Int J Comput Vis (2012) 100:122–133
DOI 10.1007/s11263-011-0489-0

Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction

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

Received: 22 December 2010 / Accepted: 1 August 2011 / Published online: 7 September 2011

Abstract The problems of *dense stereo reconstruction* and *object class segmentation* can both be formulated as Random Field labeling 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 optimize their labelings. In this work we provide a flexible framework configured via cross-validation that unifies the two problems and demonstrate that, by resolving ambiguities, which would be present in real world data if the two problems were considered separately, joint optimization of the two problems substantially improves performance. To evaluate our method, we augment the Leuven data set (http://cms.brookes.ac.uk/research/visiongroup/files/Leuven.zip), which is a stereo video shot from a car driving around the streets of Leuven, with 70 hand labeled 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. Complete source code is publicly available (http://cms.brookes.ac.uk/staff/Philip-Torr/ale.htm).

Keywords Object class segmentation · Dense stereo reconstruction · Random fields

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

P. Sturgess · S. Sengupta · P.H.S. Torr
Oxford Brookes University, Oxford, UK

P. Sturgess
e-mail: paul.sturgess@brookes.ac.uk

S. Sengupta
e-mail: ssengupta@brookes.ac.uk

P.H.S. Torr
e-mail: philiptorr@brookes.ac.uk

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

Y. Bastanlar
Izmir Institute of Technology, Izmir, Turkey
e-mail: yalinbastanlar@iyte.edu.tr

W. Clocksin
University of Hertfordshire, Hatfield, UK
e-mail: wfc@clocksin.com

1 Introduction

Many tasks require both object class and depth labeling. For an agent to interact with the world, it must be capable of recognizing both objects and their physical location. For example, camera based driverless cars must be capable of differentiating between *road* and other classes, recognizing where the road ends. Similarly, several companies (e.g. Yotta 2011) 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, localized in 3D space and an estimation of the quality of the assets made.

The problems of object class segmentation (Shotton et al. 2006; Ladicky et al. 2009), which assigns an object label such as *road* or *building* to every pixel in the image and dense stereo reconstruction, in which every pixel within an image is labeled with a disparity (Scharstein and Szeliski 2002), are well suited for being solved jointly. Both approaches formulate the problem of providing a correct labeling of an image as one of Maximum a Posteriori (MAP) estimation over a Random Field (RF), which is

typically a Potts or truncated linear model. Thus both may use graph cut based move making algorithms, such as α-expansion (Boykov et al. 2001), to solve the labeling problem. These problems *should* be solved jointly, as a correct labeling of object class can help depth labeling, and stereo reconstruction can improve object labeling. Indeed it opens the possibility for the generic stereo priors used previously to be enriched by information about the shape of specific objects. For instance, object class boundaries are more likely to occur at a sudden transition in depth and vice versa, while the height of a point above the ground plane is an extremely informative cue regarding its object class label; e.g. *road* or *sidewalk* lie in the ground plane, and pixels taking labels *pedestrian* or *car* must lie at a constrained height above the ground plane, while pixels taking label *sky* must occur at an infinite depth (zero disparity) from the camera. Figure 1 shows our model which explicitly captures these properties.

Object recognition provides substantial information about the 3D location of points in the image. This has been exploited in recent work on single view reconstruction (Hoiem et al. 2005; Ramalingam et al. 2008; Gould et al. 2009; Liu et al. 2010), in which a plausible pop-up planar model of a scene is reconstructed from a single monocular image using object recognition and prior information regarding the location of objects in typically photographed scenes. Such approaches only estimate depth from object class, assuming the object class is known. As object recognition is itself a problem full of ambiguity and often requiring knowledge of 3D such a two stage process must, in many cases, be suboptimal.

Other works have taken the converse approach of using of 3D information in inferring object class; Hoiem et al. (2006) showed how knowledge of the camera viewpoint and the typical 3D location of objects can be used to improve object detection, while Leibe et al. (2007) 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 labeling of every pixel in the image, and the final results in tracking and detection were not used to refine the *SfM*

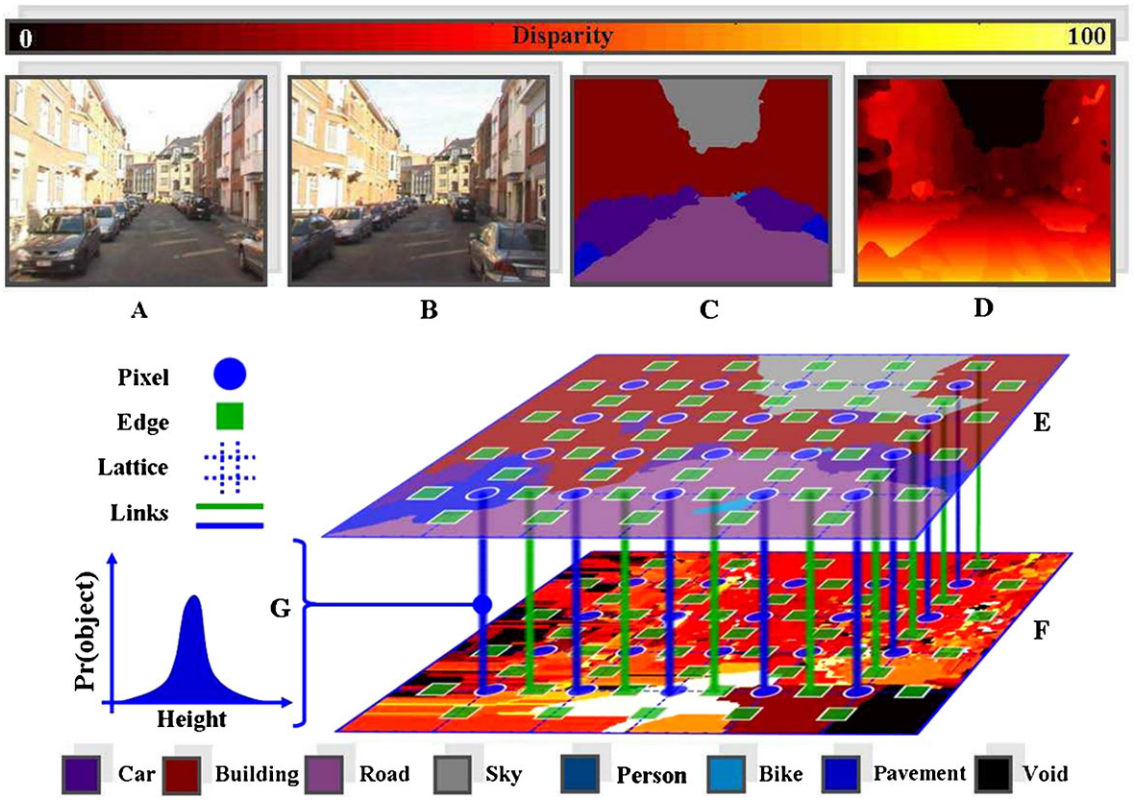

Fig. 1 (Color online) Graphical model of our joint RF. The system takes a left (**A**) and right (**B**) image from a stereo pair that has been rectified. Our formulation captures the dependencies between the object class segmentation problem (**E**, Sect. 2.1) and the dense stereo reconstruction problem (**F**, Sect. 2.2) by defining a joint energy on the recognition and disparity labels both on the unary/pixel (*blue*) and pairwise/edge variables (*green*) of both problems. The unary potentials of the joint problem encodes the fact that different objects will have different height distributions (**G**, (17)) learned from our training set containing hand labeled disparities (Sect. 5). The pairwise potentials encode that object class boundaries, and sudden changes in disparity are likely to occur together, but could also encode different shape smoothness priors for different types of object. The combined optimization results in an approximate object class segmentation (**C**) and dense stereo reconstruction (**D**). See Sects. 3 and 4 for a full treatment of our model and Sect. 6 for further results

results. The CamVid (Brostow et al. 2008) data set provides sparse *SfM* cues, which have been used by several object class segmentation approaches (Brostow et al. 2008; Sturgess et al. 2009) to generate pixel based image labeling. In these the object class segmentation was not used to refine the 3D structure.

Previous works have attempted to simultaneously solve the problems of object class detection and 3D reconstruction. Hoiem et al. (2007) fitted a 3D model to specific objects, such as *buses* or *cars* within an image by simultaneously estimating 3D location, orientation and object class, while Dick et al. (2004) fitted a 3D model of a building to a set of images by simultaneously estimating a wire-frame model and the location of assets such as *window* or *column*. In both of these papers the 3D models are intended to be plausible rather than accurate, and these models are incomplete—they do not provide location or class estimates of every pixel.

None of the discussed works perform joint inference to obtain dense stereo reconstruction and object class segmentation. In this work, we demonstrate that these problems are mutually informative, and benefit from being solved jointly. We consider the problem of scene reconstruction in an urban area (Leibe et al. 2007). 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 (Boykov et al. 2001), and range move algorithms (Kumar et al. 2011).

No real world data sets are publicly available that contain both per pixel object class and dense stereo data. In order to evaluate our method, we augmented the data set of Leibe et al. (2007) by creating hand labeled object class and disparity maps for 70 images. These annotations have been made available for download. Our experimental evaluation demonstrates that joint optimization 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 Sect. 2 we give the generic formulation of RFs for dense image labeling, 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 optimization of these two problems, while Sect. 4 shows how the optimization can be performed efficiently. The data set is described in Sect. 5 and experimental validation follows in Sect. 6.

2 Overview of the Random Field Formulations

Our joint optimization consists of two parts, object class segmentation and dense stereo reconstruction. Before we formulate our approach we give an overview of the typically used random field (RF) formulation for both problems and introduce the notation used in Sect. 3. Both problems have previously been defined as a dense RF where the set of random variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ corresponds to the set of all image pixels $i \in \mathcal{V} = \{1, 2, \ldots, N\}$. 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 *labeling* and denoted by $\mathbf{z}$, similarly we use $\mathbf{z}_c$ to denote the labeling of a clique. Each $z_i \in \mathcal{L}$, where $\mathcal{L}$ is the set of labels. Figure 1E and F depict this lattice structure as a *blue dotted grid*, the variables Z_i are shown as *blue circles*.

Random field formulation can be seen as a structured classifier minimizing the cost of the labeling $\mathbf{z}$:

$$z^* = \arg\min_{\mathbf{z}} E(\mathbf{z}) = \arg\min_{\mathbf{z}} \sum_{c \in \mathcal{C}} \psi_c(\mathbf{z}_c), \tag{1}$$

where $\mathcal{C}$ is the set of all cliques. The term $\psi_c(\mathbf{z}_c)$ is known as the potential function of the clique $c \subset \mathcal{V}$ where $\mathbf{z}_c = \{z_i : i \in c\}$. Potential functions typically take the form:

$$\psi_c(\mathbf{z}_c) = w_c \cdot \Phi_c(\mathbf{z}_c), \tag{2}$$

where $\Phi_c(\mathbf{z}_c)$ is a cost function vector containing a cost for each configuration of $\mathbf{z}_c$ and w_c a weight vector weighting the importance of each cost function. The weight vectors are constant for all cliques for each type of potential (unary, pairwise, ...). Even though there exists an underlying probabilistic distribution corresponding to any RF, the state-of-the-art algorithms for learning the potential functions $\psi_c(\cdot)$ are typically trained discriminatively and the final classifier does not have any real probabilistic interpretation. Thus, all the weights and parameters are either hand-tuned on the validation set or trained using any discriminative max-margin method (Taskar et al. 2004; Tsochantaridis et al. 2005; Alahari et al. 2010). Probabilistic interpretations whilst theoretically well grounded are hard to achieve in practise, as the probabilistic distributions are exceptionally difficult to model.

2.1 Object Class Segmentation Using a RF

We follow (Shotton et al. 2006; Kohli et al. 2008; Ladicky et al. 2009) in formulating the problem of object class segmentation as finding a minimal cost labeling of a RF defined over a set of random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$ each taking a state from the label space $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$. Each label l_j indicates a different object class such as *car*, *road*, *building*

or *sky*. These energies take the form:

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

The unary potential ψ_i^O of the RF describes the cost of a single pixel taking a particular label. We followed the approach in Ladicky et al. (2009), Sturgess et al. (2009), where the unary cost for a pixel taking certain label is based on the boosted (Torralba et al. 2004) classifier based on shape filters (Shotton et al. 2006) and multiple feature responses (Ladicky et al. 2009). We refer the reader to Shotton et al. (2006), Ladicky et al. (2009) for more details. The pairwise terms ψ_{ij}^O encourage similar neighboring pixels in the image to take the same label. These potentials are shown in Fig. 1E as *blue circles* and *green squares* respectively. $\psi_{ij}^O(x_i, x_j)$ takes the form of a contrast sensitive Potts model:

$$\psi_{ij}^O(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i,j) & \text{otherwise,} \end{cases} \quad (4)$$

where the function $g(i, j)$ is an edge feature based on the difference in colors of neighboring pixels (Boykov and Jolly 2001), typically defined as:

$$g(i,j) = \theta_p + \theta_v \exp(-\theta_\beta \|I_i - I_j\|_2^2), \quad (5)$$

where I_i and I_j are the color vectors of pixel i and j respectively. θ_p, θ_v, $\theta_\beta \geq 0$ are model parameters learned using training data. We refer the interested reader to Boykov and Jolly (2001), Rother et al. (2004), Shotton et al. (2006) for more details.

The higher order terms $\psi_c^O(\mathbf{x}_c)$ describe potentials defined over cliques containing more than two pixels. In our work we follow (Ladicky et al. 2009) and use their hierarchical potentials based upon histograms of features, evaluated on segments, obtained by unsupervised segmentation methods (Comaniciu and Meer 2002; Shi and Malik 2000). This significantly improves the results of object class segmentation method. Nearly all current RF based object class segmentation methods (Rabinovich et al. 2007; Batra et al. 2008) can be represented within this formulation via different choices for the higher order cliques (Ladicky et al. 2009; Russell et al. 2010) and can be included in the framework.

2.2 Dense Stereo Reconstruction Using a RF

We use the energy formulation of Boykov et al. (2001), Scharstein and Szeliski (2002) for the dense stereo reconstruction Sect. 2.2 part of our joint formulation. They formulated the problem as one of finding a minimal cost labeling of a RF defined over a set of random variables $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$, where each variable Y_i takes a state from the label space $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ corresponding to a set of disparities, and can be written as:

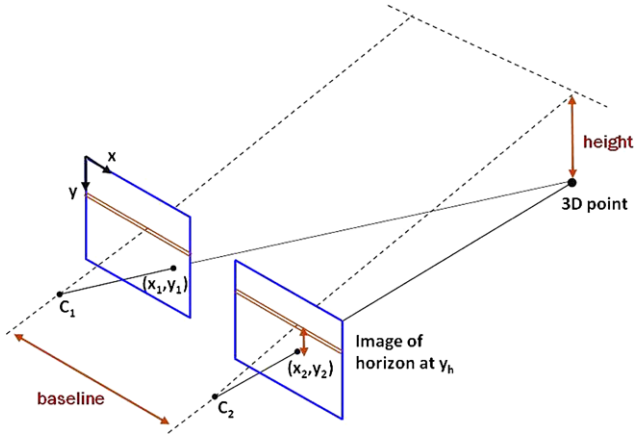

Fig. 2 An illustration of how 3D information can be reconstructed from a stereo camera rig. Also shown, the relation between disparity (the movement of a point between the pair of images) and height, once ground plane is known

$$E^D(\mathbf{y}) = \sum_{i\in\mathcal{V}} \psi_i^D(y_i) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_i} \psi_{ij}^D(y_i, y_j). \quad (6)$$

The unary potential $\psi_i^D(y_i)$ of the RF is defined as a measure of color 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 neighboring pixels in the image to have a similar disparity. The cost is a function of the distance between disparity labels:

$$\psi^D(y_i, y_j) = f(|y_i - y_j|), \quad (7)$$

where $f(.)$ usually takes the form of a linear truncated function $f(y) = \min(k_1 y, k_2)$, where $k_1, k_2 \geq 0$ are the slope and truncation respectively. The unary (*blue circles*) and pairwise (*green squares*) potentials are shown in Fig. 1F. Note that the disparity for a pixel is directly related to the depth of the corresponding 3D point (see Fig. 2). To partially resolve ambiguities in disparities for low textured objects a Gaussian filter is applied to the unary potentials.

2.3 Monocular Video Reconstruction

With minor modification, the formulation of Sect. 2.2 can also be applied to monocular video sequences, by performing stereo reconstruction over adjacent frames in the video sequence (see Fig. 3). Under the simplifying assumption that the scene remains static, the formulation remains the same. However, without a fixed baseline between the camera positions in adjacent frames the estimation of disparities, and the mapping of disparities to depths is more complex.

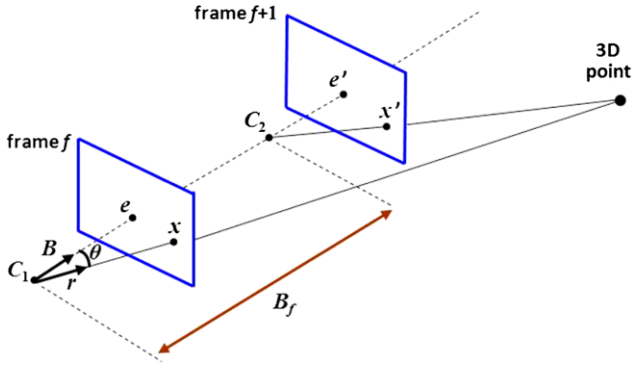

Fig. 3 An illustration of how 3D information can be reconstructed from the monocular sequence. Details of the conversion of the monocular 3D reconstruction problem into the standard stereo reconstruction are given in Sect. 2.3

We first pre-process the data, by performing SIFT matching (Lowe 2004) over adjacent frames, before using RANSAC (Fischler and Bolles 1981; Torr and Murray 1997) to simultaneously estimate the fundamental matrix, and a corresponding set of inliers from these matches. The fundamental matrix gives us both the epipoles[1] and the epipolar lines, and this allows us to solve the stereo correspondence efficiently by searching along corresponding epipolar lines for a match. Given two images 1, and 2, we write x, x' for a pair of matched points in images 1 and 2 respectively, and use e, e' for the epipoles present in each image. The disparity d is estimated as:

$$d = \big||e - x| - |e' - x'|\big|. \tag{8}$$

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

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

– Firstly, the distance traveled between frames by the camera varies with the speed of the vehicle and this implies that the baseline varies from frame to frame.
– Secondly, when the epipole lies in the image the camera can not be approximated as orthographic. The effective baseline, which we define as the component of the baseline normal to the ray, varies substantially within an image from pixel to pixel.

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

We define a ray λr, as the set of all values taken by a 3D unit vector r, multiplied by a scalar $\lambda \in \Re$. We define the baseline B_f as the 3D distance traveled by the camera between a pair of frames f and $f+1$.[2] We let θ be the angle between B and r. Then we define e the epipole, as the intersection point of the baseline and the image plane, and x as the point in the image that the ray λr passes through. Given a disparity d of a point on the ray, the distance s of that point from the camera is:

$$\begin{aligned} s &= K|(B_f - B_f \cdot r)|/d \\ &= K|B_f|\sqrt{1-\cos^2\theta}/d \\ &= K|B_f| \times |\sin\theta|/d, \end{aligned} \tag{9}$$

where K is a constant based on the internal properties of the camera.

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

$$s = K|B_f|\sqrt{\frac{\gamma^2(e-x)^2}{1+\gamma^2(e-x)^2}}\Big/d. \tag{10}$$

Solving s for a conventional stereo pair gives the related equation

$$s = K|B'|/d', \tag{11}$$

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

$$d' = \frac{|B'|}{|B_f|}\frac{d}{\sqrt{\frac{\gamma^2(e-x)^2}{1+\gamma^2(e-x)^2}}}. \tag{12}$$

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

[2] This value is a part of the standard Leuven data-set, see Sect. 5, and does not require estimating, in our application, see Sect. 6.

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

$$d' \approx \frac{|B'|\, d}{|B_f|\, \gamma\, |e - x|}. \tag{13}$$

Given this relationship, unary potentials defined over the monocular disparity d, can be mapped to unary potentials over the conventional stereo disparity d'. This allows standard stereo reconstruction on monocular sequences to be performed as in Sect. 2.2, and joint object class and 3D reconstruction from monocular sequences to be performed as described in the following section.

3 Joint Formulation of Object Class Labeling and Stereo Reconstruction

We formulate simultaneous object class segmentation and dense stereo reconstruction as an energy minimization of a dense labeling $\mathbf{z}$ over the image. Each random variable $Z_i = [X_i, Y_i]$[3] 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 RF for joint estimation can be written as:

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^J(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^J(\mathbf{z}_c), \tag{14}$$

where the terms ψ_i^J, ψ_{ij}^J and ψ_c^J are a sum of the previously mentioned terms ψ_i^O and ψ_i^D, ψ_{ij}^O and ψ_{ij}^D, and ψ_c^O and ψ_c^D respectively, plus some terms ψ_i^C, ψ_{ij}^C, ψ_c^C, which govern interactions between $\mathbf{X}$ and $\mathbf{Y}$. However, in our case $E^D(\mathbf{y})$ (see Sect. 2.2) does not contain higher order terms ψ_c^D, and the joint energy is defined as:

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^J(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^O(\mathbf{x}_c). \tag{15}$$

If the interaction terms ψ_i^C, ψ_{ij}^C are both zero, then the problems $\mathbf{x}$ and $\mathbf{y}$ are independent of one another and the energy would be decomposable into $E(\mathbf{z}) = E^O(\mathbf{x}) + E^D(\mathbf{y})$ and the two sub-problems could each be solved separately. However, in many real world data sets such as the one we describe in Sect. 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.

3.1 Joint Unary Potentials

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

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

$$h(y_i, i) = h_c + \frac{(y_h - y_i)b}{d}, \tag{16}$$

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

$$\psi_i^C([x_i, y_i]) = -\log(H(h(y_i, i)|x_i)), \tag{17}$$

where

$$H(h|l) = \frac{\sum_{i \in \mathcal{T}} \delta(x_i = l)\delta(h(y_i, i) = h)}{\sum_{i \in \mathcal{T}} \delta(x_i = l)} \tag{18}$$

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 RF is:

$$\psi_i^J([x_i, y_i]) = w_O^u \psi_i^O(x_i) + w_D^u \psi_i^D(y_i) + w_C^u \psi_i^C(x_i, y_i), \tag{19}$$

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

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

given the image. w_O^u, w_D^u, and w_C^u are weights. Fig. 1G 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.

3.2 Joint Pairwise Interactions

Pairwise potentials enforce the local consistency of object class and disparity labels between neighboring 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 neighboring pixels significantly differ. To take this information into account, we chose tractable pairwise potentials of the form:

$$\psi_{ij}^J([x_i, y_i], [x_j, y_j]) = w_O^p \psi_{ij}^O(x_i, x_j) + w_D^p \psi_{ij}^D(y_i, y_j) + w_C^p \psi_{ij}^O(x_i, x_j)\psi_{ij}^D(y_i, y_j), \quad (20)$$

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

4 Inference for the Joint RF

Optimization of the energy $E(\mathbf{z})$ is challenging. Each random variable takes a label from the set $\mathcal{L} \times D$ consequentially, in the experiments we consider (see Sect. 5) they have 700 possible states. As each image contains 316×256 random variables, there are $700^{316\times256}$ possible solutions to consider. Rather than attempting to solve this problem exactly, we use graph cut based move making algorithms to find an approximate solution.

Graph cut based move making algorithms start from an initial solution and proceed by making a series of moves or changes, each of which leads to a solution of lower energy. The algorithm is said to converge when no lower energy solution can be found. In the problem of object class labeling, the move making algorithm α-expansion can be applied to pairwise (Boykov et al. 2001) and to higher order potentials (Kohli et al. 2007, 2008; Ladicky et al. 2009) and often achieves the best results; while in dense stereo reconstruction, the truncated convex priors (see Sect. 2.2) mean that better solutions are found using range moves (Kumar et al. 2011) than with α-expansion.

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

$$\mathbf{X}_\alpha = \left\{\mathbf{x}' \in \mathcal{L}^N : x_i' = x_i \text{ or } x_i' = \alpha\right\}. \quad (21)$$

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

$$\mathbf{Y}_l = \left\{\mathbf{y}' \in \mathcal{D}^N : y_i' = y_i \text{ or } y_i' \in [l, l + r]\right\}. \quad (22)$$

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 Sect. 2.1) or a semi-metric (Boykov et al. 2001) (as in the costs of Sects. 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 minimize the energy. One single iteration of α-expansion would require $O(|\mathcal{L}||\mathcal{D}|)$ graph cuts to be computed, while $\alpha\beta$ swap requires $O(|\mathcal{L}|^2|\mathcal{D}|^2)$ resulting in slow convergence. In this 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 optimization that improves by turn first object class labeling 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 optimization in which each component is individually optimized. 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 (15), 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 optimized 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 sub-modular energy that can be efficiently solved using graph cuts (Boykov and Kolmogorov 2004); 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.

4.2 Expansion Moves in the Object Class Label Space

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

$$\mathbf{Z}_\alpha = \left\{ \begin{array}{l} \mathbf{z}' \in (\mathcal{L} \times D)^N : z'_i = [x'_i, y_i] \text{ and} \\ (x'_i = x_i \text{ or } x'_i = \alpha) \end{array} \right\}. \tag{23}$$

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

$$\begin{aligned} &\psi^J_{ij}([x_i, y_i], [x_j, y_j]) \\ &\quad = \big(w^p_O + w^p_C \psi^D_{ij}(y_i, y_j)\big)\psi^O_{ij}(x_i, x_j) \\ &\qquad + w^p_D \psi^D_{ij}(y_i, y_j) \\ &\quad = \lambda_{ij}\psi^O_{ij}(x_i, x_j) + k_{ij}. \end{aligned} \tag{24}$$

The constant k_{ij} does not affect the choice of optimal move and can safely be ignored. If $\forall y_i, y_j \ \lambda_{ij} = w^p_O + w^p_C \psi^D_{ij}(y_i, y_j) \geq 0$, the projection of the pairwise potential is a Potts model and standard α-expansion moves can be applied. For $w^p_O \geq 0$ this property holds if $w^p_O + w^p_C k_2 \geq 0$, where k_2 is defined as in Sect. 2.2. In practice we use a variant of α-expansion suitable for higher order energies (Russell et al. 2010).

4.3 Range Moves in the Disparity Label Space

For our joint optimization of disparity and object classes we propose a new move in the project disparity label space. Each pixel taking label $z_i = (x_i, y_i)$ can either keep its current label or take a new label from the range $(x_i, [l_a, l_b])$. To formalize this, given a current solution $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$ the algorithm searches through the space $\mathbf{Z}_l$ of size $(2 + r)^N$, which we define as:

$$\mathbf{Z}_l = \left\{ \begin{array}{l} \mathbf{z}' \in (\mathcal{L} \times D)^N : z'_i = [x_i, y'_i] \text{ and} \\ (y'_i = y_i \text{ or } y'_i \in [l, l + r]) \end{array} \right\}. \tag{25}$$

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 (20) as:

$$\begin{aligned} &\psi^J_{ij}([x_i, y_i], [x_j, y_j]) \\ &\quad = \big(w^p_D + w^p_C \psi^O_{ij}(x_i, x_j)\big)\psi^D_{ij}(y_i, y_j) \\ &\qquad + w^O_d \psi^O_{ij}(x_i, x_j) \\ &\quad = \lambda_{ij}\psi^D_{ij}(y_i, y_j) + k_{ij}. \end{aligned} \tag{26}$$

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

5 Data Set

We augment a subset of the Leuven stereo data set[4] of Leibe et al. (2007) with object class segmentation and disparity annotations. The Leuven data set was chosen as it provides image pairs from two cameras, 150 cm apart from each other, mounted on top of a moving vehicle, in a public urban setting. In comparison with other data sets, the larger distance between the two cameras allows better depth resolution, while the real world nature of the data set allows us to confirm our statistical model's validity. However, the data set does not contain the object class or disparity annotations, we require to learn and quantitatively evaluate the effectiveness of our approach.

To augment the data set all image pairs were rectified, and cropped to 316×256, then the subset of 70 non-consecutive frames was selected for human annotation. The annotation procedure consisted of two parts. Firstly we manually labeled each pixel in every image with one of 7 object classes: *Building, Sky, Car, Road, Person, Bike* and *Sidewalk*. An 8th label, *Void*, is given to pixels that do not obviously belong to one of these classes. Secondly disparity maps were generated by manually matching by hand the corresponding planar polygons, some examples of which are shown in the Fig. 4A, 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 (Scharstein and Szeliski 2002) data set, as it contains challenging large regions which are homogeneous in color 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 (Brostow et al. 2008) in two important ways, CamVid is a monocular sequence, and the 3D information comes in the form of a set of sparse 3D

[4] http://www.vision.ee.ethz.ch/ bleibe/cvpr07/datasets.html.

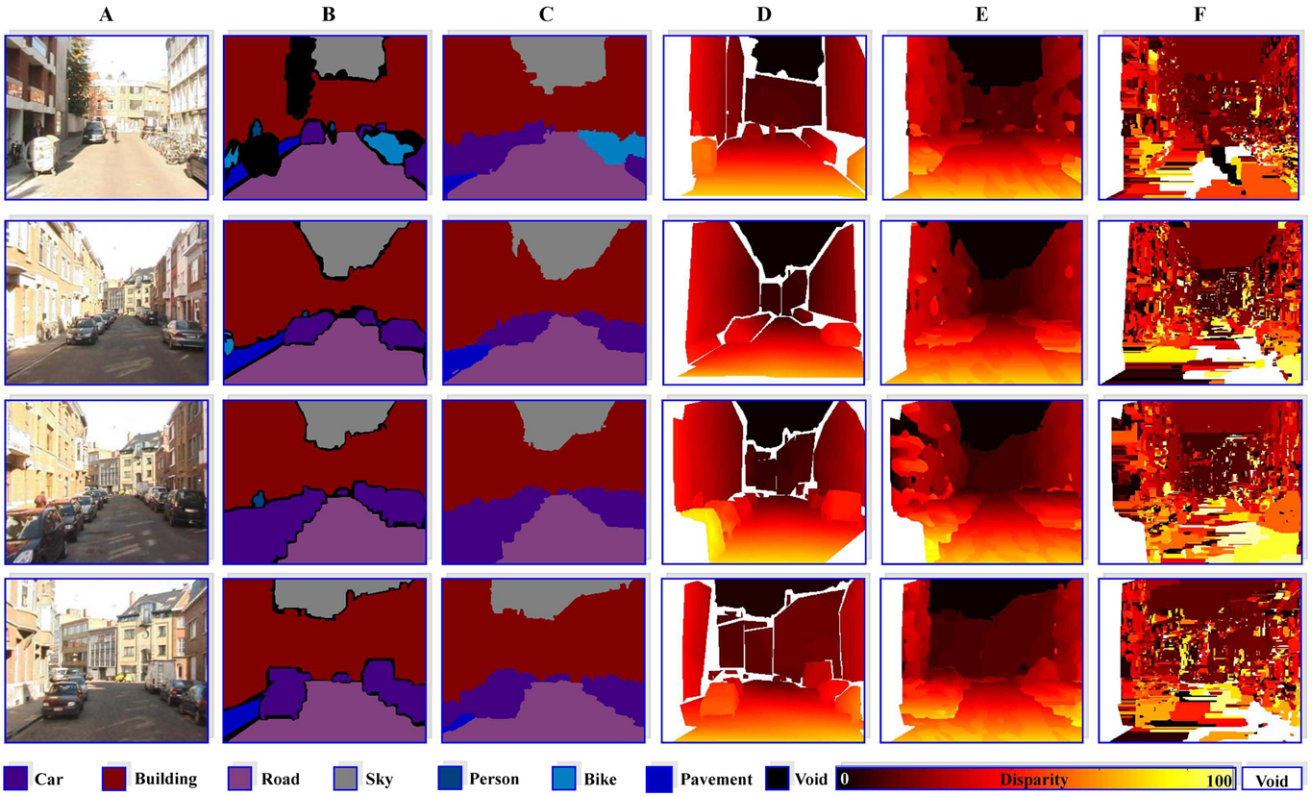

Fig. 4 (Color online) Qualitative object class and disparity results for Leuven data set. (**A**) Original Image. (**B**) Object class segmentation ground truth. (**C**) Proposed method Object class segmentation result. (**D**) Dense stereo reconstruction ground truth. (**E**) Proposed method dense stereo reconstruction result. (**F**) Stand alone dense stereo reconstruction result (LT)

points with outliers.[5] 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 Experiments

For training and evaluation of our method we split the data set (Sect. 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. Due to insufficient size of the data the class *Person* is also set to void and the parameters for object class domain were chosen the same as in Ladicky et al. (2009). The depth domain and joint parameters were learnt on the training set same as in Ladicky et al. (2009). The performance on the training set in the depth domain is not significantly better than on the test set and this approach does not lead to an over-fitting of the parameters.

[5]The outlier rejection step was not performed on the 3D point cloud in order to exploit large re-projection errors as cues for moving objects. See Brostow et al. (2008) for more details.

We quantitatively evaluate the object class segmentation by measuring the percentage of correctly predicted labels over non *void* pixels in the test sequence. The dense stereo reconstruction performance is quantified by measuring the number of pixels which satisfy $|d_i - d_i^g| \leq \delta$, where d_i is the label of ith 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.

6.1 Object Class Segmentation

The object class segmentation RF as defined in Sect. 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. Quantitative comparison of the stand alone and joint method is given in Table 1.

6.2 Dense Stereo Reconstruction

The Potts (Kolmogorov and Zabih 2001) and linear truncated (LT) baseline dense stereo reconstruction models de-

Table 1 Quantitative results for object class segmentation of stand alone and joint approach. The pixel accuracy (%) for different object classes. The 'global' measure corresponds to the total proportion of pixels labeled correctly. Per class accuracy corresponds to recall measure commonly used for this task (Shotton et al. 2006; Sturgess et al. 2009; Ladicky et al. 2009). Minor improvement were achieved for smaller classes that had fewer pixels present in the data set. We assume the difference would be larger for harder data sets

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

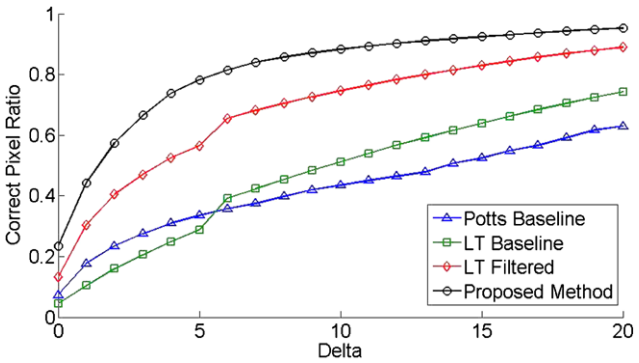

Fig. 5 Quantitative comparison of the performance of disparity RFs. We can clearly see that our joint approach Sect. 3 (Proposed Method) outperforms standard dense stereo approaches based on the Potts (Kolmogorov and Zabih 2001) (Potts Baseline), Linear truncated models described in Sect. 2.2 (LT Baseline) and Linear truncated with Gaussian filtered unary potentials (LT Filtered). The correct pixel ratio is the proportion of pixels which satisfy $|d_i - d_i^g| \leq \delta$, where d_i is the disparity label of ith pixel, d_i^g is corresponding ground truth label and δ is the allowed error. See Sect. 6 for discussion

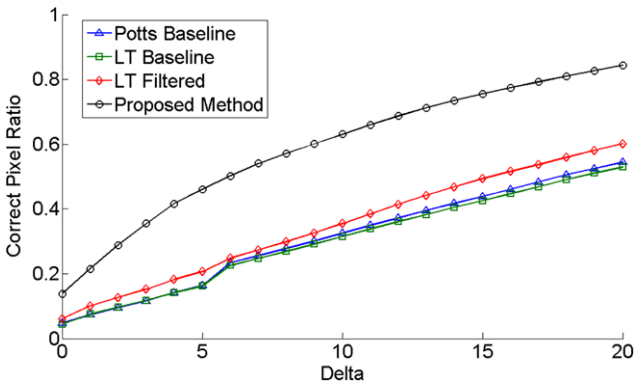

Fig. 6 Quantitative comparison of the performance of disparity RFs, on monocular sequences. As with the stereo pair, we can clearly see that our joint approach Sect. 3 (Proposed Method) outperforms the stand alone approaches with baseline Potts (Kolmogorov and Zabih 2001) (Potts Baseline), Linear truncated potentials Sect. 2.2 (LT Baseline) and Linear truncated with Gaussian filtered unary potentials (LT Filtered). The correct pixel ratio is the proportion of pixels which satisfy $|d_i - d_i^g| \leq \delta$, where d_i is the disparity label of ith pixel, d_i^g is corresponding ground truth label and δ is the allowed error. See Sect. 6.4 for discussion, and Fig. 5 to compare against conventional stereo

scribed in Sect. 2.2 performed relatively well, with large δ, considering the difficulty of the data, plotted in Fig. 5 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[6] before incorporating the potential in the RF framework with linear truncated model, as can be seen in Fig. 5 'LT Filtered'. For qualitative results see Fig. 4E.

6.3 Joint Approach

Our joint approach defined in Sects. 3 and 4 consistently outperformed the best stand-alone dense stereo reconstruction as can be seen in Fig. 5. Improvement of the object class segmentation was less dramatic, with 95.8% of predicted pixel labels agreeing with the ground truth. We expect to see a more significant improvement on more challenging data sets, and the creation of an improved data set is part of our future work. Qualitative results can be seen in Fig. 4C and E.

[6]This is a form of robust matching measure, see Sect. 3.1 of (Scharstein and Szeliski 2002) for further examples.

6.4 Monocular Reconstruction

Reconstruction from a monocular sequence is substantially harder than the corresponding stereo problem. Not only does it suffer from the same problems of varying illumination and homogeneous regions, but the effective base-line is substantially shorter making it much harder to recover 3D information with any degree of accuracy, particularly in the region around the epipole (see Sect. 2.3 and Fig. 7). Despite this, plausible 3D reconstruction is still possible, particularly when performing joint inference over object class and disparity simultaneously, quantitative results can be seen in Fig. 6. Note that the joint optimization of monocular disparity and object class out performs the pre-existing methods (*LT Baseline* and *Potts Baseline*) over conventional two camera stereo data, and is comparable to the two camera results on *LT filtered*. In Fig. 7 qualitative results can be seen. As expected, these show the quality of reconstruction improves with the distance from the epipole. Consequentially, one of the regions most successfully reconstructed is marked as *void* in the two camera disparity maps, as it is not in the

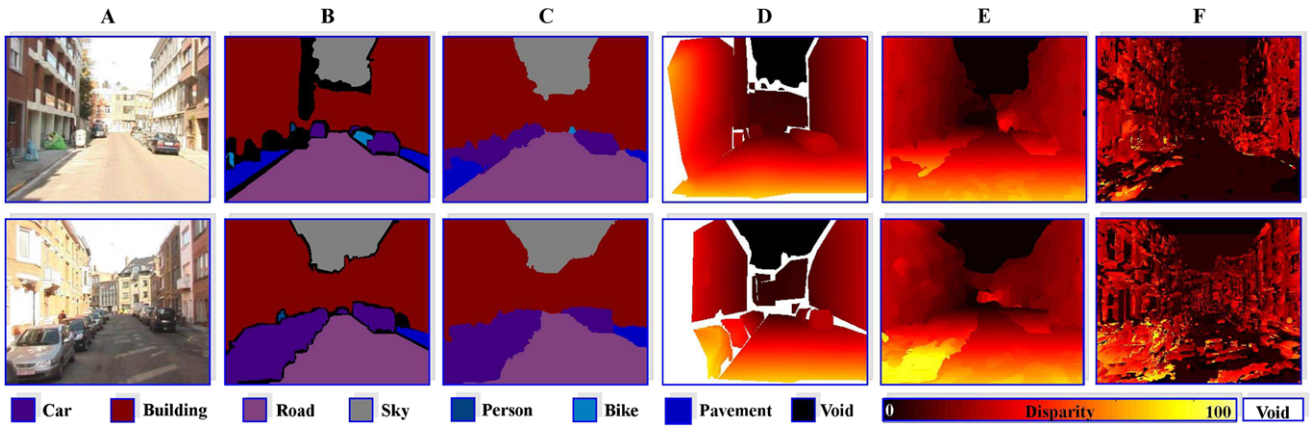

Fig. 7 (Color online) Monocular results. (**A**) Original Image. (**B**) Object class segmentation ground truth. (**C**) Proposed method Object class segmentation result. (**D**) Dense stereo reconstruction ground truth. (**E**) Proposed method dense stereo reconstruction result. (**F**) Stand alone dense stereo reconstruction result (LT). The quality of reconstruction improves with the distance from the epipole

field of view of both cameras. This suggests that the numeric evaluation of Fig. 6 may be overly pessimistic.

7 Conclusion

Traditionally the prior in stereo has been fixed to some standard tractable model such as truncated linear on disparities. Within this work we open up the intriguing possibility that the prior on shape should take in account the type of scene and object we are looking at. To do this, we provided a new formulation of the problems, a new inference method for solving this formulation and a new data set for the evaluation of our work. Evaluation of our work shows a dramatic improvement in stereo reconstruction compared to existing approaches. We assume statistically significant gain can be achieved also for object class segmentation, but it would require more challenging data set. This paper has proposed a formulation in which distributions of height maps for each object class in road scenes are used, one might also easily extend this idea to the unsupervised case, with an online learning, and this extension is investigated in Bleyer et al. (2011). The method can be generalized to any other scenes where mutual information between 3D location and object label is present and can be learnt using discriminative methods. Furthermore, it allows the incorporation of other cues commonly used in RFs such as object-class dependent pairwise potentials (Batra et al. 2008) or incorporation of occlusions (Kolmogorov and Zabih 2001) or 2nd order smoothness priors Woodford et al. (2008) in the depth domain. This work puts us one step closer to achieving complete scene understanding, and provides strong experimental evidence that the joint labeling of different problems can bring substantial gains.

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. Chris Russell was partially funded by the European Research Council under the ERC Starting Grant agreement 204871-HUMANIS.

References

Alahari, K., Russell, C., & Torr, P. H. S. (2010). Efficient piecewise learning for conditional random fields. In *Conference on computer vision and pattern recognition*.

Batra, D., Sukthankar, R., & Tsuhan, C. (2008). Learning class-specific affinities for image labelling. In *Conference on computer vision and pattern recognition*.

Bleyer, M., Rother, C., Kohli, P., Scharstein, D., & Sinha, S. (2011). Object stereo—joint stereo matching and object segmentation. In *Conference on computer vision and pattern recognition*.

Boykov, Y., & Jolly, M. (2001). Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *International conference on computer vision*.

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Transactions on Pattern Analysis and Machine Intelligence*.

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

Brostow, G. J., Shotton, J., Fauqueur, J., & Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*.

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence*.

Dick, A. R., Torr, P. H. S., & Cipolla, R. (2004). Modelling and interpretation of architecture from several images. *International Journal of Computer Vision*.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*.

Gould, S., Fulton, R., & Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In *International conference on computer vision*.

Hoiem, D., Efros, A., & Hebert, M. (2005) Automatic photo pop-up. *ACM Transactions on Graphics*.

Hoiem, D., Efros, A., & Hebert, M. (2006). Putting objects in perspective. In *Conference on computer vision and pattern recognition*.

Hoiem, D., Rother, C., & Winn, J. M. (2007). 3D layout CRF for multi-view object class recognition and segmentation. In *Conference on computer vision and pattern recognition*.

Kohli, P., Kumar, M., & Torr, P. H. S. (2007). P^3 and beyond: solving energies with higher order cliques. In *Conference on computer vision and pattern recognition*.

Kohli, P., Ladicky, L., & Torr, P. H. S. (2008). Robust higher order potentials for enforcing label consistency. In *Conference on computer vision and pattern recognition*.

Kolmogorov, V., & Zabih, R. (2001). Computing visual correspondence with occlusions via graph cuts. In *ICCV*.

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

Ladicky, L., Russell, C., Kohli, P., & Torr, P. H. S. (2009). Associative hierarchical CRFs for object class image segmentation. In *International conference on computer vision*.

Leibe, B., Cornelis, N., Cornelis, K., & Gool, L. V. (2007). Dynamic 3D scene analysis from a moving vehicle. In *Conference on computer vision and pattern recognition*.

Liu, B., Gould, S., & Koller, D. (2010). Single image depth estimation from predicted semantic labels. In *Conference on computer vision and pattern recognition*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *International conference on computer vision*.

Ramalingam, S., Kohli, P., Alahari, K., & Torr, P. H. S. (2008). Exact inference in multi-label CRFs with higher order cliques. In *Conference on computer vision and pattern recognition*.

Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut: interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*.

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

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2006). TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European conference on computer vision*.

Sturgess, P., Alahari, K., Ladicky, L., & Torr, P. H. S. (2009). Combining appearance and structure from motion features for road scene understanding. In *British machine vision conference*.

Taskar, B., Chatalbashev, V., & Koller, D. (2004). Learning associative Markov networks. In *International conference on machine learning*.

Torr, P. H. S., & Murray, D. W. (1997). The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*.

Torralba, A., Murphy, K., & Freeman, W. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Conference on computer vision and pattern recognition*.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *The Journal of Machine Learning Research*.

Woodford, O., Torr, P. H. S., Reid, I., & Fitzgibbon, A. (2008). Global stereo reconstruction under second order smoothness priors. In *Conference on computer vision and pattern recognition*.

Yotta (2011). Yotta DCL horizons. http://www.yottadcl.com/horizons/.