Are Spatial and Global Constraints Really Necessary for Segmentation?

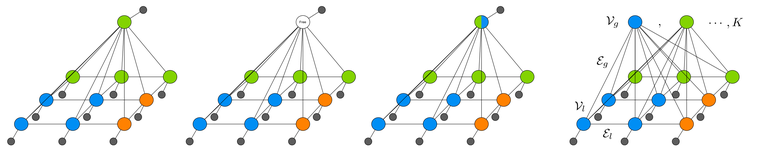
Aurélien Lucchi[1] Yunpeng Li[1] Xavier Boix[2] Kevin Smith[1] Pascal Fua[1]

[1] Computer Vision Laboratory, EPFL, Lausanne, Switzerland

[2] BIWI, ETH Zurich, Switzerland

Many state-of-the-art segmentation algorithms rely on Markov (MRF) or Conditional Random Field models (CRF) designed to enforce spatial and global consistency constraints. This is often accomplished by introducing additional latent variables to the model, which can greatly increase its complexity.
In a series of experiments on the PASCAL and the MSRC datasets, we were unable to find evidence of a significant performance increase attributable to the introduction of such constraints. On the contrary, we found that similar levels of performance can be achieved using a much simpler design that essentially ignores these constraints.

Overview of the studied CRF models



(a) Potts model (b) Robust P^N model [3] (c) Harmony model [1] (d) Class independent model (DPG)

Notations

- Set of nodes $\mathcal{V} = (\mathcal{V}_l, \mathcal{V}_g)$ where global nodes $\mathcal{V}_g$ encode high-level preferences and local nodes $\mathcal{V}_l$ represent superpixels[5].
- Set of edges $\mathcal{E} = (\mathcal{E}_l, \mathcal{E}_g)$ where $\mathcal{E}_l$ model the relationship between neighboring local nodes and $\mathcal{E}_g$ link local nodes to global nodes

$$E_w(Y|X) = \sum_{i \in \mathcal{V}_l} \underbrace{D_i(y_i)}_{\text{data term}} + \sum_{(i,j) \in \mathcal{E}_l} \underbrace{P_{ij}(y_i, y_j)}_{\text{pairwise term}} + \sum_{(i,g) \in \mathcal{E}_g} \underbrace{G_{ig}(y_i, y_g)}_{\text{global term}} \quad (1)$$

where $y_i \in \{1, \cdots, K\}$ are class labels for superpixels and $y_g \in \{0, 1\}$ represent the states of global preferences.

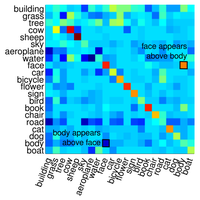
$$D_i(y_i) = \sum_{s=1}^{S} w^D_{y_i,s} c_s(x_i, y_i). \quad P_{ij}(y_i, y_j) = w^P_{y_i,y_j}. \quad G_{ig}(y_i, y_g) = w^G_{y_i,y_g}. \quad (2)$$

- The data term $D_i(y_i)$ encourages agreement between a node's label y_i and the local image evidence x_i. c_s is the output score returned by an SVM classifier.
- The pairwise term $P_{ij}(y_i, y_j)$ represents the cost of transition from class y_i to y_j, and is expressed in a non-parametric form.
- To enforce global consistency, a set of global nodes are introduced, resulting in a global term.

Parameter learning

- We express Eq. 1 in a linear form $E_w(X, Y) = w \cdot \Psi(X, Y)$ and we learn $w = \begin{bmatrix} w^D & w^P & w^G \end{bmatrix}$ using structured SVM.
- We also experimented with the learning method proposed by [1] that can only be applied to the simplest model with a few number of parameters.

The figure on the right shows the learned pairwise term $w^P(x_i, x_j, y_i, y_j)$ in matrix form where columns indicate classes y_i belonging to superpixel i and rows indicate classes y_j belonging to neighboring j for the MSRC-21 dataset. Red colors indicate that y_i is likely to appear above y_j.



Models Tested

- *D model* – Includes only the data term, consisting of the SVM classifiers scores. Equivalent to an energy function $E_w(Y|X) = \sum_{i \in \mathcal{V}_l} D_i(y_i)$.
- *DP model* – Considers both the data and pairwise terms of the energy function, i.e. $E_w(Y|X) = \sum_{i \in \mathcal{V}} D_i(y_i) + \sum_{(i,j) \in \mathcal{E}_l} P_{ij}(y_i, y_j)$.
- *DG model* – Considers the data term and the global term without the pairwise term, i.e. $E_w(Y|X) = \sum_{i \in \mathcal{V}_l} D_i(y_i) + \sum_{(i,g) \in \mathcal{E}_g} G_{ig}(y_i, y_g)$.
- *DPG model* – The full model described in Eq. 1, including the data, pairwise and global terms.
- *D-sampling* – Like the *D model*, only considers the data term. Use the sampling method of [1] to learn the parameters.

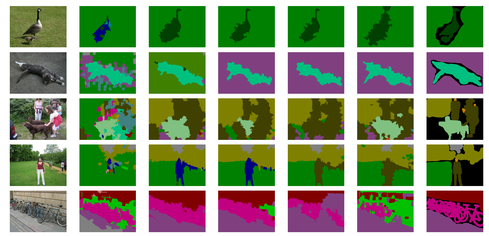
Features

- The following features are used to train $S = 6$ classifiers whose response c_s is weighted in the data term $D_i(y_i)$.
- ***Local features***: set of quantized visual words over five different **neighborhood scales**, which provides a histogram-like descriptor. Visual words are created by extracting SIFT and color histograms.
- ***Global features***: similar to the local features, but extracted **over the entire image**.

MSRC Results

- When only local features are considered, there is a clear advantage to adding spatial and global constraints (as indicated in red).
- When adding global features, this gain disappears, reflecting the **relative weakness of spatial and global constraints relative to the global features.**

		building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**Global**	**Average**
Local features S=5	*D model*	54	94	59	72	67	95	70	66	86	45	93	72	68	27	52	27	43	70	1	19	37	67	58
	DP model	53	86	72	79	75	95	93	49	85	38	81	83	64	39	63	49	68	68	38	63	9	70	64
	DG model	34	91	70	83	70	97	83	65	82	93	91	69	67	13	86	64	65	83	23	31	20	72	66
	DPG model	54	88	83	79	82	95	87	70	85	81	**97**	69	72	27	88	46	60	74	27	49	28	75	69
	D-sampling	52	83	74	50	72	89	86	68	73	69	83	67	69	22	68	25	67	54	14	46	50	69	61
Local+Global features S=6	*D model*	64	94	91	72	87	97	90	76	72	83	86	88	93	**62**	90	**89**	85	**97**	0	**83**	0	85	77
	DP model	58	87	83	73	78	94	95	78	85	68	96	89	71	41	96	83	85	87	49	52	38	80	76
	DG model	54	86	**93**	80	94	90	87	**88**	74	80	85	86	**96**	35	96	80	65	96	0	77	26	81	76
	DPG model	65	87	87	84	75	93	94	78	83	72	93	86	70	50	93	80	86	78	28	58	27	80	76
	D-sampling	50	83	87	81	84	90	**97**	72	75	79	90	95	79	52	**97**	81	80	89	51	64	**60**	79	**78**
	[6]	49	88	79	**97**	**97**	78	82	54	**87**	74	72	74	36	24	93	51	78	75	35	66	18	72	67
	[2]	53	**97**	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13	78	68
	[4]	**80**	96	86	74	87	**99**	74	87	86	**87**	82	**97**	95	30	86	31	**95**	51	**69**	66	09	**86**	75
	[1]	60	78	77	91	68	88	87	76	73	77	93	**97**	73	57	95	81	76	81	46	56	46	77	75

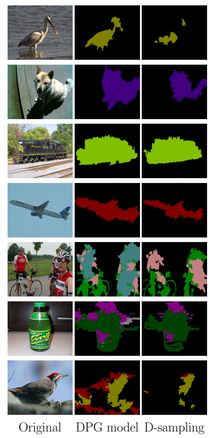


(a) (b) (c) (d) (e) (f) (g)

(a) Original images (b) *D model* with local features, (c) *DPG model* with local features, (d) *D model* with local+global features, (e) *DPG model* with local+global features, (f) *D-sampling* with local+global features, (g) Ground-truth.

VOC Results

		Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dinning Table	Dog	Horse	Motorbike	Person	Potted Plant	Sheep	Sofa	Train	TV/Monitor	**Average**
VALIDATION SET																							
Local features S=5	*D model*	31.7	9.7	1.1	9.1	0.9	3.0	27.4	12.9	14.4	0.0	21.2	0.0	0.0	3.7	26.8	22.9	0.0	30.0	9.6	23.1	0.0	11.8
	DP model	29.3	13.0	2.2	2.8	7.7	12.8	36.6	24.3	19.7	3.6	19.2	0.2	14.6	1.9	25.3	17.7	1.3	12.6	6.1	19.1	5.5	13.1
	DG model	76.2	30.8	6.8	1.6	0.0	4.4	34.1	0.7	23.5	0.4	28.3	0.3	2.6	6.5	46.2	11.4	0.0	0.0	2.2	39.9	0.1	15.0
	DPG model	67.3	22.4	15.5	18.0	15.4	0.0	28.4	24.5	18.7	0.0	30.5	4.6	0.5	6.0	28.9	23.0	3.6	0.0	12.9	33.1	7.3	17.2
	D-sampling	76.2	23.3	11.7	6.4	6.8	9.0	24.4	24.0	13.6	3.2	11.4	0.0	14.7	10.3	24.1	24.5	4.0	20.3	3.8	14.6	11.7	16.1
Local+Global features S=6	*D model*	73.0	38.9	13.3	21.1	25.7	12.2	37.7	32.7	29.0	0.4	35.5	11.5	9.5	18.8	30.1	1.8	6.5	36.1	8.1	43.7	19.9	24.1
	DP model	76.7	29.8	16.8	6.0	21.6	13.2	39.8	33.0	16.1	0.0	25.5	0.5	21.8	13.8	45.1	34.1	3.0	35.0	2.9	47.2	26.3	24.2
	DG model	74.1	27.8	0.2	22.9	20.8	16.7	32.5	29.1	25.4	7.1	30.8	14.9	15.7	16.0	45.5	29.3	5.7	32.3	17.2	42.0	0.0	24.1
	DPG model	64.3	27.1	18.2	23.1	21.0	15.0	35.3	29.2	23.7	7.8	16.9	21.5	17.3	18.8	30.7	31.5	6.7	27.8	12.2	39.5	26.7	24.5
	D-sampling	78.8	44.1	21.0	16.9	28.7	24.8	59.3	40.0	30.3	7.0	26.8	6.8	18.2	17.0	35.2	34.3	31.2	18.7	11.5	**47.3**	18.1	29.3
TEST SET																							
	BONN SVR	**84.2**	52.5	27.4	**32.3**	**34.5**	**47.4**	60.6	**54.8**	**42.6**	**9.0**	32.9	**25.2**	27.1	32.4	47.1	38.3	**36.8**	**50.3**	**21.9**	35.2	40.9	**39.7**
	BROOKES	70.1	31.0	18.8	19.5	23.9	31.3	53.5	45.3	24.4	8.2	31.0	16.4	15.8	27.3	48.1	31.1	31.0	27.5	19.8	34.8	26.4	30.3
	STANFORD	80.0	38.8	21.5	13.6	9.2	31.1	51.8	44.4	25.7	6.7	26.0	12.5	12.8	31.0	41.9	**44.4**	5.7	37.5	10.0	33.2	32.3	29.1
	UC3M	73.4	45.9	12.3	14.5	22.3	9.3	46.8	38.3	41.7	0.0	35.9	20.7	**34.1**	**34.8**	33.5	24.6	4.7	25.6	13.0	26.8	26.1	27.8
	UOCTTI	80.0	36.7	23.9	20.9	18.8	41.0	**62.7**	49.0	21.5	8.3	21.1	7.0	16.4	28.2	42.5	40.5	19.6	33.6	13.3	34.1	**48.5**	31.8
	Harmony FG-BG	80.2	**57.0**	**28.7**	29.3	31.7	27.0	57.6	48.5	35.2	8.3	29.9	22.6	25.2	33.0	52.6	35.9	25.2	39.7	16.9	43.4	24.7	35.8
	DPG model	64.8	33.4	16.6	17.8	23.4	17.2	45.7	35.0	30.3	6.0	21.5	21.0	21.9	29.6	32.6	29.6	23.3	24.9	15.7	26.4	21.1	26.6
	D-sampling	77.9	49.4	23.1	19.2	24.8	26.1	52.4	44.9	32.9	6.5	**35.8**	22.3	25.5	21.9	**58.1**	34.6	26.8	39.9	17.5	38.0	25.3	33.5



Original DPG model D-sampling

References

[1] X. Boix, J. Gonfaus, J. Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez. Harmony Potentials: Fusing Global and Local Scale for Semantic Image Segmentation. In *International Journal of Computer Vision*, 2011.

[2] J. Jiang and Z. Tu. Efficient Scale Space Auto-Context for Image Segmentation and Labeling. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] P. Kohli, M. Kumar, and P. Torr. P^3 and Beyond: Solving Energies With Higher Order Cliques. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[4] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. In *International Conference on Computer Vision*, 2009.

[5] A. Radhakrishna, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic Superpixels. Technical report, EPFL, June 2010.

[6] J. Shotton, M. Johnson, and P. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2008.