HIGHER ORDER POTENTIALS WITH SUPERPIXEL NEIGHBOURHOOD (HSN) FOR SEMANTIC IMAGE SEGMENTATION

Mostafa S. Ibrahim, Motaz El-Saban

Cairo Microsoft Innovation Lab
Microsoft Research
Cairo, Egypt

ABSTRACT

Among the approaches for solving the semantic image segmentation problem that has proven successful is in formulating an energy minimization expressed on top of a conditional random field (CRF) over image pixels. Recently, high order potentials (cliques of size greater than 2) over superpixels have been incorporated in the CRF energy function yielding promising results. These potentials encourage pixels within the same superpixel to take the same label by penalizing inconsistent labeling within the superpixel. While some of the earlier attempts modeled higher order potentials without considering the conditional dependencies between superpixels, others modeled these dependencies at the cost of oversimplified models at higher levels. In this paper, we propose incorporating superpixel neighborhood information within the high order potential, hence modeling dependencies between superpixels without the need of oversimplifying or constraining the model. Results show that the proposed method achieves state-of-the-art results on the challenging PASCAL VOC 2007 dataset.

Index Terms— Superpixels neighborhood, CRF, Object Class Image Segmentation, Higher Order Potentials.

1. INTRODUCTION

Semantic image segmentation is one of the most fundamental problems in computer vision. The goal of this problem is to assign a label for every pixel in an image out of L available labels. If such a goal can be achieved with enough accuracy, many image and video understanding applications can be enabled such as content-based image search and annotation. One of the approaches for solving the semantic image segmentation problem that has proven to be quite successful recently is in formulating the task as a discrete energy minimization over a conditional random field (CRF) [1- 4, 11, 12, 15]. CRFs allow to model shape, texture, color, and spatial relationships in a single constrained model [2] that can be efficiently minimized using graph-cut based move-making algorithms [16]. While earlier approaches were defined over pixels, with both unary and pairwise potentials [2, 9], recently energies have been defined over superpixels (pairwise CRF over segments) [11], or over both pixels and superpixels [1, 3, 12] (e.g. pairwise CRF over pixels and higher order potentials over segments). The main premise in working on the superpixel (or segment) level is that, by definition, a superpixel is a block of pixels that shares almost the same visual attributes, hence processing on the block unit is promising for both speed and segmentation quality. A standard approach in obtaining the initial superpixelization (segmentation) of images is to use unsupervised segmentation techniques such as [8, 10].

By working on the superpixel level, higher order potentials have been incorporated in the CRF function. While some approaches have not to considered the conditional dependencies between superpixels [3], hence ignored important information (such as the spatial continuity of predicted labels over neighboring superpixels), others [1] modeled these dependencies through hierarchal CRFs, but leading to oversimplified models at higher levels to maintain sub-modularity properties.

In this paper, we propose formulating the higher order potentials in terms of the superpixel neighborhood so that relationships and dependencies between superpixels are modeled but without oversimplifying or constraining the model (referred to as HSN method thereafter). The main technical contributions of this work are:

- Formulating higher order potentials in terms of likelihood of a superpixel to be assigned a certain label. This likelihood is formulated in terms of the superpixel neighborhood, hence encoding relationships between superpixels.
- Investigating the effect on segmentation accuracy if a single initial unsupervised segmentation is used as opposed to multiple initial ones.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work, section 3 explains the superpixel neighborhood framework, section 4 presents our approach and section 5 reports the experimental results. Finally section 6 draws some conclusions and highlights points of further research.

2. RELATED WORK

There are numerous approaches that have tackled the semantic image segmentation problem by minimizing an energy defined over a CRF. Broadly speaking, these methods can be classified in terms of approaches using pixel-level information and those using superpixels representations. As an example of pixel-level formulation, [2] jointly modeled the appearance, shape and context information through a CRF. For each pixel, they calculated the probability of belonging to each label through boosting weak classifiers based on set of shape filters, and augmented these probabilities as unaries in CRF energy function. An improvement over [2] was introduced in the semantic texton forests (STF) approach [13]. STFs are randomized decision forests used for both clustering and classification. In their approach, they also proposed an image level prior (ILP) that enhanced their results.

In the second category of approaches, methods relying on superpixels start by unsupervised segmentation into multiple partitions and augment the original CRF energy formulation using terms defined over segments. Some of these approaches work on one initial unsupervised segmentation of the image [11] , while others work on multiple set of overlapping segments (generated e.g. through multiple application of a segmentation technique) [1, 9, 14]. Approaches working on multiple segmentations [1, 3] advocate that no matter which method is used to partition the image into a set of non-overlapping segments, many of these segments are probably not correct in terms of aligning with object boundaries. Hence, working on multiple segmentations and in a principled manner finding the correct labeling for pixels is a promising framework.

An example work utilizing superpixels and higher-order potentials was proposed in the Robust P^N Model [3]. In their model, they robustly penalize inconsistent labeling within a segment, hence leading to a better segmentation. Unfortunately, they failed to model dependencies between segments, and as a result they missed important information between superpixels. For solving this problem, authors in [1] built associative hierarchical CRFs to allow associations in both same layer or between layers. They incorporated different higher order potentials for segments and super-segments to model segments' relationships. As indicated in [12], this model led to oversimplified representation at higher levels.

3. SUPERPIXELS NEIGHBORHOOD

Superpixels neighborhood [11] is a recent approach for object class image segmentation. In this approach, they partition each image into a set of non-overlapped superpixels by applying QuickShift [8]. They define a graph where nodes correspond to superpixels and two nodes are adjacent if they share at least an edge in the segmented image. Then, they represent each superpixel in terms of: (1) a histogram of extracted features for the superpixel (2) histograms of neighbors' superpixels where 2 superpixels are neighbors if the shortest distance in the superpixel graph between them is at most N edges. Each superpixel is assigned to the dominant label (label with highest frequency). According to [11], such representation contains spatial relationships of segments and some of their neighbors, as surrounding superpixels are either parts of the same object or some context. Finally, an SVM classifier is trained on the labeled histograms for each object category. To refine results, they use a CRF over superpixels. In our work, we incorporate the outputs of the SVM classifiers in the higher order potentials over segments in the CRF function, hence building implicit relationships among superpixels.

4. THE PROPOSED APPROACH (HSN)

Our approach starts by applying QuickShift [8], a recent method for unsupervised image segmentation. Based on experimental trials, we use six different initial segmentations by varying three parameters: λ, the trade-off between color and spatial importance, σ, density estimation scale and τ, the maximum distance in the feature space between members of the same region. For each one of the segmentation, we input the segments to the superpixel neighborhood method [11]. The outputs of the SVM classifiers are then utilized to compute the likelihood of a superpixel being assigned to each one of the labels. It is worth noting that each segmentation is obtained independently from other ones. Next, we will show how to utilize this likelihood in the CRF energy function.

We model a discrete random field X over pixels set $V = \{1, 2, 3 \ldots M\}$ with a neighborhood system $\mathcal{E}$ ($\mathcal{E}$ in our work is the set of edges in 4-connected grid). Each random variable X_i corresponds to a pixel $i \in V$ and can be assigned a label x_i out of set $L = \{l_1, l_2 \ldots l_k\}$. A clique (superpixel) c is a set of random variables $\boldsymbol{X}_c$ that are conditionally dependent on each other. S is set of the cliques for an image I, resulting from multiple segmentations. We seek finding a labeling $\pi = L^M$ that minimizes the following CRF energy function.

$$E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) + \sum_{c \in S} \psi_i(\boldsymbol{x}_c) \qquad (1)$$

In this formulation: the first term is the unary potential for pixels, the second is the traditional contrast pairwise potential term [7] which encourages similar neighbors to take the same label based on RGB differences between pixels and finally the third is the clique potential term. The unary potential is calculated as the negative log likelihood of variable X_i taking the label x_i. The SVM classifier outputs give the likelihood of a superpixel c to be assigned to one of labels. We distribute[i] the likelihood of a superpixel to its constituting pixels as in equation (2).

$$\psi_i(x_i) = -\log \Pr(X_c|I) \ \forall i \in c \quad (2)$$

As the probabilities used in equation (2) are computed only over superpixels, these pixel unaries can be computed very efficiently with respect to space and time (e.g. we don't need to compute many integral images as in [2]) which makes such potentials more suited for practical applications.

An important point to consider here is that as *S* is a set of different segmentations of the input image, then a pixel may belong to more than one superpixel. To solve this issue, we use only one segmentation for unary values using parameters described in [11] to assign each pixel to only one superpixel.

Finally, for the higher order potentials, we propose using the segment potential function as the negative log likelihood of superpixel belonging to label x_i scaled by the number of pixels of this superpixel as described by equations (3) and (4).

$$\psi_c(x) = \begin{cases} 0 & if\ x_i = l_k,\ \forall i \in c \\ |c|^{\theta} U(c) & otherwise \end{cases} \quad (3)$$

$$U(c) = -\log Pr(X_c|I) \quad (4)$$

In equation (3), θ is a parameter we learned from the training set. As stated in [3], the potential in this form is rigid, as it only scores zero if all pixels have same label. If there is only a small number of pixels different from the dominant label the same penalty will be assigned. In order to render this potential robust and to minimize it using a graph cut based moving algorithm we adopt a similar method as in [3] where the penalty is partially assigned based on the number of pixels not having the dominant label.

5. EXPERIMENTS

We evaluated the proposed method on VOC 2007 [6] dataset which represents a challenging dataset. We partition the data into training, validation and test sets in the same manner as suggested in the VOC 2007 competition, We segmented images using QuickShift using six different parameters combinations, experimentally selected such that we have superpixels ranging from small to large superpixels. The selected combinations are shown in Table 1.

Table 1. Initial unsupervised segmentation parameters where μ is average number of pixels per superpixel

ID	λ	σ	τ	μ	N
A[u]	2	.5	8	150	2
B	4	.5	8	366	2
C	2	.5	20	1230	1
D	5	.6	10	591	1
E	10	.5	8	705	1
F	7	.4	10	1431	1

In our experiments, we picked different combinations of these six divisions to investigate the effect of having only one initial unsupervised segmentation as opposed to multiple ones (see Table 1). For the parameter θ in the higher order potential, we picked a subset of training images and experimented with different values to learn the best one (θ = 0.6 was selected in our experiments). Finally, for the superpixel neighborhood approach, the maximum allowed shortest distance between 2 neighbors (*N*) was set to 2 (as suggested in [11]) for small size segments and 1 for the larger ones. Experimental results obtained using the proposed method are shown in Table 2 in terms of pixel segmentation accuracies [6].

Comparison with other state-of-the-art approaches readily shows the superiority of incorporating superpixel neighborhood information in the higher order potential formulation. Results in Table 2 also confirm the hypothesis in [1, 3] that using multiple initial segmentations actually leads to a better performance (compare the row labeled HSN_ABCDEF using all multiple initial segmentations with other rows utilizing one initial segmentation such as HSN_A which uses the "A" set of segmentation parameters from Table 1). Finally, some visual segmentation results are shown in Figure 1 to appreciate subjective gains in the proposed method.

Figure 1: From left to right: Image, ground truth, [11]'s output and our segmentation. The three top rows show success segmentation. The last row shows success in correctly finding objects boundaries but failure in detecting correct label.

6. CONCLUSION

In this paper, we presented a novel method for encoding dependencies among superpixels in higher order potentials defined over superpixels in a CRF energy-based semantic

image segmentation framework. We also used pixel unary values directly based on superpixel potentials, hence rendering the segmentation method very effective in terms of speed and memory. Experimental results suggest the promise of the presented method when compared to other recently published methods over a challenging dataset. As for future work, more investigation is warranted regarding better initial unsupervised segmentations, and applying the presented concepts for other problems like detection and classification.

Table 2. ***VOC 2007 segmentation results***. We compare our results to (a) Brookes, one of the best segmentation entries in the VOC 2007 segmentation challenge, (b) [13] [iii] with and without the ILP (Image level prior) and (C) [11]. HSN_X (where X represents the combined segmentations) refers to variants of our proposed method (please refer to text for more details).

Name	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	Motorbike	person	pottedplant	sheep	sofa	train	tv/monitor	Average
Brookes	**78**	6	0	0	0	0	9	5	10	1	2	11	0	6	6	29	2	2	0	11	1	9
[14]	59	27	1	8	2	1	32	14	14	4	8	32	9	24	15	81	11	**26**	1	28	17	20
[13]	33	46	5	14	**11**	14	**34**	8	6	3	10	39	**40**	28	23	32	19	19	8	24	9	20
[13]+ILP	20	**66**	6	15	6	**15**	32	19	7	7	**13**	**44**	31	**44**	27	29	**35**	12	7	39	23	24
[11] [iv]	40	13	11	**16**	7	10	19	34	45	20	7	15	32	16	57	40	21	17	18	34	27	24
HSN_ABCDEF	56	6	13	**16**	4	6	17	42	**60**	24	8	10	24	18	**63**	**51**	15	20	15	**46**	23	**26**
HSN_ABCD	46	7	18	14	6	8	18	40	57	24	10	12	30	21	62	43	18	18	16	44	27	**26**
HSN_A	44	9	**23**	**16**	6	8	18	37	53	23	10	11	31	20	61	41	19	18	17	43	28	25
HSN_E	42	5	4	10	4	0	28	**48**	57	**28**	3	15	28	31	46	46	14	24	**29**	34	**31**	25

7. REFERENCES

[1] L. Ladicky, C. Russell, P. Kohli and P. Torr, "Associative hierarchical CRFs for object class image segmentation", in ICCV 2009.

[2] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context", in IJCV 2009.

[3] P. Kohli, L. Ladicky, and P. Torr, "Robust higher order potentials for enforcing label consistency", in CVPR 2008.

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/maxflow algorithms for energy minimization in vision. PAMI, 2004.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph based image segmentation, in IJCV 2004.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes (VOC) Challenge in IJCV 2010.

[7] Y. Boykov, M. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images, in ICCV 2001.

[8] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking, in ECCV 2008.

[9] J. Carreira and C. Sminchisescu. Constrained parametric min cuts for automatic object segmentation, in CVPR 2010.

[10] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, in PAMI 2002.

[11] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods, in ICCV 2009.

[12] J. Gonfaus, X. Boix, J. Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez, Harmony Potentials for Joint Classification and Segmentation, in CVPR 2010.

[13] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation, in CVPR 2008.

[14] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations, in ECCV 2008.

[15] J. Laferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in ICML 2001.

[16] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. PAMI, 2001.

[i] This is inspired by the work in [11].

[ii] These are the same parameters used in [11] with N=2

[iii] We don't compare our results to TKK, an entry in VOC 2007 segmentation as it was trained using a much larger dataset. We also don't compare our work to [13] with DLP (detection level prior) as it depends on TKK.

[iv] We couldn't produce [11]'s output using declared parameters with their published code. Since our algorithm is sensitive to [11]'s output, we compare our accuracy numbers to the numbers we could reproduce rather than the one published in [11] (average segmentation accuracy of 32)