsilon^J([X, Y]; \mathbf{D}) = \Phi^{UJ}([X, Y]; \mathbf{D}) + \Phi^{PJ}([X, Y]; \mathbf{D}) \quad (4.20)$$

which has the same form as (3.5) in our layer-3, consisting of joint unary Φ^U and joint pairwise Φ^J potentials. We now discuss the configuration of these potentials for this application.

Joint Unary Potentials

The joint unary potentials, $\Phi^{UJ} = \sum \{w_j^u \phi_u^J\}_{u \in UJ}$, models the interaction between the unary potentials of both the object class segmentation and dense stereo reconstruction parts of our formulation. In order for them to interact successfully, we need to define some function that relates them 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 xz-coordinate may be informative, since we expect that *buildings* will be on both sides, *pavement* will tend to be between *building* and *road* that would take up the central portion of the image. 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 modelling like this requires a substantial amount of data.

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

$$h(y_i, i) = h_c + \frac{(y_h - y_i) \times b}{d} \quad (4.21)$$

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 modelled by estimating the a priori cost of pixel i taking label $z_i = [x_i, y_i]$ by

$$\phi_u^J([x_i, y_i]) = \log(H(h(y_i, i) \mid x_i)) \quad (4.22)$$

where

$$H(h \mid l) = \frac{\sum_{i \in \mathcal{T}} \delta(x_i = l) \delta(h(y_i, i) = h)}{\sum_{i \in \mathcal{T}} \delta(x_i = l)} \quad (4.23)$$

is a histogram based measure of the naive probability that a pixel taking label l has height h in the training set \mathcal{V} . Fig. 4-5 *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.

Joint Pairwise Potentials

The joint pairwise potentials, $\Phi^{PJ} = \sum w_J^p \phi_\rho^J$, model 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 disparity of two neighbouring pixels significantly differ. To take this information into account, we chose tractable pairwise potentials of the form:

$$\phi_\rho^J([x_i, y_i], [x_j, y_j]) = \psi_\rho^O(x_i, x_j) \psi_\rho^D(y_i, y_j). \quad (4.24)$$

Fig. 4-5 shows this linkage as *green line* between a pairwise potential (*green box*) of each part. Since there are few (4) parameters we use cross-validation to train them.

4.3.6 Inference

Inference alternates between α -expansion [5] in the object class label space, and range moves [63] in the in the disparity label space.

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}_α of size 2^N . We define \mathbf{Z}_α as:

$$\mathbf{Z}_\alpha = \{ \mathbf{z}' / (\mathcal{M} \bullet \mathcal{E})^N : z'_i = [x'_i, y_i] \text{ and } (x'_i = x_i \text{ or } x'_i = \alpha) \}. \quad (4.25)$$

One iteration of the algorithm involves making moves for all α in \mathcal{M} in some order successively. Bringing together all the pairwise terms from the object, disparity and joint parts, and under the assumption that \mathbf{y} is fixed, we have:

$$\begin{aligned}\phi_\rho^{ODJ}([x_i, y_i], [x_j, y_j]) &= (w_O^\rho + w_J^\rho \psi_\rho^D(y_i, y_j)) \psi_\rho^O(x_i, x_j) + w_D^\rho \psi_\rho^D(y_i, y_j) \\ &= \lambda_\rho \psi_\rho^O(x_i, x_j) + k_{ij}.\end{aligned}\quad (4.26)$$

The constant k_ρ does not affect the choice of optimal move and can safely be ignored. If $\mathcal{A}_{y_i, y_j} \lambda_\rho = w_O^\rho + w_J^\rho \psi_\rho^D(y_i, y_j) \leq 0$, the projection of the pairwise potential is a Potts model and standard α -expansion moves can be applied. For $w_O^\rho \leq 0$ this property holds if $w_O^\rho + w_J^\rho k_2 \leq 0$, where k_2 is defined as in §4.3.4. In practice we use a variant of α -expansion suitable for higher order energies [52].

Range moves in the disparity label space For our joint optimisation of disparity and object classes we propose a new move in the projected 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 $(x_i, [l, l + r])$, where r is the defined offset. To formalise 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:

$$\mathbf{Z}_l = \{ \mathbf{z}' \in (\mathcal{M} \times \mathcal{E})^N : z'_i = [x_i, y'_i] \text{ and } (y'_i = y_i \text{ or } y'_i \in [l, l + r]) \}. \quad (4.27)$$

Bringing together all the pairwise terms from the object, disparity and joint parts, and under the assumption that \mathbf{x} is fixed, we have:

$$\begin{aligned}\phi_\rho^{ODJ}([x_i, y_i], [x_j, y_j]) &= (w_D^\rho + w_J^\rho \psi_\rho^O(x_i, x_j)) \psi_\rho^D(y_i, y_j) + w_O^\rho \psi_\rho^O(x_i, x_j) \\ &= \lambda_\rho \psi_\rho^D(y_i, y_j) + k_\rho.\end{aligned}\quad (4.28)$$

Again, the constant k_ρ can safely be ignored, and if $\mathcal{A}_{x_i, x_j} \lambda_\rho = w_D^\rho + w_J^\rho \psi_\rho^O(x_i, x_j) \leq 0$ the projection of the pairwise potential is linear truncated and standard range expansion moves can

be applied. This property holds if $w_D^\rho + w_J^\rho(\theta_p + \theta_v) \leq 0$, where θ_p and θ_v are the weights of the Potts pairwise potential.

Chapter 5

Evaluation

In this chapter an evaluation of our applications is presented. First existing road scene datasets that are used to evaluate our contributions are summarised, along with our augmentations of their data. A description of our chosen evaluation protocols and metrics is then discussed. This is followed by the quantitative and qualitative evaluation of our applications for each layer of the proposed holistic hierarchy.

5.1 CamVid Dataset [7,8]

The Cambridge-driving Labelled Video Database (CamVid) database¹ consists of over 10 minutes of 960 • 720 resolution images captured at 30 Hz from within a driven car; the camera setup and an example of a captured frame is shown in Fig. 5-1 (top). The database addresses the need for experimental data to quantitatively evaluate emerging algorithms. Whilst most existing benchmarks are *static* such as consumer photographs, or fixed-position CCTV-style videos, the CamVid data is captured from the perspective of a driving car, giving a more *dynamic* database, and the driving scenario also increases the number and heterogeneity of the observed object classes. The database has a variety of residential, urban, and mixed road sequences. Three of the four sequences (0006R0, 0016E5, Seq05VD) are shot in daylight, and the fourth sequence

¹Available at <http://mi.eng.cam.ac.uk/research/projects/VideoRec>.

(0001TP) is captured at dusk. The camera's intrinsic and extrinsic parameters, 2D feature tracks over all frames, as well as the 3D point clouds are provided for all the sequences. This is calculated automatically and thus is not ground truth data for evaluation. The dataset ground truth is provided for the semantic image segmentation task into object classes. We augment the dataset with object bounding boxes.



Figure 5-1: **CamVid Database:** The CamVid database consists of a set a video sequences recorded from a driven car. In the top row of the figure we see how the camera is positioned within the car, along with an example frame that is captured whilst the car is driven around the Cambridge area. The database is constructed for the semantic image segmentation task with ground truth labellings; the labels and an example of a hand labelled image is shown in the bottom row. Figures reproduced from [7].

Semantic Segmentation Ground Truth A selection of 700 frames from the video sequences are manually labelled. Each pixel in these frames was labelled as one of 32 candidate classes. A small number of pixels are labelled as *void*, which do not belong to one of these classes and are ignored. The labelled images are stridden at 1 Hz, and also 15 Hz for a small subsection of one of the video sequences. The class labels with their corresponding colour codes, and an example

ground truth labelling are shown in Figure 5-1(bottom). Interested readers may refer to [7,8] for details on the database.

In practice we use the 1 Hz labelled sequences and a subset of 11 categories: *Building*, *Tree*, *Sky*, *Car*, *Sign-Symbol*, *Road*, *Pedestrian*, *Fence*, *Column-Pole*, *Sidewalk*, and *Bicyclist*, from the full set, for comparison with the work of [8], see Fig. 5-2 for summary statistics from [8].

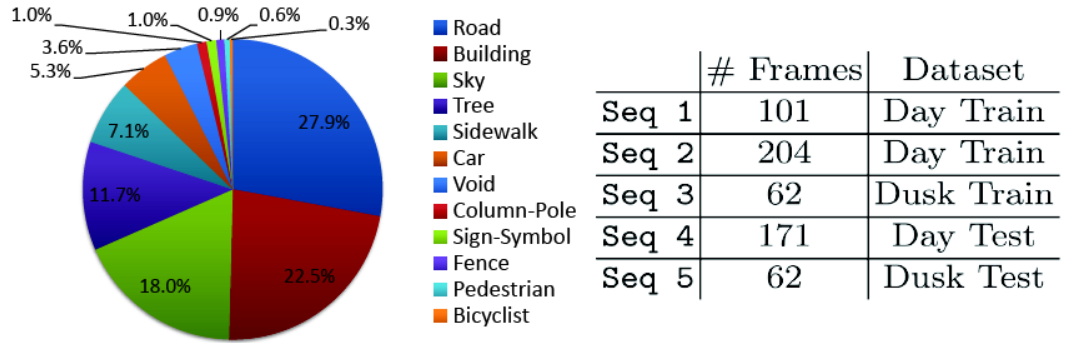


Figure 5-2: **CamVid Splits:** The CamVid database is split into several sequences of varying length, the number of ground truth labelled image per-sequence is shown on the right. In practice 11 object classes are used, these along with their percentage of their labelled pixels is shown on the left. Figure reproduced from [8].

Object Detection Ground Truth For a subset of the 11 object classes, and 1 Hz labelled framed, from the CamVid database, which are used in practice, we add bounding box annotations to enrich the dataset. The objects that are chosen are ones that whose spatial extent and localisation can be reasonably approximated by a bounding box, specifically the classes are: *Car*, *Sign-Symbol*, *Pedestrian*, *Column-Pole*, and *Bicyclist*. The bounding box labelling is performed with a bespoke tool. To label an object 4 user clicks are required, one for the top-most, bottom-most, left-most, and right-most boundaries of the object. The dominant label within the bounding box is automatically assigned, obtained from the original segmentation ground truth. Only objects in the CamVid *train* sets (Fig. 5-2) are assigned bounding boxes, and not all instances are included.

5.2 Leuven Dataset [10,42]

The Leuven Stereo Scene dataset² is a sequence of 1175 image pairs recorded, from a driven platform shown in Fig.5(top left), at 25fps and a resolution of 360 • 288 pixels over a distance of about 500m. The main difficulties for object recognition lie in the relatively low resolution, strong partial occlusion between parked cars, frequently encountered motion blur, and extreme contrast changes between brightly lit areas and dark shadows [42]. This data differs from commonly used stereo matching sets like the Middlebury [53] 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 [8] in two important ways, CamVid is a monocular sequence, and the 3D information comes in the form of an unstable³ set of sparse 3D points. 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. However, the dataset does not contain the object class or disparity annotations that are required for learning object models and for quantitative evaluation.

Disparity Ground Truth Since the Leuven stereo dataset has no disparity ground truth, we manually label a subset of data. To augment the dataset with disparity ground truth, all image pairs are first rectified automatically [17]. Then, a subset of 70 non-consecutive frames, which upon visual inspection appear to be successfully rectified, are selected for human annotation. These are then cropped to 316 • 256 such that most pixels have correspondences and the warping around the image borders is removed. The procedure for labelling these 70 rectified image pairs is carried out with a bespoke tool, targeting the left-side image. The user of the tool identifies a minimum of 3 pairs of corresponding points, between the left and right images, that belong to

²<http://www.vision.rwth-aachen.de/data/leuven-left.tgz> <http://www.vision.rwth-aachen.de/data/leuven-right.tgz>

³The 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 [8] for more details.

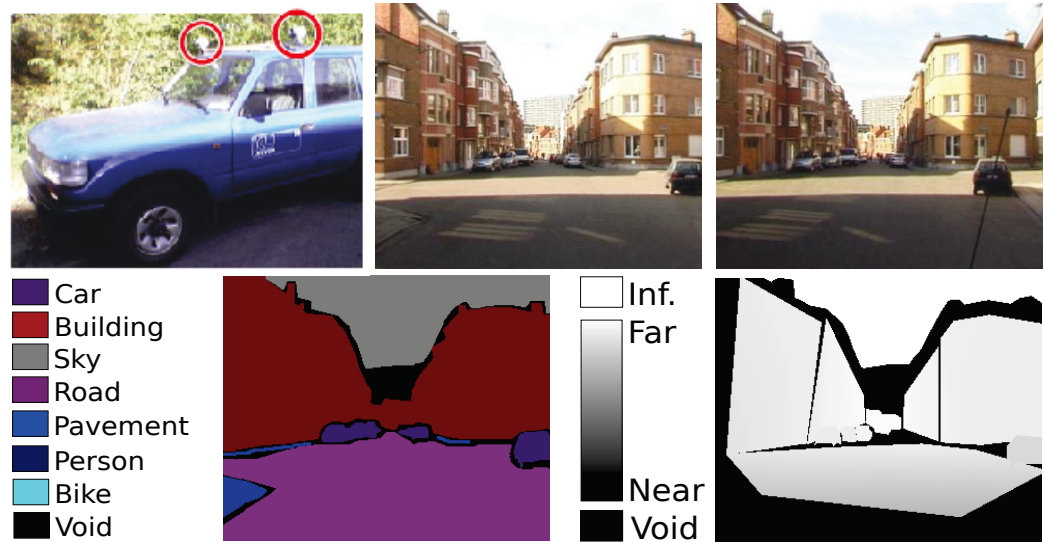


Figure 5-3: **Leuven Dataset**: The Leuven stereo dataset is captured from a driven vehicle around the Leuven area. The vehicle and stereo camera rig is shown in the top row, along with an example of a stereo pair of captured images. These images have been rectified and then labelled by hand with object-classes and disparities, creating the ground truth shown on the bottom row.

a plane (or can be approximated by a plane, such as the road, or a building facade). From each pair of points the disparity can be calculated, and since all the pairs form a plane, the disparities can be interpolated for the other pixels. The process is repeated until no more planes can be easily identified. This results in a coarse planer approximation of the disparity ground truth as depicted in Fig. 5-3 (bottom right).

Semantic Segmentation Ground Truth Since the Leuven stereo dataset has no object class segmentation ground truth, we manually label, the left-side image of, the 70 image pairs that were selected for disparity labelling. The annotation procedure consists of manually labelling every pixel of each image with 7 object classes: *Building*, *Sky*, *Car*, *Road*, *Person*, *Bike* and *Sidewalk*⁴. In order to label the images we used the layer, pencil and fill tools of the GIMP image editor.⁵ An 8th label, *void*, is given to pixels that do not obviously belong to one of the classes (this includes areas near the object boundaries that are both ambiguous, and time consuming to

⁴In practice the Person class is set to void due to an insufficient number of instances.

⁵This is a simple technique to label images, especially for those already familiar with the oddity of the GIMP interface. GIMP is freely available from <http://www.gimp.org/>

label accurately). This procedure results in a object-class segmentation ground truth as depicted in Fig. 5-3 (bottom left).

5.3 Evaluation Protocol and Metrics

By the nature of the semantic image segmentation problem, in that each pixel of an image is assigned a label, care must be taken to have reasonably independent training and test sets. For instance having a pixel location (x, y) with class z in the training set, and then a pixel $(x + 1, y)$ with the class in the test set would hardly be a challenge for modern techniques. Similarly, when dealing with videos that have frame rates up to $30fps$, having a training pixel at (x, y, t) and a test pixel at $(x, y, t + 1)$ is not a reasonable test. For this reason it is not recommended to follow the classic k -fold cross validation technique, with random splits, when dealing with highly structured data. This leads to the fixed training and test splits being the preferred technique in semantic image segmentation. Also, when comparing to other works, a direct comparison can only be accomplished by following their protocols. If there is conflict in protocols, then both should be reported.

The quality of the predicted labelling is measured w.r.t:

True Positive/Negative Pixel variable correctly/incorrectly assigned to label ℓ / \mathcal{M}

False Positive/Negative Pixel variable correctly/incorrectly not assigned to label ℓ / \mathcal{M}

For a particular label ℓ / \mathcal{M} let TP_ℓ and TN_ℓ , denote the number of true positives/negatives; FP_ℓ and FN_ℓ denote the number of false positives/negatives. Then, several metrics (graphically depicted in Fig. 5-4): recall (per-class, global, and average), and intersection over union (per-class and average) are defined as:

Recall: Measures the proportion of correctly Vs incorrectly assigned pixel variables for a particular class:

$$Recall_\ell = \frac{TP_\ell}{TP_\ell + FN_\ell}, \quad \boxed{\text{Recall}} \quad (5.1)$$

$$GlobalRecall = \frac{\sum_{\ell \in \mathcal{L}} TP_\ell}{\sum_{\ell \in \mathcal{L}} TP_\ell + \sum_{\ell \in \mathcal{L}} FN_\ell} = \frac{\sum_{\ell \in \mathcal{L}} TP_\ell}{\#Pixels} \quad \boxed{\text{Global}} \quad (5.2)$$

$$Avg.Recall = \frac{\sum_{\ell \in \mathcal{L}} Recall_\ell}{\mathcal{M}}, \quad \boxed{\text{Avg. Recall}} \quad (5.3)$$

and penalises under-estimates of the segmentation of a particular class. This is a widely reported metric for semantic image segmentation. Note that 100% recall may be achieved for a single class, so long as there are no false negative predictions.

Intersection over Union: Measures the proportion of correctly Vs incorrectly assigned *and* incorrectly un-assigned pixel variables for a particular class:

$$IoU_\ell = \frac{TP_\ell}{TP_\ell + FP_\ell + FN_\ell}, \quad \boxed{\text{Intersection over Union}} \quad (5.4)$$

$$Avg.IoU = \frac{\sum_{\ell \in \mathcal{L}} IoU_\ell}{\mathcal{M}}, \quad \boxed{\text{Avg. Intersection over Union}} \quad (5.5)$$

and penalises both over-estimates and under-estimates of the segmentation of a particular class. This is the metric used in the VOC segmentation challenge [12]. Under this measure 100% can only be achieved with no false negatives and no false positives. Note that this score will always be lower than recall, unless there are no false positives.

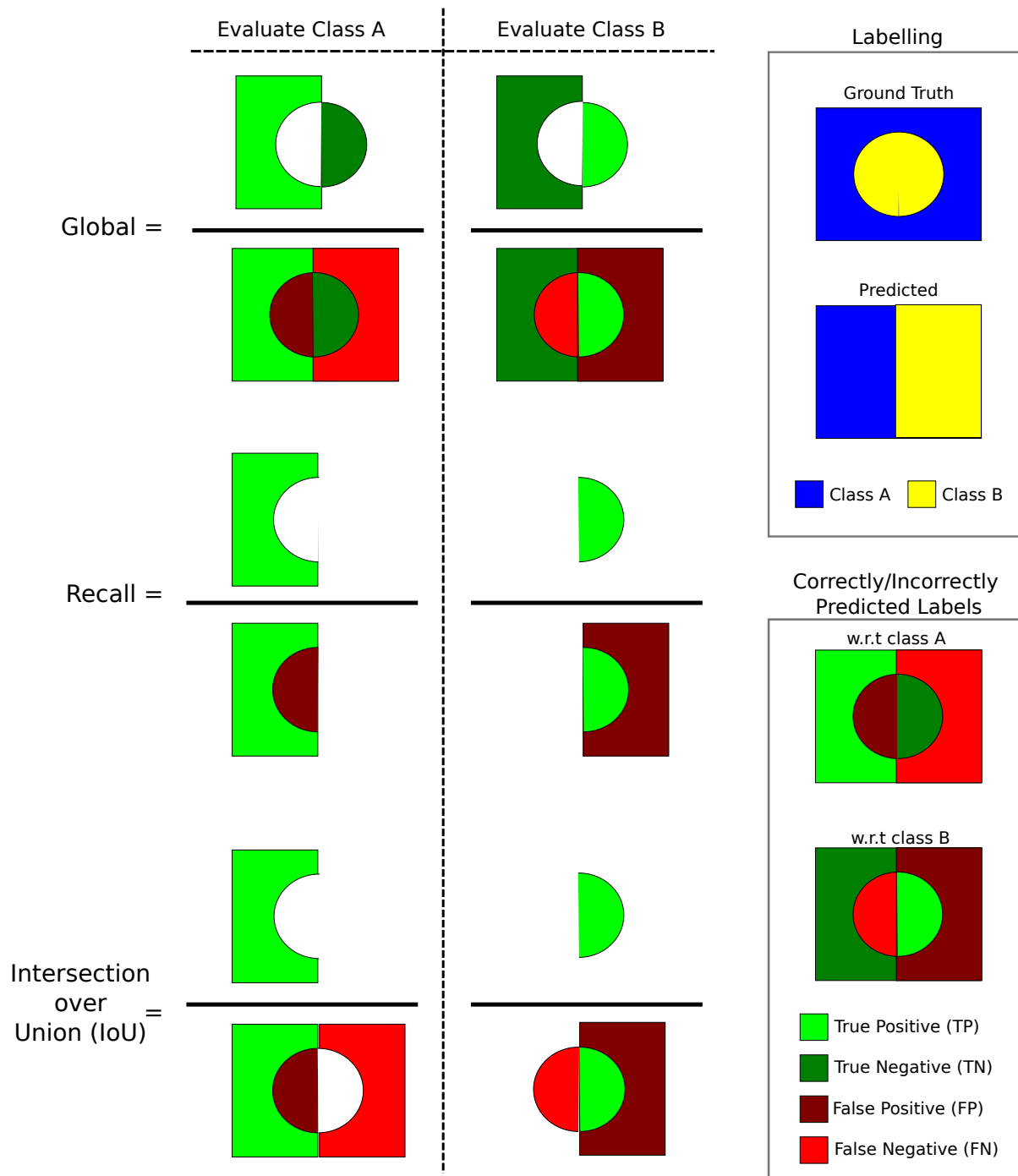


Figure 5-4: **Evaluation Metrics:** Given a ground truth, and predicted labelling, as seen in the *labelling* box, evaluation of its quality is measured using several metrics. These rely on identifying true positives/negatives, and false positives/negatives, seen in the *correctly/incorrectly predicted labels* box. Essentially these metrics are various ratios of correct Vs incorrect predictions. The main table visualises the global, recall, and intersection over union measures, w.r.t the example labelling, and each of the classes.

5.4 Feature Layer: Appearance and Geometry

In this section the qualitative and quantitative results are reported for our feature layer (§3.1) application that combines appearance and geometry cues (§4.1). We evaluated our method on the CamVid Database [7](§5.1). A comparison of our results with the state-of-the-art method of [8] is presented, where a 14.7% improvement in overall recognition accuracy is achieved.

Figure 5-5 (d, e, f) shows the qualitative results of our method on sample day and dusk images (h). The higher order results (f) have well-defined object boundaries, and are more similar to the ground truth (g) compared to the results of [8] (a, b, c). The quantitative results are summarized in Table 5.1. We achieve a global accuracy (*i.e.*, the percentage of pixels correctly classified) of 84% in comparison to 69% in [8]. A high performance ($> 50\%$ IoU) is achieved on most of the object categories. The two categories (Pedestrian, Fence) where our performance is low ($< 20\%$ IoU) is perhaps due to the lack of training data. The training dataset has less than 2% of pixels labelled as one of these categories, which appears to be insufficient to learn the potentials. In some cases the higher order CRF under-performs compared to the pairwise CRF due to objects which are only a few pixels wide in the image *e.g.*, Column-Pole. This is due to the failure of the mean-shift [9] segmenter to pick out fine structures. Figure 5-6 highlights the qualitative improvements achieved by higher order CRFs.

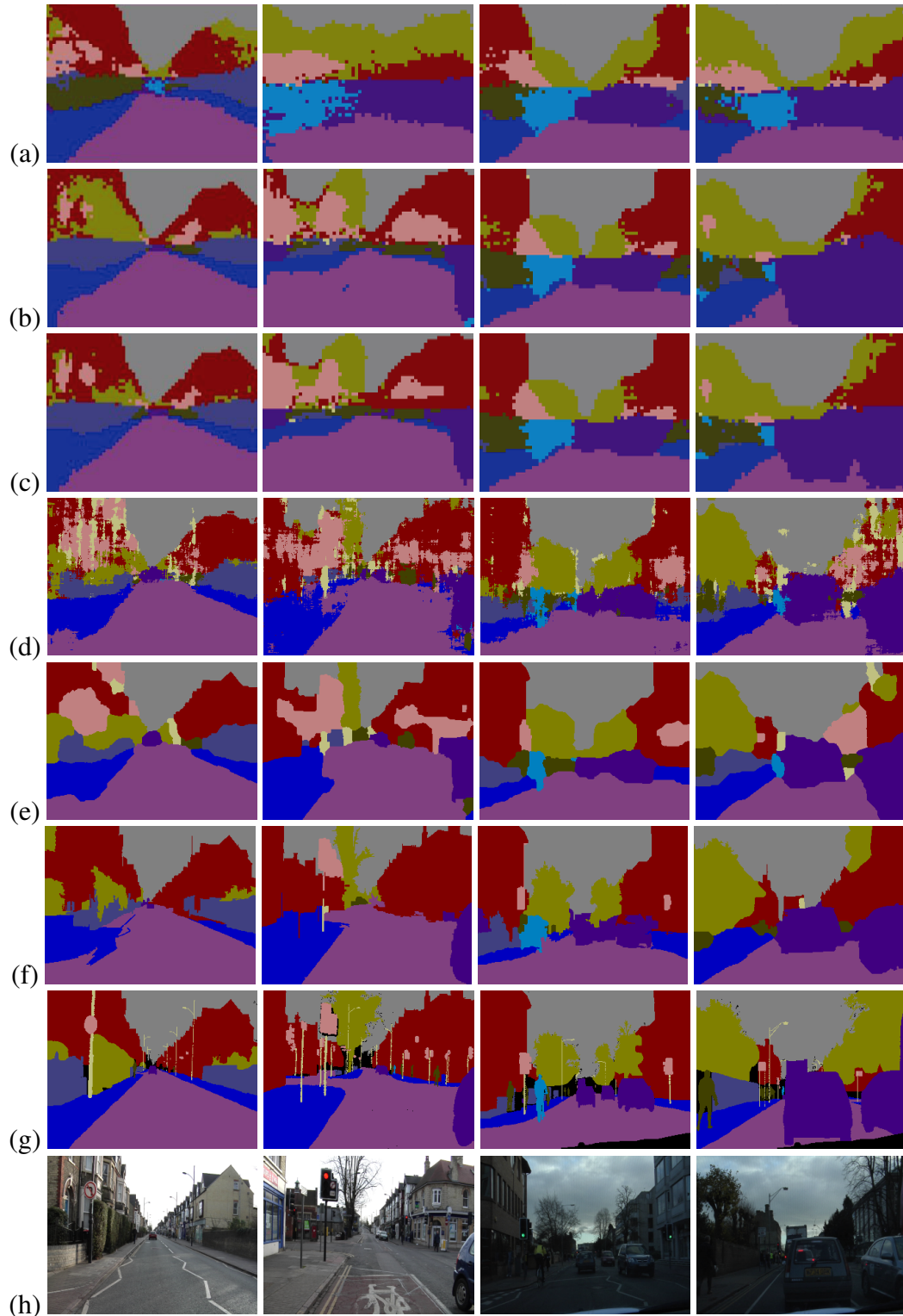


Figure 5-5: **Qualitative Results:** Sample object category segmentations of two day and two dusk images. Results from [8] are shown in: (a) Motion and structure-based segmentation, (b) Appearance-based segmentation, (c) Combined segmentation result. Our results: (d) using only unary potentials, (e) adding pairwise potentials improves the segmentation, but fails at object boundaries. The row (f) shows our combined higher order potential based segmentation, which is qualitatively better than (a) - (e). (g) Ground truth labelled image, (h) Original test image. Note that using higher order provides better segmentation, as well as clearer object boundaries.

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	Average	Global
Mot. [8]	43.9	46.2	79.5	44.6	19.5	82.5	24.4	58.8	0.1	61.8	18.0	43.6	61.8
App. [8]	38.7	60.7	90.1	71.1	51.4	88.6	54.6	40.1	1.1	55.5	23.6	52.3	66.5
Combined [8]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1
Recall													
ψ_i	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4
$\psi_i + \psi_{ij}$	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8
$\psi_i + \psi_{ij} + \psi_c$	84.5	72.6	97.5	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8
IoU													
ψ_i	55.3	54.3	84.8	51.8	11.9	85.5	15.6	27.4	7.5	60.0	15.7	42.71	NA
$\psi_i + \psi_{ij}$	63.6	58.0	87.8	55.9	13.6	86.4	16.9	27.6	6.1	61.9	18.1	45.07	NA
$\psi_i + \psi_{ij} + \psi_c$	71.6	60.4	89.5	58.3	19.4	86.6	26.1	35.0	7.2	63.8	22.6	49.15	NA

Table 5.1: **Quantitative Results:** Pixel-wise percentage accuracy on all the test sequences. Results of [8] using only motion-based (Mot.), only appearance-based (App.) and both features (Combined) are shown for comparison. We present results of our CRF-based method using the Recall and the PASCAL VOC measures. Only unary terms (ψ_i), unary and pairwise terms ($\psi_i + \psi_{ij}$), and unary, pairwise and higher order terms ($\psi_i + \psi_{ij} + \psi_c$). Note that our method, which uses all the terms, gives the best performance for almost all the classes. ‘Global’ is the percentage of pixels correctly classified, and ‘Average’ is the average of the per-class accuracies.

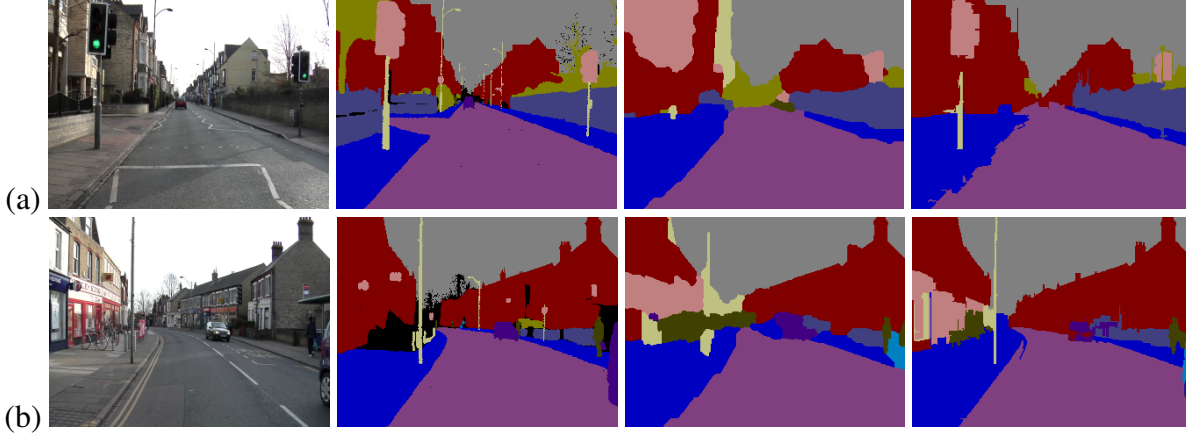


Figure 5-6: **Qualitative Results:** Qualitative improvements achieved by our higher order CRF framework. We show (left to right) the original image, the ground truth image, pairwise CRF result, and higher order CRF result for two frames from the test sequences. The higher order potentials correct the object boundary errors in the pairwise CRF results *e.g.* , traffic light, and the building in (a). They also provide accurate segmentation, which is more similar to ground truth compared to the pairwise result *e.g.* , lamp post, sidewalk in (b).

5.5 Potential Layer: Things and Stuff

In this section the qualitative and quantitative results are reported for our potential layer (§3.2) application that combines *things and stuff* (§4.2). We evaluated our method on the CamVid Database [7](§5.1). A comparison of our results with the state-of-the-art method of [8] is presented, where a 15.4% improvement in overall recognition accuracy is achieved.

Figures 4-4, 5-7 and 5-8 show qualitative results on the CamVid dataset, where we can observe that object detection artefacts, rectangular segments, do not present themselves and the precise object boundaries of the baseline (§5.4) are persevered.

Object segmentation approaches do not identify the number of instances of objects, but this information is recovered using our combined segmentation and detection model (from y_d variables, as discussed in §4.2.6), and is shown in Figure 4-4. Figure 5-7 shows the advantage of our soft constraint approach to include detection results. The false positive detection here (shown as the large green box) does not affect the final segmentation, as the other hypotheses based on pixels and segments are stronger. However, a strong detector hypothesis (shown as the purple

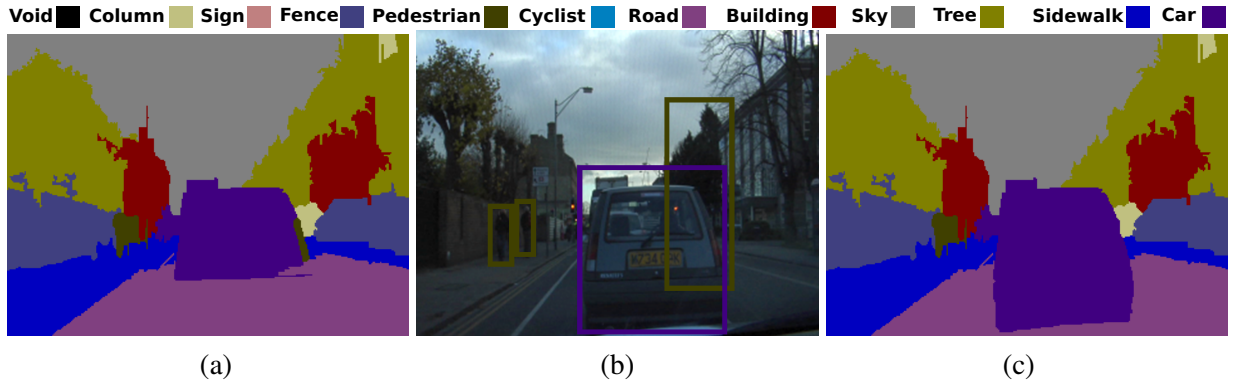


Figure 5-7: **Qualitative Results:** (a) Segmentation without object detectors, (b) Object detections for car and pedestrian shown as bounding boxes, (c) Segmentation using our method. These detector potentials act as a soft constraint. Some false positive detections (such as the large green box representing a person) do not affect the final segmentation result in (c), as it does not agree with other strong hypotheses based on pixels and segments. On the other hand, a strong detector response (such as the purple bounding box around the car) correctly relabels the road and pedestrian region as car in (c) resulting in a more accurate object class segmentation. (Best viewed in colour)

box) refines the segmentation accurately. Figure 5-8 highlights the complementary information provided by the object detectors and segment-based potentials. An object falsely missed by the detector (traffic light on the right) is recognized based on the segment potentials, while another object (traffic light on the left) overlooked by the segment potentials is captured by the detector. More details are provided in the figure captions. Quantitative results on the CamVid dataset are shown in Table 5.2. For the recall measure, our method performs the best on 5 of the classes, and shows near-best ($< 1\%$ difference in accuracy) results on 3 other classes. Accuracy of “things” classes improved by 7% on average. This measure does not consider false positives, and creates a bias towards smaller classes. Therefore, we also provide results with the intersection *vs* union measure in Table 5.2. We observe that our method shows improved results on almost all the classes in this case.

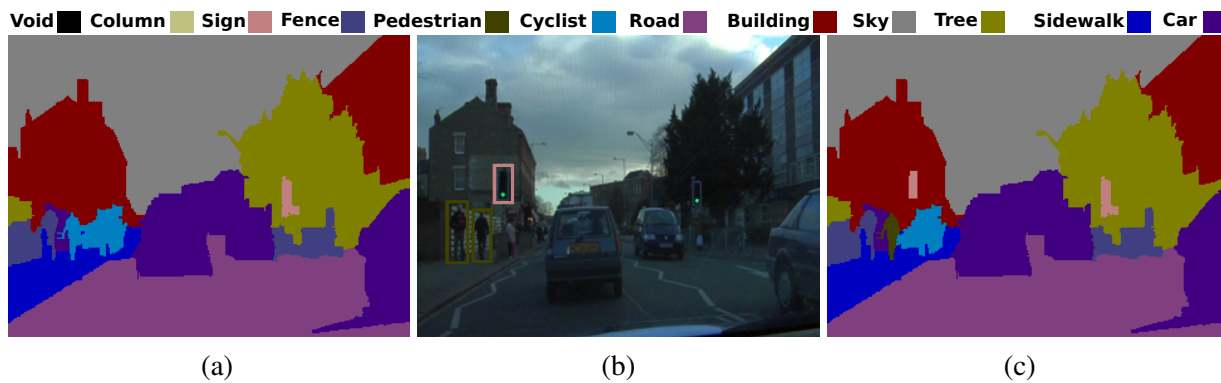


Figure 5-8: **Qualitative Results:** (a) Segmentation without object detectors, (b) Object detection results on this image showing pedestrian and sign/symbol detections, (c) Segmentation using all the detection results. Note that one of the persons (on the left side of the image) is originally labelled as bicyclist (shown in cyan) in (a). This false labelling is corrected in (c) using the detection result. We also show that unary potentials on segments (traffic light on the right), and object detector potentials (traffic light on the left) provide complementary information, thus leading to both the objects being correctly labelled in (c). Some of the regions are labelled incorrectly (the person furthest on the left) perhaps due to a weak detection response. (**Best viewed in colour**)

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	Global	Average
Recall													
[8]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	69.1	53.0
[57]	84.5	72.6	97.5	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	83.8	59.2
Without detectors	79.3	76.0	96.2	74.6	43.2	94.0	40.4	47.0	14.6	81.2	31.1	83.1	61.6
Our method	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	83.8	62.5
IoU													
[57]	71.6	60.4	89.5	58.3	19.4	86.6	26.1	35.0	7.2	63.8	22.6	-	49.2
Without detectors	70.0	63.7	89.5	58.9	17.1	86.3	20.0	35.8	9.2	64.6	23.1	-	48.9
Our method	71.5	63.7	89.4	64.8	19.8	86.8	23.7	35.6	9.3	64.6	26.5	-	50.5

Table 5.2: **Quantitative Results:** We show quantitative results on the CamVid test set on both recall and intersection vs union measures. ‘Global’ refers to the overall percentage of pixels correctly classified, and ‘Average’ is the average of the per class measures. Numbers in bold show the best performance for the respective class under each measure. Our method includes detectors trained on the 5 “thing” classes, namely Car, Sign-Symbol, Pedestrian, Column-Pole, Bicyclist. We clearly see how the inclusion of our detector potentials (‘Our method’) improves over a baseline CRF method (‘Without detectors’), which is based on [37]. For the recall measure, we perform better on 8 out of 11 classes, and for the intersection vs measure, we achieve better results on 9 classes. Note that our method was optimized for intersection vs union measure. Results, where available, of previous methods [8,57] are also shown for reference.

5.6 Energy Layer: Recognition and Reconstruction

In this section the qualitative and quantitative results are reported for our energy level (§3.3) application that combines recognition and reconstruction (§4.3). We evaluated our method on the Leuven Stereo Database [10, 42](§5.2). We quantitatively evaluate the object class segmentation by measuring the percentage of correctly predicted labels over the test sequence. The dense stereo reconstruction performance is quantified by measuring the number of pixels which satisfy $|d_i - d_i^g| \geq \delta$, 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 number of disparities used for evaluation is 100.

Object Class Segmentation The object class segmentation CRF as defined in §4.3.3 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.

Dense Stereo Reconstruction The Potts [34] and linear truncated §4.3.4 (LT) baseline dense stereo reconstruction CRFs performed relatively well, with large δ , considering the difficulty of the data, plotted in fig. 5-9 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 as can be seen in fig. 5-9 ‘LT Filtered’. For qualitative results see fig. 5-10 *F*.

<i>Recall</i>	Building	Sky	Car	Road	Pavement	Bike	Global
Stand alone	96.7	99.8	93.5	99.0	60.2	59.3	95.7
Joint approach	96.7	99.8	94.0	98.9	60.6	59.5	95.8

Table 5.3: **Quantitative Results:** Quantitative results for object class segmentation (recall) of stand alone and joint approach. 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 datasets.

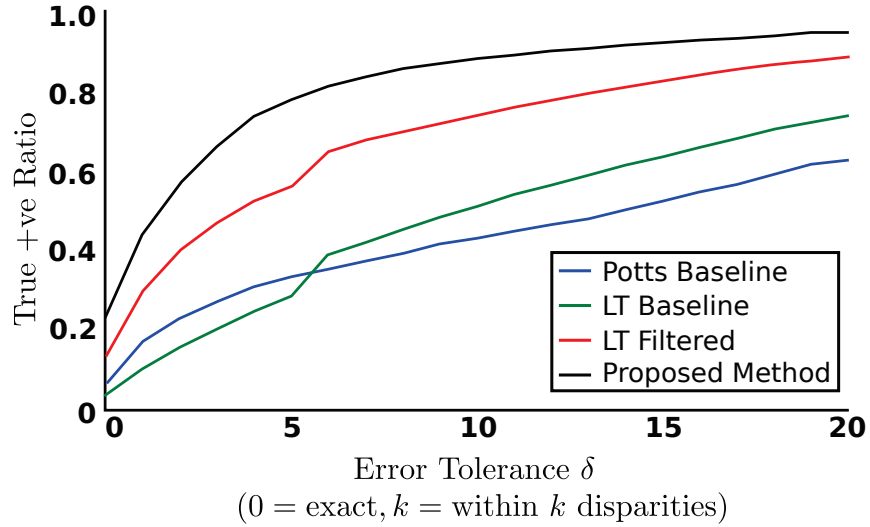


Figure 5-9: **Quantitative comparison of performance of disparity CRFs:** We can clearly see that our joint approach (Proposed Method) outperforms the stand alone approaches with baseline Potts [34] (Potts Baseline), Linear truncated potentials §4.3.4 (LT Baseline) and Linear truncated with Gaussian filtered unary potentials (LT Filtered). The true +ve ratio is the number of pixels which satisfy $d_i - d_i^g \geq \delta$, where d_i is the disparity label of i -th pixel, d_i^g is corresponding ground truth label and δ is the tolerated error.

Joint Approach Our joint approach consistently outperformed the best stand-alone dense stereo reconstruction, by a margin of up to 25%, as can be seen in fig. 5-9 ‘Proposed Method’. Improvement of the object class segmentation was incremental, with 95.8% of predicted pixel labels agreeing with the ground truth, per-class results are presented in Table 5.3. The lack of improvement can be attributed to the two mistakes being the misclassification of *person* as *building*, and the top of a uniformly white building as *sky*. Of these failure cases, 3D location is unable to distinguish between *person* and *building*, while stereo reconstruction fails on homogeneous surfaces. 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 fig 5-10 *C* and *E*.

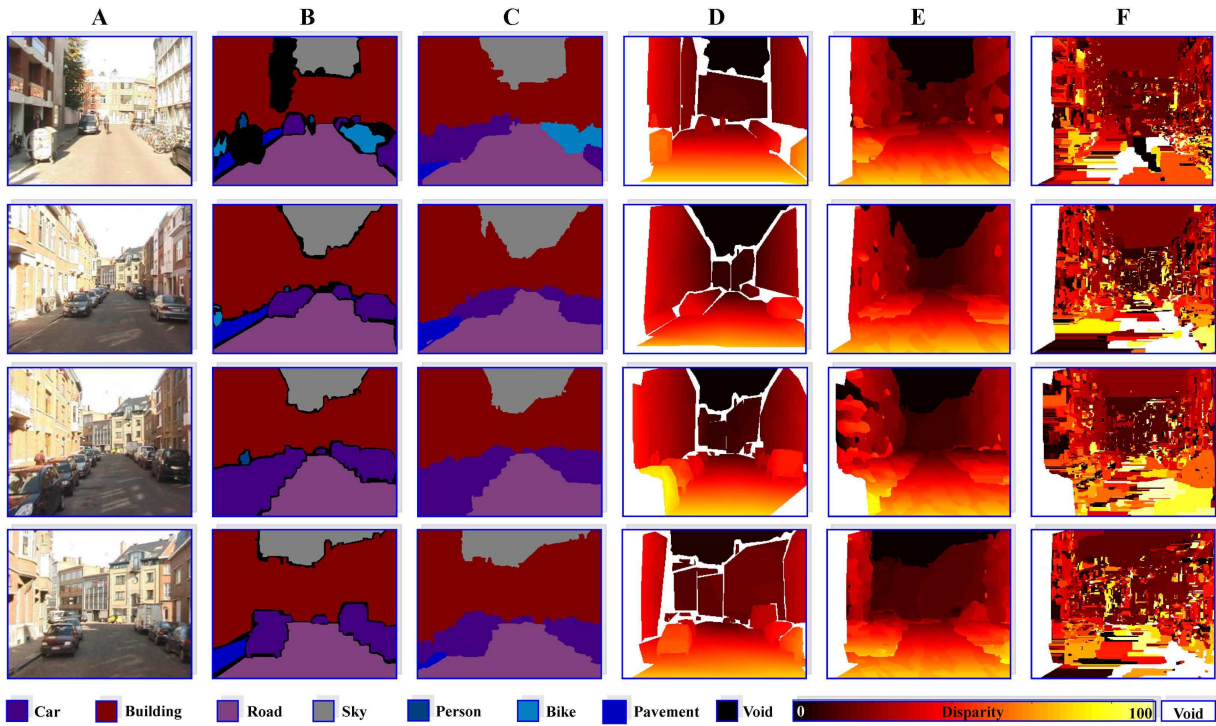


Figure 5-10: **Qualitative object class and disparity results for Leuven data set:** (A) Original Image. (B) Object class segmentation ground truth. (C) Proposed method Object class segmentation result. (D) Dense stereo reconstruction ground truth. (E) Proposed method dense stereo reconstruction result. (F) Stand alone dense stereo reconstruction result (LT Filtered). Best viewed in colour.

Chapter 6

Conclusion

In this dissertation we have specified three ways to fuse information from different sources into the AHRF graphical model with the aim of moving towards a more holistic approach to scene understanding. We concentrated on street scenes because they pose interesting challenges whilst at the same time are highly structured and well suited for experimentations in for holistic models. We empirically tested our approaches showing qualitative and quantitative results to back up our thesis. The three applications were split-up depending on how information was integrated: Fusion of feature, potential and energy functions.

6.1 Feature Layer: Geometry and Appearance

In this application, we have presented a novel principled framework to combine motion and appearance features for object class segmentation problems. Our experiments have shown both quantitative and qualitative evaluations on the challenging CamVid database. We achieve a significant increase in overall accuracy – 84% compared to 69% of the state-of-the-art method [8]. The object class boundaries in the segmentations are well-defined and also detect the fine structures in some categories. Our framework performs worst on classes with the least training data, representing less than 2% of the pixels. We also observed that objects which are a few pixels wide (*e.g.* , columns) in the image are typically merged with other neighbouring superpixel seg-

ments. We are investigating edge-based recognition methods to identify thin structures. Another interesting direction for future research would be to use temporal CRFs.

6.2 Potential Layer: Things and Stuff

In this application, we have presented a novel framework for a principled integration of detectors with CRFs. Unlike many existing methods, our approach supports the robust handling of occluded objects and false detections in an efficient and tractable manner. We believe the techniques described in this application are of interest to many working in the problem of object class segmentation, as they allow the efficient integration of any detector response with any CRF. The benefits of this approach can be seen in the results; our approach consistently demonstrated improvement over the baseline methods, under the intersection vs union measure.

This work increases the expressibility of CRFs and shows how they can be used to identify object instances, and answer the questions: “*What object instance is this?*”, “*Where is it?*”, and “*How many of them?*”.

6.3 Energy Layer: Recognition and Reconstruction

In this application, we have presented a novel approach to the problems of object class recognition and dense stereo reconstruction. 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. 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.

Overall we have shown that a holistic approach to street scene understanding is both pragmatic and effective, by implementing several applications and demonstrating superior results;

We showed that fusion of different aspects of scene understanding achieved in a principled manner through the use of graphical models, and that these graphical models have efficient MAP inference, especially if we constrict our modelling to meet the constraints of the α -expansion move making algorithm.

Bibliography

- [1] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. In *SPIE*, 2001.
- [2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.
- [3] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, 2001.
- [4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 2004.
- [5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.
- [6] G. Bradski. The OpenCV Library. *Dr. Dobbs's Journal of Software Tools*, 2000.
- [7] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *PRL*, 2009.
- [8] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008.
- [9] D. Comaniciu and P. Meer. Robust analysis of feature spaces: color image segmentation. In *CVPR*, 1997.
- [10] N. Cornelis, B. Leibe, K. Cornelis, and L. V. Gool. 3d urban scene modeling integrating recognition and reconstruction. *IJCV*, 2008.
- [11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [12] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.
- [13] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *CVPR*, 2004.
- [14] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

- [15] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. In *Object Representation in Computer Vision II*, Lecture Notes in Computer Science. Springer, 1996.
- [16] B. J. Frey, F. R. Kschischang, H. A. Loeliger, and N. Wiberg. Factor graphs and algorithms. *Proc. 35th Allerton Conf. Communications, Control, and Computing*, 1997.
- [17] A. Fusiello and L. Irsara. Quasi-euclidean uncalibrated epipolar rectification. 2008.
- [18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.
- [19] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011.
- [20] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.
- [21] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009.
- [22] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009.
- [23] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Learning and incorporating top-down cues in image segmentation. In *CVPR*, 2004.
- [24] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008.
- [25] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.
- [26] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *TOG*, 2005.
- [27] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.
- [28] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007.
- [29] D. Hoiem, A. A. Efros, and M. Hebert. Closing the loop in scene interpretation. In *CVPR*, 2008.
- [30] D. Hoiem, C. Rother, and J.M. Winn. 3D layout CRF for multi-view object class recognition and segmentation. In *CVPR*, 2007.
- [31] P. Kohli, M. P. Kumar, and P. H. S. Torr. P & beyond: Move making algorithms for solving higher order functions. *PAMI*, 2009.
- [32] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

- [33] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 2006.
- [34] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions via graph cuts. In *ICCV*, 2001.
- [35] S. Kumar and M. Hebert. Discriminative random fields. *IJCV*, 2006.
- [36] L. Ladicky. *Global Structured Models towards Scene Understanding (PhD Thesis)*. PhD thesis, Oxford Brookes University, 2011.
- [37] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.
- [38] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.
- [39] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order Markov random fields. In *ECCV*, 2006.
- [40] D. Larlus and F. Jurie. Combining appearance models and Markov random fields for category level object segmentation. In *CVPR*, 2008.
- [41] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [42] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D scene analysis from a moving vehicle. In *CVPR*, 2007.
- [43] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*, 2010.
- [44] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with RGBD cameras. In *ICCV*, 2013.
- [45] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010.
- [46] J. Matas, V. Murino, B. Rosenhahn, and L. Leal-Taixé. Holistic Scene Understanding (Dagstuhl Seminar 15081). *Dagstuhl Reports*, 2015.
- [47] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.
- [48] S. Ramalingam, P. Kohli, K. Alahari, and P. H. S. Torr. Exact inference in multi-label CRFs with higher order cliques. In *CVPR*, 2008.
- [49] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, 2005.

- [50] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *TOG*, 2004.
- [51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and Li F. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.
- [52] C. Russell, L. Ladicky, P. Kohli, and P. H. S. Torr. Exact and approximate inference in associative hierarchical networks using graph cuts. *UAI*, 2010.
- [53] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.
- [54] J. R. Shewchuk. Triangle: Engineering a 2D quality mesh generator and delaunay triangulator. In *ACM (Workshop)*, 1996.
- [55] J. Shotton and P. Kohli. *Computer Vision: A Reference Guide*, chapter Semantic Image Segmentation. Springer, 2014.
- [56] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.
- [57] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.
- [58] C. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *ICML*, 2004.
- [59] A. Torralba, K. Murphy, and W. T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.
- [60] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2005.
- [61] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 2005.
- [62] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.
- [63] O. Veksler. Graph cut based optimization for mrfs with truncated convex priors. In *CVPR*, 2007.
- [64] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
- [65] Wikipedia. Semantic gap — wikipedia, the free encyclopedia, 2014. [Online; accessed 25-August-2015].

- [66] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.
- [67] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, 2008.
- [68] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 2014.
- [69] K. Yamaguchi, D. A. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 2014.
- [70] L. Yang, P. Meer, and D. J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*, 2007.
- [71] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012.
- [72] L. Zhu, A. Rao, and A. Zhang. Theory of keyblock-based image retrieval. *ACM Trans. Inf. Syst.*, 2002.
- [73] S. Zhu, C. Guo, Y. Wang, and Z. Xu. What are textons? *IJCV*, 2005.

Appendices

Appendix A

Original Contributions

Combining Appearance and Structure from Motion Features for Road Scene Understanding

Paul Sturges

paul.sturges@brookes.ac.uk

Kartee Alahari

kartee.alahari@brookes.ac.uk

L'ubor Ladický

lladicky@brookes.ac.uk

Philip H. S. Torr

philiptorr@brookes.ac.uk

School of Technology

Oxford Brookes University

Oxford, UK

<http://cms.brookes.ac.uk/research/visiongroup>

Abstract

In this paper we present a framework for pixel-wise object segmentation of road scenes that combines motion and appearance features. It is designed to handle street-level imagery such as that on Google Street View and Microsoft Bing Maps. We formulate the problem in a CRF framework in order to probabilistically model the label likelihoods and the a priori knowledge. An extended set of appearance-based features is used, which consists of textons, colour, location and HOG descriptors. A novel boosting approach is then applied to combine the motion and appearance-based features. We also incorporate higher order potentials in our CRF model, which produce segmentations with precise object boundaries. We evaluate our method both quantitatively and qualitatively on the challenging Cambridge-driving Labeled Video dataset. Our approach shows an overall recognition accuracy of 84% compared to the state-of-the-art accuracy of 69%.

1 Introduction

One of the grand goals of computer vision is to interpret a scene semantically given an input image. This problem has manifested itself in various forms, such as object recognition [8, 16, 25], 3D scene recovery [14], and image segmentation [6, 10, 24]. With the introduction of applications such as Google Street View [2], Microsoft Bing maps [1], the problem of scene understanding has gained more importance than ever. Image sequences from such applications consist of complex scenarios involving multiple *objects*, such as people, buildings, cars, bikes. One may need to simultaneously segment and identify these objects for instance to mask out cars, or maintain highway inventories automatically [3]. This paper deals with the problem of simultaneous pixel-wise segmentation and recognition of such complex image sequences. In particular, we focus on monocular image sequences filmed from within a driven car [9].

Many methods have been proposed to address the object recognition and segmentation problems. Some of them recognize an object and provide a bounding box enclosing it, rather than a pixel-wise segmentation [12, 28]. These approaches are suited better for recognizing rigid objects, such as people and cars, rather than amorphous objects, such as sky and road. Other methods address the challenging task of combined object recognition and pixel-wise segmentation [5, 13, 17, 21]. Although they have achieved impressive results on single object classes, they tend not to scale well for multiple classes. Thus, neither approach is appropriate for complete scene understanding of road scenes consisting of multiple object classes, both rigid and amorphous.

TextonBoost proposed by Shotton *et al.* [25] combines recognition and image segmentation. They use a boosted combination of texton features to encode the shape, texture and appearance of the object classes. A conditional random field (CRF) was then used to combine the result of textons with colour and location based likelihood terms. Although their method produced promising results, the rough shape and texture model caused it to fail at object boundaries. The recent work on image categorization and segmentation using semantic texton forests [26] also suffers from this problem. Kohli *et al.* [16] proposed robust higher order potentials that improve the segmentation result considerably producing a better definition of object boundaries. Brostow *et al.* [8] recently showed that complementing appearance-based features with their motion and structure cues can improve object recognition in challenging datasets captured under varying conditions. However, their approach shares the shortcomings of TextonBoost, in that the resulting segmentation lacks clear object boundaries. Our algorithm builds on these works and addresses the object recognition and segmentation problems simultaneously to produce good object boundaries.

In this paper we present an approach to integrate motion and appearance-based features for object recognition and segmentation of challenging road scenes. The motion-based features are extracted from 3D point clouds, and appearance-based features consist of textons, colour, location, and HOG descriptors [11]. All these features are combined within a boosting framework that automatically selects the most discriminative features for each object class to generate likelihood terms. In addition to the unary likelihood and pairwise potentials, we incorporate higher order terms defined on the image segments generated using unsupervised segmentation algorithms. We perform inference in this framework using the graph cut based α -expansion algorithm [7]. Our method achieves an overall accuracy of 84% compared to the state-of-the-art accuracy of 69% [8] on the challenging new CamVid database [9]. Our paper is inspired by the work of [8] with the following major distinctions: (i) We formulate the problem in a CRF framework in order to probabilistically model the label likelihoods and our prior knowledge in a principled manner. (ii) We use a novel boosting approach to combine the motion and appearance-based features. (iii) We incorporate higher order potentials in our CRF model, which produce accurate segmentations with precise object boundaries. (iv) We use an extended set of appearance-based features. We will highlight these contributions again in the relevant sections.

Outline of the paper. In section 2 we discuss the basic theory of higher order conditional random fields and show how they can be used to model labelling problems such as object segmentation and recognition. The details of the motion and appearance-based unary potentials, computation of higher order potentials, and the inference method are given in section 3. Section 4 describes the dataset and the experimental results. These include qualitative and quantitative evaluations on the CamVid database of video sequences [9]. Concluding remarks and directions for future work are provided in section 5.

2 CRFs for Object Segmentation

Conditional Random Fields have become increasingly popular for modelling object segmentation problems [16, 25]. In this section we briefly describe the pairwise CRF model and the relevant notation.

Consider a set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, where each variable $X_i \in \mathbf{X}$ takes a value from the label set $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$. In our case labels correspond to object classes such as pedestrians, buildings, cars, trees, given in Figure 2 and pixels are the random variables. A labelling \mathbf{x} refers to any possible assignment of labels to the random variables and takes values from the set $\mathbf{L} = \mathcal{L}^N$. The random field is defined over a lattice $\mathcal{V} = \{1, 2, \dots, N\}$, where each lattice point $i \in \mathcal{V}$ is associated with its corresponding random variable X_i . Let \mathcal{N} be the neighbourhood system of the random field defined by 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 defined as a set of random variables \mathbf{X}_c which are conditionally dependent on each other.

We denote the probability of a labelling $\mathbf{X} = \mathbf{x}$ by $\Pr(\mathbf{x})$ and that of a labelling $X_i = x_i$ by $\Pr(x_i)$. A random field is said to be a Markov random field (MRF) with respect to a neighbourhood \mathcal{N} if and only if it satisfies the following two conditions: $\Pr(\mathbf{x}) > 0, \forall \mathbf{x} \in \mathbf{L}$ (positivity); and $\Pr(x_i | \{x_j : j \in \mathcal{V} - \{i\}\}) = \Pr(x_i | \{x_j : j \in \mathcal{N}_i\}), \forall i \in \mathcal{V}$ (Markovianity).

A CRF can be viewed as an MRF globally conditioned on the data \mathbf{D} . The posterior distribution $\Pr(\mathbf{x}|\mathbf{D})$ over the labellings of the CRF is a *Gibbs* distribution and is given by: $\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z} \exp(-\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c))$, where Z is a normalizing constant, and \mathcal{C} is the set of all cliques [19]. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique c , where $\mathbf{x}_c = \{x_i, i \in c\}$. The corresponding Gibbs energy $E(\mathbf{x})$ is given by: $E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$. The most probable or maximum a posteriori (MAP) labelling \mathbf{x}^* of the CRF is defined as: $\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathbf{L}} \Pr(\mathbf{x}|\mathbf{D}) = \arg \min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$.

Energy functions typically used for object segmentation consist of unary (ψ_i) and pairwise (ψ_{ij}) cliques:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j), \quad (1)$$

where \mathcal{V} is the set of image pixels and \mathcal{E} is the set of all pairs of interacting variables denoting the neighbourhood set \mathcal{N} . The labels represent the different objects, and every possible assignment of labels to the random variables (also known as a configuration of the CRF) defines a segmentation. The unary potential $\psi_i(x_i)$ gives the cost of the assignment: $X_i = x_i$. Cost functions based on colour, location, and texton features have been commonly used for object segmentation [4, 17, 25]. The pairwise potential $\psi_{ij}(x_i, x_j)$ represents the cost of the assignment: $X_i = x_i$ and $X_j = x_j$. It is also referred to as the smoothness term, and takes the form of a contrast-sensitive Potts model:

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ \theta_p + \theta_v \exp(-\theta_\beta \|I_i - I_j\|^2) & \text{otherwise,} \end{cases} \quad (2)$$

where I_i and I_j are the colours of pixels i and j respectively. The constants θ_p , θ_v and θ_β are model parameters learned using training data [6, 25].

Higher Order CRFs. There has been much interest in higher order CRFs' in the recent past. They have been successfully used to improve the results of problems such as image denoising, restoration [20, 22], texture segmentation [15], object category segmentation [16]. The improvements can be attributed to the fact that higher order potentials capture the fine details including texture and contours better than pairwise potentials (defined in equation (2) for example).

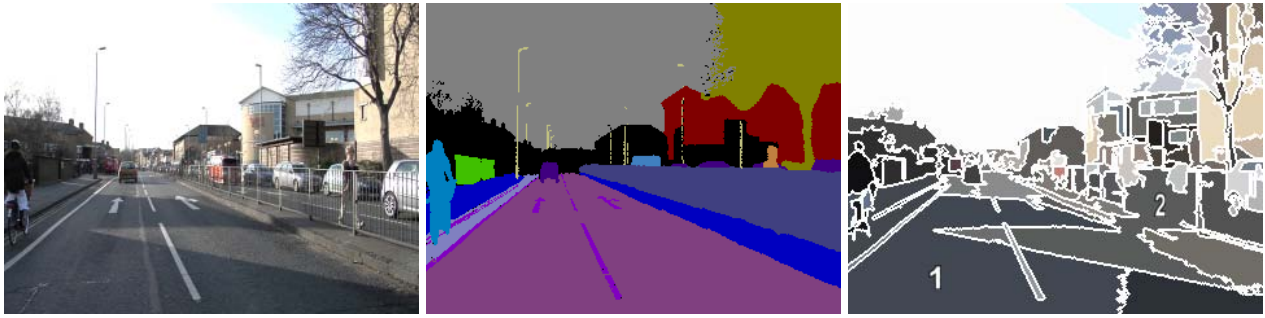


Figure 1: Assigning a single label to all the pixels of a superpixel, as a hard constraint, might produce an incorrect labelling. We show the original image (left), its ground truth labelling (centre) and the meanshift segmentation of the image (right). The segment number ‘1’ consists of all pixels with label Road (a ‘good’ segment), but the segment number ‘2’ consists of pixels with more than one label, viz. Road, Sidewalk, Fence. We use robust higher order potentials to define soft constraints on segments.

Our approach uses the robust P^n model potential defined on the segments obtained by multiple unsupervised segmentations [16]. Methods based on grouping regions for segmentation assume that all pixels constituting a segment belong to one object. Such a hard constraint on the segments is not necessarily valid as shown in Figure 1, where it can be seen that a single segment may cross multiple object-class boundaries. Unlike these methods, we use the soft constraint approach of [16], where higher order potentials are defined on the image segments generated by unsupervised segmentation algorithms. The Gibbs energy of our higher order CRF is given by:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{S}} \psi_c(\mathbf{x}_c), \quad (3)$$

where \mathcal{S} denotes the set of all segments, ψ_c refers to the higher order potential defined on them, and \mathbf{x}_c is the set of all pixels in clique c . We provide more details about the computation of the higher order potential in the next section. The segmentation is obtained by finding the lowest energy configuration of the CRF. We can minimize the energy function in (1) using approximate methods such as α -expansion [7, 16].

3 Computing the Potentials

We now describe the structure from motion and appearance-based features used for computing the energy potentials. Details of the boosting framework used to combine all these weak features and the computation of higher order potentials are also presented.

3.1 Motion and Structure Features

We use the five motion and structure features proposed by [8], namely: height above the camera, distance to the camera path, projected surface orientation, feature track density, and residual reconstruction error. They are computed using the inferred 3D point clouds¹, which are quite noisy due to the small baseline variations. These weak features are designed specifically for such point clouds. As noted by [8], the five cues are tailored for the driving application and are invariant to camera pitch, yaw and perspective distortions. A brief description of the features is given below.

Height above the camera is measured as the difference of the y coordinates of a world point and the camera centre, after aligning the car’s *up* vector as the camera’s $-y$ axis.

¹The point clouds are available as part of the dataset [9].

Distance to the camera path is computed using the entire sequence of camera centres. Let $C(t)$ denote the camera centre in frame t , and W denote a world point. This feature is defined as $\min_t \|W - C(t)\|$. The surface orientation is estimated from the 2D Delaunay triangles [23] formed using the projected world points in a frame. The intuition behind the orientation features is that although individual 3D coordinates may have inaccurate depths, the relative depths of the points gives an approximate local surface orientation. The track density feature exploits the well-known fact that objects yield sparse or dense feature tracks based on how fast they are moving, and their texture. For instance, trees, buildings, and other forms of vegetation yield dense feature tracks, while sky and roads give rise to sparse feature tracks. This cue is measured as the 2D map of the feature density. The residual reconstruction error measures the backprojection error (2D variance) of the estimated 3D world points. This residual error separates moving objects such as people and cars, from stationary ones such as buildings, vegetation, and roads.

All the features are projected from the 3D world onto the 2D image plane and clustered using the K-means algorithm. To include these features into the boosting framework, the feature value at a pixel is given by its cluster assignment. We refer the reader to [8] for more details about the motion-based features and the projection from 3D to 2D.













3.2 Appearance-based Features

We now describe the appearance-based features employed in our framework. In contrast to [8], which uses only texton histograms and localized bag of semantic texton (BOST) features, our approach uses colour, location, texton, and Histogram of Oriented Gradients (HOG) [11] features. We follow the method of [25] to learn a dictionary of textons by convolving a 17-dimensional filter bank (consisting of scaled Gaussians, derivatives of Gaussians, and Laplacians of Gaussians) with all the images and clustering the filter responses. Each pixel is then assigned to the nearest cluster centre, resulting in a texton feature map. The colour feature of a pixel is its assignment to the nearest cluster centre in the CIELuv colour space. The (x, y) pixel locations and HOG features are also clustered, and the feature value at each pixel is its cluster assignment.

3.3 Boosting for Unary Potentials

We use an adapted version [18] of the boosting approach described in TextonBoost [25] to compute the unary potentials, unlike [8] which uses a randomized decision forest. In section 4 we show that our boosting scheme performs better than their randomized decision forest approach. 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 rectangular region r and texton t pair. The feature response $v_i(r, t)$ of the shape filter for a given point i is the number of textons of type t in the region r placed relative to the point i . These filters capture the contextual relationships between objects. Each weak classifier compares the shape filter response to a threshold. The most discriminative filters are found using the Joint Boosting algorithm [27].

The classifiers defined on motion and appearance-based features (given in §3.1 and §3.2) are combined in the adapted boosting approach. The shape filters are now defined by triplets of feature type f , feature cluster t , and rectangular region r . The feature response $v_i(r, f, t)$ for a given point i is the number of features of type f belonging to cluster t in the region r . The weak classifiers compare the responses of shape filters with a set of thresholds. The feature selection and learning procedure is identical to that in [25]. The negative log likelihood given by the classifier is incorporated as the unary potential in the CRF framework.

 Road	 Void	Seq. Name	# Frames	Dataset
 Building	 Column-Pole	0006R0	101	Day Train
 Sky	 Sign-Symbol	0016E5	204	Day Train
 Tree	 Fence	0001TP_1	62	Dusk Train
 Sidewalk	 Pedestrian	Seq05VD	171	Day Test
 Car	 Cyclist	0001TP_2	62	Dusk Test

(a)

(b)

Figure 2: (a) The 11 object class names and their corresponding colours used for labelling. (b) The training and testing data split for both day and dusk sequences. The first half of the dusk sequence (0001TP_1) is used for training, and the second half (0001TP_2) for testing. The frames were extracted for ground truth labelling at a rate of 1 frame per second i.e., by considering every 30th frame. To make our data split identical to that in [8], we ignored the data extracted at 15 fps on one of the sequences (consisting of 101 frames).

3.4 Pairwise and Higher Order Potentials

In [8], the label consistency between neighbouring pixels is partially modelled by the BOST region priors, but the segmentations lack clear object boundaries. In contrast, we incorporate this consistency using pairwise and higher order potential functions. The pairwise potential is given in equation (2). A quality-sensitive higher order potential defines the label inconsistency cost i.e., the cost of assigning different labels to pixels constituting the segment, while taking the quality of a segment into account. We denote the quality of a segment c by $G(c) : c \rightarrow \mathbb{R}$. In our experiments we use the variance of colour intensity values evaluated on all constituent pixels of a segment as a quality measure. The quality-sensitive higher order potential is defined as:

$$\psi_c(\mathbf{x}_c) = \begin{cases} N_i(\mathbf{x}_c) \frac{1}{Q} \gamma_{\max} & \text{if } N_i(\mathbf{x}_c) \leq Q \\ \gamma_{\max} & \text{otherwise,} \end{cases} \quad (4)$$

where $N_i(\mathbf{x}_c)$ denotes the number of pixels in the superpixel c not taking the dominant label, $\gamma_{\max} = |c|^{\theta_\alpha} (\theta_p^h + \theta_v^h G(c))$, and Q is the truncation parameter. This potential ensures the cost of breaking a *good* segment is higher than that of a *bad* segment.

The set \mathcal{S} of segments used for defining the higher order potentials is generated by computing multiple unsupervised segmentations of an image. We choose the mean shift algorithm [10] for this purpose, as it has been shown to give good quality segments. Multiple segmentations are generated by varying the spatial and range parameters.

3.5 Inferring the Segmentation

Kohli *et al.* [16] showed that the robust higher order energy functions defined in the previous section can be efficiently solved by α -expansion and $\alpha\beta$ -swap move making algorithms. In order to compute the optimal moves for these algorithms, higher order move functions need to be minimized. They achieve this by transforming the higher order move functions to quadratic submodular functions by adding auxiliary binary variables. The transformed submodular functions are then minimized by graph cuts.

We follow this approach and use the α -expansion move making algorithm. The solution corresponding to one of the energy minima provides the object class segmentation labelling at each pixel. The class labels are represented with colours shown in Figure 2.

4 Experiments

We evaluated our method on the challenging Cambridge-driving Labelled Video Database (CamVid) [9]. We compare our results to the state-of-the-art method of [8] and achieve 14.7% improvement in overall recognition accuracy. The effectiveness of the proposed approach is shown in terms of both quantitative and qualitative evaluations.

Dataset. The CamVid database² consists of over 10 minutes of high quality 30 Hz footage. The corresponding labelled images are at 1 Hz, and also 15 Hz for one of the video sequences. The videos were captured at 960×720 resolution with a camera mounted inside a car. Several residential, urban, and mixed road sequences are included in the database. Three of the four sequences (0006R0, 0016E5, Seq05VD) were shot in daylight, and the fourth sequence (0001TP) was captured at dusk. Sample frames from the day and dusk sequences are shown in Figure 3. The camera’s intrinsic and extrinsic parameters, 2D feature tracks over all frames, as well as the 3D point clouds are provided in the database.

A selection of frames from the video sequences were manually labelled in an arduous process. Each pixel in these frames was labelled as one of the 32 candidate classes. The assigned labels were verified by a second person. We use a subset of 11 categories: Building, Tree, Sky, Car, Sign-Symbol, Road, Pedestrian, Fence, Column-Pole, Sidewalk, and Bicyclist, from this set for comparison with the work of [8]. A small number of pixels are labelled as *void*, which do not belong to one of these classes and are ignored. The class labels with their corresponding colour codes are shown in Figure 2(a). Interested readers may refer to [9] for details on the database.

Training. The ground truth labelled frames are split into distinct training and testing sets, and are identical to those used in [8]. Figure 2(b) shows the split for the 600 images. All the images are scaled by a factor 3 to speed-up the training process. The five motion and structure features (§3.1) are computed for every frame and normalized to have zero mean and unit variance. All the motion features except surface orientation are clustered³ together, with a maximum number of 150 clusters. Surface orientation features are clustered separately using the same maximum number of clusters. We observed that this clustering scheme provides stronger motion feature candidates for our joint boosting approach. Clustering the five features independently results in very weak features, and most of them are suppressed by the boosting procedure. The appearance-based features (§3.2) are also extracted, and then clustered using maximum numbers of 144, 150, 150, and 128 clusters for location, HOG, texon, and colour respectively. Every pixel is assigned to its nearest cluster centre for all the features, resulting in feature maps. The maps are used in the joint boosting framework to compute the unary likelihood (§3.3).

In our experiments we use three [spatial, range] pair values, *viz.*: {[3.0,0.1], [3.0,0.3], [3.0, 0.9]}, to generate multiple segments using the mean shift algorithm. The minimum segment size (*i.e.*, the number of pixels in a segment) is set to 200 to avoid very small segments. More segmentations using other algorithms can be easily added in our framework. However, we chose three that vary from over-segmented to under-segmented, as suggested in [16]. The higher order potentials are computed using these segments as soft constraints (§3.4). We use the parameters given in [16], because the CamVid database comprises of a subset of the class labels used in [16]. Empirically, we observed that our results were not sensitive to small changes in parameter values.

²Available at <http://mi.eng.cam.ac.uk/research/projects/VideoRec>.

³We use the K-means clustering algorithm in this paper.

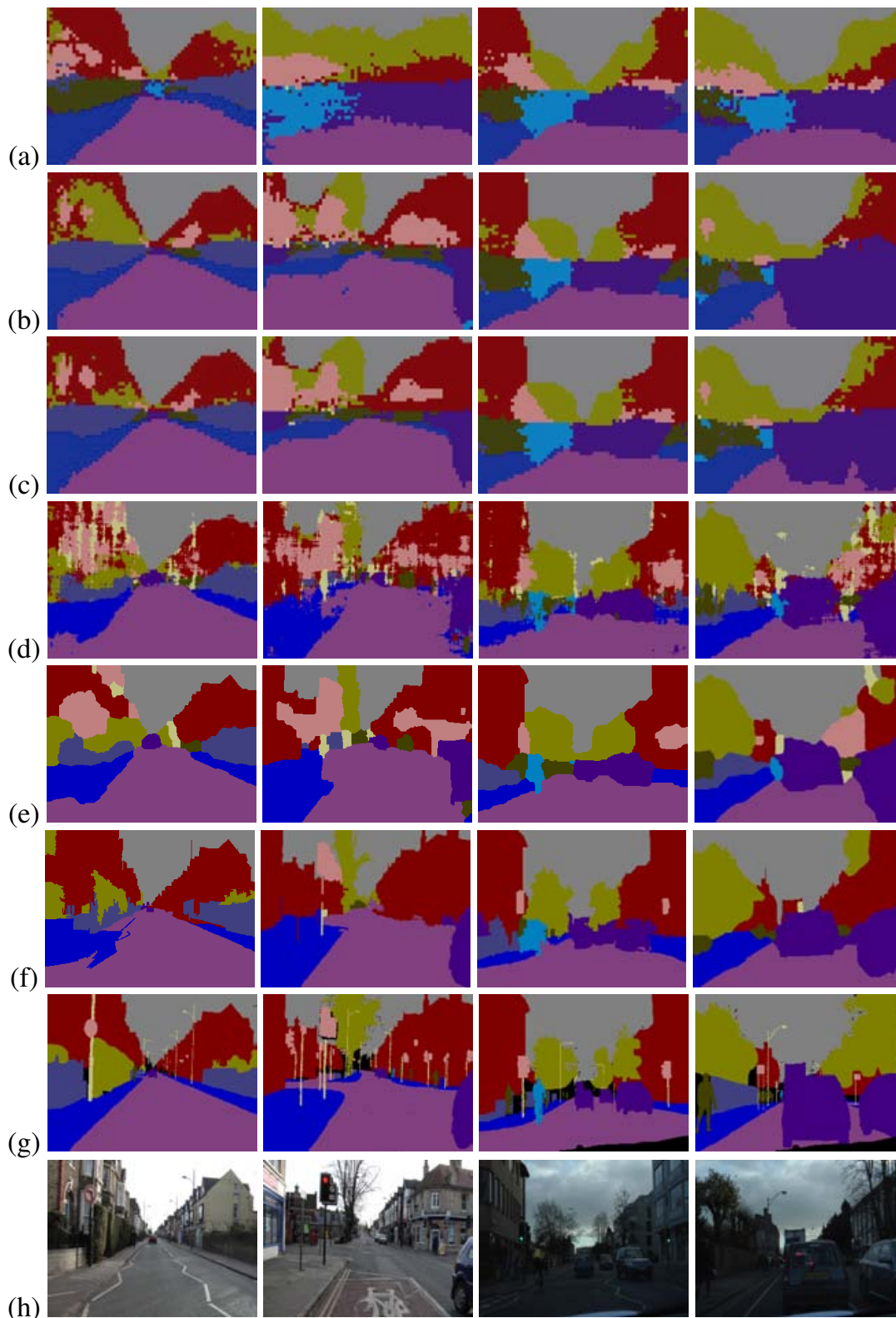


Figure 3: Sample object category segmentations of two day and two dusk images. Results from [8] are shown in: (a) Motion and structure-based segmentation, (b) Appearance-based segmentation, (c) Combined segmentation result. Our results: (d) using only unary potentials gives poor segmentation, (e) adding pairwise potentials improves the segmentation, but fails at object boundaries. The row (f) shows our combined higher order potential based segmentation, which is qualitatively better than (a) - (e). (g) Ground truth labelled image, (h) Original test image. Note that using higher order provides better segmentation, as well as clearer object boundaries.

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	Average	Global
Mot. [8]	43.9	46.2	79.5	44.6	19.5	82.5	24.4	58.8	0.1	61.8	18.0	43.6	61.8
App. [8]	38.7	60.7	90.1	71.1	51.4	88.6	54.6	40.1	1.1	55.5	23.6	52.3	66.5
Combined [8]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1
ψ_i	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4
$\psi_i + \psi_{ij}$	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8
$\psi_i + \psi_{ij} + \psi_c$	84.5	72.6	97.5	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8

Table 1: *Pixel-wise percentage accuracy on all the test sequences. Results of [8] using only motion-based (Mot.), only appearance-based (App.) and both features (Combined) are shown for comparison. We present results of our CRF-based method using only unary terms (ψ_i), unary and pairwise terms ($\psi_i + \psi_{ij}$), and unary, pairwise and higher order terms ($\psi_i + \psi_{ij} + \psi_c$). Note that our method, which uses all the terms, gives the best performance for almost all the classes. ‘Global’ is the percentage of pixels correctly classified, and ‘Average’ is the average of the per-class accuracies.*

Results. Our current implementation takes around 9 hours to train, and 30 – 40 seconds to segment and recognize a test image on a Intel Core 2, 2.4 Ghz, 3GB RAM machine. In Figure 3 we show the qualitative results of our method on sample day and dusk images. We observe that our higher order results have well-defined object boundaries, and are more similar to the ground truth compared to the results of [8]. The quantitative results are summarized in Table 4. We achieve a global accuracy (*i.e.*, the percentage of pixels correctly classified) of 84% in comparison to 69% in [8]. We perform well on most of the object categories. The two categories (Pedestrian, Fence) where our performance is bad is perhaps due to the lack of training data. The training dataset has less than 2% of pixels labelled as one of these categories, which appears to be insufficient to learn the potentials. In some cases the higher order CRF under-performs compared to the pairwise CRF due to objects which are only a few pixels wide in the image *e.g.*, Column-Pole. This is due to the failure of the mean shift segmenter to pick out fine structures. Figure 4 highlights the qualitative improvements achieved by our higher order CRF framework. Note that our method produces precise object class boundaries, and improves the pairwise CRF results significantly. Further results (in the form of a video) are available as supplementary material.

5 Discussion

In this paper we have presented a novel principled framework to combine motion and appearance features for object class segmentation problems. Our experiments have shown both quantitative and qualitative evaluations on the challenging CamVid database. We achieve a significant increase in overall accuracy – 84% compared to 69% of the state-of-the-art method [8]. The object class boundaries in the segmentations are well-defined and also detect the fine structures in some categories. Our framework performs worst on classes with the least training data, representing less than 2% of the pixels. We also observed that objects which are a few pixels wide (*e.g.*, columns) in the image are typically merged with other neighbouring superpixel segments. We are investigating edge-based recognition methods to identify thin structures. Another interesting direction for future research would be to use temporal CRFs.

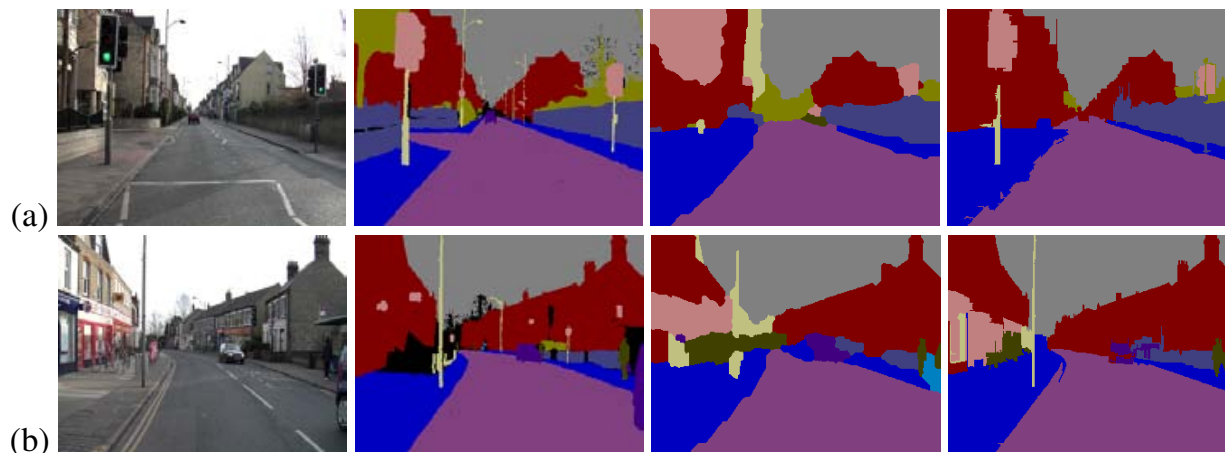


Figure 4: *Qualitative improvements achieved by our higher order CRF framework. We show (left to right) the original image, the ground truth image, pairwise CRF result, and higher order CRF result for two frames from the test sequences. The higher order potentials correct the object boundary errors in the pairwise CRF results e.g., traffic light, and the building in (a). They also provide accurate segmentation, which is more similar to ground truth compared to the pairwise result e.g., lamp post, sidewalk in (b).*

Acknowledgements. This work is supported by EPSRC research grants, HMGCC, the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. P. H. S. Torr is in receipt of Royal Society Wolfson Research Merit Award. We thank Gabriel Brostow for help with the CamVid dataset.

References

- [1] <http://www.bing.com/maps>, 2009.
- [2] <http://maps.google.com/help/maps/streetview>, 2009.
- [3] <http://www.yotta.tv>, 2009.
- [4] A. Blake, C. Rother, M. Brown, P. Perez, and P. H. S. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, volume 1, pages 428–441, 2004.
- [5] E. Borenstein and J. Malik. Shape guided object segmentation. In *CVPR*, volume 1, pages 969–976, 2006.
- [6] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, volume 1, pages 105–112, 2001.
- [7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.
- [8] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, volume 1, pages 44–57, 2008.
- [9] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.
- [10] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space. *PAMI*, 24(5):603–619, 2002.

- [11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.
- [12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.
- [13] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Learning and incorporating top-down cues in image segmentation. In *CVPR*, volume 2, pages 695–702, 2004.
- [14] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.
- [15] P. Kohli, M. P. Kumar, and P. H. S. Torr. P^3 and beyond: Solving energies with higher order cliques. In *CVPR*, 2007.
- [16] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82:302–324, 2009.
- [17] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. Obj cut. In *CVPR*, volume 1, pages 18–25, 2005.
- [18] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.
- [19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, pages 282–289, 2001.
- [20] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order Markov random fields. In *ECCV*, volume 2, pages 269–282, 2006.
- [21] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. *IJCV*, 81(1):105–118, 2009.
- [22] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, volume 2, pages 860–867, 2005.
- [23] J. R. Shewchuk. Triangle: Engineering a 2D quality mesh generator and delaunay triangulator. In *First ACM Workshop on Applied Computational Geometry*, volume 1448, LNCS, pages 203–222, 1996.
- [24] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.
- [25] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, volume 1, pages 1–15, 2006.
- [26] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.
- [27] A. Torralba, K. Murphy, and W. T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, volume 2, pages 762–769, 2004.
- [28] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, volume 2, pages 1800–1807, 2005.

What, Where & How Many?

Combining Object Detectors and CRFs

L'ubor Ladický, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip H.S. Torr *

Oxford Brookes University

<http://cms.brookes.ac.uk/research/visiongroup>

Abstract. Computer vision algorithms for individual tasks such as object recognition, detection and segmentation have shown impressive results in the recent past. The next challenge is to integrate all these algorithms and address the problem of scene understanding. This paper is a step towards this goal. We present a probabilistic framework for reasoning about regions, objects, and their attributes such as object class, location, and spatial extent. Our model is a Conditional Random Field defined on pixels, segments and objects. We define a global energy function for the model, which combines results from sliding window detectors, and low-level pixel-based unary and pairwise relations. One of our primary contributions is to show that this energy function can be solved efficiently. Experimental results show that our model achieves significant improvement over the baseline methods on CamVid and PASCAL VOC datasets.

1 Introduction

Scene understanding has been one of the central goals in computer vision for many decades [1]. It involves various individual tasks, such as object recognition, image segmentation, object detection, and 3D scene recovery. Substantial progress has been made in each of these tasks in the past few years [2–6]. In light of these successes, the challenging problem now is to put these individual elements together to achieve the grand goal — *scene understanding*, a problem which has received increasing attention recently [6, 7]. The problem of scene understanding involves explaining the whole image by recognizing all the objects of interest within an image and their spatial extent or shape. This paper is a step towards this goal. We address the problems of *what*, *where*, and *how many*: we recognize objects, find their location and spatial extent, segment them, and also provide the number of instances of objects. This work can be viewed as an integration of object class segmentation methods [3], which fail to distinguish between adjacent instances of objects of the same class, and object detection approaches [4], which do not provide information about background classes, such as grass, sky and road.

The problem of scene understanding is particularly challenging in scenes composed of a large variety of classes, such as road scenes [8] and images in the PASCAL VOC

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

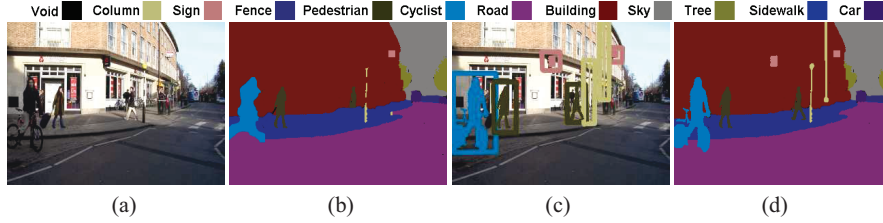


Fig. 1. A conceptual view of our method. (a) An example input image. (b) Object class segmentation result of a typical CRF approach. (c) Object detection result with foreground/background estimate within each bounding box. (d) Result of our proposed method, which jointly infers about objects and pixels. Standard CRF methods applied to complex scenes as in (a) underperform on the “things” classes, e.g. inaccurate segmentation of the bicyclist and persons, and misses a pole and a sign, as seen in (b). However, object detectors tend to perform well on such classes. By incorporating these detection hypotheses (§2.2), shown in (c), into our framework, we aim to achieve an accurate overall segmentation result as in (d) (§3.3). (*Best viewed in colour*)

dataset [9]. For instance, road scene datasets contain classes with specific shapes such as person, car, bicycle, as well as background classes such as road, sky, grass, which lack a distinctive shape (Figure 1). The distinction between these two sets of classes — referred to as *things* and *stuff* respectively — is well known [10–12]. Adelson [10] emphasized the importance of studying the properties of *stuff* in early vision tasks. Recently, these ideas are being revisited in the context of the new vision challenges, and have been implemented in many forms [12–15]. In our work, we follow the definition by Forsyth *et al.* [11], where *stuff* is a homogeneous or reoccurring pattern of fine-scale properties, but has no specific spatial extent or shape, and a *thing* has a distinct size and shape. The distinction between these classes can also be interpreted in terms of localization. *Things*, such as cars, pedestrians, bicycles, can be easily localized by bounding boxes unlike *stuff*, such as road, sky¹.

Complete scene understanding requires not only the pixel-wise segmentation of an image, but also an identification of object instances of a particular class. Consider an image of a road scene taken from one side of the street. It typically contains many cars parked in a row. Object class segmentation methods such as [3, 8, 16] would label all the cars adjacent to each other as belonging to a large car segment or blob, as illustrated in Figure 2. Thus, we would not have information about the number of instances of a particular object—car in this case. On the other hand, object detection methods can identify the number of objects [4, 17], but cannot be used for background (*stuff*) classes.

In this paper, we propose a method to jointly estimate the class category, location, and segmentation of objects/regions in a visual scene. We define a global energy function for the Conditional Random Field (CRF) model, which combines results from detectors (Figure 1(c)), pairwise relationships between mid-level cues such as superpixels, and low-level pixel-based unary and pairwise relations (Figure 1(b)). We also show that, unlike [6, 18], our formulation can be solved efficiently using graph cut based move

¹ Naturally what is classified as things or stuff might depend on either the application or viewing scale, e.g. flowers or trees might be things or stuff.



Fig. 2. (a) Object class segmentation results (without detection), (b) The detection result, (c) Combined segmentation and detection. Object class segmentation algorithms, such as [3], label all the cars adjacent to each other as belonging to one large blob. Detection methods localize objects and provide information about the number of objects, but do not give a segmentation. Our method jointly infers the number of object instances and the object class segmentation. See §2.3 for details. (*Best viewed in colour*)

making algorithms. We evaluate our approach extensively on two widely used datasets, namely Cambridge-driving Labeled Video Database (CamVid) [8] and PASCAL VOC 2009 [9], and show a significant improvement over the baseline methods.

Outline of the paper. Section 1.1 discusses the most related work. Standard CRF approaches for the object segmentation task are reviewed in Section 2.1. Section 2.2 describes the details of the detector-based potential, and its incorporation into the CRF framework. We also show that this novel CRF model can be efficiently solved using graph cut based algorithms in Section 2.3. Implementation details and the experimental evaluation are presented in Section 3. Section 4 discusses concluding remarks.

1.1 Related Work

Our method is inspired by the works on object class segmentation [3, 6, 8, 16], foreground (*thing*) object detection [4, 17], and relating *things* and *stuff* [12]. Whilst the segmentation methods provide impressive results on certain classes, they typically underperform on *things*, due to not explicitly capturing the global shape information of object class instances. On the other hand, detection methods are geared towards capturing this information, but tend to fail on *stuff*, which is amorphous.

A few object detection methods have attempted to combine object detection and segmentation sub-tasks, however they suffer from certain drawbacks. Larlus and Jurie [19] obtained an initial object detection result in the form of a bounding box, and then refined this rectangular region using a CRF. A similar approach has been followed by entries based on object detection algorithms [4] in the PASCAL VOC 2009 [9] segmentation challenge. This approach is not formulated as one energy cost function and cannot be applied to either cluttered scenes or *stuff* classes. Furthermore, there is no principled way of handling multiple overlapping bounding boxes. Tu *et al.* [15] also presented an effective approach for identifying text and faces, but leave much of the

image unlabelled. Gu *et al.* [20] used regions for object detection instead of bounding boxes, but were restricted to using a single over-segmentation of the image. Thus, their approach cannot recover from any errors in this initial segmentation step. In comparison, our method does not make such *a priori* decisions, and jointly reasons about segments and objects.

The work of layout CRF [21] also provides a principled way to integrate things and stuff. However, their approach requires that things must conform to a predefined structured layout of parts, and does not allow for the integration of arbitrary detector responses. To our knowledge, the only other existing approaches that attempt to jointly estimate segmentation and detection in one optimization framework are the works of [6, 18]. However, the minimization of their cost functions is intractable and their inference methods can get easily stuck in local optima. Thus, their incorporation of detector potentials does not result in a significant improvement of performance. Also, [6] focussed only on two classes (cars and pedestrians), while we handle many types of objects (e.g. 20 classes in the PASCAL VOC dataset). A direct comparison with this method was not possible as neither their code nor their dataset because ground truth annotations are not publicly available at the time of publication.

2 CRFs and Detectors

We define the problem of jointly estimating segmentation and detection in terms of minimizing a global energy function on a CRF model. Our approach combines the results from detectors, pairwise relationships between superpixels, and other low-level cues. Note that our framework allows us to incorporate any object detection approach into any pixel or segment based CRF.

2.1 CRFs for labelling problems

In the standard CRF formulation for image labelling problems [3] we represent each pixel as random variable. Each of these random variables takes a label from the set $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$, which may represent objects such car, airplane, bicycle. Let $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$ denote the set of random variables corresponding to the image pixels $i \in \mathcal{V} = \{1, 2, \dots, N\}$. A clique c is a set of random variables \mathbf{X}_c which are conditionally dependent on each other. A labelling \mathbf{x} refers to any possible assignment of labels to the random variables and takes values from the set $\mathbf{L} = \mathcal{L}^N$.

The posterior distribution $\Pr(\mathbf{x}|\mathbf{D})$ over the labellings of the CRF can be written as: $\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z} \exp(-\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c))$, where Z is a normalizing constant called the *partition function*, \mathcal{C} is the set of all cliques, and \mathbf{D} the given data. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique $c \subseteq \mathcal{V}$, where $\mathbf{x}_c = \{x_i : i \in c\}$. The corresponding Gibbs energy is given by: $E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$. The most probable or Maximum a Posteriori (MAP) labelling \mathbf{x}^* of the random field is defined as: $\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathbf{L}} \Pr(\mathbf{x}|\mathbf{D}) = \arg \min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$.

In computer vision labelling problems such as segmentation or object recognition, the energy $E(\mathbf{x})$ is typically modelled as a sum of unary, pairwise [3, 22], and higher order [23] potentials. The unary potentials are based on local feature responses and

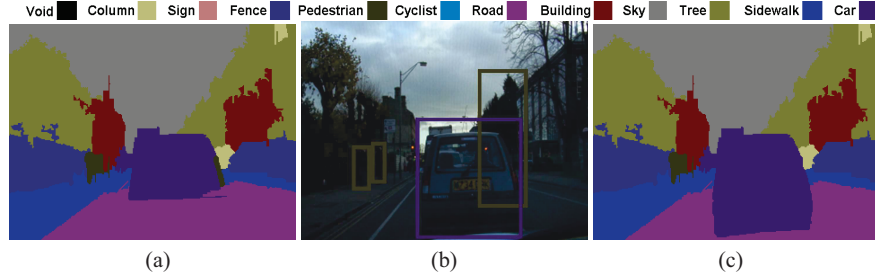


Fig. 3. (a) Segmentation without object detectors, (b) Object detections for car and pedestrian shown as bounding boxes, (c) Segmentation using our method. These detector potentials act as a soft constraint. Some false positive detections (such as the large green box representing person) do not affect the final segmentation result in (c), as it does not agree with other strong hypotheses based on pixels and segments. On the other hand, a strong detector response (such as the purple bounding box around the car) correctly relabels the road and pedestrian region as car in (c) resulting in a more accurate object class segmentation. *(Best viewed in colour)*

capture the likelihood of a pixel taking a certain label. Pairwise potentials encourage neighbouring pixels in the image to take the same label. Similarly, a CRF can be defined over segments [24, 25] obtained by unsupervised segmentation [26, 27] of the image. Recently, these models have been generalized to include pixels and segments in a single CRF framework by introducing higher order potentials [16]. All these models successfully reason about pixels and/or segments. However, they fail to incorporate the notion of object instances, their location, and spatial extent (which are important cues used by humans to understand a scene) into the recognition framework. Thus, these models are insufficient to address the problem of scene understanding. We aim to overcome these issues by introducing novel object detector based potentials into the CRF framework.

2.2 Detectors in CRF framework

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). Existing methods [16, 24, 25] are limited to such hypotheses provided by pixels and/or segments only. We introduce an additional set of hypotheses representing object detections for the recognition framework².

Some object detection approaches [4, 19] have used their results to perform a segmentation within the detected areas³. These approaches include both the true and false

² Note that our model chooses from a set of given detection hypotheses, and does not propose any new detections.

³ As evident in some of the PASCAL VOC 2009 segmentation challenge entries.

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 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. We illustrate this with one of our results shown in Figure 3. Here, the false positive detection for “person” class (shown as the large green box on the right) does not affect the segmentation result in (c). Although, the true positive detection for “car” class (shown as the purple box) refines the segmentation because it agrees with other hypotheses. This is achieved by using the object detector responses⁴ to define a clique potential over the pixels, as described below.

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 shows the inclusion of this potential graphically on a pixel-based CRF. The new energy function is given by:

$$E(\mathbf{x}) = E_{pix}(\mathbf{x}) + \sum_{d \in \mathcal{D}} \psi_d(\mathbf{x}_d, H_d, l_d), \quad (1)$$

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 $\psi_d(\cdot)$ is the minimum of the energy values provided by including ($y_d = 1$) and excluding ($y_d = 0$) the detector hypothesis, as given below:

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d \in \{0, 1\}} \phi_d(y_d, \mathbf{x}_d, H_d, l_d). \quad (2)$$

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:

⁴ This includes sliding window detectors as a special case.

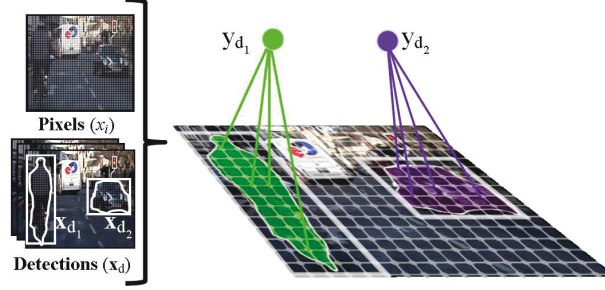


Fig. 4. Inclusion of object detector potentials into a CRF model. We show a pixel-based CRF as an example here. The set of pixels in a detection d_1 (corresponding to the bicyclist in the scene) is denoted by \mathbf{x}_{d_1} . A higher order clique is defined over this detection window by connecting 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)

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d \in \{0, 1\}} (-f(\mathbf{x}_d, H_d)y_d + g(N_d, H_d)y_d). \quad (3)$$

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:

$$f(\mathbf{x}_d, H_d) = w_d |\mathbf{x}_d| \max(0, H_d - H_t), \quad (4)$$

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 of the number of inconsistent pixels N_d of the form:

$$g(N_d, H_d) = k_d N_d, \quad k_d = \frac{f(\mathbf{x}_d, H_d)}{p_d |\mathbf{x}_d|}, \quad (5)$$

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

Detectors may be applied directly, especially if they estimate foreground pixels themselves. However, in this work, we use sliding window detectors, which provide a bounding box around objects. To obtain a more accurate set of pixels \mathbf{x}_d that belong to the object, we use a local colour model [28] to estimate foreground and background within the box. This is similar to the approach used by submissions in the PASCAL VOC 2009 segmentation challenge. Any other foreground estimation techniques may be used. See §3 for more details on the detectors used. Note that equation (1) could be defined in a similar fashion over superpixels.

2.3 Inference for detector potentials

One of the main advantages of our framework is that the associated energy function can be solved efficiently using graph cut [29] based move making algorithms (which outperform message passing algorithms [30, 31] for many vision problems). We now show that our detector potential in equation (3) can be converted into a form solvable using $\alpha\beta$ -swap and α -expansion algorithms [2]. In contrast, the related work in [6] suffers from a difficult to optimize energy. Using equations (3), (4), (5), and $N_d = \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)$, the detector potential $\psi_d(\cdot)$ can be rewritten as follows:

$$\begin{aligned} \psi_d(\mathbf{x}_d, H_d, l_d) &= \min(0, -f(\mathbf{x}_d, H_d) + k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)) \\ &= -f(\mathbf{x}_d, H_d) + \min(f(\mathbf{x}_d, H_d), k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)). \end{aligned} \quad (6)$$

This potential takes the form of a Robust P^N potential [23], which is defined as:

$$\psi_h(\mathbf{x}) = \min(\gamma_{max}, \min_l(\gamma_l + k_l \sum_{i \in \mathbf{x}} \delta(x_i \neq l))), \quad (7)$$

where $\gamma_{max} = f(\cdot)$, $\gamma_l = f(\cdot)$, $\forall l \neq d$, and $\gamma_d = 0$. Thus it can be solved efficiently using $\alpha\beta$ -swap and α -expansion algorithms as shown in [23]. The detection instance variables y_d can be recovered from the final labelling by computing y_d as:

$$y_d = \arg \min_{y'_d \in \{0,1\}} (-f(\mathbf{x}_d, H_d)y'_d + g(N_d, H_d)y'_d). \quad (8)$$

3 Experimental Evaluation

We evaluated our framework on the CamVid [8] and PASCAL VOC 2009 [9] datasets.

CamVid. The Cambridge-driving Labeled Video Database (CamVid) consists of over 10 minutes of high quality 30 Hz footage. The videos are captured at 960×720 resolution with a camera mounted inside a car. Three of the four sequences were shot in daylight, and the fourth sequence was captured at dusk. Sample frames from the day and dusk sequences are shown in Figures 1 and 3. Only a selection of frames from the video sequences are manually annotated. Each pixel in these frames was labelled as one of the 32 candidate classes. We used the same subset of 11 class categories as [8, 32] for experimental analysis. We have detector responses for the 5 *thing* classes, namely Car, Sign-Symbol, Pedestrian, Column-Pole, and Bicyclist. A small number of pixels were labelled as *void*, which do not belong to one of these classes and are ignored. The dataset is split into 367 training and 233 test images. To make our experimental setup the same as [8, 32], we scaled all the images by a factor of 3.

PASCAL VOC 2009. This dataset was used for the PASCAL Visual Object Category segmentation contest 2009. It contains 14,743 images in all, with 20 foreground (*things*)

classes and 1 background (*stuff*) class. We have detector responses for all foreground classes. Each image has an associated annotation file with the bounding boxes and the object class label for each object in the image. A subset of these images are also annotated with pixel-wise segmentation of each object present. We used only these images for training our framework. It contains 749 training, 750 validation, and 750 test images.

3.1 CRF Framework

We now describe the baseline CRF formulation used in our experiments. Note that any CRF formulation based on pixels or segments could have been used. We use the Associative Hierarchical CRF model [16], which combines features at different quantization levels of the image, such as pixels, segments, and is a generalization of commonly used pixel and segment-based CRFs. We have a base layer of variables corresponding to pixels, and a hierarchy of auxiliary variables, which encode mid-level cues from and between segments. Furthermore, it assumes that pixels in the same segment obtained using unsupervised segmentation methods, are highly correlated, but are not required to take the same label. This allows us to incorporate multiple segmentations in a principled approach.

In our experiments we used a two level hierarchy based on pixels and segments. Three segmentations are used for the CamVid dataset and six for the PASCAL VOC 2009 dataset; these were obtained by varying parameters of the MeanShift algorithm [26], similar to [16, 32].

Pixel-based potentials. The pixel-based unary potential is identical to that used in [16, 32], and is derived from *TexonBoost* [3]. It estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. Shape filters are defined by triplets of feature type, feature cluster, and rectangular region and their response for a given pixel is the number of features belonging to the given cluster in the region placed relative to the given pixel. The most discriminative filters are found using the Joint Boosting algorithm [14]. Details of the learning procedure are given in [3, 16]. To enforce local consistency between neighbouring pixels we use the standard contrast sensitive Potts model [22] as the pairwise potential on the pixel level.

Segment-based potentials. We also learn unary potentials for variables in higher layers (*i.e.* layers other than the base layer), which represent segments or super-segments (groups of segments). The segment unary potential is also learnt using the Joint Boosting algorithm [14]. The pairwise potentials in higher layers (*e.g.* pairwise potentials between segments) are defined using a contrast sensitive (based on distance between colour histogram features) Potts model. We refer the reader to [16] for more details on these potentials and the learning procedure.

3.2 Detection-based potentials

The object detections are included in the form of a higher order potential over pixels based on detector responses, as detailed in §2.2. The implementation details of this potential are described below. In order to jointly estimate the class category, location,

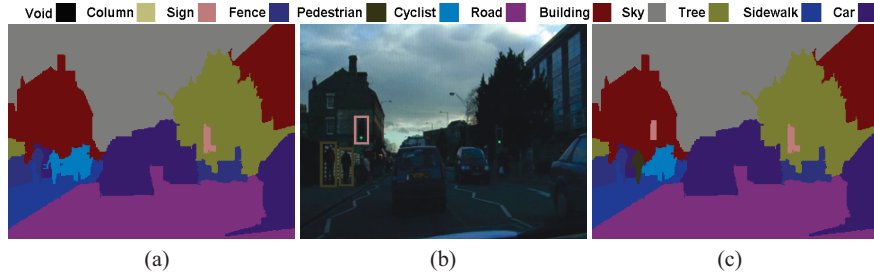


Fig. 5. (a) Segmentation without object detectors, (b) Object detection results on this image showing pedestrian and sign/symbol detections, (c) Segmentation using all the detection results. Note that one of the persons (on the left side of the image) is originally labelled as bicyclist (shown in cyan) in (a). This false labelling is corrected in (c) using the detection result. We also show that unary potentials on segments (traffic light on the right), and object detector potentials (traffic light on the left) provide complementary information, thus leading to both the objects being correctly labelled in (c). Some of the regions are labelled incorrectly (the person furthest on the left) perhaps due to a weak detection response. (*Best viewed in colour*)

and segmentation of objects, we augment the standard CRF using responses of two of the most successful detectors⁵: (i) histogram-based detector proposed in [17]; and (ii) parts-based detector proposed in [4]. Other detector methods could similarly be incorporated into our framework.

In [17], histograms of multiple features (such as bag of visual words, self-similarity descriptors, SIFT descriptors, oriented edges) were used to train a cascaded classifier composed of Support Vector Machines (SVM). The first stage of the cascade is a linear SVM, which proposes candidate object windows and discards all the windows that do not contain an object. The second and third stages are more powerful classifiers using quasi-linear and non-linear SVMs respectively. All the SVMs are trained with ground truth object instances [9]. The negative samples (which are prohibitively large in number) are obtained by bootstrapping each classifier, as follows. Potential object regions are detected in the training images using the classifier. These potential object regions are compared with the ground truth, and a few of the incorrect detections are added to the training data as negative samples. The SVM is then retrained using these negative and the positive ground truth samples.

In [4] each object is composed of a set of deformable parts and a global template. Both the global template and the parts are represented by HOG descriptors [33], but computed at a coarse and fine level respectively. The task of learning the parts and the global template is posed as a latent SVM problem, which is solved by an iterative method. The negative samples are obtained by bootstrapping the classifier, as described above.

Both these methods produce results as bounding boxes around the detected objects along with a score, which represents the likelihood of a box containing an object. A more accurate set of pixels belonging to the detected object is obtained using local

⁵ We thank the authors of [4, 17] for providing their detections on the PASCAL VOC 2009 dataset.

Table 1. We show quantitative results on the CamVid test set on both recall and intersection vs union measures. ‘Global’ refers to the overall percentage of pixels correctly classified, and ‘Average’ is the average of the per class measures. Numbers in bold show the best performance for the respective class under each measure. Our method includes detectors trained on the 5 “thing” classes, namely Car, Sign-Symbol, Pedestrian, Column-Pole, Bicyclist. We clearly see how the inclusion of our detector potentials (‘Our method’) improves over a baseline CRF method (‘Without detectors’), which is based on [16]. For the recall measure, we perform better on 8 out of 11 classes, and for the intersection vs measure, we achieve better results on 9 classes. Note that our method was optimized for intersection vs union measure. Results, where available, of previous methods [8, 32] are also shown for reference.

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	Global	Average
Recall ⁶													
[8]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	69.1	53.0
[32]	84.5	72.6	97.5	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	83.8	59.2
Without detectors	79.3	76.0	96.2	74.6	43.2	94.0	40.4	47.0	14.6	81.2	31.1	83.1	61.6
Our method	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	83.8	62.5
Intersection vs Union ⁷													
[32]	71.6	60.4	89.5	58.3	19.4	86.6	26.1	35.0	7.2	63.8	22.6	-	49.2
Without detectors	70.0	63.7	89.5	58.9	17.1	86.3	20.0	35.8	9.2	64.6	23.1	-	48.9
Our method	71.5	63.7	89.4	64.8	19.8	86.8	23.7	35.6	9.3	64.6	26.5	-	50.5

foreground and background colour models [28]. In our experiments we observed that the model is robust to change in detector potential parameters. The parameter p_d (from equation (5)) can be set anywhere in the range 10% – 40%. The parameter H_t (which defines the detector threshold, equation (4)) can be set to 0 for most of the SVM-based classifiers. To compensate the bias towards foreground classes the unary potentials of background class(es) were weighted by factor w_b . This bias weight and the detector potential weight w_d were learnt along with the other potential weights on the validation set using the greedy approach presented in [16]. The CRF was solved efficiently using the graph cut based α -expansion algorithm [2, 23].

3.3 Results

Figures 2, 3 and 5 show qualitative results on the CamVid dataset. Object segmentation approaches do not identify the number of instances of objects, but this information is recovered using our combined segmentation and detection model (from y_d variables, as discussed in §2.3), and is shown in Figure 2. Figure 3 shows the advantage of our soft constraint approach to include detection results. The false positive detection here (shown as the large green box) does not affect the final segmentation, as the other hypotheses based on pixels and segments are stronger. However, a strong detector hypothesis (shown as the purple box) refines the segmentation accurately. Figure 5 highlights the complementary information provided by the object detectors and segment-based

⁶ Defined as $\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$.

⁷ Defined as $\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative} + \text{False Positive}}$; also used in PASCAL VOC challenges.

Table 2. Quantitative analysis of VOC 2009 test dataset results [9] using the intersection vs union performance measure. Our method is ranked **third** when compared the 6 best submissions in the 2009 challenge. The method UOCTTI_L SVM-MDPM is based on an object detection algorithm [4] and refines the bounding boxes with a GrabCut style approach. The method BROOKESMSRC_AHCRF is the CRF model used as an example in our work. We perform better than both these baseline methods by 3.1% and 7.3% respectively. Underlined numbers in bold denote the best performance for each class.

	Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table	Dog	Horse	Motorbike	Person	Potted plant	Sheep	Sofa	Train	TVmonitor	Average
BONN_SVM-SEG	83.9	64.3	21.8	21.7	32.0	40.2	57.3	49.4	38.8	5.2	28.5	22.0	19.6	33.6	45.5	33.6	27.3	40.4	18.1	33.6	46.1	36.3
CVC_HOCR	80.2	67.1	26.6	30.3	31.6	30.0	44.5	41.6	25.2	5.9	27.8	11.0	23.1	40.5	53.2	32.0	22.2	37.4	23.6	40.3	30.2	34.5
UOCTTI_L SVM-MDPM	78.9	35.3	22.5	19.1	23.5	36.2	41.2	50.1	11.7	8.9	28.5	1.4	5.9	24.0	35.3	33.4	35.1	27.7	14.2	34.1	41.8	29.0
NEUJUC_CLS-DTCT	81.8	41.9	23.1	22.4	22.0	27.8	43.2	51.8	25.9	4.5	18.5	18.0	23.5	26.9	36.6	34.8	18.8	28.3	14.0	35.5	34.7	29.7
LEAR_SEGDET	79.1	44.6	15.5	20.5	13.3	28.8	29.3	35.8	25.4	4.4	20.3	1.3	16.4	28.2	30.0	24.5	12.2	31.5	18.3	28.8	31.9	25.7
BROOKESMSRC_AHCRF	79.6	48.3	6.7	19.1	10.0	16.6	32.7	38.1	25.3	5.5	9.4	25.1	13.3	12.3	35.5	20.7	13.4	17.1	18.4	37.5	36.4	24.8
Our method	81.2	46.1	15.4	24.6	20.9	36.9	50.0	43.9	28.4	11.5	18.2	25.4	14.7	25.1	37.7	34.1	27.7	29.6	18.4	43.8	40.8	32.1

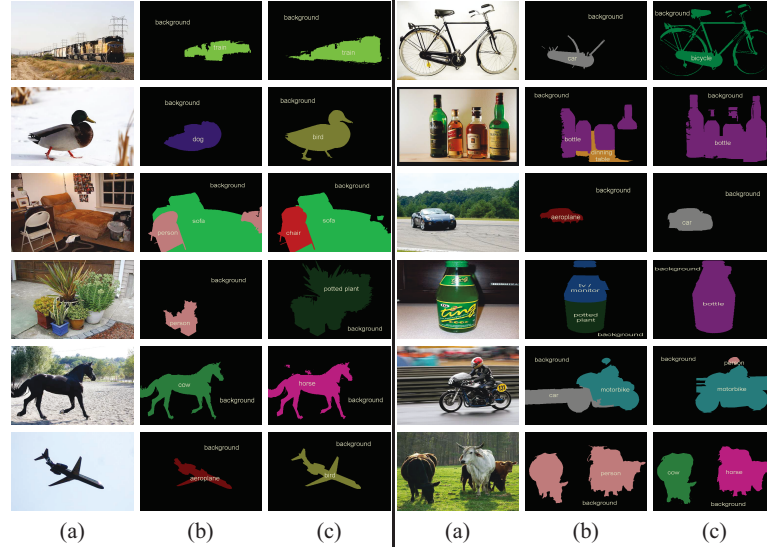


Fig. 6. (a) Original test image from PASCAL VOC 2009 dataset [9], (b) The labelling obtained by [16] without object detectors, (c) The labelling provided by our method which includes detector based potentials. Note that no groundtruth is publicly available for test images in this dataset. Examples shown 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 row. This was caused by a missed detection or incorrect detections that are very strong and dominate all other potentials. (Best viewed in colour)

potentials. An object falsely missed by the detector (traffic light on the right) is recognized based on the segment potentials, while another object (traffic light on the left) overlooked by the segment potentials is captured by the detector. More details are provided in the figure captions. Quantitative results on the CamVid dataset are shown in

Table 1. For the recall measure, our method performs the best on 5 of the classes, and shows near-best ($< 1\%$ difference in accuracy) results on 3 other classes. Accuracy of “things” classes improved by 7% on average. This measure does not consider false positives, and creates a bias towards smaller classes. Therefore, we also provide results with the intersection vs union measure in Table 1. We observe that our method shows improved results on almost all the classes in this case.

Qualitative results on PASCAL VOC 2009 test set are shown in Figure 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 dataset are provided in Table 2. We compare our results with the 6 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. Note that using the detector based work (UOCTTI_L SVM-MDPM: 29.0%) and pixel-based method (BROOKESMSRC_AHCRF: 24.8%) as examples in our framework, we improve the accuracy to 32.1%. Both the BONN [34] and CVC [35] methods can be directly placed in our work, and should lead to an increase in performance.

4 Summary

We have presented a novel framework for a principled integration of detectors with CRFs. Unlike many existing methods, our approach supports the robust handling of occluded objects and false detections in an efficient and tractable manner. We believe the techniques described in this paper are of interest to many working in the problem of object class segmentation, as they allow the efficient integration of any detector response with any CRF. The benefits of this approach can be seen in the results; our approach consistently demonstrated improvement over the baseline methods, under the intersection vs union measure.

This work increases the expressibility of CRFs and shows how they can be used to identify object instances, and answer the questions: “*What object instance is this?*”, “*Where is it?*”, and “*How many of them?*”, bringing us one step closer to complete scene understanding.

References

1. Barrow, H.G., Tenenbaum, J.M.: Computational vision. IEEE **69** (1981) 572–595
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23** (2001) 1222–1239
3. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV. (2006)
4. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
5. Hoiem, D., Efros, A., Hebert, M.: Closing the loop on scene interpretation. In: CVPR. (2008)
6. Gould, S., Gao, T., Koller, D.: Region-based segmentation and object detection. In: NIPS. (2009)

7. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: CVPR. (2009)
8. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: ECCV. Volume 1. (2008) 44–57
9. Everingham, M., et al.: The PASCAL Visual Object Classes Challenge (VOC) Results (2009)
10. Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: SPIE. Volume 4299. (2001) 1–12
11. Forsyth, D.A., et al.: Finding pictures of objects in large collections of images. In: ECCV Workshop on Object Representation in Computer Vision II. (1996) 335–360
12. Heitz, G., Koller, D.: Learning spatial context: Using stuff to find things. In: ECCV. (2008)
13. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV. (2007)
14. Torralba, A., Murphy, K., Freeman, W.T.: Sharing features: Efficient boosting procedures for multiclass object detection. In: CVPR. Volume 2. (2004) 762–769
15. Tu, Z., et al.: Image parsing: Unifying segmentation, detection, and recognition. IJCV (2005)
16. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Associative hierarchical crfs for object class image segmentation. In: ICCV. (2009)
17. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV. (2009)
18. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: ECCV. (2008)
19. Larlus, D., Jurie, F.: Combining appearance models and markov random fields for category level object segmentation. In: CVPR. (2008)
20. Gu, C., Lim, J., Arbelaez, P., Malik, J.: Recognition using regions. In: CVPR. (2009)
21. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: CVPR. (2006)
22. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: ICCV. Volume 1. (2001) 105–112
23. Kohli, P., Ladicky, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. In: CVPR. (2008)
24. He, X., Zemel, R.S., Carreira-Perpiñán, M.Á.: Learning and incorporating top-down cues in image segmentation. In: CVPR. Volume 2. (2004) 695–702
25. Yang, L., Meer, P., Foran, D.J.: Multiple class segmentation using a unified framework over mean-shift patches. In: CVPR. (2007)
26. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space. PAMI (2002)
27. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI **22** (2000) 888–905
28. Rother, C., Kolmogorov, V., Blake, A.: GrabCut. In: SIGGRAPH. (2004) 309–314
29. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI **26** (2004) 1124–1137
30. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. In: CVPR. (2004)
31. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI **28** (2006) 1568–1583
32. Sturgess, P., Alahari, K., Ladicky, L., Torr, P.H.S.: Combining appearance and structure from motion features for road scene understanding. In: BMVC. (2009)
33. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893
34. Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: CVPR. (2010)
35. Gonfaus, J.M., Boix, X., van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzalez, J.: Harmony potentials for joint classification and segmentation. In: CVPR. (2010)

Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction

L'ubor Ladický

lladicky@brookes.ac.uk

Paul Sturgess

paul.sturgess@brookes.ac.uk

Chris Russell

chris.russell@brookes.ac.uk

Sunando Sengupta

ssengupta@brookes.ac.uk

Yalin Bastanlar

yalinbastanlar@brookes.ac.uk

William Clocksin

wfc@brookes.ac.uk

Philip H. S. Torr

philiptorr@brookes.ac.uk

School of Technology

Oxford Brookes University

Oxford, UK

cms.brookes.ac.uk/research/visiongroup

This work is supported by EPSRC research grants, HMGCC, TUBITAK researcher exchange grant, the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

P. H. S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

Abstract

The problems of *dense stereo reconstruction* and *object class segmentation* can both be formulated as Conditional Random Field based labelling problems, in which every pixel in the image is assigned a label corresponding to either its disparity, or an object class such as road or building. While these two problems are mutually informative, no attempt has been made to jointly optimise their labellings. In this work we provide a principled energy minimisation framework that unifies the two problems and demonstrate that, by resolving ambiguities in real world data, joint optimisation of the two problems substantially improves performance. To evaluate our method, we augment the street view Leuven data set, producing 70 hand labelled object class and disparity maps. We hope that the release of these annotations will stimulate further work in the challenging domain of street-view analysis.

1 Introduction

The problems of object class segmentation [16, 24], 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 labelled with a disparity [12], 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) [17], which is typically a generalised Potts truncated linear model. Thus both may use graph cut

based move making algorithms, such as α -expansion [3], to solve the labelling problem. These problems should be solved jointly, as a correct labelling of object class can inform depth labelling and stereo reconstruction can also improve object labelling. To provide some intuition behind this statement, note that the object class boundaries are more likely to occur at a sudden transition in depth and vice versa. Moreover, the height of a point above the ground plane is an extremely informative cue regarding its class label, and can be computed from the depth. For example, *road* or *sidewalk* lie in the ground plane, and pixels taking labels *pedestrian* or *car* must lie above the ground plane, while pixels taking label *sky* must occur at an infinite depth from the camera. Figure 1 shows our model which explicitly captures these properties.

Object class recognition yields strong information about 3D structure as shown by the work on photo pop-up [7, 8, 19, 20]. Here a plausible pop-up or planar model of a scene was reconstructed from a single monocular image using only prior information regarding the geometry of typically photographed scenes, and knowledge of where object boundaries are likely to occur.

Beyond this, many tasks require both object class and depth labelling. For an agent to interact with the world, it must be capable of recognising both objects and their physical location. For example, camera based driverless cars must be capable of differentiating between *road* and other classes, and also of recognising where the road ends. Similarly, several companies [6] wish to provide an automatic annotation of assets (such as *street light*, *drain* or *road sign*) to local authorities. In order to provide this service, assets must be identified, localised in 3D space and an estimation of the quality of the assets made.

The use of object labellings to inform scene reconstruction is not new. The aforementioned pop-up method of [7] explicitly used object labels to aid the construction of a scene model, while 3D Layout CRF [9] matched 3D models to object instances. However, in [7] they built a plausible model from the results of object class segmentation, and neither jointly solve the two problems nor attempt to build an accurate 3D reconstruction of the scene whereas in this paper we jointly estimate both. Hoiem *et al.* [9] fit a 3D model not to the entire scene but only to specific objects, and similarly, these 3D models are intended to be plausible rather than accurate.

Leibe *et al.* [18] 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 [5] data set provides sparse *SfM* cues, which were used by several object class segmentation approaches [5, 25] to provide pixel wise labelling. In these works, no dense depth labelling was performed and the object class segmentation was not used to refine the 3D structure.

None of the discussed works perform joint inference to obtain dense stereo reconstruction and object class segmentation. In this work, we demonstrate that the problems are mutually informative, and benefit from being solved jointly. We consider the problem of scene reconstruction in an urban area [18]. 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 [3], and range move algorithms [14, 26].

No real world data sets are publicly available that contain both pixel-wise object class and dense stereo data. In order to evaluate our method, we augmented the data set of [18] by

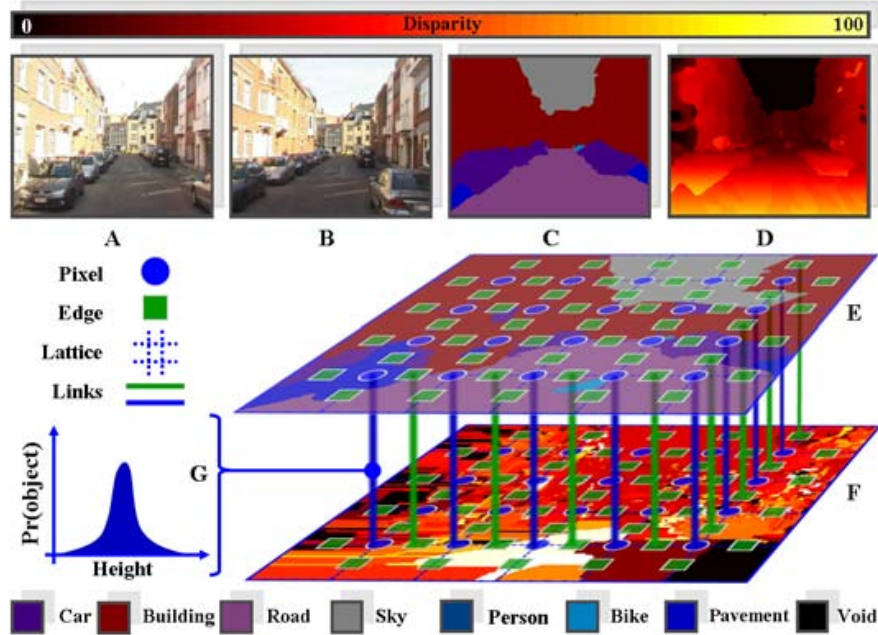


Figure 1: Graphical model of our joint CRF. The system takes a left (A) and right (B) image from a stereo pair that has been rectified. Our formulation captures the co-dependencies between the object class segmentation problem (E, §2.1) and the dense stereo reconstruction problem (F, §2.2) by allowing interactions between them. These interactions are defined to act between the unary/pixel (blue) and pairwise/edge variables (green) of both problems. The unary potentials are linked via a height distribution (G, eq. (3)) learnt from our training set containing hand labelled disparities (§5). The pairwise potentials encode that object class boundaries, and sudden changes in disparity are likely to occur together. The combined optimisation results in an approximate object class segmentation (C) and dense stereo reconstruction (D). See §3 and §4 for a full treatment of our model and §6 for further results. View in colour.

creating hand labelled object class and disparity maps for 70 images. This data set will be released to the public. Our experimental evaluation demonstrates that joint optimisation of dense stereo reconstruction and object class segmentation leads to a substantial improvement in the accuracy of final results.

The structure of the paper is as follows: In section 2 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 3 describes the formulation allowing for the joint optimisation of these two problems, while section 4 shows how the optimisation can be performed efficiently. The data set is described in section 5 and experimental validation follows in 6.

2 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 existing approaches and introduce the notations used in §3. Both problems have previously been defined as a dense CRF where the set of random variables $\mathbf{Z} = \{Z_1, Z_2, \dots, Z_N\}$ corresponds to the set of all image pixels $i \in \mathcal{V} = \{1, 2, \dots, N\}$. 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. Fig. 1 E & F depict this lattice structure as a *blue dotted grid*, the

variables Z_i are shown as *blue circles*.

2.1 Object Class Segmentation using a CRF

We follow [11, 16, 24] in formulating the problem of object class segmentation as finding a minimal cost labelling of a CRF defined over a set of random variables $\mathbf{X} = \{X_1, \dots, X_N\}$ each taking a state from the label space $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$. Each label l_j indicates a different object class such as *car*, *road*, *building* or *sky*. These energies take the form:

$$E^O(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i^O(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^O(x_i, x_j) + \sum_{c \in \mathcal{C}} \psi_c^O(\mathbf{x}_c). \quad (1)$$

The unary potential ψ_i^O of the CRF describes the cost of a single pixel taking a particular label. The pairwise terms ψ_{ij}^O encourage similar neighbouring pixels in the image to take the same label. These potentials are shown in fig. 1 E as *blue circles* and *green squares* respectively. The higher order terms $\psi_c^O(\mathbf{x}_c)$ describe potentials defined over cliques containing more than two pixels. The terms $\psi_i^O(x_i)$ are typically computed from colour, texture and location features of the individual pixels and corresponding prelearned models for each object class [1, 4, 15, 21, 24]. $\psi_{ij}^O(x_i, x_j)$ takes the form of a contrast sensitive Potts model:

$$\psi_{ij}^O(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i, j) & \text{otherwise,} \end{cases} \quad (2)$$

where the function $g(i, j)$ is an edge feature based on the difference in colours of neighbouring pixels [2], typically defined as:

$$g(i, j) = \theta_p + \theta_v \exp(-\theta_\beta \|I_i - I_j\|_2^2), \quad (3)$$

where I_i and I_j are the colour vectors of pixel i and j respectively. θ_p , θ_v , $\theta_\beta \geq 0$ are model parameters learnt using training data. We refer the interested reader to [2, 21, 24] for more details. In our work we follow [16] and use their hierarchical potentials based upon region based features, which significantly improve the results of object class segmentation. Nearly all other CRF based object class segmentation methods can be represented within this formulation via different choices for the higher order cliques, see [16, 22] for details.

2.2 Dense Stereo Reconstruction using a CRF

We use the energy formulation of [3, 12] 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 $\mathbf{Y} = \{Y_1, \dots, Y_N\}$, where each variable Y_i takes a state from the label space $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$ corresponding to a set of disparities, and can be written as:

$$E^D(\mathbf{y}) = \sum_{i \in \mathcal{V}} \psi_i^D(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^D(y_i, y_j). \quad (4)$$

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:

$$\psi^D(y_i, y_j) = f(|y_i - y_j|), \quad (5)$$

where $f(\cdot)$ usually takes the form of 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 fig. 1 F. Note that the disparity for a pixel is directly related to the depth of the corresponding 3D point.

3 Joint Formulation of Object Class Labelling and Stereo Reconstruction

We formulate simultaneous object class segmentation and dense stereo reconstruction as an energy minimisation of a dense labelling \mathbf{z} over the image. Each random variable $Z_i = [X_i, Y_i]$ ¹ 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:

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^J(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^J(\mathbf{z}_c), \quad (6)$$

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, since we use the formulation of $E^D(\mathbf{y})$ §2.2 which does not contain higher order terms ψ_c^D our energy is defined as:

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^J(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^O(\mathbf{x}_c). \quad (7)$$

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 real world data sets like ours described in §5, 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.

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 xz -coordinate may be informative, since we expect that *buildings* will be on both sides, *pavement* will tend to be between *building* and *road* that would take up the central portion of the image. 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 modelling like this requires a substantial amount of data.

In this paper we need to model these interactions with limited data. We do this by restricting our unary interaction potential to the observed fact that certain objects occupy a certain range of real world heights. We are able to obtain the height above the ground plane via the relation: $h(y_i, i) = h_c + (y_h - y_i) \cdot b/d$, 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 modelled by estimating the a priori cost of pixel i taking label $z_i = [x_i, y_i]$ by

$$\psi_i^C([x_i, y_i]) = -\log(H(h(y_i, i)|x_i)), \quad (8)$$

where

$$H(h|l) = \frac{\sum_{i \in \mathcal{T}} \delta(x_i = l) \delta(h(y_i, i) = h)}{\sum_{i \in \mathcal{T}} \delta(x_i = l)} \quad (9)$$

¹ $[X_i, Y_i]$ is the ordered pair of elements X_i and Y_i .

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:

$$\psi_i^J([x_i, y_i]) = w_O^u \psi_i^O(x_i) + w_D^u \psi_i^D(y_i) + w_C^u \psi_i^C(x_i, y_i), \quad (10)$$

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. Fig. 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.

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 disparity of two neighbouring pixels significantly differ. To take this information into account, we chose tractable pairwise potentials of the form:

$$\psi_{ij}^J([x_i, y_i], [x_j, y_j]) = w_O^p \psi_{ij}^O(x_i, x_j) + w_D^p \psi_{ij}^D(y_i, y_j) + w_C^p \psi_{ij}^O(x_i, x_j) \psi_{ij}^D(y_i, y_j), \quad (11)$$

where $w_O^p, w_D^p > 0$ and w_C^p are weights of the pairwise potential. Fig. 1 shows this linkage as *green line* between a pairwise potential (*green box*) of each part.

4 Inference of 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) they have 700 possible states. As each image contains 316×256 random variables, there are $700^{316 \times 256}$ 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 [3] and to higher order potentials [10, 11, 16] and often achieves the best results; while in dense stereo reconstruction, the truncated convex priors (see § 2.2) mean that better solutions are found using range moves [14, 26] 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 a fixed label α . More formally, given a current solution \mathbf{x} the algorithm α -expansion searches through the space \mathbf{X}_α of size 2^N , where N is the number of random variables, to find the optimal solution. Where $\mathbf{X}_\alpha = \{\mathbf{x}' \in \mathcal{L}^N : x'_i = x_i \text{ or } x'_i = \alpha\}$.

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

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 has been performed.

4.1 Projected Moves

Under the assumption that energy $E(\mathbf{z})$ is a metric (as in object class segmentation see § 2.1) or a semi-metric [3] (as in the costs of § 2.2 and § 3) over the label space $\mathcal{L} \times \mathcal{D}$, either

α -expansion or $\alpha\beta$ swap respectively can be used to minimise 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 sub-section we show 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 by 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 (7), 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 [13]; 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.

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}_α of size 2^N . We define \mathbf{Z}_α as:

$$\mathbf{Z}_\alpha = \{\mathbf{z}' \in (\mathcal{L} \times \mathcal{D})^N : z'_i = [x'_i, y_i] \text{ and } (x'_i = x_i \text{ or } x'_i = \alpha)\}. \quad (12)$$

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 (11) under the assumption that \mathbf{y} is fixed, we have:

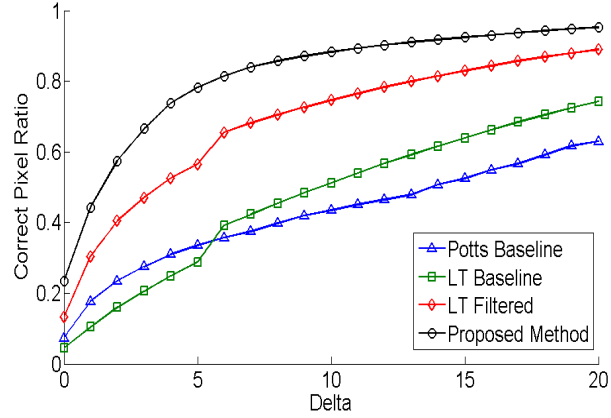
$$\begin{aligned} \psi_{ij}^J([x_i, y_i], [x_j, y_j]) &= (w_O^p + w_C^p \psi_{ij}^D(y_i, y_j)) \psi_{ij}^O(x_i, x_j) + w_D^p \psi_{ij}^D(y_i, y_j) \\ &= \lambda_{ij} \psi_{ij}^O(x_i, x_j) + k_{ij}. \end{aligned} \quad (13)$$

The constant k_{ij} does not affect the choice of optimal move and can safely be ignored. If $\forall y_i, y_j \lambda_{ij} = w_O^p + w_C^p \psi_{ij}^D(y_i, y_j) \geq 0$, the projection of the pairwise potential is a Potts model and standard α -expansion moves can be applied. For $w_O^p \geq 0$ this property holds if $w_O^p + w_C^p k_2 \geq 0$, where k_2 is defined as in §2.2. In practice we use a variant of α -expansion suitable for higher order energies [22].

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 formalise 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:

$$\mathbf{Z}_l = \{\mathbf{z}' \in (\mathcal{L} \times \mathcal{D})^N : z'_i = [x_i, y'_i] \text{ and } (y'_i = y_i \text{ or } y'_i \in [l, l + r])\}. \quad (14)$$

Figure 2: Quantitative comparison of performance of disparity CRFs. We can clearly see that our joint approach §3 (Proposed Method) outperforms the stand alone approaches with baseline Potts [12] (Potts Baseline), Linear truncated potentials §2.2 (LT Baseline) and Linear truncated with Gaussian filtered unary potentials (LT Filtered). The correct pixel ratio is the number 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 §6 for discussion.



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 (11) as:

$$\begin{aligned} \psi_{ij}^I([x_i, y_i], [x_j, y_j]) &= (w_D^p + w_C^p \psi_{ij}^O(x_i, x_j)) \psi_{ij}^D(y_i, y_j) + w_d^O \psi_{ij}^O(x_i, x_j) \\ &= \lambda_{ij} \psi_{ij}^D(y_i, y_j) + k_{ij}. \end{aligned} \quad (15)$$

Again, the constant k_{ij} can safely be ignored, and if $\forall x_i, x_j \lambda_{ij} = w_D^p + w_C^p \psi_{ij}^O(x_i, x_j) \geq 0$ the projection of the pairwise potential is linear truncated and standard range expansion moves can be applied. This property holds if $w_D^p + w_C^p(\theta_p + \theta_v) \geq 0$, where θ_p and θ_v are the weights of the Potts pairwise potential (see section §2.1).

5 Data set

We augment a subset of the Leuven stereo data set² of [18] 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 . A subset of 70 non-consecutive frames was selected for human annotation. The annotation procedure consisted of two parts. Firstly we manually labelled each pixel in every image with one of 7 object classes: *Building*, *Sky*, *Car*, *Road*, *Person*, *Bike* and *Sidewalk*. An 8th label, *void*, is given to pixels that do not obviously belong to one of these classes. Secondly a dense stereo reconstruction was generated by manually creating a disparity map i.e. matching by hand the corresponding pixels between two images. See fig. 3 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 [23] data set, as it contains challenging large regions which

²<http://www.vision.ee.ethz.ch/bleibe/cvpr07/datasets.html>

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 [5] in two important ways, CamVid is a monocular sequence, and the 3D information comes in the form of an unstable³ set of sparse 3D points. 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.

6 Results and Conclusion

For training and evaluation of our method we split the data set (§5) 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 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 number of disparities used for evaluation is 100.

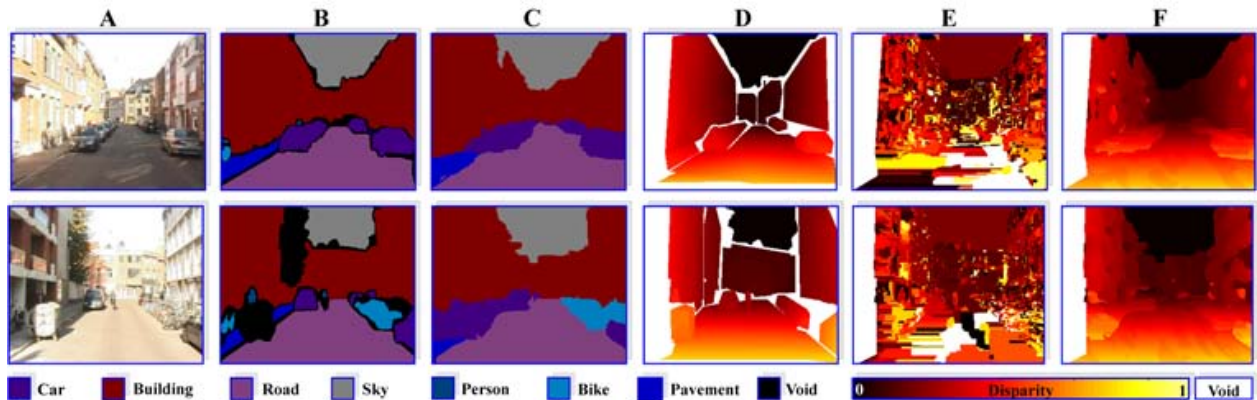


Figure 3: Qualitative object class and disparity results for Leuven data set. (A) *Original Image*. (B) *Object class segmentation ground truth*. (C) *Proposed method Object class segmentation result*. (D) *Dense stereo reconstruction ground truth*. (E) *Stand alone dense stereo reconstruction result (LT Filtered)*. (F) *Proposed method dense stereo reconstruction result*. Best viewed in colour.

Object Class Segmentation The object class segmentation CRF as defined in §2.1 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.

Dense Stereo Reconstruction The Potts [12] and linear truncated §2.2 (LT) baseline dense stereo reconstruction CRFs performed relatively well, with large δ , considering the difficulty of the data, plotted in fig. 2 as ‘Potts baseline’ and ‘LT baseline’. We found that on our data

³The 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 [5] for more details.

set a significant improvement was gained by smoothing the unary potentials with a Gaussian blur⁴ as can be seen in fig. 2 ‘LT Filtered’. For qualitative results see fig. 3 E

Joint Approach Our joint approach defined in sections §3 and §4 consistently outperformed the best stand-alone dense stereo reconstruction, by a margin of up to 25%, as can be seen in fig. 2 ‘Proposed Method’. Improvement of the object class segmentation was incremental, with 95.8% of predicted pixel labels agreeing with the ground truth. The lack of improvement can be attributed to the two mistakes being the misclassification of *person* as *building*, and the top of a uniformly white building as *sky*. Of these failure cases, 3D location is unable to distinguish between *person* and *building*, while stereo reconstruction fails on homogeneous surfaces. 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 fig 3 C and F.

Conclusion In this work, we have presented a novel approach to the problems of object class recognition and dense stereo reconstruction. 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. 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.

References

- [1] A. Blake, C. Rother, M. Brown, P. Perez, and P.H.S. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, 2004.
- [2] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pages I: 105–112, 2001.
- [3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.
- [4] M. Bray, P. Kohli, and P. H. S. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *ECCV*, 2006.
- [5] G.J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (I)*, pages 44–57, 2008.
- [6] Yotta DCL. Yotta dcl case studies. <http://www.yottadcl.com/surveys/case-studies/>, April 2010.
- [7] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.
- [8] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24(3): 577–584, 2005.
- [9] D. Hoiem, C. Rother, and J.M. Winn. 3d layout CRF for multi-view object class recognition and segmentation. In *CVPR*, 2007.

⁴This is a form of robust measure, see §3.1 of [23] for further examples.

- [10] P. Kohli, M. P. Kumar, and P. H. S. Torr. P3 & beyond: Solving energies with higher order cliques. In *CVPR*, 2007.
- [11] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.
- [12] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions via graph cuts. In *ICCV*, pages 508–515, 2001.
- [13] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts?. *PAMI*, 2004.
- [14] M. P. Kumar and P. Torr. Efficiently solving convex relaxations for map estimation. In *ICML*, 2008.
- [15] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *CVPR (I)*, pages 18–25, 2005.
- [16] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, 2009.
- [17] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.
- [18] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007.
- [19] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010.
- [20] S. Ramalingam, P. Kohli, K. Alahari, and P. H. S. Torr. Exact inference in multi-label CRFs with higher order cliques. In *CVPR*, 2008.
- [21] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, pages 309–314, 2004.
- [22] C. Russell, L. Ladicky, P. Kohli, and P. H. S. Torr. Exact and approximate inference in associative hierarchical networks using graph cuts. *UAI*, 2010.
- [23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.
- [24] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. *TexonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV (I)*, pages 1–15, 2006.
- [25] P. Sturgess, K. Alahari, L. Ladicky, and P.H.S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.
- [26] O. Veksler. Graph cut based optimization for mrfs with truncated convex priors. In *CVPR*, 2007.