Feature Space Optimization for Semantic Video Segmentation

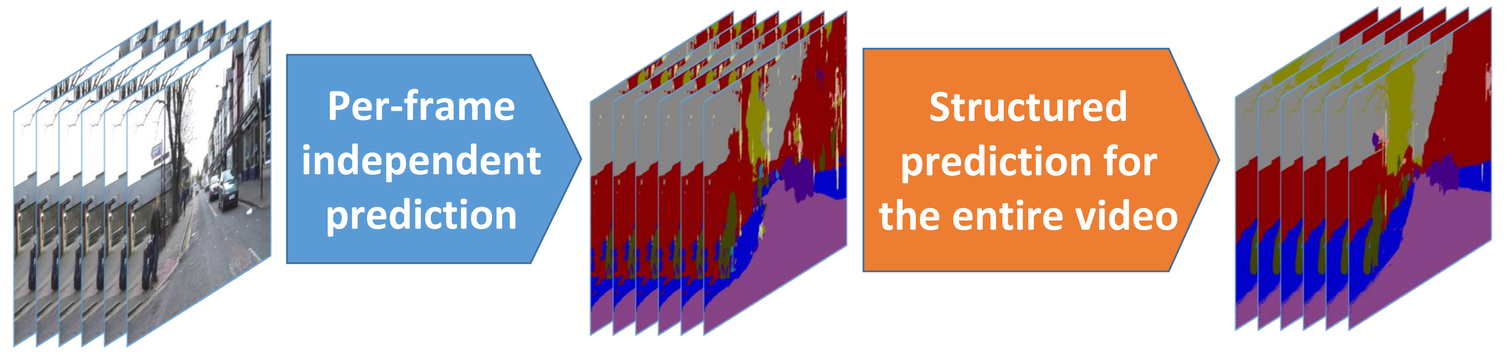
Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun

Georgia Tech College of Computing

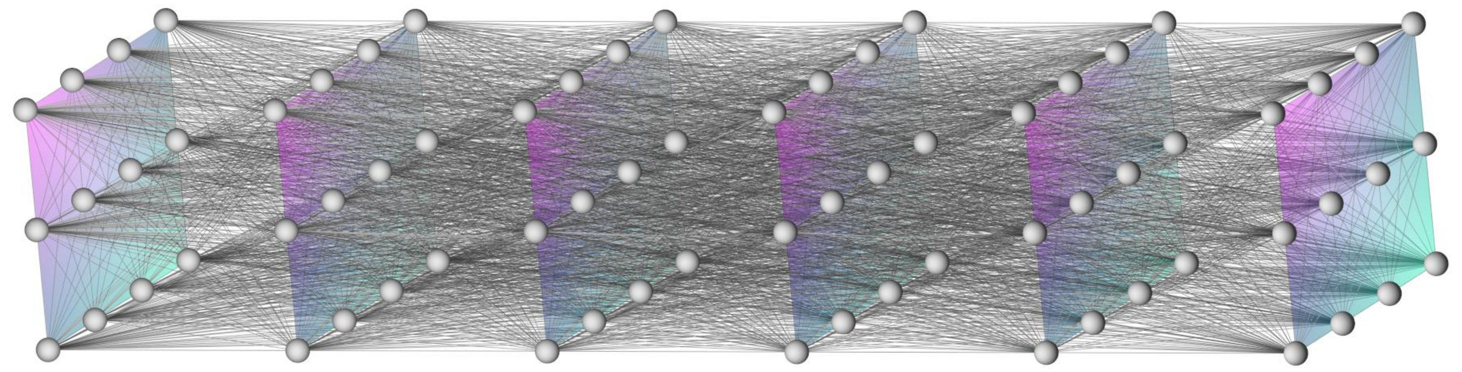
intel

Introduction

This work proposes an approach to structured prediction over video. It assigns semantic labels to all pixels in the video jointly.



Structured Prediction for Video



Our graphical model

$$E(x) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$

Unary terms: ConvNet, TextonBoost

Pairwise terms:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w^m \kappa^m(\boldsymbol{f}_i, \boldsymbol{f}_j)$$

Label compatibility function (e.g., Potts function)

Linear combination of Gaussian kernels

$$\kappa^m(\boldsymbol{f}_i, \boldsymbol{f}_j) = \exp\left(-\frac{\|\boldsymbol{f}_i - \boldsymbol{f}_j\|}{\sigma_m^2}\right)$$

$\boldsymbol{f}_i$ is some arbitrary feature space for i^{th} pixel.

Euclidean distance in feature space is used as weight for the smoothing.

Feature Space Optimization

Standard feature spaces for dense CRF:

Bilateral space: $[x, y, r, g, b]^T \in \mathbb{R}^5$

Extension to video: $[x, y, t, r, g, b]^T \in \mathbb{R}^6$

Standard feature spaces like those shown above have severe limitations in case of videos.

Desired properties of the feature space:

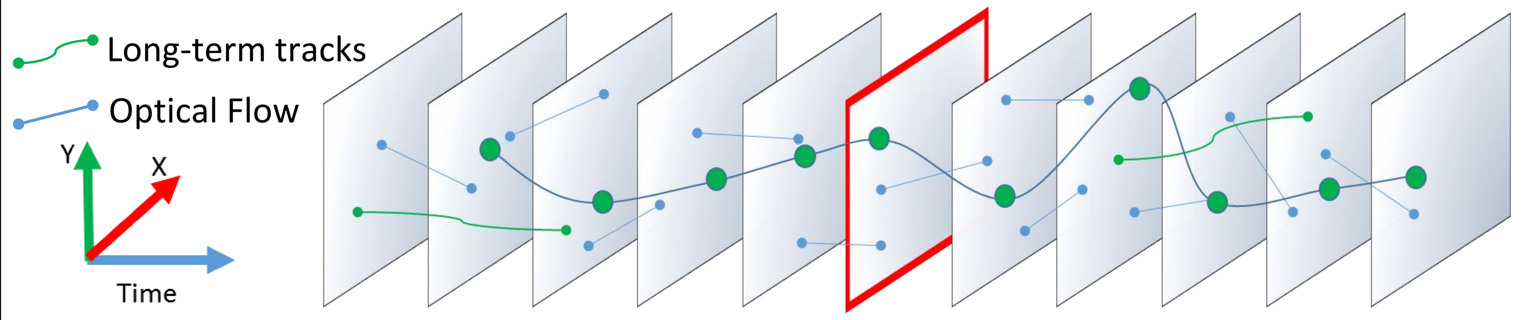
1. Corresponding pixels should map to points that are close in the feature space.
2. Features for pixels belonging to the same object should be closer than features of two pixels belonging to semantically distinct objects.
3. Needs to be low-dimensional.

Our solution:

Spatio-temporal regularization guided by optical flow and long-term trajectories.

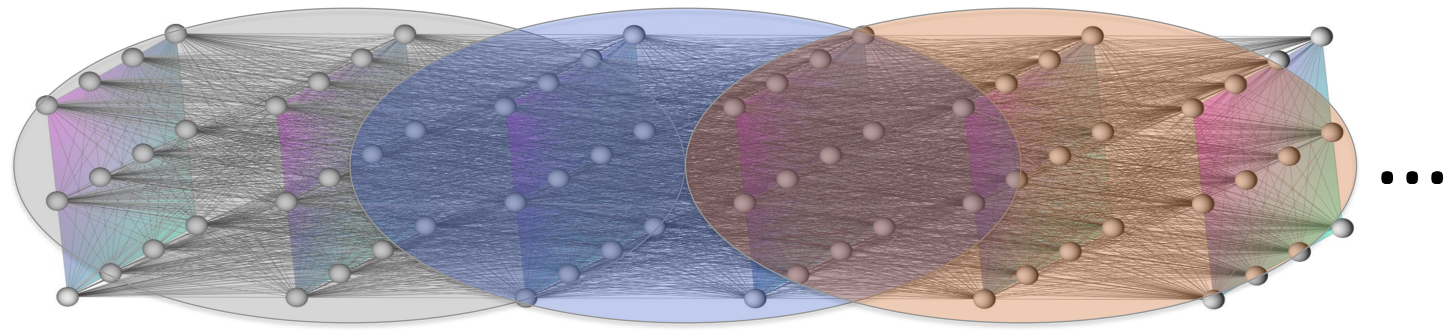
$$E(\boldsymbol{S}) = \underbrace{E_u(\boldsymbol{S})}_{\text{Data term}} + \gamma_1 \underbrace{E_s(\boldsymbol{S})}_{\text{Spatial smoothness}} + \gamma_2 \underbrace{E_t(\boldsymbol{S})}_{\text{Temporal smoothness}}$$

Optimized feature space $\boldsymbol{S}^* = \arg\min_{S} E(\boldsymbol{S})$



Large-scale Laplacian problem, can be solved using fast multi-grid solvers.

Scaling Up to Long Videos



Video sequences can be arbitrarily long.

Video is divided into overlapping blocks of frames.

Each block is a fully-connected CRF. All these blocks are solved jointly.

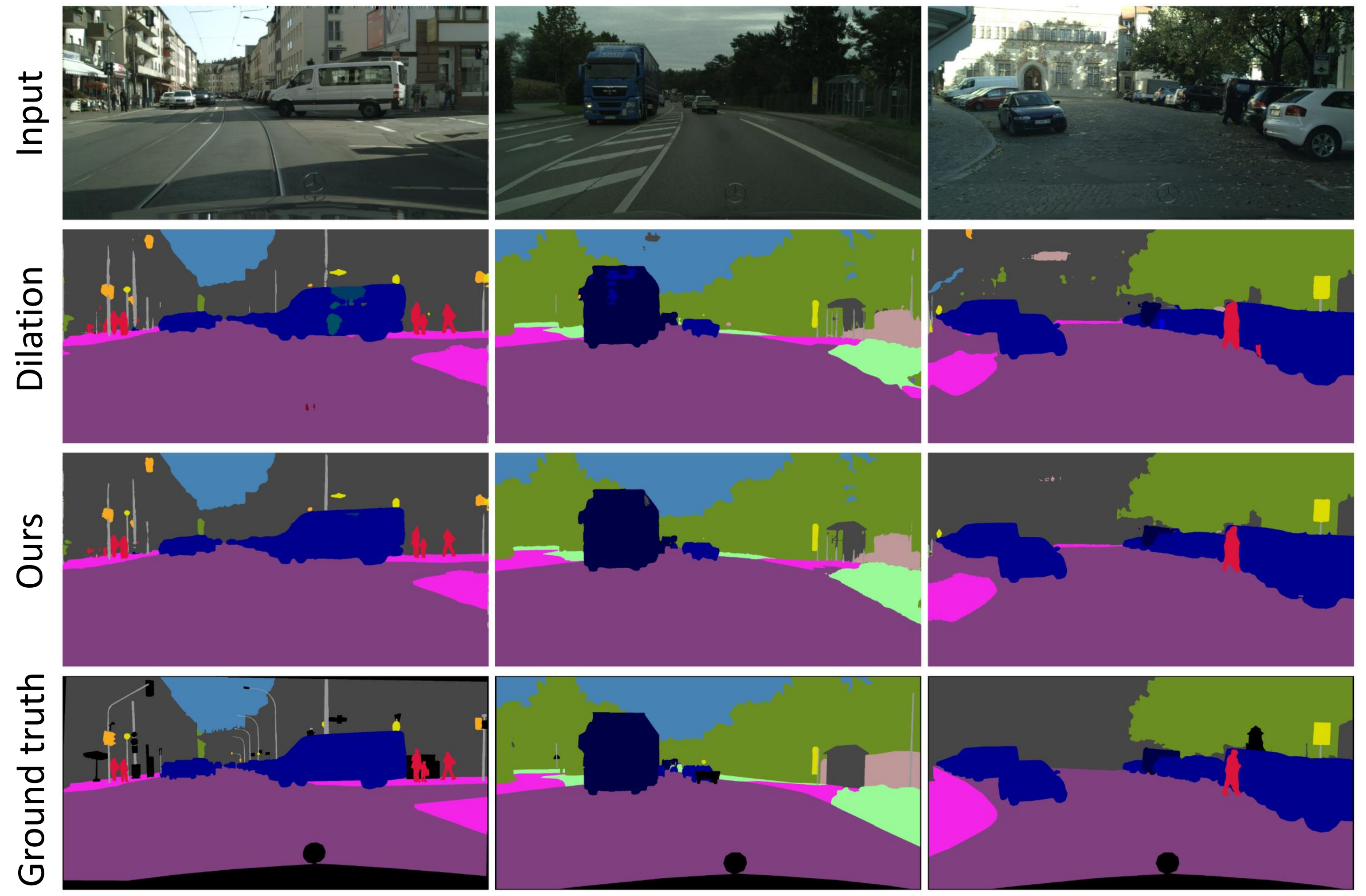
CamVid Evaluation

Ablation study	Mean IOU	Temporal consistency
TextonBoost unary	47.43	60.88
Dense 2D CRF	51.08	74.37
Dense 3D CRF	53.08	81.68
Our approach	55.23	87.33

Non-ConvNet unaries	Mean IOU	Temporal consistency
ALE (Ladicky et al. 2009)	53.59	72.2
Tighe & Lazebnik 2013	42.03	**88.8**
Tripathi et al. 2015	53.18	76.8
Liu & He 2015	47.2	77.6
TextonBoost + Our approach	**55.2**	87.3

ConvNet unaries	Mean IOU	Temporal consistency
SegNet Basic	46.4	62.5
SegNet Extended	55.6	-
Dilation (Yu & Koltun 2016)	65.29	79.0
Dilation + Our approach	**66.12**	**88.3**

Cityscapes Evaluation



Cityscapes validation	Mean IOU	Temporal consistency
Adelaide (Lin et al. 2016)	68.6	-
Dilation (Yu & Koltun 2016)	68.65	88.14
Dilation + Our approach	**70.30**	**94.71**

Conclusion

- A dense CRF model that optimizes over whole video sequences with billions of pixels.
- Exploits long-range context available in video to obtain more accurate and temporally consistent pixelwise labels.
- A low-dimensional feature space that captures correspondence information is vital for videos.
- Such a feature space can be obtained via optimization.

Code will be available soon