Combining Appearance and Structure from Motion Features for Road Scene Understanding

Paul Sturgess
paul.sturgess@brookes.ac.uk
Karteek Alahari
karteek.alahari@brookes.ac.uk
L'ubor Ladický
lladicky@brookes.ac.uk
Philip H. S. Torr
philiptorr@brookes.ac.uk

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

With the introduction of applications such as Google Street View, Microsoft Bing Maps, the problem of scene understanding has gained more importance than ever. Image sequences from such applications consist of complex scenarios involving multiple *objects*, such as people, buildings, cars, bikes. One may need to simultaneously segment and identify these objects for instance to mask out cars, or maintain highway inventories automatically. This paper deals with the problem of simultaneous pixel-wise segmentation and recognition of such complex image sequences. In particular, we focus on monocular image sequences filmed from within a driven car [2].

Many methods have been proposed to address the challenging task of combined object recognition and pixel-wise segmentation [5, 7]. Although they have achieved impressive results, they are either limited to single object classes and tend not to scale well for multiple classes [5] or fail at object boundaries [7]. Inspired by the recent work of [1] and [4], we present an approach to overcome these issues and achieve accurate segmentation and recognition of road scenes.

We formulate the problem in a Conditional Random Field (CRF) framework to probabilistically model the label likelihoods and our prior knowledge. Our approach also uses the robust P^n model potential defined on segments obtained by multiple unsupervised segmentations [4]. The energy of our higher order CRF is given by:

$$E(\mathbf{x}) = \sum_{i \in \mathscr{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathscr{E}} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathscr{S}} \psi_c(\mathbf{x}_c), \qquad (1)$$

where x_i denotes the label taken by the random variable X_i from the label set $\mathscr{L} = \{l_1, l_2, \ldots, l_k\}$. Here, the labels correspond to object classes such as pedestrians, buildings, cars, trees, and the pixels are the random variables. Every possible assignment of labels to the random variables defines a segmentation. The unary potential $\psi_i(x_i)$ gives the cost of the assignment: $X_i = x_i$, for all pixels in the set $\mathscr{V}$. The pairwise potential $\psi_{ij}(x_i, x_j)$ represents the cost of the assignment: $X_i = x_i$ and $X_j = x_j$ over the set of all pairs of interacting variables denoted by $\mathscr{E}$. The higher order potential $\psi_c(\mathbf{x}_c)$ denotes the cost of labelling the random variables $\mathbf{X}_c$, which are conditionally dependent on each other. It is defined over the set of all segments $\mathscr{S}$.

We use motion and appearance-based features to obtain the unary cost of a pixel i taking a label. The motion-based features are extracted from 3D point clouds [1], and the appearance-based features consist of textons, colour, location, and HOG descriptors. All these features are combined within a joint boosting framework [6, 8] that automatically selects the most discriminative features for each object class to generate the unary costs. The pairwise potential, also referred to as the smoothness term, takes the form of a contrast-sensitive Potts model:

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

where I_i and I_j are the colours of pixels i and j respectively. The constants θ_p, θ_v and θ_β are model parameters learned using training data [7].

The higher order potential defines the label inconsistency cost *i.e.*, the cost of assigning different labels to pixels constituting the segment, while taking the quality of a segment into account. We denote the quality of a segment c by $G(c) : c \rightarrow \mathbb{R}$. In our experiments we use the variance of colour intensity values evaluated on all constituent pixels of a segment as a quality measure. The higher order potential is defined as:

$$\psi_c(\mathbf{x}_c) = \begin{cases} N_i(\mathbf{x}_c) \frac{1}{Q} \gamma_{\max} & \text{if } N_i(\mathbf{x}_c) \leq Q \\ \gamma_{\max} & \text{otherwise,} \end{cases} \qquad (3)$$

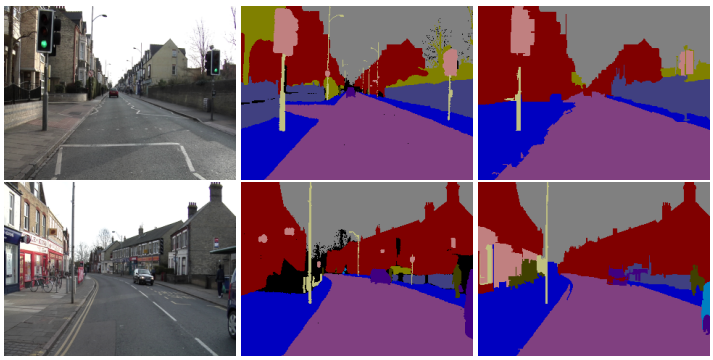

Figure 1: *Qualitative improvements achieved by our higher order* CRF *framework. We show (left to right) the original image, the ground truth image, and the higher order* CRF *result for two frames from the test sequences. The higher order potentials provide accurate segmentation* e.g., *traffic light, building (top row) and lamp post, sidewalk (bottom row).*

where $N_i(\mathbf{x}_c)$ denotes the number of pixels in the superpixel c not taking the dominant label, $\gamma_{\max} = |c|^{\theta_\alpha}(\theta_p^h + \theta_v^h G(c))$, and Q is the truncation parameter. This potential ensures the cost of breaking a *good* segment is higher than that of a *bad* segment. The set $\mathscr{S}$ of segments used for defining the higher order potentials is generated by computing multiple mean shift segmentations [3] of an image.

The segmentation is obtained by finding the lowest energy configuration of the CRF. We minimize the energy function in (1) using approximate methods such as α-expansion as described in [4].

We evaluated our method on the challenging Cambridge-driving Labelled Video Database (CamVid), which consists of 600 manually labelled frames [2]. Our method achieves an overall accuracy of 84% compared to the state-of-the-art accuracy of 69% [1]. Figure 1 highlights the qualitative improvements achieved by our higher order CRF framework.

In summary, we present an approach to integrate motion and appearance features for object recognition and segmentation of road scenes. The object class boundaries in the segmentations are well-defined and also detect the fine structures in some categories.

[1] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, volume 1, pages 44–57, 2008.

[2] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space. *PAMI*, 24(5):603–619, 2002.

[4] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82:302–324, 2009.

[5] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. Obj cut. In *CVPR*, volume 1, pages 18–25, 2005.

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

[7] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, volume 1, pages 1–15, 2006.

[8] A. Torralba, K. Murphy, and W. T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, volume 2, pages 762–769, 2004.