Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion

Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard

Department of Computer Science, University of Freiburg, Germany

Abstract. Semantic scene understanding of unstructured environments is a highly challenging task for robots operating in the real world. Deep Convolutional Neural Network (DCNN) architectures define the state of the art in various segmentation tasks. So far, researchers have focused on segmentation with RGB data. In this paper, we study the use of multispectral and multimodal images for semantic segmentation and develop fusion architectures that learn from RGB, Near-InfraRed (NIR) channels, and depth data. We introduce a first-of-its-kind multispectral segmentation benchmark that contains $15,000$ images and 325 pixel-wise ground truth annotations of unstructured forest environments. We identify new data augmentation strategies that enable training of very deep models using relatively small datasets. We show that our UpNet architecture exceeds the state of the art both qualitatively and quantitatively on our benchmark. In addition, we present experimental results for segmentation under challenging real-world conditions.

Keywords: Semantic Segmentation, Convolutional Neural Networks, Scene Understanding, Multimodal Perception

1 Introduction

Semantic scene understanding is a cornerstone for autonomous robot navigation in real-world environments. Thus far, most research on semantic scene understanding has been focused on structured environments, such as urban road scenes and indoor environments, where the objects in the scene are rigid and have distinct geometric properties. During the DARPA grand challenge, several techniques were developed for offroad perception using both cameras and lasers [13]. However, for navigation in forested environments, robots must make more complex decisions. In particular, there are obstacles that the robot can drive over, such as tall grass or bushes, but these must be distinguished safely from obstacles that the robot must avoid, such as boulders or tree trunks.

In forested environments, one can exploit the presence of chlorophyll in certain obstacles as a way to discern which obstacles can be driven over [1]. However, the caveat is the reliable detection of chlorophyll using monocular cameras. This detection can be enhanced by additionally using the NIR wavelength $(0.7 - 1.1\mu m)$, which provides a high fidelity description on the presence of vegetation. Potentially, NIR images can also enhance border accuracy and visual quality. We aim to explore the correlation and de-correlation of visible and NIR images frequencies to extract more accurate information about the scene.

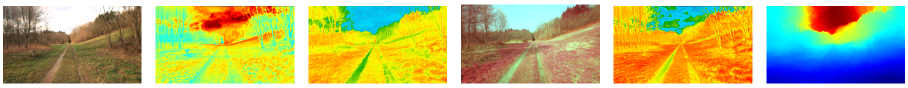
Benchmark and demo are publicly available at http://deepscene.cs.uni-freiburg.de

In this paper, we address this problem by leveraging deep up-convolutional neural networks and techniques developed in the field of photogrammetry using multispectral cameras to obtain a robust pixel-accurate segmentation of the scene. We developed an inexpensive system to capture RGB, NIR, and depth data using two monocular cameras, and introduce a first-of-a-kind multispectral and multimodal segmentation benchmark. We first evaluate the segmentation using our UpNet architecture, individually trained on various spectra and modalities contained in our dataset, then identify the best performing modalities and fuse them using various DCNN fusion architecture configurations. We show that the fused result outperforms segmentation using only RGB data.

2 Multispectral Segmentation Benchmark

We collected the dataset using our Viona autonomous mobile robot platform equipped with a Bumblebee2 stereo vision camera and a modified dashcam with the NIR-cut filter removed for acquiring RGB and NIR data respectively. We use a Wratten 25A filter in the dashcam to capture the NIR wavelength in the blue and green channels. Both cameras are time synchronized and frames were captured at 20Hz. In order to match the images captured by both cameras, we first compute SIFT [9] correspondences between the images using the Difference-of-Gaussian detector to provide similarity-invariance and then filter the detected keypoints with the nearest neighbours test, followed by requiring consistency between the matches with respect to an affine transformation. The matches are further filtered using Random Sample Consensus (RANSAC) [2] and the transformation is estimated using the Moving Least Squares method by rendering through a mesh of triangles. We then transform the RGB image with respect to the NIR image and crop to the intersecting regions of interest. Although our implementation uses two cameras, it is the most cost-effective solution compared to commercial single multispectral cameras.

We collected data on three different days to have enough variability in lighting conditions as shadows and sun angles play a crucial role in the quality of acquired images. Our raw dataset contains over $15,000$ images sub-sampled at 1Hz, which corresponds to traversing about 4.7km each day. Our benchmark contains 325 images with pixel level groundtruth annotations which were manually annotated. As there is an abundant presence of vegetation in our environment, we can compute global-based vegetation indices such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) to extract consistent spatial and global information. NDVI is resistant to noise caused due to changing sun angles, topography and shadows but is susceptible to error due to variable atmospheric and canopy background conditions [4]. EVI was proposed to compensate for these defects with improved sensitivity to high biomass regions and improved detection though decoupling of canopy background signal and reduction in atmospheric influences. For all the images in our dataset, we calculate NDVI and EVI as shown by Huete *et al.* [4].



(a) RGB (b) NIR (c) NDVI (d) NRG (e) EVI (f) DEPTH

Fig. 1. Sample images from our benchmark showing various spectra and modalities.

Although our dataset contains images from the Bumblebee stereo pair, the processed disparity images were substantially noisy due to several factors such as rectification artifacts, motion blur, etc. We compared the results from semi-global matching [3] to a DCNN approach that predicts depth from single images and found that for an unstructured environment such as ours, the DCNN approach gave better results. In our work, we use the approach from Liu *et. al,* [6] that employs a deep convolutional neural field model for depth estimation by constructing unary and pairwise potentials of conditional random fields. Fig. 1 shows some examples from our benchmark from each spectrum and modality.

3 Technical Approach

Recently, approaches that employ DCNNs for semantic segmentation have achieved state-of-the-art performance on segmentation benchmarks including PASCAL VOC, PASCAL Parts, PASCAL-Context, Sift-Flow and KITTI [8, 11]. These networks are trained end-to-end and do not require multi-stage techniques. Due to their special architecture they take the full context of the image into account while providing pixel-accurate segmentations. Our UpNet architecture follows this general architecture with two main components: contraction and expansion. Given an input image, the contraction is responsible for generating a low resolution segmentation mask. We use the 13-layer VGG [12] architecture as basis on the contraction side. The expansion side consists of 5 up-convolutional refinement segments that refine the coarse segmentation masks generated by the contraction segment. Each up-convolutional refinement is composed of one up-sampling layer followed by a convolution layer.

We represent the training set as $S = \{(X_n, Y_n), n = 1, \ldots, N\}$, where $X_n = \{x_j, j = 1, \ldots, |X_n|\}$ denotes the raw image, $Y_n = \{y_i, j = 1, \ldots, |X_n|\}, y_j \in \{0, C\}$ denotes the corresponding groundtruth mask with C classes, θ are the parameters of the network and $f(x_j; \theta)$ is the activation function. The goal of our network is to learn features by minimizing the cross-entropy $(softmax)$ loss that can be computed as $\mathcal{L}(u, y) = -\sum_k y_k \log u_k$. Using stochastic gradient decent, we then solve

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}((f(x^i; \theta)), y^i).$$

We tested two strategies to make the network learn the integration of multiple spectra and modalities: (i) one that *stacks all channels* directly at the input; (ii) a *Late-fused–convolution* of separate networks that are trained individually on each input modality. The most intuitive paradigm of fusing data using DCNNs is by stacking them into multiple channels and learning combined features end-to-end. However, previous efforts have been unsuccessful due to the difficulty in propagating gradients through the entire length of the model [8]. Contrastingly, in the late-fused-convolution approach, each model is first learned to segment using a specific spectrum/modality. Afterwards, the feature maps are summed up element-wise before a series of convolution, pooling and up-convolution layers. The later approach has the advantage as features in each model may be good at classifying a specific class and combining them may yield a better throughput, even though it necessitates heavy parameter tuning. Our experiments provide an in-depth analysis of the advantages and disadvantages of each of these approaches in the context of semantic segmentation.

4 Results and Insights

In this section, we report results using the various spectra and modalities in our benchmark. We use the Caffe [5] deep learning framework for the implementation. Training on an NVIDIA Titan X GPU took about 7 days.

Comparison to the state of the art To compare with the state-of-the-art, we train models using the *RGB RSC* set from our benchmark which contains 60,900 RGB images with Rotation, Scale and Color augmentations applied. We selected the baseline networks by choosing the top three end-to-end deep learning approaches from the PASCAL VOC 2012 leaderboard. We explored the parameter space to achieve the best baseline performance. We trained our network with both fixed and poly learning rate policies, which can be given as $\text{base\_lr} \times \left(\frac{1-\text{iter}}{\text{max\_iter}}\right)^{\text{power}}$. We found the poly learning rate policy to converge much faster and yield a slight improvement in performance. The metrics shown in Tab. 1 correspond to Mean Intersection over Union (IoU), Mean Pixel Accuracy (PA), Precision (PRE), Recall (REC), False Positive Rate (FPR), False Negative Rate (FNR) and the time reported is for a forward pass through the network. The results demonstrate that our network outperforms all the state-of-the-art approaches and with a runtime of almost twice as fast as the second best technique.

Table 1. Performance of our proposed model in comparison to the state-of-the-art

Baseline	IoU	PA	PRE	REC	FPR	FNR	Time
FCN-8 [8]	77.46	90.95	87.38	85.97	10.32	12.12	∼ 255ms
SegNet [10]	74.81	88.47	84.63	86.39	13.53	11.65	∼ 156ms
ParseNet [7]	83.65	93.43	90.07	91.57	8.94	7.41	∼ 90ms
Ours	**85.31**	**94.47**	**91.54**	**91.91**	**7.40**	**7.30**	∼ **54**ms

Parameter Estimation and Augmentation To increase the effective number of training samples, we employ data augmentations including scaling, rotation, color, mirroring, cropping, vignetting, skewing, and horizontal flipping. We evaluated the effect of augmentation using three different subsets in our benchmark: RSC (Rotation, Scale, Color), Geometric augmentation (Rotation, Scale, Mirroring, Cropping, Skewing, Flipping) and all aforementioned augmentations together. Tab. 2 shows the results from these experiments. Data augmentation helps train very large networks on small datasets. However, on the present dataset it has a smaller impact on performance than on PASCAL VOC or human body part segmentation [11]. In our network, we replace the dropout in the VGG architecture with spatial dropout which gives us an improvement of 5.7%. Furthermore, we initialize the convolution layers in the expansion part of the network with Xavier initialization, which makes the convergence faster and also enables us to use a higher learning rate. This yields a 1% improvement.

Table 2. Comparison on the effects of augmentation on our benchmark.

	Sky	Trail	Grass	Veg	Obst	IoU	PA
Ours Aug.RSC	90.46	84.51	86.72	90.66	44.39	84.90	94.47
Ours Aug.Geo	89.60	84.47	86.03	90.40	42.23	84.39	94.15
Ours Aug.All	90.39	85.03	86.78	90.90	45.31	85.30	94.51

Evaluations on Multi-Spectrum/Modality Benchmark Segmentation using RGB yields best results among all the individual spectra and modalities we experimented with. The low representational power of depth images causes poor performance in the grass, vegetation and trail classes, bringing down the mean IoU. The results in Tab. 3 demonstrate the need for fusion. Multispectrum channel fusion such as NRG (Near-Infrared, Red, Green) shows greater performance when compared to their individual counterparts and better recognition of obstacles. The best channel fusion we obtained was using a three channel input, composed of grayscaled RGB, NIR and depth data. It achieved an IoU of 86.35% and most importantly a considerable gain (over 13%) on the obstacle class, which is the hardest to segment in our benchmark. The overall best performance was from the late-fused-convolution of RGB and EVI, achieving a mean IoU of 86.9% and comparably top results in individual class IoUs as well. This approach also had the lowest false positive and false negative rates.

Table 3. Comparison of deep multispectrum and multimodal fusion approaches. D, N, E refer to depth, NIR and EVI respectively. CF and LFC refer channel fusion and late-fused-convolution.

	Sky	Trail	Grass	Veg	Obst	IoU	FPR	FNR
RGB	90.46	84.51	86.72	90.66	44.39	84.90	7.80	7.40
NIR	86.08	75.57	81.44	87.05	42.61	80.22	10.22	9.60
DEPTH	88.24	66.47	73.35	83.13	46.13	76.10	12.76	11.14
NRG	89.88	85.08	86.27	90.55	47.56	85.23	7.70	7.10
EVI	88.00	83.40	84.59	87.68	44.9	83.25	8.70	8.10
NDVI	87.79	83.86	83.57	87.45	48.19	83.39	8.62	8.00
3CF RGB-N-D	89.23	**85.86**	86.08	90.32	**61.68**	86.35	7.50	6.20
4CF RGB-N	89.64	83.37	85.83	**90.67**	59.85	85.79	**7.00**	7.20
5CF RGB-N-D	89.40	84.30	85.84	89.40	60.62	86.00	7.20	6.80
LFC RGB-N	90.67	83.31	86.19	90.30	58.82	85.94	7.50	6.56
LFC RGB-D	90.21	79.14	83.46	88.67	57.73	84.04	9.40	6.55
LFC RGB-E	**90.92**	85.75	**87.03**	90.50	59.44	**86.90**	**7.00**	**5.76**
LFC NRG-D	90.34	80.64	84.81	89.08	56.60	84.77	7.58	7.65

Robustness Analysis We collected an additional dataset in a previously unseen place in low lighting, extreme shadows and snow. Fig. 2 shows some qualitative results from this subset. It can be seen that each of the spectra performs well in different conditions. Segmentation using RGB images shows remarkable detail, although being easily susceptible to lighting changes. NIR images on the other hand show robustness to lighting changes but often show false positives between the sky and trail classes. EVI images are good at detecting vegetation but show a large amount of false positives for the sky.

5 Conclusions and Scheduled Experiments

Our best performing late-fused-convolution approach currently only has one convolution layer after the fusion, adding a pooling and up-convolution layer to it would provide more invariance and discriminability to the filters learned after the fusion layer. Recently, adaptive fusion approaches have achieved state-of-the-art performance for fusing multiple modalities for detection tasks, however they have not been explored in the

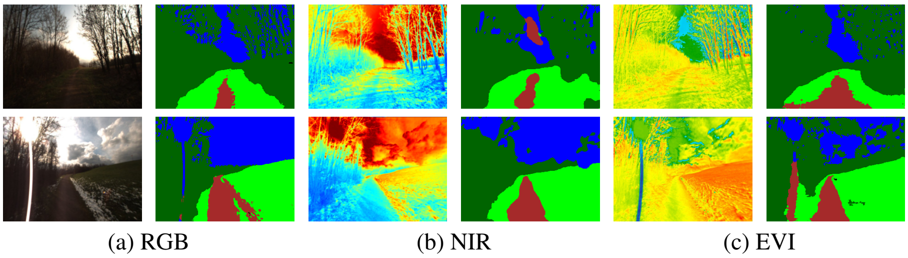


Fig. 2. Segmented examples from our benchmark. Each spectrum provides valuable information. The first row shows the image and the corresponding segmentation in highly shadowed areas. The second row shows the performance in the presence of glare and snow.

context of semantic segmentation. It would be interesting to evaluate the performance of such architectures in comparison to channel fusion and late-fused-convolution.

While evaluating our benchmark, we realized the need to add more testing images containing extreme conditions such as severely shadowed areas and glare, which would better highlight the benefits of using different spectra and modalities. Our benchmark also contains a subset of images with heavy snow and rain, evaluating the performance in such conditions will be insightful. Nevertheless, we believe that the results support our initial hypothesis of fusing the NIR wavelength with RGB to obtain a more accurate segmentation in forested environments. In conclusion, to the best of our knowledge this is the first benchmark that uses multispectral and multimodal data for semantic segmentation. We believe that the results demonstrate the benefits of fusing multiple spectrums and modalities to achieve robust segmentation in real-world environments.

References

1. D.M. Bradley, S.M. Thayer, A. Stenz, and P. Rander, "Vegetation Detection for Mobile Robot Navigation", Tech Report CMU-RI-TR-05-12, Carnegie Mellon University, 2004.
2. M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis" Comm. of the ACM, Vol 24, pp 381-395, 1981.
3. H. Hirschmüller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information", IEEE CVPR, 2005.
4. A. Huete, C.O. Justice, and W.J D. van Leeuwen, MODIS Vegetation Index (MOD 13), Algorithm Theoretical Basis Document (ATBD), Version 3.0, pp. 129, 1999.
5. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, et. al, "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv preprint arXiv:1408.5093, 2014.
6. F. Liu, C. Shen and G. Lin, "Deep Convolutional Neural Fields for Depth Estimation from a Single Image", arXiv:1411.6387, 2014.
7. W. Liu et. al, "ParseNet: Looking Wider to See Better", preprint arXiv:1506.04579, 2015.
8. J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Nov 2015.
9. D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, Vol 60, Issue 2, pp 91-110, Nov 2004.
10. V. Badrinarayanan et. al, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", arXiv preprint arXiv:1511.00561, 2015.
11. G.L. Oliveira et. al, "Deep Learning for Human Part Discovery in Images", ICRA, 2016.
12. K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv:1409.1556, 2014.
13. S. Thrun, M. Montemerlo1, H. Dahlkamp, et. al, "Stanley: The robot that won the DARPA Grand Challenge", Journal of Field Robotics, Vol 23, Issue 9, pp 661-692, 2006.