A Dynamic Conditional Random Field Model for Joint Labeling of Object and Scene Classes

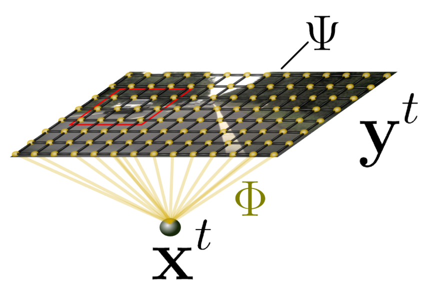
Christian Wojek, Bernt Schiele
Computer Science Department, TU Darmstadt, Germany
{wojek, schiele}@cs.tu-darmstadt.de

Objective

Pixel-wise labeling of object and scene classes in a Dynamic Conditional Random Field framework[1]

- Exploit powerful object detector in CRF framework to improve pixel-wise labeling of object classes
- Leverage temporal information
- Joint inference for objects and scene
- New Dataset with pixel-wise labels for highly dynamic scenes

Plain CRF formulation



$$\log(P_{pCRF}(\mathbf{y}^t|\mathbf{x}^t, N_1, \Theta)) = \sum_i \Phi(y_i^t, \mathbf{x}^t; \Theta_\Phi) + \sum_{(i,j)\in N_1} \Psi(y_i^t, y_j^t, \mathbf{x}^t; \Theta_\Psi) - \log(Z^t)$$

- Seven class labels: Sky, Road, Lane marking, Trees & bushes, Grass, Building, Void
- Joint boosting[2] to obtain unary potentials Softmax transform to obtain pseudo-probability:

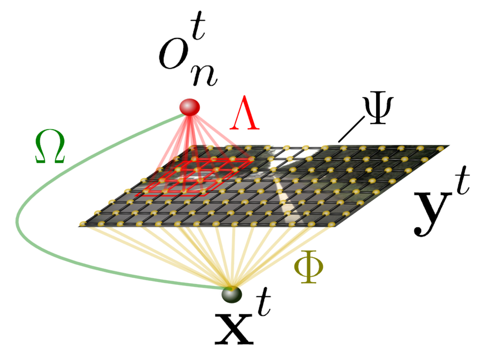
$$\Phi(y_i^t = k, \mathbf{x}^t; \Theta_\Phi) = \log \frac{\exp H(k, \mathbf{f}(x_i^t); \Theta_\Phi)}{\sum_c \exp H(c, \mathbf{f}(x_i^t); \Theta_\Phi)}$$

- Pairwise potentials with logistic classifiers (learnt with gradient descent) [3]

$$\Psi(y_i^t, y_j^t, \mathbf{x}^t; \Theta_\Psi) = \sum_{(k,l)\in C} \mathbf{w}^T \begin{pmatrix} 1 \\ \mathbf{d}_{ij}^t \end{pmatrix} \delta(y_i^t = k)\delta(y_j^t = l)$$

- Piecewise training of unary and pairwise potentials
- Distinguish east-west and north-south pairwise relations
- No dynamic information encoded
- Object classes suffer from too short range interactions

Object CRF formulation



$$\log(P_{oCRF}(\mathbf{y}^t, \mathbf{o}^t|\mathbf{x}^t, \Theta)) = \log(P_{pCRF}(\mathbf{y}^t|\mathbf{x}^t, N_2, \Theta)) + \sum_n \Omega(o_n^t, \mathbf{x}^t; \Theta_\Omega) + \sum_{(i,j,n)\in N_3} \Lambda(y_i^t, y_j^t, o_n^t, \mathbf{x}^t; \Theta_\Psi)$$

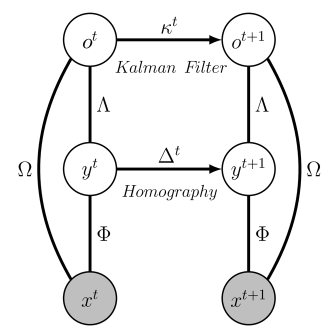
- Enrich plain CRF model with additional longer range dependency information for object classes
- Additional nodes for object hypotheses instantiated by object detector
  - Underlying pairwise cliques are extended by object node to form cliques of three
  - Object layout is learnt in discretized scale space

$$\Lambda(y_i^t, y_j^t, o_n^t, \mathbf{x}^t; \Theta_\Lambda) = \sum_{(k,l)\in C; m\in O} \mathbf{u}^T \begin{pmatrix} 1 \\ \mathbf{d}_{ij}^t \end{pmatrix} \delta(y_i^t = k)\delta(y_j^t = l)\delta(o_n^t = m)$$

- Platt's method to obtain pseudo-probability for unary object potential:

$$\Omega(o_n^t, \mathbf{x}^t; \Theta_\Omega) = \log \frac{1}{1 + \exp(s_1 \cdot (\mathbf{v}^T \cdot \mathbf{g}(\{\mathbf{x}^t\}_{o_n^t}) + b) + s_2)}$$

Dynamic CRF formulation



- Independently model scene and object motion
- Extended Kalman filter in 3D coordinate system for object classes

$$\log(P_{tCRF}(\mathbf{y}^t, \mathbf{o}^t|\mathbf{x}^t, \Theta)) = \log(P_{pCRF}(\mathbf{y}^t|\mathbf{x}^t, N_2, \Theta)) + \sum_n \kappa^t(o_n^t, \mathbf{o}^{t-1}, \mathbf{x}^t; \Theta_\kappa) + \sum_{(i,j,n)\in N_3} \Lambda(y_i^t, y_j^t, o_n^t, \mathbf{x}^t; \Theta_\Lambda)$$

- For scene classes propagate CRF posterior as prior to next time step

$$\Delta^t(y_i^t, \mathbf{y}^{t-1}; \Theta_{\Delta^t}) = \log(P_{tCRF}(y^{t-1}_{Q^{-1}(i)}|\Theta))$$

$$\log(P_{dCRF}(\mathbf{y}^t, \mathbf{o}^t, \mathbf{x}^t|\mathbf{y}^{t-1}, \mathbf{o}^{t-1}, \Theta)) = \log(P_{tCRF}(\mathbf{y}^t, \mathbf{o}^t|\mathbf{x}^t, \Theta)) + \sum_i \Delta^t(y_i^t, \mathbf{y}^{t-1}; \Theta_{\Delta^t})$$

Experiments on *TUD Dynamic Scenes* Dataset

- New dataset containing dynamic scenes
  - 176 sequences of 11 successive frames (88 sequences for training and 88 for testing)
  - Last frame of each sequence with pixel-wise labels, bounding box labels for object class *car*
- Publicly available from `http://www.mis.informatik.tu-darmstadt.de`
- Unary classification performance

		Normalization			
		on		off	
		multi-scale	single-scale	multi-scale	single-scale
Location	on	82.2%	81.1%	79.7%	79.7%
	off	69.1%	64.1%	62.3%	62.3%

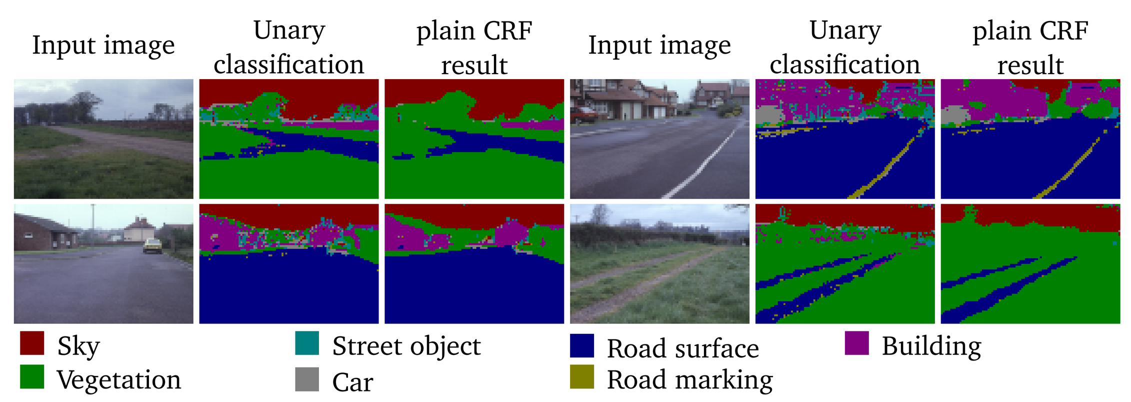
Features

- For unary and interaction potentials:
  - Gray world normalization of input images
  - Mean and Variance of 16 first Walsh-Hadamard transform coefficients from *CIE* L, a and b channel, extracted at multiple scales (8, 16 and 32 pixel windows)
  - Node coordinates in regular 2D lattice
- HOG features [4] for object node unary potentials

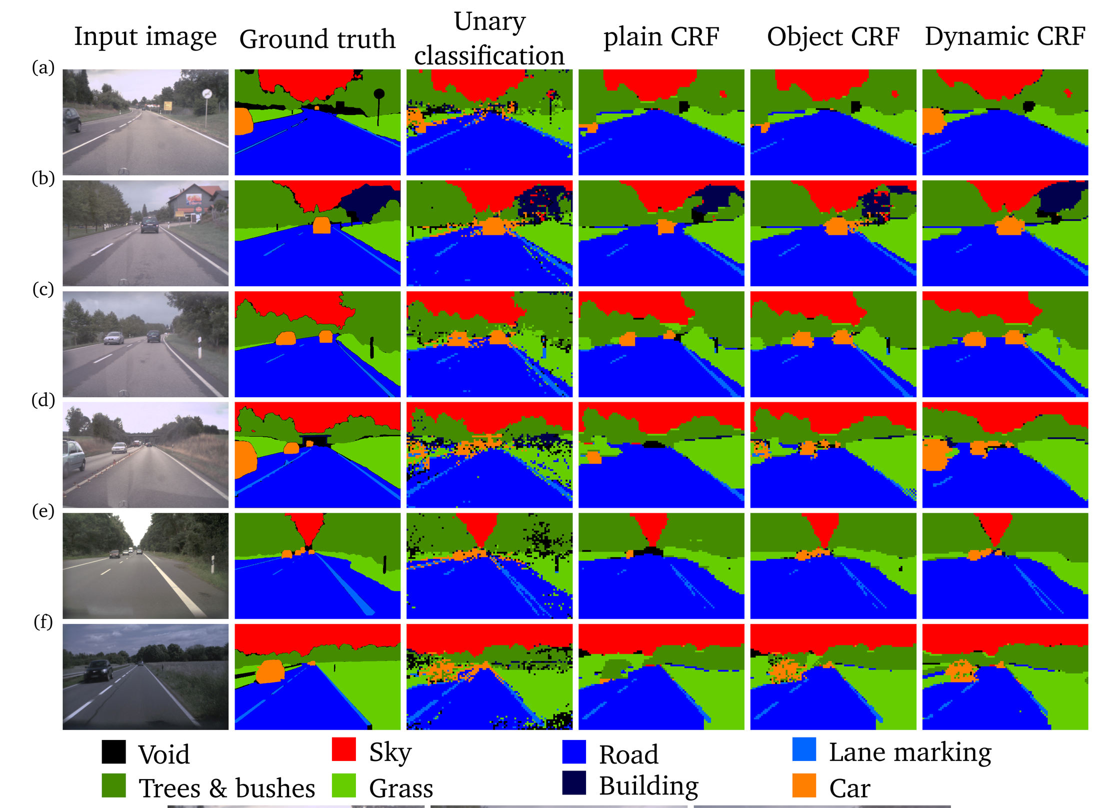
Experiments on *Sowerby* Dataset

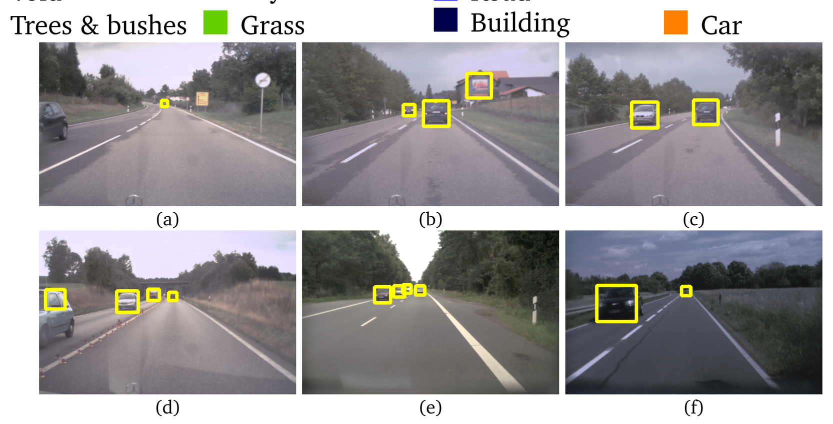
- Evaluation of plain CRF (only static images)

	Pixel-wise accuracy	
	Unary classification	plain CRF model
He *et al.* [5]	82.4%	89.5%
Kumar&Hebert [3]	85.4%	89.3%
Shotton *et al.* [6]	85.6%	88.6%
This paper	84.5%	91.1%



- Sample segmentations and detections





- Pixel-wise evaluation of object class *car*

	No objects			With object layer			Including object dynamics		
	Recall	Precision	Acc.	Recall	Precision	Acc.	Recall	Precision	Acc.
CRF	50.1%	57.7%	88.3%	62.9%	52.3%	88.6%	70.4%	57.8%	88.7%
dyn. CRF	25.5%	44.8%	86.5%	75.7%	50.8%	87.1%	78.0%	51.0%	88.1%

- Confusion matrix for all classes

True class	Fraction / Inferred	Sky	Road	Lane marking	Trees & bushes	Grass	Building	Void	Car
Sky	10.4%	**91.0**	0.0	0.0	7.7	0.5	0.4	0.3	0.1
Road	42.1%	0.0	**95.7**	1.0	0.3	1.1	0.1	0.5	1.3
Lane marking	1.9%	0.0	36.3	**56.4**	0.8	2.9	0.2	1.8	1.6
Trees & bushes	29.2%	1.5	0.2	0.0	**91.5**	5.0	0.2	1.1	0.4
Grass	12.1%	0.4	5.7	0.5	13.4	**75.3**	0.3	3.5	0.9
Building	0.3%	1.6	0.2	0.1	37.8	4.4	**48.4**	6.3	1.2
Void	2.7%	6.4	15.9	4.1	27.7	29.1	1.4	**10.6**	4.8
Car	1.3%	0.3	3.9	0.2	8.2	4.9	2.1	2.4	**78.0**

References

[1] Andrew McCallum, Khashayar Rohanimanesh, and Charles Sutton. Dynamic conditional random fields for jointly labeling multiple sequences. In *NIPS* Workshop on Syntax, Semantics*, 2003.

[2] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[3] Sanjiv Kumar and Martial Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] Xuming He, Richard S. Zemel, and Miguel Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[6] Jamie Shotton, John Winn, Carsten Rother, , and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

Acknowledgements: This work has been funded, in part, by Continental Teves AG. Further we thank Joris Mooij for publicly releasing libDAI and BAE Systems for the Sowerby Dataset.