Joint Semantic and Geometric Segmentation of Videos with a Stage Model

Buyu Liu
ANU and NICTA
Canberra, ACT, Australia
buyu.liu@anu.edu.au

Xuming He
NICTA and ANU
Canberra, ACT, Australia
xuming.he@nicta.com.au

Stephen Gould
ANU
Canberra, ACT, Australia
stephen.gould@anu.edu.au

Abstract

We address the problem of geometric and semantic consistent video segmentation for outdoor scenes. With no assumption on camera movement, we jointly model the semantic-geometric class of spatio-temporal regions (supervoxels) and geometric scene layout in each frame. Our main contribution is to propose a stage scene model to efficiently capture the dependency between the semantic and geometric labels. We build a unified CRF model on supervoxel labels and stage parameters, and design an alternating inference algorithm to minimize the resulting energy function. We also extend smoothing based on hierarchical image segmentation to spatio-temporal setting and show it achieves better performance than a pairwise random field model. Our method is evaluated on the CamVid dataset and achieves state-of-the-art per-pixel as well as per-class accuracy in predicting both semantic and geometric labels.

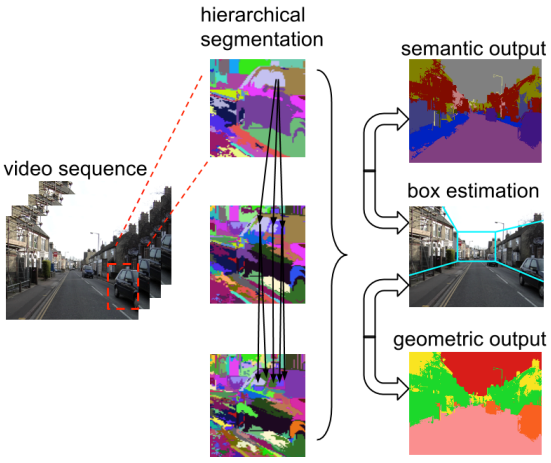

Figure 1: Overview of our approach. A hierarchical supervoxel representation is used for video sequences and we jointly predict the semantic and geometric labels based on a simple stage scene model.

1. Introduction

Scene understanding from monocular image sequences, e.g., videos taken by a moving camera, is an important problem in computer vision. Among many challenges in providing a holistic understanding of the video scenes, one core task is to infer high-level scene properties of image regions, such as semantic classes and/or geometric layout, in a consistent manner.

Most existing approaches on video segmentation focus on exploring temporal consistency in region or pixel labeling [27]. More recent work consider jointly modeling semantic class and depth of scene, and attempt to infer multiple labels of pixels consistently in spatio-temporal domain [24]. However, many of these methods require additional stereo or depth sensor as input [16], while others assume a static scene and derive the scene structure of sparse points based on structure from motion [22].

In this work, we aim to address the problem of geometric and semantic consistent video segmentation from a monocular camera for dynamic outdoor scenes. Given no assumption on camera movement and multiple moving objects in the scene, the key challenge is to integrate semantic and geometric information efficiently in a coherent framework. Inspired by [9], we consider jointly modeling the semantic class of spatio-temporal regions and a high-level description of scene geometry for each frame in video.

Specifically, we formulate video segmentation as a multi-label multi-class prediction problem, in which each over-segmented spatio-temporal volume, or supervoxel, is assigned geometric and semantic consistent labels. To efficiently capture the dependency between the semantic and geometric labels, we propose a simplified stage-like scene model [18] as an intermediate representation, which imposes additional scene dependent constraints for both semantic and geometric labeling. A typical example is shown in Figure 1, where a box-shape scene model is used to build the interaction between semantics and geometry.

We design a conditional random field (CRF) for joint modeling of the semantic class, geometric label and the

stage representation. The potential functions of our CRF encode the constraints on two types of labels based on the stage parameters. Inference can be solved by an alternating procedure between the label prediction and stage estimation. To label a video efficiently, we build on a hierarchical segmentation of a video and fuse the predictions from multi-levels to achieve both spatial and temporal consistency.

Our method is evaluated on the publicly available CamVid dataset [3], and compared with several state-of-the-art approaches. We also demonstrate the effectiveness of our scene representation for joint labeling both quantitatively and with visual results.

The main contributions of our work can be summarized as follows: (1) We propose a stage representation for jointly modeling the geometric and semantic label of a video sequence. Our results show the new representation is beneficial in predicting semantic and geometric consistent scene labeling. (2) We directly use supervoxel instead of frame-based superpixel for labeling and feature extraction, which achieves a more temporally coherent labeling. (3) We extend the smoothing method based on multilayer representation to the video setting and achieve the state-of-the-art per-class and per-pixel accuracy without any high-level object information.

2. Related Work

Holistic scene understanding is a fundamental problem in computer vision and has attracted much attention recently. Early works mostly focus on either semantic labeling [21], or the geometric layout of visual scene [13] of static images. More recent approaches attempt to jointly infer semantic and geometric property in a consistent way [9], and also consider multiple aspects of scene [14]. They found that jointly estimating these types of property can actually improve performance in predicting all of them.

Our stage model is inspired by [18] and recent work on indoor scene understanding [12]. However, the goal of [18] is to predict scene categories, and the work in [12] is focused on the layout of static indoor scenes. Our method extends the scene model used in [9] in two ways. First, we extract both appearance and geometric features rather than using the same features for prediction of two kinds of labels. More importantly, we propose a box-like stage model to efficiently capture the dependency between the semantic and geometric labels.

In video scene understanding, current methods widely exploit temporal consistency across neighboring frames, or long-term relationship, to enforce temporal label smoothness. In the following, we discuss two types of video segmentation methods. They avoid the inconsistent and independent labeling by incorporating temporal dependencies and also make better use of motion information in the video.

Unsupervised video segmentation is usually the pre-processing step for further high-level understanding [10, 5, 17, 1, 25]. Those methods are either built on optical flow or appearance in the spatio-temporal neighbors. Optical flow [4] is generally a local search and cannot capture the long-term relationship. Brox et al. [23] model long-term relationships in a video sequence by generating trackers but can be very time-consuming for a dense representation. We refer the reader to [27] for detailed comparison.

To infer high-level scene properties, video inputs usually provide more information than static images. Several works have attempted to jointly infer both semantic and 3D scene structure [16]. However, most of them rely on structure from motion [11], which assumes single relative motion, or additional stereo inputs [7]. In our model, we do not make any assumption about camera or foreground motion. The method in [22] is based on SfM with full video sequence as input and get a relatively dense reconstruction and [28] requires dense depth map as additional input. For [16], it utilizes the stereo pairs of images for static setting and relies on motion for monocular setting but achieves inferior results. Recent work also addresses the geometric context labeling in videos [19]. Tighe et al. [24] explore the semantic and geometric information in video segmentation, but geometric labeling is treated as another type of external information. Moreover, they pool over the finest layer of supervoxels, without integrating the high-level information provided by the hierarchy.

Other methods in video scene understanding make use of high-level object information to enforce object-driven label consistency [15, 26], and achieve state-of-the-art results. Unlike those works, our model does not require high level information such as object detections and our features are designed to fully exploit the video sequence.

As for the model inference part, our alternating inference shares certain similarity with [9, 16], in which two types of labeling are inferred with an alternating processes. However, our work explores two different kinds of scene properties, one is the stage model space and the other is the joint label space.

3. Our approach

We first describe a hierarchical supervoxel representation of a video clip and the features we used for joint label prediction. We then introduce the stage scene model, followed by the joint CRF for the semantic and geometric label prediction.

3.1. Supervoxel trees

Given a video input, we divide the whole sequence into smaller chunks with duration T. In the following we will focus on label prediction in each chunk and the temporal label consistency can be addressed by using overlapping chunks. For each video chunk, we employ the method proposed

Base features
Semantic and Geometric Output
S1-3. average, max and variance of semantic probability for each class
G1-3. average, max and variance of geometric probability for each class
Appearance
C1-2. mean and variance of CIE Lab value.
T1-2. mean and covariance of 17 dim filterbank response.
H1-2. mean and variance of HoG.
Optical Flow
O1-3. magnitude weighted flow direction histogram, mean flow and flow differential at 3 scales.
Shape
P1-3. mean and variance of area, ratio of perimeter and their change across time
Movement
M1. voxel start region position and end position.
M2. histogram for location change across time.

Table 1: Image feature and region statistics computed to represent supervoxels. See experiment section for details.

in [10] and obtain a hierarchical segmentation. Supervoxels are defined as spatio-temporally connected regions at the finest level. See Figure 1 for an example of hierarchical segment trees in 2D view. We denote those supervoxels as $\{v_i\}_{i=1}^N$ and associate two variables l_i^g and l_i^s for v_i's geometric and semantic labels, respectively.

We extract a set of image and motion features at each pixel, including color, texture, HOG, optical flow. We also apply method proposed in [13] and [8] in each frame to obtain the per-pixel semantic and geometric probability independently. Given these pixel-level features, we compute a feature vector $\mathbf{f}_i$ for each supervoxel as listed in Table 1.

3.2. Stage scene model

Modeling semantic and geometric label interaction at the local supervoxel level is limited as it ignores the global scene structure. To capture the long-range dependency of two types of labels, we propose an intermediate scene representation based on the stage scene model [18], and use this representation to link the semantic and geometric labels.

In particular, we focus on the urban street scene and design three types of box-like structure: frontal, turning view and view after taking a turn, as shown in Figure 2. These stage scene models cover most of the common scenarios in urban street videos taken by vehicle-mounted cameras (forward-looking). They also give us a coarse layout representation, which is used to impose global constraints on the joint prediction of semantic and geometric labeling.

We parameterize the stage scene model with the 2D position of its 8 vertices and the vertical position of the horizon line. In a video chunk, we denote the sequence of the model parameters as $S = \{s_1, \cdots, s_T\}$. Given the stage model parameters, we can extract a set of stage dependent features for each supervoxel, as illustrated in Figure 2 and Table 2. They are designed to capture the distribution of the location for different object categories in real 3D world. We denote the stage dependent features for supervoxel i as $\mathbf{g}_i(S)$.

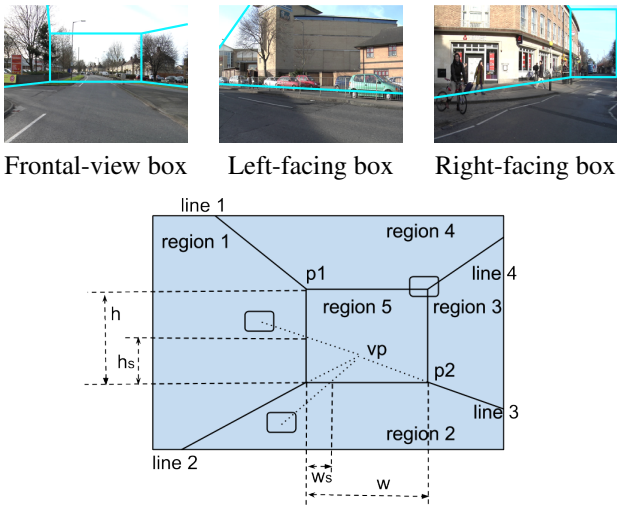

Figure 2: Top: three different box-like stage scene models. The last two cases consist of two subcases in terms of facing direction of box. Bottom: supervoxel features that reflect its geometric relationship with the box. For segment in region 1, 3 or 5, we compute the $\frac{h_s}{h}$ as its normalized height. For segment in region 2, 4 or 5, distance $\frac{w_s}{w}$ is obtained as the normalized width from road side. Vp is the estimated vanishing point.

3.3. Joint CRF for consistent labeling

To jointly predict semantic and geometric labels, we build a CRF model for each video chunk. Our model consists of three potential functions, which are described in details as follows.

3.3.1 Supervoxel potential

We model the relationship between the stage model and supervoxel labels based on a classifier taking both base and stage-dependent features. The corresponding potential function $E(l_i^s, l_i^g, S)$ can be written as

$$E(l_i^s, l_i^g, S) = -\log P_c(l_i^s, l_i^g | \mathbf{f}_i, \mathbf{g}_i(S)) \quad (1)$$

where P_c is the probabilistic score of the classifier output, and $\mathbf{f}_i$, $\mathbf{g}_i(S)$ are video and stage dependent features, respectively. In this work, we use a Random Forest [6] as the classifier.

Stage dependent features
Location and Motion
L1. mean, variance of distance between horizon and region across time
L2. mean, variance of the differential of distance between horizon and region centroid across time
L3. ratio of pixel the above the horizon
L4-6. mean, variance and the differential of distance from region centroid to bottom, side, top box region
L7. mean, variance and the differential of overlap rate of region centroid and five regions
L8. majority overlap rate with five regions
L9. normalized box location in image
L10. normalized differential of box location in time

Table 2: Stage dependent supervoxel features in video sequence setting.

3.3.2 Temporal smoothness for the stage parameters

Within each video chunk, we enforce pairwise smoothness for the stage parameters in two neighboring frames. Specifically, the potential function for the stage parameters $E(S)$ can be written as,

$$E(S) = \sum_{t=1}^{T-1} \sigma_s^2 ||s_t - s_{t+1}||^2 \quad (2)$$

where σ_s is the effective width of smoothing window in time.

3.3.3 Pairwise potentials for labels

We consider two types of smoothing for the semantic and geometric labeling. First we adopt the conventional CRF setting, in which we add a pairwise potential for neighboring supervoxels' labels.

We define a spatial and a temporal neighborhood based on the topology of supervoxels. For the spatial neighborhood N_s, we connect any two v_i and v_j that are adjacent in at least 10 frames. For the temporal neighborhood N_t, we connect v_i and v_j when v_i meets v_j's head or tail. Denote $l_i = (l_i^s, l_i^g)$, the pairwise term $E(l_i, l_j)$ is defined as

$$E(l_i, l_j) = \beta \begin{cases} e^{-\alpha_1 f_c(i,j) - \alpha_2 f_o(i,j)}, & l_i \neq l_j \\ 0, & l_i = l_j \end{cases} \quad (3)$$

where in $f_c(i,j)$ is the normalized color distance between v_i and v_j, $f_o(i,j)$ is the normalized χ^2 distance between optical flow distributions at v_i and v_j. The adjacent supervoxels are more likely to share the same label when they are similar in photometric features, or have the same trajectory.

Figure 3: Examples of smoothed stage model estimation in video sequence.

3.3.4 Smoothing with supervoxel tree

We also consider an alternative approach to enforce the smoothness between supervoxel labels. Based on the hierarchical segmentation results, we follow the same setting as in [20]. We extract all features at each level in the hierarchy and train independent classifiers for multiple layers. After obtaining the results for spatio-temporal regions in multilayer, we map to the finest level and train an additional classifier based on the concatenated individual classifier output vectors for each supervoxel. The corresponding potential has the same form as the supervoxel potential.

3.4. Model inference

To predict the semantic and geometric labels for a video chunk, we compute the MAP estimate of the CRF model. Note that our model involves both stage parameters and the supervoxel labels, which makes the inference a challenging problem. We take a greedy approach which minimizes the energy function based on coordinate descent. More specifically, we alternate between two subproblems: in one subproblem, we fix the supervoxel labeling and minimize w.r.t the stage parameters; while in the second subproblem, we fix the stage parameters and search for optimal supervoxel labeling.

3.4.1 Model initialization

We initialize the stage model based on the pixel-wise geometric labels. We can obtain the main geometric result with features in Table 1 and estimate the stage location with respect to it.

First, we generate a set of proposals for the stage parameters based on line fitting of the initial geometric labeling and image. Afterwards, we exhaustively search the proposal pool for the best stage parameter based on the overlaps between each proposal and the initial geometric labeling. After extracting the stage parameters, we perform a guassian smoothing based on the CRF model to keep the temporal consistency of the stage among frames.

Figure 3 shows several examples of our stage prediction, which are quite smooth and consistent in temporal space. Moreover, the stage representation can capture the main structure of street scene.

	Road	Building	Sky	Tree	Sidewalk	Car	Column-Pole	Fence	Pedestrian	Bicylist	Sign-symbol	Pixel	Class
Semantic Only	94.4	91.0	90.7	81.0	52.1	71.9	2.0	5.4	35.9	20.8	3.4	81.5	49.9
Static	94.2	68.7	95.5	82.4	62.5	69.0	18.1	23.6	57.2	36.1	52.5	79.9	60.0
Static + Stage	92.5	71.8	94.6	79.3	66.6	70.5	17.6	30.0	56.3	41.2	54.2	80.1	61.3
Voxel	94.4	69.0	95.4	83.3	63.6	69.1	16.2	26.4	65.2	36.9	51.2	80.2	61.0
Voxel + Stage	94.3	67.7	95.6	82.3	63.6	70.7	18.0	29.0	64.1	37.8	**55.9**	80.0	61.6
Pairwise	93.7	68.9	95.3	**82.3**	**66.5**	**70.7**	17.3	30.3	**65.6**	37.8	54.7	80.3	62.1
Multilayer	95.1	75.4	95.7	81.3	62.0	70.0	**17.9**	34.6	61.6	**46.0**	52.1	81.8	**62.8**
Tighe [24]	**95.9**	**87.0**	**96.9**	67.1	70.0	62.7	1.7	17.9	14.7	19.4	30.1	83.3	51.2
Sturgess [22]	95.3	84.5	97.5	72.6	77.6	72.7	8.1	45.7	34.2	28.5	34.1	**83.8**	59.2
Ladicky [15]	93.9	81.5	96.2	76.6	81.5	78.7	14.3	**47.6**	43.0	33.9	40.2	**83.8**	62.5

Table 3: Per-class average and Per-pixel semantic result on CamVid dataset. We show the performance of our approach with different configurations, as well as the-state-of-the-art accuracy. See text for details.

	Sky	Horizon	Vertical	Per-Pixel	Per-Class
Tighe [24]	–	–	–	94.2	94.7
Voxel	95.4	98.5	92.9	95.3	95.6
Voxel + Stage	95.2	98.4	93.0	95.3	95.5
Multilayer	95.7	98.5	93.1	95.5	95.8

Table 4: Main geometric class result on CamVid dataset.

3.4.2 Joint label prediction

Given the stage parameters, we infer the supervoxel labeling based on graph-cuts [2] if the pairwise CRF is used. In the setting of smoothing with supervoxel tree, we can predict the semantic and geometric labels of the supervoxels directly.

It is more challenging to refine the stage parameter given supervoxel labeling, due to the large state space of S. We take an approximate approach similar to the model initialization and update the stage parameters. We take a candidate of stage parameters as long as it reduces the overall energy function.

4. Experiment

4.1. Dataset and experiment setup

We evaluate our video segmentation on the standard CamVid dataset, which consists of daytime and dusk videos of street scenes. We also follow the training/test split of [3], with two daytime and one dusk for training and one daytime and one dusk for testing. In our experiment, we use a chunk that consists of 60 frames and apply the segmentation in each chunk.

Ground truth label are provided with 11 class: Sky, $Building$, $Tree$, $SideWalk$, Car, $ColumnPole$, $Fence$, $Pedestrian$, $Bicylist$ and $Signsymbol$. Although we evaluate the accuracy of output in labelled testing frames, we can obtain dense labels for all frames in the test video.

The original CamVid dataset provides only semantic classes. To obtain the ground truth geometric label, we apply a simple mapping from 11 semantic class to 5 geometric class. Note that our model is not restricted to this setting and more complicated geometric label space designing is also feasible. The geometric class in based on [13]. We have three main classes, Sky, $Horizontal$ and $Vertical$. For vertical class, we have three subclasses as $Planar$, $Porous$ and $Solid$.

4.2. Experimental results

We summarize our results on the CamVid dataset in Table 3 for the semantic segmentation and in Table 4 for the geometric segmentation. For semantic segmentation, we report four groups of results based on different configurations of our approach. The 'Semantic Only' is the pixel-wise semantic labeling results based on the Darwin system [8]. The 'Static' is the joint prediction of geometric and semantic labeling with key frame feature [9], while 'Static+Stage' adds the stage model features to predict two types of labels. The 'Voxel' and 'Voxel+Stage' are based on supervoxel representation instead of static image features. Finally, 'Pairwise' and 'Multilayer' are two versions of our full model, in which the former models spatio-temporal smoothing with pairwise terms and the latter is based on supervoxel trees. For geometric segmentation, Table 4 shows the results from three settings of our methods.

From these results, we can see that the joint supervoxel

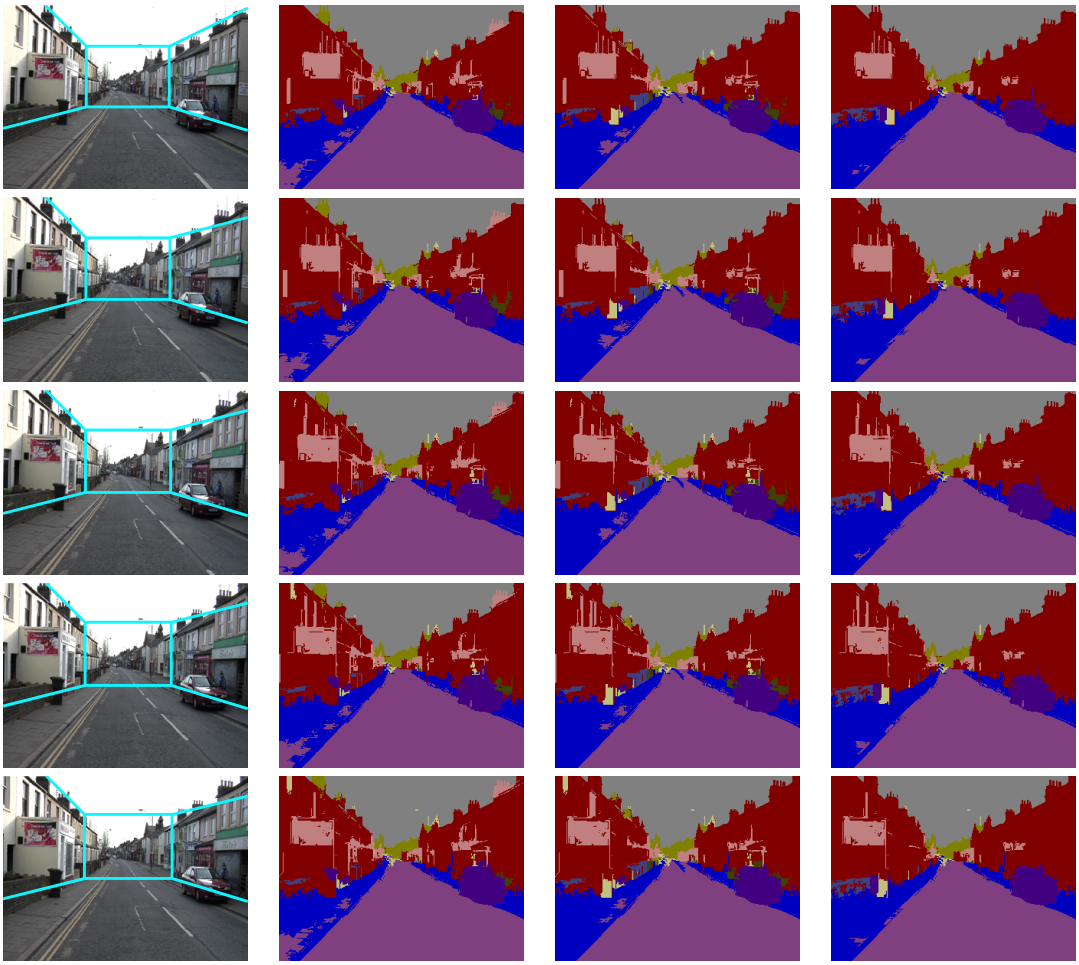

Figure 4: Examples of final prediction results from different settings. First column: input image frame with estimated stage models; Second column: joint labeling with supervoxel only; Third column: joint labeling with supervoxel plus stage features; Fourth column: full model prediction with multilayer smoothing.

labeling with the stage model achieves better performance than the baseline pixel semantic prediction. Our joint video segmentation achieves higher per-class accuracy and comparable per-pixel accuracy w.r.t the state-of-the-art methods. Note that we do not use pre-trained object models [15], nor 3D information from SfM [22]. In Table 4, we can see that our performance on geometric labeling is also superior to the state-of-the-art.

Figure 4 shows some examples of our results, which are 5 consecutive frames in a sequence. The first column is the input frames overlaid with the estimated stage models. The second and third columns are from the 'Voxel' and 'Voxel+Stage' settings. The final column is the output from our full model with supervoxel tree smoothing. The visual results demonstrate that the prediction quality becomes better after adding more model components.

4.3. Experimental analysis

We now provide detailed analysis on the main components in our method. We consider three sets of experiments in which only partial of our model is used to generate the joint label prediction. These experiments corresponds to three rows in Table 3 (from 2nd to 4th). In the following subsections, we will look deeper into these results.

4.3.1 Static scene with stage model

From Table 3, we can clearly see that the stage model improve the labeling results at key frames. In particular, the $Fence$ and $Bicyclist$ classes achieve significant improvement. We also show some qualitative results in Figure 5. We have two observations: firstly, our stage estimation is not perfect but accurate enough to be a good intermediate

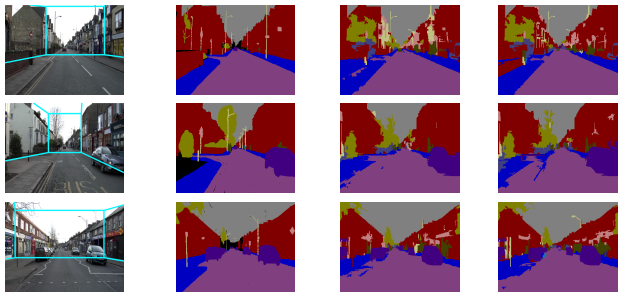

Figure 5: Examples of semantic labeling with static image features and additional stage model features.

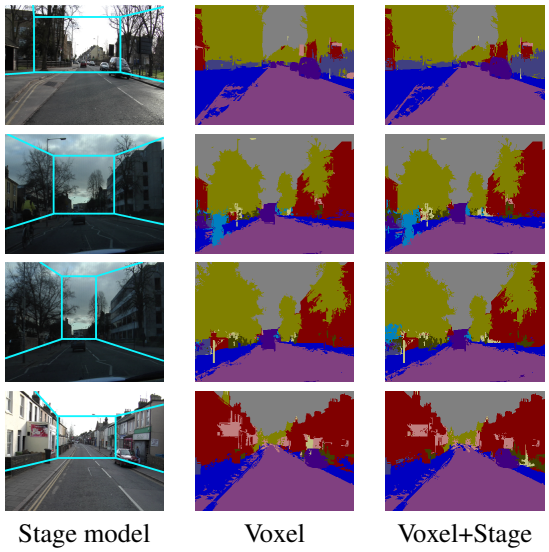

Figure 6: Examples of semantic results with supervoxel features and additional stage model features. The stage model improves semantic labeling in these video sequences.

scene representation; secondly, we can see the segmentation of $Sidewalk$, Car and $Building$ in those images are much better. For example, the first row shows that introducing stage can not only smooth the $SideWalk$ but also provide stronger information for $Building$. The main reason that the stage model is beneficial for joint label prediction is that it provide us with more geometric information such as the height of certain object in real world, the relative distance to the road side and the distribution of each category in each region.

4.3.2 Video scene with stage models

In the video setting, we can see that the stage model only slightly improves the semantic label results but has little effect on the geometric labels. One possible reason is due to the noisy estimation of stage parameters in videos. Also,

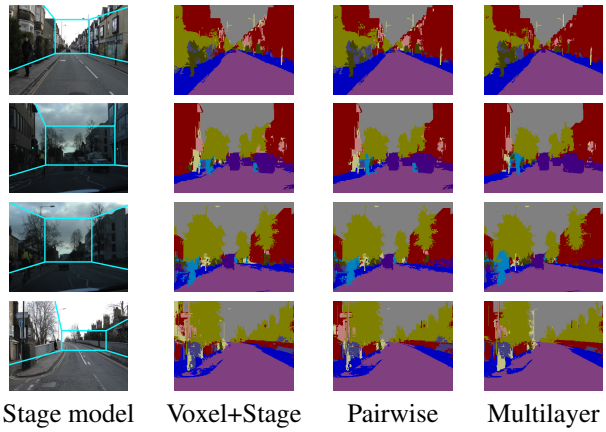

Figure 7: Examples of semantic labeling results from pairwise CRF based smoothing and multilayer supervoxel based smoothing.

the relationship between the stage and supervoxels is more challenging to capture based on simple position statistics.

Some examples of semantic labeling results are shown in Figure 6, in which we can observe the positive impact of the stage model. For example, the second and third row show that the additional information help improve the $Bicylist$ case: in the dusk, while the appearance based cue is weak, the stage information provides geometric information to boost the score of the correct classes.

4.3.3 Multilayer integration

Notice that the pairwise CRF only slightly improves the semantic labeling performance in our setting. This is likely due to the strong unary prediction based on supervoxel features, the irregular shape of supervoxels and the complex graph structure in the final pairwise CRF.

The multilayer based smoothing, on the other hand, provides better performance for both semantic and geometric labeling. The supervoxels at coarse layers can be viewed as a higher-order smoothness term, and as we extract features in each layer independently, the coarser layer can capture more information than the lower ones. It may also lead to a more stable statistical dependency between the supervoxel location with respective to stage and its label.

We compare some example results of semantic labeling in Figure 7, which are generated by the single layer model, pairwise CRF and multilayer integration. We can see that, for instance in the second and third row, both the pairwise CRF and multilayer model help improve the class $Bicylist$; but the multilayer model gives a better performance. The fourth row shows oversmoothing of pairwise CRF on $ColumnPole$, and correct prediction from the multilayer model.

5. Conclusion and Discussion

This paper has presented a novel method to combine geometric and semantic information in understanding dynamic urban street scene. We introduce a stage model as an intermediate representation of the geometric information and efficiently combine two types information. We also show that by applying the hierarchical structure, we can get a better smoothing result. Compared to state-of-the-art methods, we achieve higher average class accuracy and comparable pixel level accuracy.

Our current stage model fits street scenes from a driving perspective; however, it is still quite rigid in the general case. This can be improved by introducing more subcategories of stage models to represent a scene. Moreover, a deeper integration of the supervoxel hierarchy and the labeling might also improve the prediction performance. For future work, we intend to explore more geometric information in video and a more efficient way to combine the semantic and geometric information.

Acknowledgement

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012. 2

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 2001. 5

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008. 2, 5

[4] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *TPAMI*, 2011. 2

[5] H.-T. T. Cheng and N. Ahuja. Exploiting nonlocal spatiotemporal structure for video segmentation. In *CVPR*, 2012. 2

[6] P. Dollár. Piotr's Image and Video Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html. 3

[7] G. Floros and B. Leibe. Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *CVPR*, 2012. 2

[8] S. Gould. DARWIN: A framework for machine learning and computer vision research and development. *JMLR*, 2012. 3, 5

[9] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 1, 2, 5

[10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. *CVPR*, 2010. 2, 3

[11] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 2

[12] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. *ICCV*, 2009. 2

[13] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007. 2, 3, 5

[14] D. Hoiem, A. a. Efros, and M. Hebert. Closing the loop in scene interpretation. *CVPR*, 2008. 2

[15] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010. 2, 5, 6

[16] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. F. Clocksin, and P. H. S. Torr. Joint optimization for object class segmentation and dense stereo reconstruction. *IJCV*, 2012. 1, 2

[17] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011. 2

[18] V. Nedovi, A. W. Smeulders, A. Redert, and J.-M. Geusebroek. Stages as models of scene geometry. *PAMI*, 2010. 1, 2, 3

[19] S. H. Raza, M. Grundmann, and I. Essa. Geometric context from video. *CVPR*, 2013. 2

[20] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012. 4

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 2

[22] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009. 1, 2, 5, 6

[23] T.Brox and J.Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2

[24] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 1, 2, 5

[25] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *ECCV*, 2010. 2

[26] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, 2008. 2

[27] C. Xu. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012. 1, 2

[28] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. In *ECCV*, 2010. 2