

Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction

Lubor 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 [10], 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 [9] 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 [8] matched 3D models to object instances. However, in [20] 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 [6, 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 [24, 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

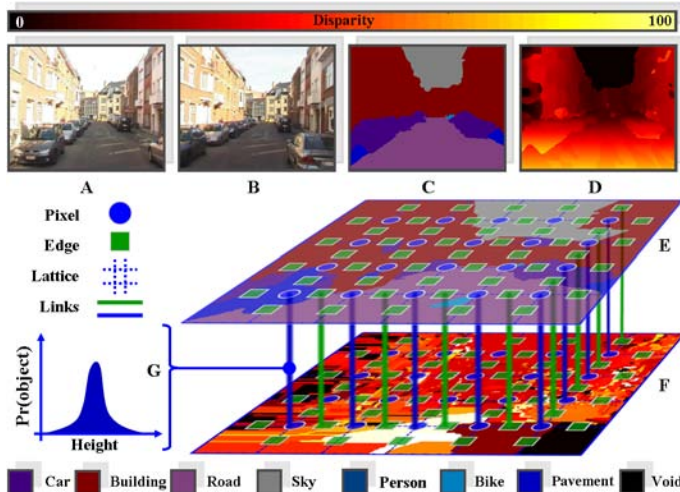


Figure 1: Graphical model of our joint CRF. The system takes a left (A) and right (B) image from a stereo pair that has been rectified. Our formulation captures the co-dependencies between the object class segmentation problem (E, §2.1) and the dense stereo reconstruction problem (F, §2.2) by allowing 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 [10, 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 [10, 16, 19, 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 [10], 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 [10, 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, 21] for details.

2.2 Dense Stereo Reconstruction using a CRF

We use the energy formulation of [8, 10] 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

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

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:

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

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, 16] 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:

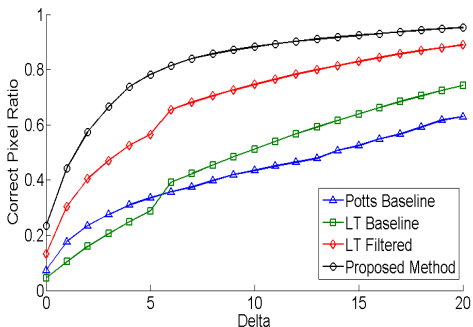
$$\begin{aligned} \Psi_{ij}^J([x_i, y_i], [x_j, y_j]) &= (w_O^p + w_C^p \Psi_{ij}^D(y_i, y_j)) \Psi_{ij}^O(x_i, x_j) + w_D^p \Psi_{ij}^D(y_i, y_j) \\ &= \lambda_{ij} \Psi_{ij}^O(x_i, x_j) + k_{ij}. \end{aligned} \quad (13)$$

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 [14].

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 [18] (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:

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

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 [9] 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.

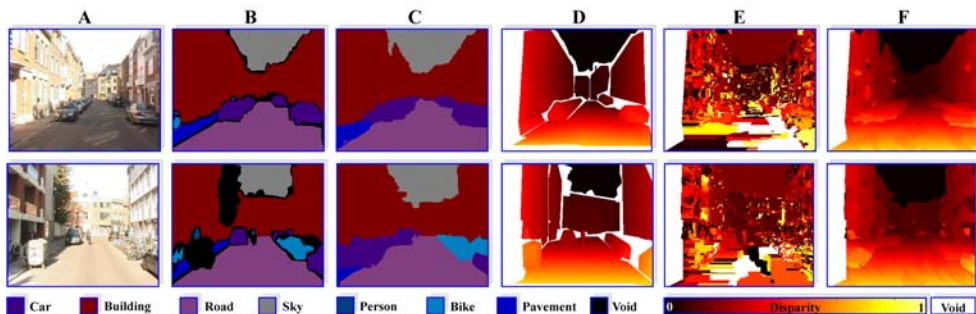


Figure 3: Qualitative object class and disparity results for Leuven data set. (A) *Original Image*. (B) *Object class segmentation ground truth*. (C) *Proposed method Object class segmentation result*. (D) *Dense stereo reconstruction ground truth*. (E) *Stand alone dense stereo reconstruction result (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 [10] 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 [9] 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 del 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 [12] 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 (1)*, 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 (1)*, 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.