WEAKLY SUPERVISED SEMANTIC SEGMENTATION WITH SUPERPIXEL EMBEDDING

Frank Z. Xing[1]*, Erik Cambria*[1]*, Win-Bin Huang*[2] *and Yang Xu*[2]

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2] Department of Information Management, Peking University, China

ABSTRACT

In this paper, we propose to use contexts of superpixels as a prior to improve semantic segmentation by the CRF framework. A graphical model is constructed on over-segmented images. Our main contribution is to take the concept of "superpixel embedding" into consideration, which is formalized as a potential item for optimizing the energy of the whole graph. We also introduce two ways of calculating this embedding potential. Experiments on several popular datasets, e.g., MRSC-21 and PASCAL VOC, illustrate that our approach enhances the performance of a previously proposed segmentation model without embedding. The accuracy results are comparable to some fully supervised methods.

Index Terms— Superpixel embedding, Image semantic parsing, Unary potential, Context feature

1. INTRODUCTION

We address the problem of semantic segmentation, which is a heavily studied, while challenging topic in computer vision research. The problem also lays the foundations for many high-level tasks, such as multimodal content analysis [1], scene recognition [2], vehicle tracking [3], image understanding [4], and more. One of the difficulties in the data collection stage of semantic segmentation is that pixel-level manually annotated training images are expensive and not always accessible. Therefore, weakly supervised methods that only employ labels presented at image level and other prior knowledge are gaining increasing attention.

A number of priors have been discussed to represent mid-level or high-level contextual cues. Co-occurrence relations and spatial layout relations introduced in [5] are effective implementations. Image Level Prior (ILP) [6] is proposed to re-weight the probability of presence of different semantic classes according to global features, such as semantic textons. Nowadays ILP becomes a popular prior in many weakly supervised semantic segmentation frameworks. Objectness is calculated from classes contained in sample windows to counter the effect of "background flooding", a phenomenon that background semantic classes tend to invade into the edge of foreground objects [7, 8].

These priors are integrated to graph-based models to add constraints for connections between superpixels [9]. While these priors have provided significant cues for segmentation, there is still much room for improvement. For instance, previous literature seldom discuss the semantic relations between classes, which is common in natural language processing [10], but overlooked in image processing. Some prior models the spatial layout relation of object classes as four parts division (upper, lower, side and center) [5], this could be extended by allowing rich information from a smaller scale of an image. Following from these former studies, we hope to look at the contextual cues from a finer scale and formalize a more balanced distributed prior.

In this paper, we attempt to improve the segmentation performance by employing an innovative prior, which is termed superpixel embedding. We report on two main contributions: 1) Application of this prior conducts to a better segmentation accuracy from the previous work. 2) This method can generate faster and more compressed contextual cues than some state-of-the-art methods, e.g., graphlet and manifold method described in [11].

The rest of the paper is organized as follows: Section 2 describes the overall framework of segmentation, explains the concept of embedding, provides two different ways of implementing and how we formulate this prior into an unary potential; Section 3 provides a schematic description of the hierarchical segmentation approach and label predicting process, inspired by [12]; experiments and comparison are reported in Section 4; finally, Section 5 summarizes the contributions of the paper.

2. SUPERPIXEL EMBEDDING

2.1. Method Overview

A synoptic illustration of our segmentation method is shown in Fig. 1. Firstly, images are segmented into superpixels with SLIC [13]. Then, a graph of superpixels is constructed and other features are extracted as priors. Next, we build connections between similar superpixels of the original image and the graph so that the labels can propagate. Finally, the embedding prior is employed to repeatedly adjust the labeling to complete the segmentation.

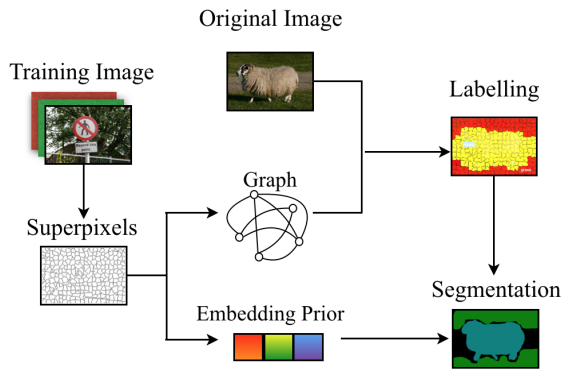

Fig. 1. Segmentation Method Overview.

2.2. Representation of superpixel context

We use the term "superpixel context" to denote the semantic classes and their spatial relations surrounding a given superpixel. It is obvious that the importance of a semantic class is inversely proportional to the distance from the given superpixel. It is not necessary to take into account superpixels in a long distance since the number of superpixels can be manipulated through tuning parameters of generation algorithm.

Therefore, we define the context of a certain superpixel as a sequence of semantic classes of its adjacent superpixels. The context sequence starts from y-axis direction and goes clockwise. If there is no superpixel in this direction, we denote it with a dummy class "boundary".

Unlike previous efforts made to investigate graph structure or edge weights, for instance in [14], we recognize the surrounding semantic classes as a variable feature to a certain superpixel.

We propose two versions of representation. The first records superpixel classes from eight main directions. For example, the context in Fig. 2 is $\mathbf{c}_8 = (6, 6, 6, 6, 5, 4, 4, 4)$. The centroid is defined as the arithmetic mean of each inner pixel coordinates, context determined by simple search. The second records both classes and their spanning angle, which indicates the context in Fig. 2 can be represented as a matrix.

$$\mathbf{c}_\infty = \begin{pmatrix} 6 & -0.083\pi & 0.375\pi \\ 5 & 0.375\pi & 0.625\pi \\ 4 & 0.625\pi & 0.917\pi \end{pmatrix}$$

2.3. Potential of embedding

To employ superpixel context in the viewpoint of semantic segmentation as an overall energy minimization problem, we propose a new potential, which gives each embedded superpixel a penalty cost value.

If x_i^j denotes the i-th superpixel in image j, y_i^j denotes the class of the i-th superpixel in image j, let $\mathcal{E}$ be the overall energy of the graph $G = (V_{x_i^j}, E_{x_i^{j(\prime)}})$, then in a classic CRF model:

$$\mathcal{E}(y_i^j) = \sum \left(\psi(y_i^j, x_i^j) + \pi(y_i^j)\right) + \sum \phi(y_i^j, x_i^j)^{(\prime)} \quad (1)$$

where the first unary potential ψ measures how much the feature of x_i^j is consistent with the class feature of y_i^j. Potential π constraints the class label in a possible pool in the training stage, and represents ILP in the test stage in [8]. Pairwise potential ϕ encourages superpixels with similar features to take the same label. Feature similarity is measured by distance of semantic texton histograms:

$$\phi(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}) = 1 - Distance(x_i^j, x_{i'}^{j'}), \forall y_i^j \neq y_{i'}^{j'} \quad (2)$$

It is difficult to optimize a high order potential, e.g., second order potential ϕ. Hence we propose the item η as an unary potential and embed adjacent contexts in each superpixel:

$$\eta(y_i, x_i) = -\log\left(P(y_i|\mathbf{c}_{x_i})\right) \quad (3)$$

where P is the probability of superpixel x_i taking label y_i condition on context $\mathbf{c}_{x_i}$. Therefore, η can be reckoned as a penalty for labeling of a superpixel that conflicts with its context. Parameter P is trained by a forward propagation neural network with the contextual data described in Section 2.2 and stored in the first output layer. Fig. 3 illustrates the training steps of a superpixel embedding network.

For test images, we initialize the inference with labelings generated by eq. (1) and iterate the process of predicting y_i^t for x_i^t until convergence of:

$$\mathcal{E}(y_i^j) = \sum_i \left(\eta(y_i^t, x_i^t)|\psi(\cdot), \pi(\cdot)\right) + \sum \phi(y_i^j, x_i^j)^{(\prime)} \quad (4)$$

where t stands for "training". Potential ψ and π are updated in each iteration. The energy is fast converging after 2 or 3 iterations according to our experiments.

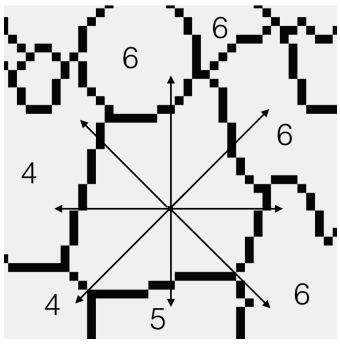

Fig. 2. Superpixel embedding representation of the central superpixel. Numbers denote different classes.

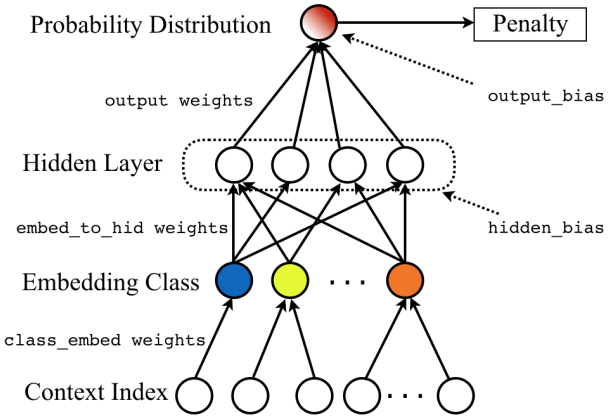

Fig. 3. The architecture of penalty training with superpixel embedding.

3. GRAPH-BASED SEGMENTATION

3.1. Construction of hierarchical superpixel graph

We empirically use SLIC [13] algorithm to generate superpixels of different scale from the same image. Scales are set as $z_s = \{20, 50, 100, 200\}$, as in [12]. Four scales interact with two mechanisms — transfer and recover. In top-down procedures, transfer mechanism assigns the semantic label of coarse superpixels to its finer counterparts in an upper scale. In bottom-up procedures, the fine scale superpixels vote for the label of its coarser counterparts in a lower scale. Recover mechanism rolls back to the previous labeling if the change increases the energy in each scale.

Superpixel features and image-level features are calculated for corresponding potential items, respectively. We use Semantic Texton Forests (STF) [6] as feature representation for both unary and pairwise potentials because of its outstanding performance. STF feature joints shape-texture, color and location cues as a whole. Therefore, no more features are required for superpixel level. Similarity of superpixels is defined as the χ^2 distance of STF features.

Furthermore, we use four image-level features: GIST [15], SIFT [16], HSOG (Histograms of the Second-Order Gradients) [17] and color histogram which is defined as $H(i) = \#\ (pixel\ value = i), i = 0, 1, ..., 255$ in one channel. We use a 512-dimensional GIST descriptor and color histogram distance is calculated as below.

$$d_{\chi^2}(H_1, H_2) = \sum_I \frac{(H_1(I) - H_2(I))^2}{H_1(I) + H_2(I)} \qquad (5)$$

The same learning method is employed as in [8] to form a linear combination of these image-level features. Thus different classes will have different combination of weights of each feature, which is believed beneficial for achieving a higher accuracy.

3.2. Semantic labels inference

In this section, we elaborate the algorithm of labels inference in the same scale. In the training stage, the probability of location distribution of each semantic class is learned as a prior. Labels with the highest probability is assigned to the superpixel. This initial mapping serves as the data for minimization.

In the test stage, the task is supervised with the aforementioned assigning result. For intra-scale labeling, eq. (4) is used in each procedure. Since the particular semantic class "void" has been taken into consideration in representation of superpixel context, the labeling result for eq. (3) includes some "void" class. In each iteration we transfer the corresponding labels from eq. (1) to substitute those "void" class. Optimizing uses graph-cut energy minimization algorithm introduced in [18].

4. EXPERIMENTS

4.1. Dataset and Analysis

We conduct our experiments on MSRC-21, which contains 591 images with 21 semantic classes. This is the most popular dataset for image semantic segmentation. PASCAL VOC 2007 and 2011 are employed for testing as well. We use the same training/testing split and parameter setting for graph construction as in [8] to make our results comparable to this method. The main difference is our use of additional priors, e.g., $\eta(y_i, x_i)$.

Since the probability of location distribution and superpixel embedding can be calculated before the label inference stage, our average time consumption for segmentation is not significantly different from the method described in [8].

Fig. 4 maps the relations of semantic classes to a lower dimensional space. Theoretically, closer classes tend to share more homogeneous surrounding context. Some semantic parallels are still preserved, for example, "road" is to "car" what "water" is to "boat".

4.2. Comparisons

We compare the proposed method with some state-of-the-art methods introduced in [8], [19] and [20]. Superpixel embedding is calculated using both context representations $\mathbf{c}_8$ and $\mathbf{c}_\infty$. Segmentation performance is evaluated by three indicators.

Node Accuracy (Node_Acc) is the most intuitive measure, which stands for the percent of superpixels correctly labeled in the testing stage. Pixel-level Accuracy (Pixel_Acc) represents the percent of pixels correctly labeled and Average per Class Accuracy (AClass_Acc) is the arithmetic mean for Pixel_Acc of 21 semantic classes. As shown in Table 1, our method achieves a circa 2% improvement on [8] for both Node_Acc and Pixel_Acc using $\mathbf{c}_8$ and 3% using $\mathbf{c}_\infty$.

Table 1. Performance comparison on MRSC-21

Method	AClass_Acc	Node_Acc	Pixel_Acc
MIM [8]	0.671	0.657	0.656
SIM[19]	0.697	–	–
Ours(using $\mathbf{c}_8$)	0.675	0.678	0.678
Ours(using $\mathbf{c}_\infty$)	**0.694**	**0.686**	**0.687**
WSDC [20]	0.714	0.529	–

A circa 2% improvement is made for AClass_Acc so that this indicator is comparable to some fully supervised method, e.g., Support Instance Machines in [19]. It is important to notice that our framework is the same with [8], which means the improvement is mainly the contribution of our additional embedding prior. This observation proves that $\eta(y_i, x_i)$ is a powerful potential. Some methods with different framework, for example, Dual Clustering [20], achieves a higher AClass_Acc, while other indicator, such as Node_Acc, is significantly lower.

Fig. 5 provides some segmentation results for comparison. We observe that the segmentation method with embedding potential usually produces a more smooth and well-shaped semantic class contour. Note that the class "sheep" in the second row has more accurate "feet" contour and the class "face" in the fourth row is more coherent. This character significantly improves the segmentation of object-centered classes like "sheep", "cow", "boat", etc. There are two type of errors that systematically appear in the results of segmentation with embedding. First, on the contrary to what is observed without embedding, object classes tend to flood into the background, though in most cases the background is just "void" class.

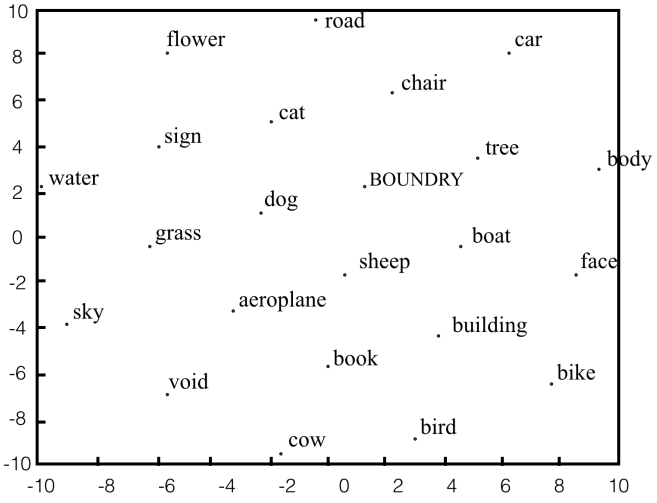

Fig. 4. A 2-dimentional mapping of semantic classes of MSRC-21 from higher dimensional space, including special classes "BOUNDRY" and "void".

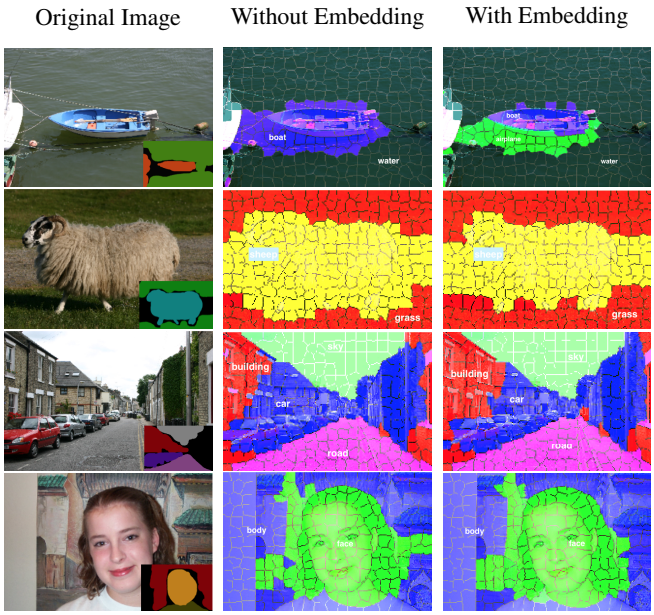

Fig. 5. Segmentation results on MSRC-21. The first column illustrates original images with ground truth laid on the right bottom corner. The second and the third columns illustrate segmentation results with and without superpixel embedding.

This phenomenon would not affect indicators because void class regions are excluded when calculating these accuracy, however, the de facto performance human perceived is seriously influenced. Second, regions tend to be labeled with a non-existent semantic class because of its shape and position. For example, the shadow below "boat" is incorrectly labeled as "airplane". This type of error, which we can term "semantic shift", is not only a counter-effect of the embedding, but also ubiquitous in human recognition, if attention only paid to a local part. To this end, as a future work we plan to exploit analogical reasoning [10] and common-sense reasoning [21] to filter out adjacent labels that are semantically irrelevant or implausible.

5. CONCLUSION

We present a better and competitive approach to represent mid-level contextual cues for images and semantic classes. A neural network is employed to transform these cues to the probability distribution of semantic class given the context. The distribution is used to form a unary potential that can be used in a graph-based model to improve semantic segmentation. Experimental results show that our superpixel embedding approach brings on a significant improvement from previous work. Further work will focus on finding faster and robust algorithm to train the prior. We also intend to reduce semantic shift through the ensemble use of common-sense knowledge and analogical reasoning.

6. REFERENCES

[1] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[2] A. Bassiouny and M. El-Saban, "Semantic segmentation as image representation for scene recognition," in *IEEE International Conference on Image Processing*, 2014, pp. 981–985.

[3] L. Liu, J. Xing, H. Ai, and S. Lao, "Semantic superpixel based vehicle tracking," in *Pattern Recognition, International Conference on*, 2012, pp. 2222–25.

[4] E. Cambria and A. Hussain, "Sentic album: Content-, concept-, and context-based online personal photo management system," *Cognitive Computation*, vol. 4, no. 4, pp. 477–496, 2012.

[5] C. Zhou and C. Liu, "Semantic image segmentation using low-level features and contextual cues," *Computers and Electrical Engineering*, vol. 40, pp. 844–857, 2014.

[6] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *ECCV*, 2008, pp. 1–8.

[7] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: label transfer via dense scene alignment," in *CVPR*, 2009, pp. 1972–79.

[8] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *ICCV*, 2011, vol. 23, pp. 643–650.

[9] S. Gould, J. Zhao, X. He, and Y. Zhang, "Superpixel graph label transfer with learned distance metric," in *ECCV*, 2014, pp. 632–647.

[10] E. Cambria, J. Fu, F. Bisio, and S. Poria, "AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis," in *AAAI*, Austin, 2015, pp. 508–514.

[11] L. Zhang, Y. Yang, Y. Gao, Y. Yu, C. Wang, and X. Li, "A probabilistic associative model for segmenting weakly supervised images," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4150–4159, 2014.

[12] S. Wang and Y. Wang, "Weakly supervised semantic segmentation with a multiscale model," *Signal Processing Letters*, vol. 22, pp. 308–312, 2015.

[13] R. Achanta, A. Shaji, K. Smith, and et al., "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[14] M. Manfredi, C. Grana, and R. Cucchiara, "Learning superpixel relations for supervised image segmentation," in *IEEE International Conference on Image Processing*, 2014, pp. 4437–41.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 143–175, 2001.

[16] D. G. Lowe, "Distinctive image features form scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] D. Huang, C. Zhu, Y. Wang, and L. Chen, "Hsog: A novel local image descriptor based on histograms of the second-order gradients," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4680–95, 2014.

[18] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 24, no. 2, pp. 147–159, 2004.

[19] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for miml instance annotation," in *ACM SIGKDD*, 2012, pp. 534–542.

[20] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly supervised dual clustering for image semantic segmentation," in *CVPR*, 2013, vol. 9, pp. 2075–82.

[21] E. Cambria and A. Hussain, *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*, Springer, Switzerland, 2015.