Combining Appearance and Structure from Motion Features for Road Scene Understanding

Paul Sturgess, Karteek Alahari, Lubor Ladicky, Philip H. S. Torr

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

Abstract

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

1 Introduction

One of the grand goals of computer vision is to interpret a scene semantically given an input image. This problem has manifested itself in various forms, such as object recognition [8, 16, 25], 3D scene recovery [14], and image segmentation [6, 10, 24]. With the introduction of applications such as Google Street View [7], Microsoft Bing maps [1], the problem of scene understanding has gained more importance than ever. Image sequences from such applications consist of complex scenarios involving multiple *objects*, such as people, buildings, cars, bikes. One may need to simultaneously segment and identify these objects for instance to mask out cars, or maintain highway inventories automatically [3]. This paper deals with the problem of simultaneous pixel-wise segmentation and recognition of such complex image sequences. In particular, we focus on monocular image sequences filmed from within a driven car [9].

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

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

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

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

2 CRFs for Object Segmentation

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

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

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

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

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

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

where $\mathscr{V}$ is the set of image pixels and $\mathscr{E}$ is the set of all pairs of interacting variables denoting the neighbourhood set $\mathscr{N}$. The labels represent the different objects, and every possible assignment of labels to the random variables (also known as a configuration of the CRF) defines a segmentation. The unary potential $\psi_i(x_i)$ gives the cost of the assignment: $X_i = x_i$. Cost functions based on colour, location, and texton features have been commonly used for object segmentation [9, 17, 25]. The pairwise potential $\psi_{ij}(x_i, x_j)$ represents the cost of the assignment: $X_i = x_i$ and $X_j = x_j$. It is also referred to as the smoothness term, and takes the form of a contrast-sensitive Potts model:

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

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

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

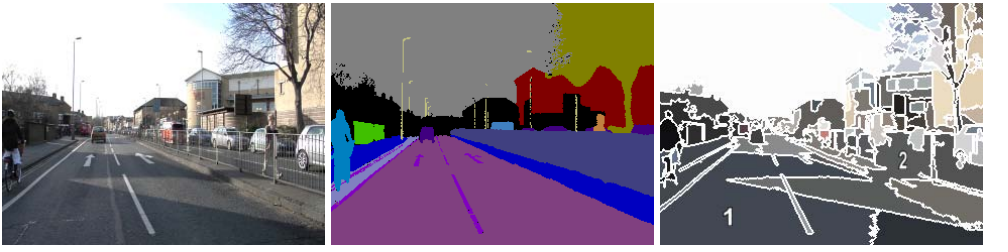


Figure 1: *Assigning a single label to all the pixels of a superpixel, as a hard constraint, might produce an incorrect labelling. We show the original image (left), its ground truth labelling (centre) and the meanshift segmentation of the image (right). The segment number '1' consists of all pixels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

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

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

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

3 Computing the Potentials

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

3.1 Motion and Structure Features

We use the five motion and structure features proposed by [8], namely: height above the camera, distance to the camera path, projected surface orientation, feature track density, and residual reconstruction error. They are computed using the inferred 3D point clouds[1], which are quite noisy due to the small baseline variations. These weak features are designed specifically for such point clouds. As noted by [8], the five cues are tailored for the driving application and are invariant to camera pitch, yaw and perspective distortions. A brief description of the features is given below.

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

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

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

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

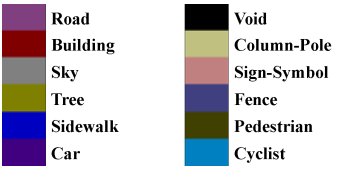
3.2 Appearance-based Features

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

3.3 Boosting for Unary Potentials

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

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



(a)

Seq. Name	# Frames	Dataset
0006R0	101	Day Train
0016E5	204	Day Train
0001TP_1	62	Dusk Train
Seq05VD	171	Day Test
0001TP_2	62	Dusk Test

(b)

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

3.4 Pairwise and Higher Order Potentials

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

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

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

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

3.5 Inferring the Segmentation

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

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

4 Experiments

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

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

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

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

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

[2] Available at http://mi.eng.cam.ac.uk/research/projects/VideoRec.

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

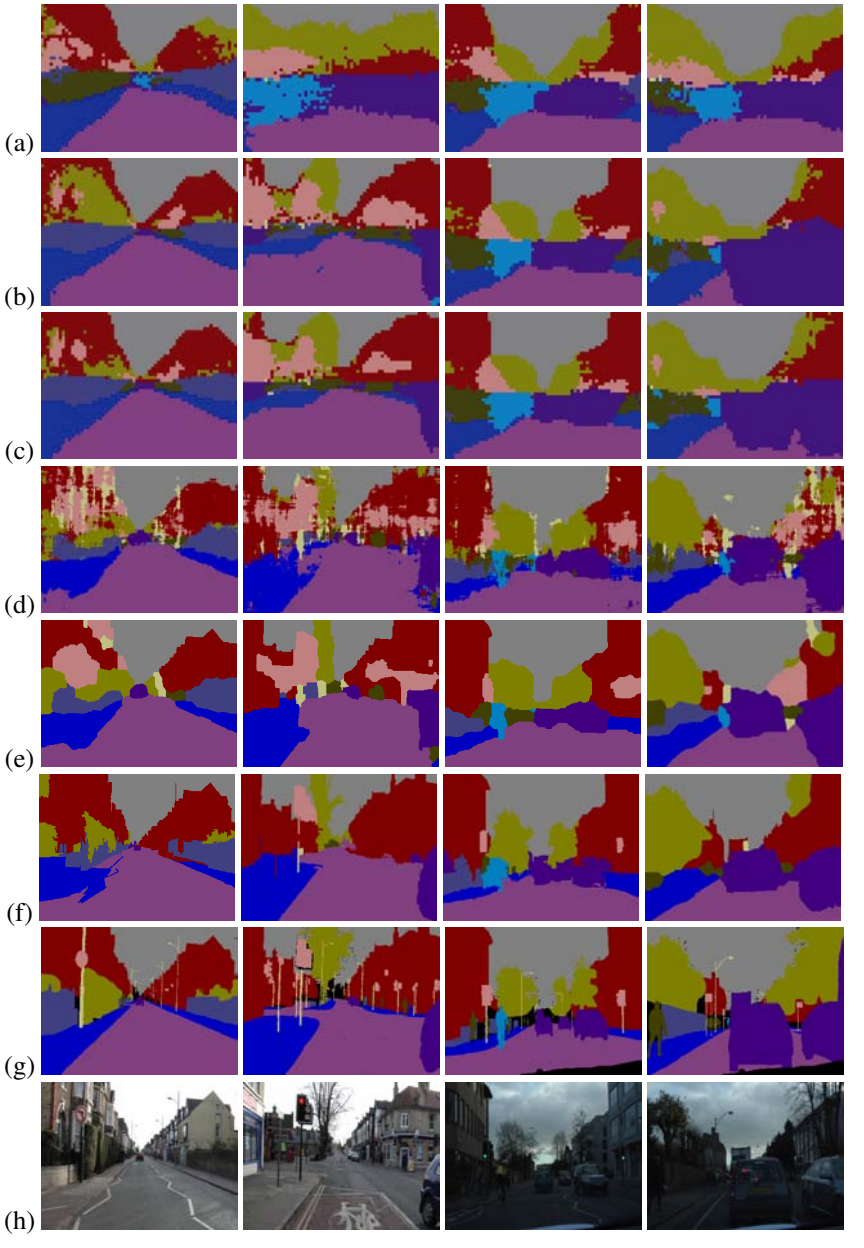


Figure 3: *Sample object category segmentations of two day and two dusk images. Results from [8] are shown in: (a) Motion and structure-based segmentation, (b) Appearance-based segmentation, (c) Combined segmentation result. Our results: (d) using only unary potentials gives poor segmentation, (e) adding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

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

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

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

5 Discussion

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

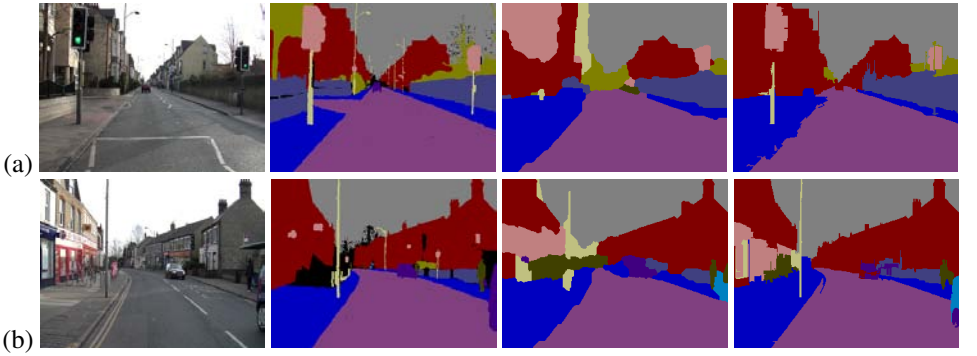


Figure 4: *Qualitative improvements achieved by our higher order* CRF *framework. We show (left to right) the original image, the ground truth image, pairwise* CRF *result, and higher order* CRF *result for two frames from the test sequences. The higher order potentials correct the object boundary errors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

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

References

[1] http://www.bing.com/maps, 2009.

[2] http://maps.google.com/help/maps/streetview, 2009.

[3] http://www.yotta.tv, 2009.

[4] A. Blake, C. Rother, M. Brown, P. Perez, and P. H. S. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, volume 1, pages 428–441, 2004.

[5] E. Borenstein and J. Malik. Shape guided object segmentation. In *CVPR*, volume 1, pages 969–976, 2006.

[6] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, volume 1, pages 105–112, 2001.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

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

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

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

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[13] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Learning and incorporating top-down cues in image segmentation. In *CVPR*, volume 2, pages 695–702, 2004.

[14] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

[15] P. Kohli, M. P. Kumar, and P. H. S. Torr. P^3 and beyond: Solving energies with higher order cliques. In *CVPR*, 2007.

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

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

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

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, pages 282–289, 2001.

[20] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order Markov random fields. In *ECCV*, volume 2, pages 269–282, 2006.

[21] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. *IJCV*, 81(1):105–118, 2009.

[22] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, volume 2, pages 860–867, 2005.

[23] J. R. Shewchuk. Triangle: Engineering a 2D quality mesh generator and delaunay triangulator. In *First ACM Workshop on Applied Computational Geometry*, volume 1448, LNCS, pages 203–222, 1996.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

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

[26] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

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

[28] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, volume 2, pages 1800–1807, 2005.