Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction

L'ubor Ladický, Paul Sturgess, Chris Russell, Sunando Sengupta,
[lladicky,paul.sturgess,chris.russell,ssengupta]@brookes.ac.uk
Yalin Bastanlar, William Clocksin, Philip H. S. Torr.
[yalinbastanlar,wfc,philiptorr]@brookes.ac.uk

School of Technology
Oxford Brookes University
Oxford, UK
http://cms.brookes.ac.uk/research/visiongroup

The problems of object class segmentation [2], 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 [1], 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). Both may use graph cut based move making algorithms to solve the labelling problem. The correct labelling of object class can inform depth labelling and stereo reconstruction can also improve object labelling. Similarly 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.

Our joint optimisation consists of two parts, object class segmentation and dense stereo reconstruction. We follow [2] in formulating the problem of object class segmentation as finding a minimal cost labelling of a CRF defined over a set of random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$ each taking a state from the label space $\mathscr{L} = \{l_1, l_2, \ldots, l_k\}$. Each label l_j indicates a different object class such as *car*, *road*, *building* or *sky*.

For the dense stereo reconstruction part of our joint formulation we use the energy formulation of [1], who formulated the problem as one of finding a minimal cost labelling of a CRF defined over a set of random variables $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$, where each variable Y_i takes a state from the label space $\mathscr{D} = \{d_1, d_2, \ldots, d_m\}$ corresponding to a set of disparities.

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 $Z_i = [X_i, Y_i]$ takes a label $z_i = [x_i, y_i]$, from the product space of object class and disparity labels $\mathscr{L} \times \mathscr{D}$ that 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:

$$E(\mathbf{z}) = \sum_{i \in \mathscr{V}} \psi_i^J(z_i) + \sum_{i \in \mathscr{V}, j \in \mathscr{N}_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in \mathscr{C}} \psi_c^J(\mathbf{z}_c), \qquad (1)$$

where the terms ψ_i^J, ψ_{ij}^J and ψ_c^J are a sum of corresponding unary, pairwise and higher order object class and disparity potentials and joint potentials modelling interractions between $\mathbf{X}$ and $\mathbf{Y}$. See fig.1 for the graphical model of our joint CRF

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 formulate our unary interaction potential on the basis of the observed fact that certain objects occupy a certain range of real world heights. This height class relationship is modelled by estimating the a priori cost of pixel i taking label z_i using histograms of height for each class in the ground truth images. 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 as an object class boundary is more likely to occur if the disparity of two neighbouring pixels significantly differ. To take this information into account, we propose tractable pairwise potentials, which can capture mutual dependencies of object class and depth boundaries.

Optimisation of the energy $E(\mathbf{z})$ is challenging due to large number of possible states of each random variable. To deal with this problem we introduce the concept of projected moves, in which the problem is iteratively projected into each domain and move making algorithms are used to find optimal move. This method is guaranteed to converge to a local optima and is significantly faster than standard move making algorithms.

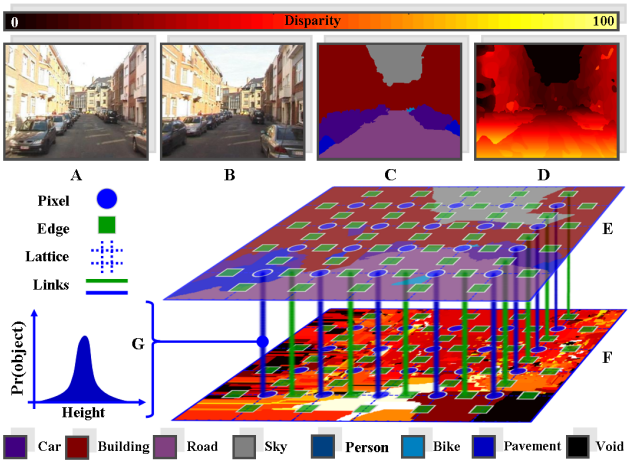

Figure 1: Graphical model of our joint CRF. *The system takes a left (**A**) and right (**B**) image from a stereo pair that has been rectified. Our formulation captures the co-dependencies between the object class segmentation problem (**E**) and the dense stereo reconstruction problem (**F**) by allowing 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 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**).*

We tested our algorithm on the challenging Leuven stereo street view data set, that we augmented with pixel level object-class and disparity ground truth. Our approach significantly outperforms existing methods for stereo reconstruction, see fig.2.

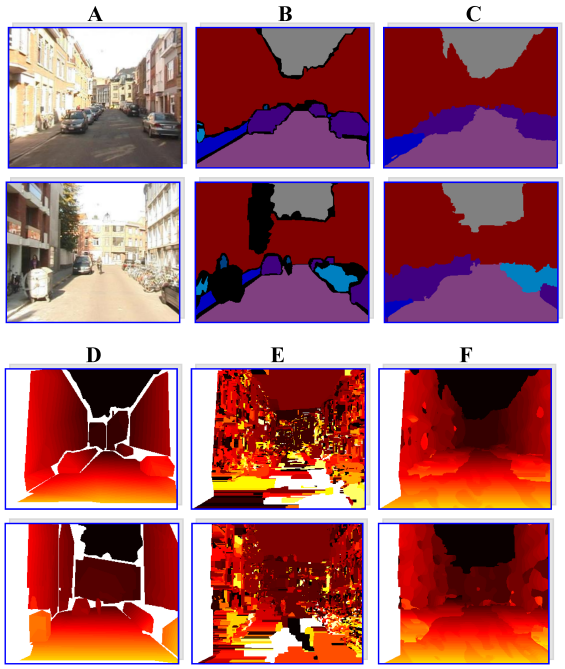

Figure 2: Qualitative object class and disparity results for Leuven data set.(**A**) *Original Image.* (**B**) *Object class segmentation ground truth.* (**C**) *Proposed method object class segmentation result.* (**D**) *Dense stereo reconstruction ground truth.* (**E**) *Stand alone dense stereo reconstruction result.* (**F**) *Proposed method dense stereo reconstruction result.*

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.

[1] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions via graph cuts. In *ICCV*, pages 508–515, 2001.

[2] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, 2009.