Lehrstuhl Informatik 8
Computergrafik und
Multimedia

Computer Vision
Group

Joint 2D-3D Temporally Consistent Semantic Segmentation of Street Scenes

Georgios Floros and Bastian Leibe

UMIC Research Centre, RWTH Aachen University

RWTH AACHEN UNIVERSITY

Abstract

We propose a novel Conditional Random Field (CRF) formulation for the semantic scene labeling problem which is able to enforce temporal consistency between consecutive video frames and take advantage of the 3D scene geometry to improve segmentation quality. The main contribution of this work lies in the novel use of a 3D scene reconstruction as a means to temporally couple the individual image segmentations, allowing information flow from 3D geometry to the 2D image space and vice versa. As our results show, the proposed framework outperforms state-of-the-art methods and opens a new perspective towards a tighter interplay of 2D and 3D information in the scene understanding problem. Quantitative and qualitative experiments show the effectiveness of the proposed method.

Main Idea

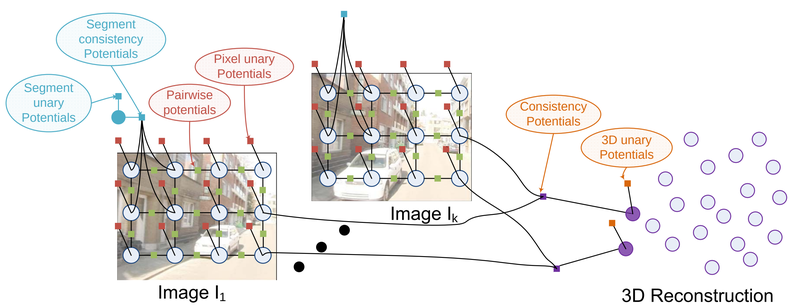
Motivation:

- Most previous approaches operate on single 2D images ignoring temporal continuity information.
- 3D information plays an important role for scene understanding.

We propose:

- Novel CRF formulation which forms temporal consistency constraints based on the underlying local 3D reconstruction.
- Principled way to incorporate local 3D geometry information into 2D semantic labeling algorithms.

CRF Model



Energy: $E(\mathbf{y}) = \sum_{i \in V} \psi_i(y_i) + \sum_{(i,j) \in E} \psi_{ij}(y_i, y_j) + \sum_{c \in \mathcal{S}} \psi_c(\mathbf{y}_c)$.

$\psi_i(y_i)$: *Unary potentials*, extended TextonBoost framework of [1].

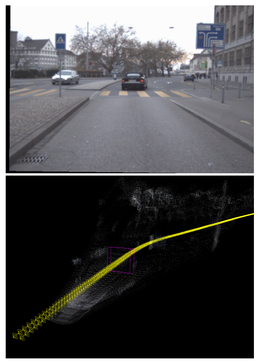
$\psi_{ij}(y_i, y_j)$: *Pairwise potentials*, contrast sensitive Potts model.

$\psi_c(y_c)$: *Higher order potentials*, robust P^N potentials, based on unsupervised mean-shift segmentations.

3D Reconstruction

3D reconstruction pipeline

- *Visual odometry:* Estimate the camera pose for each frame.
- *Stereo depth estimation:* Compute high quality depth maps efficiently for each frame.
- *Depth-map fusion:* Fuse the depth maps to a local 3D reconstruction. A zero-velocity Kalman filter keeps track of the localization uncertainty for each 3D point.



Temporal Consistency Potentials (TC)

- Take the form of **robust P^N potentials** [2].
- Force the pixels that correspond to the same 3D point to have the same labels.

$$\psi_p(y_p) = \min_{l \in \mathcal{L}} (\gamma_p^{max}, \gamma_p^l + k_p^l N_p^l) ,$$

$$\gamma_p^l = \lambda_s |p| \min(\sum_{i \in p} w_{ci} \psi_i(y_i) + K, \alpha),$$

$$\gamma_p^{max} = |p|(\lambda_p + \lambda_s \alpha),$$

α: Truncation threshold

K: Normalizing constant

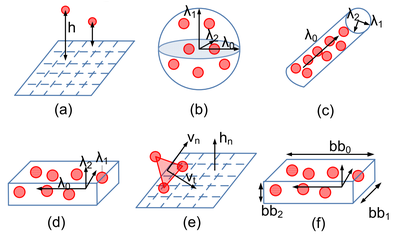
N_p^l: Number of pixels disagreeing with the majority vote

w_{ci}: Weights for the different frames' contributions to the clique

3D Features

Computed in a local neighborhood of the 3D point cloud.

(a) Height (h).
(b) *Point-ness* (λ_0).
(c) *Linear-ness* ($\lambda_0 - \lambda_1$).
(d) *Surface-ness* ($\lambda_1 - \lambda_2$).
(e) Cosines of the angles.
(f) Oriented bounding box.



3D Potentials

Bring information from the local 3D geometry.

$$\psi_p(y_p) = \min_{l \in \mathcal{L}} (\gamma_p^{max}, \gamma_p^l + k_p^l N_p^l),$$

$$\gamma_p^l = \lambda_s |p| \min \left(\sum_{i \in p} w_{ci} \psi_i(y_i) + w_{p_{3D}} (-log(H_l(p))) + K, \alpha \right).$$

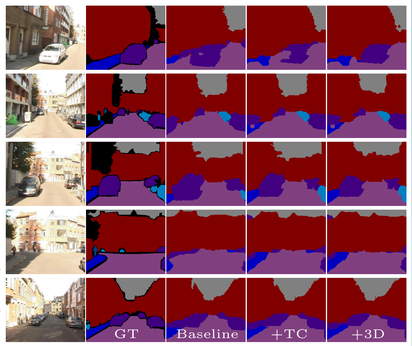
$H_l(p)$ response of a *Randomized Decision Forests* classifier

Experimental Results

Leuven:

- 1175 stereo image pairs at 25fps, 360x288 pixels.
- 70 labeled images (50 as training, 20 as testing), 7 class labels.

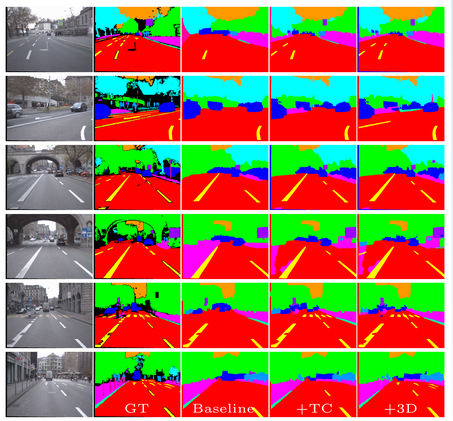
Recall	Baseline	+ TC	+ 3D	Ladicky [3]	Ours - 3D
Building	95.4	96.6	**97.8**	96.7	**97.9**
Sky	96.9	**98.7**	97.1	**99.8**	97.1
Car	84.7	**86.7**	85.4	**94.0**	85.4
Person	0.0	0.0	0.0	-	-
Road	96.3	**98.5**	98.4	**98.9**	98.4
Sidewalk	47.3	55.8	**56.1**	**60.6**	56.1
Bike	50.8	**59.6**	59.3	**59.5**	59.3
Global	93.0	94.8	**95.2**	**95.8**	95.4
Average	67.4	**70.8**	70.6	**84.9**	82.4
	7 classes			*6 classes*	



City:

- 3000 stereo image pairs at 13fps, 640x480 pixels.
- 103 labeled images (71 as training, 32 as testing), 13 class labels.

Recall	Baseline	+ TC	+ 3D	Ess [4]
Car	74.5	**78.2**	77.1	69.0
Road	97.6	**98.2**	97.8	93.0
Mark	54.1	67.3	**70.0**	73.0
Building	91.6	**92.6**	**92.6**	82.0
Sidewalk	59.6	**66.1**	65.8	41.0
Tree/Bush	82.5	**87.1**	86.9	62.0
Pole	4.2	2.8	**4.5**	**9.0**
Sign	**54.5**	50.2	50.9	17.0
Person	60.3	65.4	**66.1**	26.0
Wall	0.0	0.0	0.0	0.0
Sky	98.3	**98.6**	**98.6**	91.0
Curb	35.7	37.5	**38.3**	24.0
Grass	0.0	0.0	0.0	0.0
Global	89.5	91.1	**91.2**	-
Average	54.8	57.2	**57.6**	45.2



Conclusions

- Novel framework for enforcing semantic temporal consistency between consecutive frames.
- Incorporation of semantic information from **2D appearance** and **3D geometry** features in a single energy formulation.
- Improved segmentation performance on two stereo datasets.

References

[1] L. Ladickỳ, C. Russell, P. Kohli, P. Torr Associative hierarchical crfs for object class image segmentation In *ICCV*, 2009.

[2] P. Kohli, L. Ladickỳ, P.H.S. Torr Robust higher order potentials for enforcing label consistency In *IJCV*, 2009.

[3] L. Ladickỳ, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, P. Torr Joint optimization for object class segmentation and dense stereo reconstruction In *IJCV*, 2010.

[4] A. Ess, T. Mueller, H. Grabner, L. van Gool Segmentation-Based Urban Traffic Scene Understanding In *BMVC*, 2009.