2014 IEEE 17th International Conference on
Intelligent Transportation Systems (ITSC)
October 8-11, 2014. Qingdao, China

CRF-based semantic labeling in miniaturized road scenes (Extended Abstract)

Mario Passani, J. Javier Yebes and Luis M. Bergasa

Abstract— **This paper presents an approach for the automatic pixelwise labeling of road scenes using a Probabilistic Graphical Model (PGM). The learning stage is based upon Conditional Random Fields (CRFs) and the inference of the semantic classes is relies on Tree-Reweighted Belief Propagation (TRW). The employment of miniaturized images based on superpixels is proposed and validated to achieve real time classification, which is of interest for the integration of scene understanding capabilities into ADAS and autonomous vehicles. The evaluation is carried out using the KITTI ROAD dataset achieving top results in roads with multiple marked lanes and it is publicly ranked among the state of the art.**

I. Introduction

During the last years, a big research effort has been made to design Advanced Driver Assistance Systems (ADAS) that rely on cameras as sensing technology, taking advantage of its low cost and ease of integration. For instance, lane departure warning has been recently extended to onboard smartphones [1]. In fact, lane estimation and road detection for assisting drivers are under active research [2], [3], [4]

In addition, providing scene understanding capabilities and extracting useful semantic information from images is of great interest for autonomous vehicles. This requires the employment of computer vision and machine learning advanced techniques, which allow to learn descriptive models and to infer semantic entities (e.g. scene regions, layout and objects) from a bunch of pixels. Recent advances in discrete optimization and Probabilistic Graphical Models (PGMs) have made Conditional Random Fields (CRFs) [5] a standard tool for segmenting and labeling tasks. Thus, there are a myriad of works based on CRF to perform image understanding in road enviroments, either to detect and classify single semantic entities, like lanes [6] or curbs [7], or to predict several semantic labels contained in the scene [8], [9]. However, such approaches require relevant hardware resources to be able to work in real time.

With the aim of providing a faster prediction of the semantic categories in the scene, we reimplement the inference stage and employ miniaturized images, i.e., at reduced resolution. Indeed, this reduction has been proved to be effective in loop closure detection of SLAM systems [10]. The idea behind our proposal is to shorten the space of hypotheses using superpixels, that is, representative pixels of patches in the image.

II. Model Learning and Inference

To perform image understanding, this work employs CRFs, which is a graphical model that represents the conditional probability $p(\mathbf{y}|\mathbf{x})$ of some vector $\mathbf{y}$ given an observation $\mathbf{x}$. We propose a grid-shaped model (see Fig. 1) to define the labels of the superpixels and their relationships. This graph is not tree-structured, thus, we propose to rely on a recent work [11], which presents a parameter learning based upon approximate marginal inference instead the usual approach based on approximations of the likelihood.

More especifically, the nodes of this undirected graph correspond to a lattice of pixels in a 4-neighborhood. At each node i we introduce an output random variable y_i, taking a value from a set of labels $\mathcal{L}$ corresponding to semantic classes of interest, and the observable random variable x_i representing the value of local features.

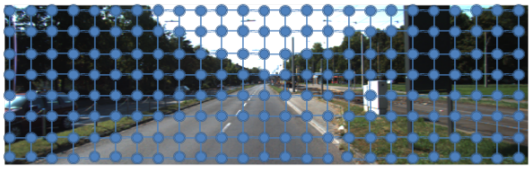

Fig. 1. Graph of a CRF aligned with the image's lattice of pixels.

Formally, the probability distribution can be factored in a product of unary (ψ_i) and pairwise (ψ_{ij}) potentials, also conditioned on the model parameters $\mathbf{w}$:

$$p(\mathbf{y}|\mathbf{x};\mathbf{w}) = \frac{1}{Z(\mathbf{x};\mathbf{w})}\prod_i \psi_i(y_i,\mathbf{x};\mathbf{w_i})\prod_{i,j}\psi_{i,j}(y_i,y_j,\mathbf{x};\mathbf{w_{ij}}) \tag{1}$$

The first product is over all individual pixels i, while the second one is over the edges in the graph which model neighboring pixels interactions. $Z(\mathbf{x};\mathbf{w})$ is known as the partition function for normalization purposes.

The optimal vector of parameters $\mathbf{w}^*$ is obtained by a supervised learning process that minimizes a loss function, in our case the marginal-based loss [11]. Once these parameters are computed from the training data, the model is applied for inferring pixelwise labels on test images. Particularly, this stage is based on Tree-Reweighted Belief Propagation (TRW) [12], truncated to a fixed and small number of iterations [11].

III. Appearance Features

For the visual description of the scenes, multiple features are incorporated into CRF model introduced above. The potentials in (1) can be defined as log-linear combinations of features extracted from the observed images $\mathbf{x}$:

**This work was supported in part by the MINECO project Smart Driving Applications (TEC2012-37104) and by the CAM project RoboCity2030 II-CM (S2009/DPI-1559).

The authors are with the Department of Electronics, UAH. Alcalá de Henares, Spain. e-mail: `mario.passani, javier.yebes, bergasa@depeca.uah.es`

978-1-4799-6078-1/14/$31.00 ©2014 European Union

$$\begin{cases} \psi_i = \exp(\mathbf{w_i}^T \cdot \mathbf{f}_i(\mathbf{x})) & \text{(2a)} \\ \psi_{i,j} = \exp(\mathbf{w_{ij}}^T \cdot \mathbf{g}_{ij}(\mathbf{x})) & \text{(2b)} \end{cases}$$

where $\mathbf{f}_i$ is a vector-valued function that defines the features of node i and similarly, $\mathbf{g}_{ij}$ is another vector that maps to pairwise features of nodes in the edge (i, j).

Node features. For each pixel in the lattice, several features are concatenated into $\mathbf{f}_i$. Firstly, we propose to incorporate color information of the scene employing HSV space due to its higher immunity to illumination changes. Then, we construct a set of feature functions from the intensities of each pixel i in the different channels.

In addition, an object position in an image is not arbitrary but maintain a certain order, for example, in the majority of driver assistance systems, the road occupies the center and bottom parts of the image. To exploit this, we define two features $\mathbf{f}_u$ and $\mathbf{f}_v$ taking the normalized position along the horizontal and vertical axes, respectively.

Moreover, since roads usually have a well-defined texture, we have opted for the inclusion of Local Binary Patterns (LBP) [13] in the extraction of road features. Besides, the local appearance around each node is captured with a Histogram of Oriented Gradients (HOG) [14] over a grid of non-overlapping 8×8 cells, with 9 orientation bins per cell, concatenating 4 cells to one block descriptor.

Edge features. The edge features consist of a concatenation of several descriptors. Firstly, the L2-norm of the difference of HSV intensities for the pixels i and j belonging to the same edge is computed. Then, a new set of ten edge features are obtained by discretizing the intensity values with a set of ten arbitrary thresholds [11]. Thus, the model complexity is augmented, the variance tends to increase and the squared bias tends to decrease in a bias-variance tradeoff.

In addition, we have empirically observed that incorporating a 'bias' feature set to one in the connected nodes, reduces the classification error. This extension captures any effects on the states of the random variables that are independent on the other features. Finally, the resultant 11D vector is doubled in size and arranged differently depending on whether the edges are vertical or horizontal. In the first case, the 11 features are kept on the first half, while the second half is filled with zeros and vice versa for horizontal links.

IV. Experiments and Conclusion

The effectiveness of our road inference approach is validated using the KITTI-ROAD dataset [3]. Particularly, we publicly rank our method (ARSL-AMI) by evaluating the performance over the test images at 15% of the original size. At this small resolution, inference is computationally efficient requiring less than 50 ms per image without a big loss in accuracy, as it is shown in Table I. Comparing our results with the reported accuracies for the other state-of-the-art works, we obtain similar values at much lower time costs, but we also rank second in *multiple marked lanes* (UMM_ROAD) evaluation. We note that our method does not require stereo vision nor 3D points. These results validate our superpixels hypothesis as the optimal way for segmenting complex images with a fair time processing without using specific hardware.

TABLE I

State of the art comparison in KITTI Urban Road [3]

Method	Setting	MaxF	Time	Environment
ProBoost	stereo	87.21%	2.5min	>8cores @3.0Ghz C/C++
SPRAY	mono	86.33%	45ms	NvidiaGTX580 Python+OpenCL
RES3D-Velo	laser	85.49%	0.36s	1core @2.5Ghz C/C++
ARSL-AMI	**mono**	**80.12%**	**0.05s**	**4cores @2.5Ghz C/C++**

Conclusion. We have shown that CRFs are one of the best alternatives in order to effectively address image labeling tasks as road detection. However, computational costs make them difficult candidates for integration into ADAS. Our work, upon reimplementation of the inference part, added to the employment of miniaturized images and easy-to-compute features facilitate complete real time integration.

References

[1] L. M. Bergasa, D. Almería, J. Almazán, J. J. Yebes, and R. Arroyo, "DriveSafe: an App for Alerting Inattentive Drivers and Scoring Driving Behaviors," in *IEEE Intelligent Vehicles Symposium (IV)*, Detroit, USA, June 2014, pp. 240–245.

[2] J. C. McCall and M. M. Trivedi, "Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation," *IEEE Trans. on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 20–37, March 2006.

[3] J. Fritsch, T. Kuehnl, and A. Geiger, "A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms," in *IEEE Intelligent Transportations Systems Conference (ITSC)*, 2013.

[4] G. B. Vitor, A. C. Victorino, and J. V. Ferreira, "A probabilistic distribution approach for the classification of urban roads in complex environments," in *ICRA'14 Workshop on Modelling, Estimation, Perception and Control of All Terrain Mobile Robots*, Hong-Kong, China, May 2014, pp. 1–6.

[5] J. D. Lafferty, A. McCallum, and F. C. N.Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *International Conference on Machine Learning (ICML)*, San Francisco, CA, USA, 2001, pp. 282–289.

[6] J. Hur, S.-N. Kang, and S.-W. Seo, "Multi-lane detection in urban driving environments using conditional random fields." in *IEEE Intelligent Vehicles Symposium (IV)*, 2013, pp. 1297–1302.

[7] J. Siegemund, D. Pfeiffer, U. Franke, and W. Förstner, "Curb reconstruction using Conditional Random Fields," in *IEEE Intelligent Vehicles Symposium (IV)*, 2010, pp. 203–210.

[8] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr, "Combining Appearance and Structure from Motion Features for Road Scene Understanding," *British Machine Vision Conf. (BMVC)*, 2009.

[9] V. Haltakov, H. Belzner, and S. Ilic, "Scene Understanding From a Moving Camera for Object Detection and Free Space Estimation," in *IEEE Intelligent Vehicles Symposium (IV)*, June 2012, pp. 105–110.

[10] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, J. J. Yebes, and S. Gámez, "Bidirectional Loop Closure Detection on Panoramas for Visual Navigation," in *IEEE Intelligent Vehicles Symposium (IV)*, Dearborn, USA, June 2014, pp. 1378–1383.

[11] J. Domke, "Learning Graphical Model Parameters with Approximate Marginal Inference," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 10, pp. 2454–2467, 2013.

[12] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-Reweighted Belief Propagation Algorithms and Approximate ML Estimation by Pseudo-Moment Matching," in *AISTATS*, 2003.

[13] T. Ojala and M. Pietikainen, "Unsupervised texture segmentation using feature distributions Pattern Recognition," in *Pattern Recognition*, 1996, pp. 477–486.

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2005, pp. 886–893.